Special thanks to Snir Ben-Ovadia, Yair Daon, Yair Hartman, Dimitris Gatzouras, Ofer Shwartz and Andreas Strömbergsson for indicating typos and mistakes in earlier versions of this set of notes. If you find additional errors please let me know! O.S.
# Contents

1 Basic definitions and constructions ........................................... 1  
  1.1 What is ergodic theory and how it came about ......................... 1  
  1.2 The abstract setup of ergodic theory ................................. 3  
  1.3 The probabilistic point of view ....................................... 4  
  1.4 Ergodicity and mixing .................................................. 5  
  1.5 Examples ............................................................... 8  
      1.5.1 Circle rotations ............................................... 8  
      1.5.2 The angle doubling map ...................................... 9  
      1.5.3 Bernoulli Schemes ............................................ 9  
      1.5.4 Finite Markov Chains ...................................... 11  
      1.5.5 The geodesic flow on a hyperbolic surface ............... 17  
  1.6 Basic constructions ................................................... 19  
      1.6.1 Skew-products ................................................. 21  
      1.6.2 Factors ....................................................... 22  
      1.6.3 The natural extension ...................................... 23  
      1.6.4 Induced transformations ................................... 28  
      1.6.5 Suspensions and Kakutani skyscrapers ................. 30  
Problems ........................................................................ 31  
References ......................................................................... 33  

2 Ergodic Theorems .................................................................... 35  
  2.1 The Mean Ergodic Theorem .............................................. 35  
  2.2 The Pointwise Ergodic Theorem ...................................... 37  
  2.3 The non-ergodic case ..................................................... 39  
      2.3.1 Conditional expectations and the limit in the ergodic theorem 40  
      2.3.2 Conditional probabilities .................................... 41  
      2.3.3 The ergodic decomposition .................................. 43  
  2.4 The Subadditive Ergodic Theorem .................................. 44  
  2.5 The Multiplicative Ergodic Theorem ................................ 49  
      2.5.1 Preparations from Multilinear Algebra .................... 49  
      2.5.2 Proof of the Multiplicative Ergodic Theorem .......... 53
## Contents

2.5.3 The Multiplicative Ergodic Theorem for Invertible Cocycles 63

Problems ....................................................... 65
References .................................................... 68

3 Spectral Theory ............................................. 69
3.1 The spectral approach to ergodic theory ................. 69
3.2 Weak mixing ............................................. 71
   3.2.1 Definition and characterization ................... 71
   3.2.2 Spectral measures and weak mixing .............. 72
3.3 The Koopman operator of a Bernoulli scheme .......... 75
Problems ....................................................... 77
References .................................................... 81

4 Entropy ...................................................... 83
4.1 Information content and entropy ......................... 83
4.2 Properties of the entropy of a partition ............... 85
   4.2.1 The entropy of $\alpha \vee \beta$ .................... 85
   4.2.2 Convexity properties ............................. 86
   4.2.3 Information and independence ................... 87
4.3 The Metric Entropy ..................................... 87
   4.3.1 Definition and meaning .......................... 87
   4.3.2 The Shannon–McMillan–Breiman Theorem ........ 90
   4.3.3 Sinai’s Generator theorem ....................... 91
4.4 Examples ................................................ 93
   4.4.1 Bernoulli schemes ................................. 93
   4.4.2 Irrational rotations ............................... 93
   4.4.3 Markov measures ................................. 93
   4.4.4 Expanding Markov Maps of the Interval .......... 94
4.5 Abramov’s Formula ..................................... 95
4.6 Topological Entropy ................................... 97
   4.6.1 The Adler–Konheim–McAndrew definition ........ 97
   4.6.2 Bowen’s definition ............................... 100
   4.6.3 The variational principle ......................... 101
Problems ....................................................... 103
References .................................................... 104

A The Monotone Class Theorem ........................... 105

Index .......................................................... 107
Chapter 1
Basic definitions and constructions

1.1 What is ergodic theory and how it came about

**Dynamical systems and ergodic theory.** Ergodic theory is a part of the theory of dynamical systems. At its simplest form, a dynamical system is a function $T$ defined on a set $X$. The *iterates* of the map are defined by induction $T^0 := id$, $T^n := T \circ T^{n-1}$, and the aim of the theory is to describe the behavior of $T^n(x)$ as $n \to \infty$.

More generally one may consider the *action of a semi-group* of transformations, namely a family of maps $T_g : X \to X$ ($g \in G$) satisfying $T_{g_1} \circ T_{g_2} = T_{g_1g_2}$. In the particular case $G = \mathbb{R}^+$ or $G = \mathbb{R}$ we have a family of maps $T_t$ such that $T_{t_1} \circ T_{t_2} = T_{t_1+t_2}$, and we speak of a *semi-flow* or a *flow*.

The original motivation was classical mechanics. There $X$ is the set of all possible states of given dynamical system (sometimes called *configuration space* or *phase space*), and $T : X \to X$ is the *law of motion* which prescribes that if the system is at state $x$ now, then it will evolve to state $T(x)$ after one unit of time. The *orbit* $\{T^n(x)\}_{n \in \mathbb{Z}}$ is simply a record of the time evolution of the system, and the understanding the behavior of $T^n(x)$ as $n \to \infty$ is the same as understanding the behavior of the system at the far future. Flows $T_t$ arise when one insists on studying continuous, rather than discrete time. More complicated group actions, e.g. $\mathbb{Z}^d$–actions, arise in material science. There $x \in X$ codes the configuration of a $d$–dimensional lattice (e.g. a crystal), and $\{T_v : v \in \mathbb{Z}^d\}$ are the symmetries of the lattice.

The theory of dynamical systems splits into subfields which differ by the structure which one imposes on $X$ and $T$:

1. *Differentiable dynamics* deals with actions by differentiable maps on smooth manifolds;
2. *Topological dynamics* deals with actions of continuous maps on topological spaces, usually compact metric spaces;
3. *Ergodic theory* deals with measure preserving actions of measurable maps on a measure space, usually assumed to be finite.
It may seem strange to assume so little on $X$ and $T$. The discovery that such meagre assumptions yield non-trivial information is due to Poincaré, who should be considered the progenitor of the field.

**Poincaré’s Recurrence Theorem and the birth of ergodic theory.** Imagine a box filled with gas, made of $N$ identical molecules. Classical mechanics says that if we know the positions $q = (q_1^i, q_2^i, q_3^i)$ and momenta $p = (p_1^i, p_2^i, p_3^i)$ of the $i$-th molecule for all $i = 1, \ldots, N$, then we can determine the positions and momenta of each particle at time $t$ by solving Hamilton’s equations

\begin{align}
\dot{q}_1^i(t) &= -\partial H / \partial q_1^i \\
\dot{p}_1^i(t) &= \partial H / \partial p_1^i.
\end{align}

$H = H(q_1, \ldots, q_N; p_1, \ldots, p_N)$, the Hamiltonian, is the total energy of the system.

It is natural to call $(q, p) := (q_1, \ldots, q_N; p_1, \ldots, p_N)$ the state of the system. Let $X$ denote the collection of all possible states. If we assume (as we may) that the total energy is bounded above, then for many reasonable choices of $H$ this is a open bounded subset of $\mathbb{R}^{6N}$. Let

$$T_t : (q, p) \mapsto (q(t), p(t))$$

denote the map which gives solution of (1.1) with initial condition $(q(0), p(0))$. If $H$ is sufficiently regular, then (1.1) had a unique solution for all $t$ and every initial condition. The uniqueness of the solution implies that $T_t$ is a flow. The law of conservation of energy implies that $\mathcal{X} \in X \Rightarrow T_t(\mathcal{X}) \in X$ for all $t$.

**Question:** Suppose the system starts at a certain state $(q(0), p(0))$, will it eventually return to a state close to $(q(0), p(0))$?

For general $H$, the question seems intractable because (1.1) is strongly coupled system of an enormous number of equations ($N \sim 10^{24}$). Poincaré’s startling discovery is that the question is trivial, if viewed from the right perspective. To understand his solution, we need to recall a classical fact, known as **Liouville’s theorem:** The Lebesgue measure $m$ on $X$ satisfies $m(T_tE) = m(E)$ for all $t$ and all measurable $E \subset X$ (problem 1.1).

Here is Poincaré’s solution. Define $T := T_1$, and observe that $T^n = T_n$. Fix $\varepsilon > 0$ and consider the set $W$ of all states $\mathcal{X} = (q, p)$ such that $d(\mathcal{X}, T^n(\mathcal{X})) > \varepsilon$ for all $n \geq 1$ (here $d$ is the Euclidean distance). Divide $W$ into finitely many disjoint pieces $W_i$ of diameter less than $\varepsilon$.

For each fixed $i$, the sets $T^{-n}(W_i)$ ($n \geq 1$) are pairwise disjoint: Otherwise $T^{-n}(W_i) \cap T^{-(n+k)}(W_i) \neq \emptyset$, so $W_i \cap T^{-k}(W_i) \neq \emptyset$, and there exists $\mathcal{X} \in W_i \cap T^{-k}(W_i)$. But this leads to a contradiction:

1. $\mathcal{X} \in T^{-k}(W_i)$ implies that $T^k(\mathcal{X}) \in W_i$ whence $d(\mathcal{X}, T^k(\mathcal{X})) \leq \text{diam}(W_i) < \varepsilon$, whereas
2. $\mathcal{X} \in W_i \subset W$ implies that $d(\mathcal{X}, T^k(\mathcal{X})) > \varepsilon$ by the definition of $W$.

So $T^{-n}(W_i)$ ($n \geq 1$) are pairwise disjoint.
1.2 The abstract setup of ergodic theory

Since \( \{T^{-n}W_i\}_{n \geq 1} \) are pairwise disjoint, \( m(X) \geq \sum_{k \geq 1} m(T^{-k}W_i) \). But \( T^{-k}(W_i) \) all have the same measure (Liouville theorem), and \( m(X) < \infty \), so we must have \( m(W_i) = 0 \). Summing over \( i \) we get that \( m(W) = 0 \). In summary, a.e. \( x \) has the property that \( d(T^n(x), x) < \varepsilon \) for some \( n \geq 1 \). Considering the countable collection \( \varepsilon = 1/n \), we obtain the following result:

**Poincaré’s Recurrence Theorem:** For almost every \( x = (q(0), p(0)) \), if the system is at state \( x \) at time zero, then it will return arbitrarily close to this state infinitely many times in the arbitrarily far future.

Poincaré’s Recurrence Theorem is a tour de force, because it turns a problem which looks intractable to a triviality by simply looking at it from a different angle.

The Ergodic Hypothesis: For certain invariant measures \( \mu \), many functions \( f : X \rightarrow \mathbb{R} \), and many states \( x = (q, p) \), the time average \( \lim_{T \to \infty} \frac{1}{T} \int_0^T f(T_t(x)) dt \) exists, and equals the space average \( \frac{1}{\mu(X)} \int f d\mu \).

(This is not Boltzmann’s original formulation.) The ergodic hypothesis is a quantitative version of Poincaré’s recurrence theorem: If \( f \) is the indicator of the \( \varepsilon \)-ball around a state \( x \), then the time average of \( f \) is the frequency of times when \( T_t(x) \) is \( \varepsilon \)-away from \( x \), and the ergodic hypothesis is a statement on its value.

1.2 The abstract setup of ergodic theory

The proof of Poincaré’s Recurrence Theorem suggests the study of the following setup.

**Definition 1.1.** A **measure space** is a triplet \( (X, \mathcal{B}, \mu) \) where

1. \( X \) is a set, sometime called the **space**.
2. \( \mathcal{B} \) is a \( \sigma \)-**algebra**: a collection of subsets of \( X \) which contains the empty set, and which is closed under complements, countable unions and countable intersections. The elements of \( \mathcal{B} \) are called **measurable sets**.
3. \( \mu : \mathcal{B} \to [0, \infty] \), called the **measure**, is a \( \sigma \)-**additive function**: if \( E_1, E_2, \ldots \in \mathcal{B} \) are pairwise disjoint, then \( \mu(\bigcup_i E_i) = \sum_i \mu(E_i) \).

If \( \mu(X) = 1 \) then we say that \( \mu \) is a **probability measure** and \( (X, \mathcal{B}, \mu) \) is a **probability space**.

In order to avoid measure theoretic pathologies, we will always assume that \( (X, \mathcal{B}, \mu) \) is the completion (see problem 1.2) of a **standard measure space**: a mea-
sure space \((X, \mathcal{B}', \mu')\), where \(X\) is a complete and separable metric space and \(\mathcal{B}'\) is its Borel \(\sigma\)-algebra.

It can be shown that such spaces are Lebesgue spaces: They are isomorphic to the union of a compact interval equipped with the Lebesgue’s \(\sigma\)-algebra and Lebesgue’s measure, and a finite or countable or empty collection of points with positive measure. See the appendix for details.

**Definition 1.2.** A measure preserving transformation (mpt) is a quartet \((X, \mathcal{B}, \mu, T)\) where \((X, \mathcal{B}, \mu)\) is a measure space, and

1. \(T\) is measurable: \(E \in \mathcal{B} \Rightarrow T^{-1}E \in \mathcal{B}\);
2. \(m\) is \(T\)-invariant: \(\mu(T^{-1}E) = \mu(E)\) for all \(E \in \mathcal{B}\).

A probability preserving transformation (ppt) is a mpt on a probability space.

This is the minimal setup needed to prove (problem 1.3):

**Theorem 1.1 (Poincaré’s Recurrence Theorem).** Suppose \((X, \mathcal{B}, \mu, T)\) is a p.p.t. If \(E\) is a measurable set, then for almost every \(x \in E\) there is a sequence \(n_k \to \infty\) such that \(T^{n_k}(x) \in E\).

Poincaré’s theorem is not true for general infinite measure preserving transformations, as the example \(T(x) = x + 1\) on \(\mathbb{Z}\) demonstrates.

Having defined the objects of the theory, we proceed to declare when do we consider two objects to be isomorphic:

**Definition 1.3.** Two mpt’s \((X_i, \mathcal{B}_i, \mu_i, T_i)\) are called measure theoretically isomorphic, if there exists a measurable map \(\pi: X_1 \to X_2\) such that

1. there are \(X'_i \in \mathcal{B}_i\) such that \(m_i(X_i \setminus X'_i) = 0\) and such that \(\pi : X'_1 \to X'_2\) is invertible with measurable inverse;
2. for every \(E \in \mathcal{B}_2\), \(\pi^{-1}(E) \in \mathcal{B}_1\) and \(m_1(\pi^{-1}E) = m_2(E)\);
3. \(T_2 \circ \pi = \pi \circ T_1\) on \(X_1\).

One of the main aims of ergodic theorists is to develop tools for deciding whether two mpt’s are isomorphic.

**1.3 The probabilistic point of view.**

Much of the power and usefulness of ergodic theory is due to the following probabilistic interpretation of the abstract set up discussed above. Suppose \((X, \mathcal{B}, \mu, T)\) is a ppt.

1. We imagine \(X\) to be a sample space: the collection of all possible outcomes \(\omega\) of a random experiment.
2. We interpret \(\mathcal{B}\) as the collection of all measurable events: all sets \(E \subset X\) such that we have enough information to answer the question “is \(\omega \in E\)?”
3. We use \(\mu\) to define the probability law: \(\Pr[\omega \in E] := \mu(E)\).
4. Measurable functions \( f : X \to \mathbb{R} \) are random variables \( f(\omega) \);
5. The sequence \( X_n := f \circ T^n \) \((n \geq 1)\) is a stochastic process, whose distribution is given by the formula

\[
\Pr[X_i \in E_{i_1}, \ldots, X_k \in E_{i_k}] := \mu \left( \bigcap_{j=1}^k \{ \omega \in X : f(T^j \omega) \in E_{i_j} \} \right).
\]

The invariance of \( \mu \) guarantees that such stochastic processes are always stationary: \( \Pr[X_{i_1+m} \in E_{i_1+m}, \ldots, X_{i_k} \in E_{i_k+m}] = \Pr[X_{i_1} \in E_{i_1}, \ldots, X_{i_k} \in E_{i_k+m}] \) for all \( m \).

The point is that we can ask what are the properties of the stochastic processes \( \{f \circ T^n\}_{n \geq 1} \) arising out of the ppt \( (X, \mathcal{B}, \mu, T) \), and bring in tools and intuition from probability theory to the study of dynamical systems.

Note that we have found a way of studying stochastic phenomena in a context which is, a priori, completely deterministic (if we know the state of the system at time zero is \( x \), then we know with full certainty that its state at time \( n \) is \( T^n(x) \)). The modern treatment of the question “how come a deterministic system can behave randomly” is based on this idea.

### 1.4 Ergodicity and mixing

Suppose \((X, \mathcal{B}, \mu, T)\) is a mpt. A measurable set \( E \in \mathcal{B} \) is called invariant, if \( T^{-1}(E) = E \). Evidently, in this case \( T \) can be split into two measure preserving transformations \( T|_E : E \to E \) and \( T|_{E^c} : E^c \to E^c \), which can be analyzed separately.

**Definition 1.4.** A mpt \((X, \mathcal{B}, \mu, T)\) is called ergodic, if every invariant set \( E \) satisfies \( \mu(E) = 0 \) or \( \mu(X \setminus E) = 0 \). We say \( \mu \) is an ergodic measure.

**Proposition 1.1.** Suppose \((X, \mathcal{B}, \mu, T)\) is a mpt on a complete measure space, then the following are equivalent:

1. \( \mu \) is ergodic;
2. if \( E \in \mathcal{B} \) and \( \mu(T^{-1}E \Delta E) = 0 \), then \( \mu(E) = 0 \) or \( \mu(X \setminus E) = 0 \);
3. if \( f : X \to \mathbb{R} \) is measurable and \( f \circ T = f \text{ a.e.} \), then there is \( c \in \mathbb{R} \) s.t. \( f = c \text{ a.e.} \).

**Proof.** Suppose \( \mu \) is ergodic, and \( E \) is measurable s.t. \( \mu(E \Delta T^{-1}E) = 0 \). We construct a measurable set \( E_0 \) such that \( T^{-1}E_0 = E_0 \) and \( \mu(E_0 \Delta E) = 0 \). By ergodicity \( \mu(E_0) = 0 \) or \( \mu(X \setminus E_0) = 0 \). Since \( \mu(E \Delta E_0) = 0 \) implies that \( \mu(E) = \mu(E_0) \) and \( \mu(X \setminus E) = \mu(X \setminus E_0) \) we get that either \( \mu(E) = 0 \) or \( \mu(X \setminus E) = 0 \).

The set \( E_0 \) we use is \( E_0 := \{ x \in X : T^k(x) \in E \text{ infinitely often} \} \). It is obvious that this set is measurable and invariant. To estimate \( \mu(E_0 \Delta E) \) note that

- (a) if \( x \in E_0 \setminus E \), then there exists some \( k \) s.t. \( x \in T^{-k}(E) \setminus E \);
- (b) if \( x \in E \setminus E_0 \), then there exists some \( k \) s.t. \( x \notin T^{-k}(E) \), whence \( x \in E \setminus T^{-k}(E) \).
Thus \( E_0 \triangle E \subset \bigcup_{k \geq 1} \triangle E T^{-k}(E) \).

We now use the following “triangle inequality”:

\[
\mu(A_1 \triangle A_3) \leq \mu(A_1 \triangle A_2) + \mu(A_2 \triangle A_3) \quad (A_i \in \mathcal{B})
\]

(This is because \( \mu(A_i \triangle A_i) = \| 1_{A_i} - 1_{A_i} \|_1 \) where \( 1_{A_i} \) is the indicator function of \( A_i \), equal to one on \( A_i \) and to zero outside \( A_i \)). The triangle inequality implies that

\[
\mu(E_0 \triangle E) \leq \sum_{k=1}^{\infty} \mu(E \triangle T^{-k}E) \leq \sum_{k=1}^{\infty} \sum_{i=0}^{k-1} \mu(T^{-i}E \triangle T^{-(i+1)}E) = \sum_{k=1}^{\infty} \mu(E \triangle T^{-1}E) (\because \mu \circ T^{-1} = \mu).
\]

Since \( \mu(E \triangle T^{-1}E) = 0 \), \( \mu(E_0 \triangle E) = 0 \) and we have shown that (1) implies (2).

Next assume (2), and let \( f \) be a measurable function s.t. \( f \circ T = f \) almost everywhere. For every \( t \), \([f \geq t] \triangle T^{-1}[f > t] \subset [f \neq f \circ T] \), so

\[
\mu([f \geq t] \triangle T^{-1}[f > t]) = 0.
\]

By assumption, this implies that either \( \mu(f > t) = 0 \) or \( \mu(f < t) = 0 \). In other words, either \( f > t \) a.e., or \( f < t \) a.e. Define \( c := \sup\{t : f > t \text{ a.e.}\} \), then \( f = c \) almost everywhere, proving (3). The implication (3) \( \Rightarrow \) (2) is obvious: take \( f = 1_{E_t} \).

An immediate corollary is that ergodicity is an invariant of measure theoretic isomorphism: If two mpt are isomorphic, then the ergodicity of one implies the ergodicity of the other.

The next definition is motivated by the probabilistic notion of independence. Suppose \((X, \mathcal{B}, \mu)\) is a probability space. We think of elements of \( \mathcal{B} \) as “events”, we interpret measurable functions \( f : X \to \mathbb{R} \) as “random variables”, and we view \( \mu \) as a “probability law” \( \mu(E) = \mathbb{P}[x \in E] \). Two events \( E, F \in \mathcal{B} \) are called independent, if \( \mu(E \cap F) = \mu(E) \mu(F) \) (because in the case \( \mu(E) \), \( \mu(F) \neq 0 \) this is equivalent to saying that \( \mu(E|F) = \mu(E), \mu(F|E) = \mu(F) \)).

**Definition 1.5.** A probability preserving transformation \((X, \mathcal{B}, \mu, T)\) is called mixing (or strongly mixing), if for all \( E, F \in \mathcal{B} \),

\[
\mu(E \cap T^{-k}F) \rightarrow \mu(E) \mu(F).
\]

In other words, \( T^{-k}(F) \) is “asymptotically independent” of \( E \). It is easy to see that strong mixing is an invariant of measure theoretic isomorphism.

It can be shown that the sets \( E, F \) in the definition of mixing can be taken to be equal (problem 1.12).

**Proposition 1.2.** Strong mixing implies ergodicity.

**Proof.** Suppose \( E \) is invariant, then \( \mu(E) = \mu(E \cap T^{-n}E) \rightarrow \mu(E)^2 \), whence \( \mu(E)^2 = \mu(E) \). It follows that \( \mu(E) = 0 \) or \( \mu(E) = 1 = \mu(X) \). \(\square\)
1.4 Ergodicity and mixing

Just like ergodicity, strong mixing can be defined in terms of functions. Before we state the condition, we recall a relevant notion from statistics. The correlation coefficient of non-constant \( f, g \in L^2(\mu) \) is defined to be

\[
\rho(f, g) := \frac{\int fg d\mu - \int f d\mu \cdot \int g d\mu}{\|f - \int f d\mu\|_2 \|g - \int g d\mu\|_2}.
\]

The numerator is equal to

\[
\text{Cov}(f, g) := \int \left[ \left( f - \int f \right) \left( g - \int g \right) \right] d\mu,
\]

called the covariance of \( f, g \). This works as follows: If \( f, g \) are weakly correlated then they will not always deviate from their means in the same direction, leading to many cancelations in the integral, and a small net result. If \( f, g \) are strongly correlated, there will be less cancelations, and a larger absolute value for the net result. Positive covariance signifies that the deviations from the mean are often in the same direction, and negative covariance indicates that that they are often in opposite directions. The denominator in the definition of \( \rho \) is not important. It is there to force \( \rho(f, g) \) to have values in \([-1, 1]\) (Cauchy-Schwarz).

**Proposition 1.3.** A ppt \( (X, \mathcal{A}, \mu, T) \) is strongly mixing iff for every \( f, g \in L^2 \),

\[
\int fg o T^n d\mu \longrightarrow \|f\|_p \|g\|_p, \text{ equivalently } \text{Cov}(f, g o T^n) \longrightarrow 0.
\]

**Proof.** We need the following trivial observations:

1. Since \( \mu \circ T^{-1} = \mu \), \( \|f o T\|_p = \|f\|_p \) for all \( f \in L^p \) and \( 1 \leq p \leq \infty \);
2. \( \text{Cov}(f, g) \) is bilinear in \( f, g \);
3. \( |\text{Cov}(f, g)| \leq 4\|f - \int f \|_2 \|g - \int g\|_2 \).

The first two statements are left as an exercise. For the third we use the Cauchy-Schwarz inequality: \( |\text{Cov}(f, g)| \leq \|f - \int f\|_2 \|g - \int g\|_2 \leq (\|f\|_2 + \|f\|_1)(\|g\|_2 + \|g\|_1) \leq (2\|f\|_2)(2\|g\|_2) \) where \( \leq \) is because \( \|\phi\|_1 = \|\phi \cdot 1\|_1 \leq \|\phi\|_2 \|1\|_2 = \|\phi\|_2 \).

Now for the proof of the proposition. The condition that \( \text{Cov}(f, g o T^n) \longrightarrow 0 \)

for all \( f, g \in L^2 \) implies mixing by substituting \( f = 1_E \), \( g = 1_F \). For the other direction, assume that \( \mu \) is mixing, and let \( f, g \) be two elements of \( L^2 \). If \( f, g \) are indicators of measurable sets, then \( \text{Cov}(f, g o T^n) \rightarrow 0 \) by mixing. If \( f, g \) are finite linear combinations of indicators, \( \text{Cov}(f, g o T^n) \rightarrow 0 \) because of the bilinearity of the covariance. For general \( f, g \in L^2 \), we can find for every \( \varepsilon > 0 \) finite linear combinations of indicators \( f_\varepsilon, g_\varepsilon \) s.t. \( \|f - f_\varepsilon\|_2, \|g - g_\varepsilon\|_2 < \varepsilon \). By the observations above,

\[
|\text{Cov}(f, g o T^n)| \leq |\text{Cov}(f - f_\varepsilon, g o T^n)| + |\text{Cov}(f_\varepsilon, g_\varepsilon o T^n)| + |\text{Cov}(f_\varepsilon, (g - g_\varepsilon) o T^n)|
\]

\[
\leq 4\|f\|_2 + o(1) + 4(\|f\|_2 + \varepsilon)|\varepsilon|, \text{ as } n \rightarrow \infty.
\]

It follows that \( \limsup |\text{Cov}(f, g o T^n)| \leq 4\|f\|_2 + \|g\|_2 + \varepsilon \). Since \( \varepsilon \) is arbitrary, the limsup, whence the limit itself, is equal to zero.  \( \square \)
1.5 Examples

1.5.1 Circle rotations

Let $T := [0, 1)$ equipped with the Lebesgue measure $m$, and define for $\alpha \in [0, 1)$ $R_\alpha : T \to T$ by $R_\alpha(x) = x + \alpha \mod 1$. $R_\alpha$ is called a circle rotation, because the map $\pi(x) = \exp[2\pi i x]$ is an isomorphism between $R_\alpha$ and the rotation by the angle $2\pi \alpha$ on the unit circle $S^1$.

Proposition 1.4.

1. $R_\alpha$ is measure preserving for every $\alpha$;
2. $R_\alpha$ is ergodic iff $\alpha \notin \mathbb{Q}$;
3. $R_\alpha$ is never strongly mixing.

Proof. A direct calculation shows that the Lebesgue measure $m$ satisfies $m(R_\alpha^{-1}I) = m(I)$ for all intervals $I \subset [0, 1)$. Thus the collection $\mathcal{M} := \{E \in \mathcal{B} : m(R_\alpha^{-1}E) = m(E)\}$ contains the algebra of finite disjoint unions of intervals. It is easy to check $\mathcal{M}$ is a monotone class, so by the monotone class theorem (see appendix) $\mathcal{M}$ contains all Borel sets. It clearly contains all null sets. Therefore it contains all Lebesgue measurable sets. Thus $\mathcal{M} = \mathcal{B}$ and $(1)$ is proved.

We prove (2). Suppose first that $\alpha = p/q$ for $p, q \in \mathbb{N}$. Then $R_\alpha^q = id$. Fix some $x \in [0, 1)$, and pick $\varepsilon$ so small that the $\varepsilon$–neighborhoods of $x + k\alpha$ for $k = 0, \ldots, q - 1$ are disjoint. The union of these neighborhoods is an invariant set of positive measure, and if $\varepsilon$ is sufficiently small then it is not equal to $T$. Thus $R_\alpha$ is not ergodic.

Next assume that $\alpha \notin \mathbb{Q}$. Suppose $E$ is an invariant set, and set $f = 1_E$. Expand $f$ to a Fourier series:

$$f(t) = \sum_{n \in \mathbb{Z}} \hat{f}(n)e^{2\pi i nt} \quad \text{(convergence in } L^2).$$

The invariance of $E$ dictates $f = f \circ R_\alpha$. The Fourier expansion of $f \circ R_\alpha$ is

$$(f \circ R_\alpha)(t) = f(t + \alpha \mod 1) = \sum_{n \in \mathbb{Z}} \hat{f}(n)e^{2\pi in\alpha}.$$

Equating coefficients, we see that $\hat{f}(n) = \hat{f}(n)\exp[2\pi in\alpha]$. Thus either $\hat{f}(n) = 0$ or $\exp[2\pi in\alpha] = 1$. Since $\alpha \notin \mathbb{Q}$, $\hat{f}(n) = 0$ for all $n \neq 0$. We obtain that $f = \hat{f}(0)$ a.e., whence $1_E = m(E)$ almost everywhere. This can only happen if $m(E) = 0$ or $m(E) = 1$, proving the ergodicity of $m$.

To show that $m$ is not mixing, we consider the function $f(x) = \exp[2\pi in x]$. This function satisfies $f \circ R_\alpha = \lambda f$ with $\lambda = \exp[2\pi i n\alpha]$ (such a function is called an eigenfunction). For every $\lambda$ there is a sequence $n_k \to \infty$ s.t. $n_k \alpha \mod 1 \to 0$ (Dirichlet theorem), thus $\|f \circ R_\alpha^n - f\|_2 = |\lambda^n - 1| \to 0$. It follows that $F := \text{Re}(f) = \cos(2\pi x)$ satisfies $\|F \circ R_\alpha^n - F\|_2 \to 0$, whence $\int F \circ R_\alpha^n F dm \to \int F^2 dm \neq (\int F)^2$, and $m$ is not mixing. □
1.5.2 The angle doubling map

Again, we work with $\mathbb{T} := [0, 1]$ equipped with the Lebesgue measure $m$, and define $T : \mathbb{T} \to \mathbb{T}$ by $T(x) = 2x \mod 1$. $T$ is called the angle doubling map, because the map $\pi(x) := \exp[2\pi i x]$ is an isomorphism between $T$ and the map $e^{i\theta} \mapsto e^{2i\theta}$ on $S^1$.

**Proposition 1.5.** The angle doubling map is probability preserving, and strong mixing, whence ergodic.

**Proof.** It is convenient to work with binary expansions $x = (0.d_1d_2d_3\ldots)$, $(d_i = 0, 1)$, because with this representation $T(0.d_1d_2\ldots) = (0.d_2d_3\ldots)$. For every finite $n$–word of zeroes and ones $(d_1, \ldots, d_n)$, define the sets (called “cylinders”)

$$[d_1, \ldots, d_n] := \{x \in [0, 1] : x = (0.d_1\ldots d_n\varepsilon_1\varepsilon_2\ldots), \text{ for some } \varepsilon_i \in \{0, 1\}\}.$$

This is a (dyadic) interval, of length $1/2^n$.

It is clear that $T^{-1}[d_1, \ldots, d_n] = [*d_1, \ldots, d_n]$ where $*$ stands for “$0$ or $1$”. Thus, $m(T^{-1}[d]) = m[0,d] + m[1,d] = 2 \cdot 2^{-(n+1)} = 2^{-n} = m[d]$. We see that $\mathcal{M} := \{E \in \mathcal{B} : m(T^{-1}E) = m(E)\}$ contains the algebra of finite disjoint unions of cylinders. Since $\mathcal{M}$ is obviously a monotone class, and since the cylinders generate the Borel $\sigma$–algebra (prove!), we get that $\mathcal{M} = \mathcal{B}$, whence $T$ is measure preserving.

We prove that $T$ is mixing. Suppose $f, g$ are indicators of cylinders: $f = 1_{[a_1, \ldots, a_n]}$, $g = 1_{[b_1, \ldots, b_m]}$. Then for all $k > n$,

$$\int f \cdot g \circ T^k \, dm = m[a_1 \ast \ldots \ast b] = m[a]m[b].$$

Thus $\text{Cov}(f, g \circ T^k) \xrightarrow{k \to \infty} 0$ for all indicators of cylinders. Every $L^2$–function can be approximated in $L^2$ by a finite linear combination of indicators of cylinders (prove!). One can therefore proceed as in the proof of proposition 1.3 to show that $\text{Cov}(f, g \circ T^k) \xrightarrow{k \to \infty} 0$ for all $L^2$ functions. $\square$

1.5.3 Bernoulli Schemes

Let $S$ be a finite set, called the alphabet, and let $X := S^{\mathbb{N}}$ be the set of all one–sided infinite sequences of elements of $S$. Impose the following metric on $X$:

$$d((x_n)_{n \geq 0}, (y_n)_{n \geq 0}) := 2^{-\min\{k : x_k \neq y_k\}}. \quad (1.2)$$

The resulting topology is generated by the collection of cylinders:

$$[a_0, \ldots, a_{n-1}] := \{x \in X : x_i = a_i \ (0 \leq i \leq n-1)\}.$$
It can also be characterized as being the product topology on $S^N$, when $S$ is given the discrete topology. In particular this topology is compact.

The left shift is the transformation $T : (x_0, x_1, x_2, \ldots) \mapsto (x_1, x_2, \ldots)$. The left shift is continuous.

Next fix a vector $p = (p_a)_{a \in S}$ of positive numbers whose sum is equal to one.

**Definition 1.6.** The Bernoulli measure corresponding to $p$ is the unique measure on the Borel $\sigma$–algebra of $X$ such that $\mu([a_0, \ldots, a_{n-1}]) = p_{a_0} \cdots p_{a_{n-1}}$ for all cylinders.

It is useful to recall why such a measure exists. Here is a review of the necessary tools from measure theory.

**Definition 1.7.** Let $X$ be a non-empty set.

1. A semi-algebra on $X$ is a collection $\mathcal{F}$ of subsets of $X$ such that
   a. $\emptyset, X \in \mathcal{F}$
   b. $\mathcal{F}$ is closed under intersections, and
   c. for every $A \in \mathcal{F}$, $X \setminus A$ is a finite disjoint union of elements from $\mathcal{F}$.

2. The $\sigma$–algebra generated by $\mathcal{F}$ is the smallest $\sigma$–algebra of subsets of $X$ which contains $\mathcal{F}$. Equivalently, this is the $\sigma$–algebra equal to the intersection of all $\sigma$–algebras which contain $\mathcal{F}$ (e.g. the power set of $X$).

3. A function $\mu : \mathcal{F} \rightarrow [0, \infty]$ is called $\sigma$–finite if $X$ is a countable disjoint union of elements $A_i \in \mathcal{F}$ such that $\mu(A_i) < \infty$ for all $i$. This always happens if $\mu(X) < \infty$.

4. A function $\mu : \mathcal{F} \rightarrow [0, \infty]$ is called $\sigma$–additive on $\mathcal{F}$ if for all pairwise disjoint countable collection of $A_i \in \mathcal{F}$, if $\biguplus A_i \in \mathcal{F}$ then $\mu(\biguplus A_i) = \sum \mu(A_i)$.

**Theorem 1.2 (Carathéodory’s Extension Theorem).** Let $X$ be a set and $\mathcal{F}$ a semi-algebra of subsets of $X$. Every $\sigma$–additive $\sigma$–finite $\mu : \mathcal{F} \rightarrow [0, \infty]$ has a unique extension to a $\sigma$–additive $\sigma$–finite function on the $\sigma$–algebra generated by $\mathcal{F}$.

In our case $X = \{(x_0, x_1, \ldots) : x_i \in S$ for all $i\}, \mathcal{F} = \{\emptyset, X\} \cup \{\text{cylinders}\}$, and $\mu : \mathcal{F} \rightarrow [0, \infty]$ is defined by $\mu(\emptyset) := 0$, $\mu(X) := 1$, $\mu([a_0, \ldots, a_{n-1}]) := p_{a_0} \cdots p_{a_{n-1}}$.

$\mathcal{F}$ is a semi-algebra, because the intersection of cylinders is empty or a cylinder, and because the complement of a cylinder $[a_0, \ldots, a_{n-1}]$ is the finite disjoint union of cylinders $[b_0, \ldots, b_{n-1}]$ such that for some $i$, $b_i \neq a_i$.

It is also clear that $\mu : \mathcal{F} \rightarrow [0, \infty]$ is $\sigma$–finite. Indeed it is finite ($\mu(X) < \infty$)! It remain to check that $\mu$ is $\sigma$–additive on $\mathcal{F}$.

Suppose $[a]$ is a countable disjoint union of cylinders $[b_i]$. Each cylinder is open and compact (prove!), so such unions are necessarily finite. Let $N$ be the maximal length of the cylinder $[b_i]$. Since $|b_i| \subseteq [a]$, we can write $[b_i] = [a, b_i] = \bigcup_{c=|a|}^{N-|b_i|} [a, \overline{b_i}, c]$, and a direct calculation shows that

$$\sum_{|c| = N-|b_i|} \mu[a, \overline{b_i}, c] = \mu[a, \overline{b_i}] \left( \sum_c p_c \right)^{N-|b_i|} = \mu[a, \overline{b_i}] = \mu[b_i].$$

Summing over $i$, we get that $\sum_i \mu[b_i] = \sum_i \sum_{|c| = N-|b_i|} \mu[a, \overline{b_i}, c]$. 


1.5 Examples

Now $[\mathbb{a}] = \{ b^j \mid |b^j| = N - |\mathbb{a}|, b^j \in \beta \}$. Thus the collection of $(\beta^j, c)$ is equal to the collection of all possible words $w$ of length $N - |\mathbb{a}|$ (otherwise the right hand side misses some sequences). Thus

$$\sum_j \mu[b^j] = \sum_{|w| = N - |\mathbb{a}|} \mu[\mathbb{a}, w] = \mu[\mathbb{a}] \left( \sum_c p_c \right)^{N - |\mathbb{a}|} = \mu[\mathbb{a}],$$

proving the $\sigma$–additivity of $\mu$ on $\mathcal{F}$.

It now follows from the Carathéodory’s Extension Theorem that $\mu : \mathcal{F} \to [0, \infty]$ has a unique extension to a probability measure on the smallest $\sigma$-algebra containing the cylinders.

Call this $\sigma$-algebra $\mathcal{B}$. Since the cylinders generate the topology of $X$, every open set is a union of cylinders. This union is countable, because there are only countably many cylinders. So every open set belongs to $\mathcal{B}$. It follows that $\mathcal{B}$ contains the Borel $\sigma$-algebra of $X$ (which equals by definition to the smallest $\sigma$-algebra containing all open sets). So $\mu$ extends uniquely to a non-negative $\sigma$-additive function on the Borel $\sigma$-algebra of $X$. Definition probability measure. Definition 1.6 is justified.

**Proposition 1.6.** Suppose $X = \{0, 1\}^\mathbb{N}$, $\mu$ is the $\left(\frac{1}{2}, \frac{1}{2}\right)$-Bernoulli measure, and $\sigma$ is the left shift, then $(X, \mathcal{B}(X), \mu, T)$ is measure theoretically isomorphic to the angle doubling map.

**Proof.** The isomorphism is $\pi(x_0, x_1, \ldots) := \sum 2^{-n} x_n$. This is a bijection between

$$X' := \{ x \in \{0, 1\}^\mathbb{N} : \exists n \text{ s.t. } x_m = 1 \text{ for all } m \geq n \}$$

and $[0, 1)$ (prove that $\mu(X') = 1$), and it is clear that $\pi \circ \sigma = T \circ \pi$. Since the image of a cylinder of length $n$ is a dyadic interval of length $2^{-n}$, $\pi$ preserves the measures of cylinders. The collection of measurable sets which are mapped by $\pi$ to sets of the same measure is a $\sigma$-algebra. Since this $\sigma$-algebra contains all the cylinders and all the null sets, it contains all measurable sets.

**Proposition 1.7.** Every Bernoulli scheme is mixing, whence ergodic.

The proof is the same as in the case of the angle doubling map. Alternatively, it follows from the mixing of the angle doubling map and the fact that the two are isomorphic.

### 1.5.4 Finite Markov Chains

We saw that the angle doubling map is isomorphic to a dynamical system acting as the left shift on a space of sequences (a Bernoulli scheme). Such representations appear frequently in the theory of dynamical systems, but more often than not, the space of sequences is slightly more complicated than the set of all sequences.
1.5.4.1 Subshifts of finite type

Let $S$ be a finite set, and $A = (a_{ij})_{S \times S}$ a matrix of zeros and ones without columns or rows made entirely of zeroes.

**Definition 1.8.** The subshift of finite type (SFT) with alphabet $S$ and transition matrix $A$ is the set $\Sigma_A = \{ x = (x_0, x_1, \ldots) \in S^\mathbb{N} : a_{x_i x_{i+1}} = 1 \text{ for all } i \}$, together with the metric $d(x, y) := 2^{-\min\{|k : x_k \neq y_k\}}$ and the action $\sigma(x_0, x_1, x_2, \ldots) = (x_1, x_2, \ldots)$.

This is a compact metric space, and the left shift map $\sigma : \Sigma_A \to \Sigma_A$ is continuous.

We think of $\Sigma_A$ as of the space of all infinite paths on a directed graph with vertices $S$ and edge $a \to b$ connecting $a, b \in S$ such that $a_{ab} = 1$.

Let $\Sigma_A^+$ be a SFT with set of states $S$, $|S| < \infty$, and transition matrix $A = (a_{ab})_{S \times S}$.

- A stochastic matrix is a matrix $P = (p_{ab})_{a,b \in S}$ with non-negative entries, such that $\sum_b p_{ab} = 1$ for all $a$, i.e. $P1 = 1$. The matrix is called compatible with $A$, if $A_{ab} = 0 \Rightarrow p_{ab} = 0$.
- A probability vector is a vector $p = (p_a : a \in S)$ of non-negative entries, s.t. $\sum_a p_a = 1$.
- A stationary probability vector is a probability vector $p = (p_a : a \in S)$ s.t. $p^T P = p$.

Given a probability vector $p$ and a stochastic matrix $P$ compatible with $A$, one can define a Markov measure $\mu$ on $\Sigma_A^+$ (or $\Sigma_A$) by

$$\mu[a_0, \ldots, a_{n-1}] := p_{a_0} p_{a_0 a_1} \cdots p_{a_{n-2} a_{n-1}},$$

where $[a_0, \ldots, a_{n-1}] = \{ x \in \Sigma_A^+ : x_i = a_i \ (i = 0, \ldots, n-1) \}$. The stochasticity of $P$ guarantees that this measure is finitely additive on the algebra of cylinders, and $\sigma$-subadditivity can be checked as for Bernoulli measures using compactness. Thus this gives a Borel probability measure on $\Sigma_A^+$.

**Proposition 1.8.** $\mu$ is shift invariant iff $p$ is stationary w.r.t. $P$. Any stochastic matrix has a stationary probability vector.

**Proof.** To see the first half of the statement, we note that $\mu$ is stationary iff $\mu[a, b] = \mu[b]$ for all $[b]$, which is equivalent to

$$\sum_a p_a p_{ab_0} p_{b_0 b_1} \cdots p_{b_{n-2} b_{n-1}} = p_{b_0} p_{b_0 b_1} \cdots p_{b_{n-2} b_{n-1}}.$$

Canceling the identical terms on both sides gives $\sum_a p_a p_{ab_0} = p_{b_0}$. Thus $\mu$ is shift invariant iff $p$ is $P$–stationary.

We now show that every stochastic matrix has a stationary probability vector. Consider the right action of $P$ on the simplex $\Delta$ of probability vectors in $\mathbb{R}^N$:

$$\Delta := \{ (x_1, \ldots, x_N) : x_i \geq 0, \sum_i x_i = 1 \}, \quad T(x) = xP.$$

We have $T(\Delta) \subseteq \Delta$, since $\sum_a (T x)_a = \sum_a x_a \sum_{b \in S} p_{ba} = \sum_b x_b \sum_a p_{ba} = \sum_b x_b = 1$. Recall Brouwer’s fixed point theorem: A continuous mapping of a compact convex
subset of $\mathbb{R}^d$ into itself has a fixed point. Applying this to $T : \Delta \to \Delta$ we find $x \in \Delta$ such that $xP = x$. This is the stationary probability vector.

Thus every stochastic matrix determines (at least one) shift invariant measure. Such measures are called Markov measures. We ask when is this measure ergodic, and when is it mixing.

### 1.5.4.2 Ergodicity and mixing of Markov measures

There are obvious obstructions to ergodicity and mixing. To state them concisely, we introduce some terminology. Suppose $P = (p_{ab})_{a,b \in S}$ is a stochastic matrix. We say that $a$ connects to $b$ in $n$ steps, and write $a \overset{n}{\to} b$, if there is a path of length $n+1$ $(a, \xi_1, \ldots, \xi_n, b) \in S^{n+1}$ s.t. $p_{a\xi_1}p_{\xi_1\xi_2}\cdots p_{\xi_n\overset{1}{b}} > 0$ (see problem 1.5).

**Definition 1.19.** A stochastic matrix $P = (p_{ab})_{a,b \in S}$ is called irreducible, if for every $a,b \in S$ there exists an $n$ s.t. $a \overset{n}{\to} b$.

**Lemma 1.1.** If $A$ is irreducible, then $p := \gcd\{n : a \overset{n}{\to} a\}$ is independent of $a$. (gcd = greatest common divisor).

**Proof.** Let $p_a := \gcd\{n : a \overset{n}{\to} a\}$, $p_b := \gcd\{n : b \overset{n}{\to} b\}$, and $\Lambda_b := \{n : b \overset{n}{\to} b\}$. Then $\Lambda_b + \Lambda_b \subset \Lambda_b$, and therefore $\Lambda_b - \Lambda_b$ is a subgroup of $\mathbb{Z}$. Necessarily $\Lambda_b - \Lambda_b = p_b\mathbb{Z}$.

By irreducibility, there are $\alpha, \beta$ s.t. $a \overset{\beta}{\to} b \overset{\beta}{\to} a$. So for all $n \in \Lambda_b$, $a \overset{n_0}{\to} b \overset{n_0}{\to} a$, whence $p_a | \gcd(\alpha + \beta + \Lambda_b)$. This clearly implies that $p_a$ divides every number in $\Lambda_b - \Lambda_b = p_b\mathbb{Z}$. So $p_a | p_b$. Similarly one shows that $p_b | p_a$, whence $p_a = p_b$. $\square$

**Definition 1.10.** The period of an irreducible stochastic matrix $P$ is the number $p := \gcd\{n : a \overset{n}{\to} a\}$ (this is independent of $a$ by the lemma). An irreducible stochastic matrix is called aperiodic if its period is equal to one.

For example, the SFT with transition matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is irreducible with period two.

If $P$ is not irreducible, then any Markov measure of $P$ and a strictly positive stationary probability vector of $P$ is non-ergodic. To see why, pick $a,b \in S$ s.t. $a$ does not connect to $b$ in any number of steps. The set

$$E := \{x \in \Sigma_A^+ : x_i \neq b \text{ for all } i \text{ sufficiently large}\}$$

is a shift invariant set which contains $[a]$, but which is disjoint from $[b]$. So $E$ is an invariant set with such that $0 < p_a \leq \mu(E) \leq 1 - p_b < 1$. So $\mu$ is not ergodic.

If $P$ is irreducible, but not aperiodic, then any Markov measure on $\Sigma_A^+$ is non-mixing, because the function $f := 1_{[a]}$ satisfies $ff \circ T^n \equiv 0$ for all $n$ not divisible by the period. This means that $\int ff \circ T^n d\mu$ is equal to zero on a subsequence, and therefore cannot converge to $\mu[a]^2$.

We claim that these are the only possible obstructions to ergodicity and mixing. The proof is based on the following fundamental fact, whose proof will be given at the next section.
**Theorem 1.3 (Ergodic Theorem for Markov Chains).** Suppose $P$ is a stochastic matrix, and write $P^n = (p_{ab}^{(n)})$, then $P$ has a stationary probability vector $\mu$ and

1. if $P$ is irreducible, then $\frac{1}{n} \sum_{k=1}^{n} P_{ab}^{(k)} \xrightarrow[n \to \infty]{} p_{ab}$ $(a, b \in S)$;

2. if $P$ is irreducible and aperiodic, then $P_{ab}^{(k)} \xrightarrow[n \to \infty]{} p_{ab}$ $(a, b \in S)$.

**Corollary 1.1.** A shift invariant Markov measure on a SFT $\Sigma^+_\Lambda$ is ergodic iff its stochastic matrix is irreducible, and mixing iff its stochastic matrix is irreducible and aperiodic.

**Proof.** Let $\mu$ be a Markov measure with stochastic matrix $P$ and stationary probability vector $\mu$. For all cylinders $[a] = [a_0, \ldots, a_{n-1}]$ and $[b] = [b_0, \ldots, b_{m-1}]$,

$$
\mu([a] \cap \sigma^{-k}[b]) = \mu \left( \bigcup_{\xi \in \mathcal{W}_{k-\Lambda}} [a, \xi, b] \right), \quad \mathcal{W}_k := \{ \xi = (\xi_1, \ldots, \xi_k) : [a, \xi, b] \neq \emptyset \}
$$

$$
= \mu[a] \cdot \sum_{\xi \in \mathcal{W}_{k-\Lambda}} P_{a_{n-1} \xi_1} \cdots P_{\xi_{k-i} b_0} \cdot \frac{\mu[b]}{p_{b_0}} = \mu[a] \mu[b] \cdot \frac{P_{a_{n-1} b_0}^{(k-n)}}{p_{b_0}}.
$$

If $P$ is irreducible, then by theorem 1.3, $\frac{1}{n} \sum_{k=0}^{n-1} \mu([a] \cap \sigma^{-k}[b]) \xrightarrow[n \to \infty]{} \mu[a] \mu[b]$.

We claim that this implies ergodicity. Suppose $E$ is an invariant set, and fix $\varepsilon > 0$, arbitrarily small. There are cylinders $A_1, \ldots, A_N \in \mathcal{F}$ s.t. $\mu \left( E \bigtriangleup \bigcup_{i=1}^{N} A_i \right) < \varepsilon$. Thus

$$
\mu(E) = \mu(E \cap \sigma^{-k}E) = \sum_{i=1}^{N} \mu(A_i \cap \sigma^{-k}E) \pm \varepsilon = \sum_{i,j=1}^{N} \mu(A_i \cap \sigma^{-k}A_j) \pm 2\varepsilon.
$$

Averaging over $k$, and passing to the limit, we get

$$
\mu(E) = \lim_{k \to \infty} \frac{1}{n} \sum_{k=1}^{n} \mu(A_i \cap \sigma^{-k}A_j) \pm 2\varepsilon = \sum_{i,j=1}^{N} \mu(A_i) \mu(A_j) \pm 2\varepsilon
$$

$$
= \left( \sum_{i=1}^{N} \mu(A_i) \right)^2 \pm 2\varepsilon = [\mu(E) \pm \varepsilon]^2 \pm 2\varepsilon.
$$

Passing to the limit $\varepsilon \to 0^+$, we obtain $\mu(E) = \mu(E)^2$, whence $\mu(E) = 0$ or 1.

Now assume that $P$ is irreducible and aperiodic. The ergodic theorem for Markov chains says that $\mu([a] \cap \sigma^{-k}[b]) \xrightarrow[k \to \infty]{} \mu[a] \mu[b]$. Since any measurable sets $E, F$ can approximated by finite disjoint unions of cylinders, an argument similar to the previous one shows that $\mu(E \cap \sigma^{-k}F) \xrightarrow[k \to \infty]{} \mu(E) \mu(F)$ for all $E, F \in \mathcal{B}$. This is is mixing.

---

1 Proof: The collection of sets $E$ satisfying this approximation property is a $\sigma$-algebra which contains all cylinders, therefore it is equal to $\mathcal{B}$. 

Remark 1. The ergodic theorem for Markov chains can be visualized as follows.
Imagine that we distribute mass on the states of $S$ according to a probability distribution $\frac{1}{2} = (q_a)_{a \in S}$. Now shift mass from one state to another using the rule that a $p_{ab}$-fraction of the mass at $a$ is moved to $b$. The new mass distribution is $qP$ (check). After $n$ steps, the mass distribution is $qP^n$. The previous theorem says that, in the aperiodic case, the mass distribution converges to the stationary distribution — the equilibrium state. It can be shown that the rate of convergence is exponential (problem 1.7).

Remark 2: The ergodic theorem for Markov chains has an important generalization to all matrices with non-negative entries, see problem 1.6.

1.5.4.3 Proof of the Ergodic Theorem for Markov chains

Suppose first that $P$ is an irreducible and aperiodic stochastic matrix. This implies that there is some power $m$ such that all the entries of $P^m$ are strictly positive.$^2$

Let $N := |S|$ denote the number of states, and consider the set of all probability vectors $\Delta := \{(x_1, \ldots, x_N) : x_i \geq 0, \sum x_i = 1\}$. Since $P$ is stochastic, the map $T(x) = xP$ maps $\Delta$ continuously into itself. By the Brouwer Fixed Point Theorem, there is a probability vector $\xi$ s.t. $pP = \xi$ (this is the stationary probability vector).

The irreducibility of $P$ implies that all the coordinates of $\xi$ are strictly positive. Indeed, $\sup p_k = 1$ so at least one coordinate $\xi_i$ is positive. For any other coordinate $p_j$, there is by irreducibility a path $(i, \xi_1, \ldots, \xi_{n-1}, j)$ such that

$$p_1\xi_1P_{\xi_1}\xi_2\cdots P_{\xi_{n-2}}\xi_{n-1}P_{\xi_{n-1}}\xi_j > 0.$$ 

So $p_j = (pP^n)_j = \sum_k p_k p_k^{(n)} \geq p_1\xi_1P_{\xi_1}\xi_2\cdots P_{\xi_{n-2}}\xi_{n-1}P_{\xi_{n-1}}\xi_j > 0$.

So $\xi$ lies in the interior of $\Delta$, and the set $C := \Delta - \xi$ is a compact convex neighborhood of the origin such that $T(C) \subset C$, $T^m(C) \subset \text{int}[C]$. (We mean the relative interior in the $(N-1)$-dimensional space $C$, not the ambient space $\mathbb{R}^N$.)

Now consider $L := \text{span}(\xi)$ (an $N-1$-dimensional space). This is an invariant space for $T$, whence for $P^T$ (the transpose of $P$), we claim that all the eigenvalues of $P^T | L$ have absolute value less than 1:

1. Eigenvalues of modulus larger than one are impossible, because $P$ is stochastic, so $\|xP\|_1 \leq \|x\|_1$, so the spectral radius of $P^T$ cannot be more than 1.

2. Roots of unity are impossible, because in this case for some $k$, $P^km$ has a real eigenvector $y$ with eigenvalue one. There is no loss of generality in assuming that $y \in \partial C$, otherwise multiply $y$ by a suitable scalar. But $P^km$ cannot have fixed points on $\partial C$, because $P^km(C) \subset \text{int}[C]$.

3. Eigenvalues $e^{i\theta}$ with $\theta \notin 2\pi \mathbb{Q}$ are impossible, because if $e^{i\theta}$ is an eigenvalue then there are two real eigenvectors $y, z \in \partial C$ such that the action of $P$ on $\text{span}\{y, z\}$

---

$^2$ Begin by proving that if $A$ is irreducible and aperiodic, then for every $a$ there is an $N_a$ s.t. $a \overset{n}{\rightarrow} a$ for all $n > N_a$. Use this to show that for all $a, b$ there is an $N_{ab}$ s.t. $a \overset{n}{\rightarrow} b$ for all $n > N_{ab}$. Take $m = \max\{N_{ab}\}$. 

is conjugate to \( \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \), an irrational rotation. This means \( \exists k_n \to \infty \) s.t. 
\( \mu^ {p^m k_n} \to \mu \in \partial C \). But this cannot be the case because \( P^n[C] \subset \text{int}[C] \), and by compactness, this cannot intersect \( \partial C \).

In summary the spectral radius of \( P^\ell |_L \) is less than one.

But \( \mathbb{R}^N = L \oplus \text{span}\{P\} \). If we decompose a general vector \( v = q + tp \) with \( q \in L \), then the above implies that 
\( v P^n = tp + O(\|P^\ell|_L\|)\|q\| \xrightarrow{n \to \infty} tp \). It follows that 
\( p_{ab}^{(n)} \xrightarrow{n \to \infty} p_b \) for all \( a, b \).

This is almost the ergodic theorem for irreducible aperiodic Markov chains, the only thing which remains is to show that \( P \) has a unique stationary vector. Suppose \( q \) is another probability vector s.t. \( qP = q \). We can write \( p_{ab}^{(n)} \to p_b \) in matrix form as follows:
\( P^n \xrightarrow{n \to \infty} Q \), where \( Q = (q_{ab})_{S \times S} \) and \( q_{ab} = p_b \).

Consider now the periodic irreducible case. Let \( A \) be the transition matrix associated to \( P \) (with entries \( t_{ab} = 1 \) when \( p_{ab} > 0 \) and \( t_{ab} = 0 \) otherwise), and let \( p \) denote the period of \( P \). Fix once and for all a state \( v \). Working with the SFT \( \Sigma^\ell \), we let 
\( S_k := \{b \in S : v \xrightarrow{\ell} b \text{ for some } n = k \mod p \} \) (\( k = 0, \ldots, p-1 \)).

\( S_k \) are pairwise disjoint, because if \( b \in S_{k_1} \cap S_{k_2} \), then \( \exists \alpha_i = k_i \mod p \) and \( \exists \beta \)
\( v \xrightarrow{\ell} b \xrightarrow{\ell} v \) for \( i = 1, 2 \). By the definition of the period, \( p|\alpha_i + \beta \) for \( i = 1, 2 \), whence \( k_1 = \alpha_1 = -\beta = \alpha_2 = k_2 \mod p \).

It is also clear that every path of length \( \ell \) which starts at \( S_k \), ends at \( S_{k'} \) where \( k' = k + \ell \mod p \). In particular, every path of length \( p \) which starts at \( S_k \) ends at \( S_k \).
This means that if \( p_{ab}^{(p)} > 0 \), then \( a, b \) belong to the same \( S_k \).

It follows that \( P^n \) is conjugate, via a coordinate permutation, to a block matrix with blocks \( (p_{ab}^{(0)})_{S_k \times S_k} \). Each of the blocks is stochastic, irreducible, and aperiodic. Let \( \pi^{(k)} \) denote the stationary probability vectors of the blocks.

By the first part of the proof, \( p_{ab}^{(p)} \xrightarrow{\ell \to \infty} \pi^{(k)} \) for all \( a, b \) in the same \( S_k \), and 
\( p_{ab}^{(n)} = 0 \) for \( n \not\equiv 0 \mod p \). More generally, if \( a \in S_{k_1} \) and \( b \in S_{k_2} \), then
\[
\lim_{\ell \to \infty} p_{ab}^{(p+k_2-k_1)} = \lim_{\ell \to \infty} \sum_{\xi \in S_{k_2}} p_{a\xi}^{(k_2-k_1+p)} \pi^{(k_2)}_b = \pi^{(k_2)}_b.
\]
\( p_{ab}^{(p+\alpha)} = 0 \) when \( \alpha \not\equiv k_2 - k_1 \mod p \). Thus
\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n} p_{ab}^{(k)} = \frac{1}{p} \pi_{b}^{(k)} \text{ for the unique } k \text{ s.t. } S_k \ni b
\]

The limiting vector \( \pi \) is a probability vector, because \( \sum_{k=1}^{p} \sum_{b \in S_k} S_k \frac{1}{n} \pi_{b}^{(k)} = 1 \).

We claim that \( \pi \) is the unique stationary probability vector of \( P \). The limit theorem for \( p_{ab}^{(n)} \) can be written in the form \( \frac{1}{n} \sum_{k=0}^{n-1} p^{k} \to Q \) where \( Q = (q_{ab})_{S \times S} \) and \( q_{ab} = \pi_{b} \). As before this implies that \( \pi P = \pi \) and that any probability vector \( q \) such that \( q P = q \), we also have \( q Q = q \), whence \( q = p \). \( \square \)

### 1.5.5 The geodesic flow on a hyperbolic surface

The hyperbolic plane is the surface \( \mathbb{H} := \{ z \in \mathbb{C} : \text{Im}(z) > 0 \} \) equipped with the Riemannian metric \( ds = |dz|/\text{Im}(z) \), which gives it constant curvature \((-1)\).

It is known that the orientation preserving isometries (i.e. distance preserving maps) are the Möbius transformations which preserve \( \mathbb{H} \). They form the group

\[
\text{Möb}(\mathbb{H}) = \left\{ \frac{az + b}{cz + d} : a, b, c, d \in \mathbb{R}, ad - bc = 1 \right\}
\]

\[
\simeq \left\{ \left( \begin{array}{cc} a & b \\ c & d \end{array} \right) : a, b, c, d \in \mathbb{R}, ad - bc = 1 \right\} / \left\{ \pm \text{id} \right\} =: \text{PSL}(2, \mathbb{R}).
\]

(We quotient by \( \left\{ \pm \text{id} \right\} \) to identify \((\begin{smallmatrix} a & b \\ c & d \end{smallmatrix})\), \((\begin{smallmatrix} -a & -b \\ -c & -d \end{smallmatrix})\) which represent the same Möbius transformation.)

The geodesics (i.e. length minimizing curves) on the hyperbolic plane are vertical half–lines, or circle arcs which meet \( \partial \mathbb{H} \) at right angles. Here is why: Suppose \( \omega \in TM \) is a unit tangent vector which points directly up, then it is not difficult to see that the geodesic at direction \( \omega \) is a vertical line. For general unit tangent vectors \( \omega \), find an element \( \varphi \in \text{Möb}(\mathbb{H}) \) which rotates them so that \( d\varphi(\omega) \) points up. The geodesic in direction \( \omega \) is the \( \varphi \)-preimage of the geodesic in direction \( d\varphi(\omega) \) (a vertical half–line). Since Möbius transformations map lines to lines or circles in a conformal way, the geodesic of \( \omega \) is a circle meeting \( \partial \mathbb{H} \) at right angles.

The geodesic flow of \( \mathbb{H} \) is the flow \( g^t \) on the unit tangent bundle of \( \mathbb{H} \),

\[
T^1 \mathbb{H} := \{ \text{tangent vectors with length one} \}
\]

which moves \( \omega \) along the geodesic it determines at unit speed.

To describe this flow it useful to find a convenient parametrization for \( T^1 \mathbb{H} \). Fix \( \omega_0 \in T^1 \mathbb{H} \) (e.g. the unit vector based at \( i \) and pointing up). For every \( \omega \), there is a unique \( \varphi_{\omega} \in \text{Möb}(\mathbb{H}) \) such that \( \omega = d\varphi_{\omega}(\omega_0) \), thus we can identify

\[
T^1 \mathbb{H} \simeq \text{Möb}(\mathbb{H}) \simeq \text{PSL}(2, \mathbb{R}).
\]

It can be shown that in this coordinate system the geodesic flow takes the form
$g^t \left( \begin{array}{cc} a & b \\ c & d \end{array} \right) = \left( \begin{array}{cc} a & b \\ c & d \end{array} \right) \left( \begin{array}{cc} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{array} \right).$

To verify this, it is enough to calculate the geodesic flow on $\omega_0 \simeq \left( \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right)$.

Next we describe the Riemannian volume measure on $T^1 \mathbb{H}$ (up to normalization). Such a measure must be invariant under the action of all isometries. In our coordinate system, the isometry $\varphi(z) = (az + b)/(cz + d)$ acts by

$$\varphi \left( \begin{array}{cc} x & y \\ z & w \end{array} \right) = \left( \begin{array}{cc} a & b \\ c & d \end{array} \right) \left( \begin{array}{cc} x & y \\ z & w \end{array} \right).$$

Since $\text{PSL}(2, \mathbb{R})$ is a locally compact topological group, there is only one Borel measure on $\text{PSL}(2, \mathbb{R})$ (up to normalization), which is left invariant by all left translations on the group: the Haar measure of $\text{PSL}(2, \mathbb{R})$. Thus the Riemannian volume measure is a left Haar measure of $\text{PSL}(2, \mathbb{R})$, and this determines it up to normalization.

It is a convenient feature of $\text{PSL}(2, \mathbb{R})$ that its left Haar measure is also invariant under right translations. It follows that the geodesic flow preserves the volume measure on $T^1 \mathbb{H}$. But this measure is infinite, and it is not ergodic (prove!).

To obtain ergodic flows, we need to pass to compact quotients of $\mathbb{H}$. These are called hyperbolic surfaces.

A hyperbolic surface is a Riemannian surface $M$ such that every point in $M$ has a neighborhood $V$ which is isometric to an open subset of $\mathbb{H}$. Recall that a Riemannian surface is called complete, if every geodesic can be extended indefinitely in both directions.

**Theorem 1.4 (Killing–Hopf Theorem).** Every complete connected hyperbolic surface is isometric to $\Gamma \backslash \mathbb{H} := \{ \Gamma z : z \in \mathbb{H} \}$, where

1. $\Gamma$ is a subgroup of $\text{Mob}(\mathbb{H})$ and $\Gamma z := \{ g(z) : g \in \Gamma \}$.
2. Every point $z \in \mathbb{H}$ is in the interior of some open set $U \subset \text{PSL}(2, \mathbb{R})$ s.t. $\{ g(U) : g \in \Gamma \}$ are pairwise disjoint. So $\Gamma z' \mapsto$ the unique point in $\Gamma z' \cap U$ is a bijection.
3. the Riemannian structure on $\{ \Gamma z' : z' \in U \}$ is the one induced by the Riemannian structure on $U$.

If we identify $\Gamma$ with a subgroup of $\text{PSL}(2, \mathbb{R})$, then we get the identification $T^1(\Gamma \backslash \mathbb{H}) \simeq \Gamma \backslash \text{PSL}(2, \mathbb{R})$. It is clear that the Haar measure on $\text{PSL}(2, \mathbb{R})$ induces a unique locally finite measure on $\Gamma \backslash \text{PSL}(2, \mathbb{R})$, and that the geodesic flow on $T^1(\Gamma \backslash \mathbb{H})$ takes the form

$$g^t(\Gamma \omega) = \Gamma \omega \left( \begin{array}{cc} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{array} \right),$$

and preserves this measure.

**Definition 1.11.** A measure preserving flow $g^t : X \to X$ is called ergodic, if every measurable set $E$ such that $g^{-1}(E) = E$ for all $t$ satisfies $m(E) = 0$ or $m(E^c) = 0$.
1.6 Basic constructions

This is equivalent to asking that all $L^2$–functions $f$ such that $f \circ g = f$ a.e. for all $t \in \mathbb{R}$ are constant (prove!).

**Theorem 1.5.** If $\Gamma \setminus \mathbb{H}$ is compact, then the geodesic flow on $T^1(\Gamma \setminus \mathbb{H})$ is ergodic.

**Proof.** Consider the following flows:

$$h^t_{st}(\Gamma \omega) = \Gamma \omega \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$$

$$h^t_{un}(\Gamma \omega) = \Gamma \omega \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}$$

If we can show that any geodesic invariant function $f$ is also invariant under these flows then we are done, because it is known that

$$\left\langle \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}, \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \right\rangle = \text{PSL}(2, \mathbb{R})$$

(prove!), and any $\text{PSL}(2, \mathbb{R})$–invariant function on $\Gamma \setminus \text{PSL}(2, \mathbb{R})$ is constant.

Since our measure is induced by the Haar measure, the flows $h^t_{un}, h^t_{st}$ are measure preserving. A matrix calculation shows:

$$g^t h^t_{st} g^{-t} = h^t_{st} \xrightarrow{t \to \infty} \text{id}$$

$$g^{-t} h^t_{un} g^t = h^t_{un} \xrightarrow{t \to -\infty} \text{id}$$

**Step 1.** If $f \in L^2$, then $f \circ h^t_{un} \xrightarrow{t \to \infty} f$.

**Proof.** Approximate by continuous functions of compact support, and observe that $h^t$ is an isometry of $L^2$. 

**Step 2.** If $f \in L^2$ and $f \circ g^t = f$, then $f \circ h^t_{un} = f$ and $f \circ h^t_{st} = f$.

**Proof.** $\|f \circ h^t_{un} - f\|_2 = \|f \circ g^t \circ h^t_{un} - f\|_2 = \|f \circ g^t \circ h^t_{st} \circ g^{-t} - f\|_2 \to 0$.

Thus $f$ is $h^t_{st}$–invariant. A similar calculation shows that it is $h^t_{un}$–invariant, and we are done. 

1.6 Basic constructions

In this section we discuss several standard methods for creating new measure preserving transformations from old ones. These constructions appear quite frequently in applications.
Products

Recall that the product of two measure spaces \((X_i, \mathcal{B}_i, \mu_i)\) \((i = 1, 2)\) is the measure space \((X_1 \times X_2, \mathcal{B}_1 \otimes \mathcal{B}_2, \mu_1 \times \mu_2)\) where \(\mathcal{B}_1 \otimes \mathcal{B}_2\) is the smallest \(\sigma\)-algebra which contains all sets of the form \(B_1 \times B_2\) where \(B_i \in \mathcal{B}_i\), and \(\mu_1 \times \mu_2\) is the unique measure such that \((\mu_1 \times \mu_2)(B_1 \times B_2) = \mu_1(B_1)\mu_2(B_2)\).

This construction captures the idea of independence from probability theory: if \((X_1, \mathcal{B}_1, \mu_1)\) are the probability models of two random experiments, and these experiments are “independent”, then \((X_1 \times X_2, \mathcal{B}_1 \otimes \mathcal{B}_2, \mu_1 \times \mu_2)\) is the probability model of the pair of the experiments, because for every \(E_1 \in \mathcal{B}_1, E_2 \in \mathcal{B}_2\),

\[
F_1 := E_1 \times X_2 \text{ is the event “in experiment } 1, E_1 \text{ happened”}
\]

\[
F_2 := X_1 \times E_2 \text{ is the event “in experiment } 2, E_2 \text{ happened”}
\]

and \(F_1 \cap F_2 = E_1 \times E_2\); so \((\mu_1 \times \mu_2)(F_1 \cap F_2) = (\mu_1 \times \mu_2)(F_1)(\mu_1 \times \mu_2)(F_2)\), showing that the events \(F_1, F_2\) are independent.

Definition 1.12. The product of two measure preserving systems \((X_i, \mathcal{B}_i, \mu_i, T_i)\) \((i = 1, 2)\) is the measure preserving system \((X_1 \times X_2, \mathcal{B}_1 \otimes \mathcal{B}_2, \mu_1 \times \mu_2, T_1 \times T_2)\), where \((T_1 \times T_2)(x_1, x_2) = (T_1x_1, T_2x_2)\).

(Check that \(S\) is measure preserving.)

Proposition 1.9. The product of two ergodic mpt is not always ergodic. The product of two mixing mpt is always mixing.

Proof. The product of two (ergodic) irrational rotations \(S := R_\alpha \times R_\alpha : T^2 \to T^2\), \(S(x, y) = (x + \alpha, y + \alpha) \mod 1\) is not ergodic: \(F(x, y) = y - x \mod 1\) is a non-constant invariant function. (See problem 1.8.)

The product of two mixing mpt is however mixing. To see this set \(\mu = \mu_1 \times \mu_2\), \(S = T_1 \times T_2\), and \(\mathcal{S} := \{A \times B : A \in \mathcal{B}_1, B \in \mathcal{B}_2\}\). For any \(E_1 := A_1 \times B_1, E_2 := A_2 \times B_2 \in \mathcal{S}\),

\[
\mu(E_1 \cap S^{-n}E_2) = \mu(\{(A_1 \times B_1) \cap (T_1 \times T_2)^{-n}(A_2 \times B_2)\}) \\
= \mu(\{(A_1 \cap T^{-n}A_2) \cap (B_1 \cap T^{-n}B_2)\}) \\
= \mu_1(A_1 \cap T^{-n}A_2)\mu_2(B_1 \cap T^{-n}B_2) \\
\to n \to \infty \mu_1(A_1)\mu_2(B_1)\mu_1(A_2)\mu_2(B_2) = \mu(A_1 \times B_1)\mu(A_2 \times B_2).
\]

Since \(\mathcal{S}\) is a semi-algebra which generates \(\mathcal{B}_1 \otimes \mathcal{B}_2\), any element of \(\mathcal{B}_1 \otimes \mathcal{B}_2\) can be approximated by a finite disjoint elements of \(\mathcal{S}\), and a routine approximation argument shows that \(\mu(E \cap S^{-n}F) \to \mu(E)\mu(F)\) for all \(E, F \in \mathcal{B}\).
1.6 Basic constructions

1.6.1 Skew-products

We start with an example. Let \( \mu \) be the \( \left( \frac{1}{2}, \frac{1}{2} \right) \)-Bernoulli measure on the two shift 
\( \Sigma^+ := \{0,1\}^\mathbb{N} \). Let \( f : \Sigma^+_2 \to \mathbb{Z} \) be the function \( f(x_0,x_1,\ldots) = (-1)^{x_0} \). Consider the transformation 
\[
T_f : \Sigma^+_2 \times \mathbb{Z} \to \Sigma^+_2 \times \mathbb{Z} \, , \, T_f(x,k) = (\sigma(x),k+f(x)),
\]
where \( \sigma : \Sigma^+_2 \to \Sigma^+_2 \) is the left shift. This system preserves the (infinite) measure \( \mu \times m_\mathbb{Z} \) where \( m_\mathbb{Z} \) is the counting measure on \( \mathbb{Z} \). The \( n \)-th iterate is 
\[
T^n_f(x,k) = (\sigma^n x,k + x_0 + \cdots + x_{n-1}) \, , \, \text{where} \, \Sigma := (-1)^n.\]

What we see in the second coordinate is the symmetric random walk on \( \mathbb{Z} \), started at \( k \), because (1) the steps \( X_i \) take the values \( \pm 1 \), and (2) \( \{X_i\} \) are independent because of the choice of \( \mu \). We say that the second coordinate is a "random walk on \( \mathbb{Z} \) driven by the noise process \( (\Sigma^+_2, \mathcal{G}, \mu, \sigma) \)."

Here is a variation on this example. Suppose \( T_0, T_1 \) are two measure preserving transformations of the same measure space \( (Y, \mathcal{C}, \nu) \). Consider the transformation \( (X \times Y, \mathcal{B} \otimes \mathcal{C}, \mu \times \nu, T_f) \), where 
\[
T_f(x,y) = (T_x, T_f(x,y)).
\]
The \( n \)-th iterate takes the form \( T^n_f(x,y) = (\sigma^n x, T_{n-1} y \cdots T_0 y) \). The second coordinate looks like the random concatenation of elements of \( \{T_0, T_1\} \). We say that \( T_f \) is a "random dynamical system driven by the noise process \( (X, \mathcal{B}, \mu, T) \)."

These examples suggest the following abstract constructions.

Suppose \( (X, \mathcal{B}, \mu, T) \) is a mpt, and \( G \) is a locally compact Polish\(^3\) topological group, equipped with a left invariant Haar measure \( m_G \). Suppose \( f : X \to G \) is measurable.

**Definition 1.13.** The skew–product with cocycle \( f \) over the basis \( (X, \mathcal{B}, \mu, T) \) is the mpt \( (X \times G, \mathcal{B} \otimes \mathcal{G}(G), \mu \times m_G, T_f) \), where \( T_f : X \times G \to X \times G \) is the transformation 
\[
T_f(x,g) = (T_x, g \cdot f(x)).
\]

(Check, using Fubini's theorem, that this is a mpt.) The \( n \)-th iterate \( T^n_f(x,g) = (T_{n-1}^{x,x} g \cdot f(T_{n-2} x) \cdots f(T_0 x)) \), is a "random walk on \( G \) driven by the noise process \( (X, \mathcal{B}, \mu, T) \)."

Now imagine that the group \( G \) is acting in a measure preserving way on some space \( (Y, \mathcal{C}, \nu) \). This means that there are measurable maps \( T_g : Y \to Y \) such that \( \nu \circ T^{-1}_g = \nu \), \( T_{g_1g_2} = T_{g_1}T_{g_2} \), and \( (g,y) \mapsto T_g(y) \) is a measurable from \( X \times G \) to \( Y \).

**Definition 1.14.** The random dynamical system on \( (Y, \mathcal{C}, \nu) \) with action \( \{T_g : g \in G\} \), cocycle \( f : X \to G \), and noise process \( (X, \mathcal{B}, \mu, T) \), is the system \( (X \times Y, \mathcal{B} \otimes \mathcal{C}, \mu \times \nu, T_f) \) given by 
\[
T_f(x,y) = (T_x, T_f(x,y)).
\]

\(^3\) "Polish" has a topology which makes it a complete separable metric space.
(Check using Fubini’s theorem that this is measure preserving.) Here the \( n \)-th iterate is \( T^n f(x,y) = (T^n x, T_f^n y) \).

It is obvious that if a skew–product (or a random dynamical system) is ergodic or mixing, then its base is ergodic or mixing. The converse is not always true. The ergodic properties of a skew–product depend in a subtle way on the interaction between the base and the cocycle.

Here are two important obstructions to ergodicity and mixing for skew–products. In what follows \( G \) is a polish group and \( \hat{G} \) is its group of characters.

\[
\hat{G} := \{ \gamma : G \to S^1 : \gamma \text{ is a continuous homomorphism} \}.
\]

**Definition 1.15.** Suppose \((X, \mathcal{B}, \mu, T)\) is a ppt and \( f : X \to G \) is Borel.

1. \( f \) is called arithmetic w.r.t. \( \mu \), if \( \exists h : X \to S^1 \) measurable, and \( \gamma \in \hat{G} \) non-constant, such that \( \gamma \circ f = h/h \circ T \) a.e.
2. \( f \) is called periodic w.r.t. \( \mu \), if \( \exists h : X \to S^1 \) measurable, \( |\lambda| = 1 \), and \( \gamma \in \hat{G} \) non-constant, such that \( \gamma \circ f = \lambda h/h \circ T \) a.e.

**Proposition 1.10.** Let \((X, \mathcal{B}, \mu, T)\) be a ppt, \( f : X \to G \) Borel, and \((X \times G, \mathcal{B} \otimes \mathcal{B}(G), \mu \times m_G, T_f)\) the corresponding skew–product. If \( f \) is arithmetic, then \( T_f \) is not ergodic, and if \( f \) is periodic, then \( T_f \) is not mixing.

**Proof.** Suppose \( f \) is arithmetic. The function \( F(x,y) := h(x)\gamma(y) \) satisfies

\[
F(Tx, yf(x)) = h(Tx)\gamma(y)\gamma(f(x)) = h(Tx)\gamma(y)h(x)/h(Tx) = F(x,y),
\]

and we have a non-constant invariant function. So arithmeticity \( \Rightarrow \) non-ergodicity. Similarly, if \( f \) is periodic, then \( F(x,y) = h(x)\gamma(y) \) satisfies \( F(Tx, yf(x)) = \lambda F(x,y) \), whence \( F \circ T_f = \lambda F \). Pick \( n_k \to \infty \) s.t. \( \lambda^{n_k} \to 1 \), then \( \text{Cov}(F,F \circ T_f^{n_k}) \to \int F^2 - (\int F)^2 \). Since \( F \neq \int F \) a.e., the limit is non-zero and we get a contradiction to mixing. So periodicity \( \Rightarrow \) non-mixing. \( \square \)

### 1.6.2 Factors

When we construct skew-products over a base, we enrich the space. A factor is a constructions which depletes the space.

**Definition 1.16.** A mpt transformation \((X, \mathcal{B}, \mu, T)\) is called a (measure theoretic) factor of a mpt transformation \((Y, \mathcal{C}, \nu, S)\), if there are sets of full measure \( X' \subset X, Y' \subset Y \) such that \( T(X') \subset X' \), \( S(Y') \subset Y' \), and a measurable onto map \( \pi : Y' \to X' \) such that \( \nu \circ \pi^{-1} = \mu \) and \( \pi \circ S = T \circ \pi \) on \( Y' \). We call \( \pi \) the factor map.
If \( (X, \mathscr{B}, \mu, T) \) is a factor of \( (Y, \mathcal{C}, \nu, S) \), then it is customary to call \( (Y, \mathcal{C}, \nu, S) \) an extension of \( (X, \mathscr{B}, \mu, T) \) and \( \pi \) the factor map.

There are three principle examples:

1. Any measure theoretic isomorphism between two mpt is a factor map between them. But some factor maps are not isomorphisms because they are not injective.
2. A skew product \( T_f : X \times Y \to X \times Y \) is an extension of its base \( T : X \to X \). The factor map is \( \pi : X \times G \to X, \pi(x, y) = x \).
3. Suppose \( (X, \mathcal{B}, \mu, T) \) is an mpt and \( T \) is measurable w.r.t. a smaller \( \sigma \)-algebra \( \mathcal{C} \subset \mathcal{B} \) (i.e. \( T^{-1}\mathcal{C} \subset \mathcal{C} \)), then \( (X, \mathcal{C}, \mu, T) \) is a factor of \( (X, \mathcal{B}, \mu, T) \). The factor map is the identity.

We dwell a bit more on the third example. In probability theory, \( \sigma \)-algebras model information: a set \( E \) is “measurable”, if we can answer the question “is \( \omega \) in \( E \)”? using the information available to use. For example, if a real number \( x \in \mathbb{R} \) is unknown, but we can “measure” \( |x| \), then the information we have on \( x \) is modeled by the \( \sigma \)-algebra \( \{ E \subset \mathbb{R} : E = [-E] \} \), because we can determined whether \( x \in E \) only for symmetric sets \( E \). By decreasing the \( \sigma \)-algebra, we are forgetting some information. For example if instead of knowing \( |x| \), we only know whether \( 0 \leq |x| \leq 1 \) or not, then our \( \sigma \)-algebra is the finite \( \sigma \)-algebra \( \{ \emptyset, \mathbb{R}, [-1, 1], \mathbb{R} \setminus [-1, 1] \} \).

Here is a typical example. Suppose we have a dynamical system \( (X, \mathcal{B}, \mu, T) \), and we cannot “measure” \( x \), but we can “measure” \( f(x) \) for some measurable \( f : X \to \mathbb{R} \). Then the information we have by observing the dynamical system is encoded in the smallest \( \sigma \)-algebra \( \mathcal{C} \subset \mathcal{B} \) with respect to which \( f \circ T^n \) are all measurable. The dynamical properties we feel in this case are those of the factor \( (X, \mathcal{C}, \mu, T) \) and not of the \( (X, \mathcal{B}, \mu, T) \). For example, it could be the case that \( \mu(E \cap T^{-n}F) = \mu(E) \mu(F) \) for all \( E, F \in \mathcal{C} \) but not for all \( E, F \in \mathcal{B} \) — and then we will observe “mixing” simply because our information is not sufficient to observe the non-mixing in the system.

### 1.6.3 The natural extension

An invertible mpt is a mpt \( (X, \mathcal{B}, \mu, T) \) such that for some invariant set \( X' \subset X \) of full measure, \( T : X' \to X' \) is invertible, and \( T, T^{-1} : X' \to X' \) are measurable. Invertible mpt are more convenient to handle than general mpt because they have the following properties: If \( (X, \mathcal{B}, \mu) \) is a complete measure space, then the forward image of a measurable set is measurable, and for any countable collection of measurable sets \( A_i, T(\bigcap A_i) = \bigcap T(A_i) \) up to a set of measure zero. This is not true in general for non-invertible maps. Luckily, every “reasonable” non-invertible mpt is the factor of an invertible mpt. Here is the construction.

---

4 Such a \( \sigma \)-algebra exists: take the intersection of all sub-\( \sigma \)-algebras which make \( f \circ T^n \) all measurable, and note that this intersection is not empty because it contains \( \mathcal{B} \).
**Definition 1.17.** Suppose \((X, \mathcal{B}, \mu, T)\) is a pmt defined on a Lebesgue measure space, and assume \(T(X) = X\). The **natural extension** of \((X, \mathcal{B}, \mu, T)\) is the system \((\tilde{X}, \mathcal{B}, \tilde{\mu}, \tilde{T})\), where

1. \(\tilde{X} := \{ \tilde{x} = (\ldots, x_{-1}, x_0, x_1, x_2, \ldots) : x_i \in X, T(x_i) = x_{i+1} \text{ for all } i \} \);

2. \(\mathcal{B}\) is the smallest \(\sigma\)-algebra which contains all sets of the form \(\{ \tilde{x} \in \tilde{X} : x_i \in E \}\) with \(i \leq 0\) and \(E \in \mathcal{B}\) (see below);

3. \(\tilde{\mu}\) is the unique probability measure on \(\tilde{\mathcal{B}}\) such that \(\mu \{ \tilde{x} \in \tilde{X} : x_i \in E_i \} = \mu(E_i)\) for all \(i \leq 0\) and \(E_i \in T^{-i}\mathcal{B}\);

4. \(\tilde{T}\) is the left shift.

**Lemma 1.2.** The measure \(\tilde{\mu}\) exists and is unique.

**Proof (the proof can be omitted on first reading).** Let \(\mathcal{J}\) denote the collection of all sets of the form \([E_{-n}, \ldots, E_0] := \{ \tilde{x} \in \tilde{X} : x_{-i} \in E_{-i} \ (i = 0, \ldots, n) \}\) where \(n \geq 0\) and \(E_{-n}, \ldots, E_{-1}, E_0 \in \mathcal{B}\). We call the elements of \(\mathcal{J}\) cylinders.

It is easy to see that \(\mathcal{J}\) is a semi-algebra. Our plan is to define \(\tilde{\mu}\) on \(\mathcal{J}\) and then apply Carathéodory’s extension theorem. To do this we first observe the following important identity:

\[
[E_{-n}, \ldots, E_0] = \{ \tilde{x} \in \tilde{X} : x_{-n} \in \bigcap_{i=0}^{n} T^{-(n-i)} E_{-i} \}. \tag{1.3}
\]

The inclusion \(\subseteq\) is because for every \(\tilde{x} \in [E_{-n}, \ldots, E_0]\), \(x_{-n} \equiv T^{-(n-i)} x_{-i} \in T^{-(n-i)} E_{-i}\) by the definition of \(\tilde{X}\); and \(\supseteq\) is because if \(x_{-n} \in T^{-(n-i)} E_{-i}\) then \(x_{-i} \equiv T^{(n-i)} (x_{-n}) \in E_{-i}\) for \(i = 0, \ldots, n\). Motivated by this identity we define \(\tilde{\mu} : \mathcal{J} \to [0, \infty]\) by

\[
\tilde{\mu}[E_{-n}, \ldots, E_0] := \mu \left( \bigcap_{i=0}^{n} T^{-(n-i)} E_{-i} \right).
\]

Since \(\tilde{\mu}(\tilde{X}) = \mu(X) = 1\), \(\tilde{\mu}\) is \(\sigma\)-finite on \(\tilde{X}\). We will check that \(\tilde{\mu}\) is \(\sigma\)-additive on \(\mathcal{J}\) and deduce the lemma from Carathéodory’s extension theorem. We begin with simpler finite statements.

**Step 1.** Suppose \(C_1, \ldots, C_n\) are pairwise disjoint cylinders and \(D_1, \ldots, D_n\) are pairwise disjoint cylinders. If \(\bigcup_{i=1}^{n} C_i = \bigcup_{i=1}^{m} D_i\), then \(\sum \tilde{\mu}(C_i) = \sum \tilde{\mu}(D_i)\).

**Proof.** Notice that \([E_{-n}, \ldots, E_0] \equiv [X, \ldots, X, E_{-n}, \ldots, E_0]\) and (by the \(T\)-invariance of \(\mu\)) \(\tilde{\mu}[E_{-n}, \ldots, E_0] = \tilde{\mu}[X, \ldots, X, E_{-n}, \ldots, E_0]\), no matter how many \(X\)’s we add to the left. So there is no loss of generality in assuming that \(C_i, D_i\) all have the same length: \(C_i = [C_{-n}^{(i)}, \ldots, C_0^{(i)}], D_j = [D_{-n}^{(j)}, \ldots, D_0^{(j)}]\).

By (1.3), since \(C_i\) are pairwise disjoint, \(\bigcap_{k=0}^{n} T^{-(n-k)} C_{-k}^{(i)}\) are pairwise disjoint. Similarly, \(\bigcap_{k=0}^{n} T^{-(n-k)} D_{-k}^{(j)}\) are pairwise disjoint. Since \(\bigcup_{i=1}^{n} C_i = \bigcup_{j=1}^{m} D_j\), the identity (1.3) also implies that \(\bigcup_{i=1}^{n} \bigcap_{k=0}^{n} T^{-(n-k)} C_{-k}^{(i)} = \bigcup_{j=1}^{m} \bigcap_{k=0}^{n} T^{-(n-k)} D_{-k}^{(j)}\). So
\[ \sum_{i=1}^{\alpha} \tilde{\mu}(C_i) = \sum_{i=1}^{\alpha} \mu \left( \bigcap_{k=0}^{n} T^{-(n-k)} C_{i}^{(k)} \right) = \sum_{j=1}^{\beta} \mu \left( \bigcap_{k=0}^{n} T^{-(n-k)} D_{j}^{(k)} \right) = \sum_{j=1}^{\alpha} \tilde{\mu}(D_j), \]

which proves our claim.

**Step 2. Suppose** \( C, C_1, \ldots, C_n \) **are cylinders. If** \( C \subseteq \bigcup_{i=1}^{n} C_i \), **then** \( \tilde{\mu}(C) \leq \sum_{i=1}^{n} \tilde{\mu}(C_i) \).

**Proof.** Notice that \( \bigcup_{i=1}^{n} C_i = \bigcup_{i=1}^{n} (C_i \cap \bigcap_{j=1}^{n-1} C_j) \). Using the fact that \( \mathcal{S} \) is a semi-algebra, it is not difficult to see that \( C_i \cap \bigcap_{j=1}^{n-1} C_j \) is a finite pairwise disjoint union of cylinders: \( C_i \cap \bigcap_{j=1}^{n-1} C_j = \bigcup_{k=1}^{n} C_{i,k} \). So

\[ \bigcup_{i=1}^{n} C_i = \bigcup_{i=1}^{n} \bigcup_{k=1}^{n} C_{i,k} \text{ where } \bigcup_{k=1}^{n} C_{i,k} \subseteq C_i. \]

Look at the inclusion \( C \subseteq \bigcup_{i=1}^{n} C_i = \bigcup_{i=1}^{n} \bigcup_{k=1}^{n} C_{i,k} \). The set difference of the two sides of the equation is a finite pairwise disjoint union of cylinders because \( \mathcal{S} \) is a semi-algebra. So by step 1 \( \tilde{\mu}(C) \leq \sum_{i=1}^{n} \tilde{\mu}(C_i) \).

Similarly, \( \bigcup_{i=1}^{n} C_{i,k} \subseteq C_i \) and the set difference of the two sides of the inclusion is a finite pairwise disjoint union of cylinders. So by step 1, \( \sum_{j=1}^{n} \tilde{\mu}(C_{i,k}) \leq \tilde{\mu}(C_i) \). In summary \( \tilde{\mu}(C) \leq \sum \tilde{\mu}(C_i) \).

We are finally ready to prove \( \sigma \)-additivity. Suppose \( C_k = [E_{-(n-k)}^{(k)}, \ldots, E_0^{(k)}] \) are a countable collection of pairwise disjoint cylinders such that \( \bigcup_{k=1}^{n} C_k = C = [E_{-n}, \ldots, E_0] \). Our aim is to show that \( \mu(C) = \sum \mu(C_k) \).

**Step 3.** \( \mu(C) \geq \sum \mu(C_k) \)

**Proof.** Without loss of generality \( n_k \geq n \) for all \( k \), otherwise we can replace \( C_k \) by the equal cylinder \( [X, \ldots, X, E_{-(n)}^{(k)}, \ldots, E_0^{(k)}] \) with \( X \) repeated \( n - n_k - 1 \) times.

For every \( K \), let \( N_K := \max \{ n_k : 1 \leq k \leq K \} \), then \( N_K \geq n \). Define \( E_{-m} := X \) for \( m > n \) and \( E_{-(n-k)}^{(k)} := X \) for \( m > n \) and \( n_k \). If \( m > N_K \) and \( 1 \leq k \leq K \) then \( C = [E_{-m}, \ldots, E_0] \) and \( C_k = [E_{-(n-k)}^{(k)}, \ldots, E_0^{(k)}] \). So \( \bigcup_{m=1}^{n} \bigcup_{k=1}^{K} [E_{-(n-k)}^{(k)}, \ldots, E_0^{(k)}] \).

It now follows from (1.3) that \( \bigcap_{i=0}^{m} T^{-i} E_{-i} \supseteq \bigcup_{k=1}^{K} \bigcap_{i=0}^{m} T^{-i} E_{-i}^{(k)} \), whence

\[ \mu \left( \bigcap_{i=0}^{m} T^{-i} E_{-i} \right) \geq \mu \left( \bigcup_{k=1}^{K} \bigcap_{i=0}^{m} T^{-i} E_{-i}^{(k)} \right) \]

Since \( \mu \) is \( \sigma \)-additive on \( \mathcal{B} \) and \( \mu \circ T^{-1} = \mu \), the left-hand-side equals \( \tilde{\mu}(C) \) and the right-hand-side equals \( \sum_{k=1}^{K} \tilde{\mu}(C_k) \).

**Step 2.** \( \mu(C) \leq \sum \mu(C_k) \)

**Proof.** It is here that we use the assumption that \( (X, \mathcal{B}, \mu) \) is a Lebesgue measure space. The assumption allows us to assume without loss of generality that \( X \) is a compact metric space and \( \mathcal{B} \) is the completion of the Borel \( \sigma \)-algebra (because any union of an interval and a countable collection of atoms can be isomorphically embedded in such a space). In this case \( \mu \) is regular: for every \( E \in \mathcal{B} \) and for every \( \varepsilon > 0 \) there is an open set \( U \) and a compact set \( F \) such that \( F \subseteq E \subseteq U \) and \( \mu(U \setminus F), \mu(U \setminus E) < \varepsilon \).
In particular, given \( \epsilon > 0 \) there is no problem in finding a compact set \( F \subset \bigcap_{i=0}^{n} T^{-(n-i)} E_{-i} \) and open sets \( U_k \supset \bigcap_{i=0}^{n_k} T^{-(n_k-i)} E_{-i}^{(k)} \) open so that

\[
\mu(F) \geq \widetilde{\mu}(C) - \epsilon \quad \text{and} \quad \mu(U_k) \leq \widetilde{\mu}(C_k) + \frac{\epsilon}{2k}.
\]

By (1.3), \([F, X, \ldots, X] \subset C = \bigcup_{k=1}^{n} C_k \subset \bigcup_{k=1}^{n_k} [U_k, X, \ldots, X] \), and with respect to the product topology on \( \tilde{X} \), \([F, X, \ldots, X] \) is compact and \([U_k, X, \ldots, X] \) are open. So there is a finite \( N \) s.t. \([F, X, \ldots, X] \subset \bigcup_{k=1}^{N} [U_k, X, \ldots, X] \). By step 2

\[
\widetilde{\mu}[F, X, \ldots, X] \leq \sum_{k=1}^{N} \widetilde{\mu}[U_k, X, \ldots, X]
\]

So \( \mu(F) \leq \sum_{k=1}^{N} \mu(U_k) \), whence \( \widetilde{\mu}(C) \leq \sum_{k=1}^{N} \widetilde{\mu}(C_k) + 2\epsilon \leq \sum_{k=1}^{N} \mu(C_k) + 2\epsilon \). The step follows by taking \( \epsilon \to 0 \).

**Theorem 1.6.** The natural extension of \((X, \mathcal{B}, \mu, T)\) is an invertible extension of \((X, \mathcal{B}, \mu, T)\), and is the factor of any other invertible extension of \((X, \mathcal{B}, \mu, T)\). \( \tilde{T} \) is ergodic iff \( T \) is ergodic, and \( \tilde{T} \) is mixing iff \( T \) is mixing.

**Proof.** It is clear that the natural extension is an invertible ppt. Let \( \pi : \tilde{X} \to X \) denote the map \( \pi(x) = x_0 \), then \( \pi \) is measurable and \( \pi \circ \tilde{T} = T \circ \pi. \) Since \( T \xi = \xi \), every point has a pre-image, and so \( \pi \) is onto. Finally, for every \( E \in \mathcal{B} \),

\[
\mu(\pi^{-1}(E)) = \tilde{\mu}(\{\tilde{x} \in \tilde{X} : \pi(\tilde{x}) \in E\}) = \mu(E)
\]

by construction. So \((\tilde{X}, \tilde{\mathcal{B}}, \tilde{\mu}, \tilde{T})\) is an invertible extension of \((X, \mathcal{B}, \mu, T)\).

Suppose \((Y, \mathcal{C}, \nu, S)\) is another invertible extension, and let \( \pi_Y : Y \to X \) be the factor map (defined a.e. on \( Y \)). We show that \((Y, \mathcal{C}, \nu, S)\) extends \((\tilde{X}, \tilde{\mathcal{B}}, \tilde{\mu}, \tilde{T})\).

Let \((\tilde{Y}, \mathcal{C}^\prime, \tilde{\nu}, \tilde{S})\) be the natural extension of \((Y, \mathcal{C}, \nu, S)\). It is isomorphic to \((Y, \mathcal{C}, \nu, S)\), with the isomorphism given by \( \tilde{\vartheta}(y) = (y_k)_{k \in \mathbb{Z}}, y_k := S^k(y) \). Thus it is enough to show that \((\tilde{X}, \tilde{\mathcal{B}}, \tilde{\mu}, \tilde{T})\) is a factor of \((\tilde{Y}, \mathcal{C}^\prime, \tilde{\nu}, \tilde{T})\). Here is the factor map: \( \vartheta : (y_k)_{k \in \mathbb{Z}} \mapsto (\pi_Y(y_k))_{k \in \mathbb{Z}} \).

If \( \tilde{T} \) is ergodic, then \( T \) is ergodic, because every \( T \)-invariant set \( E \) lifts to a \( \tilde{T} \)-invariant set \( \tilde{E} := \pi^{-1}(E) \). The ergodicity of \( \tilde{T} \) implies that \( \tilde{\mu}(\tilde{E}) = 0 \) or 1, whence \( \mu(E) = \tilde{\mu}(\pi^{-1}(E)) = \tilde{\mu}(\tilde{E}) = 0 \) or 1.

To see the converse (\( T \) is ergodic \( \Rightarrow \) \( \tilde{T} \) is ergodic) we make use of the following observation:

**Claim:** Let \( \overline{\mathcal{B}} := \{ \{\tilde{x} \in \tilde{X} : \pi(\tilde{x}) \in E\} : E \in \mathcal{B} \} \), then

1. \( \mathcal{B} \) are \( \sigma \)-algebras
2. \( \mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots \) and \( \cup_{n \geq 0} \mathcal{B}_n \) generate \( \mathcal{B} \)
3. \( \overline{T}^{-1}(\mathcal{B}_n) \subset \mathcal{B}_n \) and \((\tilde{X}, \mathcal{B}_n, \tilde{\mu}, \tilde{T})\) is a ppt
4. if \( T \) is ergodic then \((\tilde{X}, \mathcal{B}_n, \tilde{\mu}, \tilde{T})\) is ergodic.

Proof. We leave the first three items as exercises to the reader. To see the last item suppose \( T \) is ergodic and \( E \in \mathcal{B}_n \) is \( \overline{T} \)-invariant. By the definition of \( \mathcal{B}_n \), \( E = \{ \tilde{x} : \tilde{x}_{-n} \in E \} \) with \( E \in \mathcal{B} \), and it is not difficult to see that \( E \) must be \( T \)-invariant. Since \( T \) is ergodic, \( \mu(E) = 0 \) or 1. So \( \tilde{\mu}(E) = \mu(E) = 0 \) or 1. So \((\tilde{X}, \mathcal{B}_n, \tilde{\mu}, \tilde{T})\) is ergodic.

We can now prove the ergodicity of \((\tilde{X}, \mathcal{B}, \tilde{\mu}, \tilde{T})\) as follows. Suppose \( \tilde{f} \) is absolutely integrable and \( \overline{T} \)-invariant, and let

\[
\tilde{f}_n := E(f|\mathcal{B}_n)
\]

(readers who are not familiar with conditional expectations can find their definition in section 2.3.1).

We claim that \( \tilde{f}_n \circ \tilde{T} = \tilde{f}_n \). This is because for every bounded \( \mathcal{B}_n \)-measurable test function \( \phi \),

\[
\int [\phi \cdot \tilde{f}_n \circ \overline{T}^{-1}]d\tilde{\mu} \overset{(1)}{=} \int [\phi \circ \tilde{T}^{-1}]d\tilde{T} \overset{(2)}{=} \int [\phi \circ \tilde{T} \cdot \tilde{f}_n]d\tilde{\mu} \overset{(3)}{=} \int [\phi \circ \tilde{T} \cdot \tilde{f} \cdot \tilde{T}^{-1}]d\tilde{\mu} \overset{(4)}{=} \int \phi \tilde{f} d\tilde{\mu} \overset{(5)}{=} \int \phi \tilde{f}_n d\tilde{\mu}.
\]

Justifications: (1) is because \( \tilde{\mu} \circ \overline{T}^{-1} = \tilde{\mu} \); (2) is because \( \tilde{f}_n = E(f|\mathcal{B}_n) \) and \( \phi \circ \overline{T} \) is \( \mathcal{B}_n \)-measurable by part (3) of the claim; (3) is because \( f \circ \overline{T} = f \); (4) is because \( \tilde{\mu} \circ \overline{T}^{-1} = \tilde{\mu} \); and (5) is by the definition of the conditional expectation.

In summary \( \int [\phi \cdot \tilde{f}_n \circ \overline{T}^{-1}]d\tilde{\mu} = \int \phi \tilde{f}_n d\tilde{\mu} \) for all bounded \( \mathcal{B}_n \)-measurable functions, whence \( \tilde{f}_n \circ \tilde{T} = \tilde{f}_n \) a.e. as claimed.

We saw above that \((\tilde{X}, \mathcal{B}_n, \tilde{\mu}, \tilde{T})\) is ergodic. Therefore, the invariance \( \mathcal{B}_n \)-measurability of \( \tilde{f}_n \) imply that \( \tilde{f}_n = \int \tilde{f} d\tilde{\mu} \) almost everywhere. By the martingale convergence theorem,

\[
\tilde{f}_n \overset{n \to \infty}{\longrightarrow} \tilde{f} \text{ almost everywhere.}
\]

So \( \tilde{f} \) is constant almost everywhere. This shows that every \( \overline{T} \)-invariant integrable function is constant a.e., so \( \tilde{T} \) is ergodic.

Next we consider the mixing of \( T, \overline{T} \). \( \overline{T} \) mixing \( \Rightarrow \) \( T \) mixing because for every \( A, B \in \mathcal{B}, \mu(A \cap T^{-n}B) = \mu(\pi^{-1}A \cap T^{-n} \pi^{-1}B) \overset{n \to \infty}{\longrightarrow} \mu(\pi^{-1}A)\mu(\pi^{-1}B) \equiv \mu(A)\mu(B) \). We show the other direction. Suppose \( T \) is mixing and consider \( \tilde{A}, \tilde{B} \in \mathcal{B}_n \) with \( \mathcal{B}_n \) defined as before. Then \( \tilde{A} = \{ \tilde{x} : x_{-n} \in A \} \) and \( \tilde{B} = \{ \tilde{x} : x_{-n} \in B \} \) with \( A, B \in \mathcal{B} \) and using the identity \( (T\tilde{x})_i = T(x_i) \) it is easy to see that

\[
\tilde{A} \cap \overline{T}^{-k} \tilde{B} = \{ \tilde{x} \in \tilde{X} : \tilde{x}_{-n} \in A \cap \overline{T}^{-k}B \}.
\]

So \( \tilde{\mu}(\tilde{A} \cap \overline{T}^{-k} \tilde{B}) = \mu(A \cap \overline{T}^{-k}B) \overset{k \to \infty}{\longrightarrow} \mu(A)\mu(B) \equiv \tilde{\mu}(\tilde{A})\tilde{\mu}(\tilde{B}). \)
This proves mixing for $\tilde{A}, \tilde{B} \in \mathcal{B}_n$. To get mixing for general $\tilde{A}, \tilde{B} \in \mathcal{B}$ we note $\bigcup \mathcal{B}_n$ generates $\mathcal{B}$ so for any $E$ and for every $\varepsilon > 0$ there are $n$ and $\tilde{E}' \in \mathcal{B}_n$ such that $\mu(\tilde{E} \triangle \tilde{E}') < \varepsilon$ (because the collection of sets $\tilde{E}$ with such $n$ and $\tilde{E}'$ forms a $\sigma$-algebra which contains $\bigcup \mathcal{B}_n$). A standard approximation argument now shows that $\tilde{\mu}(A \cap \tilde{T}^{-k}B) \xrightarrow[k \to \infty]{} \tilde{\mu}(A)\tilde{\mu}(B)$ for all $A, \tilde{B} \in \mathcal{B}$.

\subsection{1.6.4 Induced transformations}

Suppose $(X, \mathcal{B}, \mu, T)$ is a probability preserving transformation, and let $A \in \mathcal{B}$ be a set of positive measure. By Poincaré’s Recurrence Theorem, for a.e. $x \in A$ there is some $n \geq 1$ such that $T^n(x) \in A$. Define

$$\varphi_A(x) := \min \{n \geq 1 : T^n x \in A\},$$

with the minimum of the empty set being understood as infinity. Note that $\varphi_A < \infty$ a.e. on $A$, hence $A_0 := \{x \in A : \varphi_A(x) < \infty\}$ is equal to $A$ up to a set of measure zero.

\textbf{Definition 1.18}. The \textit{induced transformation} on $A$ is $(A_0, T(A), \mu_A, T_A)$, where $A_0 := \{x \in A : \varphi_A(x) < \infty\}$, $T(A) := \{E \cap A_0 : E \in \mathcal{B}\}$, $\mu_A$ is the measure $\mu_A(E) := \mu(E|A) = \mu(E \cap A)/\mu(A)$, and $T_A : A_0 \to A_0$ is $T_A(x) = T^{\varphi_A(x)}(x)$.

\textbf{Theorem 1.7}. Suppose $(X, \mathcal{B}, \mu, T)$ is a ppt, and $A \in \mathcal{B}$ has positive finite measure.

1. $\mu_A \circ T_A^{-1} = \mu_A$;
2. if $T$ is ergodic, then $T_A$ is ergodic (but the mixing of $T \neq$ the mixing of $T_A$);
3. \textbf{Kac Formula}: If $\mu$ is ergodic, then $\int f d\mu = \int_A \sum_{k=0}^{\infty} f \circ T^k d\mu$ for every $f \in L^1(X)$. In particular $\int_A \varphi_A d\mu_A = 1/\mu(A)$.

\textbf{Proof}. Given $E \subset A$ measurable, $\mu(E) = \mu(T^{-1}E \cap A) + \mu(T^{-1}E \cap A^c) =
\begin{align*}
\mu(T^{-1}E \cap A) + \mu(T^{-2}E \cap T^{-1}A^c \cap A) + \mu(T^{-2}E \cap T^{-1}A^c \cap A^c)
\end{align*}
\begin{align*}
\mu(T^{-1}E \cap \{\varphi_A = 1\}) + \mu(T^{-1}E \cap \{\varphi_A = 2\})
\end{align*}
\begin{align*}
\cdots + \sum_{j=1}^{N} \mu(T^{-1}E \cap \{\varphi_A = j\}) + \mu(T^{-N}E \cap \bigcap_{j=0}^{N-1} T^{-j}A^c).
\end{align*}

Passing to the limit as $N \to \infty$, we see that $\mu(E) \geq \mu_A(T_A^{-1}E)$. Working with $A \setminus E$, and using the assumption that $\mu(X) < \infty$, we get that $\mu(A) - \mu(E) \leq \mu(A) - \mu(T_A^{-1}E)$ whence $\mu(E) = \mu(T_A^{-1}E)$. Since $\mu_A$ is proportional to $\mu$ on $\mathcal{B}(A)$, we get $\mu_A = \mu_A \circ T_A^{-1}$.

We assume that $T$ is ergodic, and prove that $T_A$ is ergodic. The set

$$\Omega := \{x : T^n(x) \in A \text{ for infinitely many } n \geq 0\}$$

is a T-invariant set of non-zero measure (bounded below by \( \mu(A) \)), so it must have full measure. Thus a.e. \( x \in X \) has some \( n \geq 0 \) s.t. \( T^n(x) \in A \), and

\[
\rho_A(x) := \min\{n \geq 0 : T^n x \in A\} < \infty \text{ a.e. in } X.
\]

Suppose \( f : A_0 \to \mathbb{R} \) is a \( T_A \)-invariant \( L^2 \)-function. Define

\[
F(x) := f(T^{\rho_A(x)} x).
\]

This makes sense a.e. in \( X \), because \( \rho_A < \infty \) almost everywhere. This function is \( T \)-invariant, because either \( x, Tx \in A \) and then \( F(Tx) = f(Tx) = f(T_A x) = f(x) = F(x) \) or one of \( x, Tx \) is outside \( A \) and then \( F(Tx) = f(T^{\rho_A(Tx)} x) = f(T^{\rho_A(x)} x) = F(x) \).

Since \( T \) is ergodic, \( F \) is constant a.e. on \( X \), and therefore \( f = F|_A \) is constant a.e. on \( A \). Thus the ergodicity of \( T \) implies the ergodicity of \( T_A \).

Here is an example showing that the mixing of \( T \) does not imply the mixing of \( T_A \). Let \( \Sigma^+ \) be a SFT with states \( \{a, 1, 2, b\} \) and allowed transitions

\[
a \to 1; 1 \to 1, b; b \to 2; 2 \to a.
\]

Let \( A = \{x : x_0 = a, b\} \). Any shift invariant Markov measure \( \mu \) on \( \Sigma^+ \) is mixing, because \( \Sigma^+ \) is irreducible and aperiodic (1 \( \to 1 \)). But \( T_A \) is not mixing, because \( T_A[a] = [b] \) and \( T_A[b] = [a] \), so \( [a] \cap T_A^{-n}[a] = \emptyset \) for all odd \( n \).

Next we prove the Kac formula. Suppose first that \( f \in L^\infty(X, \mathcal{B}, \mu) \) and \( f \geq 0 \).

\[
\int f d\mu = \int_A f d\mu + \int f \cdot 1_{A^c} d\mu = \int_A f d\mu + \int f \circ T \cdot 1_{T^{-1}A^c} d\mu
\]

\[
= \int_A f d\mu + \int f \circ T \cdot 1_{T^{-1}A^c \cap A} d\mu + \int f \circ T \cdot 1_{T^{-1}A \cap A^c} d\mu
\]

\[
= \int_A f d\mu + \int_A \int f \circ T \cdot 1_{\{\phi_A > 1\}} d\mu + \int f \circ T^2 \cdot 1_{T^{-2}A \cap T^{-1}A^c} d\mu
\]

\[
= \cdots = \int_A \sum_{N=0}^{N-1} f \circ T^j \cdot 1_{\{\phi_A > j\}} d\mu + \int f \circ T^N \cdot 1_{\{\phi_A > N\}} d\mu.
\]

The first term tends, as \( N \to \infty \), to

\[
\int_A \sum_{j=0}^{\infty} f \circ T^j \cdot 1_{\{\phi_A = j\}} d\mu \equiv \int_A \sum_{j=0}^{\phi_A-1} f \circ T^j d\mu.
\]

The second term is bounded by \( \|f\|_{L^\infty} \mu \{x : T^j(x) \notin A \text{ for all } k \leq N\} \). This bound tends to zero, because \( \mu \{x : T^j(x) \notin A \text{ for all } k = 0 \} = 0 \) because \( T \) is ergodic and recurrent (fill in the details). This proves the Kac formula for all \( L^\infty \) functions.

Every non-negative \( L^1 \)-function is the increasing limit of \( L^\infty \) functions. By the monotone convergence theorem, the Kac formula must hold for all non-negative \( L^1 \)-function. Every \( L^1 \)-function is the difference of two non-negative \( L^1 \)-functions \((f = f \cdot 1_{[f > 0]} - f \cdot 1_{[f < 0]})\). It follows that the Kac formula holds for all \( f \in L^1 \). □
1.6.5 Suspensions and Kakutani skyscrapers

The operation of inducing can be “inverted”, as follows. Let \((X, \mathcal{B}, \mu, T)\) be a ppt, and \(r : X \to \mathbb{N}\) an integrable measurable function.

**Definition 1.19.** The Kakutani skyscraper with base \((X, \mathcal{B}, \mu, T)\) and height function \(r\) is the system \((X_r, \mathcal{B}(X_r), \nu, S)\), where

1. \(X_r := \{(x, n) : x \in X, 0 \leq n \leq r(x) - 1\}\);
2. \(\mathcal{B}(X_r) = \{E \in \mathcal{B}(X) \otimes \mathcal{B}(\mathbb{N}) : E \subseteq X_r\}\), where \(\mathcal{B}(\mathbb{N}) = 2^\mathbb{N}\);
3. \(\nu\) is the unique measure such that \(\nu(B \times \{k\}) = \mu(B) / \int r d\mu\);
4. \(S\) is defined by \(S(x, n) = (x, n+1)\), when \(n < r(x) - 1\), and \(S(x, n) = (T^nx, 0)\), when \(n = r(x) - 1\).

(Check that this is a ppt.)

We think of \(X_r\) as a skyscraper made of stories \(\{(x, k) : r(x) > k\}\); the orbits of \(S\) climb up the skyscraper until they reach the top floor possible, and then move to the ground floor according to \(T\).

If we induce a Kakutani skyscraper on \(\{(x, 0) : x \in X\}\), we get a system which is isomorphic to \((X, \mathcal{B}, \mu, T)\).

**Proposition 1.11.** A Kakutani skyscraper over an ergodic base is ergodic, but there are non-mixing skyscrapers over mixing bases.

The proof is left as an exercise.

There is a straightforward important continuous–time version of this construction: Suppose \((X, \mathcal{B}, \mu, T)\) is a ppt, and \(r : X \to \mathbb{R}^+\) is a measurable function such that \(\inf r > 0\).

**Definition 1.20.** The suspension semi-flow with base \((X, \mathcal{B}, \mu, T)\) and height function \(r\) is the semi-flow \((X_r, \mathcal{B}(X_r), \nu, T_r)\), where

1. \(X_r := \{(x, t) \in X \times \mathbb{R} : 0 \leq t < r(x)\}\);
2. \(\mathcal{B}(X_r) = \{E \in \mathcal{B}(X) \otimes \mathcal{B}(\mathbb{R}) : E \subseteq X_r\}\);
3. \(\nu\) is the measure such that \(\int_X f d\nu = \int_X \int_0^{r(x)} f(x, t) dt\mu(x) / \int_X r d\mu\);
4. \(T_r(x, t) = (T^n x, t + s - \sum_{k=0}^{n-1} r(T^k x))\), where \(n\) is s.t. \(0 \leq t + s - \sum_{k=0}^{n-1} r(T^k x) < r(T^n x)\).

(And that this is a measure preserving semi-flow.)

Suspension flows appear in applications in the following way. Imagine a flow \(T\) on a manifold \(X\). It is often possible to find a submanifold \(S \subset X\) such that (almost) every orbit of the flow intersects \(S\) transversally infinitely many times. Such a submanifold is called a Poincaré section. If it exists, then one can define a map \(T_S : S \to S\) which maps \(x \in S\) into \(T_S x\) with \(t := \min \{s > 0 : T_s x \in S\}\). This map is called the Section map. The flow itself is isomorphic to a suspension flow over its Poincaré section.
1.6 Basic constructions

Problems

1.1. Proof of Liouville’s theorem in section 1.1
(a) Write \( x:=(q,p) \) and \( y:=T_t(q,p) \). Use Hamilton’s equations to show that the Jacobian matrix of \( y=\psi(x) \) satisfies \( \frac{\partial y}{\partial x} = I + tA + O(t^2) \) as \( t\to 0 \), where \( \text{tr}(A) = 0 \).
(b) Show that for every matrix \( A, \det(I+tA+O(t^2)) = 1 + t\text{tr}(A) + O(t^2) \) as \( t\to 0 \).
(c) Prove that the Jacobian of \( T_t \) is equal to one for all \( t \). Deduce Liouville’s theorem.

1.2. The completion of a measure space. Suppose \((X,\mathcal{B},\mu)\) is a measure space. A set \( N \subset X \) is called a null set, if there is a measurable set \( E \supseteq N \) such that \( \mu(E) = 0 \).
A measure space is called complete, if every null set is measurable. Every measure space can be completed, and this exercise shows how to do this.
(a) Let \( \mathcal{B}_0 \) denote the the collection of all sets of the form \( E \cup N \) where \( E \in \mathcal{B} \) and \( N \) is a null set. Show that \( \mathcal{B}_0 \) is a \( \sigma \)–algebra.
(b) Show that \( \mu \) has a unique extension to a \( \sigma \)–additive measure on \( \mathcal{B}_0 \).

1.3. Prove Poincaré’s Recurrence Theorem for a general probability preserving transformation (theorem 1.1).

1.4. Fill in the details in the proof above that the Markov measure corresponding to a stationary probability vector and a stochastic matrix exists, and is shift invariant measure.

1.5. Suppose \( \Sigma^+ \) is a SFT with stochastic matrix \( P \). Let \( A = (t_{ab})_{S \times S} \) denote the matrix of zeroes and ones where \( t_{ab} = 1 \) if \( p_{ab} > 0 \) and \( t_{ab} = 0 \) otherwise. Write \( A^n = (t_{ab}^{(n)}) \). Prove that \( t_{ab}^{(n)} \) is the number of paths of length \( n \) starting at \( a \) and ending at \( b \). In particular: \( a \rightarrow b \iff t_{ab}^{(n)} > 0 \).

1.6. The Perron–Frobenius Theorem\(^5\): Suppose \( A = (a_{ij}) \) is a matrix all of whose entries are non-negative, and let \( B := (b_{ij}) \) be the matrix \( b_{ij} = 1 \) if \( a_{ij} > 0 \) and \( b_{ij} = 0 \) if \( a_{ij} = 0 \). Assume that \( B \) is irreducible, then \( A \) has a positive eigenvalue \( \lambda \) with the following properties:
(i) There are positive vectors \( \mathbf{r} \) and \( \mathbf{\ell} \) s.t. \( \ell A = \lambda \mathbf{\ell}, A \mathbf{r} = \lambda \mathbf{r} \).
(ii) The eigenvalue \( \lambda \) is simple.
(iii) The spectrum of \( \lambda^{-1}A \) consists of 1, several (possibly zero) roots of unity, and a finite subset of the open unit disc. In this case the limit \( \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \lambda^{-k}A^k \) exists.
(iv) If \( B \) is irreducible and aperiodic, then the spectrum of \( \lambda^{-1}A \) consists of 1 and a finite subset of the open unit disc. In this case the limit \( \lim_{n\to\infty} \lambda^{-n}A^n \) exists.

1. Prove the Perron–Frobenius theorem in case \( A \) is stochastic, first in the aperiodic case, then in the general case.
2. Now consider the case of a non-negative matrix:

---

\(^5\) Perron first proved this in the aperiodic case. Frobenius later treated the periodic irreducible case.
a. Use a fixed point theorem to show that \( \lambda, \ell, r \) exist;
b. Set \( 1 := (1, \ldots, 1) \) and let \( V \) be the diagonal matrix such that \( V \ell = r \). Prove that \( \lambda^{-1} V^{-1} A V \) is stochastic.
c. Prove the Perron-Frobenius theorem.

1.7. Suppose \( P = (p_{ab}^{(n)})_{S \times S} \) is an irreducible aperiodic stochastic matrix. Use the spectral description of \( P \) obtained in problem 1.6 to show that \( p_{ab}^{(n)} \to p_b \) exponentially fast.

1.8. Show that the product of \( n \) irrational rotations \( R_{\alpha_1}, \ldots, R_{\alpha_n} \) is ergodic iff \( (\alpha_1, \ldots, \alpha_n) \) are independent over the irrationals.

1.9. Suppose \( g_t : X \to X \) is a measure preserving flow. The time one map of the flow is the measure preserving map \( g_1 : X \to X \). Give an example of an ergodic flow whose time one map is not ergodic.

1.10. The adding machine
Let \( X = \{0, 1\}^\mathbb{N} \) equipped with the \( \sigma \)-algebra \( \mathcal{B} \) generated by the cylinders, and the Bernoulli \((\frac{1}{2}, \frac{1}{2})\)-measure \( \mu \). The \textit{adding machine} is the ppt \( (X, \mathcal{B}, \mu, T) \) defined by the rule \( T(1^n 0^*) = (0^n 1^*), T(1^*) = 0^* \). Prove that the adding machine is invertible and probability preserving. Show that \( T(x) = x \oplus (10^*) \) where \( \oplus \) is “addition with carry to the right”.

1.11. Prove proposition 1.11.

1.12. Show that a ppt \( (X, \mathcal{B}, \mu, T) \) is mixing whenever \( \mu(A \cap T^{-n} A) \to \mu(A)^2 \) for all \( A \in \mathcal{B} \). Guidance:
1. \( \int 1_A f \circ T^n d\mu \to \mu(A) \int f d\mu \) for all \( f \in \text{span}_{\mathbb{R}} \{1, 1_A \circ T, 1_A \circ T^2, \ldots\} \).
2. \( \int 1_A f \circ T^n d\mu \to \mu(A) \int f d\mu \) for all \( f \in L^2 \).
3. \( \int g f \circ T^n d\mu \to \int g d\mu \int f d\mu \) for all \( f, g \in L^2 \).

1.13. Show that a Kakutani skyscraper over an invertible transformation is invertible, and find a formula for its inverse.

1.14. Conservativity
Let \( (X, \mathcal{B}, \mu, T) \) be a measure preserving transformation on an infinite \( \sigma \)-finite, measure space.\(^6\) A set \( W \in \mathcal{B} \) is called \textit{wandering}, if \( \{T^{-n} W : n \geq 0\} \) are pairwise disjoint. A mpt is called \textit{conservative}, if every wandering set has measure zero.

1. Show that any ppt is conservative. Give an example of a non-conservative mpt on a \( \sigma \)-finite infinite measure space.
2. Show that Poincaré’s recurrence theorem extends to conservative mpt.

\(^6\) A measure space is called \( \sigma \)-finite, if its sample space is the countable union of finite measure sets.
3. Suppose \((X, \mathcal{B}, \mu, T)\) is a conservative ergodic mpt, and let \(A\) be a set of finite positive measure. Show that the induced transformation \(T_A : A \rightarrow A\) is well-defined a.e. on \(A\), and is an ergodic ppt.

4. Prove Kac formula for conservative ergodic transformations under the previous set of assumptions.

Notes for chapter 1

The standard references for measure theory are [7] and [4]. The standard references for ergodic theory of probability preserving transformations are [6] and [8]. The standard references for ergodic theory of mpt on infinite measure spaces is [1]. Our proof of the Perron-Frobenius theorem is taken from [3]. Kac’s formula has very simple proof when \(T\) is invertible. The proof we use has the merit that it works for non-invertible transformations, and that it extends to the (conservative) infinite measure setting. It is taken from [5]. The ergodicity of the geodesic flow was first proved by E. Hopf by other means. The short proof we gave is due to Gelfand & Fomin and is reproduced in [2].

References

2.1 The Mean Ergodic Theorem

**Theorem 2.1 (von Neumann’s Mean Ergodic Theorem).** Suppose \((X, \mathcal{B}, \mu, T)\) is a ppt. If \(f \in L^2\), then \(\frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k \xrightarrow{L^2} \bar{f}\) where \(\bar{f} \in L^2\) is invariant. If \(T\) is ergodic, then \(\bar{f} = \int f \, d\mu\).

**Proof.** Observe that since \(T\) is measure preserving, then \(\int f \circ T \, d\mu = \int f \, d\mu\) for every \(f \in L^1\), and \(\|f \circ T\|_2 = \|f\|_2\) for all \(f \in L^2\) (prove this, first for indicator functions, then for all \(L^2\)-functions).

Suppose \(f = g - g \circ T\) where \(g \in L^2\) (in this case we say that \(f\) is a coboundary with transfer function \(g \in L^2\)), then it is obvious that

\[
\left\| \frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k \right\|_2 = \frac{1}{N} \| g \circ T^n - g \|_2 \leq 2 \| g \|_2 / N \xrightarrow{N \to \infty} 0.
\]

Thus the theorem holds for all elements of \(\mathcal{C} := \{g - g \circ T : g \in L^2\}\).

We claim that the theorem holds for all elements of \(\overline{\mathcal{C}}\) (\(L^2\)-closure). Suppose \(f \in \overline{\mathcal{C}}\), then for every \(\varepsilon > 0\), there is an \(F \in \mathcal{C}\) s.t. \(\|f - F\|_2 < \varepsilon\). Choose \(N_0\) such that for every \(N > N_0\), \(\| \frac{1}{N} \sum_{k=0}^{N-1} F \circ T^k \|_2 < \varepsilon\), then for all \(N > N_0\)

\[
\left\| \frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k \right\|_2 \leq \left\| \frac{1}{N} \sum_{k=0}^{N-1} (f - F) \circ T^k \right\|_2 + \left\| \frac{1}{N} \sum_{k=0}^{N-1} F \circ T^k \right\|_2
\]

\[
\leq \frac{1}{N} \sum_{k=0}^{N-1} \| (f - F) \circ T^k \|_2 + \varepsilon < 2 \varepsilon.
\]

This shows that \(\left\| \frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k \right\|_2 \xrightarrow{N \to \infty} 0\).

Next we claim that \(\overline{\mathcal{C}} = \{\text{invariant functions}\}\). Suppose \(f \perp \overline{\mathcal{C}}\), then
\[ \|f - f \circ T\|_2^2 = (f - f \circ T, f - f \circ T) = \|f\|_2^2 - 2(f, f \circ T) + \|f \circ T\|_2^2 = 2\|f\|_2^2 - 2(f, f - (f - f \circ T)) = 2\|f\|_2^2 - 2\|f\|_2^2 = 0 \implies f = f \circ T \text{ a.e.} \]

So \( \mathcal{C} \perp \subseteq \{ \text{invariant functions} \} \). Conversely, if \( f \) is invariant then for every \( g \in L^2 \)
\[ \langle f, g - g \circ T \rangle = \langle f, g \rangle - \langle f, g \circ T \rangle = \langle f \circ T, g \rangle - \langle f, g \rangle = 0, \]
so \( f \perp \mathcal{C} \), whence \( f \perp \mathcal{C} \). In summary, \( \mathcal{F} = \mathcal{C} \oplus \{ \text{invariant functions} \} \).

We saw above that the MET holds for all elements of \( \mathcal{C} \) with zero limit, and holds for all invariant functions \( f \) with limit \( f \). Therefore the MET holds for all \( L^2 \)-functions, and the limit \( \mathcal{F} \) is the orthogonal projection of \( f \) on the space of invariant functions.

In particular \( \mathcal{F} \) is invariant. If \( T \) is ergodic, then \( \mathcal{F} \) is constant and \( \mathcal{F} = \int \mathcal{F} d\mu \) almost everywhere. Also, since \( \frac{1}{N} \sum_{n=1}^{N} f \circ T^n \rightarrow f \) in \( L^2 \), then
\[ \int f d\mu = \frac{1}{N} \sum_{n=1}^{N} \langle 1, f \circ T^n \rangle = \langle 1, \frac{1}{N} \sum_{n=1}^{N} f \circ T^n \rangle \rightarrow \langle 1, \mathcal{F} \rangle = \int \mathcal{F} d\mu, \]
so \( \int \mathcal{F} d\mu = \int f d\mu \) whence \( \mathcal{F} = \int f d\mu \) almost everywhere. \( \Box \)

**Remark 1.** The proof shows that the limit \( \mathcal{F} \) is the projection of \( f \) on the space of invariant functions.

**Remark 2.** The proof only uses the fact that \( U f = f \circ T \) is an isometry of \( L^2 \). In fact it works for all linear operators \( U : H \rightarrow H \) on separable Hilbert spaces s.t. \( \|U\| \leq 1 \), see problem 2.1.

**Remark 3.** If \( f_n \overset{L^2}{\longrightarrow} f \), then \( \langle f_n, g \rangle \overset{n \rightarrow \infty}{\longrightarrow} \langle f, g \rangle \) for all \( g \in L^2 \). Specializing to the case \( f_n = \frac{1}{n} \sum_{k=0}^{n-1} 1_B \circ T^k g = 1_A \) we obtain the following corollary of the MET:

**Corollary 2.1.** A ppt \( (X, \mathcal{B}, \mu, T) \) is ergodic iff for all \( A, B \in \mathcal{B} \),
\[ \frac{1}{n} \sum_{k=0}^{n-1} \mu(A \cap T^{-k}B) \overset{n \rightarrow \infty}{\longrightarrow} \mu(A) \mu(B). \]

So ergodicity is mixing “on the average.” We will return to this point when we discuss the definition of weak mixing in the next chapter.
2.2 The Pointwise Ergodic Theorem

Theorem 2.2 (Birkhoff’s Pointwise Ergodic Theorem). Let \((X, \mathcal{B}, \mu, T)\) be a p.p.t.
If \(f \in L^1\), then the limit \(\overline{f}(x) := \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f(T^k x)\) exists for a.e. \(x\), and \(\frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k \to \overline{f}\) in \(L^1\). The function \(\overline{f}\) is \(T\)-invariant, absolutely integrable, and \(\int \overline{f} \, d\mu = \int f \, d\mu\). If \(T\) is ergodic, then \(\overline{f} = \frac{1}{\mu} \int f \, d\mu\) almost everywhere.

Proof. Since every \(f \in L^1\) is the difference of two non-negative \(L^1\)-functions, there is no loss of generality in assuming that \(f \geq 0\). Define

\[ A_n(x) := \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x), \quad \overline{A}(x) := \limsup_{n \to \infty} A_n(x), \quad \underline{A}(x) := \liminf_{n \to \infty} A_n(x). \]

\(\overline{A}(x), \underline{A}(x)\) take values in \([0, \infty]\), are measurable, and are \(T\)-invariant, as can be easily checked by taking the limits on both sides of \(A_n(x) = \frac{n-1}{n} A_{n-1}(T x) + O\left(\frac{1}{n}\right)\).

**Step 1.** \(\int f \, d\mu \geq \int \overline{A} \, d\mu\).

Proof. Fix \(\varepsilon, M > 0\), and set \(\overline{A}_M(x) := \overline{A}(x) \wedge M = \min\{\overline{A}(x), M\}\) (an invariant function). The following function is well-defined and finite everywhere:

\[ \tau(x) := \min\{n > 0 : A_n(x) > \overline{A}_M(x) - \varepsilon\}. \]

For a given \(N\), we “color” the time interval \(0, 1, 2, \ldots, N - 1\) as follows:

- If \(\tau(T^0 x) > M\), color 0 red; If \(\tau(T^0 x) \leq M\) color the next \(\tau(x)\) times blue, and move to the first uncolored \(k\).
- If \(\tau(T^k x) > M\), color \(k\) red; Otherwise color the next \(\tau(T^k x)\) times blue, and move to the first uncolored \(k\).

Continue in this way until all times colored, or until \(\tau(T^k x) > N - k\).

This partitions \(\{0, 1, \ldots, N - 1\}\) into red segments, and (possibly consecutive) blue segments of length \(\leq M\), plus (perhaps) one last segment of length \(\leq M\). Note:

- If \(k\) is red, then \(T^k x \in [\tau > M]\), so
  \[ \sum_{\text{red } k\text{'s}} 1_{[\tau > M]}(T^k x) \geq \text{number of red } k\text{'s} \]
- The average of \(f\) on each blue segment of length \(\tau(T^k x)\) is larger than \(\overline{A}_M(T^k x) - \varepsilon = \overline{A}_M(x) - \varepsilon\). So for each blue segment
  \[ \sum_{k \in \text{blue segment}} f(T^k x) \geq \text{length of segment} \times (\overline{A}_M(x) - \varepsilon). \]

Summing over all segments:

\[ \sum_{\text{blue } k\text{'s}} f(T^k x) \geq (\overline{A}_M(x) - \varepsilon) \times \text{number of blue } k\text{'s}. \]

We can combine these two estimates as follows.
\[
\sum_{k=0}^{N-1} (f + \bar{A}_M 1_{[\tau > M]}) (T^k x) \geq \#(\text{blues}) \cdot (\bar{A}_M (x) - \varepsilon) + \#(\text{reds}) \cdot \bar{A}_M (x)
\]

\[
\geq \#(\text{blues and reds}) \cdot (\bar{A}_M (x) - \varepsilon) \geq (N - M) (\bar{A}_M (x) - \varepsilon).
\]

Next we divide by \(N\), integrate, and obtain from the \(T\)-invariance of \(\mu\) that

\[
\int f d\mu + \int_{[\tau > M]} \bar{A}_M d\mu \geq (1 - \frac{M}{N}) \left( \int \bar{A}_M d\mu - \varepsilon \right) \geq \int \bar{A}_M d\mu - \frac{M^2}{N} - \varepsilon,
\]

where \(\frac{1}{M} \geq 0\). Subtracting from both sides of the inequality the (finite) quantity \(\int_{[\tau > M]} \bar{A}_M d\mu\), we find that

\[
\int f d\mu \geq \int_{[\tau \leq M]} \bar{A}_M d\mu - \frac{M^2}{N} - \varepsilon.
\]

We now take the limit \(N \to \infty\), then the limit \(M \to \infty\) (using the monotone convergence theorem and noting that \(1_{[\tau \leq M] \bar{A}_M} \uparrow \bar{A}\) as \(M \to \infty\)), and finally the limit \(\varepsilon \to 0\). The result is \(\int f d\mu \geq \int \bar{A} d\mu\).

**STEP 2.** \(\int f d\mu \leq \int \bar{A} d\mu\)

**Proof.** First observe, using Fatou’s Lemma, that \(\int \bar{A} d\mu \leq \|f\|_1 < \infty\). So \(\bar{A}(x) < \infty\) almost everywhere, and therefore, for every \(\varepsilon > 0\),

\[
\theta (x) := \min \{ n > 0 : A_n (x) < A(x) + \varepsilon \}
\]

is well-defined and finite for a.e. \(x\). We now repeat the coloring argument of step 1, with \(\theta\) replacing \(\tau\) and \(\bar{A}\) replacing \(\bar{A}_M\). Let \(f_M (x) := f(x) \wedge M\), then as before

\[
\sum_{k=0}^{N-1} f_M (T^k x) 1_{[\theta \leq M]} (T^k x) \leq \sum_{k \text{ blue}} (A(x) + \varepsilon) + \sum_{k \text{ red}} 0 + \sum_{k \text{ no color}} M
\]

\[
\leq \#(\text{blues}) (A(x) + \varepsilon) + M^2 \leq N (A(x) + \varepsilon) + M^2.
\]

Integrating and dividing by \(N\), we obtain \(\int f_M \wedge M \leq \int \bar{A} + \varepsilon + O(1/N)\). Passing to the limits \(N \to \infty\) and then \(M \to \infty\) (by the monotone convergence theorem), we get \(\int f \leq \int \bar{A} + \varepsilon\). Since \(\varepsilon\) was arbitrary, \(\int f \leq \int \bar{A}\).

**STEP 3.** \(\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) \text{ exists almost everywhere}\)

Together, steps 1 and 2 imply that \(\int (\bar{A} - \bar{A}) d\mu \leq 0\). But \(\bar{A} \geq \bar{A}\), so necessarily \(\bar{A} = \bar{A}\) \(\mu\)-almost everywhere (if \(\mu[\bar{A} > \bar{A}] > 0\), the integral would be strictly negative). So

\[
\limsup_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) = \liminf_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) \text{ a.e.,}
\]

which proves that the limit exists almost everywhere. The invariance of the limit is obvious from the invariance of \(\bar{A}(x)\).
2.3 The non-ergodic case

Step 4. \( \bar{f}(x) := \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) \) is \( T \)-invariant, absolutely integrable, and \( \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k \to \bar{f} \) in \( L^1 \). Consequently, \( \int \bar{f} \, d\mu = \int f \, d\mu \) and if \( T \) is ergodic, then \( \bar{f} = \int f \, d\mu \) almost everywhere.

Proof. The \( T \)-invariance of \( \bar{f} \) follows by taking the limit as \( n \to \infty \) in the identity

\[ A_n(x) = \frac{n-1}{n} A_{n-1}(T x) + \frac{1}{n} f(x). \]

Absolute integrability is because of step 1.

To see that \( \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k \xrightarrow{L^1} \bar{f} \), fix \( \varepsilon > 0 \) and construct \( \varphi \in L^\infty \) such that \( \|f - \varphi\|_1 < \varepsilon \). Notice that \( \|\frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ T^k\|_\infty \leq \|\varphi\|_\infty \), therefore by step 3 and the bounded convergence theorem,

\[
\frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ T^k \xrightarrow{n \to \infty} \bar{\varphi} \quad (2.1)
\]

for some bounded invariant function \( \bar{\varphi} \). So

\[
\left\| \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k - \bar{f} \right\|_1 \leq \left\| \frac{1}{n} \sum_{k=0}^{n-1} (f - \varphi) \circ T^k \right\|_1 + \left\| \frac{1}{n} \sum_{k=0}^{n-1} \varphi \circ T^k - \bar{\varphi} \right\|_1 + \|\varphi - \bar{\varphi}\|_1
\]

\[
\leq \|f - \varphi\|_1 + o(1) + \lim_{n \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} |(\varphi - f) \circ T^k| \, d\mu, \text{ by (2.1)}
\]

\[
\leq \|f - \varphi\|_1 + o(1) + \lim_{N \to \infty} \int \frac{1}{N} \sum_{k=0}^{N-1} |(\varphi - f) \circ T^k| \, d\mu, \text{ by Fatou’s Lemma}
\]

\[
\leq 2\varepsilon + o(1).
\]

Since \( \varepsilon \) was arbitrary, \( \|\frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k - \bar{f}\|_1 \to 0 \).

It is easy to see that if \( g_n \xrightarrow{L^1} g \), then \( g_n \to g \) almost surely. So \( \int \bar{f} \, d\mu = \lim_{n \to \infty} (\int f_n \, d\mu) = \int f \, d\mu \). Necessarily, in the ergodic case, \( \bar{f} = \text{const} = \int \bar{f} \, d\mu = \int f \, d\mu \). \( \square \)

2.3 The non-ergodic case

The almost sure limit in the pointwise ergodic theorem is clear when the map is ergodic: \( \frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k \xrightarrow{N \to \infty} \int f \, d\mu \). In this section we ask what is the limit in the non-ergodic case.

If \( f \) belongs to \( L^2 \), the limit is the projection of \( f \) on the space of invariant functions, because of the Mean Ergodic Theorem and the fact that every sequence of functions which converges in \( L^2 \) has a subsequence which converges almost everywhere to the same limit.\(^1\) But if \( f \in L^1 \) we cannot speak of projections. The right notion in this case is that of the *conditional expectation.*

\(^1\) Proof: Suppose \( f_n \xrightarrow{L^2} f \). Pick a subsequence \( n_k \) s.t. \( \|f_{n_k} - f\|_2 < 2^{-k} \). Then \( \sum_{k \geq 1} \|f_{n_k} - f\|_2 < \infty \). This means that \( \|\sum_{k \geq 1} f_{n_k} - f\|_2 < \infty \), whence \( \sum_{k \geq 1} (f_{n_k} - f) \) converges absolutely almost surely. It follows that \( f_{n_k} - f \to 0 \) a.e.
2.3.1 Conditional expectations and the limit in the ergodic theorem

Let \((X, \mathcal{B}, \mu)\) be a probability space. Let \(\mathcal{F} \subset \mathcal{B}\) be a \(\sigma\)-algebra. We think of \(F \in \mathcal{F}\) as of the collection of all sets \(F\) for which we have sufficient information to answer the question “is \(x \in F\)?”. The functions we have sufficient information to calculate are exactly the \(\mathcal{F}\)-measurable functions, as can be seen from the formula

\[
f(x) := \inf \{ t : x \in [f < t] \}.
\]

Suppose \(g\) is not \(\mathcal{F}\)-measurable. What is the ‘best guess’ for \(g(x)\) given the information \(\mathcal{F}\)?

Had \(g\) been in \(L^2\), then the “closest” \(\mathcal{F}\)-measurable function (in the \(L^2\)-sense) is the projection of \(g\) on \(L^2(X, \mathcal{F}, \mu)\). The defining property of the projection \(P_g\) of \(g\) is \(\langle P_g h, h \rangle = \langle g, h \rangle\) for all \(h \in L^2(X, \mathcal{F}, \mu)\). The following definition mimics this case when \(g\) is not necessarily in \(L^2\):

**Definition 2.1.** The conditional expectation of \(f \in L^1(X, \mathcal{B}, \mu)\) given \(\mathcal{F}\) is the unique \(L^1(X, \mathcal{F}, \mu)\)-element \(E(f|\mathcal{F})\) which is

1. \(E(f|\mathcal{F})\) is \(\mathcal{F}\)-measurable;
2. \(\forall \phi \in L^\infty \mathcal{F}\)-measurable, \(\int \phi E(f|\mathcal{F}) \, d\mu = \int \phi f \, d\mu\).

Note: \(E(f|\mathcal{F})\) is only determined almost everywhere.

**Proposition 2.1.** The conditional expectation exists for every \(L^1\) element, and is unique up sets of measure zero.

**Proof.** Consider the measures \(\nu_f := f \, d\mu|_\mathcal{F}\) and \(\mu|_\mathcal{F}\) on \((X, \mathcal{F})\). Then \(\nu_f \ll \mu\). The function \(E(f|\mathcal{F}) := \frac{d\nu_f}{d\mu}\) (Radon-Nikodym derivative) is \(\mathcal{F}\)-measurable, and it is easy to check that it satisfies the conditions of the definition of the conditional expectation. The uniqueness of the conditional expectation is left as an exercise. \(\square\)

**Proposition 2.2.** Suppose \(f \in L^1\).

1. \(f \mapsto E(f|\mathcal{F})\) is linear, and a contraction in the \(L^1\)-metric;
2. \(f \geq 0 \Rightarrow E(f|\mathcal{F}) \geq 0\) a.e.;
3. if \(\phi\) is convex and \(\phi \circ f \in L^1\), then \(E(\phi \circ f|\mathcal{F}) \geq \phi(E(f|\mathcal{F}))\);
4. if \(h\) is \(\mathcal{F}\)-measurable and bounded, then \(E(h|\mathcal{F}) = h E(f|\mathcal{F})\);
5. If \(\mathcal{F}_1 \subset \mathcal{F}_2\), then \(E[E(f|\mathcal{F}_1)|\mathcal{F}_2] = E(f|\mathcal{F}_2)\).

We leave the proof as an exercise.

**Theorem 2.3.** Let \((X, \mathcal{B}, \mu, T)\) be a p.p.t. and \(f \in L^1(X, \mathcal{B}, \mu)\). Then

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k = E(f|\mathcal{M}(T)) \text{ a.e. and in } L^1,
\]

where \(\mathcal{M}(T) := \{E \in \mathcal{B} : E = T^{-1}E\}\). Alternatively, \(\mathcal{M}(T)\) is the \(\sigma\)-algebra generated by all \(T\)-invariant functions.
Proof. Set \( \overline{f} := \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k \) on the set where the limit exists, and zero otherwise. Then \( \overline{f} \) is \( \mathcal{F} \)-measurable and \( T \)-invariant. For every \( T \)-invariant \( \varphi \in L^\infty \),

\[
\int \varphi \overline{f} d\mu = \int \varphi \frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k d\mu + O \left( \| \varphi \|_\infty \frac{1}{N} \sum_{k=0}^{N-1} f \circ T^k - \overline{f} \right) = \frac{1}{N} \sum_{k=0}^{N-1} \int \varphi \circ T^k f \circ T^k d\mu + o(1) \xrightarrow{N \to \infty} \int \varphi f d\mu,
\]

because the convergence in the ergodic theorem is also in \( L^1 \). \( \square \)

2.3.2 Conditional probabilities

Recall that a standard probability space is a probability space \( (X, \mathcal{B}, \mu) \) where \( X \) is a complete, metric, separable space, and \( \mathcal{B} \) is its Borel \( \sigma \)-algebra.

Theorem 2.4 (Existence of Conditional Probabilities). Let \( \mu \) by a Borel probability measure on a standard probability space \( (X, \mathcal{B}, \mu) \), and let \( \mathcal{F} \subset \mathcal{B} \) be a \( \sigma \)-algebra. There exist Borel probability measures \( \{\mu_x\}_{x \in X} \) s.t.:

1. \( x \mapsto \mu_x(E) \) is \( \mathcal{F} \)-measurable for every \( E \in \mathcal{B} \);
2. if \( f \) is \( \mu \)-integrable, then \( x \mapsto \int f d\mu_x \) is integrable, and
   \[
   \int f d\mu = \int (\int f d\mu_x) d\mu;
   \]
3. if \( f \) is \( \mu \)-integrable, then \( \int f d\mu_x = \mathbb{E}(f|\mathcal{F})(x) \) for \( \mu \)-a.e. \( x \).

Definition 2.2. The measures \( \mu_x \) are called the conditional probabilities of \( \mathcal{F} \). Note that they are only determined almost everywhere.

Proof. By the isomorphism theorem for standard spaces, there is no loss of generality in assuming that \( X \) is compact. Indeed, we may take \( X \) to be a compact interval.

Recall that for compact metric spaces \( X \), the space of continuous functions \( C(X) \) with the maximum norm is separable.\(^2\)

Fix a countable dense set \( \{f_n\}_{n=0}^\infty \subset C(X) \) s.t. \( f_0 \equiv 1 \). Let \( \mathcal{A}_Q \) be the algebra generated by these functions over \( Q \). It is still countable.

Choose for every \( g \in \mathcal{A}_Q \) an \( \mathcal{F} \)-measurable version \( \mathbb{E}(g|\mathcal{F}) \) of \( \mathbb{E}(g|\mathcal{F}) \) (recall that \( \mathbb{E}(g|\mathcal{F}) \) is an \( L^1 \)-function, namely not a function at all but an equivalence class of functions). Consider the following collection of conditions:

1. \( \forall \alpha, \beta \in \mathbb{Q}, g_{1,2} \in \mathcal{A}_Q \), \( \mathbb{E}(\alpha g_1 + \beta g_2|\mathcal{F})(x) = \alpha \mathbb{E}(g_1|\mathcal{F})(x) + \beta \mathbb{E}(g_2|\mathcal{F})(x) \)
2. \( \forall g \in \mathcal{A}_Q \), \( \min g \leq \mathbb{E}(g|\mathcal{F})(x) \leq \max g \)

\(^2\) Proof: Compact metric space are separable because they have finite covers by balls of radius \( 1/n \). Let \( \{x_n\} \) be a countable dense set of points, then \( \phi_n(\cdot) := \text{dist}(x_n, \cdot) \) is a countable family of continuous functions, which separates points in \( X \). The algebra which is generated over \( Q \) by \( \{\phi_n\} \cup \{1_X\} \) is countable. By the Stone-Weierstrass Theorem, it is dense in \( C(X) \).
This is countable collection of $\mathcal{F}$–measurable conditions, each of which holds with full $\mu$–probability. Let $X_0$ be the set of $x$’s which satisfies all of them. This is an $\mathcal{F}$–measurable set of full measure.

We see that for each $x \in X_0$, $\varphi_x[g] := \mathbb{E}(g \mid \mathcal{F})(x)$ is linear functional on $\mathcal{A}_Q$, and $\| \varphi_x \| \leq 1$. It follows that $\varphi_x$ extends uniquely to a positive bounded linear functional on $C(X)$. This is a measure $\mu_x$.

**Step 1.** $\int (\int f \, d\mu_x ) \, d\mu(x) = \int f \, d\mu$ for all $f \in C(X)$.

**Proof.** This is true for all $f \in \mathcal{A}_Q$ by definition, and extends to all $C(X)$ because $\mathcal{A}_Q$ is dense in $C(X)$. (But for $f \in L^1$ it is not even clear that the statement makes sense, because $\mu_x$ could live on a set with zero $\mu$–measure!)

**Step 2.** $x \mapsto \mu_x(E)$ is $\mathcal{F}$–measurable for all $E \in \mathcal{B}$.

**Exercise:** Prove this using the following steps

1. The indicator function of any open set is the pointwise limit of a sequence of continuous functions $0 \leq h_n \leq 1$, thus the step holds for open sets.
2. The collection of sets whose indicators are pointwise limits of a bounded sequence of continuous functions forms an algebra. The step holds for every set in this algebra.
3. The collection of sets for which step 1 holds is a monotone class which contains a generating algebra.

**Step 3.** If $f = g \mu$–a.e., then $f = g \mu_x$–a.e. for $\mu$–a.e. $x$.

**Proof.** Suppose $\mu(E) = 0$. Choose open sets $U_n \supseteq E$ such that $\mu(U_n) \to 0$. Choose continuous functions $0 \leq h_n^\varepsilon \leq 1$ s.t. $h_n^\varepsilon$ vanish outside $U_n$, $h_n^\varepsilon$ are non-zero inside $U_n$, and $h_n^\varepsilon \to 1_{U_n}$ (e.g. $h_n^\varepsilon(\cdot) := [\text{dist}(\cdot, U_n) / \text{diam}(X)]^\varepsilon$).

By construction $1_E \leq 1_{U_n} \equiv \lim \varepsilon \to 0^+ h_n^\varepsilon$, whence

$$\int \mu_x(E) \, d\mu(x) \leq \int \lim \varepsilon \to 0^+ h_n^\varepsilon \, d\mu_x \, d\mu \leq \lim \varepsilon \to 0^+ \int h_n^\varepsilon \, d\mu_x \, d\mu \leq \lim \varepsilon \to 0^+ \int h_n^\varepsilon \, d\mu \leq \mu(U_n) \to 0.$$

It follows that $\mu_x(E) = 0$ a.e.

**Step 4.** For all $f$ $\mu$–absolutely integrable, $\mathbb{E}(f \mid \mathcal{F})(x) = \int f \, d\mu_x$ $\mu$–a.e.

**Proof.** Find $g_n \in C(X)$ such that

$$f = \sum_{n=1}^\infty g_n \, \mu$–a.e., and $\sum \|g_n\|_{L^1(\mu)} < \infty$.

Then
\[ \mathbb{E}(f|\mathcal{F}) = \sum_{n=1}^{\infty} \mathbb{E}(g_n|\mathcal{F}), \text{ because } \mathbb{E}(\cdot|\mathcal{F}) \text{ is a bounded operator on } L^1 \]
\[ = \sum_{n=1}^{\infty} \int_X g_n d\mu_x \text{ a.e., because } g_n \in C(X) \]
\[ = \int_X \sum_{n=1}^{\infty} g_n d\mu_x \text{ a.e., (justification below)} \]
\[ = \int_X f d\mu_x \text{ a.e.} \]

Here is the justification: \( \int \sum |g_n| d\mu_x < \infty \), because the integral of this expression, by the monotone convergence theorem is less than \( \sum \|g_n\|_1 < \infty \). \( \square \)

2.3.3 The ergodic decomposition

**Theorem 2.5 (The Ergodic Decomposition).** Let \( \mu \) be an invariant Borel probability measure of a Borel map \( T \) on a standard probability space \( X \). Let \( \{\mu_x\}_{x \in X} \) be the conditional probabilities w.r.t. Inv(\( T \)). Then

1. \( \mu = \int_X \mu_x d\mu(x) \) (i.e. this holds when applies to \( L^1 \)–functions or Borel sets);
2. \( \mu_x \) is invariant for \( \mu \)–a.e. \( x \in X \);
3. \( \mu_x \) is ergodic for \( \mu \)–a.e. \( x \in X \).

**Proof.** By the isomorphism theorem for standard probability spaces, there is no loss of generality in assuming that \( X \) is a compact metric space, and that \( \mathcal{B} \) is its \( \sigma \)–algebra of Borel sets.

For every \( f \in L^1 \), \( \int f d\mu = \int \mathbb{E}(f|\text{Inv}(T)) d\mu(x) = \int_X \int_X f d\mu_x d\mu(x) \). This shows (1). We have to show that \( \mu_x \) is invariant and ergodic for \( \mu \)–a.e. \( x \).

Fix a countable set \( \{f_n\} \) which is dense in \( C(X) \), and choose Borel versions \( \mathbb{E}_\mu(f_n|\text{Inv}(T))(x) \). By the ergodic theorem, there is a set of full measure \( \Omega \) such that for all \( x \in \Omega \),

\[ \int f_n d\mu_x = \mathbb{E}_\mu(f_n|\text{Inv}(T))(x) = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f_n(T^k x) \text{ for all } n. \]

**Step 1.** \( \mu_x \) is \( T \)–invariant for a.e. \( x \in \Omega \).

**Proof.** For every \( n \),

\[ \int f_n \circ T d\mu_x = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f_n(T^{k+1} x) \text{ a.e. (by the PET)} \]
\[ = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} f_n(T^k x) = \mathbb{E}_\mu(f_n|\text{Inv}(T))(x) = \int f_n d\mu_x. \]
Let $\Omega'$ be the set of full measure for which the above holds for all $n$, and fix $x \in \Omega'$. Since $\{f_n\}$ is $\| \cdot \|_{\infty}$-dense in $C(X)$, we have $\int f \circ T \, d\mu_x = \int f \, d\mu_x$ for all $f \in C(X)$. Using the density of $C(X)$ in $L^1(\frac{\mu_x \circ T^{-1}}{2})$, it is routine to see that $\int f \circ T \, d\mu_x = \int f \, d\mu_x$ for all $\mu_x$-integrable functions. This means that $\mu_x \circ T^{-1} = \mu_x$ for all $x \in \Omega'$.

Step 2. $\mu_x$ is ergodic for all $x \in \Omega$.

Proof. With $\{f_n\}_{n=1}^{\infty}$ as above, let $\Omega'' := \{x \in \Omega' : \forall k, \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f_k(T^k x) = \int f_k \, d\mu_x\}$. This is a set of full measure because of the ergodic theorem. Now

\[
0 = \lim_{N \to \infty} \left\| \frac{1}{N} \sum_{k=0}^{N-1} f_n \circ T^k - \int_X f_n \, d\mu_x \right\|_{L^1(\mu)} \quad (\text{:: } L^1\text{-convergence in the PET})
\]

\[
= \lim_{N \to \infty} \int_X \left\| \frac{1}{N} \sum_{k=0}^{N-1} f_n \circ T^k - \int_X f_n \, d\mu_x \right\|_{L^1(\mu)} \, d\mu(x) \quad (\text{:: } \mu = \int_X \mu_x \, d\mu)
\]

\[
\geq \int_X \liminf_{N \to \infty} \left\| \frac{1}{N} \sum_{k=0}^{N-1} f_n \circ T^k - \int_X f_n \, d\mu_x \right\|_{L^1(\mu)} \, d\mu(x) \quad (\text{:: Fatou's Lemma})
\]

It follows that $\liminf_{n \to \infty} \left\| \frac{1}{n} \sum_{k=0}^{n-1} f_n \circ T^k - \int_X f_n \, d\mu_x \right\|_{L^1(\mu)} = 0 \mu$-a.e. Let

$$
\Omega := \left\{ x \in \Omega'' : \liminf_{n \to \infty} \left\| \frac{1}{n} \sum_{k=0}^{n-1} f_n \circ T^k - \int_X f_n \, d\mu_x \right\|_{L^1(\mu)} = 0 \text{ for all } n \right\},
$$

then $\Omega$ has full measure.

Fix $x \in \Omega$. Since $\{f_n\}_{n \geq 1}$ is dense in $C(X)$, and $C(X)$ is dense in $L^1(\mu_x)$, $\{f_n\}$ is dense in $L^1(\mu_x)$. A standard approximation argument shows that

$$
\liminf_{n \to \infty} \left\| \frac{1}{n} \sum_{k=0}^{n-1} f \circ T^k - \int_X f \, d\mu_x \right\|_{L^1(\mu_x)} = 0 \text{ for all } f \in L^1(\mu_x).
$$

In particular, every $f \in L^1(\mu_x)$ such that $f \circ T = f$ $\mu_x$-a.e. must be constant $\mu_x$-almost everywhere. So $\mu_x$ is ergodic for $x \in \Omega''$. \qed

### 2.4 The Subadditive Ergodic Theorem

We begin with two examples.

**Example 1 (Random walks on groups)** Let $(X, \mathcal{B}, \mu, T)$ be the Bernoulli scheme with probability vector $p = (p_1, \ldots, p_d)$. Suppose $G$ is a group, and $f : X \to G$ is the function $f(x_0, x_2, \ldots) = g_0$, where $g_1, \ldots, g_n \in G$. The expression

$$
f_n(x) := f(x) f(Tx) \cdots f(T^{n-1}x)
$$
describes the position of a random walk on \( G \), which starts at the identity, and whose steps have the distribution \( \Pr[\text{step} = g_i] = p_i \). What can be said on the behavior of this random walk?

In the special case \( G = \mathbb{Z}^d \) or \( G = \mathbb{R}^d \), \( f_n(x) = f(x) + f(Tx) + \cdots + f(T^{n-1}x) \), and the ergodic theorem\(^3\) says that \( \frac{1}{n} f_n(x) \) has an almost sure limit, equal to \( \int f \, d\mu = \sum p_i g_i \). So: the random walk has speed \( \| \sum p_i g_i \| \), and direction \( \sum p_i g_i / \| \sum p_i g_i \| \).

(Notice that if \( G = \mathbb{Z}^d \), the direction need not lie in \( G \).)

**Example 2 (The derivative cocycle)** Suppose \( T : V \to V \) is a diffeomorphism acting on an open set \( V \subset \mathbb{R}^d \). The derivative of \( T \) at \( x \in V \) is a linear transformation \( (dT)(x) \) on \( \mathbb{R}^d, \mathcal{V} \to ((dT)(x))_\mathcal{V} \). By the chain rule,

\[
(dT^n)(x) = (dT)(T^{n-1}x) \circ (dT)(T^{n-2}x) \circ \cdots \circ (dT)(x).
\]

If we write \( f(x) := (dT)(x) \in \text{GL}(d, \mathbb{R}) \), then we see that

\[
(dT^n)(x) = f(T^{n-1}x) f(T^{n-2}x) \cdots f(Tx) f(x)
\]

is a “random walk” on \( \text{GL}(d, \mathbb{R}) := \{ \text{invertible } d \times d \text{ matrices with real entries} \} \). Notice the order of multiplication!

What is the “speed” of this random walk? Does it have an asymptotic “direction”? The problem of describing the “direction” of random walk on a group is deep, and remain somewhat mysterious to this day, even in the case of groups of matrices. We postpone it for the moment, and focus on the conceptually easier task of defining the “speed.” Suppose \( G \) is a group of \( d \times d \) matrices with real-entries. Then \( G \) can be viewed to be group of linear operators on \( \mathbb{R}^d \), and we can endow \( A \in G \) with the operator norm \( \|A\| := \max \{ \|Av\|_2 / \|v\|_2 : 0 \neq v \in \mathbb{R}^d \} \). Notice that \( \|AB\| \leq \|A\| \|B\| \).

We will measure the speed of \( f_n(x) := f(x) f(Tx) \cdots f(T^{n-1}x) \) by analyzing

\[
g(n)(x) := \log \|f_n(x)\| \text{ as } n \to \infty.
\]

**Key observation:** \( g^{(n+m)} \leq g^{(n)} + g^{(m)} \circ T^n \), because

\[
g^{(n+m)}(x) = \log \|f_{n+m}(x)\| = \log \|f_n(x) f_m(T^n x)\| \leq \log(\|f_n(x)\| \cdot \|f_m(T^n x)\|) \\
\leq \log \|f_n(x)\| + \log \|f_m(T^n x)\| = g^{(n)}(x) + g^{(m)}(T^n x).
\]

We say that \( \{g^{(n)}\}_n \) is a *subadditive cocycle*.

**Theorem 2.6 (Kingman’s Subadditive Ergodic Theorem).** Let \((X, \mathcal{B}, m, T)\) be a probability preserving transformation, and suppose \( g^{(n)} : X \to \mathbb{R} \) is a sequence of absolutely integrable functions such that \( g^{(n+m)} \leq g^{(n)} + g^{(m)} \circ T^n \) for all \( n, m \). Then the limit \( g := \lim_{n \to \infty} g^{(n)} / n \geq -\infty \) exists almost surely, and is an invariant function.

**Proof.** We begin by observing that it is enough to treat the case when \( g^{(n)} \) are all non-positive. This is because \( h^{(n)} := g^{(n)} - (g^{(1)} + g^{(1)} \circ T + \cdots + g^{(1)} \circ T^{n-1}) \) are

\(^3\) applied to the each coordinate of the vector valued function \( f = (f^1, \ldots, f^d) \).
non–positive, sub-additive, and differ from $g^{(n)}$ by the ergodic sums of $g^{(1)}$ whose asymptotic behavior we know by Birkhoff’s ergodic theorem.

Assume then that $g^{(n)}$ are all non-negative. Define $G(x) := \liminf_{n \to \infty} g^{(n)}(x)/n$ (the limit may be equal to $-\infty$). We claim that $G \circ T = G$ almost surely.

Starting from the subadditivity inequality $g^{(n+1)} \leq g^{(n)} \circ T + g^{(1)}$, we see that $G \circ T \leq G \circ T$. Suppose there were a set of positive measure $E$ where $G \circ T > G + \epsilon$. Then for every $x \in E$, $G(T^n x) > G(T^{n-1} x) \geq \cdots \geq G(T x) > G(x) + \epsilon$. But this is impossible, because by Poincaré’s Recurrence Theorem, for a.e. $x$ there is some $n > 0$ such that $G(T^n x) = \infty$ or $|G(T^n x) - G(x)| < \epsilon$ (prove!). This contradiction shows that $G = G \circ T$ almost surely. Henceforth we work the set of full measure $X_0 := \bigcap_{n \geq 1} [G \circ T^n = G]$.

Fix $M > 0$, and define $G_M := G \vee (-M)$. This is an invariant function on $X_0$. We aim at showing $\limsup_{n} g^{(n)}_n \leq G_M$ a.s.. Since $M$ is arbitrary, this implies that $\limsup_{n} g^{(n)}_n / n \leq G = \liminf_{n} g^{(n)}_n / n$, whence the theorem.

Fix $x \in X_0$, $N \in \mathbb{N}$, and $\epsilon > 0$. Call $k \in \mathbb{N} \cup \{0\}$

- “good”, if $\exists \ell \in \{1, \ldots, N\}$ s.t. $g^{(\ell)}(T^k x)/\ell \leq G_M(T^k x) + \epsilon = G_M(x) + \epsilon$;
- “bad”, if it’s not good: $g^{(\ell)}(T^k x)/\ell > G_M(x) + \epsilon$ for all $\ell = 1, \ldots, N$.

Color the integers $0, \ldots, n - 1$ inductively as follows, starting from $k = 1$. Let $b$ be the smallest non-colored integer,

(a) If $k \leq n - N$ and $k$ is “bad”, color it red;
(b) If $k \leq n - N$ and $k$ is “good”, find the smallest $1 \leq \ell \leq N$ s.t. $g^{(\ell)}(T^k x)/\ell \leq G_M(T^k x) + \epsilon$ and color the segment $[k, k + \ell)$ blue;
(c) If $k > n - N$, color $k$ white.

Repeat this procedure until all integers $0, \ldots, n - 1$ are colored.

The “blue” part can be decomposed into segments $[\tau_i, \tau_i + \ell_i)$, with $\ell_i$ s.t. $g^{(\ell_i)}(T^k x)/\ell_i \leq G_M(x) + \epsilon$. Let $b$ denote the number of these segments.

The “red” part has size $\leq \sum_{i=1}^{b} 1_{B(N,M,\epsilon)}(T^k x)$, where

$$B(N,M,\epsilon) := \{x \in X_0 : g^{(\ell)}(x)/\ell > G_M(x) + \epsilon \text{ for all } 1 \leq \ell \leq N\}.$$ 

Let $r$ denote the size of the red part. The “white” part has size $w \leq N$.

By the sub-additivity condition

$$\frac{g^{(n)}(x)}{n} \leq \frac{1}{n} \sum_{i=1}^{b} g^{(\ell_i)}(T^k x) + \frac{1}{n} \sum_{k \text{ red}} g^{(1)}(T^k x) + \frac{1}{n} \sum_{k \text{ white}} g^{(1)}(T^k x) \leq \frac{1}{n} \sum_{i=1}^{b} g^{(\ell_i)}(T^k x) \leq \frac{1}{n} \sum_{i=1}^{b} (G_M(x) + \epsilon) \ell_i = \frac{\# \{\text{blues}\}}{n} (G_M(x) + \epsilon).$$
Now \( \#\{\text{blues}\} \leq n - (r + w) = n - \sum_{k=1}^{n} 1_{B(N,M,E)}(T^k x) + O(1) \), so by the Birkhoff ergodic theorem, for almost every \( x \), \( \#\{\text{blues}\}/n \rightarrow 1 - \mathbb{E}(1_{B(N,M,E)}|\mathcal{F}) \). Thus

\[
\limsup_{n \to \infty} \frac{g^{(n)}(x)}{n} \leq (G_M(x) + \varepsilon)(1 - \mathbb{E}(1_{B(N,M,E)}|\mathcal{F})) \text{ almost surely.}
\]

Now \( N \) was arbitrary, and for fixed \( M \) and \( \varepsilon \), \( B(N,M,\varepsilon) \downarrow \emptyset \) as \( N \uparrow \infty \), because \( G_M \geq G = \liminf_{\ell \to \infty} g^{(\ell)}/\ell \). It is not difficult to deduce from this that \( \mathbb{E}(1_{B(N,M,\varepsilon)}|\mathcal{F}) \to 0 \) almost surely.\footnote{Suppose \( 0 \leq f_n \leq 1 \) and \( f_n \downarrow 0 \). The conditional expectation is monotone, so \( \mathbb{E}(f_n|\mathcal{F}) \) is decreasing at almost every point. Let \( \varphi \) be its almost sure limit, then \( 0 \leq \varphi \leq 1 \) a.s., and by the BCT, \( \mathbb{E}(\varphi) = \mathbb{E}(\lim \mathbb{E}(f_n|\mathcal{F})) = \lim \mathbb{E}(\mathbb{E}(f_n|\mathcal{F})) = \lim \mathbb{E}(f_n) = \mathbb{E}(\lim f_n) = 0 \), whence \( \varphi = 0 \) almost everywhere.}

Thus

\[
\limsup_{n \to \infty} \frac{g^{(n)}(x)}{N(n)} \leq G_M(x) + \varepsilon \text{ almost surely.}
\]

Since \( \varepsilon \) was arbitrary, \( \limsup_{n \to \infty} g^{(n)}/n \leq G_M \) almost surely, which proves the theorem by the discussion above.

\[ \square \]

**Proposition 2.3.** Suppose \( m \) is ergodic, then the limit in Kingman’s ergodic theorem is the constant \( \inf\{(1/n) \int g^{(n)} \, dm\} \) (possibly equal to \( -\infty \)).

**Proof.** Let \( G := \lim g^{(n)}/n \). Subadditivity implies that \( G \circ T \leq G \). Recurrence implies that \( G \circ T = T \). Ergodicity implies that \( G = c \text{ a.e.} \), for some constant \( c = c(g) \geq -\infty \). We claim that \( c \leq \inf\{(1/n) \int g^{(n)} \, dm\} \). This is because

\[
c = \lim_{k \to \infty} \frac{1}{kn} g^{(kn)}(x) \leq \lim_{k \to \infty} \frac{1}{k} \left( \frac{g^{(n)}}{n} + \frac{g^{(n)}}{n} \circ T^{n} + \cdots + \frac{g^{(n)}}{n} \circ T^{n(k-1)} \right)
\]

\[
= \frac{1}{n} \int g^{(n)} \, dm \quad \text{(Birkhoff’s ergodic theorem),}
\]

proving that \( c \leq (1/n) \int g^{(n)} \, dm \) for all \( n \).

To prove the other inequality we first note (as in the proof of Kingman’s subadditive theorem) that it is enough to treat the case when \( g^{(n)} \) are all non-positive. Otherwise work with \( h^{(n)} := g^{(n)} - (g^{(1)} + \cdots + g^{(1)} \circ T^{n-1}) \). Since \( g^{(1)} \in L^1 \),

\[
\frac{1}{1-n} (g^{(1)} + \cdots + g^{(1)} \circ T^{n-1}) \to \int g^{(1)} \, dm \text{ pointwise and in } L^1.
\]

Thus \( c(g) = \lim_{n \to \infty} \frac{1}{n} \int g^{(n)} = c(h) + \int g^{(1)} = \inf(1/n) \int h^{(n)} + \int S_n g^{(1)} = \inf(1/n) \int g^{(n)} \, dm \).

Suppose then that \( g^{(n)} \) are all non-positive. Fix \( N \), and set \( g_N^{(n)} := \max\{g^{(n)}, -nN\} \).

This is, again, subadditive because
\[ g_N^{(n+m)} = \max\{g^{(n+m)}_N, -(n+m)N\} \leq \max\{g^{(n)}_N + g^{(m)}_N, -(n+m)N\} \]
\[ \leq \max\{g^{(n)}_N + g^{(m)}_N, -(n+m)N\} = g^{(n)}_N + g^{(m)}_N \circ T^n. \]

By Kingman’s theorem, \( g^{(n)}_N / n \) converges pointwise to a constant \( c(g_N) \). By definition, \( -N \leq g^{(n)}_N / n \leq 0 \), so by the bounded convergence theorem,
\[ c(g_N) = \lim_{n \to \infty} \frac{1}{n} \int g^{(n)}_N dm \geq \inf \frac{1}{n} \int g^{(n)}_N dm \geq \inf \frac{1}{n} \int g^{(n)} dm. \quad (2.2) \]

**Case 1:** \( c(g) = -\infty \). In this case \( g^{(n)} / n \to -\infty \), and for every \( N \) there exists \( N(x) \) s.t. \( n > N(x) \Rightarrow g^{(n)}_N(x) = -N \). Thus \( c(g_N) = -N \), and (2.2) implies \( \inf[(1/n) \int g^{(n)} dm] = -\infty = c(g) \).

**Case 2:** \( c(g) \) is finite. Take \( N > |c(g)| + 1 \), then for a.e. \( x \), if \( n \) is large enough, then \( g^{(n)} / n > c(g) - \varepsilon > -N \), whence \( g^{(n)}_N = g^{(n)} \). Thus \( c(g) = c(g_N) \geq \inf \frac{1}{n} \int g^{(n)} dm \) and we get the other inequality. \( \square \)

Here is a direct consequence of the subadditive ergodic theorem (historically, it predates the subadditive ergodic theorem):

**Theorem 2.7 (Furstenberg–Kesten).** Let \( (X, \mathcal{B}, \mu, T) \) be a p.p.t. and \( A : X \to \text{GL}(d, R) \) be a measurable function s.t. \( \log \|A^{\pm 1}\| \in L^1 \). If \( A_n(x) := A(T^{n-1}x) \cdots A(x) \), then the following limit exists a.e. and is invariant: \( \lim_{n \to \infty} \frac{1}{n} \log \|A_n(x)\| \).

The following immediate consequence will be used in the proof of the Oseledets theorem for invertible cocycles:

**Remark:** Suppose \( (X, \mathcal{B}, m, T) \) is invertible, and let \( g^{(n)} \) be a subadditive cocycle s.t. \( g^{(1)} \in L^1 \). Then for a.e. \( x \), \( \lim_{n \to \infty} g^{(n)} \circ T^{-n} / n \) exists and equals \( \lim_{n \to \infty} g^{(n)} / n \).

**Proof.** Since \( g^{(n)} \) is subadditive, \( g^{(n)} \circ T^{-n} \) is subadditive:
\[ g^{(n+m)} \circ T^{-(n+m)} \leq [g^{(n)} \circ T^m + g^{(m)}] \circ T^{-(n+m)} = g^{(n)} \circ T^{-n} + [g^{(m)} \circ T^{-m}] \circ T^{-n}. \]

Let \( m = \int m_y d\pi(y) \) be the ergodic decomposition of \( m \). Kingman’s ergodic theorem and the previous remark say that for \( \pi \)-a.e. \( y \),
\[ \lim_{n \to \infty} \frac{g^{(n)} \circ T^{-n}}{n} = \inf \frac{1}{n} \int g^{(n)} \circ T^{-n} dm_y = \inf \frac{1}{n} \int g^{(n)} dm_y \text{ m.a.e.} \]
\[ = \lim_{n \to \infty} \frac{g^{(n)}}{n} \text{ m.a.e.} \]

Thus the set where the statement of the remark fails has zero measure with respect to all the ergodic components of \( m \), and this means that the statement is satisfied on a set of full \( m \)-measure. \( \square \)
2.5 The Multiplicative Ergodic Theorem

2.5.1 Preparations from Multilinear Algebra

Multilinear forms. Let $V = \mathbb{R}^n$ equipped with the Euclidean inner product $\langle v, w \rangle = \sum v_i w_i$. A linear functional on $V$ is a linear map $\omega : V \to \mathbb{R}$. The set of linear functionals is denoted by $V^*$. Any $v \in V$ determines $v^* \in V^*$ via $v^* = \langle v, \cdot \rangle$. Any linear function is of this form.

A $k$-multilinear function is a function $T : V^k \to \mathbb{R}$ such that for all $i$ and $v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k \in V$, $T(v_1, \ldots, v_{i-1}, \cdot, v_{i+1}, \ldots, v_k)$ is a linear functional.

The set of all $k$-multilinear functions on $V$ is denoted by $T^k(V)$. The tensor product of $\omega \in T^k(V)$ and $\eta \in T^l(V)$ is $\omega \otimes \eta \in T^{k+l}(V)$ given by

$$(\omega \otimes \eta)(v_1, \ldots, v_{k+l}) := \omega(v_1, \ldots, v_k)\eta(v_{k+1}, \ldots, v_{k+l}).$$

The tensor product is bilinear and associative, but it is not commutative.

The dimension of $T^k(V)$ is $n^k$. Here is a basis: $\{e^*_i \otimes \cdots \otimes e^*_k : 1 \leq i_1, \ldots, i_k \leq n\}$. To see this note that every element in $T^k(\Omega)$ is completely determined by its action on $\{(e_i, \cdots, e_k) : 1 \leq i_1, \ldots, i_k \leq n\}$.

Define an inner product on $T^k(V)$ by declaring the above basis to be an orthogonal collection of vectors of length $\frac{1}{\sqrt{k!}}$ (the reason for the normalization will become clear later).

Alternating multilinear forms. A multilinear form $\omega$ is called alternating, if it satisfies $\exists i \neq j(v_i = v_j) \Rightarrow \omega(v_1, \ldots, v_n) = 0$. Equivalently,

$$\omega(v_1, \ldots, v_i, v_j, \ldots, v_n) = -\omega(v_1, \ldots, v_j, v_i, \ldots, v_n).$$

(to see the equivalence, expand $\omega(v_1, \ldots, v_i + v_j, \ldots, v_j + v_i, \ldots, v_n)$). The set of all $k$-alternating forms is denoted by $\Omega^k(V)$.

Any multilineal form $\omega$ gives rise to an alternating form $\text{Alt}(\omega)$ via

$$\text{Alt}(\omega) := \frac{1}{k!} \sum_{\sigma \in S_k} \text{sgn}(\sigma) \sigma \cdot \omega,$$

where $S_k$ is the group of $k$-permutations, and the action of a permutation $\sigma$ on $\omega \in T^k(V)$ is given by $(\sigma \cdot \omega)(v_1, \ldots, v_k) = \omega(v_{\sigma(1)}, \ldots, v_{\sigma(k)})$. The normalization $k!$ is to guarantee $\text{Alt}|_{\Omega^k(V)} = \text{id}$, and $\text{Alt}^2 = \text{Alt}$. Note that $\text{Alt}$ is linear.

Lemma 2.1. $\text{Alt}[\text{Alt}(\omega_1 \otimes \omega_2) \otimes \omega_3] = \text{Alt}(\omega_1 \otimes \omega_2 \otimes \omega_3)$.

Proof. We show that if $\text{Alt}(\omega) = 0$, then $\text{Alt}(\omega \otimes \eta) = 0$ for all $\eta$. Specializing to the case $\omega = \text{Alt}(\omega_1 \otimes \omega_2) - \omega_1 \otimes \omega_2$ and $\eta = \omega_3$, we get (since $\text{Alt}^2 = \text{Alt}$)

$$\text{Alt}[\text{Alt}(\omega_1 \otimes \omega_2) - \omega_1 \otimes \omega_2] \otimes \omega_3] = 0,$$
which is equivalent to the statement of the lemma.

Suppose \( \omega \in T^k(V) \), \( \eta \in T^l(V) \), and \( \text{Alt}(\omega) = 0 \). Let \( G := \{ \sigma \in S_{k+l} : \sigma(i) = i \text{ for all } i = k+1, \ldots, k+l \} \). This is a subgroup of \( S_{k+l} \), and there is a natural isomorphism \( \sigma \mapsto \sigma' := \sigma|_{\{1, \ldots, k\}} \) from \( G \) to \( S_k \). Let \( S_{k+l} = \bigcup_j G\sigma_j \) be the corresponding coset decomposition, then

\[
(k+l)! \text{Alt}(\omega \otimes \eta)(v_1, \ldots, v_{k+l}) =
= \sum_j \sum_{\sigma \in G} \text{sgn}(\sigma \sigma_j)(\sigma \sigma_j) \cdot (\omega \otimes \eta)(v_1, \ldots, v_{k+l})
= \sum_j \sum_{\sigma \in G} \text{sgn}(\sigma_j) \eta(v_{\sigma_j(k+1)}, \ldots, v_{\sigma_j(k+l)}) \sum_{\sigma' \in S_k} \text{sgn}(\sigma') \cdot \omega(v_{\sigma'(1)}, \ldots, v_{\sigma'(k)})
= \sum_j \sum_{\sigma \in G} \text{sgn}(\sigma_j) \eta(v_{\sigma_j(k+1)}, \ldots, v_{\sigma_j(k+l)}) \sum_{\sigma' \in S_k} \text{sgn}(\sigma') \cdot \omega(v_{\sigma_j(1)}, \ldots, v_{\sigma_j(k)})
= \sum_j \sum_{\sigma \in G} \text{sgn}(\sigma_j) \eta(v_{\sigma_j(k+1)}, \ldots, v_{\sigma_j(k+l)}) \text{Alt}(\omega)(v_{\sigma_j(1)}, \ldots, v_{\sigma_j(k)}) = 0. \quad \square
\]

Using this “antisymmetrization operator”, we define the following product, called the \textit{exterior product} or the \textit{wedge product}: If \( \omega \in \Omega^k(V) \), \( \eta \in \Omega^l(V) \), then

\[
\omega \wedge \eta := \frac{(k+l)!}{k! l!} \text{Alt}(\omega \otimes \eta).
\]

The wedge product is bilinear, and the previous lemma shows that it is associative. It is almost anti commutative: If \( \omega \in \Omega^k(V) \), \( \eta \in \Omega^l(V) \), then

\[
\omega \wedge \eta = (-1)^{kl} \eta \wedge \omega.
\]

We’ll see the reason for the peculiar normalization later.

**Proposition 2.4.** \( \{ e_{i_1} \wedge \cdots \wedge e_{i_k}^* : 1 \leq i_1 < \cdots < i_k \leq n \} \) is an orthonormal basis for \( \Omega^k(V) \), whence \( \dim \Omega^k(V) = \binom{n}{k} \).

**Proof.** Suppose \( \omega \in \Omega^k(V) \), then \( \omega \in T^k(V) \) and so \( \omega = \sum a_{i_1, \ldots, i_k} e_{i_1}^* \otimes \cdots \otimes e_{i_k}^* \), where the sum ranges over all \( k \)-tuples of numbers between 1 and \( n \). If \( \omega \in \Omega^k(V) \), then \( \text{Alt}(\omega) = \omega \) and so

\[
\omega = \sum a_{i_1, \ldots, i_k} \text{Alt}(e_{i_1}^* \otimes \cdots \otimes e_{i_k}^*).
\]

Fix \( \xi := e_{i_1}^* \otimes \cdots \otimes e_{i_k}^* \). If \( i_\alpha = i_\beta \) for some \( \alpha \neq \beta \), then the permutation \( \sigma_0 \) which switches \( \alpha \leftrightarrow \beta \) preserves \( \xi \). Thus for all \( \sigma \in S_k \),

\[
\text{sgn}(\sigma \sigma_0)(\sigma \sigma_0) \cdot \xi = -\text{sgn}(\sigma) \sigma \cdot \xi
\]

and we conclude that \( \text{Alt}(\xi) = 0 \). If, on the other hand, \( i_1, \ldots, i_k \) are all different, then it is easy to see using lemma 2.1 that \( \text{Alt}(e_{i_1}^* \otimes e_{i_2}^* \otimes \cdots \otimes e_{i_k}^*) = \frac{1}{k!} e_{i_1}^* \wedge \cdots \wedge e_{i_k}^* \).

Thus \( \omega = \frac{1}{k!} \sum a_{i_1, \ldots, i_k} e_{i_1}^* \wedge \cdots \wedge e_{i_k}^* \), and we have proved that the set of forms in the statement spans \( \Omega^k(V) \).
To see that this set is independent, we show that it is orthonormal. Suppose \( \{i_1, \ldots, i_k\} \neq \{j_1, \ldots, j_k\} \), then the sets \( \{\sigma \cdot e_{i_1}^* \otimes \cdots \otimes e_{i_k}^*\} \) and \( \{\sigma \cdot e_{j_1}^* \otimes \cdots \otimes e_{j_k}^*\} \) are disjoint, so \( \text{Alt}(e_{i_1}^* \otimes \cdots \otimes e_{i_k}^*) \perp \text{Alt}(e_{j_1}^* \otimes \cdots \otimes e_{j_k}^*) \). This proves orthogonality. Orthonormality is because \( \|e_{i_1}^* \wedge \cdots \wedge e_{i_k}^*\|_2 = k!\|\text{Alt}(e_{i_1}^* \wedge \cdots \wedge e_{i_k}^*)\|_2 = k!\|\sum_{\sigma \in S_k} \text{sgn}(\sigma) \sigma \cdot (e_{i_1}^* \otimes \cdots \otimes e_{i_k}^*)\|_2 = \sum_{\sigma \in S_k} \text{sgn}(\sigma)^2 \left(\frac{1}{\sqrt{k!}}\right)^2 = 1 \). (This explains why we chose to define \( \|e_{i_1}^* \otimes \cdots \otimes e_{i_k}^*\|_2 := \frac{1}{\sqrt{k!}} \).)

**Corollary 2.2.** \( e_1^* \wedge \cdots \wedge e_n^* \) is the determinant. This is the reason for the peculiar normalization in the definition of \( \wedge \).

**Proof.** The determinant is an alternating \( n \)-form, and \( \dim \Omega^n(V) = 1 \), so the determinant is proportional to \( e_1^* \wedge \cdots \wedge e_n^* \). Since the values of both forms on the standard basis is one (because \( e_1^* \wedge \cdots \wedge e_n^* = n!\text{Alt}(e_1^* \otimes \cdots \otimes e_n^*) \)), they are equal. \( \square \)

We define an inner product on \( \Omega^k(V) \) by declaring the basis in the proposition to be orthonormal. Let \( \| \cdot \| \) be the resulting norm.

**Lemma 2.2.** For \( \nu \in V \), let \( \nu^* := \langle \nu, \cdot \rangle \), then

(a) \( \| \omega \wedge \eta \| \leq \| \omega \| \| \eta \| \).
(b) \( \langle v_1^* \wedge \cdots \wedge v_k^*, w_1^* \wedge \cdots \wedge w_k^* \rangle = \det(\langle v_i, w_j \rangle) \).
(c) If \( \{u_1, \ldots, u_n\} \) is an orthonormal basis for \( V \), then \( \{u_1^* \wedge \cdots \wedge u_n^* : 1 \leq i_1 < \cdots < i_k \leq n\} \) is an orthonormal basis for \( \Omega^k(V) \).
(d) If \( \text{span}\{v_1, \ldots, v_k\} = \text{span}\{u_1, \ldots, u_k\} \), then \( v_1^* \wedge \cdots \wedge v_k^* \) and \( u_1^* \wedge \cdots \wedge u_k^* \) are proportional.

**Proof.** Write for \( I = (i_1, \ldots, i_k) \) such that \( 1 \leq i_1 < \cdots < i_k \leq n \), \( e_I^* := e_{i_1}^* \wedge \cdots \wedge e_{i_k}^* \).

Represent \( \omega := \sum \alpha_i e_I^* \), \( \eta := \sum \beta_j e_J^* \), then

\[
\| \omega \wedge \eta \|^2 = \left\| \sum_{I \cup J = \emptyset} \alpha_I \beta_J e_I^* \wedge e_J^* \right\|^2 = \left\| \sum_{I \cup J = \emptyset} \pm \alpha_I \beta_J e_I^* \wedge e_J^* \right\|^2 = \sum_{I \cup J = \emptyset} \alpha_I^2 \beta_J^2 \leq \| \omega \|^2 \| \eta \|^2.
\]

Take two multi indices \( I, J \). If \( I \neq J \), then the inner product matrix is the identity matrix. If \( I \neq J \), then \( \exists \alpha \in I \setminus J \) and then the \( \alpha \)-row and column of the inner product matrix will be zero. Thus the formula holds for any pair \( e_I^*, e_J^* \). Since part (b) of the lemma holds for all basis vectors, it holds for all vectors. Part (c) immediately follows.

Next we prove part (d). Represent \( v_i = \sum \alpha_{ij} w_j \), then

\[
v_1^* \wedge \cdots \wedge v_k^* = \text{const. Alt}(v_1^* \otimes \cdots \otimes v_k^*) = \text{const. Alt} \left( \sum_j \alpha_{ij} u_j^* \otimes \cdots \otimes \sum_j \alpha_{ik} u_j^* \right) = \text{const.} \sum_{\alpha} \alpha_{j_1} \cdots \alpha_{j_k} \text{Alt}(u_{j_1}^* \otimes \cdots \otimes u_{j_k}^*).
\]

The terms where \( j_1, \ldots, j_k \) are not all different are annihilated by \( \text{Alt} \). The terms where \( j_1, \ldots, j_k \) are all different are mapped by \( \text{Alt} \) to a form which proportional to \( u_1^* \wedge \cdots \wedge u_k^* \). Thus the result of the sum is proportional to \( u_1^* \wedge \cdots \wedge u_k^* \). \( \square \)
Exterior product of linear operators  Let $A: V \to V$ be a linear operator. The $k$–th exterior product of $A$ is $A^{\wedge k}: \Omega^k(V) \to \Omega^k(V)$ given by

$$ (A^{\wedge k} \omega)(v_1, \ldots, v_k) := \omega(A'v_1, \ldots, A'v_k). $$

The transpose is used to get $A^{\wedge k}(v_1^* \wedge \cdots \wedge v_k^*) = (Av_1^*)^* \wedge \cdots \wedge (Av_k^*)^*$.

**Theorem 2.8.** $\|A^{\wedge k}\| = \lambda_1 \cdots \lambda_k$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the eigenvalues of $(AA^*)^{1/2}$, listed in decreasing order with multiplicities.

**Proof.** The matrix $AA^*$ is symmetric, so it can be orthogonally diagonalized. Let $\{v_1, \ldots, v_n\}$ be an orthonormal basis of eigenvectors, listed so that $\lambda_1 \geq \cdots \geq \lambda_n$. Then $\{v_1^*: I \subseteq \{1, \ldots, d\}, |I| = k\}$ is an orthonormal basis for $\Omega^k(\mathbb{R}^d)$, where we are using the multi index notation

$$ v_I^* = v_{i_1}^* \wedge \cdots \wedge v_{i_k}^*, $$

where $i_1 < \cdots < i_k$ is an ordering of $I$.

Given $\omega \in \Omega^k(\mathbb{R}^d)$, write $\omega = \sum \omega_I v_I^*$, then

$$ \|A^{\wedge k} \omega\|^2 = \langle A^{\wedge k} \omega, A^{\wedge k} \omega \rangle = \left\langle \sum_I \omega_I A^{\wedge k} v_I^*, \sum_J \omega_J A^{\wedge k} v_J^* \right\rangle = \sum_{IJ} \omega_I \omega_J \langle A^{\wedge k} v_I^*, A^{\wedge k} v_J^* \rangle. $$

Now, $\langle A^{\wedge k} v_I^*, A^{\wedge k} v_J^* \rangle = \langle (Av_{i_1})^* \wedge \cdots \wedge (Av_{i_k})^* \rangle \langle (Av_{j_1})^* \wedge \cdots \wedge (Av_{j_k})^* \rangle = \det \left( \langle Av_{i_1}, Av_{j_1} \rangle \right) \det \left( \langle Av_{i_2}, Av_{j_2} \rangle \right) \cdots = \prod_{j=1}^k \lambda_j^2$. Thus $\|A^{\wedge k} \omega\|^2 = \sum_I \omega_I^2 \prod_{j=1}^k \lambda_j^2 \leq \|\omega\|^2 \prod_{j=1}^k \lambda_j^2$. It follows that $\|A^{\wedge k}\| \leq \lambda_1 \cdots \lambda_k$.

To see that the inequality is in fact an equality, consider the case $\omega = v_I^*$ where $I = \{1, \ldots, k\}$, then $\|A^{\wedge k} \omega\| = \langle v_I^*, v_I^* \rangle = (\lambda_1 \cdots \lambda_k)^2 = (\lambda_1 \cdots \lambda_k)^{2k} \|\omega\|^2$. \qed

**Exterior products and angles between vector spaces** The angle between vector spaces $V, W \subset \mathbb{R}^d$ is

$$ \angle(V, W) := \min \{ \arccos \langle v, w \rangle : v \in V, w \in W, \|v\| = \|w\| = 1 \}. $$

So if $V \cap W \neq \{0\}$ iff $\angle(V, W) = 0$, and $V \perp W$ iff $\angle(V, W) = \pi/2$. 


2.5 The Multiplicative Ergodic Theorem

**Proposition 2.5.** If \((w_1, \ldots, w_k)\) is a basis of \(W\), and \((v_1, \ldots, v_t)\) is a basis of \(W\), then
\[
\| (v_1^* \wedge \cdots \wedge v_t^*) \cdot (w_1^* \wedge \cdots \wedge w_k^*) \| \leq \| v_1^* \wedge \cdots \wedge v_t^* \| \cdot \| w_1^* \wedge \cdots \wedge w_k^* \| \cdot | \sin \angle(V, W)|.
\]

**Proof.** If \(V \cap W \neq \{0\}\) then both sides are zero, so suppose \(V \cap W = \{0\}\), and pick an orthonormal basis \(e_1, \ldots, e_{n+k}\) for \(V \oplus W\). Let \(w \in W, v \in V\) be unit vectors s.t. \(\angle(V, W) = \angle(v, w)\), and write \(v = \sum v_i e_i, w = \sum w_j e_j\), then
\[
\| v^* \wedge w^* \|^2 = \left\| \sum_{i,j} v_i w_j e_i^* \wedge e_j^* \right\|^2 = \left\| \sum_{i,j} (v_i w_j - v_j w_i) e_i^* \wedge e_j^* \right\|^2 = \sum_{i,j} (v_i w_j - v_j w_i)^2
\]
\[
= \frac{1}{2} \sum_{i,j} (v_i w_j - v_j w_i)^2 \quad \text{(the terms where } i = j \text{ vanish)}
\]
\[
= \frac{1}{2} \sum_{i,j} (v_i^2 w_j^2 + v_j^2 w_i^2 - 2v_i w_i \cdot v_j w_j) = \frac{1}{2} \left[ 2 \sum_i v_i^2 \sum_j w_j^2 - 2 \left( \sum_i v_i w_i \right)^2 \right]
\]
\[
= \|v\|^2 \|w\|^2 - \langle v, w \rangle^2 = 1 - \cos^2 \angle(v, w) = \sin^2 \angle(V, W).
\]

Complete \(v\) to an orthonormal basis \((v, v'_1, \ldots, v'_t)\) of \(V\), and complete \(w\) to an orthonormal basis \((w, w'_2, \ldots, w'_k)\) of \(W\). Then
\[
\| (v^* \wedge v'_1^* \wedge \cdots \wedge v'_t^*) \cdot (w^* \wedge w'_2^* \wedge \cdots \wedge w'_k^*) \|
\]
\[
\leq \|v^* \wedge w^*\| \cdot \|v'_1^* \wedge \cdots \wedge v'_t^*\| \cdot \|w'_2^* \wedge \cdots \wedge w'_k^*\| = | \sin \angle(V, W) | \cdot 1 \cdot 1,
\]
because of orthonormality. By lemma 2.2
\[
v_i^* \wedge \cdots \wedge v_t^* = \pm \|v_1^* \wedge \cdots \wedge v_t^*\| \cdot v_1^* \wedge v_2^* \wedge \cdots \wedge v_t^*
\]
\[
w_1^* \wedge \cdots \wedge w_k^* = \pm \|w_1^* \wedge \cdots \wedge w_k^*\| \cdot w_1^* \wedge w_2^* \wedge \cdots \wedge w_k^*
\]
and the proposition follows. \(\Box\)

### 2.5.2 Proof of the Multiplicative Ergodic Theorem

Let \((X, \mathcal{B}, m, f)\) be a ppt, and \(A : X \to \text{GL}(d, \mathbb{R})\) some Borel map. We define \(A_n := A \circ f^{n-1} \cdots A\), then the cocycle identity holds: \(A_{n+m}(x) = A_n(f^m x) A_m(x)\).

**Theorem 2.9 (Multiplicative Ergodic Theorem).** Let \((X, \mathcal{B}, T, m)\) be a ppt, and \(A : X \to \text{GL}(d, \mathbb{R})\) a Borel function s.t. \(\ln \|A(x)^{\pm 1}\| \in L^1(m)\), then
\[
\Lambda(x) := \lim_{n \to \infty} \left[ A_n(x) A_n(x)^* \right]^{1/2n}
\]
exists a.e., and \(\lim_{n \to \infty} \frac{1}{n} \ln \| A_n(x) \Lambda(x)^{-n} \| = \lim_{n \to \infty} \frac{1}{n} \ln \| (A_n(x) \Lambda(x)^{-n})^{-1} \| = 0 \text{ a.s.}\)

**Proof.** The matrix \(B_n(x) := \sqrt{A_n(x)^* A_n(x)}\) is symmetric, therefore it can be orthogonally diagonalized. Let \(\lambda_n^1(x) < \cdots < \lambda_n^{n(x)}(x)\) be its different eigenvalues, and...
$$\mathbb{R}^d = W_n^{\lambda_1}(x) \oplus \cdots \oplus W_n^{\lambda_m}(x)$$

the orthogonal decomposition of $\mathbb{R}^d$ into the corresponding eigenspaces. The proof has the following structure:

**Part 1**: Let $t_n^i(x) \leq \cdots \leq t_n^d(x)$ be a list of the eigenvalues of $B_n(x) := \sqrt{A_n(x)^i A_n(x)}$

*with multiplicities*, then for a.e. $x$, there is a limit $t_i(x) = \lim_{n \to \infty} [t_n^i(x)]^{1/n}$, $i = 1, \ldots, d$.

**Part 2**: Let $\lambda_1(x) < \cdots < \lambda_{s(x)}(x)$ be a list of the different values of $\{t_i(x)\}_{i=1}^d$.

Divide $\{t_n^i(x)\}_{i=1}^d$ into $s(x)$ subsets of values $\{t_n^i(x) : i \in I_n^j\}$, $1 \leq j \leq s(x)$ in such a way that $t_n^i(x)^{1/n} \to \lambda_j(x)$ for all $i \in I_n^j$. Let

$$U_n^j(x) := \text{sum of the eigenspaces of } t_n^i(x), i \in I_n^j.$$ 

= the part of the space where $B_n(x)$ dilates by approximately $\lambda_j(x)^n$.

We show that the spaces $U_n^j(x)$ converge as $n \to \infty$ to some limiting spaces $U^j(x)$ (in the sense that the orthogonal projections on $U_n^j(x)$ converge to the orthogonal projection on $U^j(x)$).

**Part 3**: The theorem holds with $\Lambda(x) : \mathbb{R}^d \to \mathbb{R}^d$ given by $\nu \mapsto \lambda_i(x) \nu$ on $U^j(x)$.

Part 1 is proved by applying the subadditive ergodic theorem for a cleverly chosen sub-additive cocyle (“Raghunathan’s trick”). Parts 2 and 3 are (non-trivial) linear algebra.

**Part 1**: Set $g_i^{(n)}(x) := \sum_{d-i} \ln t_n^i(x)$. This quantity is finite, because $A_n^i A_n$ is invertible, so none of its eigenvalues vanish.

The sequence $g_i^{(n)}$ is subadditive! This is because the theory of exterior products says that $\exp g_i^{(n)} = \text{product of the } i \text{ largest e.v.'s of } \sqrt{A_n(x)^i A_n(x)} = \|A_n(x)^i\|$, so

$$\exp g_i^{(n+m)}(x) = \|A_{n+m}^{\lambda_i}(x)\| = \|A_m(T^n)^{\lambda_i} A_n(x)^{\lambda_i} \| \leq \|A_m(T^n)^{\lambda_i}\| \|A_n(x)^{\lambda_i}\|$$

$$= \exp [g_i^{(m)}(T^n x) + g_i^{(n)}(x)],$$

whence $g_i^{(n+m)} \leq g_i^{(n)} + g_i^{(m)} \circ T^n$.

We want to apply Kingman’s subadditive ergodic theorem. First we need to check that $g_i^{(1)} \in L^1$. We use the following fact from linear algebra: if $\lambda$ is an eigenvalue of a matrix $B$, then $\|B^{-1}\|^{-1} \leq |\lambda| \leq \|B\|$. Therefore

---

5 Proof: Let $v$ be an eigenvector of $\lambda$ with norm one, then $|\lambda| = \|Bv\| \leq \|B\|$ and $1 = \|B^{-1}Bv\| \leq \|B^{-1}\|\|Bv\| = \|B^{-1}\| |\lambda|.$
The Pointwise Ergodic Theorem imply that

\[ V \cdot \text{vector in the sense that their orthogonal projections converge.} \]

Thus, \( \tilde{\cdot} \)

The linear spaces \( U \)

\[ \text{Technical lemma: Denote the projection of a} \]

Taking differences, we see that the following limit exists a.e.:

\[ \ln t_i(x) := \lim_{n \to \infty} \frac{1}{n} [g_{d-i+1}^{(n)}(x) - g_{d-i}^{(n)}(x)] = \lim_{n \to \infty} \frac{1}{n} \ln t_i^n(x). \]

Thus \( [t_i^n(x)]^{1/n} \to t_i(x) \) almost surely, for some \( t_i(x) \in \mathbb{R} \).

**Part 2:** Fix \( x \) s.t. \( [t_i^n(x)]^{1/n} \to t_i(x) \) for all \( 1 \leq i \leq d \). Henceforth we work with this \( x \) only, and write for simplicity \( A_n = A_n(x), t_i = t_i(x) \) etc.

Let \( s = s(x) \) be the number of the different \( t_i \). List the different values of these quantities an increasing order: \( \lambda_1 < \lambda_2 < \cdots < \lambda_s \). Set \( \chi_i := \log \lambda_i \). Fix \( 0 < \delta < \frac{1}{2} \min \{ \chi_{j+1} - \chi_j \} \). Since for all \( i \) there is a \( j \) s.t. \( (t_i^n)^{1/n} \to \lambda_j \), the following sets eventually stabilize and are independent of \( n \):

\[ I_j := \{ i : |(t_i^n)^{1/n} - \lambda_j| < \delta \} \quad (j = 1, \ldots, s). \]

Define, relative to \( \sqrt{A_n^t}A_n \),

- \( U_{d}^i := \sum_{j \in I_j} \text{eigenspace of } t_i^n(x) \) (this sum is not necessarily direct);
- \( V_{d}^i := \oplus_{j \in I_j} U_{d}^j \)
- \( V_{d}^i := \oplus_{j \not\in I_j} U_{d}^j \)

The linear spaces \( U_{1}^d, \ldots, U_{d}^s \) are orthogonal, since they are eigenspaces of different eigenvalues for a symmetric matrix \( \sqrt{A_n^t}A_n \). We show that they converge as \( n \to \infty \), in the sense that their orthogonal projections converge.

The proof is based on the following technical lemma. Denote the projection of a vector \( v \) on a subspace \( W \) by \( v \parallel W \), and write \( \chi_i := \log \lambda_i \).

**Technical lemma:** For every \( \delta > 0 \) there exists constants \( K_1, \ldots, K_s > 1 \) and \( N \) s.t. for all \( n > N, t = 1, \ldots, s, k \in \mathbb{N}, \) and \( u \in V_{d}^i \),
\[ \|u|V^r_{n+1}\| \leq K_1 \|u\| \exp(-n(\chi_r+\delta - \delta)) \]

We give the proof later. First we show how it can be used to finish parts 2 and 3.

We show that \( V^r_n \) converge as \( n \to \infty \). Since the projection on \( U^r_n \) is the projection on \( V^r_n \) minus the projection on \( V^r_{n-1} \), it will then follow that the projections of \( U^r_n \) converge.

Fix \( N \) large. We need it to be so large that
1. \( I_j \) are independent of \( n \) for all \( n > N \);
2. The technical lemma works for \( n > N \) with \( \delta \) as above.

There will be other requirements below.

Fix an orthonormal basis \( \{v^1_n, \ldots, v^d_n\} \) for \( V^r_n \) (\( d_r = \dim(V^r_n) = \sum_{|J| = r} |J| \)). Write

\[ v^j_n = \alpha^j_n w^j_{n+1} + u^j_{n+1}, \quad \text{where} \quad w^j_{n+1} \in V^r_{n+1}, \|w^j_{n+1}\| = 1, \ u^j_{n+1} \in \tilde{V}^{r+1}_{n+1}. \]

Note that \( \|w^j_{n+1}\| = \|v^j_n|\tilde{V}^{r+1}_{n+1}\| \leq K_1 \exp(-n(\chi_r+\delta - \delta)) \). Using the identity \( \alpha^j_n = \sqrt{1 - \|w^j_{n+1}\|^2} \), it is easy to see that for some constants \( C_1 \) and \( 0 < \theta < 1 \) independent of \( n \) and \( (v^j_n) \),

\[ \|v^j_n - w^j_{n+1}\| \leq C_1 \theta^n. \]  

(\( \theta := \max, \exp[-(\chi_r+\delta - \delta)] \) and \( C_1 := 2K_1 \) should work.)

The system \( \{w^j_{n+1}\} \) is very close to being orthonormal:

\[ \langle w^j_{n+1}, w^k_{n+1} \rangle = \langle v^j_n, v^k_n \rangle + v^j_n, v^k_n \rangle = \delta_j + O(\theta^n), \]

because \( \{v^j_n\} \) is an orthonormal system. It follows that for all \( n \) large enough, \( w^j_{n+1} \) are linearly independent. A quick way to see this is to note that

\[ \|(w^1_{n+1})^* \wedge \cdots \wedge (w^d_{n+1})^*\|^2 = \det(\langle w^j_{n+1}, w^k_{n+1} \rangle) \quad \text{(lemma 2.2)} \]

\[ = \det(I + O(\theta^n)) \neq 0, \quad \text{provided} \ n \ \text{is large enough}, \]

and to observe that wedge produce of a linearly dependent system vanishes.

It follows that \( \{w^1_{n+1}, \ldots, w^d_{n+1}\} \) is a linearly independent subset of \( V^r_{n+1} \). Since \( \dim(V^r_{n+1}) = \sum_{|J| = r} |J| = d_r \), this is a basis for \( V^r_{n+1} \).

Let \( \{v^j_{n+1}\} \) be the orthonormal basis obtained by applying the Gram–Schmidt procedure to \( \{w^j_{n+1}\} \). We claim that there is a global constant \( C_2 \) such that

\[ \|v^j_n - v^j_{n+1}\| \leq C_2 \theta^n. \]  

(2.4)

Write \( v_i = v^i_{n+1}, w_i = w^i_{n+1} \), then the Gram–Schmidt process is to set \( v_1 = u_i/\|u_i\| \), where \( u_i \) are defined by induction by \( u_1 := w_1, u_i := w_i - \sum_{j<i} \langle w_i, v_j \rangle v_j \). We construct by induction global constants \( C_i \) s.t. \( \|v_i - w_i\| \leq C_i \theta^n \), and then take \( C_2 := \max\{C_2\} \). When \( i = 1 \), we can take \( C_1 := C_1 \), because \( v_1 = w_1 \), and \( \|w_1 - v_1\| \leq C_1 \theta^n \). Suppose we have constructed \( C_2, \ldots, C_{i-2} \). Then
\[ \|u_i - w_i\| \leq \sum_{j<i} |\langle w_i, v_j \rangle| \leq \sum_{j<i} |\langle w_j, w_j \rangle| + \|w_j - v_j\| \leq \left(2C_1(i-1) + \sum_{j<i} C_j^2\right) \theta^n, \]

because \(|\langle w_i, v_j \rangle| = |\langle w_i - v_i, w_j \rangle| + |\langle v_i, w_j - w_j \rangle| \leq 2C_1 \theta^n\). Call the term in the brackets \(K\), and assume \(n\) is so large that \(K \theta^n < 1/2\), then

\[ \|u_i - w_i\| \leq \left(2C_1(i-1) + \sum_{j<i} C_j^2\right) \theta^n. \]

and we can take \(C_i^2 := 4K\). This proves (2.4).

Starting from the orthonormal basis \(\{v_{i,n}\}\) for \(V_{r,n}\), we have constructed an orthonormal basis \(\{v_{i,n+1}\}\) for \(V_{r,n+1}\) such that \(\|v_{i,n} - v_{i,n+1}\| \leq C_2 \theta^n\). Continue this procedure by induction, and construct the orthonormal bases \(\{v_{i,n+k}\}\) for \(V_{r,n+k}\). By (2.4), these bases form Cauchy sequences: \(v_{i,n+k} \to v_i\).

The limit vectors must also be orthonormal. Denote their span by \(V^r\). The projection on \(V^r\) takes the form

\[ \sum_{i=1}^d \langle v^i, \cdot \rangle v^i = \lim_{k \to \infty} \sum_{i=1}^d \langle v^i_{n+k}, \cdot \rangle v^i_{n+k} = \lim_{k \to \infty} \text{proj}_{V^r} v^i_{n+k}. \]

Thus \(V^r_{n+k} \to V^r\).

**Part 3:** We saw that \(\text{proj}_{U_i(x)} \to \text{proj}_{U_i(x)}\) for some linear spaces \(U_i(x)\). Set \(\Lambda(x) \in \text{GL}(\mathbb{R}^d)\) to be the matrix representing

\[ \Lambda(x) = \sum_{j=1}^{s(x)} e^{Z_j(x)} \text{proj}_{U_j(x)}. \]

Since \(U_i(x)\) are limits of \(U^i_n\), they are orthogonal, and they sum up to \(\mathbb{R}^d\). It follows that \(\Lambda\) is invertible, symmetric, and positive.

Choose an orthogonal basis \(\{v^i_n(x), \ldots, v^d_n(x)\}\) of \(B_n(x) := \sqrt{A_n(x)} A_n(x)\) so that

\[ B_n v^i_n = t^i_n v^i_n \quad \text{for all } i,\]

and let \(W_n := \text{span}\{v^i_n\}\). Then for all \(v \in \mathbb{R}^d\),
\[(A_n\Lambda_0^{-1/2})^{1/2n}v = (\sqrt{A_n\Lambda_0})^{1/2n}v = \sum_{i=1}^d t_i(x)^{1/n}\text{proj}_{V_i}(v)\]
\[= \sum_{j=1}^s \sum_{i \in I_j} t_i(x)^{1/n}\text{proj}_{V_i}(v)\]
\[= \sum_{j=1}^s e^{X_j(x)} \sum_{i \in I_j} \text{proj}_{W_i}(v) + o(\|v\|),\]
where \(o(\|v\|)\) denotes a vector with norm \(o(\|v\|)\).

Thus \((A_n\Lambda_0^{-1/2})^{1/2n} \to \Lambda_0\).

We show that \(\frac{1}{n} \log \|A_n\Lambda_0^{-n}\| \to 0\) a.e. It’s enough to show that
\[\lim_{n \to \infty} \frac{1}{n} \log \|A_n\| = \chi_r := \log \Lambda_0 \text{ uniformly on the unit ball in } U_r.\] (2.5)

To see that this is enough, note that \(A_n = \sum_{i=1}^s e^{X_i} v_i(U_r)\); for all \(\delta > 0\), if \(n\) is large enough, then for every \(v\),
\[\|A_n\Lambda_0^{-n}v\| \leq \sum_{i=1}^s e^{-n\delta} \|A_n(v)U_r\| = \sum_{i=1}^s e^{-n\delta} \|e^{n\delta} \|v\| \leq s e^{n\delta} \|v\| \quad (v \in \mathbb{R}^d)\]
\[\|A_n\Lambda_0^{-n}v\| = e^{n\delta} \|A_n\| \to e^{2n\delta} \|v\| \quad (v \in U_r)\]

Thus \(\|A_n\Lambda_0^{-n}\| \to e^{-n\delta}\) for all \(\delta\), whence \(\frac{1}{n} \log \|A_n\Lambda_0^{-n}\| \to 0\) a.e.

To see that \(\frac{1}{n} \log \|A_n\Lambda_0^{-n}\| \to 0\), we use a duality trick.

Define for a matrix \(C, C^\#: = (C^{-1})^t\), then \((C_1C_2)^\# = C_1^\#C_2^\#\). Thus \((\Lambda)^\# = (\Lambda_n)^\#\),
and \(B_n^\# := \sqrt{(\Lambda^\#)_{n}^{-1}}(\Lambda^\#)_{n} = (\sqrt{A_n^\# A_n^\#})^{-1} = \sqrt{A_n^\# A_n^\#}^{-1}\). Thus we have the following relation between the objects associated to \(A^\#\) and \(A\):

1. the eigenvalues of \(B_n^\#\) are \(1/t_1^\# \leq \cdots \leq 1/t_s^\#\) (the order is flipped)
2. the eigenspaces of \(1/t_1^\#\) for \(B_n^\#\) is the eigenspace of \(t_1^\#\) for \(B_n\)
3. 
4. \(U_r^\# = U_n^{r+j} V_r^{-r+1}, V_r^\# = V_n^{r-j+1}\)
5. \(\Lambda^\# = \Lambda^{-1}\).

Thus \(\|(\Lambda^\# A_n^{-1})\| = \|(\Lambda^\# A_n^{-1})^\#\| = \|(A_n^\# \Lambda^{-n})\|\), so the claim \(\frac{1}{n} \log \|A_n\Lambda_0^{-n}\| \to 0\) a.e. follows from what we did above, applied to \(A^\#\).

Here is another consequence of this duality: There exist \(K_1^\#, \ldots, K_t^\#\) s.t. for all \(\delta\), there is an \(N\) s.t. for all \(n > N\), if \(u \in U_n^\#\), then for all \(k\)
\[\|u|V_n^{r-k}\| \leq K_t^\# \exp[-n(\chi_r - \chi_{r-\delta})],\] (2.6)
2.5 The Multiplicative Ergodic Theorem

To see this note that \( V^{r-t}_{n+k} = (V^{s-r-t+1}_{n+k})^\# \) and \( U^r_n \subset (V^{s-r+1}_n)^\# \), and apply the technical lemma to the cocycle generated by \( A^\# \).

We prove (2.5). Fix \( \delta > 0 \) and \( N \) large (we see how large later), and assume \( n > N \). Suppose \( v \in U^r_n \) and \( ||v|| = 1 \). Write \( v = \lim v_{n+k} \) with \( v_{n+k} = v|U^r_{n+k} \subset U^r_{n+k} \).

Note that \( ||v_{n+k}|| \leq 1 \). We decompose \( v_{n+k} \) as follows

\[
v_{n+k} = (v_{n+k}|V^{r-1}_n) + (v_{n+k}|U^r_n) + \sum_{t=1}^{r-t} (v_{n+k}|U^t_{n+k}),
\]

and estimate the size of the image of each of the summands under \( A_n \).

**First summand:**

\[
\|A_n(v_{n+k}|V^{r-1}_n)\| = \langle B_n^2(v_{n+k}|V^{r-1}_n), (v_{n+k}|V^{r-1}_n) \rangle
\]

\[
= e^{2n(\chi_{r-1} + o(1))} \|v_{n+k}|V^{r-1}_n\| \leq e^{2n(\chi_{r-1} + o(1))}.
\]

Thus the first summand is less than \( \exp[n(\chi_{r-1} + \delta)] \).

**Second Summand:**

\[
\|A_n(v_{n+k}|U^r_n)\| = \langle B_n(v_{n+k}|U^r_n), (v_{n+k}|U^r_n) \rangle
\]

\[
= e^{2n(\chi_{r} + o(1))} \|v_{n+k}|U^r_n\| \leq e^{2n(\chi_{r} + o(1))} (\|v|U^r_n\| + (\|v_{n+k} - v\|)U^r_n)^2
\]

\[
= e^{2n(\chi_{r} + \delta)} (\|v|U^r_n\| + \|v_{n+k} - v\|)^2 = e^{2n(\chi_{r} + \delta)} [1 + o(1)] \text{ uniformly in } v.
\]

Thus the second summand is \( [1 + o(1)] \exp[n(\chi_{r} + \delta)] \) uniformly in \( v \in U^r, \|v\| = 1 \).

**Third Summand:** For every \( t \),

\[
\|A_n(v_{n+k}|U^t_{n+k})\| = \langle B_n(v_{n+k}|U^t_{n+k}), (v_{n+k}|U^t_{n+k}) \rangle
\]

\[
\leq e^{2n(\chi_{r+t} + o(1))} \|v_{n+k}|U^t_{n+k}\|^2
\]

\[
= e^{2n(\chi_{r+t} + o(1))} \left( \sup_{u \in U^t_{n+k}, ||u|| = 1} \langle v_{n+k}, u \rangle \right)^2, \text{ because } \|x|W\| = \sup_{w \in W, ||w|| = 1} \langle x, w \rangle
\]

\[
\leq e^{2n(\chi_{r+t} + o(1))} \left( \sup_{u \in U^t_{n+k}, ||u|| = 1} \sup_{v \in V^r_{n+k}, ||v|| \leq 1} \langle v, u \rangle \right)^2
\]

\[
= e^{2n(\chi_{r+t} + o(1))} \sup_{u \in U^t_{n+k}, ||u|| = 1} ||u|V^r_{n+k}\|^2
\]

\[
\leq (K^\#)^2 e^{2n(\chi_{r+t} + o(1))} \exp[-2n(\chi_r - \chi_{r+t} - o(1))], \text{ by (2.6)}
\]

\[
= (K^\#)^2 e^{2n(\chi + o(1))}.
\]

Note that the cancellation of \( \chi_{r+t} \) — this is the essence of the technical lemma. We get: \( \|A_n(v_{n+k}|U^t_{n+k})\| = O(\exp[n(\chi_r + o(1))]) \). Summing over \( t = 1, \ldots, s-r \), we get that third summand is \( O(\exp[n(\chi_r + o(1))]) \).
Putting these estimates together, we get that
\[ \|A_n v_{n+k}\| \leq \text{const. } \exp[n(\chi_r + o(1))] \] uniformly in \( k \), and on the unit ball in \( U_r \).

“Uniformity” means that the \( o(1) \) can be made independent of \( v \) and \( k \). It allows us to pass to the limit as \( k \to \infty \) and obtain
\[ \|A_n v\| \leq \text{const. } \exp[n(\chi_r + o(1))] \] uniformly on the unit ball in \( U_r \).

On the other hand, an orthogonality argument shows that
\[ \|A_n v_{n+k}\|^2 = \langle B^2_n v_{n+k}, v_{n+k} \rangle \]
\[ = \|1\text{st summand}\|^2 + \|2\text{nd summand}\|^2 + \|3\text{rd summand}\|^2 \]
\[ \geq \|2\text{nd summand}\|^2 = [1 + o(1)] \exp[2n(\chi_r + o(1))]. \]

Thus \( \|A_n v_{n+k}\| \geq [1 + o(1)] \exp[n(\chi_r + o(1))] \) uniformly in \( v, k \). Passing to the limit as \( k \to \infty \), we get \( \|A_n v\| \geq \text{const. } \exp[n(\chi_r + o(1))] \) uniformly on the unit ball in \( U_r \). These estimates imply (2.5).

**Proof of the technical lemma:** We are asked to estimate the norm of the projection of a vector in \( V_n^r \) on \( V_{n+k}^{r+t} \). We do this in three steps:

1. \( V_n^r \to V_{n+1}^{r+t} \), all \( t > 0 \);
2. \( V_n^r \to V_{n+k}^r \), all \( k > 0 \);
3. \( V_n^r \to V_{n+k}^{r+t} \), all \( t, k > 0 \).

**Step 1.** The technical lemma for \( k = 1 \): Fix \( \delta > 0 \), then for all \( n \) large enough and for all \( i' > r \), if \( u \in V_n^r \), then \( \|u|V_{n}^{r+t}\| \leq \|u\| \exp(-n(\chi_{r'} - \chi_r - \delta)) \).

**Proof.** Fix \( \varepsilon \), and choose \( N = N(\varepsilon) \) so large that \( t'_n = e^{\pm n\varepsilon t_i} \) for all \( n > N, i = 1, \ldots, d \). For every \( t = 1, \ldots, s \), if \( u \in V_n^r \), then
\[ \|A_{n+1} u\| = \sqrt{\langle A'_{n+1} A_{n+1} u, u \rangle} \]
\[ = \sqrt{\langle A'_{n+1} A_{n+1} u|V_{n+1}^{r+t}\rangle, (u|V_{n+1}^{r+t}) \rangle + \langle A'_{n+1} A_{n+1} (u|V_{n+1}^{r+t-1}), (u|V_{n+1}^{r+t}) \rangle \}
\]
(because \( V_{n+1}^{r+t-1}, V_{n+1}^{r+t} \) are orthogonal, \( A'_{n+1} A_{n+1} \)-invariant, and \( \mathbb{R}^d = V_{n+1}^{r+t-1} \oplus V_{n+1}^{r+t} \))
\[ = \sqrt{\|A_{n+1} (u|V_{n+1}^{r+t})\|^2 + \|A_{n+1} (u|V_{n+1}^{r+t-1})\|^2} \]
\[ \geq \|A_{n+1} (u|V_{n+1}^{r+t})\| = \exp(\chi_{r+t/n}) (n+1) \|u|V_{n+1}^{r+t}\|. \]

On the other hand
\[ \|A_{n+1}u\| = \|A(T^n x)A_n(x)u\| \leq \|A(T^n x)\| \sqrt{\langle A_n A_n u, u \rangle} \]
\[ \leq \|A(T^n x)\| e^{n(\chi \pm \varepsilon)} \|u\| \]
\[ = e^{n(\chi \pm \varepsilon) + o(n)} \|u\| , \]

because by the ergodic theorem
\[ \frac{1}{n} \log \|A(T^n x)\| = \frac{1}{n} \sum_{k=0}^{n-1} \log \|A(T^k x)\| - \frac{1}{n} \sum_{k=0}^{n-1} \log \|A(T^k x)\| \xrightarrow{n \to \infty} 0 \text{ a.e.} \]

By further increasing \( N \), we can arrange \(|o(n)| < ne\), which gives
\[ e^{(\chi + \varepsilon)(n+1)} \|u\| \|
\[ \leq e^{\varepsilon(\chi + 2\varepsilon)} , \]

whence \( \|u\| \|
\[ \leq e^{n(\chi - 3\varepsilon)} \). Now take \( \varepsilon := \delta / 3 \).

**Step 2.** Fix \( \delta > 0 \). Then for all \( n \) large enough and for all \( k \), if \( u \in V_r^n \), then
\[ \|u\| \|
\[ \leq \|u\| \sum_{j=0}^{n-1} \exp(-(n+j)(\chi_{r+1} - \chi_r - \delta)) \]. Thus \( \exists K_1 \) s.t.
\[ \|u\| \|
\[ \leq K_1 \|u\| \exp[-n(\chi_{r+1} - \chi_r - \delta)]. \]

**Proof.** We use induction on \( k \). The case \( k = 1 \) is dealt with in step 1. We assume by induction that the statement holds for \( k - 1 \), and prove it for \( k \). Decompose
\[ u_{n+k} = [u_{n+k}^r] + [u_{n+k}] . \]

- **First summand:** \( u_{n+k}^r \in V_{n+k-1} \), so by step 1 the norm of the first summand is less than \( \|u\| \|
\[ \exp[-(n+k-1)(\chi_{r+1} - \chi_r - \delta)] \), whence less than \( \|u\| \|
\[ \exp[-(n+k-1)(\chi_{r+1} - \chi_r - \delta)]. \]

- **Second summand:** The norm is at most \( \|u\| \|
\[ \sum_{j=0}^{n-1} \exp(-(n+j)(\chi_{r+1} - \chi_r - \delta)). \]

We get the statement for \( k \), and step 2 follows by induction.

As a result, we obtain the existence of a constant \( K_1 > 1 \) for which \( u \in V_r^n \) implies
\[ \|u\| \|
\[ \leq K_1 \|u\| \exp[-n(\chi_{r+1} - \chi_r - \delta)] . \]

**Step 3.** \( \exists K_1, \ldots, K_{s-1} > 1 \) s.t. for all \( n \) large enough and for all \( k \), if \( u \in V_r^n \) implies
\[ \|u\| \|
\[ \leq K_s \|u\| \exp[-n(\chi_{r+s-1} - \chi_r - s\delta)] \). 

**Proof.** We saw that \( K_1 \) exists. We assume by induction that \( K_1, \ldots, K_{s-1} \) exist, and construct \( K_s \). Fix \( 0 < \delta_0 < \min_{j} \{ \chi_{j+1} - \chi_j \} - \delta \); the idea is to first prove that if \( u \in V_r^n \), then
\[ \|u\| \|
\[ \leq \|u\| \left( \sum_{j=0}^{k-1} (\sum_{s=1}^{\delta_0} e^{-s}) \left( \sum_{j=0}^{k-1} \exp[-(n+j)(\chi_{r+1} - \chi_r - \delta)] \right) \right) . \]

Once this is done, step 3 follows with \( K_s := \left( \sum_{s=1}^{\delta_0} e^{-s} \right)^2 . \)
We prove (2.7) using induction on $k$. When $k = 1$ this is because of step 1. Suppose, by induction, that (2.7) holds for $k - 1$. Decompose:

$$u |V_{n+k}^{r+1} = u |V_{n+k}^{r} + \sum_{r' < r < t} u |V_{n+k}^{r} + u |V_{n+k}^{r+1}$$

\(A \quad B \quad C\)

- **Estimate of \(\|A\|\):** By step 1, \(\|A\| \leq \|u\| \exp(-n(k-1)(\chi_{r+t} - \chi_{r} - \delta))\).
- **Estimate of \(\|B\|\):** By step 1, and the induction hypothesis (on $t$):

\[
\|B\| \leq \sum_{r < r' < t} \|u\| \exp(-(n + k - 1)(\chi_{r+t} - \chi_{r'} - \delta))
\]

\[
\leq \sum_{r < r' < t} \|u\| \exp(-(n + k - 1)(\chi_{r+t} - \chi_{r'} - \delta))
\]

\[
\leq \sum_{r < r' < t} K_{r-r} \|u\| \exp(-(n(\chi_{r'} - \chi_{r} - (r'-r)\delta)) \times \exp(-(n + k - 1)(\chi_{r+t} - \chi_{r'} - \delta))
\]

\[
\leq \|u\| \left( \sum_{r' = 1}^{t-1} K_{r'} \right) e^{-(k-1)(\chi_{r+t} - \chi_{r} - \delta)} \exp(-(n(\chi_{r+t} - \chi_{r} - t\delta))
\]

\[
\leq \|u\| \left( \sum_{r = 1}^{t-1} K_{r} \right) e^{-(k-1)} \exp(-(n(\chi_{r+t} - \chi_{r} - t\delta)).
\]

- **Estimate of \(\|C\|\):** \(\|C\| \leq \|u\| \|V_{n+k}^{r+1}\|\). By the induction hypothesis on $k$,

\[
\|C\| \leq \|u\| \left( \sum_{r = 1}^{t-1} K_{r} \right) \left( \sum_{j = 0}^{k-2} e^{-\delta_{0}j} \right) \left( \sum_{j = 0}^{k-2} \exp(-(n + j)(\chi_{r+t} - \chi_{r} - t\delta)) \right).
\]

It is not difficult to see that when we add these bounds for \(\|C\|, \|B\|\) and \(\|A\|\), the result is smaller than the RHS of (2.7) for $k$. This completes the proof by induction of (2.7). As explained above, step 3 follows by induction. \(\Box\)

**Corollary 2.3.** Let $\chi_{1}(x) < \cdots < \chi_{d(x)}(x)$ denote the logarithms of the (different) eigenvalues of $\Lambda(x)$. Let $U_{\chi_{i}}$ be the eigenspace of $\Lambda(x)$ corresponding to $\exp \chi_{i}$. Set $V_{\chi} := \bigoplus_{\chi_{i} \leq \chi} U_{\chi_{i}}$.

1. $\chi(x, v) := \lim_{n \to \infty} \frac{1}{n} \log \|A_{n}(x)v\|$ exists a.s, and is invariant.
2. $\chi(x, v) = \chi_{i}$ on $V_{\chi_{i}} \setminus V_{\chi_{i-1}}$.
3. If $\|A^{-1}\|$, $\|A\| \in L^{\infty}$, then $\frac{1}{n} \log |\det A_{n}(x)| = \sum k_{i} \chi_{i}$, where $k_{i} = \dim U_{\chi_{i}}$.

\{\chi(x)\} are called the Lyapunov exponents of $x$. \(\{V_{\chi}\}\) is called the Lyapunov filtration of $x$. Property (2) implies that \(\{V_{\chi}\}\) is $A$–invariant: $\Lambda(x)V_{\chi}(x) = V_{\chi}(Tx)$. Property (3) is sometimes called regularity.
2.5 The Multiplicative Ergodic Theorem

Remark: \( V_{\mathcal{L}_0} \setminus V_{\mathcal{L}_{t-1}} \) is \( A \)-invariant, but if \( A(x) \) is not orthogonal, then \( U_{\mathcal{L}_t} \) doesn’t need to be \( A \)-invariant. When \( T \) is invertible, there is a way of writing \( V_{\mathcal{L}_t} = \bigoplus_{j \leq t} H_j \) so that \( A(x)H_j(x) = H_j(Tx) \) and \( \chi(x, \cdot) = \chi_j \) on \( H_j(x) \), see the next section.

2.5.3 The Multiplicative Ergodic Theorem for Invertible Cocycles

Suppose \( A : X \to \text{GL}(n, \mathbb{R}) \). There is a unique extension of the definition of \( A_n(x) \) to non-positive \( n \)'s, which preserves the cocycle identity: \( A_0 := \text{id} \), \( A_{-n} := (A_n \circ T^{-n})^{-1} \). (Start from \( A_{n-n} = A_0 = \text{id} \) and use the cocycle identity.) The following theorem establishes a compatibility between the Lyapunov spectra and filtrations of \( A_n \) and \( A_{-n} \).

**Theorem 2.10.** Let \( (X, \mathcal{B}, m, T) \) be an invertible probability preserving transformation, and \( A : X \to \text{GL}(d, \mathbb{R}) \) a Borel function s.t. \( \ln ||A(x)^{\pm 1}|| \). There are invariant Borel functions \( p(x), \chi_i(x) < \cdots < \chi_{p(x)}(x) \), and a splitting \( \mathbb{R}^d = \bigoplus_{i=1}^{p(x)} H_i(x) \) s.t.

1. \( A_n(x)H_i(x) = H_i(T^n x) \) for all \( n \in \mathbb{Z} \)
2. \( \lim_{n \to \pm \infty} \log ||A_n(x)v|| = \pm \chi_i(x) \) on the unit sphere in \( H_i(x) \).
3. \( \frac{1}{n} \log \sin \angle (H_i(T^n x), H_j(T^n x)) \to 0 \).

**Proof.** Fix \( x \), and let \( \{t_n^1 \leq \cdots \leq t_n^d \} \) be the eigenvalues of \((A_n^d A_n)^{-1/2} \) and \((A_n^d A_{-n})^{-1/2} \). Let \( t_i := \lim(t_i^1)^{1/n}, t_i = \lim(t_i^d)^{1/n} \). These limits exist almost surely, and \( \{t_i \} \) are lists of the Lyapunov exponents of \( A_n \) and \( A_{-n} \), repeated with multiplicity. The proof of the Oseledets theorem shows that

\[
\sum_{k=d-i+1}^{d} \log t_i \equiv \lim_{n \to \infty} \frac{1}{n} \log ||A_{-n}^{i,n}||
\]

\[
= \lim_{n \to \infty} \frac{1}{n} \log \left( |\det A_{-n}| \cdot ||(A_{-n})^{-1} \cdot (A_{n}^{-1})^{d-i}|| \right) \quad \text{(write using e.v.'s)}
\]

\[
\equiv \lim_{n \to \infty} \frac{1}{n} \log \left( |\det A_n \circ T^{-n}| \cdot ||A_{n}^{d-i} \circ T^{-n}|| \right)
\]

\[
\equiv \lim_{n \to \infty} \frac{1}{n} \log \left( |\det A_n|^{-1} \cdot ||A_{n}^{d-i}|| \right) \quad \text{(remark after Kingman’s Theorem)}
\]

\[
= \sum_{k=d-i+1}^{d} \log t_i - \sum_{k=1}^{d-i} \log t_i = - \sum_{k=1}^{i} \log t_i.
\]

Since this is true for all \( i \), \( \log t_i = -\log t_{d-i+1} \).

It follows that if the Lyapunov exponents of \( A_n \) are \( \chi_1 < \cdots < \chi_s \), then the Lyapunov exponents of \( A_{-n} \) are \( -\chi_s < \cdots < -\chi_1 \).

Let \( V^1(x) \subset V^2(x) \subset \cdots \subset V^d(x) \) be the Lyapunov filtration of \( A_n \):

\[
V^i(x) := \{ v : \lim_{n \to \infty} \frac{1}{n} \log ||A_n(x)v|| \leq \chi_i(x) \}.
\]
Let $\mathcal{V}^1(x) \supset \mathcal{V}^2(x) \supset \cdots \supset \mathcal{V}^r(x)$ be the following decreasing filtration, given by

$$\mathcal{V}^r(x) := \{ v : \lim_{n \to \infty} \frac{1}{n} \log \| A_{-n}(x)v \| \leq -\chi_i(x) \}.$$ 

These filtrations are invariant: $A(x)\mathcal{V}^i(x) = \mathcal{V}^i(Tx)$, $A(x)\mathcal{V}^i(x) = \mathcal{V}^i(Tx)$.

Set $H^i(x) := \mathcal{V}^i(x) \cap \mathcal{V}^i(x)$. We must have $A(x)H^i(x) = H^i(Tx)$.

We claim that $\mathbb{R}^d = \bigoplus H^i(x)$ almost surely. It is enough to show that for a.e. $x$,

$$\mathbb{R}^d = \mathcal{V}^i(x) \oplus \mathcal{V}^{i+1}(x),$$

because

$$\begin{align*}
\mathbb{R}^d &\equiv \mathcal{V}^1 = \mathcal{V}^1 \cap [\mathcal{V}^1 \oplus \mathcal{V}^2] \\
&= H^1 \oplus [\mathcal{V}^1 \cap \mathcal{V}^2] = H^1 \oplus \mathcal{V}^2 \\
&= H^1 \oplus [\mathcal{V}^2 \cap (\mathcal{V}^2 \oplus \mathcal{V}^3)] = (\mathcal{V}^2 \oplus \mathcal{V}^3) = \mathcal{V}^d.
\end{align*}$$

Since the spectra of $\Lambda_h$ agree with matching multiplicities, $\dim \mathcal{V}^i + \dim \mathcal{V}^{i+1} = d$. Thus it is enough to show that $E := \{ x : \mathcal{V}^i(x) \cap \mathcal{V}^{i+1}(x) \neq \{0\} \}$ has zero measure for all $i$.

Assume otherwise, then by the Poincaré recurrence theorem, for almost every $x \in E$ there is a sequence $n_k \to \infty$ for which $T^{n_k}(x) \in E$. By the Oseledets theorem, for every $\delta > 0$, there is $N_\delta(x)$ such that for all $n > N_\delta(x)$,

$$\begin{align*}
||A_n(x)u|| &\leq ||u|| \exp[n(\chi_i + \delta)] \quad \text{for all } u \in \mathcal{V}^i \cap \mathcal{V}^{i+1}, \\
||A_{-n}(x)u|| &\leq ||u|| \exp[-n(\chi_{i+1} - \delta)] \quad \text{for all } u \in \mathcal{V}^i \cap \mathcal{V}^{i+1}.
\end{align*}$$

If $n_k > N_\delta(x)$, then $A_{n_k}(x)u \in \mathcal{V}^i(T^{n_k}x) \cap \mathcal{V}^{i+1}(T^{n_k}x)$ and $T^{n_k}(x) \in E$, so

$$||u|| = ||A_{-n_k}(T^{n_k}x)A_{n_k}(x)u|| \leq ||A_{n_k}(x)u|| \exp[-n_k(\chi_{i+1} - \delta)],$$

whence $||A_{n_k}(x)u|| \geq ||u|| \exp[n_k(\chi_{i+1} - \delta)]$. By (2.8),

$$\exp[n_k(\chi_{i+1} + \delta)] \leq \exp[n_k(\chi_i + \delta)],$$

whence $|\chi_{i+1} - \chi_i| < 2\delta$. But $\delta$ was arbitrary, and could be chosen to be much smaller than the gaps between the Lyapunov exponents. With this choice, we get a contradiction which shows that $m(E) = 0$.

Thus $\mathbb{R}^d = \bigoplus H^i(x)$. Evidently, $\mathcal{V}^i \supset \bigoplus_{j \leq i} H^j$ and $\mathcal{V}^{i+1} \cap \bigoplus_{j \geq i} H^j \subset \mathcal{V}^i \cap \mathcal{V}^{i+1} = \{0\}$, so $\mathcal{V}^i = \bigoplus_{j \leq i} H^j$. In the same way $\mathcal{V}^i = \bigoplus_{j \geq i} H^j$. It follows that $H^i \subset (\mathcal{V}^i \setminus \mathcal{V}^{i-1}) \cap (\mathcal{V}^i \setminus \mathcal{V}^{i+1})$. Thus $\lim_{n \to \infty} \frac{1}{n} \log ||A_{n}v|| = \pm \chi_i$ on the unit sphere in $H^i$.

Next we study the angle between $H^i(x)$ and $\widetilde{H}^i(x) := \bigoplus_{j \neq i} H^j(x)$. Pick a basis $(v'_1, \ldots, v'_m)$ for $H^i(x)$. Pick a basis $(w'_1, \ldots, w'_m)$ for $\widetilde{H}^i(x)$. Since $A_{n}(x)$ is invertible, $A_{n}(x)$ maps $(v'_1, \ldots, v'_m)$ onto a basis of $H^i(T^nx)$, and $(w'_1, \ldots, w'_m)$ onto a basis of
2.5 The Multiplicative Ergodic Theorem

\( \bar{H}^i(T^k x) \). Thus if \( v := \wedge v_j, w := \wedge w_j \), then

\[
\sin \angle(H^i(T^k x), \bar{H}(T^k x)) \geq \frac{\|A_n(x)^{\wedge d}(v \wedge w)\|}{\|A_n(x)^{\wedge m} v\| \cdot \|A_n(x)^{\wedge (d-m)} w\|}.
\]

We view \( A_n^{\wedge p} \) as an invertible matrix acting on \( \text{span}\{e_{i_1}^{e_j} \wedge \cdots \wedge e_{i_p}^{e_j} : i_1 < \cdots < i_p \} \) via

\[
(A_n(x)e_{i_1})^* \wedge \cdots \wedge (A_n(x)e_{i_p})^*.
\]

It is clear

\[
A_p(x) := \lim_{n \to \infty} (A_n^{\wedge p})^*(A_n^{\wedge p})^{1/2n} = \left( \lim_{n \to \infty} (A_n^* A_n)^{1/2n} \right)^{\wedge p} = A(x)^{\wedge p},
\]

thus the eigenspaces of \( A_p(x) \) are the tensor products of the eigenspaces of \( A(x) \). This determines the Lyapunov filtration of \( A_p(x)^{\wedge p} \), and implies – by Oseledets theorem – that if \( v_j \in V_{\lambda(j)} \setminus V_{\lambda(j)-1} \), and \( v_1, \ldots, v_k \) are linearly independent, then

\[
\lim_{n \to \infty} \frac{1}{n} \log \|A_n(x)^{\wedge p} \omega\| = \sum_{j=1}^p \chi_{k(j)}, \quad \text{for } \omega := v_1 \wedge \cdots \wedge v_p.
\]

It follows that \( \lim_{n \to \infty} \frac{1}{n} \log |\sin \angle(H^i(T^n x), \bar{H}(T^n x))| \geq 0 \). \( \square \)

Problems

2.1. The Mean Ergodic Theorem for Contractions
Suppose \( H \) is a Hilbert space, and \( U : H \to H \) is a bounded linear operator such that \( \|U\| \leq 1 \). Prove that \( \frac{1}{N} \sum_{k=0}^{N} U^k f \) converges in norm for all \( f \in H \), and the limit is the projection of \( f \) on the space \( \{f : U f = f\} \).

2.2. Ergodicity as a mixing property
Prove that a ppt \((X, \mathcal{B}, \mu, T)\) is ergodic, iff for every \( A, B \in \mathcal{B}, \frac{1}{N} \sum_{k=0}^{N-1} \mu(A \cap T^{-k} B) \xrightarrow{N \to \infty} \mu(A) \mu(B) \).

2.3. Use the pointwise ergodic theorem to show that any two different ergodic invariant probability measures for the same transformation are mutually singular.

2.4. Ergodicity and extremality
An invariant probability measure \( \mu \) is called extremal, if it cannot be written in the form \( \mu = t \mu_1 + (1-t) \mu_2 \) where \( \mu_1, \mu_2 \) are different invariant probability measures, and \( 0 < t < 1 \). Prove that an invariant measure is extremal iff it is ergodic, using the following steps.

1. Show that if \( E \) is a \( T \)-invariant set of non-zero measure, then \( \mu(\cdot \mid E) \) is a \( T \)-invariant measure. Deduce that if \( \mu \) is not ergodic, then it is not extremal.
2. Show that if \( \mu \) is ergodic, and \( \mu = t \mu_1 + (1-t) \mu_2 \) where \( \mu_i \) are invariant, and \( 0 < t < 1 \), then
a. For every $E \in \mathcal{B}$, $\frac{1}{N} \sum_{k=0}^{N-1} 1_{E} \circ T^{k} \overset{N \to \infty}{\longrightarrow} \mu(E)$ $\mu$-a.e. $(i = 1, 2)$.

b. Conclude that $\mu_i(E) = \mu(E)$ for all $E \in \mathcal{B}$ $(i = 1, 2)$. This shows that ergodicity implies extremality.

2.5. Prove that the Bernoulli $(\frac{1}{2}, \frac{1}{2})$–measure is the invariant probability measure for the adding machine (Problem 1.10), by showing that all cylinders of length $n$ must have the same measure as $[0^n]$. Deduce from the previous problem that the adic machine is ergodic.

2.6. Explain why when $f \in L^2$, $E(f|\mathcal{F})$ is the projection of $f$ on $L^2(\mathcal{F})$. Prove:

1. If $\mathcal{F} = \{\varnothing, A, X \setminus A\}$, then $\mathbb{E}(1_B|\mathcal{F}) = \mu(B|A)1_A + \mu(B|A^c)1_{A^c}$.
2. If $\mathcal{F} = \{\varnothing, X\}$, then $\mathbb{E}(f|\mathcal{F}) = \int f d\mu$.
3. If $X = [-1, 1]$ with Lebesgue measure, and $\mathcal{F} = \{A: A$ is Borel and $-A = A\}$, then $\mathbb{E}(f|\mathcal{F}) = \frac{1}{2}[f(x) + f(-x)]$.

2.7. Prove:

1. $f \mapsto \mathbb{E}(f|\mathcal{F})$ is linear, and a contraction in the $L^1$–metric.
2. $f \geq 0 \Rightarrow \mathbb{E}(f|\mathcal{F}) \geq 0$ a.e.
3. If $\varphi$ is convex, then $\mathbb{E}(\varphi \circ f|\mathcal{F}) \leq \varphi(\mathbb{E}(f|\mathcal{F}))$.
4. If $h$ is $\mathcal{F}$–measurable, then $\mathbb{E}(h|\mathcal{F}) = h\mathbb{E}(f|\mathcal{F})$.
5. If $\mathcal{F}_1 \subset \mathcal{F}_2$, then $\mathbb{E}[E(f|\mathcal{F}_2)|\mathcal{F}_1] = E(f|\mathcal{F}_1)$.

2.8. The Martingale Convergence Theorem (Doob)

Suppose $(X, \mathcal{B}, \mu)$ is a probability space, and $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$ are $\sigma$–algebras all of which are contained in $\mathcal{B}$. Let $\mathcal{F} := \sigma(\bigcup_{n \geq 1} \mathcal{F}_n)$ (the smallest $\sigma$–algebra containing the union). If $f \in L^1$, then $\mathbb{E}(f|\mathcal{F}_n) \overset{n \to \infty}{\longrightarrow} \mathbb{E}(f|\mathcal{F})$ a.e. and in $L^1$.

Prove this theorem, using the following steps (W. Parry). It is enough to consider non-negative $f \in L^1$.

1. Prove that $\mathbb{E}(f|\mathcal{F}_n) \overset{n \to \infty}{\longrightarrow} \mathbb{E}(f|\mathcal{F})$ using the following observations:

   a. The convergence holds for all elements of $\bigcup_{n \geq 1} L^1(X, \mathcal{F}_n, \mu)$;
   b. $\bigcup_{n \geq 1} L^1(X, \mathcal{F}_n, \mu)$ is dense in $L^1(X, \mathcal{F}, \mu)$.

2. Set $E_n := \{x : \max_{1 \leq n \leq N} \mathbb{E}(f|\mathcal{F}_n)(x) > a\}$. Show that $\mu(E_n) \leq \frac{1}{n} \int f d\mu$. (Hint: $E = \bigcup_{n \geq 1} \{x : \mathbb{E}(f|\mathcal{F}_n)(x) > \lambda\}$, and $\mathbb{E}(f|\mathcal{F}_n)(x) \leq \lambda$ for $k = 1, \ldots, n-1$.)

3. Prove that $\mathbb{E}(f|\mathcal{F}_n) \overset{n \to \infty}{\longrightarrow} \mathbb{E}(f|\mathcal{F})$ a.e. for every non-negative $f \in L^1$, using the following steps. Fix $f \in L^1$. For every $\epsilon > 0$, choose $n_0$ and $g \in L^1(X, \mathcal{F}_n_0, \mu)$ such that $\|\mathbb{E}(f|\mathcal{F}) - g\|_1 < \epsilon$.

   a. Show that $|\mathbb{E}(f|\mathcal{F}_n) - \mathbb{E}(f|\mathcal{F})| \leq \mathbb{E}(|f - g| |\mathcal{F}_n|) + |\mathbb{E}(f|\mathcal{F}) - g|$ for all $n \geq n_0$. Deduce that
2.5 The Multiplicative Ergodic Theorem

\[ \mu \left[ \limsup_{n \to \infty} |\mathbb{E}(f \mid \mathcal{F}_n) - \mathbb{E}(f \mid \mathcal{F})| > \sqrt{\varepsilon} \right] \leq \mu \left[ \sup_n |\mathbb{E}(|f - g| \mid \mathcal{F}_n)| > \frac{1}{2} \sqrt{\varepsilon} \right] \\
\quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad + \mu \left[ |\mathbb{E}(f \mid \mathcal{F}) - g| > \frac{1}{2} \sqrt{\varepsilon} \right] \]

b. Show that \( \mu \left[ \limsup_{n \to \infty} |\mathbb{E}(f \mid \mathcal{F}_n) - \mathbb{E}(f \mid \mathcal{F})| > \sqrt{\varepsilon} \right] \xrightarrow{\varepsilon \to 0^+} 0 \). (Hint: Prove first that for every \( L^1 \) function \( F \), \( \mu(|F| > a) \leq \frac{1}{a} \|F\|_1 \).)

c. Finish the proof.

2.9 Hopf’s ratio ergodic theorem
Let \((X, \mathcal{B}, \mu, T)\) be a conservative ergodic mpt on a \( \sigma \)-finite measure space. If \( f, g \in L^1 \) and \( f \neq 0 \), then \( \frac{\sum_{n=0}^{\infty} f \circ T^n}{\sum_{n=0}^{\infty} g \circ T^n} \xrightarrow{n \to \infty} \frac{\int f d\mu}{\int g d\mu} \) almost everywhere.

Prove this theorem using the following steps (R. Zweimüller). Fix a set \( A \in \mathcal{B} \) s.t. \( 0 < \mu(A) < \infty \), and let \((A, \mathcal{B}_A, T_A, \mu_A)\) denote the induced system on \( A \) (problem 1.14). For every function \( F \), set

\[ S_n F := F + F \circ T + \cdots + F \circ T^{n-1} \]
\[ S_n^A F := F + F \circ T_A + \cdots + F \circ T_A^{n-1} \]

1. Read problem 1.14, and show that a.e. \( x \) has an orbit which enters \( A \) infinitely many times. Let \( 0 < \tau_1(x) < \tau_2(x) < \cdots \) be the times when \( T^T(x) \in A \).

2. Suppose \( f \geq 0 \). Prove that for every \( n \in (\tau_{k-1}(x), \tau_k(x)] \) and a.e. \( x \in A \),

\[ \frac{(S_{k-1}^A f)(x)}{(S_k^A 1_A)(x)} \leq \frac{(S_n f)(x)}{(S_n 1_A)(x)} \leq \frac{(S_k^A f)(x)}{(S_k^A 1_A)(x)} \]

3. Verify that \( S_k^A 1_A = j \) a.e. on \( A \), and show that \( (S_n f)(x)/(S_n 1_A)(x) \xrightarrow{n \to \infty} \frac{1}{\mu(A)} \int f d\mu \) a.e. on \( A \).

4. Finish the proof.

Notes for chapter 2

For a comprehensive reference to ergodic theorems, see [2]. The mean ergodic theorem was proved by von Neumann. The pointwise ergodic theorem was proved by Birkhoff. By now there are many proofs of this theorem. The one we use is taken from [1], where it is attributed to Kamae — who apparently found it using ideas from nonstandard analysis. The subadditive ergodic theorem was first proved by Kingman. The proof we give is due to Steele [5]. The multiplicative ergodic theorem is due to Oseledets. The proof we use is due to Raghunathan and Ruelle, and is taken from [4]. The Martingale convergence theorem (problem 2.8) is due to Doob. The proof sketched in problem 2.8 is taken from [2]. The proof of Hopf’s ratio er-
Ergodic theorem sketched in problem 2.9 is due to R. Zweimüller and is taken from [6].

References


Chapter 3
Spectral Theory

3.1 The spectral approach to ergodic theory

A basic problem in ergodic theory is to determine whether two ppt are measure theoretically isomorphic. This is done by studying invariants: properties, quantities, or objects which are equal for any two isomorphic systems. The idea is that if two ppt have different invariants, then they cannot be isomorphic. Ergodicity and mixing are examples of invariants for measure theoretic isomorphism.

An effective method for inventing invariants is to look for a weaker equivalence relation, which is better understood. Any invariant for the weaker equivalence relation is automatically an invariant for measure theoretic isomorphism. The spectral point of view is based on this approach.

The idea is to associate to the ppt $(X, B, \mu, T)$ the operator $U_T : L^2(X, B, \mu) \to L^2(X, B, \mu)$, $U_T f = f \circ T$. This is an isometry of $L^2$ (i.e. $\|U_T f\|_2 = \|f\|_2$ and $\langle U_T f, U_T g \rangle = \langle f, g \rangle$). It is useful here to think of $L^2$ as a Hilbert space over $\mathbb{C}$.

**Definition 3.1.** Two ppt $(X, B, \mu, T)$, $(Y, C, \nu, S)$ are called spectrally isomorphic, if their associated $L^2$-isometries $U_T$ and $U_S$ are unitarily equivalent, namely if there exists a linear operator $W : L^2(X, B, \mu) \to L^2(Y, C, \nu)$ s.t.

1. $W$ is invertible;
2. $\langle W f, W g \rangle = \langle f, g \rangle$ for all $f, g \in L^2(X, B, \mu)$;
3. $W U_T = U_S W$.

It is easy to see that any two measure theoretically isomorphic ppt are spectrally isomorphic, but we will see later that there are Bernoulli schemes which are spectrally isomorphic but not measure theoretically isomorphic.

**Definition 3.2.** A property of ppt is called a spectral invariant, if whenever it holds for $(X, B, \mu, T)$, it holds for all ppt which are spectrally isomorphic to $(X, B, \mu, T)$.

**Proposition 3.1.** Ergodicity and mixing are spectral invariants.
Proof. Suppose \((X, \mathcal{B}, \mu, T)\) is a ppt, and let \(U_T\) be as above. The trick is to phrase ergodicity and mixing in terms of \(U_T\).

Ergodicity is equivalent to the statement “all invariant \(L^2\)-functions are constant”, which is the same as saying that \(\dim \{f : U_T f = f\} = 1\). Obviously, this is a spectral invariant.

Mixing is equivalent to the following statement: \(\dim \{f : U_T f = f\} = 1\), and
\[
\left\langle f, U^n_T g \right\rangle \longrightarrow_{n \to \infty} \langle f, 1 \rangle \overline{\langle g, 1 \rangle} \text{ for all } f, g \in L^2.
\]

To see that this property is preserved by spectral isomorphisms, note that if \(\dim \{f : U_T f = f\} = 1\), then any unitary equivalence \(W\) satisfies \(W1 = c\) with \(|c| = 1\).

The spectral point of view immediately suggests the following invariant.

**Definition 3.3.** Suppose \((X, \mathcal{B}, \mu, T)\) is a ppt. If \(f : X \to \mathbb{C}, f \in L^2\) satisfies \(f \circ T = \lambda f\), then we say that \(f\) is an *eigenfunction* and that \(\lambda\) is an *eigenvalue*. The *point spectrum* \(T\) is the set \(H(T) := \{\lambda \in \mathbb{C} : f \circ T = \lambda f\}\).

\(H(T)\) is a countable subgroup of the unit circle (problem 3.1). Evidently \(H(T)\) is a spectral invariant of \(T\).

It is easy to see using Fourier expansions that for the irrational rotation \(R_\alpha\), \(H(R_\alpha) = \{\alpha^k : k \in \mathbb{Z}\}\) (problem 3.2), thus irrational rotations by different angles are non-isomorphic.

Here are other related invariants:

**Definition 3.4.** Given a ppt \((X, \mathcal{B}, \mu, T)\), let \(V_d := \operatorname{span}\{\text{eigenfunctions}\}\). We say that \((X, \mathcal{B}, \mu, T)\) has
1. *discrete spectrum* (sometime called *pure point spectrum*), if \(V_d = L^2\),
2. *continuous spectrum*, if \(V_d = \{\text{constants}\}\) (i.e. is smallest possible),
3. *mixed spectrum*, if \(V_d \neq L^2, \{\text{constants}\}\).

Any irrational rotation has discrete spectrum (problem 3.2). Any mixing transformation has continuous spectrum, because a non-constant eigenfunction \(f \circ T = \lambda\) satisfies
\[
\langle f, f \circ T^n \rangle \longrightarrow_{n \to \infty} \|f\|^2_2 \neq (\int f)^2
\]
along any \(n_k \to \infty\) s.t. \(\lambda^{n_k} \to 1\). (To see that \(\|f\|^2_2 \neq (\int f \, d\mu)^2\) for all non-constant functions, apply Cauchy-Schwarz to \(f - \int f\), or note that non-constant \(L^2\) functions have positive variance.)

The invariant \(H(T)\) is tremendously successful for transformations with discrete spectrum:

**Theorem 3.1 (Discrete Spectrum Theorem).** Two ppt with discrete spectrum are measure theoretically isomorphic if they have the same group of eigenvalues.

But this invariant cannot distinguish transformations with continuous spectrum. In particular - it is unsuitable for the study of mixing transformations.
3.2 Weak mixing

3.2.1 Definition and characterization

We saw that if a transformation is mixing, then it does not have non-constant eigenfunctions. But the absence of non-constant eigenfunctions is not equivalent to mixing (see problems 3.8–3.10 for an example). Here we study the dynamical significance of this property. First we give it a name.

Definition 3.5. A ppt is called weak mixing, if every \( f \in L^2 \) s.t. \( f \circ T = \lambda f \) a.e. is constant almost everywhere.

Theorem 3.2. The following are equivalent for a ppt \( (X, \mathcal{B}, \mu, T) \) on a Lebesgue space:

1. weak mixing;
2. for all \( E, F \in \mathcal{B} \), \( \frac{1}{N} \sum_{k=0}^{N-1} |\mu(E \cap T^{-n}F) - \mu(E)\mu(F)| \xrightarrow{N \to \infty} 0 \);
3. for every \( E, F \in \mathcal{B} \), \( \exists \mathcal{N} \subset \mathbb{N} \) of density zero (i.e. \( |\mathcal{N} \cap [1,N]|/N \xrightarrow{N \to \infty} 0 \)) s.t. \( \mu(E \cap T^{-n}F) \xrightarrow{N \not\in \mathcal{N} \to \infty} \mu(E)\mu(F) \);
4. \( T \times T \) is ergodic.

Proof. We prove \((2) \Rightarrow (3) \Rightarrow (4) \Rightarrow (1)\). The remaining implication \((1) \Rightarrow (2)\) requires additional preparation, and will be shown later.

The implication \((2) \Rightarrow (3)\) is a general fact from calculus (Koopman–von Neumann Lemma): If \( a_n \) is a bounded sequence of non-negative numbers, then \( \frac{1}{N} \sum_{n=1}^{N} a_n \to 0 \) if there is a set of zero density \( \mathcal{N} \subset \mathbb{N} \) s.t. \( a_n \xrightarrow{N \not\in \mathcal{N} \to \infty} 0 \) (Problem 3.3).

We show that \((3) \Rightarrow (4)\). Let \( \mathcal{I} \) be the semi-algebra \( \{E \times F : E, F \in \mathcal{B}\} \) which generates \( \mathcal{B} \otimes \mathcal{B} \), and fix \( E_i \times F_i \in \mathcal{I} \). By (3), \( \exists \mathcal{N} \subset \mathbb{N} \) of density zero s.t.

\[
\mu(E_i \cap T^{-n}F_i) \xrightarrow{\mathcal{N} \not\in \mathcal{N} \to \infty} \mu(E_i)\mu(F_i) \quad (i = 1, 2).
\]

The set \( \mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2 \) also has zero density, and

\[
\mu(E_i \cap T^{-n}F_i) \xrightarrow{\mathcal{N} \not\in \mathcal{N} \to \infty} \mu(E_i)\mu(F_i) \quad (i = 1, 2).
\]

Writing \( m = \mu \times \mu \) and \( S = T \times T \), we see that this implies that

\[
m[(E_1 \times E_2) \cap S^{-n}(F_1 \times F_2)] \xrightarrow{\mathcal{N} \not\in \mathcal{N} \to \infty} m(E_1 \times F_1)m(E_2 \times F_2),
\]

whence \( \frac{1}{N} \sum_{k=0}^{N-1} m[(E_1 \times F_1) \cap S^{-n}(E_2 \times F_2)] \xrightarrow{N \to \infty} m(E_1 \times F_1)m(E_2 \times F_2) \). In summary, \( \frac{1}{N} \sum_{k=0}^{N-1} m[A \cap S^{-n}B] \xrightarrow{N \to \infty} m(A)m(B) \) for all \( A, B \in \mathcal{I} \).

Since \( \mathcal{I} \) generates \( \mathcal{B} \otimes \mathcal{B} \) the above holds for all \( A, B \in \mathcal{B} \otimes \mathcal{B} \), and this implies that \( T \times T \) is ergodic.
Proof that (4)⇒ (1): Suppose \( T \) were not weak mixing, then \( T \) has an non-constant eigenfunction \( f \) with eigenvalue \( \lambda \). The eigenvalue \( \lambda \) has absolute value equal to one, because \( |\lambda|\|f\|_2 = ||f| \circ T||_2 = \|f\|_2 \). Thus
\[
F(x,y) = f(x)f(y)
\]
is \( T \times T \)-invariant. Since \( f \) is non-constant, \( F \) is non-constant, and we get a contradiction to the ergodicity of \( T \times T \).

The proof that (1)⇒ (2) is presented in the next section. \( \square \)

### 3.2.2 Spectral measures and weak mixing

It is convenient to introduce the following notation \( U^n_T := (U^n_T)^{|n|} \) where \( n < 0 \), where \( U^n_T \) is the unique operator s.t. \( \langle U^n_T f, g \rangle = \langle f, U_T g \rangle \) for all \( g \in L^2 \). This makes sense even if \( U_T \) is not invertible. The reader can check that when \( U_T \) is invertible, \( U^{-1}_T = (U_T^{-1}) \) so that there is no risk of confusion.

We are interested in the behavior of \( U^n_T f \) as \( n \to \pm \infty \). To study it, it is enough to study \( U_T: H_f \to H_f \), where \( H_f := \text{span} \{ U^n_T f : n \in \mathbb{Z} \} \).

It turns out that \( U_T : H_f \to H_f \) is unitarily equivalent to the operator \( M : g(z) \mapsto z g(z) \) on \( L^2(S^1, \mathcal{B}(S^1), \nu_f) \) where \( \nu_f \) is some finite measure on \( S^1 \), called the spectral measure of \( f \), which contains all the information on \( U_T : H_f \to H_f \).

To construct it, we need the following important tool from harmonic analysis. Recall that The \( n \)-th Fourier coefficient of \( \mu \) is the number \( \hat{\mu}(n) = \int_{S^1} z^n d\mu(z) \).

**Theorem 3.3 (Herglotz).** A sequence \( \{r_n\}_{n \in \mathbb{Z}} \) is the sequence of Fourier coefficients of a positive Borel measure on \( S^1 \) iff \( r_{-n} = \overline{r_n} \) and \( \{r_n\} \) is positive definite:
\[
\sum_{n,m=-N}^N r_{n-m}a_m\overline{a_n} \geq 0 \text{ for all sequences } \{a_n\} \text{ and } N. \text{ This measure is unique.}
\]

It is easy to check that \( r_n = \langle U^n_T f, f \rangle \) is positive definite (to see this expand \( \langle \sum_{n=-N}^N a_n U^n_T f, \sum_{m=-N}^N a_m U^m_T f \rangle \) noting that \( \langle U^n_T f, U^m_T f \rangle = \langle U^{-n}_T f, f \rangle \)).

**Definition 3.6.** Suppose \( (X, \mathcal{B}, \mu, T) \) is a ppt, and \( f \in L^2 \setminus \{0\} \). The spectral measure of \( f \) is the unique measure \( \nu_f \) on \( S^1 \) s.t. \( (f \circ T^n, f) = \int_{S^1} z^n d\nu_f \) for \( n \in \mathbb{Z} \).

**Proposition 3.2.** Let \( H_f := \text{span} \{ U^n_T f : n \in \mathbb{Z} \} \), then \( U_T : H_f \to H_f \) is unitarily equivalent to the operator \( g(z) \mapsto zg(z) \) on \( L^2(S^1, \mathcal{B}(S^1), \nu_f) \).

**Proof.** By the definition of the spectral measure,
\[
\left\| \sum_{n=-N}^N a_n z^n \right\|_{L^2(\nu_f)}^2 = \left\langle \sum_{n=-N}^N a_n z^n, \sum_{m=-N}^N a_m z^m \right\rangle = \sum_{n,m=-N}^N a_n\overline{a_m} \int_{S^1} z^{n-m} d\nu_f(z)
\]
\[
= \sum_{n,m=-N}^N a_n\overline{a_m} \langle U^n_T f, U^m_T f \rangle = \sum_{n=-N}^N a_n \overline{a_n} \langle U^n_T f, f \rangle = \left\| \sum_{n=-N}^N a_n U^n_T f \right\|_{L^2(\mu)}^2
\]
3.2 Weak mixing

In particular, if \( \sum_{n=-N}^{N} a_n U_n^k f = 0 \) in \( L^2(\mu) \), then \( \sum_{n=-N}^{N} a_n z^n = 0 \) in \( L^2(\nu_f) \). It follows that \( W: U_n^k f \mapsto z^n \) extends to a linear map from \( \text{span}\{U_n^k f : n \in \mathbb{Z}\} \) to \( L^2(\nu_f) \).

This map is an isometry, and it is bounded. It follows that \( W \) extends to a linear isometry \( W: H_f \to L^2(\nu_f) \). The image of \( W \) contains all the trigonometric polynomials, therefore \( W(H_f) \) is dense in \( L^2(\nu_f) \). Since \( W \) is an isometry, its image is closed (exercise). It follows that \( W \) is an isometric bijection from \( H_f \) onto \( L^2(\nu_f) \).

Since \( \langle Wu_f | g(z) \rangle = z \langle Wu_g(z) \rangle \) on \( \text{span}\{U_n^k f : n \in \mathbb{Z}\} \), \( Wu_f(z) = zg(z) \) on \( H_f \), and so \( W \) is the required unitary equivalence.

**Proposition 3.3.** If \( T \) is weak mixing ppt on a Lebesgue space, then all the spectral measures of \( f \in L^2 \) s.t. \( \int f = 0 \) are non-atomic (this explains the terminology “continuous spectrum”).

**Proof.** Suppose \( f \in L^2 \) has integral zero and that \( \nu_f \) has an atom \( \lambda \in S^1 \). We construct an eigenfunction (with eigenvalue \( \lambda \)). Consider the sequence \( \frac{1}{N} \sum_{n=0}^{N-1} \lambda^{-n} U_n^k f \).

This sequence is bounded in norm, therefore has a weakly convergent subsequence (here we use the fact that \( L^2 \) is separable — a consequence of the fact that \( (X, \mathcal{B}, \mu) \) is a Lebesgue space):

\[
\frac{1}{N_k} \sum_{n=0}^{N_k-1} \lambda^{-n} U_n^k f \xrightarrow{w_{N \to \infty}} g.
\]

The limit \( g \) must satisfy \( \langle U_T g, h \rangle = \langle \lambda g, h \rangle \) (check!), therefore it must be an eigenfunction with eigenvalue \( \lambda \).

But it could be that \( g = 0 \). We rule this out using the assumption that \( \nu_f \{ \lambda \} \neq 0 \):

\[
\langle g, f \rangle = \lim_{k \to \infty} \frac{1}{N_k} \sum_{n=0}^{N_k-1} \lambda^{-n} \langle U_n^k f, f \rangle = \lim_{k \to \infty} \frac{1}{N_k} \sum_{n=0}^{N_k-1} \int \lambda^{-n} z^n d\nu_f(z)
\]

\[
= \nu_f\{\lambda\} + \lim_{k \to \infty} \frac{1}{N_k} \sum_{n=0}^{N_k-1} \int_{S^1 \setminus \{\lambda\}} \lambda^{-n} z^n d\nu_f(z)
\]

\[
= \nu_f\{\lambda\} + \lim_{k \to \infty} \int_{S^1 \setminus \{\lambda\}} \frac{1 - 1 - \lambda^{-N_k} z^{N_k}}{1 - \lambda^{-1} z} d\nu_f(z).
\]

The limit is equal to zero, because the integrand tends to zero and is uniformly bounded (by one). Thus \( \langle g, f \rangle = \nu_f\{\lambda\} \neq 0 \), whence \( g \neq 0 \). \( \square \)

**Lemma 3.1.** Suppose \( T \) is a ppt on a Lebesgue space. If \( T \) is weak mixing, then for every \( f \in L^2 \), \( \frac{1}{N} \sum_{k=0}^{N-1} | \int f \cdot T^n d\mu - (\int f d\mu)^2 | \xrightarrow{N \to \infty} 0 \).

**Proof.** It is enough to treat the case when \( \int f d\mu = 0 \). Let \( \nu_f \) denote the spectral measure of \( f \), then
Every constant $g$.

If $T$ is weak mixing, then for all $f$

**Proposition 3.4.** We can now complete the proof of the theorem in the previous section:

$$
\frac{1}{N} \sum_{k=0}^{N-1} \left| \int f \circ T^k d\mu \right|^2 = \frac{1}{N} \sum_{k=0}^{N-1} \left| \int (U_T^n f, f) \right|^2 = \frac{1}{N} \sum_{k=0}^{N-1} \left| \int_s \varphi^k d\nu_f(z) \varphi^k d\nu_f(z) \right|
$$

$$
= \frac{1}{N} \int_s \left( \sum_{k=0}^{N-1} \int_s \varphi^k d\nu_f(z) \right) \left( \sum_{k=0}^{N-1} \int_s \varphi^k d\nu_f(z) \right)
$$

$$
= \frac{1}{N} \sum_{k=0}^{N-1} \int_s \int_s \frac{1}{N} \left( \sum_{k=0}^{N-1} \varphi^k w^k \right) d\nu_f(z) d\nu_f(w)
$$

The integrand tends to zero and is bounded outside $\Delta := \{(z,w) : z = w\}$. If we can show that $(\nu_T \times \nu_T)(\Delta) = 0$, then it will follow that $\frac{1}{N} \sum_{k=0}^{N-1} \int f \circ T^k d\mu \rightarrow 0$.

This is indeed the case: $T$ is weak mixing, so by the previous proposition $\nu_T$ is non-atomic, whence $(\nu_T \times \nu_T)(\Delta) = \int \mu_T \nu_T \{w \} d\nu_T(w) = 0$ by Fubini-Tonelli.

It remains to note that by the Koopman - von Neumann theorem, for every bounded non-negative sequence $a_n$, $\frac{1}{N} \sum_{k=1}^{N} a_n^2 \rightarrow 0$ iff $\frac{1}{N} \sum_{k=1}^{N} a_n \rightarrow 0$, because both conditions are equivalent to saying that $a_n$ converges to zero outside a set of indices of density zero.

We can now complete the proof of the theorem in the previous section:

**Proposition 3.4.** If $T$ is weak mixing, then for all $f, g \in L^2$,

$$
\frac{1}{N} \sum_{k=0}^{N-1} \left| \int g \cdot f \circ T^k d\mu - \left( \int f d\mu \right) \left( \int g d\mu \right) \right| \rightarrow 0. \quad (3.1)
$$

**Proof.** Assume first $T$ is invertible, then $U_T$ is invertible, with a bounded inverse (equal to $U_T^{-1}$). Fix $f \in L^2$, and set

$$
S(f) := \text{span}\{U_T^k f : k \in \mathbb{Z}\}.
$$

Write $L^2 = S(f) + \{\text{constants}\} + [S(f) + \{\text{constants}\}]^\perp$.

1. Every $g \in S(f)$ satisfies (3.1), because $S(f)$ is generated by functions of the form $g := U_T^k f$, and these functions satisfy (3.1) by Lemma 3.1.
2. Every constant $g$ satisfies (3.1) trivially.
3. Every $g \perp S(f) + \{\text{constants}\}$ satisfies (3.1) because $\langle g, f \circ T^n \rangle$ is eventually zero.

It follows that every $g \in L^2$ satisfies (3.1).

Now consider the case of a non-invertible ppt. Let $(\tilde{X}, \tilde{\mathcal{B}}, \tilde{\mu}, \tilde{T})$ be the natural extension. A close look at the definition of $\tilde{\mathcal{B}}$ shows that if $\tilde{f} : \tilde{X} \rightarrow \mathbb{R}$ is $\tilde{\mathcal{B}}$-measurable, then the value of $\tilde{f}(\ldots, x_{-1}, x_0, x_1, \ldots)$ is completely determined by $x_0$. Moreover, $\tilde{f} : \tilde{X} \rightarrow \mathbb{C}$ is of the form $f \circ \pi$ where $f$ is $\mathcal{B}$-measurable. Thus every eigenfunction for $\tilde{T}$ is a lift of an eigenfunction for $T$. It follows that if $T$ is weak mixing, then $\tilde{T}$ is weak mixing.
By the first part of the proof, \( \tilde{T} \) satisfies (3.1). Since \( T \) is a factor of \( T \), it also satisfies (3.1). \( \square \)

### 3.3 The Koopman operator of a Bernoulli scheme

In this section we analyze the Koopman operator of an invertible Bernoulli scheme. The idea is to produce an orthonormal basis for \( L^2 \) which makes the action of \( U_T \) transparent.

We cannot expect to diagonalize \( U_T \): Bernoulli schemes are mixing, so they have no non-constant eigenfunctions. But we shall see that we can get the following nice structure:

**Definition 3.7.** An invertible ppt is said to have *countable Lebesgue spectrum* if \( L^2 \) has an orthonormal basis of the form \( \{1\} \cup \{f_{\lambda,j} : \lambda \in \Lambda, j \in \mathbb{Z}\} \) where \( \Lambda \) is countable, and \( U_T f_{\lambda,j} = f_{\lambda,j+1} \) for all \( i,j \).

The reason for the terminology is that the spectral measure of each \( f_{\lambda,j} \) is proportional to the Lebesgue measure on \( S^1 \) (problem 3.6).

**Example.** The invertible Bernoulli scheme with probability vector \((\frac{1}{2}, \frac{1}{2})\) has countable Lebesgue spectrum.

**Proof.** The phase space is \( X = \{0, 1\}^\mathbb{Z} \). Define for every finite non-empty \( A \subset \mathbb{Z} \) the function \( \phi_A(x) := \prod_{j \in A} (-1)^{x_j} \). Define \( \phi_\emptyset := 1 \). Then,

1. if \( A \neq B \), then \( \phi_A \perp \phi_B \);
2. \( \text{span}\{\phi_A : A \subset \mathbb{Z} \text{ finite}\} \) is algebra of functions which separates points, and contains the constants.

By the Stone-Weierstrass theorem, \( \text{span}\{\phi_A : A \subset \mathbb{Z} \text{ finite}\} = L^2 \), so \( \{\phi_A\} \) is an orthonormal basis of \( L^2 \). This is called the *Fourier–Walsh system*.

Note that \( U_T \phi_A = \phi_{A+1} \), where \( A+1 := \{a+1 : a \in A\} \). Take \( \Lambda \) the set of equivalence classes of the relation \( A \sim B \iff \exists c \text{ s.t. } A = c + B \). Let \( A_\lambda \) be a representative of \( \lambda \in \Lambda \). The basis is \( \{1\} \cup \{\phi_{A_\lambda + n} : \lambda \in \Lambda, n \in \mathbb{Z}\} = \{\text{Fourier Walsh functions}\} \).

It is not easy to produce such bases for other Bernoulli schemes. But they exist. To prove this we introduce the following sufficient condition for countable Lebesgue spectrum, which turns out to be satisfied by many smooth dynamical systems:

**Definition 3.8.** An invertible ppt \((X, \mathcal{B}, \mu, T)\) is called a *K automorphism* if there is a \( \sigma \)-algebra \( \mathcal{A} \subset \mathcal{B} \) s.t.

1. \( T^{-1} \mathcal{A} \subset \mathcal{A} \);
2. \( \mathcal{A} \) generates \( \mathcal{B} \): \( \sigma(\bigcup_{n \in \mathbb{Z}} T^{-n} \mathcal{A}) = \mathcal{B} \text{ mod } \mu \);\(^1\)
3. the tail of \( \mathcal{A} \) is trivial: \( \bigcap_{n=0}^{\infty} T^{-n} \mathcal{A} = \{\emptyset, X\} \text{ mod } \mu \).

\(^1\) \( \mathcal{F}_1 \subset \mathcal{F}_2 \text{ mod } \mu \) is for all \( \mathcal{F}_1 \subset \mathcal{F}_2 \) there is a set \( F_2 \subset \mathcal{F}_2 \) s.t. \( \mu(F_1 \triangle F_2) = 0 \), and \( \mathcal{F}_1 = \mathcal{F}_2 \text{ mod } \mu \) iff \( \mathcal{F}_1 \subset \mathcal{F}_2 \text{ mod } \mu \) and \( \mathcal{F}_2 \subset \mathcal{F}_1 \text{ mod } \mu \).
Proposition 3.5. Every invertible Bernoulli scheme has the K property.

Proof. Let \((S^2, \mathcal{B}(S^2), \mu, T)\) be a Bernoulli scheme, i.e. \(\mathcal{B}(S^2)\) is the sigma algebra generated by cylinders \(-\ell | a_{-\ell}, \ldots, a_0| := \{x \in S^2 : x_i = a_i \ (−\ell \leq i \leq 0)\}, T\) is the left shift map, and \(\mu(k[a_{−\ell}, \ldots, a_0]) = p_{a_{−\ell}} \cdots p_{a_0}\).

Call a cylinder non-negative, if it is of the form \(\{a_0, \ldots, a_n\}\). Let \(\mathcal{A}\) be the sigma algebra generated by all non-negative cylinders. It is clear that \(T^{-1} \mathcal{A} \subset \mathcal{A}\) and that \(\bigcup_{n \in \mathbb{Z}} T^{-n} \mathcal{A}\) generates \(\mathcal{B}(S^2)\). We show that the measure of every element of \(\bigcap_{n=0}^{\infty} T^{-n} \mathcal{A}\) is either zero or one. Probabilists call the elements of this intersection tail events. The fact that every tail event for a sequence of independent identically distributed random variables has probability zero or one is called “Kolmogorov’s zero–one law”.

Two measurable sets \(A, B\) are called independent, if \(\mu(A \cap B) = \mu(A) \mu(B)\). For Bernoulli schemes, any two cylinders with non-overlapping set of indices is independent (check). Thus for every cylinder \(B\) of length \(|B|\),

\[
B \text{ is independent of } T^{-|B|} A \text{ for all non-negative cylinders } A.
\]

It follows that \(B\) is independent of every element of \(T^{-|B|} \mathcal{A}\) (a monotone class theorem argument). Thus every cylinder \(B\) is independent of every element of \(\bigcap_{n \geq 1} T^{-n} \mathcal{A}\). Thus every element of \(\mathcal{B}\) is independent of every element of \(\bigcap_{n \geq 1} T^{-n} \mathcal{A}\) (another monotone class theorem argument).

This means that every \(E \in \bigcap_{n \geq 1} T^{-n} \mathcal{A}\) is independent of itself. Thus \(\mu(E) = \mu(E \cap E) = \mu(E)^2\), whence \(\mu(E) = 0\) or 1.

\[\square\]

Proposition 3.6. Every K automorphism on a non-atomic standard probability space has countable Lebesgue spectrum.

Proof. Let \((X, \mathcal{B}, \mu, T)\) be a K automorphism of a non-atomic standard probability space. Since \((X, \mathcal{B}, \mu)\) is a non-atomic standard space, \(L^2(X, \mathcal{B}, \mu)\) is (i) infinite dimensional, and (ii) separable.

Let \(\mathcal{A}\) be a sigma algebra in the definition of the K property. Set \(V := L^2(X, \mathcal{A}, \mu)\). This is a closed subspace of \(L^2(X, \mathcal{B}, \mu)\), and

1. \(U_T(V) \subseteq V\), because \(T^{-1} \mathcal{A} \subset \mathcal{A}\);
2. \(\bigcup_{n \in \mathbb{Z}} U_T^n(V)\) is dense in \(L^2(X, \mathcal{B}, \mu)\), because \(\bigcup_{n \in \mathbb{Z}} T^{-n} \mathcal{A}\) generates \(\mathcal{B}\), so every \(B \in \mathcal{B}\) can be approximated by a finite disjoint union of elements of \(\bigcup_{n \in \mathbb{Z}} T^{-n} \mathcal{A}\);
3. \(\bigcap_{n=1}^{\infty} U_T^n(V) = \{\text{constant functions}\}\), because \(\bigcap_{n \geq 1} T^{-n} \mathcal{A} = \{\emptyset, X\}\) mod \(\mu\).

Now let \(W := V \oplus U_T(V)\) (the orthogonal complement of \(U_T(V)\) in \(V\)). For all \(n > 0\), \(U_T^n(W) \subset U_T^n(V) \subset U_T^n(W) \perp W\). Thus \(W \perp U_T^n(W)\) for all \(n > 0\). Since \(U_T^{-1}\) is an isometry, \(W \perp U_T^{-1}(W)\) for all \(n < 0\). It follows that

\[
L^2(X, \mathcal{B}, \mu) = \{\text{constants}\} \oplus \bigoplus_{n \in \mathbb{Z}} U_T^n(W) \quad (\text{orthogonal sum}).
\]

If \(\{f_\lambda : \lambda \in \Lambda\}\) is an orthonormal basis for \(W\), then the above implies that
is an orthonormal basis of \( L^2(X, \mathcal{B}, \mu) \) (check!).

This is almost the full countable Lebesgue spectrum property. It remains to show that \(|\Lambda| = \aleph_0\), \(|\Lambda| \leq \aleph_0\) because \( L^2(X, \mathcal{B}, \mu) \) is separable. We show that \(|\Lambda| = \aleph_0\).

\[ \forall N \exists A_1, \ldots, A_N \in \mathcal{A} \text{ pairwise disjoint sets, with positive measure.} \]  \hspace{1cm} (3.2)

Suppose we know this. Pick \( f \in W \setminus \{0\} \) (otherwise \( L^2 = \{ \text{constants} \}\) and \((X, \mathcal{B}, \mu)\) is atomic). Set \( w_i := f1_{A_i} \circ T \) with \( A_1, \ldots, A_N \) as above, then (i) \( w_i \) are linearly independent (because they have disjoint supports); (ii) \( w_i \in V \) (because \( T^{-1}A_i \subseteq T^{-1}\mathcal{A}_\mu \), so \( w_i \) is \( \mathcal{A}_\mu \)-measurable); and (iii) \( w_i \perp U_T(V) \) (check, using \( f \in W \)). It follows that \( \dim(W) \geq N \). Since \( N \) was arbitrary, \( \dim(W) = \infty \).

Here is the proof of (3.2). Since \((X, \mathcal{B}, \mu)\) is non-atomic, \( \exists B_1, \ldots, B_N \in \mathcal{B} \) pairwise disjoint with positive measure. By assumption, \( \bigcup_{n \in \mathbb{Z}} T^n\mathcal{A} \) generates \( \mathcal{B} \), thus we can approximate \( B_i \) arbitrarily well by elements of \( \bigcup_{n \in \mathbb{Z}} T^n\mathcal{A} \). By assumption, \( \mathcal{A} \subseteq T\mathcal{A} \). This means that we can approximate \( B_i \) arbitrarily well by sets from \( T^n\mathcal{A} \) by choosing \( n \) sufficiently large. It follows that \( L^2(X, T^n\mathcal{A}, \mu) \) has dimension at least \( N \). This forces \( T^n\mathcal{A} \) to contain at least \( N \) pairwise disjoint sets of positive measure. It follows that \( \mathcal{A} \) contains at least \( N \) pairwise disjoint sets of positive measure.

**Corollary 3.1.** All systems with countable Lebesgue spectrum, whence all invertible Bernoulli schemes, are spectrally isomorphic.

**Proof.** Problem 3.7. \hspace{1cm} \square

But it is not true that all Bernoulli schemes are measure theoretically isomorphic. To prove this one needs new (non-spectral) invariants. Enter the *measure theoretic entropy*, which we discuss in the next chapter.

**Problems**

3.1. Suppose \((X, \mathcal{B}, \mu, T)\) is an ergodic ppt on a Lebesgue space, and let \( H(T) \) be its group of eigenvalues.

1. show that if \( f \) is an eigenfunction, then \(|f| = \text{const. a.e.}\), and that if \( \lambda, \mu \in H(T) \), then so do \( \lambda \mu, \lambda / \mu \).

2. Show that eigenfunctions of different eigenvalue are orthogonal. Deduce that \( H(T) \) is a countable subgroup of the unit circle.

3.2. Prove that the irrational rotation \( R_\alpha \) has discrete spectrum, and calculate \( H(R_\alpha) \).

3.3. **Koopman - von Neumann Lemma**

Suppose \( a_n \) is a bounded sequence of non-negative numbers. Prove that \( \frac{1}{N} \sum_{n=1}^{N} a_n \rightarrow \)
0 iff there is a set of zero density \( \mathcal{N} \subset \mathbb{N} \) s.t. \( a_n \xrightarrow{\mathcal{N} \not
ot\ni n \to \infty} 0 \). Guidance: Fill in the details in the following argument.

1. Suppose \( \mathcal{N} \subset \mathbb{N} \) has density zero and \( a_n \xrightarrow{\mathcal{N} \not
ot\ni n \to \infty} 0 \), then \( \frac{1}{N} \sum_{n=1}^{N} a_n \to 0 \).

2. Now assume that \( \frac{1}{N} \sum_{n=1}^{N} n = \frac{1}{a_n} \to 0 \).
   a. Show that \( \mathcal{M}_m := \{ k : a_k > 1/m \} \) form an increasing sequence of sets of density zero.
   b. Fix \( \varepsilon_i \downarrow 0 \), and choose \( k_i \uparrow \infty \) such that if \( n > k_i \), then \( (1/n) |_{\mathcal{N} \cap (k_i, k_{i+1})} | < \varepsilon_i \).
   c. Show that \( \mathcal{N} := \bigcup_{i} \mathcal{N}_i \cap (k_i, k_{i+1}) \) has density zero.

3.4. Here is a sketch of an alternative proof of proposition 3.4, which avoids natural extensions (B. Parry). Fill in the details.

1. Set \( H := L^2 \), \( V := \bigcap_{n \geq 0} U^T_n(H) \), and \( W := H \ominus U^T H := \{ g \in H, g \perp U^T H \} \).
   a. \( H = V \oplus \bigcup_{k} (U^T_k)^{\perp} + \bigcup_{k} (U^T_2)^{\perp} + \cdots \)
   b. \( \{ U^T_k \} \) is decreasing, \( \{ (U^T_k)^{\perp} \} \) us increasing.
   c. \( H = V \oplus \bigoplus_{k=1}^{\infty} U^T_k W \) (orthogonal space decomposition).

2. \( U_T : V \to V \) has a bounded inverse (hint: use the fact from Banach space theory that any bounded linear operator between mapping one Banach space onto another Banach space which is one-to-one, has a bounded inverse).
3. (3.1) holds for any \( f, g \in V \).
4. if \( g \in U^T_k W \) for some \( k \), then (3.1) holds for all \( f \in L^2 \).
5. if \( g \in V \), but \( f \in U^T_k W \) for some \( k \), then (3.1) holds for \( f, g \).
6. (3.1) holds for all \( f, g \in L^2 \).

3.5. Show that every invertible ppt with countable Lebesgue spectrum is mixing, whence ergodic.

3.6. Suppose \( (X, \mathcal{B}, \mu, T) \) has countable Lebesgue spectrum. Show that \{ \( f \in L^2 : \int f = 0 \) \} is spanned by functions \( f \) whose spectral measures \( \nu_f \) are equal to the Lebesgue measure on \( S^1 \).

3.7. Show that any two ppt with countable Lebesgue spectrum are spectrally isomorphic.

3.8. Cutting and Stacking and Chacon’s Example

This is an example of a ppt which is weak mixing but not mixing. The example is a certain map of the unit interval, which preserves Lebesgue’s measure. It is constructed using the method of “cutting and stacking” which we now explain.

Let \( A_0 = [1, \frac{2}{3}) \) and \( R_0 := [\frac{2}{3}, 1] \) (thought of as reservoir).

Step 1: Divide \( A_0 \) into three equal subintervals of length \( \frac{2}{9} \). Cut a subinterval \( B_0 \) of length \( \frac{2}{9} \) from the left end of the reservoir.
3.3 The Koopman operator of a Bernoulli scheme

- Stack the three thirds of $A_0$ one on top of the other, starting from the left and moving to the right.
- Stick $B_0$ between the second and third interval.
- Define a partial map $f_1$ by moving points vertically in the stack. The map is defined everywhere except on $R \setminus B_0$ and the top floor of the stack. It can be viewed as a partially defined map of the unit interval.

Update the reservoir: $R_1 := R \setminus B_0$. Let $A_1$ be the base of the new stack (equal to the rightmost third of $A_0$).

**Step 2:** Cut the stack vertically into three equal stacks. The base of each of these thirds has length $\frac{1}{3} \times \frac{2}{9}$. Cut an interval $B_1$ of length $\frac{1}{3} \times \frac{2}{9}$ from the left side of the reservoir $R_1$.

- Stack the three stacks one on top of the other, starting from the left and moving to the right.
- Stick $B_1$ between the second stack and the third stack.
- Define a partial map $f_2$ by moving points vertically in the stack. This map is defined everywhere except the union of the top floor floor and $R_1 \setminus B_1$.

Update the reservoir: $R_2 := R_1 \setminus B_1$. Let $A_2$ be the base of the new stack (equal to the rightmost third of $A_1$).

**Step 3:** Cut the stack vertically into three equal stacks. The base of each of these thirds has length $\frac{1}{3} \times \frac{2}{9}$. Cut an interval $B_2$ of length $\frac{1}{3} \times \frac{2}{9}$ from the left side of the reservoir $R_2$.

- Stack the three stacks one on top of the other, starting from the left and moving to the right.
- Stick $B_2$ between the second stack and the third stack.
- Define a partial map $f_3$ by moving points vertically in the stack. This map is defined everywhere except the union of the top floor floor and $R_2 \setminus B_2$.

Update the reservoir: $R_3 := R_2 \setminus B_2$. Let $A_3$ be the base of the new stack (equal to the rightmost third of $A_2$).

Continue in this manner, to obtain a sequence of partially defined maps $f_n$. There is a canonical way of viewing the intervals composing the stacks as of subintervals of the unit interval. Using this identification, we may view $f_n$ as partially defined maps of the unit interval.

1. Show that $f_n$ is measure preserving where it is defined (the measure is Lebesgue’s measure). Calculate the Lebesgue measure of the domain of $f_n$.
2. Show that $f_{n+1}$ extends $f_n$ (i.e. the maps agree on the intersection of their domains). Deduce that the common extension of $f_n$ defines an invertible probability preserving map of the open unit interval. This is Chacon’s example. Denote it by $(I, \mathcal{F}, m, T)$.
3. Let $\ell_n$ denote the height of the stack at step $n$. Show that the sets $\{T^i(A_n) : i = 0, \ldots, \ell_n, n \geq 1\}$ generate the Borel $\sigma$–algebra of the unit interval.
3.9. (Continuation) Prove that Chacon’s example is weak mixing using the following steps. Suppose $f$ is an eigenfunction with eigenvalue $\lambda$.

1. We first show that if $f$ is constant on $A_n$ for some $n$, then $f$ is constant everywhere. ($A_n$ is the base of the stack at step $n$.)
   a. Let $\ell_n$ denote the height of the stack at step $k$. Show that $A_{n+1} \subset A_n$, and $T^{\ell_n}(A_{n+1}) \subset A_n$. Deduce that $\lambda^{\ell_n} = 1$.
   b. Prove that $\lambda^{\ell_{n+1}} = 1$. Find a recursive formula for $\ell_n$. Deduce that $\lambda = 1$.
   c. The previous steps show that $f$ is an invariant function. Show that any invariant function which constant on $A_n$ is constant almost everywhere.

2. We now consider the case of a general $L^2$–eigenfunction.
   a. Show, using Lusin’s theorem, that there exists an $n$ such that $f$ is nearly constant on most of $A_n$. (Hint: part 3 of the previous question).
   b. Modify the argument done above to show that any $L^2$–eigenfunction is constant almost everywhere.

3.10. (Continuation) Prove that Chacon’s example is not mixing, using the following steps.

1. Inspect the image of the top floor of the stack at step $n$, and show that for every $n$ and $0 \leq k \leq \ell_{n-1}$, $m(T^k A_n \cap T^{k+\ell_n} A_n) \geq \frac{1}{3} m(T^k A_n)$. 
2. Use problem 3.8 part 3 and an approximation argument to show that for every Borel set $E$ and $\varepsilon > 0$, $m(E \cap T^n E) \geq \frac{1}{3} m(E) - \varepsilon$ for all $n$. Deduce that $T$ cannot be mixing.

**Notes to chapter 3**

The spectral approach to ergodic theory is due to von Neumann. For a thorough modern introduction to the theory, see Nadkarni’s book [1]. Our exposition follows in parts the books by Parry [2] and Petersen [1]. A proof of the discrete spectrum theorem mentioned in the text can be found in Walters’ book [3]. A proof of Herglotz’s theorem is given in [2].

**References**

Chapter 4
Entropy

In the end of the last chapter we saw that every two Bernoulli schemes are spectrally isomorphic (because they have countable Lebesgue spectrum). The question whether any two Bernoulli schemes are measure theoretically isomorphic was a major open question in the field. It was solved by Kolmogorov and Sinai, through the invention of a new invariant: entropy. Later, Ornstein proved that this invariant is complete within the class of Bernoulli schemes.

4.1 Information content and entropy

Let $\alpha = \{A_1, \ldots, A_N\}$ be a measurable partition of $(X, \mathcal{B}, \mu)$, and suppose $T : X \to X$ is measurable. Let

$$\alpha(x) := \text{The element of } \alpha \text{ which contains } x.$$

The itinerary of $x$ is $(\alpha(x), \alpha(Tx), \alpha(T^2x), \ldots)$, a sequence taking values $A_1, \ldots, A_N$.

Suppose $x$ is not known, but $(\alpha(x), \alpha(Tx), \ldots, \alpha(T^{n-1}x))$ is known; How much uncertainty do we have regarding $\alpha(T^nx)$?

**Example 1: Irrational Rotations.** $R_\theta(z) = e^{i\theta}z$ with $0 < \theta < \frac{1}{100}$ irrational, and $\alpha := \{A_0, A_1\}$ where $A_0 := \{e^{i\theta} : 0 \leq \theta < \pi\}$, and $A_1 := \{e^{i\theta} : \pi \leq \theta < 2\pi\}$. If the five first symbols are

$$(1, 1, 1, 0, 0, \ldots)$$

then we are certain that the next one is 0. This is the case whenever there is a 1 there. In the case $(0, 0, 0, 0, 0, \ldots)$ we can guess that the next one is 0 with certainty of 99%. Thus knowing $\alpha(T^3x)$ isn’t ‘worth much’, if we already know $(\alpha(x), \alpha(Tx), \ldots, \alpha(T^4x))$. We can guess it with high certainty anyway.

**Example 2: Angle Doubling.** $T(z) = z^2$, same partition. Knowing the first five symbols tells us nothing on the sixth one: It is zero or one with probability 50%. So the ‘information content’ of $\alpha(T^3x)$ given $(\alpha(x), \alpha(Tx), \ldots, \alpha(T^4x))$ is constant.
The following question arises: Let \((X, \mathcal{B}, \mu)\) be a probability space. Suppose \(x \in X\) is unknown. How to quantify the ‘information content’ \(I(A)\) of the statement ‘\(x\) belongs to \(A\)?’

Our guiding principle is to think of the information content of an event \(E\) as of the uncertainty lost when learning that \(x \in A\). Thus the information content of an event of small probability is large. Here are some intuitively clear requirements that a good definition of \(I(A)\) should satisfy:

1. \(I(A)\) should be a continuous function of the probability of \(A\);
2. \(I(A)\) should be non-negative, decreasing in \(\mu(A)\), and if \(\mu(A) = 1\) then \(I(A) = 0\);
3. If \(A, B\) are independent, i.e. \(\mu(A \cap B) = \mu(A)\mu(B)\), then \(I(A \cap B) = I(A) + I(B)\).

**Proposition 4.1.** The only functions \(\varphi : [0, 1] \rightarrow \mathbb{R}^+\) such that \(I(A) = \varphi[\mu(A)]\) satisfies the above axioms for all probability spaces \((X, \mathcal{B}, \mu)\) are \(c \log\) with \(c < 0\).

We leave the proof as an exercise. This leads to the following definition.

**Definition 4.1 (Shannon).** Let \((X, \mathcal{B}, \mu)\) be a probability space.

1. The Information Content of a set \(A \in \mathcal{B}\) is \(I_\mu(A) := - \log \mu(A)\)
2. The Information Function of a finite measurable partition is
   \[
   I_\mu(\alpha)(x) := \sum_{A \in \alpha} I_\mu(A)1_A(x) = - \sum_{A \in \alpha} \log \mu(A)1_A(x)
   \]
3. The Entropy of a finite measurable partition is the average of the information content of its elements:
   \[
   H_\mu(\alpha) := \int_X I_\mu(\alpha)d\mu = - \sum_{A \in \alpha} \mu(A) \log \mu(A).
   \]

Conventions: The base of the log is 2; \(0 \log 0 = 0\).

The are important conditional versions of these notions:

**Definition 4.2.** Let \((X, \mathcal{B}, \mu)\) be a probability space, and suppose \(\mathcal{F}\) is a sub-\(\sigma\)-algebra of \(\mathcal{B}\). We use the notation \(\mu(A|\mathcal{F})(x) := \mathbb{E}(1_A|\mathcal{F})(x)\) (as \(L^1\)-elements).

1. The information content of \(A\) given \(\mathcal{F}\) is \(I_\mu(A|\mathcal{F})(x) := - \log \mu(A|\mathcal{F})(x)\)
2. The information function of a finite measurable partition \(\alpha\) given \(\mathcal{F}\) is \(I_\mu(\alpha|\mathcal{F}) := \sum_{A \in \alpha} I_\mu(A|\mathcal{F})1_A\)
3. The conditional entropy of \(\alpha\) given \(\mathcal{F}\) is \(H_\mu(\alpha|\mathcal{F}) := \int I_\mu(\alpha|\mathcal{F})d\mu\).

**Convention:** Let \(\alpha, \beta\) be partitions; We write \(H_\mu(\alpha|\beta)\) for \(H_\mu(\alpha|\sigma(\beta))\), where \(\sigma(\beta) := \text{smallest } \sigma\text{-algebra which contains } \beta\).

The following formulæ are immediate:

\[
H_\mu(\alpha|\mathcal{F}) = - \int_X \sum_{A \in \alpha} \mu(A|\mathcal{F})(x) \log \mu(A|\mathcal{F})(x)d\mu(x)
\]
\[
H_\mu(\alpha|\beta) = - \sum_{B \in \beta} \mu(B) \sum_{A \in \alpha} \mu(A|B) \log \mu(A|B), \text{ where } \mu(A|B) = \frac{\mu(A \cap B)}{\mu(B)}.
\]
4.2 Properties of the entropy of a partition

We need some notation and terminology. Let \( \alpha, \beta \) be two countable partitions.

1. \( \sigma(\alpha) \) is the smallest \( \sigma \)-algebra which contains \( \alpha \);
2. \( \alpha \leq \beta \) means that \( \alpha \subseteq \sigma(\beta) \mod \mu \), i.e. every element of \( \alpha \) is equal up to a set of measure zero to an element of \( \sigma(\beta) \). Equivalently, \( \alpha \leq \beta \) if every element of \( \alpha \) is equal up to a set of measure zero to a union of elements of \( \beta \). We say that \( \beta \) is finer than \( \alpha \), and that \( \alpha \) is coarser than \( \beta \).
3. \( \alpha = \beta \mod \mu \) iff \( \alpha \subseteq \beta \mod \mu \) and \( \beta \subseteq \alpha \mod \mu \).
4. \( \alpha \vee \beta \) is the smallest partition which is finer than both \( \alpha \) and \( \beta \). Equivalently, \( \alpha \vee \beta := \{ A \cap B : A \in \alpha, B \in \beta \} \).

If \( \mathcal{F}_1, \mathcal{F}_2 \) are two \( \sigma \)-algebras, then \( \mathcal{F}_1 \vee \mathcal{F}_2 \) is the smallest \( \sigma \)-algebra which contains \( \mathcal{F}_1, \mathcal{F}_2 \).

4.2.1 The entropy of \( \alpha \vee \beta \)

It is useful to think of a partition \( \alpha = \{ A_1, \ldots, A_n \} \) as of the “information” which element of \( \alpha \) contains an unknown \( x \).

We state and prove a formula which says that the information content of \( \alpha \) and \( \beta \) is the information content of \( \alpha \) plus the information content of \( \beta \) given the knowledge \( \alpha \).

**Theorem 4.1 (The Basic Identity).** Suppose \( \alpha, \beta \) are measurable countable partitions, and assume \( H_\mu(\alpha), H_\mu(\beta) < \infty \), then

1. \( I_\mu(\alpha \vee \beta | \mathcal{F}) = I_\mu(\alpha | \mathcal{F}) + I_\mu(\beta | \mathcal{F} \vee \sigma(\alpha)) \);
2. \( H_\mu(\alpha \vee \beta) = H_\mu(\alpha) + H_\mu(\beta | \alpha) \).

**Proof.** We calculate \( I_\mu(\beta | \mathcal{F} \vee \sigma(\alpha)) \):

\[
I_\mu(\beta | \mathcal{F} \vee \sigma(\alpha)) = -\sum_{B \in \beta} 1_B \log \mu(B | \mathcal{F} \vee \sigma(\alpha))
\]

**Claim:** \( \mu(B | \mathcal{F} \vee \sigma(\alpha)) = \sum_{A \in \alpha} 1_A \frac{\mu(B \cap A | \mathcal{F})}{\mu(A | \mathcal{F})} \).

1. This expression is \( \mathcal{F} \vee \sigma(\alpha) \)-measurable
2. Observe that \( \mathcal{F} \vee \sigma(\alpha) = \{ \mu_{A \in \alpha} A \cap F_A : F_A \in \mathcal{F} \} \) (this is a \( \sigma \)-algebra which contains \( \alpha \) and \( \mathcal{F} \)). Thus every \( \mathcal{F} \vee \sigma(\alpha) \)-measurable function is of the form \( \sum_{A \in \alpha} 1_A \varphi_A \) with \( \varphi_A \mathcal{F} \)-measurable. It is therefore enough to check test functions of the form \( 1_A \varphi \) with \( \varphi \in L^\infty(\mathcal{F}) \). For such functions
This proves the first part of the theorem.

Lemma 4.1. Let

\[ \mu(B \cap A|\mathcal{F}) \frac{d\mu}{\mu(A|\mathcal{F})} \]

then

\[ \sum_{A \in \alpha} 1_A \phi \]

Proof. This is because \( \phi(x) = -\log t \) is strictly concave. Let \( m := \sum p_i x_i \). If \( m = 0 \) then the lemma is obvious, so suppose \( m > 0 \). It is an exercise in calculus to see that \( \phi(t) \leq \phi(m) + \phi'(m)(t - m) \) for \( t \in [0,1] \), with equality iff \( t = m \). In the particular case \( m = \sum p_i x_i \) and \( t = x_i \), we get

\[ p_i \phi(x_i) \leq p_i \phi(m) + \phi'(m)(p_i x_i - p_i m) \]

with equality iff \( p_i = 0 \) or \( x_i = m \).

Summing over \( i \), we get

\[ \sum p_i \phi(x_i) \leq \phi(m) + \phi'(m)(\sum p_i x_i - m) = \phi(m) \]. There is an equality iff for every \( i \) \( p_i = 0 \) or \( x_i = m \). \( \square \)

4.2.2 Convexity properties

Lemma 4.1. Let \( \phi(t) := -t \log t \), then for every probability vector \( (p_1, \ldots, p_n) \) and \( x_1, \ldots, x_n \in [0,1] \), \( \phi(p_1 x_1 + \cdots + p_n x_n) \geq p_1 \phi(x_1) + \cdots + p_n \phi(x_n) \), with equality iff all the \( x_i \) with \( i \) s.t. \( p_i \neq 0 \) are equal.

Proof. This is because \( \phi(t) \) is strictly concave. Let \( m := \sum p_i x_i \). If \( m = 0 \) then the lemma is obvious, so suppose \( m > 0 \). It is an exercise in calculus to see that \( \phi(t) \leq \phi(m) + \phi'(m)(t - m) \) for \( t \in [0,1] \), with equality iff \( t = m \). In the particular case \( m = \sum p_i x_i \) and \( t = x_i \), we get

\[ p_i \phi(x_i) \leq p_i \phi(m) + \phi'(m)(p_i x_i - p_i m) \]

with equality iff \( p_i = 0 \) or \( x_i = m \).

Summing over \( i \), we get \( \sum p_i \phi(x_i) \leq \phi(m) + \phi'(m)(\sum p_i x_i - m) = \phi(m) \). There is an equality iff for every \( i \) \( p_i = 0 \) or \( x_i = m \). \( \square \)

Proposition 4.2 (Convexity properties). Let \( \alpha, \beta, \gamma \) be countable measurable partitions with finite entropies, then

1. \( \alpha \leq \beta \Rightarrow H_\mu(\alpha|\gamma) \leq H_\mu(\beta|\gamma) \)
2. \( \alpha \leq \beta \Rightarrow H_\mu(\gamma|\alpha) \geq H_\mu(\gamma|\beta) \)
4.3 The Metric Entropy

**Proof.** The basic identity shows that $\beta \vee \gamma$ has finite entropy, and so $H_\mu(\beta|\gamma) = H_\mu(\alpha|\gamma) + H_\mu(\beta|\gamma) \geq H_\mu(\alpha|\gamma)$.

For the second inequality, note that $\phi(t) = -t \log t$ is strictly concave (i.e. its negative is convex), therefore by Jensen’s inequality

\[
H_\mu(\gamma|\alpha) = \int C \in \gamma \phi[\mathbb{E}(C|\sigma(\gamma))|\sigma(\alpha)]d\mu \geq \int C \in \gamma \mathbb{E}([\phi[\mathbb{E}(1_C|\sigma(\beta))]|\sigma(\alpha)]|d\mu = \sum C \in \gamma \phi[\mathbb{E}(1_C|\sigma(\beta))]|d\mu = H_\mu(\gamma|\beta),
\]

proving the inequality. \qed

4.2.3 Information and independence

We say that two partitions are independent, if $\forall A \in \alpha, B \in \beta, \mu(A \cap B) = \mu(A) \mu(B)$. This the same as saying that the random variables $\alpha(x), \beta(x)$ are independent.

**Proposition 4.3 (Information and Independence).** $H_\mu(\alpha \vee \beta) \leq H_\mu(\alpha) + H_\mu(\beta)$ with equality iff $\alpha, \beta$ are independent.

**Proof.** $H_\mu(\alpha \vee \beta) = H_\mu(\alpha) + H_\mu(\beta)$ iff $H_\mu(\alpha|\beta) = H_\mu(\alpha)$. But

\[
H_\mu(\alpha|\beta) = -\sum_{B \in \beta} \mu(B) \sum_{A \in \alpha} \mu(A|B) \log \mu(A|B).
\]

Let $\phi(t) = -t \log t$. We have:

\[
\sum_{A \in \alpha} \sum_{B \in \beta} \mu(B) \phi[\mu(A|B)] = \sum_{A \in \alpha} \phi[\mu(A)].
\]

But $\phi$ is strictly concave, so $\sum_{B \in \beta} \mu(B) \phi[\mu(A|B)] \leq \phi[\mu(A)]$, with equality iff $\mu(A|B)$ are equal for all $B \in \beta$ s.t. $\mu(B) \neq 0$.

We conclude that $\mu(A|B) = c(A)$ for all $B \in \beta$ s.t. $\mu(B) \neq 0$. For such $B$, $\mu(A \cap B) = c(A) \mu(B)$. Summing over $B$, gives $c(A) = \mu(A)$ and we obtain the independence condition. \qed

4.3 The Metric Entropy

4.3.1 Definition and meaning

**Definition 4.3 (Kolmogorov, Sinai).** The metric entropy of a ppt $(X, \mathcal{B}, \mu, T)$ is defined to be
\[ h_\mu(T) := \sup \{ h_\mu(T, \alpha) : \alpha \text{ is a countable measurable partition s.t. } H_\mu(\alpha) < \infty \}, \]

where \( h_\mu(T, \alpha) := \lim_{n \to \infty} \frac{1}{n} H_\mu(\bigvee_{i=0}^{n-1} T^{-i} \alpha). \)

**Proposition 4.4.** The limit which defines \( h_\mu(T, \alpha) \) exists.

It can be shown that the supremum is attained by finite measurable partitions (problem 4.9).

**Proof.** Write \( \alpha_n := \bigvee_{i=0}^{n-1} T^{-i} \alpha. \) Then \( a_n := H_\mu(\alpha_n) \) is subadditive, because \( a_{n+m} := H_\mu(\alpha_{n+m}) \leq H_\mu(\alpha_n) + H_\mu(T^{-n} \alpha_m) = a_n + a_m. \)

We claim that any sequence of numbers \( \{a_n\}_{n \geq 1} \) which satisfies \( a_{n+m} \leq a_n + a_m \) converges to a limit (possibly equal to minus infinity), and that this limit is \( \inf[a_n/n] \).

Fix \( n \). Then for every \( m, m = kn + r, 0 \leq r \leq n - 1 \), so

\[ a_m \leq k a_n + a_r. \]

Dividing by \( m \), we get that for all \( m > n \)

\[ \frac{a_m}{m} \leq \frac{k a_n + a_r}{kn + r} \leq \frac{a_n}{n} + \frac{a_r}{m}, \]

whence \( \limsup(a_m/m) \leq a_n/n \). Since this is true for all \( n \), \( \limsup a_m/m \leq \inf a_n/n \).

But it is obvious that \( \liminf a_m/m \geq \inf a_n/n \), so the limsup and liminf are equal, and their common value is \( \inf a_n/n \).

We remark that in our case the limit is not minus infinity, because \( H_\mu(\bigvee_{i=0}^{n-1} T^{-i} \alpha) \) are all non-negative.

\( H_\mu(\alpha_n) \) is the average information content in the first \( n \)–digits of the \( \alpha \)–itinerary. Dividing by \( n \) gives the average ‘information per unit time’. Thus the entropy measure the maximal rate of information production the system is capable of generating.

It is also possible to think of entropy as a measure of unpredictability. Let’s think of \( T \) as moving backward in time. Then \( \alpha^* := \sigma(\bigcup_{n=1}^{\infty} T^{-n} \alpha) \) contains the information on the past of the itinerary. Given the future, how unpredictable is the present, on average? This is measured by \( H_\mu(\alpha | \alpha^*_1) \).

**Theorem 4.2.** If \( H_\mu(\alpha) < \infty \), then \( h_\mu(T, \alpha) = H_\mu(\alpha | \alpha^*_1) \), where \( \alpha^*_1 = \sigma(\bigcup_{n=1}^{\infty} T^{-1} \alpha) \).

**Proof.** We show that \( h_\mu(T, \alpha) = H_\mu(\alpha | \alpha^*_1) \). Observe that

\[ H_\mu(\alpha | \alpha^*_0) = H_\mu(\alpha^*_0) - H_\mu(T^{-1} \alpha^*_0) = H_\mu(\alpha^*_0) - H_\mu(\alpha^*_0 - 1). \]

Summing over \( n \), we obtain

\[ H_\mu(\alpha_n) - H_\mu(\alpha) = \sum_{k=1}^{n} H_\mu(\alpha | \alpha^*_k) \]
4.3 The Metric Entropy

Dividing by \( n \) and passing to the limit we get

\[
\hat{h}_\mu(T, \alpha) = \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} H_\mu(\alpha|\alpha_k^T)
\]

It is therefore enough to show that \( H_\mu(\alpha|\alpha_k^T) \xrightarrow{k \to \infty} H_\mu(\alpha|\alpha_1^\infty) \).

This is dangerous!! It is true that \( H_\mu(\alpha|\alpha_k^T) = \int f_\mu(\alpha|\alpha_k^T)d\mu \) and that by the martingale convergence theorem

\[
I_\mu(\alpha|\alpha_k^T) \xrightarrow{k \to \infty} I_\mu(\alpha|\alpha_1^\infty)
\]

But the claim that the integral of the limit is equal to the limit of the integrals requires justification.

If \( |\alpha| < \infty \), then we can bypass the problem by writing

\[
H_\mu(\alpha|\alpha_k^T) = \int \sum_{A \in \alpha} \varphi(A|\alpha_k^T)d\mu, \text{ with } \varphi(t) = -t \log t,
\]

and noting that this function is bounded (by \( |\alpha| \max \varphi \)). Thus the BCT applies and gives \( H_\mu(\alpha|\alpha_k^T) \xrightarrow{k \to \infty} H_\mu(\alpha|\alpha_1^\infty) \).

If \( |\alpha| = \infty \) (but \( H_\mu(\alpha) < \infty \)) then we need to be more clever, and appeal to the following lemma (proved below):

**Lemma 4.2 (Chung–Neveu).** Suppose \( \alpha \) is a countable measurable partition with finite entropy, then the function \( f^\ast := \sup_{t \geq 1} I_\mu(\alpha|\alpha_k^T) \) is absolutely integrable.

The result now follows from the dominated convergence theorem. \( \square \)

Here is the proof of the Chung Neveu Lemma. Fix \( A \in \alpha \), then we may decompose \( A \cap \{f^\ast > t\} \equiv \bigcup_{m \geq 1} A \cap B_m(t; A) \), where \( B_m(t; A) := \{x \in X : m \text{ is the minimal natural number s.t. } -\log_2 \mu(A|\alpha_m^T) > t\} \).

We have

\[
\mu(A \cap B_m(t; A)) = E_\mu(1_A 1_{B_m(t; A)}) = E_\mu(1_A 1_{B_m(t; A)}|\sigma(\alpha_m^T))
\]

\[
= E_\mu(1_{B_m(t; A)} \cdot 1_A|\sigma(\alpha_m^T)), \text{ because } B_m(t; A) \in \sigma(\alpha_m^T)
\]

\[
= E_\mu(1_{B_m(t; A)} \cdot 2^{-\log_2 \mu(A|\alpha_m^T)})
\]

\[
\leq E_\mu(1_{B_m(t; A)} \cdot 2^{-t}) = 2^{-t} \mu(B_m(t; A)).
\]

Summing over \( m \) we see that \( \mu(A \cap \{f^\ast > t\}) \leq 2^{-t} \). Of course we also have \( \mu(A \cap \{f^\ast > t\}) \leq \mu(A) \). Thus \( \mu(A \cap \{f^\ast > t\}) \leq \min\{\mu(A), 2^{-t}\} \).

We now use the following fact from measure theory: If \( g \geq 0 \), then \( \int g d\mu = \int_0^\infty \mu|g > t|dt \).

---

1 Proof: \( f_1 g d\mu = \int g \int_0^\infty 1_{|g| < e(x)}(x,t)dtd\mu(x) = \int_0^\infty \int 1_{|g| < e}(x,t)d\mu(x)dt = \int_0^\infty \mu|g > t|dt \).
\[ \int_A f^* d\mu = \int_0^\infty \mu(A \cap [f^* > t]) dt = \int_0^\infty \min\{\mu(A), 2^{-t}\} dt \]
\[ \leq \int_0^{-\log_2 \mu(A)} \mu(A) dt + \int_{-\log_2 \mu(A)}^\infty 2^{-t} dt = -\mu(A) \log_2 \mu(A) - \frac{2^{-t}}{\ln 2} \int_{-\log_2 \mu(A)}^\infty dt \]
\[ = -\mu(A) \log_2 \mu(A) + \mu(A) / \ln 2. \]

Summing over \( A \in \alpha \) we get that \( \int f^* d\mu \leq H_\mu(\alpha) + (\ln 2)^{-1} < \infty \).

### 4.3.2 The Shannon–McMillan–Breiman Theorem

**Theorem 4.3 (Shannon–McMillan–Breiman).** Let \((X, \mathcal{B}, \mu, T)\) be an ergodic ppt, and \(\alpha\) a countable measurable partition of finite entropy, then

\[ \frac{1}{n} I_\mu(\alpha_0^{-1}) \longrightarrow h_\mu(T, \alpha) \text{ a.e.} \]

In particular, if \(\alpha_n(x) := \text{element of } \alpha_n \text{ which contains } x\), then \( -\frac{1}{n} \log \alpha_n(x) \longrightarrow h_\mu(T, \alpha) \text{ a.e.} \).

**Proof.** We start with the basic identity \( I_\mu(\alpha_0^{-1}) = I_\mu(\alpha_1^{-1} \cup \alpha) = I_\mu(\alpha_1^{-1}) + I_\mu(\alpha|\alpha_1^{-1}). \) This gives

\[ I_\mu(\alpha_n) = I_\mu(\alpha|\alpha_1^n) + I_\mu(\alpha_0^{-1}) \circ T \]
\[ = I_\mu(\alpha|\alpha_1^n) + [I_\mu(\alpha|\alpha_2^n) + I_\mu(\alpha_0^{-2}) \circ T] \circ T \]
\[ = \cdots = \sum_{k=0}^{n-1} I_\mu(\alpha|\alpha_1^{n-k}) \circ T^k \]
\[ = \sum_{k=1}^n I_\mu(\alpha|\alpha_1^k) \circ T^{n-k} \]

By the Martingale Convergence Theorem, \( I_\mu(\alpha|\alpha_1^k) \longrightarrow I_\mu(\alpha|\alpha_1^\infty) \). The idea of the proof is to use this to say

\[ \lim_{n \to \infty} \frac{1}{n} I_\mu(\alpha_n) \equiv \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^n I_\mu(\alpha|\alpha_1^k) \circ T^{n-k} \]
\[ \equiv \lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^n I_\mu(\alpha|\alpha_1^\infty) \circ T^{n-k} \equiv \frac{1}{n} \sum_{k=0}^{n-1} I_\mu(\alpha|\alpha_1^\infty) \circ T^k \]
\[ = \int I_\mu(\alpha|\alpha_1^\infty) d\mu \quad \text{(Ergodic Theorem)} \]
\[ = H_\mu(\alpha|\alpha_1^\infty) = h_\mu(T, \alpha). \]
The point is to justify the question mark. Write \( f_n := I_\mu(\alpha|\alpha_n^\infty) \) and \( f_\infty = I_\mu(\alpha|\alpha_\infty^\infty) \). It is enough to show

\[
\int \limsup_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} |f_k - f_\infty| \circ T^{n-k} d\mu = 0.
\]

(This implies that the limsup is zero almost everywhere.) Set \( F_n := \sup_{k>n} |f_k - f_\infty| \). Then \( F_n \to 0 \) almost everywhere. We claim that \( F_n \to 0 \) in \( L^1 \). This is because of the dominated convergence theorem and the fact that \( F_n \leq 2f^* := 2 \sup_m f_m \in L^1 \) (Chung–Neveu Lemma). Fix some large \( N \), then

\[
\int \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} |f_{n-k} - f_\infty| \circ T^k d\mu = \\
\leq \int \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} F_N \circ T^k d\mu + \int \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} (\sum_{k=0}^{n-1} 2f^* \circ T^k) \circ T^{-n} d\mu \\
= \int F_N d\mu + \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{N-1} 2f^* \circ T^k d\mu = \int F_N d\mu.
\]

Since \( F_N \to 0 \) in \( L^1 \), \( \int F_N d\mu \to 0 \), and this proves that the integral of the limsup is zero. \( \square \)

### 4.3.3 Sinai’s Generator theorem

Let \( \mathcal{F}_1, \mathcal{F}_2 \) be two sub-\( \sigma \)-algebras of a probability space \((X, \mathcal{B}, \mu)\). We write \( \mathcal{F}_1 \subseteq \mathcal{F}_2 \mod \mu \), if \( \forall F_1 \in \mathcal{F}_1, \exists F_2 \in \mathcal{F}_2 \) s.t. \( \mu(F_1 \triangle F_2) = 0 \). We write \( \mathcal{F}_1 = \mathcal{F}_2 \mod \mu \), if both inclusions hold \( \mod \mu \). For example, \( \mathcal{B}(\mathbb{R}) = \mathcal{B}_0(\mathbb{R}) \mod \text{Lebesgue’s measure} \). For every partition \( \alpha \), let

\[
\alpha^\infty = \bigvee_{i=-\infty}^\infty T^{-i} \alpha, \quad \alpha^0 = \bigvee_{i=0}^\infty T^{-i} \alpha
\]

denote the smallest \( \sigma \)-algebras generated by, respectively, \( \bigcup_{i=-\infty}^\infty T^{-i} \alpha \) and \( \bigcup_{i=0}^\infty T^{-i} \alpha \).

**Definition 4.4.** A countable measurable partition \( \alpha \) is called a **generator** for an invertible \((X, \mathcal{B}, \mu, T)\) if \( \bigvee_{i=-\infty}^\infty T^{-i} \alpha = \mathcal{B} \mod \mu \), and a **strong generator**, if \( \bigvee_{i=0}^\infty T^{-i} \alpha = \mathcal{B} \mod \mu \).

(This latter definition makes sense in the non-invertible case as well)
Example: \( \alpha = \{ [0, \frac{1}{2}), [\frac{1}{2}, 1] \} \) is a strong generator for \( T_x = 2x \mod 1 \), because \( \bigvee_{i=0}^{\infty} T^{-i} \alpha = \sigma(\bigcup_{i} \alpha^{n-1}_0) \) is the Borel \( \sigma \)-algebra (it contains all dyadic intervals, whence all open sets).

**Theorem 4.4 (Sinai’s Generator Theorem).** Let \( (X, \mathcal{B}, \mu, T) \) be a ppt. If \( \alpha \) is a generator of finite entropy, then \( h_{\mu}(T) = h_{\mu}(T, \alpha) \).

**Proof.** Fix a finite measurable partition \( \beta \); Must show that \( h_{\mu}(T, \beta) \leq h_{\mu}(T, \alpha) \).

**Step 1.** \( h_{\mu}(T, \beta) \leq h_{\mu}(T, \alpha) + H_{\mu}(\beta | \alpha) \)

\[
\frac{1}{n} H_{\mu}(\beta_0^{n-1}) = \frac{1}{n} \left[ H_{\mu}(\alpha_0^{n-1}) + H_{\mu}(\beta_0^{n-1} | \alpha_0^{n-1}) \right] \\
\leq \frac{1}{n} \left[ H_{\mu}(\alpha_0^{n-1}) + \sum_{k=0}^{n-1} H_{\mu}(T^{-k} \beta | \alpha_0^{n-1}) \right] \\
\leq \frac{1}{n} \left[ H_{\mu}(\alpha_0^{n-1}) + \sum_{k=0}^{n-1} H_{\mu}(T^{-k} \beta | T^{-k} \alpha) \right] \\
= \frac{1}{n} H_{\mu}(\alpha_0^{n-1}) + H_{\mu}(\beta | \alpha).
\]

Now pass to the limit.

**Step 2.** For every \( N \), \( h_{\mu}(T, \beta) \leq h_{\mu}(T, \alpha) + H_{\mu}(\beta | \alpha^N_N) \)

Repeat the previous step with \( \alpha^N_N \) instead of \( \alpha \), and check that \( h_{\mu}(T, \alpha^N_N) = h_{\mu}(T, \alpha) \).

**Step 3.** \( H_{\mu}(\beta | \alpha^N_N) \xrightarrow{N \to \infty} H_{\mu}(\beta | \mathcal{B}) = 0 \).

\[
H_{\mu}(\beta | \alpha^N_N) = \int I_{\mu}(\beta | \alpha^N_N) d\mu = -\sum_{B \in \beta} \int 1_B \log \mu(B | \alpha^N_N) d\mu \\
= -\sum_{B \in \beta} \int \mu(B | \alpha^N_N) \log \mu(B | \alpha^N_N) d\mu \\
= \sum_{B \in \beta} \int \varphi[\log \mu(B | \alpha^N_N)] d\mu \xrightarrow{N \to \infty} \sum_{B \in \beta} \int \varphi[\log \mu(B | \mathcal{B})] d\mu = 0,
\]

because \( \mu(B | \mathcal{B}) = 1_B, \varphi[1_B] = 0 \), and \( |B| < \infty \).

This proves that \( h_{\mu}(T, \alpha) \geq \sup \{ h_{\mu}(T, \beta) : |\beta| < \infty \} \). Problem 4.9 says that this supremum is equal to \( h_{\mu}(T) \), so we are done. \( \square \)
4.4 Examples

4.4.1 Bernoulli schemes

Proposition 4.5. The entropy of the Bernoulli shift with probability vector $p$ is $- \sum p_i \log p_i$. Thus the $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$-Bernoulli scheme and the $(\frac{1}{2}, \frac{1}{2})$-Bernoulli scheme are not isomorphic.

Proof. $\alpha = \{[1], \ldots, [n]\}$ is a strong generator, and

$$H_\mu(\alpha_0^{n-1}) = - \sum_{x_0, \ldots, x_{n-1}} p_{x_0} \cdots p_{x_{n-1}} \left( \log p_{x_0} + \cdots + \log p_{x_{n-1}} \right) = -n \sum p_i \log p_i.$$ 

$\square$

4.4.2 Irrational rotations

Proposition 4.6. The irrational rotation has entropy zero w.r.t. the Haar measure.

Proof. The reason is that it is an invertible transformation with a strong generator. We first explain why any invertible map with a strong generator must have zero entropy. Suppose $\alpha$ is such a strong generator. Then

$$h_\mu(T, \alpha) = H_\mu(\alpha \alpha_0) = H_\mu(T \alpha(T(\alpha_0))) = H_\mu(T \alpha | \alpha_0) = H_\mu(T \alpha | \mathcal{B}) = 0,$$

because $T \alpha \subset \mathcal{B}$.

We now claim that $\alpha := \{A_0, A_1\}$ (the two halves of the circle) is a strong generator. It is enough to show that for every $\epsilon$, $\bigcup_{k \geq 1} \alpha_0^{k-1}$ contains open covers of the circle by open arcs of diameter $< \epsilon$ (because this forces $\alpha_0^w$ to contain all open sets).

It is enough to manufacture one arc of diameter less than $\epsilon$, because the translations of this arc by $k\alpha$ will eventually cover the circle.

But such an arc necessarily exits: Choose some $n$ s.t. $n\alpha \mod 1 \in (0, \epsilon)$. Then $A_1 \setminus T^{-n}A_1 = (A_1 \setminus [A_1 - n\alpha]$ is an arc of diameter less than $\epsilon$.

4.4.3 Markov measures

Proposition 4.7. Suppose $\mu$ is a translation invariant Markov measure with transition matrix $P = (p_{ij})$ and probability vector $(p_i)$. Then $h_\mu(\sigma) = - \sum p_i p_{ij} \log p_{ij}$.

Proof. The natural partition $\alpha = \{[a] : a \in S\}$ is a strong generator.
\[ H_{\mu}(\alpha^0_n) = - \sum_{\xi_0, \ldots, \xi_n \in S} \mu(\xi) \log \mu(\xi) \]
\[ = - \sum_{\xi_0, \ldots, \xi_n \in S} p_{\xi_0} p_{\xi_0 \xi_1} \cdots p_{\xi_{n-1} \xi_n} \log p_{\xi_0} + \log p_{\xi_0 \xi_1} + \cdots + \log p_{\xi_{n-1} \xi_n} \]
\[ = - \sum_{j=0}^{n-1} \sum_{\xi_0, \ldots, \xi_n \in S} p_{\xi_0} p_{\xi_0 \xi_1} \cdots p_{\xi_{j-1} \xi_j} \log p_{\xi_0 \xi_1} + \cdots + \log p_{\xi_{n-1} \xi_n} \]
\[ = - \sum_{j=0}^{n-1} \sum_{\xi_0, \ldots, \xi_n \in S} p_{\xi_0} p_{\xi_0 \xi_1} \cdots p_{\xi_{j-1} \xi_j} p_{\xi_j \xi_{j+1}} \log p_{\xi_0 \xi_1} + \cdots + \log p_{\xi_{n-1} \xi_n} \]
\[ = - \sum_{j=0}^{n-1} \sum_{\xi_0, \ldots, \xi_n \in S} \mu(\sigma^{-j}(\xi_0)) \cdot p_{\xi_0} p_{\xi_0 \xi_1} \cdots p_{\xi_{j-1} \xi_j} p_{\xi_j \xi_{j+1}} \log p_{\xi_0 \xi_1} + \cdots + \log p_{\xi_{n-1} \xi_n} \]
\[ = - \sum_{j=0}^{n-1} \sum_{\xi_0, \ldots, \xi_n \in S} \mu(\sigma^{-j}(\xi_0)) \cdot p_{\xi_0} p_{\xi_0 \xi_1} \cdots p_{\xi_{j-1} \xi_j} p_{\xi_j \xi_{j+1}} \log p_{\xi_0 \xi_1} + \cdots + \log p_{\xi_{n-1} \xi_n} \]
\[ = - \sum_{j=0}^{n-1} \sum_{\xi_0, \ldots, \xi_n \in S} \mu(\sigma^{-j}(\xi_0)) \cdot p_{\xi_0} p_{\xi_0 \xi_1} \cdots p_{\xi_{j-1} \xi_j} p_{\xi_j \xi_{j+1}} \log p_{\xi_0 \xi_1} + \cdots + \log p_{\xi_{n-1} \xi_n} \]
\[ = n \left( - \sum_{i,j} p_{ij} \log p_{ij} \right) - \sum_i p_i \log p_i \]
Now divide by \( n + 1 \) and pass to the limit. \( \square \)

### 4.4.4 Expanding Markov Maps of the Interval

**Theorem 4.5 (Rokhlin formula).** Suppose \( T : [0, 1] \to [0, 1] \) and \( \alpha = \{ I_1, \ldots, I_N \} \) is a partition into intervals s.t.

1. \( \alpha \) is a Markov partition
2. The restriction of \( T \) to \( \alpha \) is \( C^1 \), monotonic, and \( |T'| > \lambda > 1 \)
3. \( T \) has an invariant measure \( \mu \).

Then \( h_\mu(T) = - \int \log \frac{d\mu}{\mu \circ T} d\mu \), where \( (\mu \circ T)(E) = \sum_{A \in \alpha} \mu(T(A \cap E)) \).
4.5 Abramov’s Formula

Proof. One checks that the elements of \( \alpha_0^{n-1} \) are all intervals of length \( O(\lambda^{-n}) \). Therefore \( \alpha \) is a strong generator, whence

\[
h_\mu(T) = h_\mu(T, \alpha) = H_\mu(\alpha|\alpha^\infty_1) = \int I_\mu(\alpha|\alpha^\infty_1) d\mu.
\]

We calculate \( I_\mu(\alpha|\alpha^\infty_1) \). First note that \( \alpha^\infty_1 = T^{-1}(\alpha^\infty_0) = T^{-1}\mathcal{B} \), thus \( I_\mu(\alpha|\alpha^\infty_1) = -\sum_{A \in \alpha} 1_A \log \mu(A|T^{-1}\mathcal{B}) \).

We need to calculate \( \mathbb{E}(|T^{-1}\mathcal{B}) \). For this purpose, introduce the operator \( \hat{T} : L^1 \to L^1 \) given by

\[
(\hat{T} f)(x) = \sum_{y \equiv x} \frac{d\mu}{d\mu \circ T} (y)f(y).
\]

Exercise: Verify: \( \forall \varphi \in L^\infty \) and \( f \in L^1 \), \( \int \varphi \hat{T} f d\mu = \int \varphi \circ T \cdot f d\mu \).

We claim that \( E(f|T^{-1}\mathcal{B}) = (\hat{T} f) \circ T \). Indeed, the \( T^{-1}\mathcal{B} \)–measurable functions are exactly the functions of the form \( \varphi \circ T \) with \( \varphi \mathcal{B} \)–measurable; Therefore \( (\hat{T} f) \circ T \) is \( T^{-1}\mathcal{B} \)–measurable, and

\[
\int \varphi \circ T \cdot \hat{T} f \circ T d\mu = \int \varphi \cdot \hat{T} f d\mu = \int \varphi \circ T \cdot f d\mu,
\]

proving the identity.

We can now calculate and see that

\[
I_\mu(\alpha|\alpha^\infty_1) = -\sum_{A \in \alpha} 1_A(x) \log \mathbb{E}(1_A|T^{-1}\mathcal{B})(x)
\]

\[
= -\sum_{A \in \alpha} 1_A(x) \log \sum_{y \equiv x} \frac{d\mu}{d\mu \circ T} (y)1_A(y) \equiv -\sum_{A \in \alpha} 1_A(x) \log \frac{d\mu}{d\mu \circ T}(x)
\]

\[
= -\log \frac{d\mu}{d\mu \circ T}(x).
\]

We conclude that \( h_\mu(T) = -\int \log \frac{d\mu}{d\mu \circ T}(x) d\mu(x). \) \( \Box \)

4.5 Abramov’s Formula

Suppose \((X, \mathcal{B}, \mu, T)\) is a ppt. A set \( A \) is called spanning, if \( X = \bigcup_{n=0}^{\infty} T^{-n}A \mod \mu \). If \( T \) is ergodic, then every set of positive measure is spanning.

**Theorem 4.6 (Abramov).** Suppose \((X, \mathcal{B}, \mu, T)\) is a ppt on a Lebesgue space, let \( A \) be a spanning measurable set, and let \((A, \mathcal{B}_A, \mu_A, T_A)\) be the induced system, then \( h_{\mu_A}(T_A) = \frac{1}{\mu(A)} h_{\mu}(T) \).

**Proof.** (Scheller) We prove the theorem in the case when \( T \) is invertible. The non-invertible case is handled by passing to the natural extension.
The idea is to show, for as many partitions $\alpha$ as possible, that $h_\mu(T, \alpha) = \mu(A) h_{\mu_A}(T_A, \alpha \cap A)$, where $\alpha \cap A := \{E \cap A : E \in \alpha\}$. As it turns out, this is the case for all partitions s.t. (a) $H_\mu(\alpha) < \infty$; (b) $A^c \in \alpha$; and (c) $\forall n, T_A[\phi_A = n] \in \sigma(\alpha)$.

Here, as always, $\phi_A(x) := 1_A(x) \inf\{n \geq 1 : T^n x \in A\}$ (the first return time).

To see that there are such partitions, we let

$$\xi_A := \{A^c\} \cup T_A \eta_A, \text{ where } \eta_A := \{|\varphi_A = n| : n \in \mathbb{N}\}$$

(the coarsest possible) and show that $H_\mu(\xi_A) < \infty$. A routine calculation shows that $H_\mu(\xi_A) = H_\mu(\{A, A^c\}) + \mu(A) H_\mu(T_A, \eta_A) \leq 1 + H_\mu(\eta_A)$. It is thus enough to show that $-\sum p_n \log p_n < \infty$, where $p_n := \mu\{\varphi_A = n\}$. This is because $\sum n p_n = 1/\mu(A)$ (Kac formula) and the following fact from calculus: probability vectors with finite expectations have finite entropy.

Assume now that $\alpha$ is a partition which satisfies (a)-(c) above. We will use throughout the following fact:

$$A, A^c, [\varphi_A = n] \in \alpha_{1}^{\infty}.$$  \hfill (4.1)

Here is why: $[\varphi_A = n] = T^{-n}T_A[\varphi_A = n] \in T^{-n} \alpha \subset \alpha_{1}^{\infty}$. Since $AA = \bigcup_{n \geq 1} [\varphi_A = n]$, we automatically have $A, A^c \in \alpha_{1}^{\infty}$.

Let $\alpha$ be a finite entropy countable measurable partition of $X$ such that $A^c$ is an atom of $\alpha$ and such that $\alpha \supseteq \xi_A$. In what follows we use the notation $A \cap \alpha := \{B \cap A : B \in \alpha\}$, $\alpha_{1}^{\infty} \cap A := \{B \cap A : B \in \alpha_{1}^{\infty}\}$. Since $H_\mu(\alpha) < \infty$,

$$h_\mu(T, \alpha) = H_\mu(\alpha|\alpha_{1}^{\infty})$$

$$= \int \sum_{B \in A^c|\alpha} 1_B \log \mu(\alpha_{1}^{\infty})d\mu + \int 1_{A^c} \log \mu(\alpha_{1}^{\infty})d\mu$$

$$= \int \sum_{B \in A^c|\alpha} 1_B \log \mu(\alpha_{1}^{\infty})d\mu, \text{ because } A^c \in (\xi_A)_{1}^{\infty} \subseteq \alpha_{1}^{\infty}$$

$$= \mu(A) \int \sum_{A \in A^c|\alpha} 1_B \log \mu_A(B|\alpha_{1}^{\infty})d\mu_A,$$

because $A \in \alpha_{1}^{\infty}$ and $B \subset A$ imply $\mathbb{E}_\mu(1_B|\mathcal{F}) = 1_A \mathbb{E}_\mu(1_B|A \cap \mathcal{F})$.

It follows that $h_\mu(T, \alpha) = \mu(A) H_{\mu_A}(A \cap \alpha|A \cap \alpha_{1}^{\infty})$. We will show later that

$$A \cap \alpha_{1}^{\infty} = \bigvee_{i=1}^{\infty} T_{-i}^{-1}(A \cap \alpha)$$  \hfill (4.2)

This implies that $h_\mu(T, \alpha) = \mu(A) h_{\mu_A}(T_A, A \cap \alpha)$. Passing to the supremum over all $\alpha$ which contain $A^c$ as an atom, we obtain

---

2 Proof: Enumerate $(p_n)$ in a decreasing order: $p_{n_1} \geq p_{n_2} \geq \cdots$. If $C = n p_n$, then $C \geq \sum_{i=1}^{k} n_i p_n \geq p_{n_1} (1 + \cdots + k)$, whence $p_{n_2} = O(k^{1-\epsilon})$. Since $-x \log x = O(x^{1-\epsilon})$ as $x \to 0^+$, this means that $-p_{n_2} \log p_{n_2} = O(k^{1-2\epsilon})$, and so $-\sum p_{n_2} \log p_{n_2} = -\sum p_{n_2} \log p_{n_2} < \infty$. 

The same proof shows that
\[ \mu(A)h_{\mu}(T_A) = \sup \{ h_{\mu}(T, \alpha) : \alpha \geq \xi_A , A^c \in \alpha, H_\mu(\alpha) < \infty \}. \]

(See problem 4.11).

Now \( R' \equiv R \mod \mu \), because \( A \) is spanning, so \( \forall E \in R, E = \bigcup_{n=0}^{\infty} T^{-n}(T^n E \cap A) \mod \mu \), whence \( E \in R' \mod \mu \). This shows Abramov’s formula, given (4.2).

The proof of (4.2):

Proof of \( \subseteq \): Suppose \( B \) is an atom of \( A \cap \bigvee_{j=1}^{n} T^{-j} \alpha \), then \( B \) has the form \( A \cap \bigcap_{j=1}^{n} T^{-j} A_j \) where \( A_j \in \alpha \). Let \( j_1 < j_2 < \cdots < j_N \) be an enumeration of the \( j \)'s s.t. \( A_j \subset A \) (possibly an empty list). Since \( A^c \) is an atom of \( \alpha \), \( A_j = A^c \) for \( j \) not in this list, and so \( B = \bigcap_{k=1}^{N-1} T^{-k} (A_j \cap [\varphi_k = j_{k+1} - j_k]) \cap T^{-N} \varphi_A \). Since \( \eta_A \leq \eta \cap A, B \in \bigvee_{i=1}^{\infty} T_A^{-1} (\alpha \cap A) \).

Proof of \( \supseteq \): \( T_A^{-1}(\alpha \cap A) \leq A \cap \bigvee_{n=1}^{\infty} T^{-i} \alpha \), because if \( B \in \alpha \cap A \), then
\[
T_A^{-1} B = \bigcup_{n=1}^{\infty} T^{-n} (B \cap T_A [\varphi_A = n]) \in \bigvee_{n=1}^{\infty} T^{-n} \alpha \quad \therefore T_A \eta \leq \xi_A \leq A \cap \alpha.
\]

The same proof shows that \( T_A^{-1}(T^{-n} \alpha \cap A) \leq A \cap \bigvee_{i=1}^{\infty} T^{-i} \alpha \). It follows that
\[
T_A^{-2}(\alpha \cap A) \leq T_A^{-1} \left( A \cap \bigvee_{i=1}^{\infty} T^{-i} \alpha \right) \subseteq A \cap \bigvee_{i=1}^{\infty} T_A^{-1} (A \cap T^{-i} \alpha) \subseteq A \cap \bigvee_{i=1}^{\infty} T^{-i} \alpha.
\]

Iterating this procedure we see that \( T_A^{-n}(\alpha \cap A) \leq A \cap \bigvee_{i=1}^{\infty} T^{-i} \alpha \) for all \( n \), and \( \supseteq \) follows.

\[ \square \]

### 4.6 Topological Entropy

Suppose \( T : X \rightarrow X \) is a continuous mapping of a compact topological space \( (X, d) \). Such a map can have many different invariant Borel probability measures. For example, the left shift on \( \{0, 1\}^\mathbb{N} \) has an abundance of Bernoulli measures, Markov measures, and there are many others.

Different measures may have different entropies. What is the largest value possible? We study this question in the context of continuous maps on topological spaces which are compact and metric.

#### 4.6.1 The Adler–Konheim–McAndrew definition

Let \( (X, d) \) be a compact metric space, and \( T : X \rightarrow X \) a continuous map. Some terminology and notation:
1. an open cover of $X$ is a collection of open sets $\mathcal{U} = \{U_\alpha : \alpha \in \Lambda\}$ s.t. $X = \bigcup_{\alpha \in \Lambda} U_\alpha$.

2. if $\mathcal{U} = \{U_\alpha : \alpha \in \Lambda\}$ is an open cover, then $T^{-k}\mathcal{U} := \{T^{-k}U_\alpha : \alpha \in \Lambda\}$. Since $T$ is continuous, this is another open cover.

3. if $\mathcal{U}, \mathcal{V}$ are open covers, then $\mathcal{U} \vee \mathcal{V} := \{U \cap V : U \in \mathcal{U}, V \in \mathcal{V}\}$.

Since $X$ is compact, every open cover of $X$ has a finite subcover. Define

$$N(\mathcal{U}) := \min\{\#\mathcal{V} : \mathcal{V} \subseteq \mathcal{U} \text{ is finite, and } X = \bigcup \mathcal{V}\}.$$ 

It easy to check that $N(\cdot)$ is subadditive in the following sense:

$$N(\mathcal{U} \vee \mathcal{V}) \leq N(\mathcal{U}) + N(\mathcal{V}).$$

**Definition 4.5.** Suppose $T : X \to X$ is a continuous mapping of a compact metric space $(X,d)$, and let $\mathcal{U}$ be an open cover of $X$. The topological entropy of $\mathcal{U}$ is

$$h_{\text{top}}(T,\mathcal{U}) := \lim_{n \to \infty} \frac{1}{n} \log_2 N(\mathcal{U}_{n-1}^0),$$ where $\mathcal{U}_{n-1}^0 := \bigvee_{i=0}^{n-1} T^{-k}\mathcal{U}$. 

The limit exists because of the subadditivity of $N(\cdot)$: $a_n := \log N(\mathcal{U}_{n-1}^0)$ satisfies $a_{m+n} \leq a_m + a_n$, so $\lim a_n/n$ exists.

**Definition 4.6.** Suppose $T : X \to X$ is a continuous mapping of a compact metric space $(X,d)$, then the topological entropy of $T$ is the (possibly infinite)

$$h_{\text{top}}(T) := \sup\{h_{\text{top}}(T,\mathcal{U}) : \mathcal{U} \text{ is an open cover of } X\}.$$ 

The following theorem was first proved by Goodwyn.

**Theorem 4.7.** Suppose $T$ is a continuous mapping of a compact metric space, then every invariant Borel probability measure $\mu$ satisfies $h_\mu(T) \leq h_{\text{top}}(T)$.

**Proof.** Eventually everything boils down to the following inequality, which can be checked using Lagrange multipliers: For every probability vector $(p_1, \ldots, p_k)$,

$$-\sum_{i=1}^{k} p_i \log_2 p_i \leq \log k,$$

with equality iff $p_1 = \cdots = p_k = 1/k$.

Suppose $\mu$ is an invariant probability measure, and let $\alpha := \{A_1, \ldots, A_k\}$ be a measurable partition.

We approximate $\alpha$ by a partition into sets with better topological properties. Fix $\epsilon > 0$ (to be determined later), and construct compact sets

$$B_j \subset A_j \text{ s.t. } \mu(A_j \setminus B_j) < \epsilon \quad (j = 1, \ldots, k).$$
Let $B_0 := X \setminus \bigcup_{j=1}^{b} B_j$ be the remainder (of measure less than $k\epsilon$), and define $\beta = \{B_0; B_1, \ldots, B_k\}$.

Step 1 in the proof of Sinai’s theorem says that $h_\mu(T, \alpha) \leq h_\mu(T, \beta) + H_\mu(\alpha|\beta)$. We claim that $H_\mu(\alpha|\beta)$ can be made uniformly bounded by a suitable choice of $\epsilon = \epsilon(\alpha)$:

$$H_\mu(\alpha|\beta) = - \sum_{A \in \alpha} \sum_{B \in \beta} \mu(A \cap B) \log_2 \mu(A|B)$$

$$= - \sum_{B \in \beta \setminus \{B_0\}} \sum_{A \in \alpha} \mu(A \cap B) \log_2 \mu(A|B) - \sum_{A \in \alpha} \mu(A \cap B_0) \log_2 \mu(A|B_0)$$

$$= \sum_{i=1}^{k} \mu(B_i) \log_2 1 - \sum_{A \in \alpha} \mu(A \cap B_0) \log_2 \mu(A|B_0)$$

$$= - \mu(B_0) \sum_{A \in \alpha} \mu(A|B_0) \log_2 \mu(A|B_0) \leq - \mu(B_0) \log(\#\alpha) \leq k\epsilon \cdot \log_2 k.$$  

If we choose $\epsilon < 1/(k \log_2 k)$, then we get $H_\mu(\alpha|\beta) \leq 1$, and

$$h_\mu(T, \alpha) \leq h_\mu(T, \beta) + 1.$$  

(4.4)

We now create an open cover from $\beta$ by setting $\mathcal{U} := \{B_0 \cup B_1, \ldots, B_0 \cup B_k\}$. This is a cover. To see that it is open note that

$$B_0 \cup B_j = B_0 \cup (A_j \setminus B_0) \quad (\because A_j \cap B_0 = A_j \setminus B_j)$$

$$= B_0 \cup A_j = B_0 \cup \left( X \setminus \bigcup_{i \neq j} A_i \right) = B_0 \cup \left( X \setminus \bigcup_{i \neq j} B_i \right).$$

We compare the number of elements in $\mathcal{V}_0^{n-1}$ to the number of elements in $\beta_0^{n-1}$. Every element of $\mathcal{V}_0^{n-1}$ is of the form

$$(B_0 \cup B_0) \cap T^{-1}(B_0 \cup B_1) \cap \cdots \cap T^{-(n-1)}(B_0 \cup B_{n-1}).$$

This can be written as a pairwise disjoint union of $2^n$ elements of $\beta_0^{n-1}$ (some of which may be empty sets). Thus every element of $\mathcal{V}_0^{n-1}$ contains at most $2^n$ elements of $\beta_0^{n-1}$. Forming the union over a sub cover of $\mathcal{V}_0^{n-1}$ with cardinality $N(\mathcal{V}_0^{n-1})$, we get that $\#\beta_0^{n-1} \leq 2^n N(\mathcal{V}_0^{n-1}).$

We no appeal to (4.3): $H_\mu(\beta_0^{n-1}) \leq \log_2(\#\beta_0^{n-1}) \leq H(\mathcal{V}_0^{n-1}) + n$. Dividing by $n$ and passing to the limit as $n \to \infty$, we see that $h_\mu(T, \beta) \leq h_{\text{top}}(\mathcal{U}) + 1$. By (4.4), $h_\mu(T, \alpha) \leq h_{\text{top}}(\mathcal{U}) + 2 \leq h_{\text{top}}(T) + 2$.

Passing to the supremum over all $\alpha$, we get that $h_\mu(T) \leq h_{\text{top}}(T) + 2$, and this holds for all continuous mappings $T$ and invariant Borel measures $\mu$. In particular, this holds for $T^n$ (note that $\mu \circ (T^n)^{-1} = \mu$): $h_\mu(T^n) \leq h_{\text{top}}(T^n) + 2$. But $h_{\text{top}}(T^n) = nh_\mu(T)$ and $h_{\text{top}}(T^n) = nh_\mu(T)$ (problems 4.4 and 4.13). Thus we get upon division by $n$ that $h_\mu(T) \leq h_{\text{top}}(T) + (2/n) \xrightarrow{n \to \infty} 0$, which proves the theorem. □
In fact, \( h_{\text{top}}(T) = \sup \{ h_\mu(T) : \mu \text{ is an invariant Borel probability measure} \} \). But to prove this we need some more preparations. These are done in the next section.

### 4.6.2 Bowen’s definition

We assume as usual that \((X, d)\) is a compact metric space, and that \(T : X \to X\) is continuous. For every \(n\) we define a new metric \(d_n\) on \(X\) as follows:

\[
d_n(x, y) := \max_{0 \leq i \leq n-1} d(T^i x, T^i y).
\]

This is called Bowen’s metric. It depends on \(T\). A set \(F \subset X\) is called \((n, \varepsilon)\)-separated, if for every \(x, y \in F\) s.t. \(x \neq y\), \(d_n(x, y) > \varepsilon\).

**Definition 4.7.**

1. \( s_n(\varepsilon, T) := \max \{ \#(F) : F \text{ is } (n, \varepsilon)\text{-separated} \} \)
2. \( s(\varepsilon, T) := \limsup_{n \to \infty} \frac{1}{n} \log s_n(\varepsilon, T) \)
3. \( \overline{h}_{\text{top}}(T) := \lim_{\varepsilon \to 0^+} s(\varepsilon, T) \).

**Theorem 4.8 (Bowen).** Suppose \(T\) is a continuous mapping of a compact metric space \(X\), then \(h_{\text{top}}(T) = \overline{h}_{\text{top}}(T)\).

**Proof.** Suppose \(\mathcal{U}\) is an open cover all of whose elements have diameters less than \(\varepsilon\). We claim that \(N(\mathcal{U}^{n-1}) \geq s_n(\varepsilon, T)\) for all \(n\). To see this suppose \(F\) is an \((n, \varepsilon)\)-separated set of maximal cardinality. Each \(x \in F\) is contained in some \(U_x \in \mathcal{U}^{n-1}\). Since the \(d\)-diameter of every element of \(\mathcal{U}\) is less than \(\delta\), the \(d_n\)-diameter of every element of \(\mathcal{U}^{n-1}\) is less than \(\delta\). Thus the assignment \(x \mapsto U_x\) is one-to-one, whence

\[
N(\mathcal{U}^{n-1}) \geq s_n(\varepsilon, T).
\]

It follows that \(s(\varepsilon, T) \leq h_{\text{top}}(T, \mathcal{U}) \leq h_{\text{top}}(T)\), whence \(\overline{h}_{\text{top}}(T) \leq h_{\text{top}}(T)\).

To see the other inequality we use Lebesgue numbers: a number \(\delta\) is called a Lebesgue number for an open cover \(\mathcal{U}\), if for every \(x \in X\), the ball with radius \(\delta\) and center \(x\) is contained in some element of \(\mathcal{U}\). (Lebesgue numbers exist because of compactness.)

Fix \(\varepsilon\) and let \(\mathcal{U}\) be an open cover with Lebesgue number bigger than or equal to \(\varepsilon\). It is easy to check that for every \(n\), \(\varepsilon\) is a Lebesgue number for \(\mathcal{U}^{n-1}\) w.r.t. \(d_n\).

Let \(F\) be an \((n, \varepsilon/2)\)-separated set of maximal cardinality, i.e. \(#F = s_n(\varepsilon)\). Then any point \(y\) we add to \(F\) will break its \((n, \varepsilon)\)-separation property, and so

\[
\forall y \in X \exists x \in F \text{ s.t. } d_n(x, y) \leq \varepsilon/2.
\]

It follows that the sets \(B_n(x; \varepsilon/2) := \{ y : d_n(x, y) \leq \varepsilon/2 \} (x \in F)\) cover \(X\).
4.6 Topological Entropy

Every \( B_n(x, \varepsilon/2) \) (\( x \in F \)) is contained in some element of \( \mathcal{U}_0^{n-1} \), because \( \mathcal{U}_0^{n-1} \) has Lebesgue number \( \varepsilon \) w.r.t \( d_n \). The union of these elements covers \( X \). We found a sub cover of \( \mathcal{U}_0^{n-1} \) of cardinality at most \( \#F = s_n(\varepsilon) \). This shows that

\[
N(\mathcal{U}_0^{n-1}) \leq s_n(\varepsilon).
\]

We just proved that for every open cover \( \mathcal{U} \) with Lebesgue number at least \( \varepsilon \),

\[
h_{\text{top}}(T, \mathcal{U}) \leq s(\varepsilon).
\]

It follows that

\[
\sup \{ h_{\text{top}}(T, \mathcal{U}) : \mathcal{U} \text{ has Lebesgue number at least } \varepsilon \} \leq s(\varepsilon).
\]

We now pass to the limit \( \varepsilon \to 0^+ \). The left hand side tends to the supremum over all open covers, which is equal to \( h_{\text{top}}(T) \). We obtain

\[
h_{\text{top}}(T) \leq \lim_{\varepsilon \to 0^+} s(\varepsilon). \tag*{\blacksquare}
\]

**Corollary 4.1.** Suppose \( T \) is an isometry, then all its invariant probability measures have entropy zero.

**Proof.** If \( T \) is an isometry, then \( d_n = d \) for all \( n \), therefore \( s(\varepsilon, T) = 0 \) for all \( \varepsilon > 0 \), so \( h_{\text{top}}(T) = 0 \). The theorem says that \( h_{\text{top}}(T) = 0 \). The corollary follows from Goodwyn’s theorem. \( \Tag*{\blacksquare} \)

### 4.6.3 The variational principle

The following theorem was first proved under additional assumptions by Dinaburg, and then in the general case by Goodman. The proof below is due to Misiurewicz.

**Theorem 4.9 (Variational principle).** Suppose \( T : X \to X \) is a continuous map of a compact metric space, then \( h_{\text{top}}(T) = \sup \{ h_{\mu}(T) : \mu \text{ is an invariant Borel measure} \} \).

**Proof.** We have already seen that the topological entropy is an upper bound for the metric entropies. We just need to show that this is the least upper bound.

Fix \( \varepsilon \), and let \( F_n \) be a sequence of \( (n, \varepsilon) \)–separated sets of maximal cardinality (so \( \#F_n = s_n(\varepsilon, T) \)). Let

\[
\nu_n := \frac{1}{\#F_n} \sum_{x \in F_n} \delta_x,
\]

where \( \delta_x \) denotes the Dirac measure at \( x \) (i.e. \( \delta_x(E) = 1_{E}(x) \)). These measure are not invariant, so we set

\[
\mu_n := \frac{1}{n} \sum_{k=0}^{n-1} \nu_n \circ T^{-k}.
\]

Any weak star limit of \( \mu_n \) will be \( T \)–invariant (check).

Fix some sequence \( n_k \to \infty \) s.t. \( \mu_{n_k} \overset{w^*}{\to} \mu \) and s.t. \( \frac{1}{n_k} \log s_{n_k}(\varepsilon, T) \overset{n \to \infty}{\to} s(\varepsilon, T) \).

We show that the entropy of \( \mu \) is at least \( s(\varepsilon, T) \). Since \( s(\varepsilon, T) \overset{\varepsilon \to 0^+}{\to} h_{\text{top}}(T) \), this will prove the theorem.
Let $\alpha = \{A_1, \ldots, A_N\}$ be a measurable partition of $X$ s.t. (1) $\text{diam}(A_i) < \varepsilon$; and (2) $\mu(\partial A_i) = 0$. (Such a partition can be generated from a cover of $X$ by balls of radius less than $\varepsilon/2$ and boundary of zero measure.) It is easy to see that the $d_n$-diameter of $\alpha_0^{n-1}$ is also less than $\varepsilon$. It is an exercise to see that every element of $\alpha_0^{n-1}$ has boundary with measure $\mu$ equal to zero.

We calculate $H_{\nu_0}(\alpha_0^{n-1})$. Since $F_n$ is $(n, \varepsilon)$-separated and every atom of $\alpha$ has $d_n$-diameter less than $\varepsilon$, $\alpha_0^{n-1}$ has $\# F_n$ elements whose $\nu_n$ measure is $1/\# F_n$, and the other elements of $\alpha_0^{n-1}$ have measure zero. Thus

$$H_{\nu_0}(\alpha_0^{n-1}) = \log_2 \# F_n = \log_2 s_n(\varepsilon, T).$$

We now “play” with $H_{\nu_0}(\alpha_0^{n-1})$ with the aim of bounding it by something involving a sum of the form $\sum_{i=0}^{q-1} H_{\nu_0 \circ T^{-i}}(\alpha_0^{q-1})$. Fix $q$, and $j \in \{0, \ldots, q-1\}$, then

$$\log_2 s_n(\varepsilon, T) = H_{\nu_0}(\alpha_0^{n-1}) \leq H_{\nu_0}(\alpha_0^{j-1}) \vee \sum_{i=0}^{\lfloor n/q \rfloor - 1} T^{-(q+i)} \alpha_0^{q-1} \vee \alpha_0^{n-1} + q \log_2 \# \alpha.$$

Summing over $j = 0, \ldots, q-1$, we get

$$q \log_2 s_n(\varepsilon, T) \leq n \cdot \sum_{k=0}^{n-1} H_{\nu_0 \circ T^{-k}}(\alpha_0^{q-1}) + 2q \log_2 \# \alpha,$$

because $\mu_n = \frac{1}{n} \sum_{i=0}^{n-1} \nu_n \circ T^{-i}$ and $\phi(t) = -t \log_2 t$ is concave. Thus

$$\frac{1}{n_k} \log_2 s_{n_k}(\varepsilon, T) \leq \frac{1}{q} H_{\mu_{n_k}}(\alpha_0^{q-1}) + \frac{2}{n_k} \log_2 \# \alpha, \quad (4.5)$$

where $n_k \to \infty$ is the subsequence chosen above.

Since every $A \in \alpha_0^{n-1}$ satisfies $\mu(\partial A) = 0$, $\mu_{n_k}(A) \xrightarrow[k \to \infty]{w^*} \mu(A)$ for all $A \in \alpha_0^{n-1}$. It follows that $H_{\mu_{n_k}}(\alpha_0^{q-1}) \xrightarrow[k \to \infty]{} H_{\mu}(\alpha_0^{q-1})$. Passing to the limit $k \to \infty$ in (4.5), we have $s(\varepsilon, T) \leq \frac{1}{q} H_{\mu}(\alpha) \xrightarrow[q \to \infty]{q} h_{\mu}(T, \alpha) \leq h_{\mu}(T)$. Thus $h_{\mu}(T) \geq s(\varepsilon, T)$. Since $s(\varepsilon, T) \xrightarrow[\varepsilon \to 0^+]{} h_{\text{top}}(T)$ the theorem is proved. \quad $\square$
## Problems

4.1. Prove: \( H_\mu(\alpha | \beta) = - \sum_{B \in \beta} \mu(B) \sum_{A \in \alpha} \mu(A | B) \log \mu(A | B) \), where \( \mu(A | B) = \frac{\mu(A \cap B)}{\mu(B)} \).

4.2. Prove: if \( H_\mu(\alpha | \beta) = 0 \), then \( \alpha \subseteq \beta \mod \mu \).

4.3. Prove that \( h_\mu(T) \) is an invariant of measure theoretic isomorphism.

4.4. Prove that \( h_\mu(T^n) = nh_\mu(T) \).

4.5. Prove that if \( T \) is invertible, then \( h_\mu(T^{-1}) = h_\mu(T) \).

### 4.6. Entropy is affine

Let \( T \) be a measurable map on \( X \), and \( \mu_1, \mu_2 \) be two \( T \)-invariant probability measures. Set \( \mu = t \mu_1 + (1 - t) \mu_2 \). Show: \( h_\mu(T) = th_{\mu_1}(T) + (1 - t)h_{\mu_2}(T) \).

**Guidance:** Start by showing that for all \( 0 \leq x, y, t \leq 1 \),

\[
0 \leq \phi(tx + (1 - t)y) - [\phi(x) + (1 - t)\phi(y)] \leq -tx\log t - (1 - t)y\log(1 - t)
\]

4.7. Let \( (X, \mathcal{B}, \mu) \) be a probability space. If \( \alpha, \beta \) are two measurable partitions of \( X \), then we write \( \alpha = \beta \mod \mu \) if \( \alpha = \{A_1, \ldots, A_n\} \) and \( \beta = \{B_1, \ldots, B_n\} \) where \( \mu(A_i \triangle B_i) = 0 \) for all \( i \). Let \( \mathcal{P} \) denote the set of all countable measurable partitions of \( X \), modulo the equivalence relation \( \alpha = \beta \mod \mu \). Show that

\[
\rho(\alpha, \beta) := H_\mu(\alpha | \beta) + H_\mu(\beta | \alpha)
\]

induces a metric on \( \mathcal{P} \).

4.8. Let \( (X, \mathcal{B}, \mu, T) \) be a ppt. Show that \( |h_\mu(T, \alpha) - h_\mu(T, \beta)| \leq H_\mu(\alpha | \beta) + H_\mu(\beta | \alpha) \).

4.9. Use the previous problem to show that the supremum in the definition of metric entropy is attained by finite measurable partitions.

4.10. Suppose \( \alpha = \{A_1, \ldots, A_n\} \) is a finite measurable partition. Show that for every \( \varepsilon, \), there exists \( \delta = \delta(\varepsilon, n) \) such that if \( \beta = \{B_1, \ldots, B_n\} \) is measurable partition s.t. \( \mu(A_i \triangle B_i) < \delta \), then \( \rho(\alpha, \beta) < \varepsilon \).

4.11. **Entropy via generating sequences of partitions**

Suppose \( (X, \mathcal{B}, \mu) \) is a probability space, and \( \mathcal{A} \) is an algebra of \( \mathcal{F} \)-measurable subsets (namely a collection of sets which contains \( \emptyset \) and which is closed under finite unions, finite intersection, and forming complements). Suppose \( \mathcal{A} \) generates \( \mathcal{B} \) (i.e. \( \mathcal{B} \) is the smallest \( \sigma \)-algebra which contains \( \mathcal{A} \)).

1. For every \( F \in \mathcal{F} \) and \( \varepsilon > 0 \), there exists \( \epsilon \in \mathcal{A} \) s.t. \( \mu(A \triangle F) < \varepsilon \).
2. For every \( \mathcal{F} \)-measurable finite partition \( \beta \) and \( \varepsilon > 0 \), there exists an \( \mathcal{A} \)-measurable finite partition \( \alpha \) s.t. \( \rho(\alpha, \beta) < \varepsilon \).
3. If $T : X \to X$ is probability preserving, then
\[ h_\mu(T) = \sup \{ h_\mu(T, \alpha) : \alpha \text{ is an } \mathcal{A} \text{-measurable finite partition} \} . \]

4. Suppose $\alpha_1 \leq \alpha_2 \leq \cdots$ is an increasing sequence of finite measurable partitions such that $\sigma(\bigcup_{n \geq 1} \alpha_n) = \emptyset \mod \mu$, then $h_\mu(T) = \lim_{n \to \infty} h_\mu(T, \alpha_n)$.

4.12. Show that the entropy of the product of two ppt is the sum of their two entropies.

4.13. Show that $h_{\text{top}}(T^n) = nh_{\text{top}}(T)$.

### Notes to chapter 4

The notion of entropy as a measure of information is due to Shannon, the father of the modern theory of information. Kolmogorov had the idea to adapt this notion to the ergodic theoretic context for the purposes of inventing an invariant which is able to distinguish Bernoulli schemes. This became possible once Sinai has proved his generator theorem — which enables the calculation of this invariant for Bernoulli schemes. Later, in the 70’s, Ornstein has proved that entropy is a complete invariant for Bernoulli schemes: they are isomorphic iff they have the same entropy. The maximum of the possible entropies for a topological Markov shift was first calculated by Parry, who also found the maximizing measure. The material in this chapter is all classical, [3] and [1] are both excellent references. For an introduction to Ornstein’s isomorphism theorem, see [2].

### References


Appendix A
The Monotone Class Theorem

Definition A.1. A sequence of sets \{A_n\} is called increasing (resp. decreasing) if \(A_n \subseteq A_{n+1}\) for all \(n\) (resp. \(A_n \supseteq A_{n+1}\) for all \(n\)).

Notation: \(A_n \uparrow A\) means that \(\{A_n\}\) is an increasing sequence of sets, and \(A = \bigcup A_n\). \(A_n \downarrow A\) means that \(\{A_n\}\) is a decreasing sequence of sets, and \(A = \bigcap A_n\).

Proposition A.1. Suppose \((X, \mathcal{B}, \mu)\) is a measure space, and \(A_n \in \mathcal{B}\).

1. if \(A_n \uparrow A\), then \(\mu(A_n) \rightarrow \mu(A)\);
2. if \(A_n \downarrow A\) and \(\mu(A_n) < \infty\) for some \(n\), then \(\mu(A_n) \rightarrow \mu(A)\).

Proof. For (1), observe that \(A = \bigcup_{n \geq 1} (A_{n+1} \setminus A_n)\) and use \(\sigma\)-additivity. For (2), fix \(n_0\) s.t. \(\mu(A_{n_0}) < \infty\), and observe that \(A_n \downarrow A\) implies that \((A_{n_0} \setminus A_n) \uparrow (A_{n_0} \setminus A)\). \(\square\)

The example \(A_n = (n, \infty)\), \(\mu = \text{Lebesgue measure on } \mathbb{R}\), shows that the condition in (2) cannot be removed.

Definition A.2. Let \(X\) be a set. A monotone class of subsets of \(X\) is a collection \(\mathcal{M}\) of subsets of \(X\) which contains the empty set, and such that if \(A_n \in \mathcal{M}\) and \(A_n \uparrow A\) or \(A_n \downarrow A\), then \(A \in \mathcal{M}\).

Recall that an algebra of subsets of a set \(X\) is a collection of subsets of \(X\) which contains the empty set, and which is closed under finite unions, finite intersections, and forming the complement.

Theorem A.1 (Monotone Class Theorem). A monotone class which contains an algebra, also contains the sigma–algebra generated by this algebra.

Proof. Let \(\mathcal{M}\) be a monotone class which contains an algebra \(\mathcal{A}\). Let \(\mathcal{M}(\mathcal{A})\) denote the intersection of all the collections \(\mathcal{M}' \subset \mathcal{M}\) such that (a) \(\mathcal{M}'\) is a monotone class, and (b) \(\mathcal{M}' \supseteq \mathcal{A}\). This is a monotone class (check!). In fact it is the minimal monotone class which contains \(\mathcal{A}\). We prove that it is a \(\sigma\)-algebra. Since \(\mathcal{M}(\mathcal{A}) \subset \mathcal{M}\), this completes the proof.
We begin by claiming that $\mathcal{M}(\mathcal{A})$ is closed under forming complements. Suppose $E \in \mathcal{M}(\mathcal{A})$. The set
\[
\mathcal{M}' := \{ E' \in \mathcal{M}(\mathcal{A}) : (E')^c \in \mathcal{M}(\mathcal{A}) \}
\]
contains $\mathcal{A}$ (because $\mathcal{A}$ is an algebra), and it is a monotone class (check!). But $\mathcal{M}(\mathcal{A})$ is the minimal monotone class which contains $\mathcal{A}$, so $\mathcal{M}' \supset \mathcal{M}(\mathcal{A})$. It follows that $E \in \mathcal{M}'$, whence $E^c \in \mathcal{M}(\mathcal{A})$.

Next we claim that $\mathcal{M}(\mathcal{A})$ has the following property:

$$E \in \mathcal{M}(\mathcal{A}), A \in \mathcal{A} \implies E \cup A \in \mathcal{M}(\mathcal{A}).$$

Again, the reason is that the collection $\mathcal{M}'$ of sets with this property contains $\mathcal{A}$, and is a monotone class.

Now fix $E \in \mathcal{M}(\mathcal{A})$, and consider the collection
\[
\mathcal{M}' := \{ F \in \mathcal{M}(\mathcal{A}) : E \cup F \in \mathcal{M}(\mathcal{A}) \}.
\]

By the previous paragraph, $\mathcal{M}'$ contains $\mathcal{A}$. It is clear that $\mathcal{M}'$ is a monotone class. Thus $\mathcal{M}(\mathcal{A}) \subseteq \mathcal{M}'$, and as a result $E \cup F \in \mathcal{M}(\mathcal{A})$ for all $F \in \mathcal{M}(\mathcal{A})$. But $E \in \mathcal{M}(\mathcal{A})$ was arbitrary, so this means that $\mathcal{M}(\mathcal{A})$ is closed under finite unions.

Since $\mathcal{M}(\mathcal{A})$ is closed under finite unions, and countable increasing unions, it is closed under general countable unions.

Since $\mathcal{M}(\mathcal{A})$ is closed under forming complements and taking countable unions, it is a sigma algebra. By definition this sigma algebra contains $\mathcal{A}$ and is contained in $\mathcal{M}$. \qed
Index

| Abramov formula, 93  |
| Action, 1  |
| adding machine, 32, 64  |
| algebra, 101  |
| Alt, 48  |
| arithmeticity, 22  |
| Bernoulli scheme, 9, 73  |
| entropy of, 91  |
| Bowen’s metric, 98  |
| Carathéodory Extension Theorem, 10  |
| Chacon’s example, 76  |
| Chung-Neveu Lemma, 87  |
| coboundary, 35  |
| complete  |
| measure space, 31  |
| Riemannian surface, 18  |
| conditional  |
| entropy, 82  |
| expectation, 39  |
| probabilities, 40  |
| configuration, 1  |
| conservative mpt, 32  |
| covariance, 7  |
| cutting and stacking, 76  |
| cylinders, 9  |
| Dynamical system, 1  |
| eigenfunction, 68  |
| eigenvalue, 68  |
| entropy  |
| conditional, 82  |
| of a measure, 85  |
| of a partition, 82  |
| topological, 95, 96  |
| ergodic, 5  |
| decomposition, 42  |
| hypothesis, 3  |
| ergodicity and countable Lebesgue spectrum, 76  |
| ergodicity and extremality, 64  |
| ergodicity and mixing, 64  |
| flows, 18, 32  |
| theory, 1  |
| Ergodic Theorem  |
| Mean, 35, 64  |
| Multiplicative, 52  |
| Pointwise, 36  |
| Ratio, 65  |
| Subadditive, 44  |
| ergodicity and mixing, 36  |
| extension, 23  |
| exterior product, 49  |
| and angles, 51  |
| of linear operators, 50  |
| extremal measure, 64  |
| factor, 22  |
| factor map, 23  |
| flow, 1  |
| Fourier Walsh system, 73  |
| Furstenberg-Kesten theorem, 47  |
| generator, 89  |
| geodesic flow, 17  |
| Goodwyn’s theorem, 96  |
| Herglotz theorem, 70  |
| hyperbolic  |
| plane, 17  |
| surface, 18  |
| independence, 6  |
Index

for partitions, 85
induced transformation, 28
Abramov formula, 93
entropy of, 93
for infinite mpt, 33
Kac formula, 28
Kakutani skyscraper, 30
information
conditional, 82
content, 82
function, 82
invariant set, 5
invertibility, 23
isomorphism, 4
measure theoretic, 4
spectral, 67
iterates, 1
itinerary, 81
K automorphism, 73
Kac formula, 28, 33
Kakutani skyscraper, 30
Lebesgue number, 98
Liouville’s theorem, 2
Markov measure, 13
ergodicity and mixing, 13
Markov measures
Entropy, 91
Martingale convergence theorem, 65
measure, 3
measure preserving transformation, 4
measure space, 3
Lebesgue, 4
sigma finite, 32
standard, 3
mixing
and countable Lebesgue spectrum, 76
weak, 69
mixing and ergodicity, 36
monotone class, 103
mpt, 4
multilinear function, 48
alternating, 48
natural extension, 23
orbit, 1
partition
finer or coarser, 83
wedge product of, 83
periodicity, 22
Perron-Frobenius Theorem, 31
Phase space, 1
Poincaré Recurrence Theorem, 3
Poincaré section, 30
Polish, 21
polish space, 21
positive definite, 70
probability
preserving transformation, 4
stationary probability vector, 12
measure, 3
space, 3
vector, 12
product
of measure spaces, 20
of mpt, 20
PSL(2,R), 17
Rokhlin formula, 92
rotation, 8
Rotations, 32
Entropy of, 91
section map, 30
semi algebra, 10
semi-flow, 1
Shannon–McMillan–Breiman Theorem, 88
sigma algebra, 3
sigma finiteness, 32
Sinai’s Generator Theorem, 90
skew product, 21
spanning set, 93
spectral
invariant, 67
isomorphism, 67
measure, 70
spectrum
and the K property, 74
continuous, 68, 71
countable Lebesgue, 73
discrete, 68
Lebesgue, 76
mixed, 68
point, 68
pure point, 68
stationary
probability vector, 12
stationary stochastic process, 5
stochastic matrix, 12
stochastic process, 5
subadditive
cocycle, 44
ergodic theorem, 44
subshift of finite type, 12
Index

suspension, 30

period, 13

tail events, 74

tensor product, 48

time one map, 32

unitary equivalence, 67

Topological entropy, 95

definition using separated sets, 98

Topological entropy, 96

definition using separated sets, 98

Topological entropy, 96

variational principle, 99

Topological entropy, 95

variational principle, 99

topological entropy, 96

wandering set, 32

transition matrix

wedge product

aperiodic, 13

of $\sigma$–algebras, 83

irreducible, 13

of multilinear forms, 49

Zero one law, 74

of partitions, 89