LECTURE NOTES FOR THE COURSE TOPICS IN ERGODIC THEORY, MICHAELMAS 2018

PÉTER VARJÚ

These are brief lecture notes for my course. Please send comments to: pv270@dpmms.cam.ac.uk.

1. Measure preserving systems

Ergodic Theory is the study of measure preserving systems.

Definition 1. A **measure preserving system** (MPS) is a quadruple $(X, \mathcal{B}, \mu, T)$, where

- X is a set
- $\mathcal{B}$ is a σ-algebra
- μ is a probability measure, that is, $\mu(X) = 1$ and $\mu(A) \geq 0$ for all $A \in \mathcal{B}$
- $T : X \to X$ is a measure preserving transformation, that is a measurable transformation such that $\mu(T^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{B}$.

Example 2 (Circle rotation)**.** Fix a number $\alpha \in \mathbf{R}/\mathbf{Z}$. Let $X = \mathbf{R}/\mathbf{Z}$, let $\mathcal{B}$ be the collection of Borel sets in $[0, 1]$, μ the Lebesgue measure, and $T = R_\alpha$, where R_α is the map

$$R_\alpha : x \mapsto x + \alpha.$$

Example 3 ("Times 2" map)**.** Let $X, \mathcal{B}, \mu$ be as in the previous example and let $T = T_2$, where

$$T_2(x) = 2x.$$

These examples can be generalized to more general compact groups. In the first case the map is addition (or multiplication) by a fixed element of the group, in the second case it is an endomorphism of the group.

Proof that T_2 is measure preserving. Easy to see for intervals:

$$\mu(a, b) = b - a = \mu((a/2, b/2) \cup (a/2 + 1/2, b/2 + 1/2)) = \mu(T_2^{-1}(a, b)).$$

An open set $U \subset X$ is the disjoint union of intervals $I_1 \cup I_2 \cup \ldots$, hence

$$\mu(U) = \sum \mu(I_j) = \sum \mu(T_2^{-1}(I_j)) = \mu(T_2^{-1}(U)).$$

A compact set K is the complement of an open set U, hence

$$\mu(K) = 1 - \mu(U) = 1 - \mu(T_2^{-1}(U)) = \mu(X \backslash T_2^{-1}(U)) = \mu(T_2^{-1}(K)).$$

Finally a general Borel set $B \in \mathcal{B}$ is approximated by an open set U and a compact set K such that $K \subset B \subset U$ and $\mu(U) - \mu(K) \leq \varepsilon$. On the other hand

$$\mu(T_2^{-1}(B)) \leq \mu(T_2^{-1}(U)) = \mu(U)$$

and similarly $\mu(T_2^{-1}(B)) \geq \mu(K)$, so we must have $|\mu(T_2^{-1}(B)) - \mu(B)| \leq \varepsilon$, which proves the claim letting $\varepsilon \to 0$. □

Ergodic theory studies the long term behaviour of orbits in MPS's. The **orbit** of a point $x \in X$ is the sequence $x, Tx, T^2x, \ldots$. In particular the following questions are asked:

- Let $A \in \mathcal{B}$ and $x \in A$. Does the orbit of x visit A infinitely often?
- What is the proportion of the n's such that $T^n x \in A$?
- What is the measure of $\{x \in A : T^n x \in A\}$ for some large n?

Example 4. Let $A = [0, 1/4)$. Then $T_2^n x \in A$ if and only if the $n+1$'th and $n+2$'th binary digits of x are both 0. Hence the orbit of

$$x = \frac{1}{6} = 0.00101010101010\ldots_{(2)}$$

never visits A again. However, the opposite is true for "most points". In addition:

$$\mu(\{x \in A : T^n x \in A\}) = \frac{1}{16} \quad \text{for all } n \geq 2.$$

We mention one further example of a MPS.

Example 5 (Markov shifts)**.** Let $n \geq 2$ be an integer, let $p_1, \ldots, p_n$ be a probability vector and $A = (a_{i,j}) \in \mathbf{R}_{\geq 0}^{n \times n}$ a matrix, called the matrix of transition probabilities. We assume

$$A \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad (p_1, \ldots, p_n)A = (p_1, \ldots, p_n).$$

The following MPS is called a **Markov shift**.

- $X = \{1, 2, \ldots, n\}^{\mathbf{Z}}$,
- $\mathcal{B}$ is the Borel σ-algebra generated by the product topology,
- $T = \sigma$ is the shift map $(\sigma x)_m = x_{m+1}$,
- the measure μ is defined by the following formula

$$\begin{aligned} \mu(\{x \in X : x_m = i_0, x_{m+1} =& i_1, \ldots, x_{m+k} = i_k\}) \\ =& p_{i_0} a_{i_0 i_1} a_{i_1 i_2} \cdots a_{i_{k-1} i_k}, \end{aligned} \tag{1}$$

which is required to hold for all $m, k, i_0, \ldots, i_k$.

It requires a proof that the set-function defined by (1) is additive and that it can be extended to a probability measure on $\mathcal{B}$ that is σ-invariant. This will appear in the first example sheet.

An important special case of Markov shifts is when $a_{i,j} = p_j$ for all i, j. Inspection of (1) shows that μ is the infinite product of the measure on $\{1, \dots, k\}$ given by the probability vector $(p_1, \dots, p_n)$. In this special case the MPS is called a **Bernoulli shift**.

2. Furstenberg's correspondence principle

Theorem 6 (Szemerédi)**.** *Let $S \subset \mathbf{Z}$ be a set of positive upper Banach density, that is:*

$$\overline{d}(S) := \limsup_{N-M\to\infty} \frac{1}{N-M} |\{x \in S : M \le x < N\}|.$$

Then S contains an arbitrarily long arithmetic progression, that is: for each $l \in \mathbf{Z}_{>0}$, there is $a \in \mathbf{Z}$ and $d \in \mathbf{Z}_{>0}$ such that

$$\{a + jd : j = 0, \dots, l-1\} \subset S.$$

Fix a set S with positive upper Banach density. We construct below an MPS and show that Szemerédi's theorem follows from the so-called multiple recurrence property of this system.

We define $X = \{0,1\}^{\mathbf{Z}}$ and take $\mathcal{B}$ to be the Borel σ-algebra generated by the product topology. We consider the shift transformation

$$\sigma : X \to X, (\sigma(x))_n = x_{n+1}.$$

The invariant measure will be constructed below, and it will encode the set S.

Define the element $x^S \in X$ by

$$x_n^S := \begin{cases} 1 & \text{if } n \in S \\ 0 & \text{if } n \notin S. \end{cases}$$

and consider the set

$$A := \{x \in X : x_0 = 1\}.$$

The following observation will be used to relate the set S to the dynamics: For each $n \in \mathbf{Z}$ we have

$$n \in S \quad \text{if and only if} \quad \sigma^n x^S \in A.$$

Indeed,

$$n \in S \Leftrightarrow 1 = x_n^S = (\sigma^n x^S)_0 \Leftrightarrow \sigma^n x^S \in A.$$

The invariant measure in the MPS will be constructed as the limit of normalized counting measures on longer and longer segments of the orbit of x^S. Before we give the details, we recall the notion and properties of weak limits of probability measures.

2.1. **Weak convergence of measures.** In this section X is a compact metric space.

Definition 7. Let μ_n be a sequence of Borel probability measures on X and let μ be another Borel probability measure. We say that μ_n weakly converges [(1)] to μ if

$$\lim_{n\to\infty} \int f d\mu_n = \int f d\mu \quad \text{for all } f \in C(X).$$

Here $C(X)$ stands for the space of continuous functions on X. In our notation, we will write

$$\text{lim-w}_{n\to\infty}\, \mu_n = \mu.$$

This concept is useful thanks to the following result, which is an analogue of the Bolzano-Weierstrass theorem.

Theorem 8 (Banach-Alaoglu/Helly)**.** *The space of Borel probability measures $M(X)$ on a compact metric space X endowed with the topology of weak convergence is compact and metrizable. In other words, any sequence of probability measures has a weakly convergent subsequence.*

Remark: If we drop the condition that X is metric (but it should be Hausdorff, at least), then the compactness of $M(X)$ will still hold, but we may lose metrizability and the sequential version needs to be reformulated in terms of nets.

2.2. **The invariant measure.** Let S, X, $\mathcal{B}$, σ and x^S be the same as above. We construct now the invariant measure on the MPS that will be used to prove Szemerédi's theorem.

We write δ_x for the probability measure defined by

$$\delta_x(B) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{if } x \notin B. \end{cases}$$

We choose two sequences of integers $\{M_m\}, \{N_m\} \subset \mathbf{Z}$ such that

$$\overline{d}(S) = \lim_{m\to\infty} \frac{1}{N_m - M_m} |\{x \in S : M_m \le x < N_m\}|.$$

These sequences exist by the definition of $\overline{d}(S)$.

We define the measures $\mu_m \in M(X)$ by

$$\mu_m = \frac{1}{N_m - M_m} \sum_{i=M_m}^{N_m-1} \delta_{\sigma^i x^S}.$$

Interpretation: $\mu_m(B)$ is the proportion of the points in the orbit segment $\sigma^{M_m} x^S, \dots, \sigma^{N_m-1} x^S$ that are in the set B.

We define μ to be the weak limit of a subsequence of μ_m.

Lemma 9. *The above defined $(X, \mathcal{B}, \mu, \sigma)$ is a MPS.*

[(1)] This is also called weak-$*$ convergence.

Proof sketch. Later in the course we prove a general result, which contains this lemma. For this reason, we do not give a rigorous argument now.

Without loss of generality, we assume that lim-w $\mu_m = \mu$. (Otherwise we pass to the suitable subsequence.)

We can write for a set $B \in \mathcal{B}$:

$$\mu_m(B) = \frac{1}{N_m - M_m} |\{i : M_m \leq i < N_m, \sigma^i x^S \in B\}|$$

and similarly

$$\begin{aligned}\mu_m(\sigma^{-1}(B)) =& \frac{1}{N_m - M_m} |\{i : M_m \leq i < N_m, \sigma^i x^S \in \sigma^{-1}(B)\}| \\ =& \frac{1}{N_m - M_m} |\{i : M_m + 1 \leq i < N_m + 1, \sigma^i x^S \in B\}|.\end{aligned}$$

Hence

$$|\mu_m(B) - \mu_m(\sigma^{-1}B)| < \frac{1}{N_m - M_m}$$

and therefore

$$\lim_{m\to\infty} \mu_m(B) = \lim_{m\to\infty} \mu_m(\sigma^{-1}B)$$

provided the limits exist. What remains to verify is that we can pass to the limit and conclude $\mu(B) = \mu(\sigma^{-1}B)$, but we omit this for now. □

Remark 10. If B is a cylinder set[(2)], i.e. a set of the form

$$B = \{x \in X : (x_{-N}, \ldots, x_N) \in \overline{B}\}$$

for some $N \in \mathbf{Z}_{\geq 0}$ and $\overline{B} \subset \{0,1\}^{2N+1}$, then B is both open and closed, hence χ_B is continuous.

By the definition of weak convergence, then

$$\begin{aligned}\mu(\sigma^{-1}(B)) = \int \chi_{\sigma^{-1}B} d\mu = \lim_{m\to\infty} \int \chi_{\sigma^{-1}B} d\mu_m = \lim_{m\to\infty} \mu_m(\sigma^{-1}B) \\ = \lim_{m\to\infty} \mu_m(B) = \lim_{m\to\infty} \int \chi_B d\mu_m = \int \chi_B d\mu = \mu(B)\end{aligned}$$

holds for any cylinder set B.

It is possible to make the proof of Lemma 9 rigorous by approximating arbitrary Borel sets with cylinder sets in a suitable way.

Proposition 11. *Let $S \subset \mathbf{Z}$ be a set of positive upper Banach density and let $(X, \mathcal{B}, \mu, \sigma)$ be the measure preserving system constructed above. Let $A \subset X$ be the set $A = \{x \in X : x_0 = 1\}$. Let $l \geq 1$ be an integer. Suppose that*

$$\mu(A \cap \sigma^{-n}(A) \cap \sigma^{-2n}(A) \cap \ldots \cap \sigma^{-(l-1)n}(A)) > 0$$

(2)Some sources use a more restrictive terminology and call only sets of the form $\{x \in X : x_{-N} = a_{-N}, \ldots, x_N = a_N\}$ for some $a_{-N}, \ldots, a_N \in \{0,1\}$ cylinder sets.

for an integer $n \geq 1$. Then S contains an arithmetic progression of length l.

Proof. The set

$$B := A \cap \sigma^{-n}(A) \cap \sigma^{-2n}(A) \cap \ldots \cap \sigma^{-(l-1)n}(A)$$

is a cylinder set, hence $\mu(B) = \lim \mu_m(B)$, as we have seen above. This means that for some m large enough $\mu_m(B) > 0$. By the definition of μ_m, we must have $\sigma^j x^S \in B$ for some $M_m \leq j \leq N_m$. Then $\sigma^j x^S \in \sigma^{-in}(A)$, hence $\sigma^{j+in} x^S \in A$ for all $0 \leq i \leq l-1$. As we observed in the beginning of the lecture, this translates as

$$\{j + in : 0 \leq i \leq l-1\} \subset S,$$

which was to be shown. □

We note that $\mu(A) = \overline{d}(S)$ as can be seen easily from the construction of μ. We can now conclude that the following theorem of Furstenberg implies Szemerédi's theorem.

Theorem 12 (Multiple Recurrence, Furstenberg)**.** *Let $(X, \mathcal{B}, \mu, T)$ be a MPS and $A \in \mathcal{B}$ with $\mu(A) > 0$. Then for any integer $l \geq 1$ we have*

$$\liminf \frac{1}{N} \sum_{n=1}^{N} \mu(A \cap T^{-n}A \cap T^{-2n}A \cap \ldots \cap T^{-(l-1)n}A) > 0.$$

3. Poincaré recurrence, Ergodicity

The following lemma can be considered the pigeon hole principle of Ergodic Theory. Putting together with the Furstenberg correspondence principle it implies that every set of integers of positive upper Banach density contains an arithmetic progression of length 2, (which is not very impressive, yet).

Lemma 13. *Let $(X, \mathcal{B}, \mu, T)$ be an MPS and let $A \in \mathcal{B}$ with $\mu(A) > 0$. Then there is some $n \geq 1$ such that $\mu(T^{-n}(A) \cap A) > 0$.*

Proof. Assume to the contrary that $\mu(A \cap T^{-n}(A)) = 0$ for all $n \in \mathbf{Z}_{>0}$. Let $i < j \in \mathbf{Z}_{\geq 0}$. Then

$$\mu(T^{-i}(A) \cap T^{-j}(A)) = \mu(T^{-i}(A \cap T^{-(j-i)}(A))) = 0.$$

We fix a number N and write

$$\begin{aligned}
&\mu(A \cup T^{-1}(A) \cup \ldots \cup T^{-N}(A)) \\
&\quad = \mu(A) + \mu(T^{-1}A) - \mu(T^{-1}A \cap A) \\
&\qquad + \mu(T^{-2}A) - \mu(T^{-2}A \cap (A \cup T^{-1}A)) + \ldots \\
&\qquad + \mu(T^{-N}A) - \mu(T^{-N}(A) \cap (A \cup \ldots \cup T^{-N}A)) \\
&\quad = N\mu(A),
\end{aligned}$$

which is a contradiciton if $N > \mu(A)^{-1}$. □

Theorem 14 (Poincaré recurrence)**.** *Let $(X, \mathcal{B}, \mu, T)$ be an MPS and let $A \in \mathcal{B}$ with $\mu(A) > 0$. Then almost every point of A goes back to A infinitely often. That is to say*

$$\mu\Big(A \setminus \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} T^{-n}(A)\Big) = 0.$$

Note that $T^{-n}(A)$ consists of the set of points x that satisfy $T^n x \in A$, $\bigcup_{n=N}^{\infty} T^{-n}(A)$ is the set of points that visit A at least once after the Nth iteration of T and $\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} T^{-n}(A)$ is the set of points that visit A infinitely often.

Proof. We denote by A_0 the set of points in A that never come back to A. Then $A_0 \cap T^{-n} A_0 \subset A \cap T^{-n} A_0 = \varnothing$ for all $n \geq 1$, hence $\mu(A_0) = 0$ by the lemma.

Let $x \in A$ be a point that does not come back to A infinitely often. Let n be the largest integer such that $T^n x \in A$. Then $T^n x \in A_0$. This shows that

$$\Big(A \setminus \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} T^{-n}(A)\Big) \subset \bigcup_{n=0}^{\infty} T^{-n}(A_0).$$

This latter set is of measure 0, because it is the countable union of measure 0 sets. □

Definition 15 (Ergodicity)**.** An MPS $(X, \mathcal{B}, \mu, T)$ is called ergodic if $A = T^{-1}(A)$ implies $\mu(A) = 0$ or $\mu(A) = 1$ for any $A \in \mathcal{B}$.

Ergodicity is a notion of indecomposibility for MPS's. If $A \in \mathcal{B}$ is invariant (that is $T^{-1}(A) = A$), then we can restrict $\mathcal{B}$, μ and T to A and after renormalizing the measure, we obtain a new MPS.

The following lemma gives a few alternative characterization of ergodicity.

Lemma 16. *The following are equivalent*

1. *The MPS $(X, \mathcal{B}, \mu, T)$ is ergodic.*
2. *For any $A \in \mathcal{B}$ with $\mu(A) > 0$, we have $\mu(\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} T^{-n}(A)) = 1$.*
3. *For any $A \in \mathcal{B}$, $\mu(A \triangle T^{-1}(A)) = 0$ implies $\mu(A) = 0$ or $\mu(A) = 1$.*
4. *For any bounded measurable function $f : X \to \mathbf{R}$, $f \circ T = f$ holds almost everywhere implies that f is constant almost everywhere.*
5. *For any measurable function $f : X \to \mathbf{C}$, $f \circ T = f$ holds almost everywhere implies that f is constant almost everywhere.*

The second item shows that for ergodic systems Poincaré recurrence holds in a stronger form: not only almost every point in A but also almost every point in X visit A infinitely often. The third, fourth and fifth items give characterizations that are useful in practice. In

particular, we will use the fourth item to show that the circle rotation R_α is ergodic if and only if α is irrational.

Proof. (1) $\Rightarrow$ (2) : The set $B := \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} T^{-n}(A)$ is invariant. Indeed, $x \in B$ if and only if its orbit visits A infinitely often. This is true for x if and only if it is true for Tx. This shows that $B = T^{-1}(B)$, as claimed. By Poincaré recurrence $\mu(B) \geq \mu(A) > 0$, hence $\mu(B) = 1$ by ergodicity.

(2) $\Rightarrow$ (3) : Let A satisfy $\mu(A \triangle T^{-1}(A)) = 0$ and let B be as in the previous paragraph. We show that $\mu(B) \leq \mu(A)$. For $n \geq 1$, we define D_n as the set of points $x \in B \backslash A$ such that n is the smallest integer with $T^n x \in A$. The sets D_n clearly partition $B \backslash A$. We can write

$$\mu(D_n) \leq \mu(T^{-n}(A) \backslash T^{-n+1}(A)) = \mu(T^{-1}A \backslash A) \leq \mu(T^{-1}A \triangle A) = 0.$$

Hence $\mu(\bigcup_{n=1}^{\infty} D_n) = 0$ and $\mu(B) \leq \mu(A)$, as required. If $\mu(A) \neq 0$, then $\mu(A) \geq \mu(B) = 1$, which proves (3).

(3) $\Rightarrow$ (4) : For $t \in \mathbf{R}$ write

$$A_t := \{x \in X : f(x) \leq t\}.$$

Clearly $\mu(A_t \triangle T^{-1} A_t) = 0$, hence $\mu(A_t) = 0$ or $\mu(A_t) = 1$ for each t. Observe that the sets A_t are increasing with t, hence the function $t \mapsto \mu(A_t)$ is a monotone increasing function.

Since f is bounded, we have $\mu(A_t) = 0$ if t is sufficiently small, and $\mu(A_t) = 1$ if t is sufficiently large. Therefore, there is a finite real number c such that $\mu(A_t) = 0$ for $t < c$ and $\mu(A_t) = 1$ for $t > c$. By the previous observations,

$$\{x : f(x) = c\} = \bigcap_{n=1}^{\infty} (A_{c+1/n} \backslash A_{c-1/n})$$

is of measure 1.

(4) $\Rightarrow$ (1) : Let $A \in \mathcal{B}$ be such that $T^{-1}(A) = A$. Then $\chi_A \circ T = \chi_A$, hence χ_A is constant almost everywhere. If this constant is 0, then $\mu(A) = 0$, if it is 1, then $\mu(A) = 1$. Hence the system is ergodic, indeed. $\square$

Example 17. The circle rotation R_α is ergodic if and only if α is irrational.

Indeed: Let $f : \mathbf{R}/\mathbf{Z} \to \mathbf{R}$ be a bounded measurable function. We can expand f in Fourier series

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

Similarly for $f \circ R_\alpha$:

$$f \circ R_\alpha = \sum_{n \in \mathbf{Z}} a_n \exp(2\pi i n (x + \alpha)) = \sum_{n \in \mathbf{Z}} (a_n \exp(2\pi i n \alpha)) \exp(2\pi i n x).$$

We see that f is R_α-invariant if and only if $a_n = a_n \exp(2\pi i n\alpha)$ holds for all n by the uniqueness of Fourier series.

If α is irrational, then $\exp(2\pi i n\alpha) \neq 1$ for all $n \neq 0$. Hence $a_n = 0$ for all $n \neq 0$, if f is invariant. This proves that $f(x) = a_0$ almost everywhere.

If α is rational, then $n_1\alpha$ is an integer for some $n_1 \neq 0$, hence the n_1'th condition holds irrespective of the value of a_{n_1}. Thus $f(x) = \cos(2\pi i n_1 x)$ is R_α invariant and not constant almost everywhere.

4. Ergodic theorems

In the previous section, we discussed recurrence, which is concerened with orbits visiting a given set infinitely often. We have not yet addressed the question how frequent these visits are, which is what we pursue now.

Theorem 18 (Mean Ergodic Theorem, von Neumann)**.** *Let $(X, \mathcal{B}, \mu, T)$ be an MPS and denote by*

$$I := \{f \in L^2(X) : f \circ T = f \ \text{a.e.}\}$$

the closed subspace of T-invariant functions. Denote by $P_T : L^2(X) \to I$ the orthogonal projection. Then for every $f \in L^2(X)$:

$$\frac{1}{N}\sum_{n=0}^{N-1} f \circ T^n \to P_T f \quad \text{in } L^2(X).$$

Theorem 19 (Pointwise Ergodic Theorem, Birkhoff)**.** *Let $(X, \mathcal{B}, \mu, T)$ be an MPS. Then for every $f \in L^1(X)$, there is a T-invariant function $f^* \in L^1(X)$ such that*

$$\frac{1}{N}\sum_{n=0}^{N-1} f(T^n x) \to f^*(x) \quad \text{pointwise } \mu\text{-almost everywhere.}$$

When $f \in L^2(X)$, then of course the function f^* in the Pointwise Ergodic Theorem is $P_T f$. There are also L^p variants of the Mean Ergodic Theorem for $1 \leq p < \infty$, see the example sheet. The quantity $(1/N)\sum_{n=0}^{N-1} f \circ T^n$ is called an **ergodic average**.

To understand the meaning of these results, we look at the case, when the system is ergodic. In that case, the limit function f^* (or $P_T f$ in case of the Mean Ergodic Theorem) is constant almost everywhere being T invariant. To figure out the value of this constant we look at the integral of the ergodic averages. We see that

$$\int \frac{1}{N}\sum_{n=0}^{N-1} f \circ T^n d\mu = \int f d\mu$$

for all N. Using the L^1 version of the Mean Ergodic Theorem (or assuming $f \in L^2(X)$) we get $\int f^* d\mu = \int f d\mu$. Thus $f^* = \int f d\mu$ almost everywhere.

Hence we can write the conclusion of the ergodic theorems as

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

in the appropriate sense (i.e. almost everywhere, or in L^2.) The quantities on the left are called time averages, because in a physical system they correspond to averages of the results of measurements at time $0, 1, \ldots, N-1$. (The measurement is the evaluation of the function f at a given point, which represent a state of the physical system.) The quantity on the right is called the space average, as it corresponds to the average of the result of the measurement taken at all possible states of the system with respect to the invariant measure. In this language the ergodic theorems are interpreted as the convergence of the time averages to the space average.

We begin with some background material concerning the operator $f \mapsto f \circ T$.

Lemma 20. *Let $(X, \mathcal{B}, \mu)$ be a probability space and let $T : X \to X$ be a measurable transformation. Then T is measure preserving if and only if*

$$\int_X f \circ T d\mu = \int_X f d\mu \tag{2}$$

for all $f \in L^1(X)$.

Proof. The "If part": Let $A \in \mathcal{B}$ and note that $x \in T^{-1}(A)$ is equivalent to $Tx \in A$. Hence we can write:

$$\mu(T^{-1}(A)) = \int \chi_{T^{-1}A} d\mu = \int \chi_A \circ T d\mu = \int \chi_A d\mu = \mu(A),$$

which proves that T is measure preserving.

The "Only if part": Similarly to the argument of the "If part" we can show that (2) holds for characteristic functions. Then by linearity of integration, it follows also for simple functions, that is functions that take only finitely many different values.

Now let $f \in L^1(X)$ and assume that it is non-negative. Let $f_1, f_2, \ldots$ be a monotone increasing sequence of simple functions with $f = \lim f_n$ almost everywhere. Then $f \circ T = \lim f_n \circ T$ almost everywhere. (Here we used the measure preserving property.) Hence using the definition of integration and (2) for simple functions:

$$\int f \circ T d\mu = \lim \int f_n \circ T d\mu = \lim \int f_n d\mu = \int f d\mu.$$

If $f \in L^1(X)$ is not non-negative, then we can write it as the difference of two non-negative functions and conclude (2) by linearity of integration. □

Definition 21. Let $(X, \mathcal{B}, \mu, T)$ be an MPS. If $f : X \to \mathbf{C}$ is a measurable function, we write

$$U_T f = f \circ T$$

and call U_T the **Koopman operator**.

Lemma 22. *The Koopman operator U_T is an isometry on the Hilbert space $L^2(X)$. That is to say*

$$\langle f, g \rangle = \langle U_T f, U_T g \rangle$$

for all $f, g \in L^2(X)$.

Proof. Apply the previous lemma to the function $f \cdot \overline{g} \in L^1(X)$. □

Definition 23. An MPS $(X, \mathcal{B}, \mu, T)$ is said to be invertible, if there is a measure preserving map $S : X \to X$ such that $T \circ S = S \circ T = Id_X$ almost everywhere. If such a map exists, it is customary to denote it by T^{-1}.

Example 24. The circle rotation is invertible, but the times 2 map is not.

Lemma 25. *If the system $(X, \mathcal{B}, \mu, T)$ is invertible, then U_T is unitary on $L^2(X)$ and $U_T^* = U_{T^{-1}}$.*

Proof. We use the first lemma of the lecture to the function $f \cdot \overline{(g \circ T^{-1})}$:

$$\langle f, U_{T^{-1}} g \rangle = \int f \cdot \overline{(g \circ T^{-1})} d\mu = \int (f \cdot \overline{(g \circ T^{-1})}) \circ T d\mu = \int f \circ T \cdot \overline{g} d\mu = \langle U_T f, g \rangle.$$

This shows that $U_T^* = U_{T^{-1}}$, indeed. Clearly $U_T U_{T^{-1}} = U_{T^{-1}} U_T = Id_{L^2(X)}$, which proves that U_T is unitary. □

The proof of both ergodic theorems (at least the proofs that we will give) rely on the observation that the convergence can be proved easily for certain special functions. First, the ergodic averages of a T-invariant function $f \in I$ are equal to f, so they converge to f.

If $f = U_T g - g$ for some g, then the ergodic averages become telescopic sums and only the boundary terms remain, that is:

$$\frac{1}{N} \sum_{n=0}^{N-1} U_T^n (U_T g - g) = \frac{1}{N} (U_T^N g - g)$$

and the right hand side is easily seen to converge to 0.

The following lemma shows that these two type of functions are enough to look at.

Lemma 26. *Write*

$$B := \{U_T g - g : g \in L^2(X)\}.$$

Then $I = B^\perp$.

It is important to note that the space B is not closed. So we have $L^2(X) = I \oplus \overline{B}$, but not every function in $L^2(X)$ is the sum of a T-invariant function and a cocycle. (The elements of B are called cocycles.)

Proof. We can write

$$f \in B^\perp \Leftrightarrow \langle f, U_T g - g\rangle = 0 \text{ for all } g \in L^2(X)$$
$$\Leftrightarrow \langle f, g\rangle = \langle f, U_T g\rangle = \langle U_T^* f, g\rangle \text{ for all } g \in L^2(X) \Leftrightarrow f = U_T^* f.$$

If the system was invertible, then we could finish the proof by applying $U_T = (U_T^*)^{-1}$ to both sides of the last equation.

We prove that $f = U_T f \Leftrightarrow f = U_T^* f$ holds in the general case, too. We can write

$$\begin{aligned} f =& U_T f \Leftrightarrow \langle f - U_T f, f - U_T f\rangle = 0 \\ \Leftrightarrow& \|f\|_2^2 + \|U_T f\|_2^2 - \langle f, U_T f\rangle - \langle U_T f, f\rangle = 0 \\ \Leftrightarrow& \|f\|_2^2 + \|U_T^* f\|_2^2 - \langle U_T^* f, f\rangle - \langle f, U_T^* f\rangle + (\|U_T f\|_2^2 - \|U_T^* f\|_2^2) = 0 \\ \Leftrightarrow& \|f - U_T^* f\|_2^2 + (\|U_T f\|_2^2 - \|U_T^* f\|_2^2) = 0 \end{aligned}$$

We note that both terms in the left hand side of the last equation are non-negative. This is clear about the first term. To show it for the second term, we observe that $\|U_T f\|_2 = \|f\|_2$, since U_T is an isometry, and $\|U_T^* f\|_2 \le \|f\|_2$, as $\|U_T^*\| = \|U_T\| = 1$.

It follows then that

$$f = U_T f \Leftrightarrow f = U_T^* f \text{ and } \|f\|_2 = \|U_T^* f\|_2.$$

The second statement on the right follows from the first one, hence $f = U_T f \Leftrightarrow f = U_T^* f$, as required. □

Proof of the Mean Ergodic Theorem. Let $f \in L^2(X)$ and fix an $\varepsilon > 0$. By the lemma, we can write $f = P_T f + U_T g - g + e$, for some $e, g \in L^2(X)$ such that $\|e\|_2 < \varepsilon$. Thus

$$\limsup_{N\to\infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} U_T^n f - P_T f \right\|_2 \le \limsup_{N\to\infty} \left\| \frac{1}{N}(U_T^N g - g) + \frac{1}{N} \sum_{n=0}^{N-1} U_T^n e \right\|_2 < \varepsilon.$$

We can conclude the proof by taking $\varepsilon \to 0$. □

5. Proof of the pointwise ergodic theorem

Theorem 27 (Wiener, Maximal Ergodic Theorem)**.** *Let $(X, \mathcal{B}, \mu, T)$ be a MPS and let $f \in L^1(X)$ be a function. Write*

$$E_\alpha = \left\{ x \in X : \sup_{N \in \mathbf{Z}_{>0}} \frac{1}{N} \sum_{n=0}^{N-1} U_T^n f(x) > \alpha \right\}.$$

Then $\mu(E_\alpha) \le \alpha^{-1} \|f\|_1$.

Proposition 28. *Let $(X, \mathcal{B}, \mu, T)$ be a MPS and let $f \in L^1(X)$ be a function. Write*

$$f_0 = 0,\ f_1 = f,\ f_2 = f + U_T f, \ldots,$$
$$f_n = f + U_T f + \ldots + U_T^{n-1} f, \ldots$$

and

$$F_N(x) = \max_{0 \le i \le N-1} \{f_i(x)\}.$$

Then

$$\int_{\{x \in X : F_N(x) > 0\}} f d\mu \ge 0.$$

Proof. Fix an $x \in X$ is such that $F_N(x) > 0$. Then $F_N(x) = f_j(x)$ for some $1 \le j \le N-1$. Hence

$$F_N(x) = U_T f_{j-1}(x) + f(x) \le U_T F_N(x) + f(x).$$

Thus $F_N(x) > 0$ implies that

$$f(x) \ge F_N(x) - U_T F_N(x).$$

We can write

$$\begin{aligned}\int_{\{x \in X : F_N(x) > 0\}} f d\mu &\ge \int_{\{x \in X : F_N(x) > 0\}} (F_N(x) - U_T F_N(x)) d\mu \\ &\ge \int_X (F_N(x) - U_T F_N(x)) d\mu = 0.\end{aligned}$$

The second inequality holds because $U_T F_n$ is always non-negative and hence $F_N(x) = 0$ implies that the integrand $F_N(x) - U_T F_N(x)$ is non-positive. □

Proof of the Maximal Ergodic Theorem. Put

$$\begin{aligned}E_{\alpha,M} &= \left\{ x \in X : \max_{1 \le N \le M} \frac{1}{N} \sum_{n=0}^{N-1} U_T^n f(x) > \alpha \right\} \\ &= \left\{ x \in X : \max_{1 \le N \le M} \sum_{n=0}^{N-1} (U_T^n f(x) - \alpha) > 0 \right\}.\end{aligned}$$

By the proposition applied to the function $f - \alpha$, we have

$$0 \le \int_{E_{\alpha,M}} (f - \alpha) d\mu \le \int_{E_{\alpha,M}} f d\mu - \alpha\mu(E_{\alpha,M}) \le \|f\|_1 - \alpha\mu(E_{\alpha,M}).$$

This implies $\mu(E_{\alpha,M}) \le \alpha^{-1}\|f\|_1$. We can conclude the proof by noting $E_\alpha = \bigcup E_{\alpha,M}$. (The sets $E_{\alpha,M}$ are increasing with M.) □

Proof of the Pointwise Ergodic Theorem. We fix a number $\varepsilon > 0$. We can write $f = f_\varepsilon + e_{1,\varepsilon}$ such that $f_\varepsilon \in L^2$ and $\|e_{1,\varepsilon}\|_1 < \varepsilon$.

Recall that $I = \{f \in L^2(X) : U_T f = f\}$ denotes the space of invariant functions and $B = \{U_T g - g : g \in L^2(X)\}$. As we have seen

in the proof of the Mean Ergodic Theorem, we have $L^2(X) = I \oplus \overline{B}$. Hence we can write

$$f_\varepsilon = P_T f_\varepsilon + U_T g_\varepsilon - g_\varepsilon + e_{2,\varepsilon}$$

such that $\|e_{2,\varepsilon}\|_1 \leq \|e_{2,\varepsilon}\|_2 < \varepsilon$. (Recall that P_T denotes the orthogonal projection to the space I.)

Furthermore, we can write $g_\varepsilon = h_\varepsilon + e_{3,\varepsilon}$ such that $\bar{h}_\varepsilon \in L^\infty(X)$ and$\|e_{3,\varepsilon}\|_1 < \varepsilon$. Putting everything together, we obtain

$$f = P_T f_\varepsilon + U_T h_\varepsilon - h_\varepsilon + e_\varepsilon,$$

where $\|e_\varepsilon\|_1 \leq 4\varepsilon$. (Here we used that the L^2-norm always dominates the L^1-norm on a probability space.)

We note that for every $x \in X$, we have

$$\Big|\frac{1}{N}\sum_{n=0}^{N-1} U_T^n f(x) - P_T f_\varepsilon(x)\Big| \leq \frac{2}{N}\|h_\varepsilon\|_\infty + \Big|\frac{1}{N}\sum_{n=0}^{N-1} U_T^n e_\varepsilon(x)\Big|. \tag{3}$$

We estimate the measure of the set of points x, where the right hand side is large using the Maximal Ergodic Theorem. We consider the set

$$E_{m,\varepsilon} := \Big\{x \in X : \limsup_{N\to\infty} |\frac{1}{N}\sum_{n=0}^{N-1} U_T^n f(x) - P_T f_\varepsilon(x)| > \frac{1}{m}\Big\}.$$

Using (3) and the Maximal Ergodic theorem for e_ε and $-e_\varepsilon$ with $\alpha = 1/m$ we obtain:

$$\mu(E_{m,\varepsilon}) \leq 2 \cdot m \cdot 4\varepsilon.$$

At this point, we intend to take the limit $\varepsilon \to 0$, however, the fact that $P_T f_\varepsilon$ depends on ε presents a difficulty. One way to overcome this problem is to first focus on the convergence of the ergodic averages only without identifying the limit.

We denote by F the set of points x such that the ergodic averages $(1/N)\sum_{n=0}^{N-1} U_T^n f(x)$ does not converge at x. We aim to show that $\mu(F) = 0$. Write for all $m \in \mathbf{Z}_{>0}$

$$F_m := \Big\{x \in X : \limsup_{N_1,N_2\to\infty} \Big|\frac{1}{N_1}\sum_{n=0}^{N_1-1} U_T^n f(x) - \frac{1}{N_2}\sum_{n=0}^{N_2-1} U_T^n f(x)\Big| > \frac{2}{m}\Big\}.$$

It is clear that $F \subset \bigcup F_m$, so it is enough to show that $\mu(F_m) = 0$ for all m.

We observe that $F_m \subset E_{m,\varepsilon}$ for all m, ε, hence $\mu(F_m) \leq 8m\varepsilon$. We take $\varepsilon \to 0$ and conclude that $\mu(F_m) = 0$, as required.

We proved that the ergodic averages convere to a limit function f^* almost everywhere. By Fatou's Lemma, $f^* \in L^1$.

To prove T-invariance, we observe that for almost all x, we have

$$f^*(x) = \lim \frac{1}{N+1} \sum_{n=0}^{N} U_T^n f(x) = \lim \frac{1}{N} \sum_{n=0}^{N} U_T^n f(x),$$

$$f^*(Tx) = \lim \frac{1}{N} \sum_{n=0}^{N-1} U_T^n f(Tx) = \lim \frac{1}{N} \sum_{n=1}^{N} U_T^n f(x).$$

Subtracting these two equations, we find

$$f^*(Tx) - f^*(x) = \lim \frac{-f(x)}{N} = 0.$$

□

Definition 29. Let $x \in [0,1)$ be a number and $K \geq 2$ an integer. Write $x = 0.a_1a_2\ldots_{(K)}$ for the base K expansion of x. We say that the number x is normal in base K, if for all integer $M \geq 1$ and $b_1, \ldots, b_M \in \{0, \ldots, K-1\}$ we have

$$\frac{|\{1 \leq n \leq N : a_{n+j} = b_j \text{ for all } 1 \leq j \leq M\}|}{N} \to \frac{1}{K^M}.$$

Theorem 30. *Almost every number $x \in [0,1)$ is normal in every base $K \geq 2$.*

Proof. We apply the pointwise ergodic theorem for the system $(\mathbf{R}/\mathbf{Z}, \mathcal{B}, m, T_K)$, where m is the Lebesgue measure and T_K is the map $x \mapsto K \cdot x$. This system is ergodic; see the first example sheet.

Put $A = [\sum_{j=1}^{M} b_j K^{-j}, \sum_{j=1}^{M} b_j K^{-j} + K^{-M})$. Observe that $T^n x \in A$ if and only if $a_{n+j} = b_j$ for all $1 \leq j \leq M$. Then the pointwise ergodic theorem applied for $f = \chi_A$, for every set A of this form, implies that almost every $x \in [0,1)$ is normal in base K. Here we use that there is a countable collection of sets A that we need to consider, hence the union of the countably many sets of exceptional points is of measure 0.

The union for K of the set of non-normal numbers in base K is of measure 0, since there are only countably many K. □

6. Unique ergodicity

In the previous lecture, we applied the pointwise ergodic theorem to show that a property holds for almost every numbers. This result gives no insight whether the property holds for specific numbers, e.g. e, π or $2^{1/3}$. In this lecture, we study the question: In which situation does the pointwise ergodic theorem hold for all points?

To make sense of this question, we work in the continuous category (measurable functions are not even defined everywhere).

If there are more than one T-invariant measures for a transformation T, then we can apply the Pointwise Ergodic Theorem for both measures, which may predict different limits for the ergodic averages

in each case. This is not a contradiction, because the set, where convergence holds in one case may be a null-set with respect to the other measure. However, this suggests, that the everywhere convergence of ergodic averages prohibits the presence of multiple invariant measures. The next result confirms this intuition.

Definition 31. Let $T : X \to X$ be a continuous map on a compact metric space. We say that the topological dynamical system (X, T) is uniquely ergodic if there is exactly one T-invariant Borel probability measure.

Theorem 32. *Let X be a compact metric space and let $T : X \to X$ be a continuous map. The following are equivalent*

(1) *The system (X, T) is uniquely ergodic.*
(2) *For every $f \in C(X)$ there is a number $c_f \in \mathbf{C}$ such that*

$$\frac{1}{N}\sum_{n=0}^{N-1} U_T^n f(x) \to c_f$$

uniformly in $x \in X$.
(3) *There is a dense subset $A \subset C(X)$ such that for every $f \in A$ there is a number $c_f \in \mathbf{C}$ such that*

$$\frac{1}{N}\sum_{n=0}^{N-1} U_T^n f(x) \to c_f$$

for every $x \in X$ not necessarily uniformly.

We begin by recalling the following result.

Theorem 33 (Riesz Representation Theorem)**.** *Let X be a compact metric space and denote by $C(X)$ the space of continuous functions. To a complex Borel measure μ, we associate a bounded functional L_μ on $C(X)$ as follows:*

$$L_\mu f := \int f d\mu.$$

Then the operation $\mu \to L_\mu$ is a bijection (in fact a Banach space isomorphism) between the space of complex Borel measures on X and the space of bounded linear functionals on $C(X)$.

Most of the time we will only need the following corollary, which is much simpler than the theorem itself.

Corollary 34. *If two probability measures $\mu_1, \mu_2 \in \mathcal{M}(X)$ satisfy*

$$\int f d\mu_1 = \int f d\mu_2$$

for all $f \in C(X)$, then $\mu_1 = \mu_2$.

For the rest of this lecture, X denotes a compact metric space. We denote by $\mathcal{M}(X)$ the space of Borel probability measures on X. If $T : X \to X$ is a continuous map, then we define the push-forward of a measure $\mu \in \mathcal{M}(X)$ as

$$T_*\mu(A) := \mu(T^{-1}(A))$$

for all Borel sets $A \in \mathcal{B}$. We note that T preserves the measure μ (or in other words μ is T-invariant) if and only if $T_*\mu = \mu$.

Lemma 35. *We have*

$$\int f dT_*\mu = \int f \circ T d\mu$$

for every bounded measurable functions f.

Proof Sketch. First we consider a characteristic function $f = \chi_A$ for a Borel set $A \in \mathcal{B}$. Then

$$\int \chi_A dT_*\mu = T_*\mu(A) = \mu(T^{-1}A)$$
$$\int \chi_A \circ T d\mu = \int \chi_{T^{-1}A} d\mu = \mu(T^{-1}A),$$

and we see that the claim holds. The claim can be verified for simple functions and then for general Borel functions in the same way as we did before. □

The following is a useful characterization of the measure preserving property of continuous maps.

Lemma 36. *Let $T : X \to X$ be a continuous transformation. A measure $\mu \in \mathcal{M}(X)$ is T-invariant if and only if*

$$\int f d\mu = \int f \circ T d\mu \tag{4}$$

for all $f \in C(X)$.

Proof. We have already seen that (4) holds even for all functions in L^1 if the measure is T-invariant.

For the other direction, we note that by the previous lemma

$$\int f dT_*\mu = \int f \circ T d\mu = \int f d\mu$$

for all $f \in C(X)$. We have then $T_*\mu = \mu$, by the Riesz representation theorem. Thus μ is T-invariant, indeed. □

The next result is a powerful tool to construct invariant measures. In particular it shows that there is at least one always if T is a continuous map on a compact metric space.

Theorem 37. *Let $T : X \to X$ be a continuous map on a compact metric space. Let $\nu_j \subset \mathcal{M}(X)$ be a sequence of probability measures on X and let $\{N_j\} \subset \mathbf{Z}_{\geq 0}$ be a sequence such that $N_j \to \infty$. Then any weak limit point of the sequence of measures*

$$\frac{1}{N_j} \sum_{n=0}^{N_j-1} T_*^n \nu_j$$

is invariant for T.

Proof. Assume that

$$\mu = \text{lim-w} \frac{1}{N_j} \sum_{n=0}^{N_j-1} T_*^n \nu_j$$

for otherwise we can pass to a subsequence. Then for every $f \in C(X)$ we have

$$\begin{aligned}\int f \circ T d\mu - \int f d\mu &= \lim \Big[\frac{1}{N_j} \sum_{n=0}^{N_j-1} \int f \circ T dT_*^n \nu_j - \frac{1}{N_j} \sum_{n=0}^{N_j-1} \int f dT_*^n \nu_j \Big] \\ &= \lim \Big[\frac{1}{N_j} \sum_{n=0}^{N_j-1} \int f \circ T^{n+1} d\nu_j - \frac{1}{N_j} \sum_{n=0}^{N_j-1} \int f \circ T^n d\nu_j \Big] \\ &= \lim \frac{1}{N_j} \Big[\int f \circ T^{N_j} d\nu_j - \int f d\nu_j \Big] = 0.\end{aligned}$$

□

Proof of Theorem 32. (1) $\Rightarrow$ (2) : Denote by $\mu \in \mathcal{M}(X)$ the unique invariant measure. Suppose to the contrary that (2) fails for $c_f = \int f d\mu$. Then there is a sequence of points $\{x_j\} \subset X$ and a sequence of positive integers $\{N_j\} \subset \mathbf{Z}_{\geq 0}$ with $N_j \to \infty$ such that

$$\Big| \frac{1}{N_j} \sum_{n=0}^{N_j-1} U_T^n f(x_j) - \int f d\mu \Big| \geq \varepsilon$$

for some $\varepsilon > 0$ and all j. Moreover, we can assume that

$$\lim \frac{1}{N_j} \sum_{n=0}^{N_j-1} U_T^n f(x_j) = a$$

for some number $a \in \mathbf{C}$, otherwise we pass to a subsequence. Necessarily $a \neq \int f d\mu$, indeed $\big|a - \int f d\mu\big| \geq \varepsilon$.

We let ν to be any weak limit point of the sequence of measures

$$\frac{1}{N_j} \sum_{n=0}^{N_j-1} T_*^n \delta_{x_j}.$$

We know that it is T invariant, but $\int f d\nu = a \neq \int f d\mu$. Hence $\nu \neq \mu$, which is a contradiction.

(3) $\Rightarrow$ (1) : Let μ and ν be two T-invariant probability measures. By dominated convergence,

$$\lim_{N\to\infty} \int \frac{1}{N} \sum_{n=0}^{N-1} f \circ T^n d\mu = c_f$$

for all $f \in A$. Since

$$\int \frac{1}{N} \sum_{n=0}^{N-1} f \circ T^n d\mu = \int f d\mu$$

for all N, this implies $\int f d\mu = c_f$.

Similarly, $\int f d\nu = c_f$ for all $f \in A$. Then $\int f d\mu = \int f d\nu$ holds for all $f \in A$. It also holds for all $f \in C(X)$, because A is dense in $C(X)$. Hence $\mu = \nu$, as required. □

This theorem illustrates the importance of understanding the structure of measures invariant for a given continuous map. In many cases, the invariant measures are hopeless to classify, but sometimes they have nice structure. The following is a very important question raised by Furstenberg.

Open Problem 38. What are the Borel probability measures on $\mathbf{R}/\mathbf{Z}$ that are invariant under both $T_2 : x \mapsto 2x$ and $T_3 : x \mapsto 3x$. The Lebesgue measure is one example, and it is also easy to give examples that are supported on finitely many rational points. Any known examples are convex combinations of these.

Later in the course we will discuss Rudolph's theorem, which claims that a measure that is invariant and ergodic under the semigroup generated by T_2 and T_3 and that has positive entropy with respect to T_2 is necessarily the Lebesgue measure.

Example 39. If α is irrational, then the circle rotation $(\mathbf{R}/\mathbf{Z}, \mathcal{B}, m, R_\alpha)$ is uniquely ergodic. Indeed, let μ be an invariant measure. Then for all $n \in \mathbf{Z}$

$$\begin{aligned}\int \exp(-2\pi i n x) d\mu &= \int \exp(-2\pi i n (x+\alpha)) d\mu \\ &= \exp(-2\pi i n \alpha) \cdot \int \exp(-2\pi i n x) d\mu.\end{aligned}$$

Since $n\alpha$ is not an integer for any $n \in \mathbf{Z}$ except for $n = 0$, we have $\exp(-2\pi i n\alpha) \neq 1$. Thus

$$\widehat{\mu}(n) = \int \exp(-2\pi i n x) d\mu = 0$$

for all $n \neq 0$. This implies that μ is the Lebesgue measure. (If you have not seen this before, look up the Stone Weierstrass theorem and show that the set of trigonometric polynomials (i.e. finite linear combinations of the functions $\exp(-2\pi i n x)$) form a dense subset of $C(\mathbf{R}/\mathbf{Z})$.)

Definition 40. A sequence $\{x_n\} \subset [0,1]$ is **equidistributed** if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \int f d\mu$$

holds for all $f \in C(\mathbf{R}/\mathbf{Z})$.

Remark 41. If $\{x_n\} \subset [0,1]$ is **equidistributed**, then

$$\frac{|\{1 \le n \le N : a \le x_n < b\}|}{N} \to b - a$$

for every numbers $0 \le a < b \le 1$.

The following is an immediate corollary of the unique ergodicity of circle rotations.

Theorem 42. *The sequence $\{\alpha n\}$ is equidistributed if α is irrational.*

7. Equidistribution of polynomials

The goal of this section is to generalize the equidistribution result for $\{\alpha n\}$ to polynomial sequences. We begin with a theorem of Furstenberg, which we will use to construct uniquely ergodic systems that are suited to that application.

Theorem 43 (Furstenberg)**.** *Let X be a compact metric space and let $T : X \to X$ be a continuous transformation. Suppose (X, T) is uniquely ergodic and denote by μ the invariant measure. Let $c : X \to \mathbf{R}/\mathbf{Z}$ be a continuous function. Define the map S on $X \times \mathbf{R}/\mathbf{Z}$ by*

$$S(x, y) = (Tx, c(x) + y).$$

Then $\mu \times m$ is S-invariant. If the measure $\mu \times m$ is ergodic, then the system $(X \times \mathbf{R}/\mathbf{Z}, S)$ is uniquely ergodic.

Remark 44. This result holds with an arbitrary compact metrizable topological group in place of $\mathbf{R}/\mathbf{Z}$, which need not be Abelian. The name of this construction is **skew-product**.

One of the key notions used in the proof of Furstenberg's theorem is genericity that we define now.

Definition 45. Let X be a compact metric space, let $T : X \to X$ be a continuous transformation and let μ be a T-invariant Borel probability measure. A point $x \in X$ is called generic with respect to μ, if

$$\text{lim-w}_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \delta_{T^n x} = \mu.$$

Proposition 46. *Let $(X, \mathcal{B}, \mu, T)$ be as above, and suppose that it is an ergodic MPS. Then μ-almost every $x \in X$ is generic with respect to μ.*

Proof. Let $F \subseteq C(X)$ be a dense countable set. By the pointwise ergodic theorem, for every $f \in F$, there is $X_f \in \mathcal{B}$ with $\mu(X_f) = 1$ such that

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

Fix $x \in \bigcap_{f\in F} X_f$. Let $g \in C(X)$. Fix $\varepsilon > 0$ and choose $f \in F$ such that $\|g - f\|_\infty < \varepsilon$. Then

$$\Big|\frac{1}{N} \sum_{n=0}^{N-1} f(T^n x) - \frac{1}{N} \sum_{n=0}^{N-1} g(T^n x)\Big| < \varepsilon,$$
$$\Big|\int f d\mu - \int g d\mu\Big| < \varepsilon.$$

Thus

$$\limsup_{N\to\infty} \Big|\frac{1}{N} \sum_{n=0}^{N-1} g(T^n x) - \int g d\mu\Big| < 2\varepsilon.$$

We take $\varepsilon \to 0$ and conclude that x is generic with respect to μ, since g was arbitrary. □

Proof of Furstenberg's theorem on skew products. For showing that S preserves $\mu \times m$, we choose a testfunction $f \in C(X \times \mathbf{R}/\mathbf{Z})$ and write

$$\begin{aligned}\int_{\mathbf{R}/\mathbf{Z}} \int_X f(S(x,y)) d\mu(x) dy &= \int_X \int_{-c(x)}^{1-c(x)} f(Tx, c(x)+y) dy d\mu(x) \\ &= \int_X \int_0^1 f(Tx, y) dy d\mu(x) \\ &= \int_{\mathbf{R}/\mathbf{Z}} \int_X f(x,y) d\mu(x) dy.\end{aligned}$$

Assume that $\mu \times m$ is ergodic, and we prove unique ergodicity. Denote by $E \subset X \times \mathbf{R}/\mathbf{Z}$ the set of generic points with respect to $\mu \times m$. Our strategy is to show that E is very large.

We first observe $\mu \times m(E) = 1$. Then we prove the following.

Claim. If $(x, y) \in E$, then $(x, y + t) \in E$, also for every $t \in \mathbf{R}/\mathbf{Z}$.

Proof. For each $t \in \mathbf{R}/\mathbf{Z}$, we consider the transformation $U_t : X \times \mathbf{R}/\mathbf{Z} \to X \times \mathbf{R}/\mathbf{Z}$ defined as $U_t(x, y) = (x, y + t)$. The proof relies on the observation that $U_t \circ S = S \circ U_t$. Indeed,

$$U_t \circ S(x,y) = U_t(x, c(x)+y) = (x, c(x)+y+t) = S(x, y+t) = S \circ U_t(x,y).$$

To show that $(x, y+t)$ is generic, we fix a testfunction $f \in C(X \times \mathbf{R}/\mathbf{Z})$ and write

$$\begin{aligned}\frac{1}{N}\sum_{n=0}^{N-1} f(S^n(x,y+t)) =& \frac{1}{N}\sum_{n=0}^{N-1} f(S^n U_t(x,y)) \\ =& \frac{1}{N}\sum_{n=0}^{N-1} f\circ U_t(S^n(x,y)) \to \iint f\circ U_t d\mu(x)dy \\ =& \iint f(x,y)d\mu(x)dy.\end{aligned}$$

We used that $(x,y) \in E$ and the definition of genericity applied to the testfunction $f \circ U_t$ and then that U_t preserves $\mu \times m$, which is easy to check. □

By the above claim, there is a set $A \in \mathcal{B}(X)$ such that $E = A \times \mathbf{R}/\mathbf{Z}$. We also have $\mu(A) = \mu \times m(E) = 1$.

Let ν be an arbitrary S invariant Borel probability measure. Denote by $P : X \times \mathbf{R}/\mathbf{Z} \to X$ the coordinate projection. Then $P_*\nu$ is T-invariant, hence it is equal to μ. Indeed,

$$\begin{aligned}P_*\nu(T^{-1}B) =& \nu((T^{-1}B) \times \mathbf{R}/\mathbf{Z}) = \nu(S^{-1}(B \times \mathbf{R}/\mathbf{Z})) \\ =& \nu(B \times \mathbf{R}/\mathbf{Z}) = P_*\nu(B)\end{aligned}$$

for any $B \in \mathcal{B}(X)$.

Then $P_*\nu(A) = \mu(A) = 1$, hence $\nu(E) = \nu(A \times \mathbf{R}/\mathbf{Z}) = 1$. Fix $f \in C(X \times \mathbf{R}/\mathbf{Z})$. For all $x \in E$, hence for ν-almost every x, we have

$$\frac{1}{N}\sum_{n=0}^{N-1} f(S^n(x,y)) \to \int f d\mu \times m.$$

By dominated convergence, we have

$$\int \frac{1}{N}\sum_{n=0}^{N-1} f(S^n(x,y))d\nu \to \int f d\mu \times m.$$

The left hand side is equal to $\int f d\nu$ for all N, hence

$$\int f d\nu = \int f d\mu \times m.$$

since f was arbitrary, this proves that $\nu = \mu \times m$, hence $(X \times \mathbf{R}/\mathbf{Z}, S)$ is indeed uniquely ergodic. □

Corollary 47. *Let $\alpha \in \mathbf{R}/\mathbf{Z}$ be irrational and define the map S on $X = (\mathbf{R}/\mathbf{Z})^d$ by*

$$S(x_1, \dots, x_d) = (x_1 + \alpha, x_2 + x_1, \dots, x_d + x_{d-1}).$$

Then the system (X, S) is uniquely ergodic.

Proof. We argue by induction. If $d = 1$, then the system is an irrational circle rotation, hence it is uniquely ergodic.

We assume that $d > 1$ and the claim holds for $d - 1$. We can apply Furstenberg's above theorem, and we only need to show that the system is ergodic.

To that end, we fix an invariant bounded measurable function f on X and aim to show it is constant. We expand f in Fourier series and write:

$$\begin{aligned} f(x) &= \sum_{n\in\mathbf{Z}^d} a_n \exp(2\pi i n\cdot x),\\ f(Sx) &= \sum_{n\in\mathbf{Z}^d} a_n \exp(2\pi i n\cdot Sx)\\ &= \sum_{n\in\mathbf{Z}^d} a_n \exp(2\pi i(n_1(x_1+\alpha)+n_2(x_2+x_1)+\ldots+n_d(x_d+x_{d-1})))\\ &= \sum_{n\in\mathbf{Z}^d} a_n \exp(2\pi i n_1\alpha)\exp(2\pi i((n_1+n_2)x_1+\ldots\\ &\qquad\qquad +(n_{d-1}+n_d)x_{d-1}+n_d x_d)). \end{aligned}$$

Writing

$$\widehat{S}(n) = (n_1+n_2,\ldots,n_{d-1}+n_d,n_d)$$

for $n\in\mathbf{Z}^d$, we obtain

$$a_n\exp(2\pi i n_1\alpha) = a_{\widehat{S}(n)}.$$

This means that $|a_n| = |a_{\widehat{S}(n)}|$, in particular.

Let $m\in\mathbf{Z}^d$ be such that $a_m\neq 0$. By Parseval's formula,

$$\sum_{n\in\mathbf{Z}^d}|a_n|^2 = \|f\|_2^2.$$

Hence the number of $n\in\mathbf{Z}^d$ such that $|a_n| = |a_m|$ must be finite. This means that the sequence $\widehat{S}^k(m)$ must be periodic. Since $(\widehat{S}^k m)_{d-1} = m_{d-1} + km_d$, we have $m_d = 0$. By a similar argument, we can prove $m_2 = \ldots = m_d = 0$.

Finally, we consider the case $m_2 = \ldots = m_d = 0$ and $a_m\neq 0$. Then $\widehat{S}(m) = m$, hence $a_m\exp(2\pi i m_1\alpha) = a_m$, which implies $\exp(2\pi i m_1\alpha) = 1$. Since α is irrational, this implies $m_1 = 0$.

We showed that $a_n = 0$ unless $n = (0,\ldots,0)$, which proves that f is constant, which is precisely what we wanted to show. □

Theorem 48 (Weyl)**.** *Let $p(x) = a_d x^d + \ldots + a_1 x + a_0$ be a polynomial such that at least one of the coefficients a_i is irrational. Then the sequence $\{p(n)\}$ is equidistributed.*

Proof. We first consider the special case when a_d is irrational. We consider the system (X, S) defined in the corollary given above. It is easily proved by induction that

$$S^n \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} \binom{n}{0}x_1 + \binom{n}{1}\alpha \\ \binom{n}{0}x_2 + \binom{n}{1}x_1 + \binom{n}{2}\alpha \\ \vdots \\ \binom{n}{0}x_d + \binom{n}{1}x_{d_1} + \ldots + \binom{n}{d-1}x_1 + \binom{n}{d}\alpha \end{pmatrix}.$$

Here we wrote the vector vertically just for notational convenience.

Write

$$q_i(t) = \frac{t(t-1)\cdots(t-i+1)}{i!}.$$

Since $q_i(t)$ for $i = 0, \ldots, d$ is a basis in the vectorspace of polynomials of degree at most d, there are some real numbers α, x_1, x_2, $\ldots$, x_d such that

$$p(t) = q_0(t)x_d + q_1(t)x_{d-1} + \ldots + q_{d-1}(t)x_1 + q_d\alpha.$$

Moreover, $\alpha = a_d \cdot d!$ is irrational. Hence there are $\alpha, x_1, \ldots, x_d \in \mathbf{R}/\mathbf{Z}$ such that

$$p(n) = \binom{n}{0}x_d + \binom{n}{1}x_{d-1} + \ldots + \binom{n}{d-1}x_1 + \binom{n}{d}\alpha \mod \mathbf{Z} \tag{5}$$

for each $n \in \mathbf{Z}_{>0}$ and α is irrational.

Fix $f \in C(\mathbf{R}/\mathbf{Z})$ and define $g \in C((\mathbf{R}/\mathbf{Z})^d)$ by $g(x) = f(x_d)$. Then

$$\frac{1}{N}\sum_{n=0}^{N-1} f(p(n)) = \frac{1}{N}\sum_{n=0}^{N-1} g(S^n x), \tag{6}$$

where α and $x = (x_1, \ldots, x_d)$ are chosen in such a way that (5) holds. By unique ergodicity of the system (X, S), we conclude that (6) converges to

$$\int g dm^d = \int f dm,$$

as required.

Finally, we consider the general case and reduce it to the special case considered above. Let k be maximal such that a_k is irrational. Let $q \in \mathbf{Z}_{>0}$ be such that $qa_j \in \mathbf{Z}$ for all $j = d, \ldots, k+1$. Then

$$p(qn+b) = a_d b^d + \ldots + a_{k+1}b^{k+1} + a_k(qn+b)^k + \ldots a_1(qn+b) + a_0 \mod \mathbf{Z},$$

where the right hand side is a polynomial in n, whose leading coefficient is irrational. By the special case, we know then that $p(qn + b)$ is equidistributed for each $b = 0, \ldots, q-1$. Then the original sequence is also equidistributed. □

8. Mixing properties

Definition 49. A MPS $(X, \mathcal{B}, \mu, T)$ is said to be **mixing** if for every measurable sets $A, B \in \mathcal{B}$

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

We note that $\mu(T^{-n}(A) \cap B) = \mu(x \in X : x \in B, T^n x \in A)$, hence mixing can be interpreted as the property that the state of the system at time n becomes independent of the state at time 0 as n grows. This notion can be generalized to multiple sets and times as follows.

Definition 50. A MPS $(X, \mathcal{B}, \mu, T)$ is said to be **mixing on k sets** if the following holds. Let $A_0, \ldots, A_{k-1} \in \mathcal{B}$ and let $\varepsilon > 0$. Then there is a number $N \in \mathbf{Z}_{>0}$ such that for all $n_1, n_2, \ldots, n_{k-1} \in \mathbf{Z}_{>0}$ that satisfy

$$n_1 \geq N, n_2 - n_1 \geq N, \ldots, n_{k-1} - n_{k-2} \geq N$$

we have

$$|\mu(A_0 \cap T^{-n_1} A_1 \cap \ldots \cap T^{-n_{k-1}} A_{k-1}) - \mu(A_0) \cdots \mu(A_{k-1})| \leq \varepsilon.$$

It is clear that a MPS is mixing if and only if it is mixing on 2 sets. The following is a long standing open problem in ergodic theory.

Open Problem 51. Is there a MPS that is mixing on 2 sets but not mixing on 3 sets?

This basic problem being still open underlines the fact that mixing is not an easy notion to work with. We will introduce now the notion of weak mixing, which may look less natural at first sight but turns out to be more useful for the theory. This notion replaces the limit in the definition of mixing with another kind of convergence, which we define now.

Definition 52. We say that a set $S \subset \mathbf{Z}_{>0}$ has **full density** if

$$\lim \frac{1}{N} |S \cap [1, N]| = 1.$$

We say that a sequence of complex numbers $\{a_n\} \subset \mathbf{C}$ **converges in density** to a complex number $a \in \mathbf{C}$ if the set

$$\{n \in \mathbf{Z}_{>0} : |a - a_n| < \varepsilon\}$$

has full density for all $\varepsilon > 0$. We denote this by D-$\lim a_n = a$.

We say that a sequence of complex numbers $\{a_n\} \subset \mathbf{C}$ **Cesàro converges** to a complex number $a \in \mathbf{C}$ if

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} a_n = a.$$

We denote this by C-$\lim a_n = a$.

Definition 53. A MPS $(X, \mathcal{B}, \mu, T)$ is said to be **weak mixing** if

$$\text{D-lim}_{n\to\infty} \mu(T^{-n}(A) \cap B) = \mu(A) \cdot \mu(B)$$

for all $A, B \in \mathcal{B}$.

Lemma 54. *Let $\{a_n\} \subset \mathbf{R}$ be a bounded sequence and $a \in \mathbf{R}$. The following are equivalent.*

(1) $\text{D-lim}\, a_n = a$,
(2) $\text{C-lim}\, |a_n - a| = 0$,
(3) $\text{C-lim}(a - a_n)^2 = 0$,
(4) $\text{C-lim}\, a_n = a$ *and* $\text{C-lim}\, a_n^2 = a^2$.

Proof. (1) $\Rightarrow$ (2): Fix $\varepsilon > 0$ and denote by M the supremum of $|a_n|$. Let N be sufficiently large so that

$$\frac{1}{N}|\{n \in \{1, \dots, N\} : |a - a_n| < \varepsilon\}| > 1 - \varepsilon.$$

Then

$$\sum_{n=1}^{N} |a - a_n| \leq 2MN\varepsilon + \varepsilon N = (2MN + N)\varepsilon.$$

We finish the proof by taking $\varepsilon \to 0$.

(2) $\Rightarrow$ (1): Fix $\varepsilon > 0$. Then

$$\varepsilon|\{n \in \{1, \dots, N\} : |a - a_n| > \varepsilon\}| \leq \sum_{n=1}^{N} |a - a_n|.$$

Thus

$$\lim_{N\to\infty} \frac{1}{N}|\{n \in \{1, \dots, N\} : |a - a_n| > \varepsilon\}| = 0.$$

This proves the claim.

(1) $\Leftrightarrow$ (3) is very similar to (1) $\Leftrightarrow$ (2).

(1), (2) $\Rightarrow$ (4):

$$\lim_{N\to\infty} \left|\frac{1}{N}\sum_{n=1}^{N} a_n - a\right| = \lim_{N\to\infty} \left|\frac{1}{N}\sum_{n=1}^{N} (a_n - a)\right| \leq \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} |a_n - a| = 0$$

showing the first part of the claim. The second part can be proved similarly by noting that $\text{D-lim}\, a_n = a$ implies $\text{D-lim}\, a_n^2 = a^2$, which can be seen directly from the definition.

(4) $\Rightarrow$ (3):

$$\begin{aligned}\text{C-lim}(a - a_n)^2 = &\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} (a^2 - 2aa_n + a_n^2) \\ = &a^2 - 2a\,\text{C-lim}\, a_n + \text{C-lim}\, a_n^2 = a^2 - 2a^2 + a^2 = 0.\end{aligned}$$

□

Theorem 55. *Let $(X, \mathcal{B}, \mu, T)$ be an MPS. Then the following are equivalent.*

(1) $(X, \mathcal{B}, \mu, T)$ *is weak mixing,*
(2) $(X \times Y, \mathcal{B} \times \mathcal{C}, \mu \times \nu, T \times S)$ *is ergodic for all ergodic MPS* $(Y, \mathcal{C}, \nu, S)$*,*
(3) $(X \times X, \mathcal{B} \times \mathcal{B}, \mu \times \mu, T \times T)$ *is ergodic,*
(4) $(X \times X, \mathcal{B} \times \mathcal{B}, \mu \times \mu, T \times T)$ *is weak mixing,*
(5) *the operator* U_T *has no non-constant eigenfunctions.*

The following lemma, which appears in the second example sheet, will be used in the proof of the theorem.

Lemma 56. *Let* $(X, \mathcal{B}, \mu, T)$ *be a MPS. Let* $\mathcal{S} \subset \mathcal{B}$ *be a semi-algebra (or* π*-system) generating* $\mathcal{B}$*.*

The system $(X, \mathcal{B}, \mu, T)$ *is weak mixing if and only if*

$$\text{D-lim}_{n\to\infty} \mu(T^{-n}(A) \cap B) = \mu(A) \cdot \mu(B)$$

for all $A, B \in \mathcal{S}$*.*

The system $(X, \mathcal{B}, \mu, T)$ *is ergodic if and only if*

$$\text{C-lim}_{n\to\infty} \mu(T^{-n}(A) \cap B) = \mu(A) \cdot \mu(B)$$

for all $A, B \in \mathcal{S}$*.*

Proof of the theorem. (1) $\Rightarrow$ (2): We consider the semi-algebra of measurable rectangles

$$\mathcal{S} := \{B \times C : B \in \mathcal{B}, C \in \mathcal{C}\},$$

which generates the σ-algebra $\mathcal{B} \times \mathcal{C}$. We check that the property characterizing ergodicity in the previous lemma holds. Let $B_1 \times C_1, B_2 \times$

$C_2 \in \mathcal{S}$. Then

$$\begin{aligned}
&\Big|\frac{1}{N}\sum_{n=0}^{N-1} \mu\times\nu((T\times S)^{-n}(B_1\times C_1)\cap B_2\times C_2) \\
&\qquad\qquad - \mu\times\nu(B_1\times C_1)\mu\times\nu(B_2\times C_2)\Big| \\
&= \Big|\frac{1}{N}\sum_{n=0}^{N-1}[\mu(T^{-n}(B_1)\cap B_2)\nu(S^{-n}(C_1)\cap C_2) \\
&\qquad\qquad - \mu(B_1)\nu(C_1)\mu(B_2)\nu(C_2)]\Big| \\
&= \Big|\frac{1}{N}\sum_{n=0}^{N-1}[\mu(T^{-n}(B_1)\cap B_2)\nu(S^{-n}(C_1)\cap C_2) \\
&\qquad\qquad - \mu(B_1)\mu(B_2)\nu(S^{-n}(C_1)\cap C_2)]\Big| \\
&\quad + \Big|\frac{1}{N}\sum_{n=0}^{N-1}[\mu(B_1)\mu(B_2)\nu(S^{-n}(C_1)\cap C_2) - \mu(B_1)\nu(C_1)\mu(B_2)\nu(C_2)]\Big| \\
&\le \frac{1}{N}\sum_{n=0}^{N-1}|\mu(T^{-n}(B_1)\cap B_2) - \mu(B_1)\mu(B_2)| \\
&\quad + \mu(B_1)\mu(B_2)\Big|\frac{1}{N}\sum_{n=0}^{N-1}[\nu(S^{-n}(C_1)\cap C_2) - \nu(C_1)\nu(C_2)]\Big|
\end{aligned}$$

In the last inequality, we used the triangle inequality to estimate the first sum and then estimated $\nu(S^{-n}(C_1)\cap C_2)$ above by 1 in each term. Thus

$$\begin{aligned}
&\limsup_{N\to\infty}\Big|\frac{1}{N}\sum_{n=0}^{N-1} \mu\times\nu((T\times S)^{-n}(B_1\times C_1)\cap B_2\times C_2) \\
&\qquad\qquad - \mu\times\nu(B_1\times C_1)\mu\times\nu(B_2\times C_2)\Big| \\
&\qquad \le \lim_{N\to\infty}\frac{1}{N}\sum_{n=0}^{N-1}|\mu(T^{-n}(B_1)\cap B_2) - \mu(B_1)\nu(B_2)| \\
&\qquad\quad + \lim_{N\to\infty}\Big|\frac{1}{N}\sum_{n=0}^{N-1}[\nu(S^{-n}(C_1)\cap C_2) - \nu(C_1)\nu(C_2)]\Big| = 0.
\end{aligned}$$

The first sum converges to 0 since $(X,\mathcal{B},\mu,T)$ is weak mixing. The second sum converges to 0 since $(Y,\mathcal{C},\nu,S)$ is ergodic. This proves that the product system is indeed ergodic.

(2) $\Rightarrow$ (3): First, we apply (2) with Y being the one point space to see that $(X,\mathcal{B},\mu,T)$ is ergodic. Then (3) follows as a special case of (2).

(3) $\Rightarrow$ (1): Fix $B, C \in \mathcal{B}$. We first apply the characterization of ergodicity in the lemma for the sets $B \times X, C \times X \subset X \times X$:

$$\text{C-lim}\, \mu \times \mu((T \times T)^{-n}(B \times X) \cap C \times X) = \mu \times \mu(B \times X)\mu \times \mu(C \times X).$$

Unwinding the definitions, we write

$$\text{C-lim}\, \mu(T^{-n}(B) \cap C)\mu(T^{-n}(X) \cap X) = \mu(B)\mu(X)\mu(C)\mu(X)$$

and hence

$$\text{C-lim}\, \mu(T^{-n}(B) \cap C) = \mu(B)\mu(C).$$

We do the same with the sets $B \times B, C \times C \subset X \times X$:

$$\text{C-lim}\, \mu \times \mu((T \times T)^{-n}(B \times B) \cap C \times C) = \mu \times \mu(B \times B)\mu \times \mu(C \times C).$$

Thus

$$\text{C-lim}\, \mu(T^{-n}(B) \cap C)^2 = \mu(B)^2\mu(C)^2.$$

This together with the previous paragraph implies

$$\text{D-lim}\, \mu(T^{-n}(B) \cap C) = \mu(B)\mu(C),$$

which was to be proved.

(1) $\Leftrightarrow$ (4): This is very similar to the above arguments.

(3) $\Rightarrow$ (5): Let f be a non-constant eigenfunction of the operator U_T. We show that the function $\widetilde{f}(x_1, x_2) = f(x_1) \times \overline{f}(x_2) : X \times X \to \mathbf{C}$ is $T \times T$ invariant. By ergodicity of $(X \times X, \mathcal{B} \times \mathcal{B}, \mu \times \mu, T \times T)$, $\widetilde{f}$ is constant almost everywhere, hence f must be constant almost everywhere, also.

Now we show the claim. Denote by λ the eigenvalue of U_T corresponding to f. Since U_T is an isometry, $|\lambda| = 1$ necessarily. We can write

$$\widetilde{f}(Tx_1, Tx_2) = f(Tx_1)\overline{f}(Tx_2) = \lambda f(x_1)\overline{\lambda}\,\overline{f}(x_2) = f(x_1)\overline{f}(x_2) = \widetilde{f}(x_1, x_2).$$

This completes the proof. □

The proof of the implication that (5) implies weak mixing requires some knowledge of operator theory and it is not examinable. Two possible approaches are outlined in the second example sheet.

9. Multiple recurrence for weak mixing systems

Our next goal is to prove the following result.

Theorem 57. *Let $(X, \mathcal{B}, \mu, T)$ be a weak mixing MPS. Then for any $k \in \mathbf{Z}_{>0}$, and $f_1, \ldots, f_k \in L^\infty(X)$, we have*

$$\frac{1}{N}\sum_{n=0}^{N-1} U_T^n f_1 \cdot U_T^{2n} f_2 \cdots U_T^{kn} f_k \underset{in\ L^2}{\longrightarrow} \int f_1 d\mu \cdot \int f_2 d\mu \cdots \int f_k d\mu. \tag{7}$$

Corollary 58. *In the setting of the theorem, let $f_0 \in L^\infty$ be another function. Then*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} \int f_0 \cdot U_T^n f_1 \cdot U_T^{2n} f_2 \cdots U_T^{kn} f_k d\mu = \int f_0 d\mu \cdots \int f_k d\mu. \tag{8}$$

In particular, taking $f_0 = f_1 = \ldots = f_k = \chi_A$ for a set $A \in \mathcal{B}$ we obtain

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

This proves Furstenberg's multiple recurrence theorem for weak mixing systems.

Observe that (8) follows immediately from (7). Indeed, (7) claims that a certain sequence of functions converge in norm, while (8) claims that the inner products with any fixed function f_0 converge.

The following lemma is very useful in showing that the averages of a sequence of vectors converge to 0 in a Hilbert space. In the proof of Theorem 57 we will use it in the special case when $\int f_k d\mu = 0$.

Lemma 59 (van der Corput)**.** *Let $u_1, u_2, \ldots$ be a bounded sequence of vectors in a Hilbert space. Write*

$$s_h = \limsup_{N\to\infty} \frac{1}{N}\Big|\sum_{n=1}^{N} \langle u_n, u_{n+h}\rangle\Big|.$$

Suppose that D-lim $s_h = 0$*. Then*

$$\lim_{N\to\infty} \frac{1}{N}\Big\|\sum_{n=1}^{N} u_n\Big\| = 0.$$

Note that D-lim $s_h = 0$ is equivalent C-lim $s_h = 0$, since $s_h \geq 0$ for all h. The following would be a natural approach to prove the lemma. We can write

$$\begin{aligned}\Big\|\frac{1}{N}\sum_{n=1}^{N} u_n\Big\|^2 =& \frac{1}{N^2}\sum_{n_1=1}^{N}\sum_{n_2=1}^{N}\langle u_{n_1}, u_{n_2}\rangle \\ =& \frac{1}{N^2}\Big[\sum_{n=1}^{N}\|u_n\|^2 + 2\sum_{h=1}^{N-1}\sum_{n=1}^{N-h}\mathrm{Re}\langle u_n, u_{n+h}\rangle\Big]. \end{aligned} \tag{9}$$

Unfortunately, the expression $\sum_{n=1}^{N-h}\mathrm{Re}\langle u_n, u_{n+h}\rangle$ can be estimated by s_h only for h fixed and $N \to \infty$. Observe how we overcome this problem in the following proof.

Proof. We fix $\varepsilon > 0$ and let H be sufficiently large so that

$$\frac{1}{H}\sum_{h=0}^{H-1} s_h < \varepsilon.$$

If N is sufficiently large (depending on H, ε and $\sup \|u_n\|$) then

$$\Big\|\frac{1}{N}\sum_{n=1}^{N} u_n - \frac{1}{NH}\sum_{n=1}^{N}\sum_{h=1}^{H} u_{n+h}\Big\| \leq \frac{1}{N}\sum_{n=1}^{H}\|u_n\| + \frac{1}{N}\sum_{n=N+1}^{N+H}\|u_n\| < \varepsilon.$$

To see this, note that the contribution of u_n for $H < n \leq N$ is the same to both averages.

The above inequality allows us to turn our attention to the double average. We first use the triangle inequality:

$$\Big\|\frac{1}{NH}\sum_{n=1}^{N}\sum_{h=1}^{H} u_{n+h}\Big\| \leq \frac{1}{N}\sum_{n=1}^{N}\Big\|\frac{1}{H}\sum_{h=1}^{H} u_{n+h}\Big\|.$$

We use now the Cauchy-Schwartz inequality to turn the above into an average of norm squares that can be expanded similarly to (9). Note that this allows us to avoid inner products of vectors whose indices differ by more than H.

$$\begin{aligned}\Big\|\frac{1}{NH}\sum_{n=1}^{N}\sum_{h=1}^{H} u_{n+h}\Big\|^2 \leq& \frac{1}{N}\sum_{n=1}^{N}\Big\|\frac{1}{H}\sum_{h=1}^{H} u_{n+h}\Big\|^2\\ =&\frac{1}{N}\sum_{n=1}^{N}\frac{1}{H^2}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\langle u_{n+h_1}, u_{n+h_2}\rangle\\ \leq&\frac{1}{H^2}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\Big|\frac{1}{N}\sum_{n=1}^{N}\langle u_{n+h_1}, u_{n+h_2}\rangle\Big|.\end{aligned}$$

Keeping H fixed, we let $N \to \infty$:

$$\begin{aligned}\limsup_{N\to\infty}\Big\|\frac{1}{NH}\sum_{n=1}^{N}\sum_{h=1}^{H} u_{n+h}\Big\|^2 \leq&\frac{1}{H^2}\sum_{h_1=1}^{H}\sum_{h_2=1}^{H} s_{|h_1-h_2|}\\ \leq&\frac{1}{H^2}\sum_{h=1}^{H} 2H s_h < 2\varepsilon.\end{aligned}$$

We combine this with our first inequality and obtain

$$\limsup_{N\to\infty}\Big\|\frac{1}{N}\sum_{n=1}^{N} u_n\Big\| \leq \varepsilon + \sqrt{2\varepsilon}.$$

Letting $\varepsilon \to 0$, we can complete the proof. □

We recall two auxiliary results from the second example sheet that we need for the proof of the theorem.

Lemma 60. *Let $(X, \mathcal{B}, \mu, T)$ be a weak mixing MPS. Then for any $f, g \in L^2(X)$ we have*

$$\text{D-lim}\langle U_T^n f, g\rangle = \int f d\mu \cdot \int g d\mu.$$

Note that when f and g are characteristic functions of sets, then the displayed equation is precisely the definition of weak mixing.

Lemma 61. *If the MPS $(X, \mathcal{B}, \mu, T)$ is weak mixing than so is $(X, \mathcal{B}, \mu, T^k)$ for all $k \in \mathbf{Z}_{>0}$.*

Proof of Theorem 57. The proof is by induction on k. The claim for $k = 1$ is the mean ergodic theorem. We assume that $k > 1$ and that the theorem and the corollary holds for $k - 1$.

We first consider the special case when $\int f_k d\mu = 0$. We set

$$u_n = U_T^n f_1 \cdot U_T^{2n} f_2 \cdots U_T^{kn} f_k$$

and prove $\|(1/N)\sum_{n=1}^N u_n\|_2 \to 0$ using the van der Corput lemma. We compute

$$\begin{aligned}\langle u_n, u_{n+h}\rangle &= \int U_T^n f_1 \cdot U_T^{2n} f_2 \cdots U_T^{kn} f_k \\ &\qquad \cdot U_T^{n+h} f_1 \cdot U_T^{2(n+h)} f_2 \cdots U_T^{k(n+h)} f_k d\mu \\ &= \int U_T^n [(f_1 \cdot U_T^h f_1) \cdot (U_T^n f_2 \cdot U_T^{n+2h} f_2) \cdots \\ &\qquad \cdot (U_T^{(k-1)n} f_k \cdot U_T^{(k-2)n+kh} f_k)] d\mu \\ &= \int (f_1 \cdot U_T^h f_1) \cdot U_T^n (f_2 \cdot U_T^{2h} f_2) \cdots U_T^{(k-1)n} (f_k \cdot U_T^{kh} f_k) d\mu.\end{aligned}$$

The last equation used the measure preserving property.

We use now the induction hypothesis, more precisely (8) for $k - 1$ and $f_i U_T^{ih} f_i$ in place of f_{i-1}:

$$\lim_{n\to\infty} \frac{1}{N} \sum_{n=1}^N \langle u_n, u_{n+h}\rangle = \int f_1 \cdot U_T^h f_1 d\mu \cdot \int f_2 \cdot U_T^{2h} f_2 d\mu \cdots \int f_k \cdot U_T^{kh} f_k d\mu.$$

We note that

$$\Big| \int f_i \cdot U_T^{ih} f_i d\mu \Big| \le \|f_i\|_\infty^2$$

for all $1 \le i \le k$. We apply this for $1 \le i \le k - 1$ and write

$$s_h \le \|f_1\|_\infty^2 \cdots \|f_{k-1}\|_\infty^2 \Big| \int f_k \cdot U_T^{kh} f_k d\mu \Big|.$$

We apply now Lemma 60 for the system $(X, \mathcal{B}, \mu, T^k)$, which is weak mixing by Lemma 61. We conclude that D-lim $s_h = 0$. Hence van der Corput's lemma applies and implies (7) in the special case $\int f_k d\mu = 0$.

In the general case we write $a := \int f_k d\mu$ and $f_k' := f_k - a$. We can write thus

$$\begin{aligned}\frac{1}{N}\sum_{n=0}^{N-1} U_T^n f_1 \cdot U_T^{2n} f_2 \cdots U_T^{kn} f_k \\ &= \frac{1}{N}\sum_{n=0}^{N-1} U_T^n f_1 \cdot U_T^{2n} f_2 \cdots U_T^{kn}(f_k' + a) \\ &= \frac{1}{N}\sum_{n=0}^{N-1} U_T^n f_1 \cdot U_T^{2n} f_2 \cdots U_T^{kn} f_k' \\ &\quad + a\frac{1}{N}\sum_{n=0}^{N-1} U_T^n f_1 \cdot U_T^{2n} f_2 \cdots U_T^{(k-1)n} f_{k-1}.\end{aligned}$$

The first sum converges to 0 by the special case, the second converges to

$$a\int f_1 d\mu \cdots \int f_{k-1} d\mu = \int f_1 d\mu \cdots \int f_k d\mu$$

by the induction hypothesis. This proves the theorem in the general case. □

10. Entropy

Entropy is a number attached to measure preserving systems that quantifies the following related intuitive properties.

- What is the information content of a typical orbit $x, Tx, T^2x, \ldots$?
- How chaotic is a typical orbit $x, Tx, T^2x, \ldots$?
- 'Knowing' $x, Tx, \ldots, T^n x$, to what extent can we 'predict' $T^{n+1}x$?

The original motivation for the introduction of entropy to ergodic theory was the following problem.

Definition 62. Let $(p_1, \ldots, p_n)$ be a probability vector, i.e. $p_i \geq 0$ and $p_1 + \ldots + p_n = 1$. The $(p_1, \ldots, p_n)$-Bernoulli shift is the measure preserving system $(\{1, \ldots, n\}^{\mathbf{Z}}, \mathcal{B}, \mu, \sigma)$, where $\mathcal{B}$ is the Borel σ-algebra, μ is the infinite-fold product of the measure $p_1\delta_1 + \ldots + p_n\delta_n$ on $\{1, \ldots, n\}$ and σ is the shift $(\sigma x)_n = x_{n+1}$.

Problem 63. Are the $(1/2, 1/2)$ and $(1/3, 1/3, 1/3)$ Bernoulli shifts isomorphic?

This problem may not seem very deep at first sight, but it was open for a decade or so until it was solved by Kolmogorov, and the following facts suggest that it is subtle.

- The two systems have the same mixing properties. (They are both mixing on k sets for all $k \in \mathbf{Z}_{>0}$.)

- They are also spectrally isomorphic. That is to say, there is a unitary transformation $U : L^2(\{1,2\}^{\mathbf{Z}}) \to L^2(\{1,2,3\}^{\mathbf{Z}})$ such that
$$UU_{\sigma_2} = U_{\sigma_3}U,$$
where σ_i is the shift map on $\{1,\dots,i\}^{\mathbf{Z}}$.
- Meshalkin proved that the $(1/4,1/4,1/4,1/4)$ and the $(1/2,1/8,1/8,1/8,1/8)$ Bernoulli shifts are isomorphic.

For more details on these facts, see the third example sheet.

In the next lectures we develop entropy theory and prove that the $(1/2,1/2)$ and $(1/3,1/3,1/3)$ Bernoulli shifts are not isomorphic.

10.1. **Jensen's inequality.** We recall an inequality that will be used frequently to derive elementary properties of entropy.

Definition 64. We say that a continuous function $f : [a,b] \to \mathbf{R}$ is **convex** if there is a number α_x for each $a < x < b$ such that
$$f(y) \geq f(x) + \alpha_x(y-x)$$
for all $y \in (a,b)$. The function is called strictly convex, if the inequality is strict for all $y \neq x$.

We note that if $f \in C^2$ and $f''(x) > 0$ for all $x \in (a,b)$, then f is strictly convex, as can be seen easily from the Taylor expansion of f.

Theorem 65 (Jensen's inequality)**.** *Let $f : [a,b] \to \mathbf{R}$ be a convex function. Let $x_1, x_2, \dots \in [a,b]$ and let $p_1, p_2, \dots$ be a (possibly infinite) probability vector. Then*
$$f(p_1x_1 + p_2x_2 + \dots) \leq p_1f(x_1) + p_2f(x_2) + \dots. \tag{10}$$
If the function is strictly convex, then equality occurs in (10) *if and only if the numbers x_i coincide for all i such that $p_i > 0$.*

10.2. **Entropy of partitions.** Let $(X, \mathcal{B}, \mu)$ be a probability space. A **countable measurable partition** is a countable collection of pairwise disjoint measurable sets, whose union covers the whole space X. We call the sets in this collection the **atoms** of the partition. If $\xi, \eta \subset \mathcal{B}$ are two finite partitions, then their **join** or **coarsest common refinement** is the finite partition
$$\xi \vee \eta = \{A \cap B : A \in \xi, B \in \eta\}.$$

We define the function H on the set of probability vectors (of possibly infinite length) by
$$H(p_1, p_2, \dots) = -\sum_j p_j \log p_j,$$
where we use the convention $p \log p = 0$ if $p = 0$.

We define the **entropy of a partition** $\xi = (A_1, A_2, \dots)$ as
$$H_\mu(\xi) = H(\mu(A_1), \mu(A_2), \dots).$$

We define the **conditional entropy** of ξ relative to another countable measurable partition η as

$$H_\mu(\xi|\eta) = \sum_{B\in\eta} \mu(B) H\Big(\frac{\mu(A_1\cap B)}{B}, \frac{\mu(A_2\cap B)}{\mu(B)}, \dots\Big).$$

One should think about conditional entropy as follows. The partition η subdivides the whole space as the union of probability spaces. Indeed, we can endow each atom B with a probability measure given by $\mu(C)/\mu(B)$ for each measurable set $C \subset B$. The partition ξ induces a partition on each of these spaces: $\{A_1\cap B, A_2\cap B, \dots\}$. Now the conditional entropy is the average of the entropies of these partitions on each of the probability spaces weighted by the measure of the corresponding atom B_j of η.

The following intuition for the meaning of entropy is useful. One may think about the space X as the possible states of a physical system and about the partition ξ as an experiment; the atoms are the possible outcomes. Then we may think of entropy as the amount of uncertainty in the experiment, i.e. how difficult it is to predict its outcome. Or equivalently, we may think of entropy as the amount of information we gain when we learn the outcome of the experiment.

With this interpretation, the two partitions are two experiments, and conditional entropy measures the amount of new information that we gain when we learn the outcome of the second experiment if we already knew the outcome of the first one.

The choice of the function

$$H(p_1, p_2, \dots) = \sum_j -p_j \log p_j \tag{11}$$

may seem arbitrary at first, but the following properties are reassuring.

Lemma 66. *The following holds.*

1. $H_\mu(\xi) \geq 0$ *for any partition* ξ.
2. *The maximum of* $H_\mu(\xi)$ *over all partitions with* k *atoms is obtained when each atoms have the same probability* $1/k$. *Interpretation: 'Uncertainity is maximal, when each outcome is equally likely.'*
3. $H_\mu(A_1, \dots, A_k) = H_\mu(A_{\sigma 1}, \dots, A_{\sigma k})$ *for any permutation* σ.
4. $H_\mu(\xi\vee\eta) = H_\mu(\xi) + H_\mu(\eta|\xi)$. *Interpretation: 'The amount of information contained together in the two experiments* ξ *and* η *is the same as the amount of information contained in* ξ *plus the amount of new information contained in* η *when we learn the outcome with the prior knowledge of the result of* ξ.'

We note that Khinchin proved that the function H given in (11) is unique upto scalar multiples such that the resulting notion of entropy enjoys the properties listed in the lemma.

Proof. (1): This holds because $-p\log p \geq 0$ for all $p \in [0,1]$.

(2): This follows from Jensen's inequality applied to the strictly convex function $x \mapsto x\log x$ with $p_i = 1/k$ and $x_i = \mu(A_i)$. Indeed, we note

$$\sum p_i x_i = \sum \frac{1}{k}\mu(A_i) = \frac{1}{k},$$

hence

$$-\frac{1}{k}\log k \leq \sum_i -\frac{1}{k}\mu(A_i)\log(\mu(A_i)),$$

which in turn gives

$$\log k \geq H(\xi).$$

Analysing the equality case in Jensen's inequality, we find that it occurs if and only if $\mu(A_1) = \ldots = \mu(A_k) = 1/k$.

(3): Immediate from the definition. □

We will prove item (4) below in greater generality

10.3. **The information function.** We introduce a notion, which will be a very useful gadget in our calculations with entropy. Let $(X, \mathcal{B}, \mu)$ be a probability space and let $\xi \subset \mathcal{B}$ be a countable partition. The **information function** of ξ is the function $X \mapsto \mathbf{R} \cup \{\infty\}$

$$I_\mu(\xi)(x) = -\log \mu([x]_\xi),$$

where $[x]_\xi$ denotes the atom of ξ that contains x. If $\eta \subset \mathcal{B}$ is another countable partition, then the **conditional information function** of ξ relative to η is defined as

$$I_\mu(\xi|\eta)(x) = -\log \frac{\mu([x]_{\xi\vee\eta})}{\mu([x]_\eta)}.$$

The link with entropy is given in the next lemma.

Lemma 67. *In the above setting, we have*

$$H_\mu(\xi) = \int I_\mu(\xi) d\mu,$$

$$H_\mu(\xi|\eta) = \int I_\mu(\xi|\eta) d\mu.$$

Proof. The first claim is a special case of the second one when we consider $\eta = \{X\}$, therefore, we only consider the second claim.

If $A \in \xi$ and $B \in \eta$, then $I_\mu(\xi|\eta)$ is constant on $A \cap B$ and its value is $-\log(\mu(A\cap B)/\mu(B))$, which is immediate from the definition. Thus

$$\begin{aligned}\int I_\mu(\xi|\eta) d\mu &= \sum_{A\in\xi, B\in\eta} -\mu(A\cap B)\log\frac{\mu(A\cap B)}{\mu(B)} \\ &= \sum_{B\in\eta}\mu(B)\sum_{A\in\xi} -\frac{\mu(A\cap B)}{\mu(B)}\log\frac{\mu(A\cap B)}{\mu(B)},\end{aligned}$$

as required. □

Lemma 68 (Chain rule)**.** *Let $\xi, \eta, \lambda \subset \mathcal{B}$ be countable partitions in a probability space $(X, \mathcal{B}, \mu)$. Then*

$$\begin{aligned} I_\mu(\xi \vee \eta|\lambda) =& I_\mu(\xi|\lambda) + I_\mu(\eta|\xi \vee \lambda) \\ H_\mu(\xi \vee \eta|\lambda) =& H_\mu(\xi|\lambda) + H_\mu(\eta|\xi \vee \lambda) \end{aligned}$$

Note that this lemma generalizes item (4) from the first lemma of the section.

Proof. The second line follows from the first one upon integration thanks to the previous lemma.

Unwinding the definitions, we can write

$$\begin{aligned} I_\mu(\xi|\lambda)(x) &= -\log \frac{\mu([x]_{\xi\vee\lambda})}{\mu([x]_\lambda)} \\ I_\mu(\eta|\xi \vee \lambda)(x) &= -\log \frac{\mu([x]_{\xi\vee\eta\vee\lambda})}{\mu([x]_{\xi\vee\lambda})} \\ I_\mu(\xi \vee \eta|\lambda)(x) &= -\log \frac{\mu([x]_{\xi\vee\eta\vee\lambda})}{\mu([x]_\lambda)}. \end{aligned}$$

It is clear that the sum of the first two lines yields the third line, as required. □

Lemma 69. *Let $\xi, \eta, \lambda \subset \mathcal{B}$ be countable partitions in a probability space $(X, \mathcal{B}, \mu)$. Then*

$$H_\mu(\xi|\eta \vee \lambda) \leq H_\mu(\xi|\lambda).$$

Interpretation: 'The new information content in the experiment ξ after we know the outcomes of η and λ is at most as much as its new information content if we know the outcome of λ alone.'

Proof. We note the definitions

$$H_\mu(\xi|\eta \vee \lambda) = \sum_{A\in\xi, B\in\eta, C\in\lambda} -\mu(A \cap B \cap C) \log \tfrac{\mu(A\cap B\cap C)}{\mu(B\cap C)} \tag{12}$$

$$H_\mu(\xi|\lambda) = \sum_{A\in\xi, C\in\lambda} -\mu(A \cap C) \log \tfrac{\mu(A\cap C)}{\mu(C)}. \tag{13}$$

Therefore, it is enough to show that

$$\sum_{B\in\eta} \mu(A \cap B \cap C) \log \frac{\mu(A \cap B \cap C)}{\mu(B \cap C)} \geq \mu(A \cap C) \log \frac{\mu(A \cap C)}{\mu(C)}$$

for each $A \in \xi$ and $C \in \lambda$.

In order to show this, we fix $A \in \xi$, $C \in \lambda$ and apply Jensen's inequality for the function $x \mapsto x \log x$ at the points $\mu(A \cap B \cap C)/\mu(B \cap C)$ with weights $\mu(B \cap C)/\mu(C)$ as B runs through the atoms of η. We note that

$$\sum_{B\in\eta} \frac{\mu(B \cap C)}{\mu(C)} \cdot \frac{\mu(A \cap B \cap C)}{\mu(B \cap C)} = \frac{\mu(A \cap C)}{\mu(C)}.$$

Hence Jensen's inequality yields

$$\sum_{B\in\eta} \frac{\mu(B\cap C)}{\mu(C)} \cdot \frac{\mu(A\cap B\cap C)}{\mu(B\cap C)} \log \frac{\mu(A\cap B\cap C)}{\mu(B\cap C)} \geq \frac{\mu(A\cap C)}{\mu(C)} \log \frac{\mu(A\cap C)}{\mu(C)},$$

which in turn yields

$$\sum_{B\in\eta} \mu(A\cap B\cap C) \log \frac{\mu(A\cap B\cap C)}{\mu(B\cap C)} \geq \mu(A\cap C) \log \frac{\mu(A\cap C)}{\mu(C)},$$

as required. □

The above inequality together with the chain rule can be used to prove a range of useful inequalities for entropies of partitions. For instance, we note the following.

Corollary 70. *Let $\xi, \eta \subset \mathcal{B}$ be countable partitions in a probability space $(X, \mathcal{B}, \mu)$. Then*

$$H_\mu(\eta) \leq H_\mu(\xi \vee \eta) \leq H_\mu(\xi) + H_\mu(\eta).$$

Proof. By the chain rule, we have

$$H_\mu(\xi \vee \eta) = H_\mu(\eta) + H_\mu(\xi|\eta).$$

Since $-p \log p \geq 0$ for all $p \in [0,1]$, $H_\mu(\xi|\eta) \geq 0$ similarly as we saw in item (1) of the first lemma of the section. This proves the inequality on the left.

By the previous lemma (applied with $\lambda = \{X\}$), we have $H_\mu(\xi|\eta) \leq H_\mu(\xi)$, which proves the inequality on the right. □

10.4. **Entropy of measure preserving systems.** In what follows $(X, \mathcal{B}, \mu, T)$ is a measure preserving system and $\xi, \eta \subset \mathcal{B}$ are countable partitions.

Notation. We write $T^{-1}\xi$ for the partition, whose atoms are $T^{-1}([x]_\xi)$.

Lemma 71 (Invariance)**.** *In the above setting, we have*

$$I_\mu(T^{-1}\xi|T^{-1}\eta)(x) = I_\mu(\xi|\eta)(Tx),$$
$$H_\mu(T^{-1}\xi|T^{-1}\eta) = H_\mu(\xi|\eta).$$

Proof. By the definition, we have

$$I_\mu(T^{-1}\xi|T^{-1}\eta)(x) = -\log \frac{\mu([x]_{T^{-1}\xi\vee T^{-1}\eta})}{\mu([x]_{T^{-1}\xi})}.$$

We note that

$$T^{-1}\xi \vee T^{-1}\eta = T^{-1}(\xi \vee \eta),$$

and

$$[x]_{T^{-1}(\xi\vee\eta)} = T^{-1}([Tx]_{\xi\vee\eta}).$$

Indeed, the set on the right hand side is an atom of $T^{-1}(\xi \vee \eta)$ and it contains x.

By the measure preserving property, we have

$$\mu([x]_{T^{-1}(\xi\vee\eta)}) = \mu([Tx]_{\xi\vee\eta}),$$

and similarly $\mu([x]_{T^{-1}(\xi)}) = \mu([Tx]_\xi)$, which proves the first claim.

The second claim follows from the first one by integration and the measure preserving property. □

Before we can give the definition of entropy of a measure preserving system, we need two more auxiliary results.

Lemma 72. *Let $a_1, a_2, \ldots$ be a subadditive sequence, i.e. $a_n + a_m \geq a_{n+m}$ for all n, m. Then*

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf \frac{a_n}{n},$$

in particular, the limit exists.

Proof. Fix $n_0 \in \mathbf{Z}_{>0}$. For each $n \in \mathbf{Z}_{>0}$, write $n = i(n)n_0 + j(n)$, where $0 \leq j(n) < n_0$. By (repeated use of) subadditivity,

$$a_n \leq i(n)a_{n_0} + a_{j(n)},$$

hence

$$\begin{aligned}\limsup_{n\to\infty} \frac{a_n}{n} &\leq \limsup_{n\to\infty} \frac{i(n)a_{n_0} + a_{j(n)}}{n} \\ &= \limsup_{n\to\infty} \frac{i(n)}{n} a_{n_0} + \limsup_{n\to\infty} \frac{a_{j(n)}}{n} \\ &= \limsup_{n\to\infty} \frac{n - j(n)}{n} \cdot \frac{a_{n_0}}{n_0} = \frac{a_{n_0}}{n_0}.\end{aligned}$$

Since n_0 is arbitrary, we can write

$$\inf \frac{a_n}{n} \geq \limsup_{n\to\infty} \frac{a_n}{n} \geq \liminf_{n\to\infty} \frac{a_n}{n} \geq \inf \frac{a_n}{n},$$

which proves the claim. □

Notation. We write

$$\xi_m^n = T^{-m}\xi \vee T^{-(m+1)}\xi \vee \ldots \vee T^{-n}\xi$$

for any integers $0 \leq m \leq n$. When the system is invertible, we allow m or n to be negative.

Lemma 73. *In the above setting, we have*

$$H_\mu(\xi_0^{n-1}) + H_\mu(\xi_0^{m-1}) \geq H_\mu(\xi_0^{n+m-1}),$$

i.e. the sequence $n \mapsto H_\mu(\xi_0^{n-1})$ is subadditive.

Proof. We note that $\xi_0^{n+m-1} = \xi_0^{n-1} \vee \xi_n^{n+m-1}$ and $T^{-n}\xi_0^{m-1} = \xi_n^{n+m-1}$. By the above corollary and invariance, we can write

$$H_\mu(\xi_0^{n+m-1}) \leq H_\mu(\xi_0^{n-1}) + H_\mu(\xi_n^{n+m-1}) = H_\mu(\xi_0^{n-1}) + H_\mu(\xi_0^{m-1}),$$

as required. □

Definition 74. Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system, and let $\xi \subset \mathcal{B}$ be a countable partition. The **entropy** of the system **with respect to the partition** ξ is

$$h_\mu(T, \xi) = \lim_{n\to\infty} \frac{H_\mu(\xi_0^{n-1})}{n} = \inf \frac{H_\mu(\xi_0^{n-1})}{n}.$$

The **entropy** of the system is

$$h_\mu(T) = \sup_{\xi: H(\xi)<\infty} h_\mu(T, \xi).$$

The following intuition is useful to bear in mind. If we think about the measure preserving system as a physical system, the transformation T is the progression of time and ξ is an experiment, then the partition ξ_0^{n-1} corresponds to the experiment ξ performed repeatedly at time $0, 1, \ldots, n-1$. Therefore, the entropy of the system with respect to ξ is the amount of information gained per one experiment when we perform it repeatedly at regular time intervals. The entropy of the system is the maximal amount of information we can gain per one experiment optimized over all possible experiments we can perform.

10.5. **The Kolmogorov-Sinai theorem and the entropy of Bernoulli shifts.** One of the difficulties in calculating the entropy of a measure preserving system is that the definition involves a supremum over all partitions. The Kolmogorov-Sinai theorem addresses this issue.

Definition 75. Let $(X, \mathcal{B}, \mu, T)$ be an invertible measure preserving system. For a countable partition $\xi \subset \mathcal{B}$, we denote by $\mathcal{B}(\xi)$ the σ-algebra generated by the atoms of ξ. A countable partition $\xi \subset \mathcal{B}$ is called a **two-sided generator**, if for every $A \in \mathcal{B}$ and for every $\varepsilon > 0$, there are a number $n \in \mathbf{Z}_{>0}$ and a set $A' \in \mathcal{B}(\xi_{-n}^n)$ such that $\mu(A \triangle A') < \varepsilon$.

Example 76. Let $(X = \{1, \ldots, k\}^{\mathbf{Z}}, \mathcal{B}, \mu, \sigma)$ be the $(p_1, \ldots, p_k)$- Bernoulli shift. Then the partition

$$\xi = \{\{x \in X : x_0 = i\} : i = 1, \ldots, k\}$$

is a two-sided generator.

Indeed, it is easy to check that the collection of sets

$$\{A \in \mathcal{B} : \text{ for every } \varepsilon > 0, \text{ there are } n \text{ and } A' \in \mathcal{B}(\xi_{-n}^n) \text{ with } \mu(A \triangle A') < \varepsilon\}$$

is a σ-algebra and it contains all cylinder sets, hence it equals $\mathcal{B}$.

Theorem 77 (Kolmogorov-Sinai)**.** *Let $(X, \mathcal{B}, \mu, T)$ be an invertible measure preserving system and let $\xi \subset \mathcal{B}$ be a countable partition that is a two-sided generator. Then*

$$h_\mu(T) = h_\mu(T, \xi).$$

There is a version of this theorem with a similar proof for non-invertible measure preserving systems, where a two-sided generator is replaced by the analogous notion of a one-sided generator. However, we will not need that result, therefore we omit the details.

Example 78. Let $(X = \{1, \dots, k\}^{\mathbf{Z}}, \mathcal{B}, \mu, \sigma)$ be the $(p_1, \dots, p_k)$- Bernoulli shift, and let ξ be the partition we considered in the previous example.

We prove that $h_\mu(T, \xi) = H(p_1, \dots, p_k)$, which implies $h_\mu(T) = H(p_1, \dots, p_k)$ by the Kolmogorov-Sinai theorem.

We first show that

$$H_\mu(\xi|\xi_1^n) = H(p_1, \dots, p_k) \tag{14}$$

for all $n \in \mathbf{Z}_{>0}$. To this end we calculate the information function

$$I_\mu(\xi|\xi_1^n)(x) = -\log \frac{\mu([x]_{\xi_0^n})}{\mu([x]_{\xi_1^n})}.$$

We note that

$$[x]_{\xi_0^n} = \{y : y_0 = x_0, \dots, y_n = x_n\},$$

which has measure $p_{x_0} \cdots p_{x_n}$ by the definition of the product measure. Similarly,

$$\mu([x]_{\xi_1^n}) = p_{x_1} \cdots p_{x_n}.$$

Thus

$$I_\mu(\xi|\xi_1^n)(x) = -\log p_{x_0},$$

and (14) follows by integration.

Using the chain rule, invariance and then (14), we obtain

$$\begin{aligned} H_\mu(\xi_0^{n-1}) =& H_\mu(\xi_{n-1}^{n-1}) + H_\mu(\xi_{n-2}^{n-2}|\xi_{n-1}^{n-1}) + \ldots + H_\mu(\xi_0^0|\xi_1^{n-1}) \\ =& H_\mu(\xi) + H_\mu(\xi|\xi_1^1) + \ldots + H_\mu(\xi|\xi_1^{n-1}) \\ =& nH(p_1, \dots, p_k). \end{aligned}$$

We divide both sides by n and take the limit to conclude

$$h_\mu(T, \xi) = H(p_1, \dots, p_k),$$

as required.

Since isomorphic systems have the same entropy, this shows that the $(1/2, 1/2)$ and $(1/3, 1/3, 1/3)$ Bernoulli shifts are not isomorphic. We add that a deep theorem of Ornstein shows that two Bernoulli shifts are isomorphic if and only if they have the same entropy.

The proof of the Kolmogorov-Sinai theorem relies on the following three Lemmata.

Lemma 79. *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and let $\xi \subset \mathcal{B}$ be a countable measurable partition. Then*

$$h_\mu(T, \xi) = h_\mu(T, \xi_{-n}^n)$$

for all $n \in \mathbf{Z}_{>0}$.

Lemma 80. *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and let $\xi, \eta \subset \mathcal{B}$ be countable measurable partitions. Then*

$$h_\mu(T, \eta) \leq h_\mu(T, \xi) + H_\mu(\eta|\xi).$$

Lemma 81. *For every $\varepsilon > 0$ and for every $k \in \mathbf{Z}_{>0}$, there is $\delta > 0$ such that the following holds. Let $(X, \mathcal{B}, \mu)$ be a probability space, and let $\xi \subset \mathcal{B}$ be a countable and let $\eta \subset \mathcal{B}$ be a finite partition. Suppose that η has k atoms and for each $A \in \eta$ there is $A' \in \mathcal{B}(\xi)$ such that $\mu(A \triangle B) \leq \delta$.*

Then $H_\mu(\eta|\xi) \leq \varepsilon$.

Proof of Lemma 84. We can write

$$h_\mu(T, \xi) = \lim_{m\to\infty} \frac{H_\mu(\xi_0^{m-1})}{m}$$

and

$$\begin{aligned} h_\mu(T, \xi_{-n}^n) &= \lim_{m\to\infty} \frac{H_\mu(\xi_{-n}^n \vee T^{-1}\xi_{-n}^n \vee \ldots \vee T^{-(m-1)}\xi_{-n}^n)}{m} \\ &= \lim_{m\to\infty} \frac{H_\mu(\xi_{-n}^{m+n-1})}{m} \\ &= \lim_{m\to\infty} \frac{H_\mu(\xi_0^{m+2n-1})}{m} \\ &= \lim_{m\to\infty} \frac{H_\mu(\xi_0^{m+2n-1})}{m+2n-1}. \end{aligned}$$

Hence the entropies with respect to the two partitions are equal, indeed. □

Proof of Lemma 85. We can write

$$\begin{aligned} H_\mu(\eta_0^{m-1}) \leq & H_\mu(\eta_0^{m-1} \vee \xi_0^{m-1}) \\ = & H_\mu(\xi_0^{m-1}) + H_\mu(\eta_0^{m-1}|\xi_0^{m-1}) \\ = & H_\mu(\xi_0^{m-1}) + \sum_{j=0}^{m-1} H_\mu(\eta_j^j|\xi_0^{m-1} \vee \eta_{j+1}^{m-1}) \\ \leq & H_\mu(\xi_0^{m-1}) + \sum_{j=0}^{m-1} H_\mu(\eta_j^j|\xi_j^j) \\ = & H_\mu(\xi_0^{m-1}) + m \cdot H_\mu(\eta|\xi). \end{aligned}$$

Here, we used the Corollary of Lemma 74, then the chain rule twice, then Lemma 74, and finally invariance.

We divide this inequality by m and take the limit to obtain the claim. □

Proof of Lemma 86. Let $\eta = \{A_1, \ldots, A_k\}$, and for each $1 \leq i \leq k$, let $B_i \in \mathcal{B}(\xi)$ be such that $\mu(A_i \triangle B_i) < \delta$. We consider the partition λ

that has the following $k+1$ atoms:

$$C_0 := \bigcup \Big(A_i \cap B_i \backslash \bigcup_{j \neq i} B_j \Big)$$

and $C_i := A_i \backslash C_0$.

The set C_0 has the following two key properties. First, it has large measure, i.e.

$$\mu(C_0) \geq \sum_{i=1}^{k} (\mu(A_i) - k\delta) = 1 - k^2\delta.$$

Second, for $x \in C_0$, we have $x \in B_i$ if and only if $x \in A_i$.

We note that

$$\begin{aligned} H_\mu(\lambda) &= -\mu(C_0)\log\mu(C_0) - \sum_{i=1}^{k} \mu(C_i)\log\mu(C_i) \\ &\leq -\mu(C_0)\log\mu(C_0) - \sum_{i=1}^{k} \mu(C_i)\log\frac{\sum_{i=1}^{k}\mu(C_i)}{k}. \end{aligned}$$

Indeed, this follows from Jensen's inequality applied for the function $x \mapsto x\log x$ at the points $\mu(C_i)$ for $i = 1, \ldots, k$ with weights $1/k$ similarly to the proof of (1) in Lemma 71. If δ is sufficiently small, then $\mu(C_0)$ is arbitrarily close to 1 and $\sum_{i=1}^{k}\mu(C_i)$ is arbitrarily close to 0. Hence $H_\mu(\lambda) < \varepsilon$, provided δ is small enough.

We prove that $H_\mu(\eta|\xi \vee \lambda) = 0$, and hence

$$H_\mu(\eta|\xi) \leq H_\mu(\eta \vee \lambda|\xi) = H_\mu(\lambda|\xi) + H_\mu(\eta|\xi \vee \lambda) < \varepsilon,$$

as required.

To that end, we prove $I_\mu(\eta|\xi \vee \lambda)(x) = 0$ and consider two cases. If $x \in C_0$, then $[x]_{\xi\vee\lambda} \subset C_0$, because $C_0 \in \mathcal{B}(\xi \vee \lambda)$. Moreover, there is a unique $1 \leq i \leq k$ such that $x \in A_i$, hence $x \in B_i$, because $C_0 \cap B_i = C_0 \cap A_i$. Then $[x]_{\xi\vee\lambda} \subset B_i$, because $B_i \in \mathcal{B}(\xi \vee \lambda)$. Therefore

$$[x]_{\xi\vee\lambda} \subset A_i = [x]_\eta$$

and $I_\mu(\eta|\xi \vee \lambda)(x) = 0$ in this case.

The second case is when $x \in C_i$ for some $i > 1$. Then $[x]_{\xi\vee\lambda} \subset C_i$, because $C_i \in \mathcal{B}(\xi \vee \lambda)$. Since $C_i \subset A_i$, we have

$$[x]_{\xi\vee\lambda} \subset A_i = [x]_\eta.$$

Again, we find that $I_\mu(\eta|\xi \vee \lambda)(x) = 0$ in this case. This completes the proof. □

Proof of the Kolmogorov Sinai theorem. We first show that

$$\sup_{\eta:\eta \text{ is finite}} h_\mu(T,\eta) = \sup_{\xi:H_\mu(\eta)<\infty} h_\mu(T,\eta).$$

Let $\eta \subset \mathcal{B}$ be a countable partition with $H_\mu(\eta) < \infty$, and let $\varepsilon > 0$. We show that there is a finite partition $\widetilde{\eta} \subset \mathcal{B}$ such that $H_\mu(\eta|\widetilde{\eta}) < \varepsilon$. Then Lemma 85 implies

$$h_\mu(T, \eta) \leq h_\mu(T, \widetilde{\eta}) + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrarily small, this proves the claim.

Write $A_1, A_2, \ldots$ for the atoms of η and put $\widetilde{\eta} = \{A_1, \ldots, A_k, \bigcup_{j>k} A_j\}$, where k is sufficiently large, to be chosen later. Since $\eta = \eta \vee \widetilde{\eta}$,

$$H_\mu(\eta) = H_\mu(\eta \vee \widetilde{\eta}) = H_\mu(\widetilde{\eta}) + H_\mu(\eta|\widetilde{\eta}),$$

hence

$$H_\mu(\eta|\widetilde{\eta}) = H_\mu(\eta) - H_\mu(\widetilde{\eta}) \leq \sum_{j>k} -\mu(A_j)\log(\mu(A_j)).$$

The right hand side is arbitrarily small if k is sufficiently large, since the series $\sum -\mu(A_j)\log(\mu(A_j))$ is convergent.

We now need to show that $h_\mu(T|\eta) \leq h_\mu(T|\xi)$ for any finite partition $\eta \subset \mathcal{B}$.

Fix $\varepsilon > 0$. Since ξ is a two-sided generator, we have $H_\mu(\eta|\xi_{-n}^n) < \varepsilon$ for n large enough by Lemma 86. By Lemma 85, we have

$$h_\mu(T, \eta) \leq h_\mu(T, \xi_{-n}^n) + H_\mu(\eta|\xi) \leq h_\mu(T, \xi_{-n}^n) + \varepsilon.$$

By Lemma 84, $h_\mu(T, \xi_{-n}^n) = h_\mu(T, \xi)$, hence $h_\mu(T|\eta) \leq h_\mu(T|\xi) + \varepsilon$. This is sufficient, because ε is arbitrarily small. □

11. The Shannon McMillan Breiman theorem

Theorem 82. *Let $(X, \mathcal{B}, \mu, T)$ be an ergodic measure preserving system and let $\xi \subset \mathcal{B}$ be a countable partition with $H_\mu(\xi) < \infty$. Then*

$$\frac{1}{N} I_\mu(\xi_0^{N-1}) \to h_\mu(T, \xi) \text{ as } N \to \infty$$

pointwise μ-almost everywhere and in $L^1(X, \mu)$.

Recall the definition of the information function

$$I_\mu(\xi_0^{N-1})(x) = -\log([x]_{\xi_0^{N-1}}).$$

Therefore, the theorem states that the atom of a typical point in the partition ξ_0^{N-1} is approximately $\exp(-Nh_\mu(T, \xi))$.

The Shannon McMillan Breiman theorem is also useful in calculating the entropy of measure preserving systems. See the last example sheet for some examples.

11.1. **Conditional expectation.** The proof of the Shannon McMillan Breiman theorem is based on the following idea. Using the chain rule and invariance, we can write

$$I_\mu(\xi_0^{N-1})(x) = \sum_{n=0}^{N-1} I_\mu(\xi_n^n|\xi_{n+1}^{N-1})(x) = \sum_{n=0}^{N-1} I_\mu(\xi|\xi_1^{N-1-n})(T^n x).$$

The sum on the right hand side looks like an ergodic sum and it would be tempting to apply the pointwise ergodic theorem to conclude the proof. However, in this sum, at each point a different function is evaluated and the ergodic theorem applies only if we evaluate the same function. The idea to overcome this issue is to generalize the notion of conditional information functions to σ-algebras in place of partitions and show that the functions $I_\mu(\xi|\xi_1^{N-1-n})$ converges to such a general conditional information function. Then we can replace the functions in the sum by this limit and obtain a genuine ergodic sum.

This generalization requires some preparation. We begin by recalling the definition of conditional expectation and its basic properties.

Theorem 83. *Let $(X, \mathcal{B}, \mu)$ be a probability space and let $\mathcal{A} \subset \mathcal{B}$ be a σ-algebra. Then for every $f \in L^1(X, \mu)$, there is $f^* \in L^1(X, \mu)$ such that the following holds.*

(1) *f^* is $\mathcal{A}$-measurable,*
(2) *$\int_A f d\mu = \int_A f^* d\mu$ for all $A \in \mathcal{A}$.*

If $f_1^*, f_2^* \in L^1(X, \mu)$ are two functions that satisfy both items (1) *and* (2) *in the role of f^*, then $f_1^* = f_2^*$ holds μ-almost everywhere.*

Definition 84. The function f^* in the above theorem, which is defined up to a set of μ-measure 0 is called the conditional expectation of f with respect to $\mathcal{A}$ and is denoted by $\mathbf{E}(f|\mathcal{A})$.

We note that

$$V_\mathcal{A} = \{f \in L^2(X, \mu) : \ f \text{ is } \mathcal{A} \text{ measurable}\}$$

is a closed subspace of $L^2(X, \mu)$ and for $f \in L^2(X, \mu)$, $\mathbf{E}(f|\mathcal{A})$ is the orthogonal projection of f to $V_\mathcal{A}$.

Example 85. Suppose that $\mathcal{A}$ is generated by a countable partition ξ. Then a function is $\mathcal{A}$-measurable if and only if it is constant on the atoms of ξ, in particular, $\mathbf{E}(f|\mathcal{A})$ is constant on the atoms of ξ. Let A be an atom of ξ. Then for μ-almost every $x \in A$, we have

$$\mathbf{E}(f|\mathcal{A})(x) = \mu(A)^{-1} \int_A \mathbf{E}(f|\mathcal{A}) d\mu = \mu(A)^{-1} \int_A f d\mu.$$

The following theorem summarizes the basic properties of conditional expectation.

Theorem 86. *Let $(X, \mathcal{B}, \mu)$ be a probability space, and let $\mathcal{A}, \mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{B}$ be σ-algebras, and let $f, f_1, f_2 \in L^1(X, \mu)$. Then*

(1) $\mathbf{E}(f_1 + f_2|\mathcal{A}) = \mathbf{E}(f_1|\mathcal{A}) + \mathbf{E}(f_2|\mathcal{A})$,
(2) $\mathbf{E}(f_1 f_2|\mathcal{A}) = f_1\mathbf{E}(f_2|\mathcal{A})$ *if* f_1 *is* $\mathcal{A}$*-measurable,*
(3) $\mathbf{E}(f|\mathcal{A}_2) = \mathbf{E}(\mathbf{E}(f|\mathcal{A}_1)|\mathcal{A}_2)$ *if* $\mathcal{A}_2 \subset \mathcal{A}_1$.

We will also need the following result.

Theorem 87 (Increasing and decreasing martingale theorems)**.** *Let* $(X, \mathcal{B}, \mu)$ *be a probability space and let* $\mathcal{A}_1, \mathcal{A}_2, \ldots \subset \mathcal{B}$ *and* $\mathcal{A} \subset \mathcal{B}$ *be* σ*-algebras. Suppose that either*

- $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \ldots$ *and* $\mathcal{A}$ *is the* σ*-algebra generated by* $\bigcup \mathcal{A}_i$ *or*
- $\mathcal{A}_1 \supset \mathcal{A}_2 \supset \ldots$ *and* $\mathcal{A} = \bigcap \mathcal{A}_i$.

Let $f \in L^1(X, \mu)$.

Then

$$\lim_{n\to\infty} \mathbf{E}(f|\mathcal{A}_i) = \mathbf{E}(f|\mathcal{A})$$

pointwise μ*-almost everywhere and in* $L^1(X, \mu)$.

We omit the proof, which can be found in standard textbooks. It is a useful exercise to prove this theorem for L^2-convergence (in the case $f \in L^2$), however, understanding the proof of this result is not required in what follows.

As we have seen above, conditional expectation is particularly easy to compute and work with when the σ-algebra is generated by a finite (or countable) partition. In many situations, a more complicated σ-algebra $\mathcal{A}$ can be approximated by ones generated by finite partitions in the sense of the martingale theorem, i.e. there is a sequence of finer and finer finite partitions whose union generate $\mathcal{A}$. In that case the conditional expectation with respect to $\mathcal{A}$ can be approximated by conditional expectations with respect to finite partitions thanks to the martingale theorem.

11.2. **Conditional entropy with respect to σ-algebras.**

Definition 88. Let $(X, \mathcal{B}, \mu)$ be a probability space, let $\xi \subset \mathcal{B}$ be a countable partition with $H_\mu(\xi) < \infty$ and let $\mathcal{A} \subset \mathcal{B}$ be a σ-algebra. We define

$$I_\mu(\xi|\mathcal{A}) = \sum_{A\in\xi} -\chi_A \cdot \log \mathbf{E}(\chi_A|\mathcal{A}),$$

where χ_A is the characteristic function of the set A. We also set

$$H_\mu(\xi|\mathcal{A}) = \int I_\mu(\xi|\mathcal{A}) d\mu.$$

This definition may look arbitrary. To illustrate that this is the right one, we first verify that it coincide with the original definition, when $\mathcal{A} = \mathcal{B}(\eta)$ for a countable partition $\eta \subset \mathcal{B}$.

Example 89. Let $\eta \subset \mathcal{B}$ be a countable partition and let $\mathcal{A}$ be the σ-algebra it generates. Then

$$\begin{aligned} I_\mu(\xi|\mathcal{A})(x) &= \sum_{A\in\xi} -\chi_A(x) \log \mathbf{E}(\chi_A|\mathcal{A})(x) \\ &= -\log \mathbf{E}(\chi_{[x]_\xi}|\mathcal{A})(x) = -\log \frac{\int_{[x]_\eta} \chi_{[x]_\xi} d\mu}{\mu([x]_\eta)} \\ &= -\log \frac{\mu([x]_\xi \cap [x]_\eta)}{\mu([x]_\eta)} = I_\mu(\xi|\eta). \end{aligned}$$

As a second reassurance, we prove that the conditional information function with respect to finer and finer σ-algebras converge to the conditional information function with respect to the σ-algebra generated by them. This will also be important in the proof of the Shannon-McMillan-Breiman theorem.

Proposition 90. *Let $(X, \mathcal{B}, \mu)$ be a probability space and let $\mathcal{A}_1, \mathcal{A}_2, \ldots \subset \mathcal{B}$ be a sequence of countable partitions such that $\mathcal{A}_i \subset \mathcal{A}_{i+1}$ for all i. Let $\mathcal{A} = \mathcal{B}(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots)$. Let $\eta \in \mathcal{B}$ be a countable partition with finite entropy. Then*

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

in L^1 and pointwise μ-almost everywhere.

Moreover,

$$I^*(x) := \sup_{n\in\mathbf{Z}_{>0}} I_\mu(\eta|\mathcal{A}_n)(x) \in L^1(X,\mu). \tag{15}$$

Proof. We only prove the proposition in the special case, when η is a finite partition to avoid some technicalities. The general case is dealt with in an example sheet.

First we prove the pointwise convergence. This follows easily form the martingale theorem and the definition of the conditional information function. Indeed,

$$\begin{aligned} I_\mu(\eta|\mathcal{A})(x) &= -\log \mathbf{E}(\chi_{[x]_\eta}|\mathcal{A})(x) \\ &= \lim_{n\to\infty} -\log \mathbf{E}(\chi_{[x]_\eta}|\mathcal{A}_n)(x) \\ &= \lim_{n\to\infty} I_\mu(\eta|\mathcal{A}_n)(x) \end{aligned}$$

holds almost everywhere.

Next, we prove (15). This together with pointwise convergence imply the L^1 convergence by the dominated convergence theorem.

Fix $\alpha \in \mathbf{R}_{>0}$ and $A \in \eta$. For each $x \in X$, we define $n(x)$ as the smallest n such that

$$-\log \mathbf{E}(\chi_A|\mathcal{A}_n)(x) \geq \alpha.$$

We put $n(x) = \infty$ if there is no such n.

For each n, the set

$$\begin{aligned} B_n :=& \{x \in X : n(x) = n\} \\ =& \{x \in X : \mathbf{E}(\chi_A|\mathcal{A}_n)(x) \leq \exp(-\alpha) \text{ and} \\ & \qquad \mathbf{E}(\chi_A|\mathcal{A}_j)(x) > \exp(-\alpha) \text{ for every } j < n\} \end{aligned}$$

is $\mathcal{A}_n$-measurable. Therefore,

$$\mu(B_n)\exp(-\alpha) \geq \int_{B_n} \mathbf{E}(\chi_A|\mathcal{A}_n)(x) = \int_{B_n} \chi_A d\mu = \mu(A \cap B_n).$$

Note that the sets B_n are pairwise disjoint and they cover the set

$$A^* := \{x \in A : I^*(x) \geq \alpha\} = \{x \in A : \inf \mathbf{E}(\chi_A|\mathcal{A}_n) \leq \exp(-\alpha)\}.$$

Thus

$$\mu(A^*) = \sum_{n \in \mathbf{Z}_{>0}} \mu(A \cap B_n) \leq \sum_{n \in \mathbf{Z}_{>0}} \exp(-\alpha)\mu(B_n) \leq \exp(-\alpha).$$

Now

$$\mu(x \in X : I^*(x) \geq \alpha) \leq \sum_{A \in \eta} \mu(A^*) \leq |\eta| \exp(-\alpha).$$

Thus

$$\int I^* d\mu \leq \sum_{n \in \mathbf{Z}_{>0}} n\mu(x : n-1 \leq I^*(x) \leq n) \leq |\eta| \sum_{n \in \mathbf{Z}_{>0}} n \cdot \exp(-(n-1)) < \infty,$$

as claimed. □

11.3. **Proof of the theorem.**

Lemma 91. *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and let $\xi \subset \mathcal{B}$ be a countable partition with $H_\mu(\xi) < \infty$. Then*

$$\lim_{n \to \infty} H_\mu(\xi|\xi_1^n) = h_\mu(T, \xi).$$

Proof. Similarly to the calculations we did when we computed the entropy of Bernoulli shifts, we can write using the chain rule and invariance

$$H_\mu(\xi_0^{n-1}) = H_\mu(\xi|\xi_1^{n-1}) + \ldots + H_\mu(\xi|\xi_1^1) + H_\mu(\xi).$$

Therefore

$$h_\mu(T, \xi) = \text{C-lim}_{n \to \infty} H_\mu(\xi|\xi_1^{n-1}).$$

Since $n \to H_\mu(\xi|\xi_1^{n-1})$ is a monotone non-increasing sequence, it converges and its limit must equal to its Cesàro limit, which was to be proved. □

Proof of the Shannon-McMillan-Breiman theorem. Recall from the beginning of the section that

$$\frac{1}{N} I_\mu(\xi_0^{N-1})(x) = \frac{1}{N} \sum_{n=0}^{N-1} I_\mu(\xi|\xi_1^{N-1-n})(T^n x).$$

We can thus write

(16)

$$\frac{1}{N} I_\mu(\xi_0^{N-1})(x) = \frac{1}{N}\sum_{n=0}^{N-1} I_\mu(\xi|\mathcal{B}(\xi_1^\infty))(T^n x) + \frac{1}{N}\sum_{n=0}^{N-1}(I_\mu(\xi|\xi_1^{N-n-1})(T^n x) - I_\mu(\xi|\mathcal{B}(\xi_1^\infty))(T^n x)),$$

where $\mathcal{B}(\xi_1^\infty)$ denotes the smallest σ-algebra that contains ξ_1^N for all N. By the pointwise ergodic theorem, we have

$$\frac{1}{N}\sum_{n=0}^{N-1} I_\mu(\xi|\mathcal{B}(\xi_1^\infty))(T^n x) \to \int I_\mu(\xi|\mathcal{B}(\xi_1^\infty))d\mu.$$

We prove that $\int I_\mu(\xi|\mathcal{B}(\xi_1^\infty))d\mu = h_\mu(T,\xi)$. We have

$$\int I_\mu(\xi|\mathcal{B}(\xi_1^\infty))d\mu = \lim_{N\to\infty}\int I_\mu(\xi|\xi_1^N)d\mu = \lim_{N\to\infty} H_\mu(\xi|\xi_1^N) = h_\mu(T,\xi).$$

Therefore, it is enough to show that the second term in (16) converges to 0. To that end, we write

$$I_K^*(x) = \sup_{k\geq K} |I_\mu(\xi|\mathcal{B}(\xi_1^\infty))(x) - I_\mu(\xi|\xi_1^k)(x)|.$$

By the previous proposition, we have $I_K^* \in L^1(X,\mu)$ for all K, and $I_K^* \to 0$ in $L^1(X,\mu)$ and pointwise μ-almost everywhere. We fix an integer K and write

$$\Big|\frac{1}{N}\sum_{n=0}^{N-1}(I_\mu(\xi|\xi_1^{N-n-1})(T^n x) - I_\mu(\xi|\mathcal{B}(\xi_1^\infty))(T^n x))\Big| \leq \frac{1}{N}\sum_{n=0}^{N-K-1} I_K^*(T^n x) + \frac{1}{N}\sum_{n=N-K}^{N-1} I_0^*(T^n x).$$

We estimate these two sums individually. First we write

$$\frac{1}{N}\sum_{n=0}^{N-K-1} I_K^*(T^n x) \leq \frac{1}{N}\sum_{n=0}^{N} I_K^*(T^n x) \to \int I_K^* d\mu$$

as $N\to\infty$. For the second sum, we write

$$\frac{1}{N}\sum_{n=N-K}^{N-1} I_0^*(T^n x) = \frac{1}{N}\sum_{n=0}^{N-1} I_0^*(T^n x) - \frac{N-K-1}{N}\cdot\frac{1}{N-K-1}\sum_{n=0}^{N-K-1} I_0^*(T^n x) \to \int I_0^* d\mu - \int I_0^* d\mu = 0,$$

as $N\to\infty$.

We obtained that

$$\limsup_{N\to\infty}\left|\frac{1}{N}\sum_{n=0}^{N-1}(I_\mu(\xi|\xi_1^{N-n-1})(T^nx)-I_\mu(\xi|\mathcal{B}(\xi_1^\infty))(T^nx))\right|\le\int I_K^*d\mu.$$

Since K was arbitrary, and

$$\lim_{K\to\infty}\int I_K^*d\mu=0,$$

this proves the theorem. □

12. Mixing and entropy

Earlier in the course we discussed weak mixing as a more convenient alternative for mixing. Now we discuss a similarly convenient notion that is stronger than mixing.

Definition 92. A measure preserving system $(X,\mathcal{B},\mu,T)$ is called K-mixing (K for Kolmogorov) if the following holds. Let $A\in\mathcal{B}$ and let $\xi\subset\mathcal{B}$ be a finite partition. Then for every $\varepsilon>0$, there is $N\in\mathbf{Z}_{>0}$ such that for every $B\in\mathcal{B}(\xi_N^\infty)$, we have

$$|\mu(A\cap B)-\mu(A)\mu(B)|<\varepsilon.$$

Recall that $\mathcal{B}(\xi_N^\infty)$ denotes the σ-algebra generated by the partitions ξ_N^M for $M\in\mathbf{Z}_{\ge N}$.

We note that we can take $\xi=\{C,X\backslash C\}$ in the definition for an arbitrary fixed set $C\in\mathcal{B}$. This allows us to take $B=T^{-n}(C)$ for $n\ge N$ and obtain

$$|\mu(A\cap T^{-n}C)-\mu(A)\mu(C)|<\varepsilon.$$

This shows that K-mixing indeed implies mixing. Moreover, it also implies mixing on k-sets for any $k\in\mathbf{Z}_{\ge2}$, which is the content of an exercise on the last example sheet.

Some authors refer to invertible K-mixing systems as K-automorphisms.

The notion of K-mixing can be characterized using entropy or tail σ-algebras. We first define the latter.

Definition 93. Let $(X,\mathcal{B},\mu,T)$ be a measure preserving system and let $\xi\subset\mathcal{B}$ be a finite partition. The **tail σ-algebra** of ξ is

$$\mathcal{T}(\xi)=\bigcap_{n=0}^{\infty}\mathcal{B}(\xi_n^\infty).$$

In our usual interpretation, we think about the σ-algebra $\mathcal{B}(\xi_n^\infty)$ as the knowledge of the result of an experiment for time $n,n+1,\ldots$. In this manner we think about $\mathcal{T}(\xi)$ as the knowledge of the result of the experiment in the distant future.

Theorem 94. *Let $(X,\mathcal{B},\mu,T)$ be a measure preserving system. The following are equivalent.*

(1) *The system is K-mixing.*
(2) *For each finite partition* $\xi \subset \mathcal{B}$, $\mathcal{T}(\xi)$ *is trivial, that is it contains only sets of measure* 0 *and* 1.
(3) *The system is of totally positive entropy, that is, we have* $h_\mu(T, \xi) > 0$ *for any finite partition* $\xi \subset \mathcal{B}$ *with* $H_\mu(\xi) > 0$.

We note that the Kolmogorov $0-1$ law states that $\mathcal{T}(\xi)$ is trivial if the system is a Bernoulli shift and $\xi = \{\{x : x_0 = i\} : i = 1, \dots, k\}$. In fact, this is true for any finite partitions $\xi \subset \mathcal{B}$, hence Bernoulli shifts are examples of K-mixing systems.

We begin with two easier implications in the theorem.

Proof of (1) $\Rightarrow$ (2). Suppose the system is K-mixing. Fix a finite partition ξ, and let $A \in \mathcal{T}(\xi)$. Then $A \in \mathcal{B}(\xi_N^\infty)$ for all N, hence

$$|\mu(A \cap A) - \mu(A)^2| < \varepsilon$$

for any $\varepsilon > 0$. Then $\mu(A) = \mu(A)^2$, hence $\mu(A) = 0$ or 1, as required. □

Proof of (2) $\Rightarrow$ (1). Fix $A \in \mathcal{B}$ and fix a finite partition $\xi \subset \mathcal{B}$.

Suppose $\mathcal{T}(\xi)$ is trivial. We first show that $\mathbf{E}(\chi_A|\mathcal{T}(\xi)) = \mu(A)$ almost everywhere. If this fails, then

$$B := \{x : \mathbf{E}(\chi_A|\mathcal{T}(\xi)) \geq \mu(A) + \varepsilon\}$$

is of positive measure for some $\varepsilon > 0$. Since $B \in \mathcal{T}(\xi)$, $\mu(B) = 1$, and we have

$$\mu(A) = \int_B \chi_A d\mu = \int_B \mathbf{E}(\chi_A|\mathcal{T}(\xi))d\mu \geq \mu(A) + \varepsilon,$$

a contradiction.

Fix $\varepsilon > 0$. By the martingale convergence theorem, we have

$$\|\mathbf{E}(\chi_A|\mathcal{T}(\xi)) - \mathbf{E}(\chi_A|\mathcal{B}(\xi_N^\infty))\|_1 < \varepsilon$$

provided N is sufficiently large, which we may and will assume.

Let now $B \in \mathcal{B}(\xi_N^\infty)$. Then

$$\mu(A \cap B) = \int_B \chi_A d\mu = \int_B \mathbf{E}(\chi_A|\mathcal{B}(\xi_N^\infty))d\mu.$$

Therefore

$$|\mu(A \cap B) - \mu(A)\mu(B)| = \Big| \int_B (\mathbf{E}(\chi_A|\mathcal{B}(\xi_N^\infty)) - \mathbf{E}(\chi_A|\mathcal{T}(\xi)))d\mu \Big| < \varepsilon,$$

as required. □

The equivalence with item (3) requires some preparation. We also need the following generalization of the basic properties of entropy that we have proved for finite partitions. This lemma is included in the last example sheet.

Lemma 95. *Let $(X, \mathcal{B}, \mu, T)$ be a MPS. Let $\xi, \eta \subset \mathcal{B}$ be finite partitions and let $\mathcal{A} \subset \mathcal{B}$ be a σ-algebra. Suppose that there is a sequence of finite partitions $\tau_1, \tau_2, \ldots \subset \mathcal{B}$ such that $\mathcal{A}$ is the smallest σ-algebra that contains the atoms of τ_i for all i. We write $\eta \vee \mathcal{A}$ for the smallest σ-algebra that contains $\mathcal{A}$ and the atoms of η. Then*

- $I_\mu(T^{-1}\xi|T^{-1}\mathcal{A})(x) = I_\mu(\xi|\mathcal{A})(Tx)$,
- $H_\mu(T^{-1}\xi|T^{-1}\mathcal{A}) = H_\mu(\xi|\mathcal{A})$,
- $I_\mu(\xi \vee \eta|\mathcal{A})(x) = I_\mu(\xi|\mathcal{A})(x) + I_\mu(\eta|\xi \vee \mathcal{A})(x)$,
- $H_\mu(\xi \vee \eta|\mathcal{A}) = H_\mu(\xi|\mathcal{A}) + H_\mu(\eta|\xi \vee \mathcal{A})$,
- $H_\mu(\xi|\eta \vee \mathcal{A}) \leq H_\mu(\xi|\mathcal{A})$.

We also recall the following result from the last example sheet, which is a variant of Proposition 95.

Proposition 96. *Let $(X, \mathcal{B}, \mu)$ be a probability space and let $\mathcal{A}_1, \mathcal{A}_2, \ldots \subset \mathcal{B}$ be a sequence of σ-algebras such that $\mathcal{A}_i \supset \mathcal{A}_{i+1}$ for all i. Let $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2 \cap \cdots$. Let $\eta \subset \mathcal{B}$ be a finite partition. Then*

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

in L^1 and pointwise μ-almost everywhere.

The next two lemmata will be used in the proof of the implication (2) $\Rightarrow$ (3).

Lemma 97. *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and let $\xi \subset \mathcal{B}$ be a finite partition. Then*

$$h_\mu(T, \xi) = H_\mu(\xi|\mathcal{B}(\xi_1^\infty)).$$

Proof. In the previous section, we showed that

$$h_\mu(T, \xi) = \lim_{n\to\infty} H_\mu(\xi|\xi_1^n).$$

Now the claim follows from the convergence of entropy conditioned on partitions to the entropy conditioned on the σ-algebra generated by the partitions, which was proved in Proposition 95. □

Lemma 98. *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and let $\xi \subset \mathcal{B}$ be a finite partition. If $h_\mu(T|\xi) = 0$, then $H_\mu(\xi|\mathcal{T}(\xi)) = 0$.*

Proof. Suppose $h_\mu(T|\xi) = 0$. By the previous lemma, we have

$$H_\mu(\xi|\mathcal{B}(\xi_1^\infty)) = 0.$$

Using the chain rule and invariance, we can write

$$\begin{aligned} H_\mu(\xi|\mathcal{B}(\xi_n^\infty)) \leq & H_\mu(\xi_0^{n-1}|\mathcal{B}(\xi_n^\infty)) = \sum_{j=0}^{n-1} H_\mu(\xi_j^j|\mathcal{B}(\xi_{j+1}^\infty)) \\ = & \sum_{j=0}^{n-1} H_\mu(\xi|\mathcal{B}(\xi_1^\infty)) = 0. \end{aligned}$$

Now the lemma follows from Proposition 101. □

Proof of (2) $\Rightarrow$ (3). Suppose (2) holds and $h_\mu(T,\xi) = 0$ for some finite partition $\xi \subset \mathcal{B}$. We show that $H_\mu(\xi) = 0$, which completes the proof.

By the previous lemma, we have $H_\mu(\xi|\mathcal{T}(\xi)) = 0$.

By (2), we know that $\mathcal{T}(\xi)$ is trivial, hence

$$I_\mu(\xi|\mathcal{T}(\xi)) = -\log(\mathbf{E}(\chi_{[x]_\xi}, \mathcal{T}(\xi))) = -\log \mu([x]_\xi),$$

and $H_\mu(\xi|\mathcal{T}(\xi)) = H_\mu(\xi) = 0$, as required. □

The proof of the implication (3) $\Rightarrow$ (2) requires the following proposition.

Proposition 99. *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and let $\xi, \eta \subset \mathcal{B}$ be two finite partitions. Then*

$$h_\mu(T,\xi) = H_\mu(\xi|\mathcal{B}(\xi_1^\infty) \vee \mathcal{T}(\eta)).$$

We begin with two lemmata.

Lemma 100. *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and let $\xi \subset \mathcal{B}$ be a finite partition. Then*

$$h_\mu(T,\xi) = \frac{1}{n} H_\mu(\xi_0^{n-1}|\mathcal{B}(\xi_n^\infty))$$

for each n.

Proof. By the chain rule and invariance, we have

$$H_\mu(\xi_0^{n-1}|\mathcal{B}(\xi_n^\infty)) = \sum_{j=0}^{n-1} H_\mu(\xi_j^j|\mathcal{B}(\xi_{j+1}^\infty)) = nH_\mu(\xi|\mathcal{B}(\xi_1^\infty)).$$

The claim now follows from Lemma 102. □

Lemma 101. *Let $(X, \mathcal{B}, \mu, T)$ be a measure preserving system and let $\xi, \eta \subset \mathcal{B}$ be two finite partitions. Then*

$$h_\mu(\xi,T) = \lim_{n\to\infty} \frac{1}{n} H_\mu(\xi_0^{n-1}|\mathcal{B}(\xi_n^\infty) \vee \mathcal{B}(\eta_n^\infty)).$$

Proof. We note that

$$H_\mu(\xi_0^{n-1}|\mathcal{B}(\xi_n^\infty) \vee \mathcal{B}(\eta_n^\infty)) \leq H_\mu(\xi_0^{n-1})$$

for each n. Therefore,

$$h_\mu(\xi,T) \geq \limsup_{n\to\infty} \frac{1}{n} H_\mu(\xi_0^{n-1}|\mathcal{B}(\xi_n^\infty) \vee \mathcal{B}(\eta_n^\infty)).$$

In order to prove

$$h_\mu(\xi,T) \leq \liminf_{n\to\infty} \frac{1}{n} H_\mu(\xi_0^{n-1}|\mathcal{B}(\xi_n^\infty) \vee \mathcal{B}(\eta_n^\infty)), \tag{17}$$

we suppose to the contrary that it does not hold.

We can write

$$\begin{aligned} H_\mu(\xi_0^{n-1} \vee \eta_0^{n-1} | \mathcal{B}(\xi_n^\infty) \vee \mathcal{B}(\eta_n^\infty)) =& H_\mu(\xi_0^{n-1} | \mathcal{B}(\xi_n^\infty) \vee \mathcal{B}(\eta_n^\infty)) \\ &+ H_\mu(\eta_0^{n-1} | \mathcal{B}(\xi_0^\infty) \vee \mathcal{B}(\eta_n^\infty)) \\ H_\mu(\xi_0^{n-1} \vee \eta_0^{n-1} | \mathcal{B}(\xi_n^\infty)) =& H_\mu(\xi_0^{n-1} | \mathcal{B}(\xi_n^\infty)) + H_\mu(\eta_0^{n-1} | \mathcal{B}(\xi_0^\infty)). \end{aligned}$$

We note that

$$H_\mu(\eta_0^{n-1} | \mathcal{B}(\xi_0^\infty)) \geq H_\mu(\eta_0^{n-1} | \mathcal{B}(\xi_0^\infty) \vee \mathcal{B}(\eta_n^\infty))$$

for each n, and the failure of (17) yields (using the previous lemma)

$$\lim_{n\to\infty} \frac{1}{n} H_\mu(\xi_0^{n-1} | \mathcal{B}(\xi_n^\infty)) > \liminf_{n\to\infty} \frac{1}{n} H_\mu(\xi_0^{n-1} | \mathcal{B}(\xi_n^\infty) \vee \mathcal{B}(\eta_n^\infty)).$$

Therefore

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} & H_\mu(\xi_0^{n-1} \vee \eta_0^{n-1} | \mathcal{B}(\xi_n^\infty) \vee \mathcal{B}(\eta_n^\infty)) \\ &< \limsup_{n\to\infty} \frac{1}{n} H_\mu(\xi_0^{n-1} \vee \eta_0^{n-1} | \mathcal{B}(\xi_n^\infty)) \\ &\leq \limsup_{n\to\infty} \frac{1}{n} H_\mu(\xi_0^{n-1} \vee \eta_0^{n-1}) \\ &= h_\mu(T, \xi \vee \eta). \end{aligned}$$

This contradicts, however, the previous lemma, hence (17) and the lemma hold. □

Proof of the proposition. Using the chain rule and invariance, we can write

$$\begin{aligned} H_\mu(\xi_0^{n-1} | \mathcal{B}(\xi_n^\infty) \vee \mathcal{B}(\eta_n^\infty)) &= \sum_{j=0}^{n-1} H_\mu(\xi_j^j | \mathcal{B}(\xi_{j+1}^\infty) \vee \mathcal{B}(\eta_n^\infty)) \\ &= \sum_{j=0}^{n-1} H_\mu(\xi | \mathcal{B}(\xi_1^\infty) \vee \mathcal{B}(\eta_{n-j}^\infty)). \end{aligned}$$

Therefore, it follows from the previous lemma that

$$h_\mu(\xi, T) = \text{C-lim}_{n\to\infty} H_\mu(\xi | \mathcal{B}(\xi_1^\infty) \vee \mathcal{B}(\eta_n^\infty)).$$

By Proposition 101, we have

$$\lim_{n\to\infty} H_\mu(\xi | \mathcal{B}(\xi_1^\infty) \vee \mathcal{B}(\eta_n^\infty)) = H_\mu(\xi | \mathcal{B}(\xi_1^\infty) \vee \mathcal{T}(\eta)),$$

and the claim follows. □

Proof of (3) ⇒ (2). Suppose that (3) holds. Suppose to the contrary that there is a finite partition $\eta \subset \mathcal{B}$ such that $\mathcal{T}(\eta)$ is not trivial, that is, it contains a set $A \in \mathcal{T}(\eta)$ with $0 < \mu(A) < 1$.

Let $\xi = \{A, A^c\}$, so $H_\mu(\xi) > 0$. We prove that

$$H_\mu(\xi | \mathcal{B}(\xi_1^\infty) \vee \mathcal{T}(\eta)) = 0, \tag{18}$$

which implies $h_\mu(T,\xi) = 0$ by the proposition. This contradicts (3), finishing the proof of the theorem.

We calculate

$$I_\mu(\xi|\mathcal{B}(\xi_1^\infty) \vee \mathcal{T}(\eta))(x) = -\log \mathbf{E}(\chi_{[x]_\xi}|\mathcal{B}(\xi_1^\infty) \vee \mathcal{T}(\eta))(x).$$

Since $[x]_\xi$ is either A or A^c, which are both contained in $\mathcal{T}(\eta)$, we see that $\chi_{[x]_\xi}$ is $\mathcal{B}(\xi_1^\infty) \vee \mathcal{T}(\eta)$-measurable, hence

$$\mathbf{E}(\chi_{[x]_\xi}|\mathcal{B}(\xi_1^\infty) \vee \mathcal{T}(\eta)) = \chi_{[x]_\xi}.$$

This proves that

$$I_\mu(\xi|\mathcal{B}(\xi_1^\infty) \vee \mathcal{T}(\eta))(x) = 0$$

for μ-almost all x, hence (18) holds. □

13. Rudolph's theorem

In this section, we discuss the problem posed by Furstenberg of classifying probability measures on $\mathbf{R}/\mathbf{Z}$ that are invariant under both $T_2 : x \mapsto 2 \cdot x$ and $T_3 : x \mapsto 3 \cdot x$. As demonstrated by the examples on the exmaple sheets, there is a plentiful supply of measures that are invariant under either one of these maps. However, the collection of measures invariant under both is expected to be more restricted, and the next result confirms this partially.

Theorem 102 (Rudolph)**.** *Let μ be a T_2 and T_3 invariant ergodic measure. That is, if $A \subset [0,1]$ is measurable and $T_2^{-1}(A) = T_3^{-1}(A) = A$, then $\mu(A) \in \{0,1\}$. Suppose $h_\mu(T_2) > 0$ or $h_\mu(T_3) > 0$. Then $\mu = m$ is the Lebesgue measure.*

A word of caution. In the theorem, we assume that μ is ergodic with respect to the joint action of T_2 and T_3, but this is a weaker assumption than assuming that it is ergodic as a T_2 or T_3 invariant measure. Indeed, there may be a set $A \in \mathcal{B}$ with $0 < \mu(A) < 1$, such that A is either T_2 or T_3 invariant. We only assume that it cannot be invariant under both maps.

We discuss a proof given by Host soon after Rudolph's paper. In this approach, the following slightly stronger result is proved.

Theorem 103. *Let μ be a T_3-invariant ergodic measure. Suppose that $h_\mu(T_3) > 0$. Then μ-almost every $x \in \mathbf{R}/\mathbf{Z}$ is normal in base 2, that is, the sequence $T_2^n x$ is equidistributed in $\mathbf{R}/\mathbf{Z}$.*

Moreover, without the assumption of ergodicity, we can conclude

$$\mu(x \in \mathbf{R}/\mathbf{Z} : \{T_2^n x\} \text{ is equidistributed}) \geq h_\mu(T_3)/\log(3).$$

Remark 104. This result is of independent interest. An example for an ergodic T_3 invariant measure is the so-called Cantor-Lebesgue measure supported on the middle-third Cantor set. In the usual construction of the Cantor set, for each $n \in \mathbf{Z}_{\geq 0}$, one gives a collection

of 2^n disjoint intervals of equal length 3^{-n}. Using the standard extension theorems, it is possible to show that there is a unique probability measure, which gives 2^{-n} mass to each interval in the n'th stage of the construction of the Cantor set. This measure is called the Cantor-Lebesgue measure and it can be proved that it is T_3-invariant, ergodic and has entropy $\log(2)$. Therefore, Theorem 108 implies that almost-every element of the Cantor set with respect to this measure is normal in base 2. This does not follow from our earlier results, because the Cantor set is of Lebesgue measure 0.

We begin with the proof of Theorem 108. It is based on the following simple algebraic observation.

Lemma 105. *Fix $k \in \mathbf{Z}_{>0}$. The order of 2 modulo 3^k, that is the smallest integer $n \in \mathbf{Z}_{>0}$ such that $3^k | 2^n - 1$ is $2 \cdot 3^{k-1}$.*

Proof. We denote by $\mathrm{ord}_{3^k}(2)$ the order of 2 modulo 3^k. We prove by induction that

$$\mathrm{ord}_{3^k}(2) = 2 \cdot 3^{k-1} \text{ and } 3^{k+1} \nmid 2^{2 \cdot 3^{k-1}} - 1.$$

If $k = 1$, then this holds, because $3 \nmid 2 - 1$, $3 | 2^2 - 1$ and $3^2 \nmid 2^n - 1$.

Suppose that $k > 1$ and the claim holds for $k - 1$. If $3^k | 2^n - 1$ for some integer n, then $3^{k-1} | 2^n - 1$, too, hence $\mathrm{ord}_{3^{k-1}}(2) | n$. By the induction hypothesis, $\mathrm{ord}_{3^k}(2)$ is then of the form $a \cdot 2 \cdot 3^{k-2}$ for some positive integer a. Furthermore, by the second half of the induction hypothesis, we have

$$2^{2 \cdot 3^{k-2}} = b \cdot 3^{k-1} + 1$$

for some $b \in \mathbf{Z}$ with $3 \nmid b$.

By the binomial theorem, we can write

$$2^{2 \cdot 2 \cdot 3^{k-2}} = b^2 \cdot 3^{2k-2} + 2 \cdot b \cdot 3^{k-1} + 1$$

$$2^{3 \cdot 2 \cdot 3^{k-2}} = b^3 \cdot 3^{3k-3} + 3 \cdot b^2 \cdot 3^{2k-2} + 3 \cdot b \cdot 3^{k-1} + 1.$$

This shows that $3^k \nmid 2^{2 \cdot 2 \cdot 3^{k-2}} - 1$ and $3^k | 2^{3 \cdot 2 \cdot 3^{k-2}} - 1$, thus $\mathrm{ord}_{3^k}(2) = 2 \cdot 3^{k-1}$. Moreover, we also see from the last equation that $3^{k+1} \nmid 2^{3 \cdot 2 \cdot 3^{k-2}} - 1$, which completes the induction. □

Remark 106. Before continuing the proof of Theorem 108, we explain the general strategy and the role of the above lemma.

Let $\frac{a}{3^k} \in [0, 1]$ be a number such that $a \in \mathbf{Z}$ and $3 \nmid a$. It follows from the lemma that the orbit of $\frac{a}{3^k}$ under T_2 is the set

$$\left\{ \frac{b}{3^k} : b \in [1, \dots, 3^k], 3 \nmid b \right\}.$$

This set is – vaguely speaking – very uniformly distributed at scales larger than 3^{-k}.

Note also that $T_2^n(x + y) = T_2^n x + T_2^n y$. Intuitively, this suggests that if the distribution of the T_2 orbit of x is very non-uniform, for

example, it spends most of its time near 0, then the orbit of $x + \frac{a}{3^k}$ is quite uniform.

In the next lemma, we formalize this idea that among the points $x + \frac{a}{3^k}$ only a few can have a non-uniformly distributed T_2-orbit. Then, in the second half of the proof, we show that a T_3-invariant measure with positive entropy cannot concentrate on a few elements among the points $\{x + \frac{a}{3^k}\}$. This last statement will require clarification, because the set $\{x + \frac{a}{3^k}\}$ is finite, hence probably have measure 0.

To prove Theorem 108, we need to show that for μ-almost every x, we have

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T_2^n x) = \int f dm \tag{19}$$

for all $f \in C(\mathbf{R}/\mathbf{Z})$. Since linear combinations of the functions $x \mapsto \exp(2\pi i m x)$ is dense in $C(\mathbf{R}/\mathbf{Z})$, it is enough to prove (19) for $f(x) = \exp(2\pi i m x)$ for all $m \in \mathbf{Z}$. The claim is trivial for $m = 0$, as in this case $f(x) = 1$ for all x.

When $f = \exp(2\pi i m x)$ the left hand side of (19) is a trigonometric polynomial that plays a special role in the proof, therefore we introduce the notation

$$P_{N,m}(x) := \frac{1}{N} \sum_{n=0}^{N-1} \exp(2\pi i m \cdot 2^n x).$$

In this notation, (19) can be rewritten as

$$\lim_{N\to\infty} P_{N,m}(x) = 0$$

for all $m \neq 0$ and for μ-almost every x.

Lemma 107. *Fix $k \in \mathbf{Z}_{>0}$ and $m \in \mathbf{Z}\backslash\{0\}$. Let α be the largest exponent such that $3^\alpha | m$. Let $N \leq 2 \cdot 3^{k-\alpha-1}$ be an integer. Then*

$$\sum_{a=0}^{3^k-1} |P_{N,m}(x + a/3^k)|^2 = \frac{3^k}{N}.$$

This lemma will be applied with N being not much smaller than 3^k, hence the conclusion shows that $|P_{N,m}(x + a/3^k)| < \varepsilon$ except for a few choices for a for any previously fixed $\varepsilon > 0$. This is one way to formalize the statement that the T_2-orbit of $x + a/3^k$ of length N is uniformly distributed at the scale $1/m$ for most choices of a.

Proof. Using the definition of $P_{N,m}$, we can compute

$$\begin{aligned} |P_{N,m}(x)|^2 =& \frac{1}{N^2} \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} \exp(2\pi i m \cdot 2^{n_1} x) \overline{\exp(2\pi i m \cdot 2^{n_2} x)} \\ =& \frac{1}{N^2} \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} \exp(2\pi i m (2^{n_1} - 2^{n_2}) x). \end{aligned} \tag{20}$$

We also note that if $b \in \mathbf{Z}$ and $3^k \nmid b$, then

$$\sum_{a=0}^{3^k-1} \exp(2\pi i b(x+a/3^k)) = \exp(2\pi i b x) \sum_{a=0}^{3^k-1} \exp(2\pi i b a/3^k)$$
$$= \exp(2\pi i b x) \frac{\exp(2\pi i b \cdot 3^k/3^k) - \exp(2\pi i b)}{\exp(2\pi i b/3^k) - 1} = 0.$$

On the other hand if $3^k | b$, then clearly $\exp(2\pi i b(x+a/3^k)) = \exp(2\pi i b x)$ for all a, hence

$$\sum_{a=0}^{3^k-1} \exp(2\pi i b(x+a/3^k)) = 3^k \cdot \exp(2\pi i b x).$$

We combine these observations with (20) and obtain

(21)
$$\sum_{a=0}^{3^k-1} |P_{N,m}(x+a/3^k)|^2 = \frac{3^k}{N^2} \sum_{n_1,n_2 : 3^k | m(2^{n_1}-2^{n-2})} \exp(2\pi i m(2^{n_1}-2^{n_2})x).$$

Let $0 \le n_1 < n_2 \le N-1$. By the previous lemma, the largest exponent β such that

$$3^\beta | 2^{n_1} - 2^{n_2} = 2^{n_1}(2^{n_2-n_1}-1)$$

is at most $k-\alpha-1$, because $n_2 - n_1 < N \le 2 \cdot 3^{k-\alpha-1}$. A similar argument applies if $n_1 > n_2$. Thus $3^k | m(2^{n_1}-2^{n_2})$ is equivalent to $n_1 = n_2$, and the lemma follows from (21). □

We turn to the second half of the proof of Theorem 108. In this part we utilize the entropy assumption and aim to show that a T_3-invariant ergodic measure with positive entropy may not concentrate on a few elements among the points $\{x + \frac{a}{3^k}\}$. This vague statement is formalized in the next result.

Proposition 108. *Let μ be a T_3-invariant ergodic measure on $\mathbf{R}/\mathbf{Z}$ and let $k \in \mathbf{Z}_{>0}$ be a number. Define the measure μ_k on $\mathbf{R}/\mathbf{Z}$ by*

$$\mu_k(A) := \sum_{a=0}^{3^k-1} \mu(A + a/3^k).$$

Write $\xi = \{[0,1/3), [1/3,2/3), [2/3,1)\}$.

For every $\varepsilon > 0$ and for μ-almost every $x \in \mathbf{R}/\mathbf{Z}$ we have

$$\mu_k([x]_{\xi_0^K}) \ge \exp((h_\mu(T_3) - \varepsilon)k) \mu([x]_{\xi_0^K})$$

provided k is sufficiently large depending on ε and x and K is sufficiently large depending on ε, x and k.

Remark 109. We note that $[x]_{\xi_0^K}$ is an interval of length $3^{-(K+1)}$ containing x and $\mu_k([x]_{\xi_0^K})$ is the μ measure of the union of intervals of the same length around the points $\{x + b/3^k\}$. The proposition claims that

the μ measure of these 3^k intervals is much larger than the μ-measure of any single interval.

The proposition is deduced from the following.

Lemma 110. *Let μ be a T_3-invariant ergodic measure on $\mathbf{R}/\mathbf{Z}$ and let $\xi = \{[0,1/3),[1/3,2/3),[2/3,1)\}$.*

Then

$$\lim_{k\to\infty} \frac{1}{k} I_\mu(\xi_0^{k-1}|\xi_k^\infty)(x) = h_\mu(T_3)$$

holds μ-almost everywhere.

Proof. Using the chain rule and invariance, we write

$$I_\mu(\xi_0^{k-1}|\xi_k^\infty)(x) = \sum_{j=0}^{k-1} I_\mu(\xi_j^j|\xi_{j+1}^\infty)(x) = \sum_{j=0}^{k-1} I_\mu(\xi|\xi_1^\infty)(T^j x).$$

By the pointwise ergodic theorem and ergodicity, we have

$$\lim_{k\to\infty} \frac{1}{k} I_\mu(\xi_0^{k-1}|\xi_k^\infty)(x) = \int I_\mu(\xi|\xi_1^\infty) d\mu = H_\mu(\xi|\xi_1^\infty) = h_\mu(T_3,\xi).$$

The claim follows from the next lemma and a modification of the Kolmogorov-Sinai theorem adapted for non-invertible systems. □

Lemma 111. *Let μ be a T_3-invariant ergodic measure on $\mathbf{R}/\mathbf{Z}$ and let $\xi = \{[0,1/3),[1/3,2/3),[2/3,1)\}$.*

Then the partition ξ is a one sided generator, that is, for every $A \in \mathcal{B}$ and for every $\varepsilon > 0$, there is a number n and a set $B \in \mathcal{B}(\xi_0^n)$ such that $\mu(B \triangle A) < \varepsilon$.

Proof. The set

$$\mathcal{C} := \{A \in \mathcal{B} : \forall \varepsilon > 0,\ \exists n \text{ and } \exists B \in \xi \text{ such that } \mu(A \triangle B) < \varepsilon\}$$

is a σ-algebra. Therefore it is enough to show that it contains all open sets.

Let $U \in \mathcal{B}$ be open. The atoms of ξ_0^n are clearly contained in $\mathcal{C}$. So we are done if we show that U is a union of countably many such atoms. For any $x \in U$, there is $\delta > 0$ such that $[x-\delta, x+\delta] \subset U$. If n is sufficiently large so that $3^{-n} < \delta$, then $[x]_{\xi_0^{n-1}} \subset [x-\delta, x+\delta]$. This means that the set of all atoms of ξ_0^{n-1} for $n \in \mathbf{Z}_{\geq 0}$ that are contained in U cover U. This is clearly a countable collection, as there are only finitely many atoms for each n. □

Proof of Proposition 113. By Lemma 115, for all ε and for μ-almost all x, we have

$$I_\mu(\xi_0^{k-1}|\xi_k^\infty)(x) > k(h_\mu(T_3) - \varepsilon)$$

provided k is sufficiently large depending in ε and x. By Proposition 95, we have

$$\log\Big(\frac{\mu([x]_{\xi_k^K})}{\mu([x]_{\xi_0^K})}\Big) = I_\mu(\xi_0^{k-1}|\xi_k^K)(x) > k(h_\mu(T_3) - \varepsilon)$$

provided K is sufficiently large depending on x, ε and k.

Writing $x = 0.x_1x_2\ldots_{(3)}$ for the base 3 expansion of x, we have

$$\begin{aligned}
[x]_{\xi_0^K} =& \{y = 0.y_1y_2\ldots_{(3)} : y_j = x_j \text{ for } j = 1,\ldots,K+1\}\\
=& \Big[\frac{[3^{K+1}x]}{3^{K+1}}, \frac{\lfloor 3^{K+1}x\rfloor + 1}{3^{K+1}}\Big),\\
[x]_{\xi_k^K} =& \{y = 0.y_1y_2\ldots_{(3)} : y_j = x_j \text{ for } j = k+1,\ldots,K+1\}\\
=& \bigcup_{a=0}^{3^k-1} \Big[\frac{[3^{K+1}x]}{3^{K+1}} + \frac{a}{3^k}, \frac{\lfloor 3^{K+1}x\rfloor + 1}{3^{K+1}} + \frac{a}{3^k}\Big).
\end{aligned}$$

This means that $\mu([x]_{\xi_k^K}) = \mu_k([x]_{\xi_0^K})$, hence

$$\frac{\mu_k([x]_{\xi_0^K})}{\mu([x]_{\xi_0^K})} \geq \exp(k(h_\mu(T_3) - \varepsilon))$$

and the claim follows. □

Proposition 113 can be used to compare the μ and μ_k measures, but it only applies to sets of very special form. In the next corollary we obtain a more general estimate.

Corollary 112. *For each $k \in \mathbf{Z}_{>0}$, there is a set $A_k \in \mathcal{B}$ such that for μ-a.e. $x \in X$ we have $x \in A_k$ for all sufficiently (depending on x) large k and*

$$\mu(A_k \cap U) \leq \exp(-h_\mu(T_3)k/2)\mu_k(U).$$

Proof. We set

$$A_k := \{x \in \mathbf{R}/\mathbf{Z} : \mu_k([x]_{\xi_0^K}) \geq \exp(h_\mu(T_3)k/2)\mu([x]_{\xi_0^K}) \text{ for all sufficiently large } K\}.$$

Proposition 113 applied with $\varepsilon = h_\mu(T_3)/2$, yields that for μ-almost all x, $x \in A_k$ for all sufficiently large k.

To prove the second claim, we note that for any $x \in A_k \cap U$, we have

$$[x]_{\xi_0^K} \subset U \quad \text{and} \quad \mu_k([x]_{\xi_0^K}) \geq \exp(h_\mu(T_3)k/2)\mu([x]_{\xi_0^K}) \tag{22}$$

for all large enough K, because U is open and because of the definition of A_k.

For each K we denote by $\mathcal{A}_K$ the collection of the atoms of ξ_0^K for which (22) holds. Since the elements of $\mathcal{A}_K$ are disjoint and are contained in U, we have $\mu_k(\bigcup_{A\in\mathcal{A}_K} A) \leq \mu_k(U)$. Furthermore, using the second half of (22), we obtain

$$\mu\Big(\bigcup_{A\in\mathcal{A}_K} A\Big) \leq \exp(-h_\mu(T_3)k/2)\mu_k(U).$$

Then

$$\mu\Big(\bigcap_{K\geq L}\bigcup_{A\in\mathcal{A}_K} A\Big) \leq \exp(-h_\mu(T_3)k/2)\mu_k(U).$$

for all $L \in \mathbf{Z}_{>1}$. The sets $\bigcap_{K \geq L} \bigcup_{A \in \mathcal{A}_K} A$ increase with respect to containment as L grows, and their union covers $U \cap A_k$, hence the claim is proved. □

We also need the next lemma, whose proof is available in standard textbooks on probability, and it is also a useful exercise.

Lemma 113 (Borel-Cantelli)**.** *Let $(X, \mathcal{B}, \mu)$ be a probability space and let $B_1, B_2, \ldots \in \mathcal{B}$. If $\sum_{n=1}^{\infty} \mu(B_n) < \infty$, then*

$$\mu(x : x \in B_n \text{ for infintely many } n) = 0.$$

In the proof of Theorem 108, we wish to use the Borel-Cantelli lemma to show that for μ-almost all x, we have $|P_{N,m}(x)| > \varepsilon$ only for finitely many N. However, the sum of the measures of the corresponding sets may be divergent. To overcome this problem, we show that it is enough to prove $\lim P_{N,m} = 0$ through a suitable subsequence.

Lemma 114. *For any $C \in \mathbf{Z}_{>0}$ and for any $x \in \mathbf{R}/\mathbf{Z}$,*

$$\lim_{N \to \infty} P_{N,m}(x) = 0 \quad \text{is equivalent to} \quad \lim_{L \to \infty} P_{L^C,m}(x) = 0.$$

Proof. To prove this lemma, we observe from the definition of $P_{N,m}$ that

$$|NP_{N,m}(x) - L^C P_{L^C,m}(x)| \leq |L^C - N|.$$

Thus

$$\begin{aligned} |P_{N,m}(x) - P_{L^C,m}(x)| \leq & \frac{1}{N}\big(|NP_{N,m}(x) - L^C P_{L^C,m}(x)| + |N - L^C| P_{L^C,m}(x)\big) \\ \leq & \frac{2|L^C - N|}{N}. \end{aligned}$$

As N grows, the ratio between N and the nearest C-power converges to 1, hence this proves the claim. □

Proof of Theorem 108. We give the proof only in the ergodic case. After the proof, we will discuss how the proof can be adapted in the non-ergodic case.

As we noted above, we need to show that $P_{N,m}(x) \to 0$ for all $m \neq 0$ and for μ-almost all x. By Lemma 119, this is equivalent to $\lim_{L \to \infty} P_{L^C,m}(x) = 0$, where C is any fixed integer.

We fix $m \in \mathbf{Z}_{\neq 0}$, $C \in \mathbf{Z}_{>0}$ and $\varepsilon > 0$. We denote by B_ε the set of points x such that $\limsup_{L \to \infty} |P_{L^C,m}(x)| > \varepsilon$ and we aim to prove $\mu(B_\varepsilon) = 0$. This will complete the proof of the theorem.

Let α be the largest exponent such that $3^\alpha | m$. For each $N \in \mathbf{Z}_{>0}$, let $k = k(N)$ be an integer such that $2 \cdot 3^{k-\alpha-2} < N \leq 2 \cdot 3^{k-\alpha-1}$. We first estimate $\int P_{N,m} d\mu_k$ using Lemma 112. Then we use this to estimate the μ_k-measure of the set, where $P_{N,m}$ is large. Finally, we use Corollary 117 to derive an estimate for the μ-measure of the same set.

By Lemma 112, we have

$$\sum_{a=0}^{3^k-1} |P_{N,m}(x+a/3^k)|^2 = \frac{3^k}{N} < \frac{3^{\alpha+2}}{2}.$$

Observe that the right hand side depends only on m. Using the definition of μ_k, we can write

$$\int |P_{N,m}(x)|^2 d\mu_k(x) = \int \sum_{a=0}^{3^k-1} |P_{N,m}(x+a/3^k)|^2 d\mu < \frac{3^{\alpha+2}}{2},$$

hence

$$\mu_k(x \in \mathbf{R}/\mathbf{Z} : |P_{N,m}(x)| > \varepsilon) < \frac{3^{\alpha+2}}{2\varepsilon^2}$$

for any $\varepsilon > 0$.

We apply Corollary 117 for the set $U = \{x \in \mathbf{R}/\mathbf{Z} : |P_{N,m}(x)| > \varepsilon\}$, and obtain

$$\mu(x \in A_k : |P_{N,m}(x)| > \varepsilon) < \frac{3^{\alpha+2}}{2\varepsilon^2} \exp(-h_\mu(T_3)k/2). \tag{23}$$

For almost every $x \in \mathbf{R}/\mathbf{Z}$, we have $x \in A_k$ for all sufficiently large k and for all $x \in B_\varepsilon$, we have $|P_{L^C,m}(x)| > \varepsilon$ for infinitely many L. We write

$$D_{L,\varepsilon} = \{x \in \mathbf{R}/\mathbf{Z} : x \in A_{k(L^C)} \text{ and } |P_{L^C,m}(x)| > \varepsilon\}.$$

Then almost all $x \in B_\varepsilon$ is contained in $D_{L,\varepsilon}$ for infinitely many L. The Borel-Cantelli lemma implies that $\mu(B_\varepsilon) = 0$ provided

$$\sum_{L=1}^{\infty} \mu(D_{L,\varepsilon}) < \infty.$$

We note that (23) yields

$$\mu(D_{L,\varepsilon}) < \frac{3^{\alpha+2}}{2\varepsilon^2} \exp(-h_\mu(T_3)k(L^C)/2).$$

By the definition of $k(L^C)$, we have $2 \cdot 3^{k(L^C)-\alpha-1} \geq L^C$. Choosing C sufficiently large, this implies $k(L^C) > 4\log(L)/h_\mu(T_3)$, hence

$$\sum_{L=1}^{\infty} \mu(D_{L,\varepsilon}) < \frac{3^{\alpha+2}}{2\varepsilon^2} \sum_{L=1}^{\infty} L^{-2} < \infty.$$

□

Remark 115. Using a bit more sophisticated measure theory, the part of the proof that compares the measure μ and μ_k can be made more conceptual. One can deduce from the proof of Proposition 113 that the Radon-Nikodym derivative satisfies

$$\lim_{k\to\infty} \log\Big(\frac{d\mu_k}{d\mu}(x)\Big) = h_\mu(T_3) \tag{24}$$

for almost all x. (We do not claim that μ_k is absolutely continuous with respect to μ, and usually it is not.) Using this property, one can directly estimate $\int P_{N,m} d\mu$ in terms of $\int P_{N,m} d\mu_k$, which eliminates most of the unpleasant technicalities of the proof. However, the proof via (24) is less elementary.

Remark 116. The hypothesis that the measure μ is ergodic was used only once in the proof of Lemma 115, where we used it to compute the limit in the pointwise ergodic theorem. In the general case, we still obtain a limit function f and it is possible to show that f takes values in $[0, \log(3)]$ and $\int f d\mu = h_\mu(T_3)$. Writing $A_\delta \in \mathcal{B}$ for the set of x such that $f(x) > \delta$, we see that for every $\varepsilon > 0$ there is $\delta > 0$ such that $\mu(A_\delta) > h_\mu(T_3)/\log(3) - \varepsilon$.

Using this in the proof of Proposition 113, one can conclude that for μ-almost every $x \in A_\delta$, we have

$$\mu_k([x]_{\xi_0^K}) \geq \exp(\delta k)\mu([x]_{\xi_0^K})$$

provided k is sufficiently large depending on x and δ and K is sufficiently large depending on x, δ and k. This version of Proposition 113 is sufficient for the purposes of the proof of Theorem 108 if we restrict our attention to elements of A_δ.

Finally, we deduce Theorem 107.

Proof of Theorem 107. Let μ be a T_2 and T_3 invariant ergodic measure. The role of T_2 and T_3 are interchangeable, so we assume $h_\mu(T_3) > 0$. We write A for the set of points x such that the sequence $T_2^n x$ is equidistributed.

By Theorem 108, we know that $\mu(A) > 0$. We aim to use the ergodicity assumption to show that $\mu(A) = 1$. We first note that $T_2^{-1}(A) = A$. Indeed, $T_2^n x$ is equidistributed if and only if $T_2^n(T_2 x) = T_2^{n+1} x$ is.

We now show that $A \subset T_3^{-1}(A)$. This is based on the fact that T_2 and T_3 commute, exploiting this fact similarly to the proof of Furstenberg's theorem on unique ergodicity of skew products, where we showed that the set of generic points is invariant under translations in the second coordinate. The details are as follows. Let $x \in A$. Then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T_2^n x) = \int_{\mathbf{R}/\mathbf{Z}} f(x) dx, \tag{25}$$

for any $f \in C(\mathbf{R}/\mathbf{Z})$. We can write

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T_2^n(T_3 x)) &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T_3 T_2^n x) \\ &= \int_{\mathbf{R}/\mathbf{Z}} f(T_3 x) dx = \int_{\mathbf{R}/\mathbf{Z}} f(x) dx, \end{aligned}$$

where we used (25) for $f \circ T_3$ in place of f and then that the Lebesgue measure is T_3 invariant. This proves that $T_3 x \in A$, hence $x \in T_3^{-1}(A)$.

We consider the set $B = \bigcup_{n=0}^{\infty} T_3^{-n}(A)$. We have $T_3^{-1}(B) = \bigcup_{n=1}^{\infty} T_3^{-n}(A)$, hence $T_3^{-1}(B) \subset B$. On the other hand, $A \subset T_3^{-1}(A)$ implies $B \subset T_3^{-1}(B)$, hence $B = T_3^{-1}(B)$. We note finally that

$$T_2^{-1}(B) = \bigcup_{n=0}^{\infty} T_2^{-1} T_3^{-n}(A) = \bigcup_{n=0}^{\infty} T_3^{-n} T_2^{-1}(A) = B,$$

because $T_2^{-1}(A) = A$. Using ergodicity, and $\mu(B) \geq \mu(A) > 0$, we find $\mu(B) = 1$.

On the other hand, we know $\mu(T_3^{-1}(A) \backslash A) = \mu(T_3^{-1}(A)) - \mu(A) = 0$, which implies $\mu(T_3^{-(n+1)}(A) \backslash T_3^{-n}(A)) = 0$ for all n. Thus

$$\mu(B \backslash A) \leq \sum_{n=0}^{\infty} \mu(T_3^{-(n+1)}(A) \backslash T_3^{-n}(A)) = 0.$$

Therefore $\mu(A) = 1$.

For all $x \in A$ and for all $f \in C(\mathbf{R}/\mathbf{Z})$, we have

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T_2^n x) = \int f(x) dx.$$

Using $\mu(A) = 1$ and dominated convergence, we then have

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \int f(T_2^n x) d\mu = \int f(x) dx.$$

Since μ is T_2-invariant, we have $\int f(T_2^n x) d\mu = \int f d\mu$. Thus

$$\int f d\mu = \int f(x) dx.$$

This proves that μ is the Lebesgue measure, which we wanted to prove. □