Notes for the course: Ergodic theory and entropy.

Barney Bramham

Bochum
Summer Semester 2014[*]

Contents

[*] This version: 9 Oct 2014

Course Information

Tentative syllabus:

- Invariant measures of a transformation, ergodicity, Birkhoff ergodic theorem.
- Oseledec's theorem (in 2D), subadditive ergodic theorem, Lyapunov exponents.
- Metric (measure theoretic) entropy and topological entropy.
- Pesin-Ruelle inequality/formula.
- Introduction to Pesin's theory of non-uniformly hyperbolic systems.
- Interlude on *uniform* hyperbolicity: basic examples, properties such as shadowing, closing lemmas.
- Back to non-uniform hyperbolicity: shadowing and closing, Katok's theorem relating entropy and periodic points.

Literature:

The course will be based loosely around:

> Mark Pollicott: "Lectures on ergodic theory and Pesin theory on compact manifolds".

Here are some other sources:

- Denker: "Einführung In Die Analysis Dynamischer Systeme".
- Einsiedler-Schmidt: "Dynamische Systeme: ergodentheorie und topologische dynamik".
- Walters: "An introduction to ergodic theory".
- Mañe: "Ergodic theory and differentiable dynamics".
- Katok-Hasselblatt: "Introduction to the modern theory of dynamical systems"

- Barreira-Pesin: “Introduction to smooth ergodic theory“

- Fomin-Kornfeld-Sinai: ”Ergodic theory“.

Some ”alternative” sources - more playful or of a survey nature.

- Katok-Hasselblatt: ”A First Course in Dynamics: with a Panorama of Recent Developments”.

- ”Ergodic theory, symbolic dynamics and hyperbolic space.” edited by Bedford, Keane, Series.

- Lai-Sang Young: survey papers on entropy at www.cims.nyu.edu/∼lsy/

- Pugh and Shub: A nice survey paper (with lots of pictures). This will make sense once we have defined ergodicity. www.ams.org/journals/bull/2004-41-01/S0273-0979-03-00998-4/S0273-0979-03-00998-4.pdf

- Amie Wilkinson: survey papers on hyperbolicity at http://math.uchicago.edu/∼wilkinso/

- David Ruelle: ”Elements of differentiable dynamics and bifurcation theory“.

- Arnold-Avez: ”Ergodic problems of classical mechanics”.

- Terence Tao’s blog: he has a discussion of the ergodic theorem, although he is heading more in the direction of number theory.

1 Preliminaries from measure theory and functional analysis

This section is intended to be a quick review. The only thing we prove is how to use Riesz' representation theorem to reduce the statement of compactness of the space of probability measures (in the weak, or vague, topology) to that of Alaoglu's theorem about the weak-$*$ compactness of the unit ball in the dual of a Banach space. Bare in mind that for most of the course we will be working with regular Borel probability measures on smooth compact manifolds, or at least compact metric spaces, so that many of the following statements are stated in far less than full generality.

1.1 Measures

First assume X is an arbitrary set.

Definition 1.1. A **σ-algebra** on X is a collection of subsets $\mathcal{B} \subset X$ such that $\emptyset \in \mathcal{B}$ and is closed under countable unions, and closed under complements. We call the pair $(X, \mathcal{B})$ a **measurable space**.

Definition 1.2. A **measure** is a function $\mu : \{\sigma\text{-algebra}\} \to [0, \infty]$ satisfying $\mu(\emptyset) = 0$ and countable additivity:

$$B_1, B_2, \ldots \in \mathcal{B} \;\text{ pairwise disjoint } \implies \mu\left(\bigcup_i B_i\right) = \sum_i \mu(B_i).$$

When a σ-algebra is large, it is often useful to establish properties on a smaller collection of subsets of X.

- An **algebra** is a collection of subsets $\mathcal{O} \subset X$ such that $\emptyset, X \in X$ and is closed under finite unions and finite intersections.
- An algebra $\mathcal{O}$ determines/generates a unique σ-algebra $\mathcal{B}(\mathcal{O})$ by taking all possible countable unions and complements of sets in $\mathcal{O}$. Formally, $\mathcal{B}(\mathcal{O})$ is the intersection of all σ-algebras on X that contain $\mathcal{O}$.

- For example any topology on X is an algebra

$$\mathcal{O} = \{U \subset X \text{ open}\}.$$

  The σ-algebra $\mathcal{B}(X)$ generated by the open sets is called the **Borel σ-algebra** on X.

- Any measure defined on the Borel σ-algebra of X is called a **Borel measure** on X. We will additionally assume that this includes the condition $\mu(X) < \infty$.

Theorem 1.3. *(Caratheodory Extension theorem) Suppose that $\mathcal{O}$ is an algebra on X and $\mathcal{B} = \mathcal{B}(\mathcal{O})$ is the generated σ-algebra. Then any map $m : \mathcal{O} \to [0, +\infty]$ such that*

1. *$m(\emptyset) = 0$, and*

2. *$A_1, \dots A_n \in \mathcal{O}$ pairwise disjoint $\implies$ $m(\cup_i A_i) = \sum_i m(A_i)$, i.e. finitely additive,*

3. *$m(X) < \infty$,*

extends uniquely to a measure on X.

Corollary 1.4. *If m_1, m_2 are two finite measures defined on the same σ-algebra $\mathcal{B}$, then to see if $m_1 = m_2$ it suffices to check on any generating algebra.*

To compare measures the following notions are handy.

Definition 1.5. If m_1, m_2 are two measures on a measurable space $(X, \mathcal{B})$, then we say:

1. $m_1 << m_2$ if $m_2(B) = 0$ implies $m_1(B) = 0$ for all $B \in \mathcal{B}$. (Say m_1 is **absolutely continuous with respect to** m_2.)

2. $m_1 \sim m_2$ if $m_1 << m_2$ and $m_2 << m_1$. (Say m_1 and m_2 are **equivalent**.)

3. $m_1 \perp m_2$ if there exists $B \in \mathcal{B}$ such that $m_1(B) = 0 = m_2(X \backslash B)$. (Say m_1 and m_2 are **mutually singular**.)

Theorem 1.6 (Radon-Nikodym)**.** *If $\mu_1 << \mu_2$ are finite measures on $(X, \mathcal{B})$ then there exists $f : X \to \mathbb{R}$ μ_2-integrable such that*

$$\mu_1(B) \ = \ \int_B f \, d\mu_2$$

for all $B \in \mathcal{B}$. The function f is unique μ_2-a.e, and so determines a unique element in $L^1(X, \mathcal{B}, \mu_2)$ called the ***Radon-Nikodym derivative*** *and denoted $f = d\mu_1/d\mu_2$. Moreover $f \geq 0$.*

We recall the definition of measurable and integrable functions in the following remark:

Remark 1.7.

- A function $f : (X, \mathcal{B}) \to \mathbb{R}$ is **measurable** if $f^{-1}(U) \in \mathcal{B}$ for all $U \subset \mathbb{R}$ open. Then f is **integrable** w.r.t. a given measure μ on $(X, \mathcal{B})$ if

  $$\int_X |f| \, d\mu \ < \ \infty.$$

  To recall the definition of the left hand side see your favourite book on measure theory or analysis. The space of equivalence classes of μ-integrable functions is denoted

  $$L^1(X, \mathcal{B}, \mu) := \{f : X \to \mathbb{R} \text{ integrable w.r.t. } \mu \ \}/\sim$$

  where $f_1 \sim f_2$ if they are equal μ a.e. Similarly, for $1 \leq p < \infty$ we define

  $$L^p(X, \mathcal{B}, \mu) := \{f : X \to \mathbb{R} \text{ integrable w.r.t. } \mu \mid \int_X |f|^p \, d\mu < \infty\}/\sim .$$

  This is a Banach space with norm $\|f\|_{L^p} := \left(\int_X |f|^p \, d\mu\right)^{1/p}$.

- $L^\infty(X, \mathcal{B}, \mu)$ is defined to be those μ-measurable functions bounded μ-almost everywhere. This is a Banach space with norm $\|f\|_{L^\infty} = \inf c$ over all $c \in \mathbb{R}$ for which there exists $E \in \mathcal{B}$ with $\mu(E) = 0$ and $|f(x)| \leq c$ for all $x \notin E$.

1.2 Monotone convergence and approximation theorems

Let $(X, \mathcal{B})$ be a measurable space.

Call a sequence of measurable functions $f_j : (X, \mathcal{B}) \to \mathbb{R} \cup \{+\infty\}$, for $j \geq 1$, a **monotone increasing sequence** if

$$f_j(x) \leq f_{j+1}(x) \qquad \forall x \in X,\ \forall j \geq 1.$$

Note that such a sequence converges pointwise to the measurable function $f : (X, \mathcal{B}) \to \mathbb{R} \cup \{+\infty\}$ given by

$$f(x) = \sup_{j \geq 1} f_j(x)$$

for all $x \in X$.

Theorem 1.8 (Monotone convergence theorem)**.** *Suppose $f_j : (X, \mathcal{B}) \to [0, +\infty]$, $j \geq 1$, is a monotone increasing sequence converging pointwise to f. Then for* any *measure μ on $(X, \mathcal{B})$ we have*

$$\lim_{j \to \infty} \int_X f_j \, d\mu \;=\; \int_X f \, d\mu \;\in\; [0, +\infty].$$

An important point here is that the assumptions on the sequence $(f_j)_j$ do not involve a specific choice of measure, only a σ-algebra, while the conclusion then holds for all measures on the σ-algebra.

This theorem often works well in tandem with:

Theorem 1.9 (Monotone approximation theorem)**.** *Suppose $f : (X, \mathcal{B}) \to [0, +\infty]$ is measurable. Then there exists a monotone increasing sequence of simple functions $\varphi_j : (X, \mathcal{B}) \to [0, +\infty]$ converging pointwise to f. That is,*

$$\lim_{j \to \infty} \varphi_j(x) = f(x)$$

for all $x \in X$.

Remark 1.10. In both theorems the assumptions of non-negativity of the functions is essential.

Remark 1.11. By a **simple function** we mean a measurable function $\varphi : (X, \mathcal{B}) \to \mathbb{R} \cup \{+\infty\}$ that takes on only a finite number of values. Equivalently φ is a finite linear sum of characteristic functions:

$$\varphi = \sum_{i=1}^{n} a_i \chi_{A_i}$$

some $A_i \in \mathcal{B}$ and $a_i \in \mathbb{R} \cup \{+\infty\}$ and $n \in \mathbb{N}$.

1.3 Extra structure from a topology

Now suppose that X is a compact metric space and let $\mathcal{B}$ be the induced Borel σ-algebra, i.e. the σ-algebra generated by open sets in X.

The first nice fact is that any *continous* $f : X \to \mathbb{R}$ is automatically measurable. Secondly, if f is continous and m is a *finite* measure on $(X, \mathcal{B})$ then f is m-integrable because continuous functions on compact spaces are bounded.

Definition 1.12. A **regular Borel measure** on X is a Borel measure μ such that

$$\mu(E) \;=\; \inf_{\substack{U \text{ open} \\ E \subset U}} \mu(U) \;=\; \sup_{\substack{K \text{ compact} \\ K \subset E}} \mu(K)$$

for all $E \in \mathcal{B}$. i.e. can be approximated from outside by open sets and inside by compact sets.

One consequence of regularity is the following:

Theorem 1.13. *$C(X, \mathbb{R}) \subset L^1(X, \mathcal{B}, \mu)$ is a dense subset for any regular Borel measure μ on $(X, \mathcal{B})$.*

In other words, for any $f : (X, \mathcal{B}) \to \mathbb{R}$ measurable with $\int_X |f| \, d\mu < \infty$, there exists a sequence $f_j \in C(X, \mathbb{R})$ so that

$$\lim_{j \to \infty} \int_X |f_j - f| \, d\mu \;=\; 0.$$

Sketch Proof. To see where the regularity property of the measure enters let's look at the proof when $X = [0, 1]$. One uses the following four steps: (1) By the monotone convergence and approximation theorems above it suffices to approximate any characteristic function by continous functions in the L^1 sense. (2) By regularity of the measure μ any Borel set $B \in \mathcal{B}$ can be approximated by an open set U, so it suffices to approximate any characteristic function χ_U, with U open, by continous functions, in the L^1 sense. (3) After removing a set of measure ε, U has only a finite number of components, so it suffices to assume U is connected and therefore is just an interval $(a,b) \in [0,1]$. (4) It is now easy to approximate $\chi_{(a,b)}$ by continous functions in the L^1 sense with the following picture:

Insert picture

For a general compact metric space X the argument is similar, see for example [3].

Remark 1.14. The last theorem generalizes in two directions: for any $1 \leq p < \infty$ the inclusion $C(X, \mathbb{R}) \subset L^p(X, \mathcal{B}, \mu)$ is also dense, and is false for $p = \infty$. One can also replace continuous by C^∞-smooth functions.

Definition 1.15. A **Probability measure** on X is a measure m such that $m(X) = 1$ (implicitely X is non-empty). We set

$$\mathcal{M}(X) \ = \ \{\text{regular Borel probability measures on } X\}.$$

Definition 1.16. The **support** of $m \in \mathcal{M}(X)$, denoted $\operatorname{supp}(m)$, is the smallest closed set $C \subset X$ such that $m(C) = 1$.

Equivalently, $x \in \operatorname{supp}(m)$ iff $m(U) > 0$ for all open U containing x.

Definition 1.17 (Vague/weak-$*$ topology)**.** Say that a sequence of probability measures $m_k \in \mathcal{M}(X)$, $k \geq 1$, converges **weak-$*$**, or **vaguely**, to $m \in \mathcal{M}(X)$ if

$$\int_X f \, dm_k \to \int_X f \, dm$$

for $k \to \infty$, for all $f \in C(X)$.

It's a fact that this indeed defines a metrizable and hence Hausdorff topology on $\mathcal{M}(X)$ when X is compact. Moreover, the space $\mathcal{M}(X)$ itself is compact in the vague topology:

Theorem 1.18. *To emphasize, X is a compact metric space. Then the space of regular Borel probability measures $\mathcal{M}(X)$ with the weak-$*$ topology is:*

1. *metrizable,*
2. *compact.*

Let us outline how to prove this using Riesz' representation theorem and Alaoglu's theorem from functional analysis. Precise proofs can be found in [3],[12], [13].

First recall:

- $C(X) := C^0(X, \mathbb{R})$, the space of continuous functions from $X \to \mathbb{R}$, is a Banach space with the supremum norm and is separable when X is compact (i.e. has a countable dense subset).
- A **signed measure** is a function $\mu : \{\sigma\text{-algebra}\} \to \mathbb{R} \cup \{+\infty\}$ satisfying the axioms of a measure, i.e. $\mu(\emptyset) = 0$ and countably additive.
- The **variation** of a signed measure μ is
$$|\mu| \;:=\; \sup \sum_i \mu(E_i)$$
taken over all finite partitions $\{E_i\}_i$ of X. The notion of regular for a signed measure is that $|\mu| < \infty$.

Theorem 1.19 (Riesz representation theorem)**.** *Let X be a compact metric space. Then*

$$C(X)^* \simeq \{\mu \mid \mu \text{ regular signed Borel measure}\}$$

with norm $\|\mu\| = |\mu|$ equal to the variation of μ. The isomorphism is given by

$$l_\mu(f) \;=\; \int_X f\, d\mu.$$

Remark 1.20. We observe:

- This isomorphism identifies the **measures** on X with the **positive linear functionals** on $C(X)$; $l \in C(X)^*$ is positive if $l(f) \geq 0$ whenever $f(x) \geq 0$ for all $x \in X$.
- The isomorphism identifies the **probability measures** on X with the **positive functionals** l **which satisfy** $l(1) = 1 = \|l\|$.

Theorem 1.21 (Alaoglu)**.** *Let V be a separable Banach space. Then the unit ball $B^1_{V^*} := \{\|l\| \leq 1\}$ in the dual space V^* is sequentially weak-$*$ compact.*

Remark 1.22. Recall the comparison of weak versus weak-$*$ convergence in V^*. Consider a sequence $l_k \in V^*$, $k \geq 1$. Then l_k converges **weakly** to $l \in V^*$ if $\lambda(l_k) \to \lambda(l)$, as $k \to \infty$, for all $\lambda \in V^{**}$. In contrast l_k converges **weak-$*$** to $l \in V^*$ if $l_k(v) \to l(v)$, as $k \to \infty$, for all $v \in V$.

Steps in proof of compactness Theorem 1.18. Let

$$B^1 := \left\{ \; l \in C(X)^* \mid \|l\| \leq 1 \; \right\}$$

be the unit ball in $C(X)^*$ (with respect to the dual space norm). Then X compact implies $C(X)$ is separable implies B^1 is metrizable and compact in the weak-$*$ topology.

In particular, metrizable means that we don't have to worry about the distinction between sequential and topological compactness. The Riesz isomorphism takes $\mathcal{M}(X)$ bijectively onto

$$\mathcal{C} := \{l \mid l \geq 0, \; \|l\| = 1, \; l(1) = 1\} \; \subset \; B_1 \tag{1}$$

Moreover, the isomorphism takes weak-$*$ convergent sequences in the space of measures to weak-$*$ convergent sequences in the space of functionals, and therefore $\mathcal{M}(X) \simeq \mathcal{C}$ are homeomorphic with the weak-$*$ topologies. In particular $\mathcal{M}(X)$ is metrizable.

It remains to check that $\mathcal{C} \subset B_1$ is a closed subspace (in the weak-$*$ topology) and it will follow from Alaoglu that $\mathcal{C}$ is compact. We examine the conditions defining the right hand side of (1). The conditions $l \geq 0$ and $l(1) = 1$ are pointwise and therefore closed under weak-$*$ convergence. The condition $\|l\| = 1$ alone is not closed under weak-$*$ convergence, apriori the norm can drop under weak-$*$ limits. However $l(1) = 1$ together with $\|l\| \leq 1$ implies $\|l\| = 1$. This ends our sketch of the proof.

Lemma 1.23. *The space $\mathcal{M}(X)$ is convex. That is, $m_1, m_2 \in \mathcal{M}(X)$ implies $\alpha m_1 + (1-\alpha)m_2 \in \mathcal{M}(X)$ for all $\alpha \in [0,1]$.*

Proof. From the definitions. ■

1.4 Exploiting convexity

We state two general results from functional analysis which apply to compact convex subsets of locally convex topological vector spaces. These apply to the probability measures $\mathcal{M}(X)$ which we can view as a subset of $C(X)^*$ via the Riesz isomorphism.

Suppose V is a topological vector space. $C \subset V$ is convex.

Definition 1.24.

- An **extreme point** of C is a point $x \in C$ such that $C \backslash \{x\}$ is still convex. E.g. the extreme points of a closed convex polygon are the vertices. (Not the whole boundary).
- Let $E \subset V$ be an arbitrary subset. The **convex hull** of E, denoted $\mathrm{co}(E)$, is the set of all finite convex combinations of points in E:
$$\mathrm{co}(E) := \left\{ \sum_i^n \alpha_i e_i \;\middle|\; e_i \in E,\ \alpha_i \in [0,1],\ \sum_i \alpha_i = 1,\ 1 \leq n < \infty \right\}.$$
- It's an easy exercise to check that an equivalent definition of $\mathrm{co}(E)$ is the intersection of all convex C such that $E \subset C \subset V$.

Finite dimensions. Let $C \subset \mathbb{R}^n$ be non-empty, compact, convex. Intuitively:

A. $\mathrm{ex}(C) \neq \emptyset$.

B. every point $x \in C$ is a finite convex sum of extreme points.

This is true, infact a theorem of Minkowski states that in $\mathbb{R}^n$ every point x in C can be written as a convex sum of $n+1$ extreme points. For example the centre point of an equilateral triangle in R^2. One can think of the Krein-Milman theorem below as a generalisation of A and B to infinite dimensions.

Infinite dimensions. For example, if V is a normed vector space (possibly $\dim V = +\infty$) then the open and closed unit balls

$$B_1 = \{v \in V \mid \|v\| \leq 1\} \qquad \dot{B}_1 = \{v \in V \mid \|v\| < 1\}$$

are convex (triangle inequality) and $\mathrm{ex}(\dot{B}_1) = \emptyset$ and $\mathrm{ex}(B_1) \subset \{v \in V \mid \|v\| = 1\}$. In infinite dimensions it is possible that even $\mathrm{ex}(B_1) = \emptyset$. For example, the closed unit ball in $L^1([0,1])$ has no extremal points[*], while in L^p for $1 < p < \infty$ every point in the unit sphere is an extremal point!

Remark 1.25. In the two theorems we wish to state below, Krein-Milman and Choquet, it is convenient to state them in a more general context than just a normed vector space or Banach space (because the weak-$*$ topology does not come from a norm). Therefore we will use the notion of a **locally convex topological vector space (LCS)** without recalling exactly what this means (see for example [3]). All that is important for us is that, for a Banach space V, the weak-$*$ topology on V^* gives it the structure of a locally convex topological vector space.

Theorem 1.26 (Krein-Milman)**.** *Let V be a LCS. Then for any non-empty compact convex subset $C \subset V$ we have:*

[*] It suffices to show that any $f \in L^1$ with $\|f\|_1 = 1$ is a non-trivial convex sum of two elements $g, h \in B_1$. To do this simply partition the domain $[0,1]$ into two pieces I_1, I_2 (measurable, e.g. intervals) and consider the restriction of f to I_1 and I_2. Choose the partition so that both restricted functions have L^1 norm strictly less than 1 (e.g. 1/2), then normalize both functions to have L^1 norm equal to 1. Then f is a (non-trivial) convex sum of the resulting two functions.

1. *The set of extreme points is non-empty:* $\mathrm{ex}(C) \neq \emptyset$*.*

2. *C equals the closure of the convex hull of its extremal points. That is,* $C = \overline{\mathrm{co}}(\mathrm{ex}(C))$*.*

For a proof and further discussion and applications see Conway's book [3]. Note that one nice application is the following: the closed unit ball in the dual of a Banach space is a non-empty convex and compact subset with respect to the weak-$*$ topology. Since the weak-$*$ topology is also a LCS we can apply Krein-Milman and conclude that the closed unit ball in a dual Banach space always has many extremal points. By the footnote earlier, the closed unit ball in $L^1[0,1]$ has no extremal points. We must conclude that there exists no Banach space V for which $V^* = L^1[0,1]$.

In case one is tempted to think that C compact implies $\mathrm{ex}(C)$ is compact, we note that this need not be true even in finite dimensions:

Insert example where C =two cones in $\mathbb{R}^3$

Here is an even stronger generalisation of properties A and B to infinite dimensions.

Theorem 1.27 (Choquet)**.** *Let V be a LCS and $C \subset V$ a non-empty metrizable compact convex subset. Then for each $x \in C$ there exists a regular Borel probability measure $\mu = \mu_x$ on C, such that*

1. $\mathrm{supp}(\mu) \subset \mathrm{ex}(C)$*,*

2. *and*
$$l(x) = \int_C l(x')\, d\mu(x') \tag{2}$$
for every $l \in V^*$.

A proof of Choquet is in [14].

Remark 1.28.

- How should we think of (2)? We can think of it as giving meaning to the relation
$$x = \int_C x' \, d\mu(x'). \tag{3}$$
The right hand side of (3) is some kind of integral of a function with values in a locally convex topological vector space and requires some justification, while (2) is just the usual notion of integral for real valued functions. We can think of (3) as expressing x as an *infinite* convex sum of extremal points of C. Indeed, if x happens to be a finite sum of extremal points: $x = a_1e_1 + \ldots + a_ne_n$, where $e_i \in \mathrm{ex}(C)$ and $a_i \in [0,1]$ with $a_1 + \ldots + a_n = 1$, then formally (3) holds with
$$\mu = \sum_{i=1}^{n} a_i \delta_{e_i}$$
where δ_{e_i} is the delta measure (or "point mass") at e_i.

- The advantage of Choquet over Krein-Milman is that the measure μ is supported in $\mathrm{ex}(C)$ and not just in the closure $\overline{\mathrm{ex}}(C)$. This comes at the expense of assuming C is metrizable, without which it is possible $\mathrm{ex}(C)$ is not a Borel set. But in our applications $C = \mathcal{M}(X)$, for some compact metric space X, and we saw that this is metrizable in theorem 1.18.

- In general the measure μ_x in the Choquet theorem will not be unique. (Think of a convex set in $\mathbb{R}^2$ having lots of extremal points such as the closed unit ball w.r.t. the Euclidean norm, and take any x with $|x| < 1$; each straight line through x gives us a different way to write x as a convex sum of two points on the boundary.)

2 Invariant measures and the Birkhoff ergodic theorem

Now we begin the course in earnest.

The starkest definition of a (discrete) dynamical system is a set X and a transformation $T : X \to X$, and one is interested in how points and subsets of X get moved around the space by T. Even if T itself can be described in very simple terms the behavior of orbits of points

$$x,\ Tx,\ T^2x,\ \dots\ T^nx,$$

or more generally of subsets $A \subset X$,

$$A,\ TA,\ T^2A,\ \dots\ T^nA$$

can be so complicated as to be seemingly impossible to describe.

In ergodic theory the basic idea is to examine the orbits of T on (probability) measures $\mathcal{M}(X)$ on X. We can think of measures as a generalization of the points and subsets of X (there's a natural inclusion $X \hookrightarrow \mathcal{M}(X)$ taking x to its point mass δ_x). For example suppose T has a periodic point $x \in X$ such that

$$T^nx = x$$

for some $n \in \mathbb{N}$. Then it's easy to see that there is a (unique) T-invariant probability measure μ supported on the orbit of x, namely

$$\mu = \frac{1}{n}\Big(\delta_x + \delta_{Tx} + \cdots + \delta_{T^{n-1}x}\Big).$$

It turns out, under very general assumptions, that even if T has no periodic points at all it will still have invariant probability measures and often these tell us interesting things about the long term behavior of T.

To explore this approach seriously we need some structure on our space X and on our transformation T.

2.1 Invariant measures

Throughout, unless stated otherwise, $X =$ a compact metric space, and $\mathcal{B} =$ the Borel σ-algebra on X. Then we say $T : X \to X$ is a **measurable transformation** if

$$T^{-1}B \in \mathcal{B}$$

for all $B \in \mathcal{B}$. Let

$$\mathcal{M}(X) = \{\text{regular probability measures on } (X, \mathcal{B})\}.$$

Then T determines a map

$$T^* : \mathcal{M} \to \mathcal{M}$$

by $(T^*m)(B) = m(T^{-1}B)$.

Lemma 2.1 (Change of variables)**.** *Suppose $T : X \to X$ is a measurable transformation, $f : (X, \mathcal{B}) \to \mathbb{R} \cup \{+\infty\}$ is a measurable function, and $m \in \mathcal{M}(X)$. Then, $f \circ T \in L^1(X, \mathcal{B}, m)$ if and only if $f \in L^1(X, \mathcal{B}, T^*m)$, and in either case we have the relation*

$$\int f \circ T \, dm \;=\; \int f \, d(T^*m). \tag{4}$$

Proof. Any measurable f can be written as a difference $f = f_+ - f_-$ of two *positive* (meaning non-negative) measurable functions, and relation (4) is linear in f. It therefore suffices to prove that for any measurable $f : (X, \mathcal{B}) \to [0, +\infty]$ we have

$$\int f \circ T \, dm \;=\; \int f \, d(T^*m) \;\in\; [0, +\infty] \tag{5}$$

meaning that one side is finite iff both sides are finite and equal. Notice that it's much easier to work with positive measurable functions because their integrals are always well defined - even if infinite.

To prove (5) we follow the usual steps: (1) verify for $f = \chi_B$, $B \in \mathcal{B}$, (2) by linearity for $f = \varphi$ a simple function, (3) by the monotone approximation theorem and the monotone convergence theorem (see Section 1) we can take limits and extend to $f : (X, \mathcal{B}) \to [0, +\infty]$ measurable. Check the details. ∎

The fixed points of T^* are of particular interest:

Definition 2.2. A probability measure m on $(X, \mathcal{B})$ is T**-invariant** if $m(T^{-1}B) = m(B)$ for all $B \in \mathcal{B}$. Equivalently $T^*m = m$.

We will denote the set of all T-invariant probability measures by

$$\mathcal{M}_{\text{inv}}(X, \mathcal{B}, T)$$

or just $\mathcal{M}_{\text{inv}}$ when the rest is clear.

Lemma 2.3 (Characterizing invariant measures)**.** *For $T : X \to X$ measurable and $m \in \mathcal{M}(X)$, the following are equivalent.*

1. $m \in \mathcal{M}_{\text{inv}}(X, T)$.
2. *For all* $f \in C^0(X, \mathbb{R})$
$$\int f \circ T \, dm \;=\; \int f \, dm. \tag{6}$$
3. *For all* $f : (X, \mathcal{B}) \to \mathbb{R} \cup \{+\infty\}$ *measurable both of the following hold:*
   (a) $f \in L^1(X, \mathcal{B}, m) \iff f \circ T \in L^1(X, \mathcal{B}, m)$.
   (b) *For all* $f \in L^1(X, \mathcal{B}, m)$,
$$\int f \circ T \, dm \;=\; \int f \, dm. \tag{7}$$

Proof.
3. $\implies$ ***2.*** Because a continuous function f on X is bounded, so $f \circ T$ is bounded, and so both are in $L^1(X, \mathcal{B}, m)$.

2. $\implies$ ***1.*** Let $B \in \mathcal{B}$. Since m is regular we can approximate B by an open set U. By suitably approximating χ_U by continuous functions in the L^1 sense we conclude that (6) holds for $f = \chi_U$ and then for $f = \chi_B$. This means that $m(T^{-1}B) = m(B)$.

1.* $\Longrightarrow$ *3. Suppose $m \in \mathcal{M}_{\text{inv}}(X,T)$. Since every measurable function is a difference of two positive measurable functions, and (7) is linear in f, *3.(a)* and *3.(b)* together are equivalent to the following: for every measurable $f : (X, \mathcal{B}) \to [0, +\infty]$ we have

$$\int f \circ T \, dm \; = \; \int f \, dm \; \in \; [0, +\infty] \tag{8}$$

meaning that one side is finite iff both sides are finite and equal. To prove (8) we use

$$\int f \circ T \, dm \; = \; \int f \, d(T^* m) \; = \; \int f \, dm \; \in \; [0, +\infty]$$

where the first equality is from lemma 2.1, and the second because $T^* m = m$. ■

Example 2.4. $X = T^2 = \mathbb{R}^2/\mathbb{Z}^2$ the 2-torus, and $T : T^2 \to T^2$ is

$$T(x, y) = (x + \alpha \mod 1,\ y + \beta \mod 1).$$

Clearly the "Lebesgue" measure m inherited from $\mathbb{R}^2$ is an invariant (regular Borel) probability measure. Suppose $\alpha \neq 0$. Then,

- β/α =irrational $\Longrightarrow$ no further invariant measures.
- β/α =rational $\Longrightarrow$ many invariant measures. (Every point is periodic, each periodic orbit supports an invariant measure. Also each line in $\mathbb{R}^2$ with slope β/α descends to an invariant closed curve in T^2, which will also support an invariant measure. The invariant annulus inbetween two invariant closed curves supports an invariant measure, etc.)

Example 2.5 (Bernoulli processes)**.** Fix a probability vector $p = (p_1, \ldots, p_s)$ of length $s \in \mathbb{N}$, meaning each $p_i \in [0, 1]$ and $\sum_i p_i = 1$. Set

$$X \; := \; \left\{ \underline{x} = (x_i)_{i \in \mathbb{N}} \,\middle|\, x_i \in \{1, \ldots, s\} \right\}$$

be the space of one-sided sequences in $\{1, \ldots, s\}$, and let $T : X \to X$ be the shift operator

$$T(x_1, x_2, x_3 \ldots) = (x_2, x_3, \ldots).$$

Give X the product topology from the discrete topology on $\{1, \dots, s\}$, and let $\mathcal{B}$:= the Borel σ-alegbra. (Note that this is the topology of a metric d that makes X a compact metric space, and T is continuous.) There is a unique Borel measure m on $(X, \mathcal{B})$ which satisfies

$$m\Big(\{\, \underline{x} \mid x_1 = i \,\}\Big) \;=\; p_i$$

for $i = 1, \dots, s$, and

$$m\Big(\{\, \underline{x} \mid x_1 = i,\ x_2 = j \,\}\Big) \;=\; p_i p_j$$

etc. These are the so called "cylinder sets", they form a semi-algebra $\mathcal{S}$ meaning that $X, \emptyset \in \mathcal{S}$ and $\mathcal{S}$ is closed under finite intersections and for all $A \in \mathcal{S}$ we can write A^c as a finite disjoint union of elements of $\mathcal{S}$. For example, the collection of all intervals in $\mathbb{R}$ forms a semi-algebra but is not an algebra since it's not closed under finite unions. One checks that m is additive under finite disjoint unions and therefore extends uniquely to the whole Borel σ-algebra $\mathcal{B}$ by Caratheodory's extension theorem, see Theorem 1.3. m is called the **Bernoulli** $(p_1, \dots, p_s)$**-measure**.

m is a T-invariant measure, for example consider

$$T^{-1}\Big(\{\, \underline{x} \mid x_1 = i \,\}\Big) \;=\; \{\, \underline{x} \mid x_2 = i \,\}.$$

Let us verify that the measure of $A := \{\, \underline{x} \mid x_2 = i \,\}$ is indeed p_i. A is not a cylinder set, but

$$A = \bigcup_{l=1}^{s} \{\, \underline{x} \mid x_1 = l,\ x_2 = i \,\}$$

is a disjoint union of cylinder sets, so

$$m(A) \;=\; \sum_{l=1}^{s} p_l p_i \;=\; p_i$$

as we require.

Question: when does a transformation T have invariant measures in general?

Exercise 2.6. Show that the following dynamical system has no invariant probability measures: $T : [0,1] \to [0,1]$ given by

$$T(x) = \begin{cases} 1 & \text{if } x = 0 \\ x/2 & \text{if } x \neq 0. \end{cases}$$

In this example T is not continuous. We will prove in a moment that T *must* have an invariant probability measure if it is continuous. First a lemma.

Lemma 2.7. *If $T : X \to X$ is continuous then the induced map $T^* : \mathcal{M}(X) \to \mathcal{M}(X)$ is continous in the weak-$*$ topology.*

Proof. Consider a sequence $m_j \in \mathcal{M}(X)$, $j \geq 1$, converging to $m \in \mathcal{M}(X)$. We wish to show that $T^*m_j \to T^*m$.

Fix $f \in C(X)$. Then $f \circ T$ is also continous, so $m_j \to m$ implies

$$\int_X f \circ T \, dm_k \to \int_X f \circ T \, dm \qquad \text{as } k \to \infty.$$

By lemma 2.1 this is equivalent to

$$\int_X f \, d(T^*m_k) \to \int_X f \, d(T^*m) \qquad \text{as } k \to \infty.$$

Hence by definition $T^*m_j \to T^*m$. ■

Remark 2.8. It looks like the converse is also true. Check!

Theorem 2.9. *(Existence of invariant measures) Let X be a non-empty compact metric space. Then a continuous map $T : X \to X$ has at least one invariant probability measure. More precisely, $\mathcal{M}_{\text{inv}}(X) \neq \emptyset$.*

Proof. Recall that $\mathcal{M}(X)$, the set of regular Borel probability measures on X, is non-empty and compact with the weak-$*$ topology/convergence. Moreover it is convex, so starting from any $m \in \mathcal{M}(X)$ each convex sum

$$m_k \ := \ \frac{1}{k}\Big(m + T^*m + (T^*)^2m + \cdots + (T^*)^{k-1}m\Big)$$

is an element in $\mathcal{M}(X)$. By compactness we find a weak-$*$ convergent subsequence

$$m_{k_j} \to \hat{m}$$

to some $\hat{m} \in \mathcal{M}(X)$. It is now easy to check that $T^*\hat{m} = \hat{m}$. Indeed, by continuity of T^* (lemma 2.7)

$$T^*\hat{m} = \lim_{j\to\infty} T^* m_{k_j}.$$

And,

$$\begin{aligned}
T^* m_{k_j} &= \frac{1}{k_j}\Big(T^*m + (T^*)^2 m + \cdots + (T^*)^{k_j} m\Big) \\
&= \frac{1}{k_j}\Big(m + T^*m + (T^*)^2 m + \cdots + (T^*)^{k_j-1} m\Big) \\
&\qquad\qquad + \frac{1}{k_j}\Big(-m + (T^*)^{k_j} m\Big).
\end{aligned}$$

Sending $j \to \infty$ this converges to $\hat{m} + 0$. Thus $T^*\hat{m} = \hat{m}$. ■

Exercise 2.10. Assume $(X, \mathcal{B})$ is a compact metric space equipped with the Borel σ-algebra. $T : X \to X$ is a continuous transformation.

1. If T is a homeomorphism show that $\text{supp}(m)$ is a T-invariant set for each $m \in \mathcal{M}_{\text{inv}}$.

2. If T is only continuous prove the same or find a counterexample.

What can we do with invariant measures?

Theorem 2.11 (Poincaré recurrence)**.** *Let $(X, \mathcal{B})$ be an arbitrary measurable space, and $T : X \to X$ a measurable transformation with an invariant probability measure m. Then for any set $B \in \mathcal{B}$ almost all points $x \in B$ return to B under some (positive) iterate of T.*

Remark 2.12. Note that “almost all” in this statement cannot be improved to “all”. Consider for example the following Hamiltonian system (modelling the motion of a pendulum) on the cylinder $C = \mathbb{R} \times \mathbb{R}/2\pi\mathbb{Z}$. $H : C \to \mathbb{R}$ is the smooth map $H(x, y) = y^2 - \cos(x)$. The corresponding Hamiltonian vector field X_H on $\mathbb{R}^2$ is defined by

$$X_H(x, y) = J\nabla H(x, y)$$

where J is the 2×2 matrix

$$i = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

that rotates vectors in $\mathbb{R}^2$ anti-clockwise by 90 degrees. The flow generated by X_H is the 1-parameter family of diffeomorphisms $\phi^t : C \to C$, $t \in \mathbb{R}$, satisfying $\phi^0 = \mathrm{id}$ and for each $(x, y) \in \mathbb{R}^2$ solving the equation

$$\frac{d}{dt}\phi^t(x, y) = X_H\big(\phi^t(x, y)\big).$$

Each path $t \mapsto \phi^t(x, y)$ remains in a level curve of H. A theorem due to Liouville states that for any smooth H the induced diffeomorphisms will preserve the area form $dx \wedge dy$ on C. We can easily convert this into an invariant probability measure by setting

$$m(E) \ := \ \int_E f(H(x, y))\, dx \wedge dy$$

for an appropriate smooth function $f : \mathbb{R} \to [0, \infty)$ chosen so that $m(C) = 1$. The level curves of H look like:

Let $p \in C$ be a point on one of the homoclinic connections, and B be any small ball about p. Then the path $t \mapsto \phi^t(p)$ never returns to B for $t > 0$.

However, for any point $q \in B$ that does not lie on a homoclinic connection the path $t \mapsto \phi^t(q)$ will return to B for infinitely many values of $t > 0$. All of this fits in with Poincaré's recurrence theorem, because the two homoclinic connections fill up a set of measure zero with respect to the measure m.

Proof. (Of Poincaré's recurrence theorem) Fix $B \in \mathcal{B}$ and let F be the points in B that "never return":

$$F = \left\{ b \in B \;\middle|\; T^n b \notin B \text{ for all } n \geq 1 \right\}.$$

We wish to show $m(F) = 0$.

Observe that for each $n \geq 1$, $T^{-n}F \cap B = \emptyset$ by choice of F. In particular,

$$T^{-n}F \cap F = \emptyset$$

for all $n \geq 1$ (as $F \subset B$). Applying inverses of T to both sides we find that the sequence of sets $T^{-n}F$ for $n \geq 1$ are pairwise disjoint. Since they must all have equal measure and $m(X) = 1 < \infty$ it follows that $m(F) = 0$. ■

Now we turn our attention to a particularly important type of invariant measure, namely the ergodic ones.

2.2 Ergodic invariant measures

To motivate this important definition let's consider the following example.

Example 2.13. $X = D =$ closed unit disk in $\mathbb{R}^2$. $T : D \to D$ is rotation through irrational angle $\alpha \in \mathbb{R} \backslash \mathbb{Q}$:

$$Tz \;=\; e^{2\pi i \alpha} z.$$

Clearly $m_{\text{leb}} := \frac{1}{\pi}$(Lebesgue measure) is an invariant (regular Borel) probability measure on D. The fixed point at the origin supports an invariant delta measure. The disk is filled by invariant circles $C_r := \{z \;\mid\; |z| = r\}$. Each C_r also supports an invariant measure in $\mathcal{M}$.

Clearly we can find many more invariant probability measures, for example on the the disk-like regions enclosed by C_r, the annular-like regions imbetween, etc. However, one gets the feeling that all further invariant measures we can cook up are merely "built up" out of the invariant measures sitting on the circles $\{C_r\}_{r\in(0,1]}$ and the invariant point mass at the origin. Moreover, we get the feeling that we cannot "build" one of these measures merely from the others, except perhaps by taking limits. That is, they are in some sense indecomposable.

What is special about the measures supported on circles in this example? The abstract answer is contained in the next definition.

Definition 2.14 (Ergodicity)**.** An invariant measure $m \in \mathcal{M}_{\text{inv}}(X)$ is said to be **ergodic** (w.r.t. T) if whenever $T^{-1}B = B$ we have $m(B) = 0$ or $m(B) = 1$.

We will denote the set of ergodic invariant measures in $\mathcal{M}(X)$ by

$$\mathcal{M}_{\text{erg}}(X) \ := \ \left\{ \ m \in \mathcal{M}_{\text{inv}}(X) \ \middle| \ m \text{ is ergodic w.r.t. } T \right\}$$

Example 2.15. Find (without details) the ergodic measures in example 2.13.

Exercise 2.16. Show that m is ergodic if and only if it satisfies the following apriori weaker condition: every $B \in \mathcal{B}$ for which $m(T^{-1}B \,\Delta\, B) = 0$ has either $m(B) = 0$ or $m(B) = 1$. (Where $A \,\Delta\, B$ denotes the symmetric difference.)

Here is a cute statement that uses the notion of an ergodic invariant measure, that strengthens the Poincaré Recurrence theorem. It is unusual in that it does not require one of the "ergodic theorems".

Let $(X, \mathcal{B})$ be an arbitrary measurable space and $T : X \to X$ a measurable transformation with an invariant measure $m \in \mathcal{M}_{\text{inv}}(X)$. Fix any set $B \in \mathcal{B}$ with $m(B) > 0$ and define the "first return time function"

$$r_B : B \to \mathbb{N} \cup \{+\infty\}$$

to take each point $x \in B$ to the minimal $n \geq 1$ for which $T^n x \in B$. Note that r_B is measureable (check). The Poincaré recurrence theorem is equivalent to the statement that r_B is finite almost everywhere (for every $B \in \mathcal{B}$). The following argument shows that infact r_B is also integrable. If we assume moreover that m is *ergodic* then we obtain a stronger quantitative statement:

Theorem 2.17 (Kac's formula). *If $m \in \mathcal{M}_{\text{erg}}(X, T)$ then*

$$\int_B r_B \, dm \;=\; 1.$$

In other words the expected first return time for a point starting in B is $1/m(B)$.

Proof. Let us prove this under the additional assumption that T is invertible. For the proof without assuming invertibility see Sarig's notes [17] (Incidentally, Sarig references Royden's book on Real Analysis).

Set X_0 equal to the collection of points which came from B at some point in the past and which return to B at some point in the present or future. More precisely

$$X_0 := \left\{ \, x \in X \;\middle|\; \exists n \geq 0 \text{ s.t. } T^n x \in B \;\text{ and }\; \exists m \geq 1 \text{ s.t. } T^{-m} x \in B \, \right\}.$$

Then clearly $T^{-1} X_0 = X_0$ and $m(X_0) \geq m(B) > 0$ so by ergodicity $m(X_0) = 1$.

We partition B according to first return time. Set

$$B_n := \left\{ \, x \in B \;\middle|\; r_B(x) = n \, \right\}.$$

Then

$$\begin{aligned}
\int_B r_B\, dm &= \sum_{n=1}^{\infty} \int_{B_n} r_B\, dm \\
&= \sum_{n=1}^{\infty} n\, m(B_n) \\
&= \sum_{n=1}^{\infty} \Big(m(TB_n) + m(T^2B_n) + \cdots + m(T^nB_n) \Big) \\
&= \sum_{n=1}^{\infty} \sum_{j=1}^{n} m(T^jB_n).
\end{aligned}$$

We claim that the sets $\{T^jB_n\}_{j,n}$ in the double sum are disjoint. Before examining this, it is clear that the union is equal to X_0. Indeed, the union of $\{T^jB_n\}_{j,n}$ is precisely those points $x \in X$ for which $x = T^kb$ some $b \in B$, $k \geq 1$, and for which $T^lx \in B$, some $l \geq 0$. Thus the double sum is $\leq m(X_0) = 1$ with equality if the sets in the double sum are pair wise disjoint. It only remains then to show that the sets $\{T^jB_n\}_{j,n}$ in the double sum are disjoint.

Arguing indirectly suppose that there exist integers $1 \leq n_1 \leq n_2$ and $1 \leq j_1 \leq n_1$ and $1 \leq j_2 \leq n_2$ with $(j_1, n_1) \neq (j_2, n_2)$ such that

$$T^{j_1}B_{n_1} \cap T^{j_2}B_{n_2} \neq \emptyset.$$

Consider a point p in this intersection. Then the minimal $k \geq 0$ such that $T^kp \in B$ is both $n_1 - j_1$ and $n_2 - j_2$. Thus $n_1 - j_1 = n_2 - j_2$. On the other hand, the minimal $k \geq 1$ such that $T^{-k}p \in B$ is both j_1 and j_2. Thus $j_1 = j_2$. We conclude that $(j_1, n_1) = (j_2, n_2)$ after all. ■

Example 2.18. We include the following trivial example to illustrate a difference between using preimages $T^{-1}E$ instead of images TE when T is not invertible: take $T : [0, 1) \to [0, 1)$ to be

$$Tx = 4x \mod 1.$$

Let m_{leb} denote Lebesgue measure on $[0,1)$. Then the pushforward of T does not preserve m_{leb} due to the "stretching" factor of 4. However the pullback of T *does* preserve m_{leb}. That is, for every Borel set $E \subset [0,1)$

$$m_{\text{leb}}\big(T^{-1}E\big) = m_{\text{leb}}\big(E\big).$$

See the figure.

Before proving that ergodic measures exist, let us note the following reformulation of ergodicity for later.

Lemma 2.19 (Characterizing ergodic measures)**.** *The following are equivalent.*

1. $m \in \mathcal{M}_{\text{erg}}$.
2. *The only invariant measurable functions are constant almost everywhere. i.e. for all* $f : (X, \mathcal{B}) \to \mathbb{R} \cup \{+\infty\}$ *measurable,*

$$f \circ T = f \quad m\text{-}a.e. \quad \Longrightarrow \quad f \text{ is constant } m\text{-}a.e. \tag{9}$$

Proof.
2. $\Longrightarrow$ ***1.*** Suppose $B \in \mathcal{B}$ is invariant, i.e. $T^{-1}B = B$. Then $f \circ T = f$ where $f = \chi_B$. Therefore χ_B is constant m-a.e. which means that $m(B) = 0$ or $m(B) = 1$. Since $B \in \mathcal{B}$ was arbitrary, m is ergodic.

1. $\Longrightarrow$ ***2.*** Suppose $m \in \mathcal{M}_{\text{erg}}$. Let $f : (X, \mathcal{B}) \to \mathbb{R} \cup \{+\infty\}$ be a fixed measurable function with

$$f \circ T = f \quad m\text{-a.e.}$$

We wish to show that f is constant m-a.e. Arguing indirectly we find some $c \in \mathbb{R}$ such that

$$B_c \;=\; \left\{\; x \in X \;\middle|\; f(x) \geq c \;\right\} \tag{10}$$

satisfies $0 < m(B_c) < 1$. However B_c is m-“almost” T-invariant. More precisely

$$T^{-1}B_c \;=\; \{\; x \in X \;\big|\; f(Tx) \geq c \;\}. \tag{11}$$

Since $f(Tx) = f(x)$ outside of a null set, comparing (10) and (11) we conclude that $m(T^{-1}B_c \,\Delta\, B_c) = 0$. Thus by ergodicity (see exercise2.16) $m(B_c) = 0$ or $m(B_c) = 1$ giving a contradiction. ■

Corollary 2.20. *If T is continuous then the non-empty set of invariant measures $\mathcal{M}_{\mathrm{inv}}(X)$ besides being convex is also compact in the weak-$*$ topology.*

Proof. $T : X \to X$ continuous implies $T^* : \mathcal{M}(X) \to \mathcal{M}(X)$ continuous implies that the set of fixed points $\mathcal{M}_{\mathrm{inv}}(X,T)$ of T^* is a closed subset of $\mathcal{M}(X)$. Thus compactness of $\mathcal{M}_{\mathrm{inv}}(X,T)$ follows from compactness of $\mathcal{M}(X)$. Convexity is also obvious. ■

Thus by Krein-Milman $\mathcal{M}_{\mathrm{inv}}(X)$ has extremal points (if T is continuous). What are they? It turns out that these are the ergodic invariant measures. To prove this we will use:

Lemma 2.21. *Let $(X, \mathcal{B})$ be an arbitrary measurable space and $T : X \to X$ a measurable transformation. If $\nu, \mu \in \mathcal{M}_{\mathrm{inv}}$ and $\nu << \mu$ then the Radon-Nikodym derivative $f = d\nu/d\mu$ is T-invariant in the sense that*

$$f \circ T = f \tag{12}$$

holds μ-almost everywhere.

Recall that $f : (X, \mathcal{B}) \to [0, \infty)$ is a measurable function such that

$$\nu(B) \;=\; \int_B f \, d\mu \tag{13}$$

for all $B \subset \mathcal{B}$.

Proof. Consider the sublevel sets

$$B_c := \left\{ x \,\middle|\, f(x) < c \right\}$$

for $c \in \mathbb{R}$. Clearly (12) holds if $\mu(T^{-1}B_c \,\Delta\, B_c) = 0$ for all $c \in \mathbb{R}$. Since $B_c = \emptyset$ for all $c \le 0$ we need only consider $c > 0$. So fix $B = B_c$ for some $c > 0$. We have $T^{-1}B \,\Delta\, B = T^{-1}B \backslash B \cup B \backslash T^{-1}B$. Observe that

$$\mu\big(T^{-1}B \backslash B\big) = \mu\big(B \backslash T^{-1}B\big). \tag{14}$$

Indeed $\mu(T^{-1}B \backslash B) = \mu(T^{-1}B) - \mu(T^{-1}B \cap B)$ while $\mu(B \backslash T^{-1}B) = \mu(B) - \mu(B \cap T^{-1}B)$ and the T-invariance of μ. Now let us explain from the figure why “obviously“ $\mu(T^{-1}B \,\Delta\, B) = 0$. (It is straight forward to convert the picture into a string of rigorous equations, but hopefully easier to remember this way.)

In view of (13) the area under the graph of f over a region $E \in \mathcal{B}$ is $\nu(E)$. Thus $T^*\nu = \nu$ means that the area (under f) over B equals the area over $T^{-1}B$. Thus the regions Ω_1 and Ω_2 have equal areas in $X \times \mathbb{R}$, that is

$$\int_{T^{-1}B \backslash B} f \, d\mu \;=\; \int_{B \backslash T^{-1}B} f \, d\mu.$$

However in the one region $f \ge c$ while in the other $f < c$, so from (14) the only possibility is that both areas vanish, that is both integrals are zero. Since f is strictly positive on the left hand side this means $\mu(T^{-1}B \backslash B) = 0$. Thus by (14) $\mu(B \backslash T^{-1}B) = 0$ also and we are done. ■

Proposition 2.22 (Extremal invariant measures are ergodic)**.** *The extremal points of $\mathcal{M}_{\text{inv}}(X)$ are precisely the ergodic invariant measures $\mathcal{M}_{\text{erg}}(X)$.*

Proof.
One direction: $m \notin \mathcal{M}_{\text{erg}} \implies m \notin \text{ex}\big(\mathcal{M}_{\text{inv}}\big)$.

If m is not ergodic there exists $B \in \mathcal{B}$ with $T^{-1}B = B$ and $0 < m(B) < 1$. Define measures m_1, m_2 by

$$m_1(E) := \frac{m(E \cap B)}{m(B)} \qquad\qquad m_2(E) := \frac{m(E \cap B^c)}{m(B^c)}$$

for all $E \in \mathcal{B}$. Then m_1, m_2 are T-invariant probability measures and $m = m(B)m_1 + (1 - m(B))m_2$. Thus m is not an extremal point.

The other direction: $m \notin \text{ex}\big(\mathcal{M}_{\text{inv}}\big) \implies m \notin \mathcal{M}_{\text{erg}}$.

Since m is not extremal we can write

$$m \;=\; \alpha m_1 \;+\; (1-\alpha)m_2$$

for some $m_1, m_2 \in \mathcal{M}_{\text{inv}}$ both *distinct from* m, and some $0 < \alpha < 1$. Observe that $m(N) = 0$ implies $m_1(N) = 0$ so that $m_1 << m$. Thus by Radon-Nikodym there exists a measurable "density" function $f : X \to [0, \infty)$ so that

$$m_1(E) \;=\; \int_E f\, dm$$

for all $E \in \mathcal{B}$. By lemma 2.21 f is T-invariant m-a.e.

Claim: f is not constant m-a.e.

From the claim we will be done because we will have produced a non-constant T-invariant function (everything m-a.e.) which by lemma 2.19 means m is not ergodic. We prove the claim indirectly: suppose f is constant m-a.e. Then the constant must be 1 since

$$\int_X f_1\, dm \;=\; m_1(X) \;=\; 1.$$

But $f = 1$ m-a.e. $\implies m_1 = m$ which is impossible. ■

Corollary 2.23 (Existence of ergodic measures)**.** $\mathcal{M}_{\text{erg}}(X, T) \neq \emptyset$ *if* $T : X \to X$ *is a continuous transformation on a compact metric space.*

Proof. Since $\mathcal{M}_{\text{inv}}$ is a non-empty compact convex set, by Krein-Milman it has extremal points. We just saw that the extremal points are ergodic invariant measures. ■

Remark 2.24. It also follows from Krein-Milman that the invariant (probability) measures are equal to the closure of the convex hull of the ergdodic invariant measures. Therefore there is only one invariant measure iff there is only one extremal point iff there is only one ergodic invariant measure. A transformation T is called **uniquely ergodic** if it has precisely one invariant measure, (equivalently exactly one ergodic invariant measure). For example the discrete version of any irrational flow on a torus is uniquely ergodic with Lebesgue being the unique invariant measure. Uniquely ergodic transformations enjoy some surprisingly strong properties and have stronger relations to the topology of the space. For example, if the topology of X forces every continuous T to have at least two fixed points then as any such T would support at least two invariant measures (delta's at the fixed points) it could not be uniquely ergodic.

Applying Choquet instead of Krein-Milman yields the *Ergodic Decomposition Theorem.* (Which in particular captures precisely our discussion of the irrational rotation in example 2.13.)

Theorem 2.25 (Ergodic decomposition theorem)**.** *Let $(X, \mathcal{B})$ be a compact metric space with the Borel σ-algebra, and $T : X \to X$ a continous transformation. Given $m \in \mathcal{M}_{\text{inv}}$ there exists a regular Borel probability measure μ on $\mathcal{M}_{\text{inv}}$ such that*

1. $\text{supp}(\mu) \subset \mathcal{M}_{\text{erg}}$*, and*
2. *for all $f \in C^0(X, \mathbb{R})$,*

$$\int_X f\,dm \;=\; \int_{m' \in \mathcal{M}_{\text{erg}}} \left(\int_X f\,dm' \right) d\mu. \tag{15}$$

Proof. We saw that the ergodic invariant measures are precisely the extremal points of $\mathcal{M}_{\text{inv}}$. Thus by Choquet's theorem there exists a regular Borel probability measure μ on $\mathcal{M}_{\text{inv}}$ with support in $\mathcal{M}_{\text{erg}}$ such that

$$l(m) \;=\; \int_{m' \in \mathcal{M}_{\text{erg}}} l(m')\, d\mu$$

for all continuous linear functionals $l : \{\text{reg. Borel signed prob. measures on } X\} \to \mathbb{R}$. In particular, to each $f \in C^0(X, \mathbb{R})$ we have the associated functional

$$l_f(m') \;=\; \int_X f\, dm'.$$

Applying this for each f yields (15). ■

2.3 Birkhoff's ergodic theorem

We will first prove a special case of the ergodic theorem, theorem 2.26, as a warm up. This is conceptually easier and contains the main idea in the proof of the general result, theorem 2.30. Infact it also contains the main idea of the proof of the yet more general "subadditive ergodic theorem" which we will encounter in the next section.

Let $(X, \mathcal{B})$ be an arbitrary measurable space, $T : X \to X$ a measurable transformation, and $\mu \in \mathcal{M}_{\text{inv}}(X, T)$. Suppose $B \in \mathcal{B}$. For each $x \in X$ and $n \geq 1$ consider the total and average number of visits to B:

$$S_n(x) \;:=\; \#\Big\{\; 0 \leq i \leq n-1 \;\Big|\; T^i x \in B \;\Big\}$$

and

$$A_n(x) \;:=\; \frac{1}{n} S_n(x) \;=\; \frac{1}{n}\#\Big\{\; 0 \leq i \leq n-1 \;\Big|\; T^i x \in B \;\Big\}.$$

So $0 \leq A_n(x) \leq 1$ for all x, n. It is sometimes useful to write

$$A_n(x) \;=\; \frac{1}{n}\sum_{j=0}^{n-1} \chi_B\big(T^j x\big). \tag{16}$$

Theorem 2.26 (Birkhoff's ergodic theorem: version I)**.** *The pointwise limit*

$$A(x) \ := \ \lim_{n\to\infty} A_n(x) \tag{17}$$

exists μ-almost everywhere, and defines a T-invariant function. More precisely, the full measure set of points E on which the above limit exists satisfies $TE = E = T^{-1}E$, and $A(Tx) = A(x)$ for all $x \in E$. Moreover,

$$\int_X A\, d\mu \ = \ \mu(B). \tag{18}$$

Remark 2.27. It follows from the point wise convergence plus the dominated convergence theorem that $A_n \to A$ in $L^p(X, \mathcal{B}, \mu)$ for all $1 \leq p < \infty$. (The function $x \mapsto 1$ serves as a dominating function.)

Remark 2.28. Integrating (16) term by term gives

$$\int_X A_n\, d\mu \ = \ \mu(B)$$

for each $n \geq 1$.

Proof. (Of the Birkhoff ergodic theorem, version I) Set

$$\begin{cases} \overline{A}(x) \ := \ \limsup\limits_{n\to\infty} A_n(x) \\ \underline{A}(x) \ := \ \liminf\limits_{n\to\infty} A_n(x) \end{cases}$$

for each $x \in X$. Since $\overline{A} \geq \underline{A}$ the limit $A(x) = \lim_{n\to+\infty} A_n(x)$ will exist precisely at those $x \in X$ for which

$$\overline{A}(x) \ \leq \ \underline{A}(x). \tag{19}$$

It is easy to see that the set of points for which $A(x)$ exists is T invariant. Indeed, for each $x \in X$ and $n \geq 1$ we have

$$\frac{n+1}{n} A_{n+1}(x) \ = \ A_n(Tx) \ + \ \frac{1}{n}\chi_B(x). \tag{20}$$

Thus the limit $A(x)$ exists if and only if $A(T(x))$ exists and they are equal.

Therefore, to prove the theorem, it suffices only to prove that (19) holds μ-a.e and the rest of the statement will follow from Remarks (2.27) and (2.28) above.

One final remark, before we properly begin the proof, is that the inequalities

$$\overline{A} \circ T \geq \overline{A} \tag{21}$$

$$\underline{A} \circ T \leq \underline{A} \tag{22}$$

hold pointwise everywhere. Indeed, for fixed $x \in X$ let $n_j \to +\infty$ be so that $A_{n_j}(x) \to \overline{A}(x)$. Then from (20) we have $A_{n_j-1}(Tx) \to \overline{A}(x)$, thus $\overline{A} \circ T(x) \geq \overline{A}(x)$. The same argument shows (22)[*].

Now we begin the proof properly, our goal being to establish (19) μ-almost everywhere. Fix $\varepsilon > 0$. By definition of $\underline{A}$ we have that $\forall x \in X\ \exists n \geq 1$ such that

$$A_n(x) \ \leq \ \underline{A}(x) \ + \ \varepsilon.$$

Let $\tau(x)$ denote the minimum such n. More precisely, define $\tau : X \to \mathbb{N}$ by

$$\tau(x) \ := \ \min\left\{ \ n \geq 1 \ \middle| \ A_n(x) \ \leq \ \underline{A}(x) \ + \ \varepsilon \ \right\}.$$

(τ is everywhere defined and measurable.) Now we consider two cases.

Case 1: Suppose $\exists N \in \mathbb{N}$ such that $\tau(x) \leq N$ for all $x \in X$. To prove (19) we argue as follows.

Consider any $x \in X$, and the length-n orbit of x for some large n.

Decompose the orbit into pieces with initial points $x_0 = x, x_1, x_2, \ldots, x_q$, so

[*] We could at this point argue that (21) and (22) are equalities μ-almost everywhere since the integrals of both sides over X yield the same values. We won't require this here, but it will be useful in our proof of the subadditive ergodic theorem in section 3.5.

for $0 \leq j \leq q-1$ the orbit piece starting at x_j has length $\tau(x_j)$ and the last piece (which starts at x_q) has length $0 \leq l \leq \tau(x_q)$.

This means that on each piece, with the possible exception of the last one, the average number of visits to B is $\leq \underline{A}(x_j)+\varepsilon \leq \underline{A}(x)+\varepsilon$ by (22). For the last piece we don't have this upper bound, on the other hand we know that its length is $l \leq N$.

Using this decomposition we can estimate the average number of visits to B along the *whole* orbit from x to $T^n x$ by

$$\begin{aligned} S_n(x) \ &= \ \sum_{j=0}^{q-1} S_{\tau(x_j)}(x_j) \ + \ S_l(x_q) & (23) \\ &= \ \sum_{j=0}^{q-1} \tau(x_j) A_{\tau(x_j)}(x_j) \ + \ l A_l(x_q) & \\ &\leq \ \sum_{j=0}^{q-1} \tau(x_j)\Big(\underline{A}(x)+\varepsilon\Big) \ + \ N & \\ &\leq \ n\Big(\underline{A}(x)+\varepsilon\Big) \ + \ N. & (24) \end{aligned}$$

Dividing by n yields

$$A_n(x) \ \leq \ \underline{A}(x) \ + \ \varepsilon \ + \ \frac{N}{n}$$

for all $x \in X$ and $n \geq 1$. Taking lim sup of both sides we have shown

$$\overline{A}(x) \ \leq \ \underline{A}(x)+\varepsilon$$

for all $x \in X$ and all $\varepsilon > 0$. In particular (19) follows *everywhere* i.e. the limit $A_n \to A$ exists point wise everywhere!

Case 2: In which τ is unbounded. Pick $N \in \mathbb{N}$ large enough that

$$\mu\left(\left\{ \ x \ \middle| \ \tau(x) > N \ \right\}\right) \ < \ \varepsilon.$$

(Which we can do because the sets $E_N := \{x \,|\, \tau(x) > N\}$ for $N \geq 1$ are nested and the intersection of all of them is empty because τ is everywhere finite, so $\mu(E_N) \to 0$ as $N \to +\infty$.)

Naively one would like to throw out these points to use the proof in case 1. But this obviously doesn't work because we need T to be a dynamical system. Instead modify B to be

$$B' \;:=\; B \;\cap\; \left\{\, x \;\middle|\; \tau(x) \leq N \,\right\}$$

so $\mu(B') \geq \mu(B) - \varepsilon$. Then define the modified averages A'_n (which counts visits to B') and its liminf function $\underline{A}'$. Clearly $\underline{A}' \leq \underline{A}$ so the function

$$\tau'(x) \;:=\; \min\left\{\; n \geq 1 \;\middle|\; A'_n(x) \;\leq\; \underline{A}(x) \;+\; \varepsilon \;\right\}$$

is everywhere finite. Note that since this uses $\underline{A}(x)+\varepsilon$, rather than $\underline{A}'(x)+\varepsilon$, the function τ' is not precisely analogous to τ but some hybrid of the original data (associated to B) and the unmodified data (associated to B').

Now we observe that τ' is bounded! (by N). Indeed, since $A'_n \leq A_n$ for each $n \geq 1$, we have $\tau' \leq \tau$. Thus for $x \in B'$ we have $\tau'(x) \leq \tau(x) \leq N$, while for $x \notin B'$ we have $\tau'(x) = 1$ because $A'_1(x) = 0 \leq \underline{A}(x) + \varepsilon$.

Following the proof of Case 1 we decompose an arbitrary orbit $x, Tx, \ldots, T^n x$ into pieces whose length is determined by the function τ'. This leads to the estimate

$$A'_n(x) \;\leq\; \underline{A}(x) \;+\; \varepsilon \;+\; \frac{N}{n}$$

for all $x \in X$ and $n \geq 1$. To relate the left hand side with the unmodified data we integrate before letting $n \to +\infty$ to get

$$\mu(B') \;\leq\; \int_X \underline{A}\, d\mu \;+\; \varepsilon.$$

But $\mu(B')$ and $\mu(B)$ differ by at most ε, so $\mu(B) \;\leq \int_X \underline{A}\, d\mu \;+\; 2\varepsilon$, and letting $\varepsilon \to 0$ we have

$$\mu(B) \;\leq \int_X \underline{A}\, d\mu. \tag{25}$$

To complete the proof observe that it suffices to establish the inequality

$$\mu(B) \;\geq \int_X \overline{A}\, d\mu. \tag{26}$$

Indeed it would then follow that

$$\int_X \overline{A} - \underline{A}\, d\mu \;\leq\; 0$$

and since the integrand is non-negative, that $\overline{A} = \underline{A}$ μ-almost everywhere as we require.

There are two ways to show (26). One way is to follow the proof of (25) reversing the roles of $\overline{A}$ and $\underline{A}$. The proof is not absolutely identical, but no additional issues arise. Alternatively, one can deduce (26) by applying (25) to the set $B^c = X \setminus B$. In either case, this completes the proof of our first version of the Birkhoff ergodic theorem. ■

Corollary 2.29. *If μ is an* ergodic *invariant measure for $T : X \to X$ and $B \in \mathcal{B}$ then the averages*

$$A_n(x) \;:=\; \frac{1}{n}\#\bigl\{\, 0 \leq i \leq n-1 \bigm| T^i x \in B \,\bigr\}$$

converge point wise μ-a.e. to the value $\mu(B)$.

Proof. The limiting function A is T-invariant so by ergodicity it must be constant μ-a.e. Since it also has integral over X equal to $\mu(B)$ this constant must be $\mu(B)$. ■

Question 1. What is wrong with the following argument? Suppose that $\mu \in \mathcal{M}_{\mathrm{erg}}(X,T)$. Let $B \in \mathcal{B}$. By the ergodic theorem (in the form of corollary 2.29), the functions A_n (which at x equals the average number of visits of the orbit $x, Tx, \ldots, T^n x$ to B) converges μ-almost everywhere to the constant $\mu(B)$. But now suppose that $\nu \in \mathcal{M}_{\mathrm{erg}}(X,T)$ is a second ergodic invariant measure. Then by the same reasoning the averages converge almost everywhere to the constant $\nu(B)$. The functions A_n are independent of μ and ν, and therefore they must converge to the same limit and we conclude that $\mu(B) = \nu(B)$. Since this holds for all $B \in \mathcal{B}$ we must have $\mu = \nu$ and so there is always at most 1 ergodic invariant measure.

Theorem 2.26 is just a special case of the ”real” Birkhoff ergodic theorem. We discuss this now.

Let $(X, \mathcal{B})$ be an arbitrary measurable space, $T : X \to X$ a measurable transformation, and $\mu \in \mathcal{M}_{\text{inv}}(X, T)$.

Theorem 2.30 (Birkhoff's ergodic theorem: version II)**.** *For $f : (X, \mathcal{B}) \to \mathbb{R}$ representing an element of $L^1(X, \mathcal{B}, \mu)$ we define average functions*

$$f_n(x) \ := \ \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) \tag{27}$$

for $n \geq 1$. Then the limit

$$f_*(x) \ := \ \lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x) \tag{28}$$

exists μ-almost everywhere. Moreover,

1. *$f_* \in L^p(X, \mathcal{B}, \mu)$ and $\|f_n - f_*\|_{L^p(X,\mathcal{B},\mu)} \to 0$ for all $1 \leq p < \infty$.*
2. *f_* is T-invariant. Indeed, if $f_*(x)$ exists then so does $f_*(Tx)$, and $f_*(Tx) = f_*(x)$.*
3. *$\int_X f_* \, d\mu \ = \ \int_X f \, d\mu$.*

Remark 2.31. Our first version of the Birkhoff ergodic theorem follows by setting $f = \chi_B$, since then the average of f along an orbit segment is simply the average number of visits to B;

$$f_n(x) \ = \ \frac{1}{n} \sum_{k=0}^{n-1} \chi_B(T^k x) \ = \ A_n(x).$$

Moreover $\int_X f \, d\mu = \mu(B)$.

Proof. (Of the Birkhoff ergodic theorem, version II) First observe that without loss of generality $f \geq 0$ since every L^1 function is a difference of two non-negative L^1 functions. Set

$$\begin{cases} \overline{f}(x) & := \limsup\limits_{n\to\infty} f_n(x) \\ \underline{f}(x) & := \liminf\limits_{n\to\infty} f_n(x) \end{cases}$$

for each $x \in X$. Since $\overline{f} \geq \underline{f}$ the limit $f_*(x) = \lim_{n\to+\infty} f_n(x)$ will exist precisely at those $x \in X$ for which

$$\overline{f}(x) \leq \underline{f}(x). \tag{29}$$

Therefore we will prove that (29) holds μ-a.e. Before doing so, observe the inequalities

$$\overline{f} \circ T \geq \overline{f} \tag{30}$$

$$\underline{f} \circ T \leq \underline{f} \tag{31}$$

hold point wise everywhere, which can be shown analogously to the functions $\overline{A}$ and $\underline{A}$ in the proof of theorem 2.26.

Fix $\varepsilon > 0$. By definition of $\underline{f}$ we have that $\forall x \in X$ $\exists n \geq 1$ such that

$$f_n(x) \leq \underline{f}(x) + \varepsilon.$$

Let $\tau(x)$ denote the minimum such n. More precisely, define $\tau : X \to \mathbb{N}$ by

$$\tau(x) := \min\left\{ n \geq 1 \;\middle|\; f_n(x) \leq \underline{f}(x) + \varepsilon \right\}.$$

(τ is everywhere defined and measurable.) Now we consider two cases[*].

Case 1: Suppose $\exists N, M \in \mathbb{N}$ such that $\tau(x) \leq N$ and $f(x) \leq M$ for all $x \in X$.

[*] It is actually not necessary to split the proof into two cases, but I think it is conceptually helpful. Indeed the simpler case 1 conveys the one key idea which is contained in lines (32) to (33). In case 2 one has to use more epsilons to get essentially the same idea to work.

Consider any $x \in X$, and the length-n orbit of x for some large n. Decompose the orbit into pieces with initial points $x_0 = x, x_1, x_2, \ldots, x_q$, such that the average of f along the piece of orbit with initial point x_j is $\leq \underline{f}(x_j) + \varepsilon \leq \underline{f}(x) + \varepsilon$ (using (31)), with the exception of the last piece beginning at x_q, which may not be long enough. To summarize, for $0 \leq j \leq q-1$, the average of f along the piece of orbit with initial point x_j is $\leq \underline{f}(x) + \varepsilon$.

From this decomposition we can estimate the average of f along the *whole* orbit from x to $T^n x$, just as we did in the proof of theorem 2.26, by

$$\begin{aligned}
\sum_{k=0}^{n-1} f(T^k x) \;&=\; \sum_{j=0}^{q-1} \sum_{k=0}^{\tau(x_j)-1} f(T^k x_j) \quad + \quad \sum_{k=0}^{l-1} f(T^k x_q) && (32)\\
&\leq\; \sum_{j=0}^{q-1} \tau(x_j) \left(\frac{1}{\tau(x_j)} \sum_{k=0}^{\tau(x_j)-1} f(T^k x_j) \right) \quad + \quad lM \\
&\leq\; \sum_{j=0}^{q-1} \tau(x_j) \Big(\underline{f}(x) + \varepsilon \Big) \quad + \quad NM \\
&\leq\; n \Big(\underline{f}(x) + \varepsilon \Big) \quad + \quad NM. && (33)
\end{aligned}$$

Divide by n:

$$f_n(x) \;\leq\; \underline{f}(x) \;+\; \varepsilon \;+\; \frac{NM}{n}$$

for all $x \in X$ and $n \geq 1$. Taking lim sup of both sides, as $n \to +\infty$, yields

$$\overline{f}(x) \;\leq\; \underline{f}(x) + \; \varepsilon$$

for all $x \in X$. We conclude that $\underline{f} = \overline{f}$ *everywhere*, and Case 1 follows.

Case 2: In which τ and or f are unbounded. In this case, rather than directly proving (29), we will prove the two integral inequalities

$$\int_X f\,d\mu \;\leq\; \int_X \underline{f}\,d\mu \qquad \text{and} \qquad \int_X \overline{f}\,d\mu \;\leq\; \int_X f\,d\mu. \tag{34}$$

Indeed together these will imply that $\int_X \overline{f} - \underline{f}\,d\mu = 0$ and therefore that $\overline{f} = \underline{f}$ μ-a.e. as we require.

We concentrate first on the left hand inequality in (34), which is slightly the more involved of the two. Fix $\varepsilon > 0$, we will show that

$$\int_X f\,d\mu \;\leq \int_X \underline{f}\,d\mu \;+\; \varepsilon.$$

There are two kinds of potential "bad" points for us which prevent us from directly applying the argument in case 1; those for which τ is large, and those for which f are large. In our first version of the ergodic theorem the function f corresponded to χ_B for some subset B, which is bounded by 1, and so only one kind of "bad“ point entered the proof.

With this in mind pick $N, M \geq 1$ sufficiently large that the set

$$E \;:=\; \left\{\; x \in X \;\middle|\; \tau(x) > N \text{ or } f(x) > M \;\right\}$$

is small in the sense that

$$\int_E \overline{f}\,d\mu \;<\; \varepsilon.$$

We can do this because $f, \overline{f} \in L^1(X, \mathcal{B}, \mu)$ and τ is almost everywhere finite. Now modify f on the "bad“ set E by setting

$$f'(x) \;:=\; \begin{cases} f(x) & \text{if } x \notin E \\ \min\left\{f(x), \underline{f}(x), M\right\} & \text{if } x \in E. \end{cases}$$

Observe that $f' \leq f$, and moreover f' is uniformly bounded since on the "bad" set E it is now bounded (by M). It is less transparent why we also asked for f' to be bounded by $\underline{f}$ on E. The reason for this will appear in a moment (it will imply $\tau' = 1$ on E). For $n \geq 1$ let f'_n denote the sequence of average functions corresponding to f'. Clearly $f' \leq f$ implies $f'_n \leq f_n$, which means that the function $\tau' : X \to \mathbb{N}$ given by

$$\tau'(x) \ := \ \min \left\{ \ n \geq 1 \ \middle| \ f'_n(x) \ \leq \ \underline{f}(x) \ + \ \varepsilon \ \right\}$$

is everywhere finite since $\tau' \leq \tau$. However, unlike τ, the modified function τ' is uniformly bounded! (by N) since for every $x \in E$ there holds

$$f'(x) \leq \underline{f}(x)$$

implying $\tau'(x) = 1$. Thus for our modified data f' and τ' there are no "bad" points, and we can fairly closely follow the argument from case 1.

Decomposing an arbitrary orbit $x, Tx, \ldots, T^n x$ into pieces whose length is determined by the function τ' leads to the estimate

$$f'_n(x) \ \leq \ \underline{f}(x) \ + \ \varepsilon \ + \ \frac{NM}{n}$$

for all $x \in X$ and $n \geq 1$. Taking $\limsup$ as we did in case 1 is awkward because we have the modified function on the left hand side. However, integrating first and then letting $n \to +\infty$ gives

$$\int_X f' \, d\mu \ \leq \ \int_X \underline{f} \, d\mu \ + \ \varepsilon$$

(note that $\int f'_n d\mu = \int f' d\mu$ for all n). It is now easy to compare the left hand side with the unmodified data, arriving at

$$\int_X f \, d\mu \ \leq \ \int_X \underline{f} \, d\mu \ + \ 2\varepsilon.$$

This now holds for all $\varepsilon > 0$, proving the left hand inequality in (34).

It remains to prove the right hand inequality in (34). Essentially the same argument works, that is one fixes $\varepsilon > 0$ and defines $\tau : X \to \mathbb{N}$ by

$$\tau(x) \ := \ \min \left\{ \ n \geq 1 \ \middle| \ f_n(x) \ \geq \ \overline{f}(x) \ - \ \varepsilon \ \right\}.$$

This time we are no longer troubled if f is unbounded from above, but rather whether it is unbounded from below. But this is trivially the case because from the outset we can assume that f is non-negative. In this small sense then things are a fraction easier.

The proof continues almost word for word as above but with all the inequalities reversed. Roughly, one modifies f to a new function f' if necessary (i.e. if τ is unbounded) so that a corresponding τ' is uniformly bounded by N say. Then the most important estimate, corresponding to lines (32) to (33), becomes

$$\begin{aligned}\sum_{k=0}^{n-1} f'(T^k x) \;&\geq\; \sum_{j=0}^{q-1} \tau'(x_j)\left(\frac{1}{\tau'(x_j)}\sum_{k=0}^{\tau'(x_j)-1} f'(T^k x_j)\right)\\ &\geq\; \sum_{j=0}^{q-1} \tau'(x_j)\Big(\overline{f}(x)-\varepsilon\Big)\\ &\geq\; (n-N)\Big(\overline{f}(x)-\varepsilon\Big),\end{aligned}$$

and on dividing by n we get

$$f'_n(x) \;\geq\; \left(1-\frac{N}{n}\right)\Big(\overline{f}(x)-\varepsilon\Big)$$

for all $x \in X$ and $n \geq 1$. Integrating and letting $n \to +\infty$ gives

$$\int_X f'\,d\mu \;\geq\; \int_X \overline{f}\,d\mu \;-\; \varepsilon.$$

By construction the integral over X of the modified function f' differs from that of f by at most ε. This leads to the right hand inequality in (34). ■

Corollary 2.32. *If μ is an* ergodic *invariant measure for $T : X \to X$ and $f \in L^1(X, \mathcal{B}, \mu)$ then the average functions f_n, for $n \geq 1$, given by*

$$f_n(x) \;:=\; \frac{1}{n}\sum_{k=0}^{n-1} f(T^k x) \qquad \forall x \in X,$$

converge point wise μ-a.e. and in L^1 to the value $\int_X f\,d\mu$.

Proof. The limiting function f_* is T-invariant so by ergodicity it must be constant μ-a.e. Since its' integral over X equals $\int_X f\,d\mu$ this must be the constant. ■

Exercise 2.33. Let $(X, \mathcal{B})$ be a compact metric space with the Borel σ-algebra, and suppose there exists $\mu \in \mathcal{M}_{\text{erg}}(T, X)$ where $\mu(U) > 0$ for every non-empty open $U \subset X$. Show that T has a dense forward orbit. Show moreover that almost every orbit is dense, i.e. that the set of initial conditions $x \in X$ for which the forward orbit $x, Tx, T^2x, \ldots$ is dense in X has measure 1.

Exercise 2.34. Suppose $Q = [0, 1] \times [0, 1]$ is the unit "box" and Q_L, Q_R denote the left and right halves (which half shares the boundary is unimportant). Let $T : Q \to Q$ be a measurable transformation preserving Lebesgue measure μ_{leb}. Suppose that we find a subset $A \subset Q_L$ of strictly positive (Lebesgue) measure such that for all $x \in A$ the asymptotic average number of visits of the orbit $x, Tx, \ldots$ to Q_L is less than $1/4$.

By the ergodic theorem it would follow that μ_{leb} is not ergodic for T, (because $1/4 < 1/2 = \mu_{\text{leb}}(Q_L)$). Therefore by definition there must exist a non-trivial T-invariant subset $B \subset Q$ (i.e. one having Lebesgue measure strictly between 0 and 1). The question is, how could be "find" such a set B - not in a constructive or explicit way, but how do we see a set B existing (e.g. extract it from the proof)? In particular it is not A or any of the other subsets mentioned.

3 Liapunov exponents

To describe roughly what we hope to accomplish in this section consider a (positive) hyperbolic fixed point $x \in \mathbb{R}^2$ of a smooth map $f : \mathbb{R}^2 \to \mathbb{R}^2$. This means that $Df(x)$ has only real, positive, eigenvalues $\mu_1 \leq \mu_2 \in (0, 1) \cup (1, +\infty)$. Suppose for example the eigenvalues are e^{10}, e^3. (I deliberately chose 10 and 3 to have the same sign; this is not necessary but better illustrates the general behavior.) Then $Df(x)$ can be represented by the matrix

$$\begin{pmatrix} e^{10} & 0 \\ 0 & e^3 \end{pmatrix}$$

with respect to a basis v_1, v_2 of $T_x R^2$. This gives us a splitting

$$T_x\mathbb{R}^2 \;=\; E^1 \oplus E^2$$

where E^i is spanned by v_i. Clearly

$$Df(x)(E^1) = E^1 \qquad Df(x)(E^2) = E^2$$

and for all $n \geq 1$

$$\begin{cases} Df^n(x)v_1 \;=\; e^{10n}v_1 & \forall v_1 \in E_1 \\ Df^n(x)v_2 \;=\; e^{3n}v_2 & \forall v_2 \in E_2. \end{cases}$$

These latter imply

$$\begin{cases} \displaystyle\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)v_1\| \;= 10 & \forall v_1 \in E_1\backslash\{0\} \\ \displaystyle\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)v_2\| \;= 3 & \forall v_2 \in E_2\backslash\{0\}. \end{cases}$$

How about in directions not in E_1 or E_2? Clearly if $v \in T_x\mathbb{R}^2$ has a non-zero E_1 component then $\|Df^n(x)v_1\|$ will blow up asymptotically exponentially like e^{10n}. Thus we arrive at the following for non-zero $v \in T_x\mathbb{R}^2$:

$$\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)v\| \;=\; \begin{cases} 10 & \text{if } v \notin E^2_x \\ 3 & \text{if } v \in E^2_x. \end{cases} \tag{35}$$

Liapunov exponents generalize this concept of an invariant filtration of the linear space on which the linearized dynamics behaves asymptotically exponentially at different rates, to points x that are not necessarily fixed points or periodic points of the diffeomorphism f.

What could this mean? Roughly speaking, given almost any point x we will find a 1-dimensional linear subspace of the tangent space at each point along the orbit of x:

$$E_{f^n(x)} \;\subset\; T_{f^n(x)}\mathbb{R}^2 \qquad \forall n \geq 0,$$

that is invariant under f, meaning that,

$$Df^n(x)(E_x) = E_{f^n(x)}$$

for all $n \geq 0$, (and hence by chain rule $Df^n(x)(E_{f^k(x)}) = E_{f^{n+k}(x)}$ for all $n, k \geq 0$) such that there exist real numbers $\lambda_1 > \lambda_2$ so that for non-zero $v \in T_x\mathbb{R}^2$ there holds

$$\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)v\| \;=\; \begin{cases} \lambda_1 & \forall v \notin E_x \\ \lambda_2 & \forall v \in E_x. \end{cases} \tag{36}$$

When such structure can be found we say that λ_1, λ_2 are the Liapunov exponents for f at the point x, and $E_x \subset T_x\mathbb{R}^2$ the Liapunov filtration. (Note: if x is a periodic point then we easily find such splittings along the orbit of x and real numbers λ_1, λ_2.) There is another possibility, namely that there is a single real number $\lambda \in \mathbb{R}$ and for all non-zero $v \in T_x\mathbb{R}^2$

$$\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)v\| \;=\; \lambda. \tag{37}$$

In this case λ is called the Liapunov exponent for f at x and the filtration is simply the whole space $T_x\mathbb{R}^2$.

It turns out to be a difficult question whether for a given $x \in M$ the Liapunov exponents at x exist. However a theorem of Oseledec basically says that the set of such points has full measure with respect to any f-invariant probability measure $\mu \in \mathcal{M}_{\text{inv}}$.

3.1 The subadditive ergodic theorem: statement and discussion

In this chapter and for most of the course, we will work with a transformation that is at least C^1 so that the differential makes sense and varies continuously. Let M be a smooth manifold, $f : M \to M$ a C^1-diffeomorphism. Fix an auxilliary Riemannian metric on M, all lengths of tangent vectors etc will implicitly be with respect to this metric.

Suppose $x \in M$. A limit resembling those in (36) drops easily out of the Birkhoff ergodic theorem as follows: By the chain rule, for $n \geq 1$, $Df^n(x) = Df(f^{n-1}(x)) \circ Df(f^{n-2}(x)) \circ \;\cdots\cdots\; \circ Df(x)$. Thus,

$$\|Df^n(x)\| \;\leq\; \|Df(f^{n-1}(x))\| \cdot \|Df(f^{n-2}(x))\| \;\; \cdots \;\; \|Df(x)\|$$

and so

$$\log \|Df^n(x)\| \;\leq\; \sum_{j=0}^{n-1} \log \|Df(f^j(x))\|. \tag{38}$$

Let $\mu \in \mathcal{M}_{\text{erg}}(M, f)$. By Birkhoff's ergodic theorem $1/n$ times the right hand side converges μ-almost everywhere. Moreover, as the invariant measure is also ergodic we get

$$\lim_{n \to +\infty} \frac{1}{n} \sum_{k=0}^{n-1} \log \|Df(f^k x))\| \;=\; \int_M \log \|Df\| \, d\mu.$$

How about $1/n$ times the left hand side of (38), i.e. does the limit

$$\lim_{n \to +\infty} \frac{1}{n} \log \|Df^n(x)\| \;= ? \tag{39}$$

exist μ-almost everywhere? We've just seen that the Birkhoff ergodic theorem implies

$$\limsup_{n \to +\infty} \frac{1}{n} \log \|Df^n(x)\| \;\leq\; \int_M \log \|Df\| \, d\mu$$

μ-almost everywhere.

This is about as far as we can go with the Birkhoff ergodic theorem. To handle (39) we need a stronger result known as the subadditive ergodic theorem. To motivate this first observe that the sequence of functions

$$F_n(x) \;:=\; \log \|Df^n(x)\| \qquad\qquad \forall n \geq 1$$

have the following important property, known as *subadditivity.*

$$\begin{aligned} F_{k+n}(x) \;&=\; \log \|Df^{k+n}(x)\| \\ &=\; \log \|Df^k(f^n(x)) \circ \|Df^n(x)\| \\ &\leq\; \log \|Df^k(f^n(x))\| \;+\; \log \|Df^n(x)\| \\ &=\; F_k(f^n(x)) \;+\; F_n(x). \end{aligned} \tag{40}$$

Theorem 3.1 (Subadditive ergodic theorem)**.** *Suppose $T : (X, \mathcal{B}) \to (X, \mathcal{B})$ is a measurable transformation on an arbitrary measurable space, and $\mu \in \mathcal{M}_{\text{erg}}(X, T)$. Let $(F_n)_{n\geq 1}$ be a sequence of functions in $L^1(X, \mathcal{B}, \mu)$ satisfying the subadditivity condition*

$$F_{n+k}(x) \;\leq\; F_n(x) \;+\; F_k(T^n(x)) \qquad \mu - a.e. \tag{41}$$

for all $n, k \geq 1$. Then there exists $\lambda \in \mathbb{R} \cup \{-\infty\}$ such that

$$\lim_{n\to+\infty} \frac{1}{n} F_n(x) \;=\; \lambda \tag{42}$$

μ-almost everywhere. Moreover,

$$\lambda \;=\; \inf_{n\geq 1} \frac{1}{n} \int_X F_n \, d\mu. \tag{43}$$

This was first proved by Kingman in 1963. We postpone the proof to section 3.5. Let us prove the following useful fact for later.

Lemma 3.2. *Suppose that $a_n \in \mathbb{R}$, for $n \geq 1$, is a sequence satisfying the subadditivity condition $a_{n+m} \leq a_n + a_k$ for all $n, m \in \mathbb{N}$. Then*

$$\lim_{n\to+\infty} \frac{1}{n} a_n = \inf_{n\geq 1} \frac{1}{n} a_n$$

exists in $[-\infty, \infty)$.

Proof. Fix $k \in \mathbb{N}$ arbitrary. Then subadditivity implies that $a_{nk} \leq n a_k$ for all $n \geq 1$, and so

$$\frac{a_{nk}}{nk} \leq \frac{a_k}{k}$$

for all $n \geq 1$. Now for arbitrary $m \in \mathbb{N}$ we can write $m = nk + l$ for some $l \in \{0, 1, \dots, k-1\}$. Then

$$\frac{a_m}{m} \leq \frac{a_{nk}}{nk + l} + \frac{a_l}{nk + l} \leq \frac{a_k}{k} + \frac{a_l}{nk + l}.$$

Letting $m \to +\infty$ implies $n \to +\infty$, so that

$$\limsup_{m\to+\infty} \frac{a_m}{m} \leq \frac{a_k}{k} \tag{44}$$

for all $k \geq 1$. Hence

$$\limsup_{m \to +\infty} \frac{a_m}{m} \leq \liminf_{k \to +\infty} \frac{a_k}{k}.$$

Therefore the limit $\lim_{n \to +\infty} a_n/n = L$ exists within $[-\infty, \infty)$. Moreover by (44) $L = \inf_{n \geq 1} a_n/n$. ■

Remark 3.3. The Birkhoff ergodic theorem 2.30 corresponds to the "additive" case of theorem 3.1. That is, given $f \in L^1(X, \mathcal{B}, \mu)$, the sequences of functions $F_n(x) := \sum_{j=0}^{n-1} f(T^j x)$ satisfy

$$F_{n+k}(x) \; = \; F_n(x) \; + \; F_k(T^n(x)) \qquad \mu - a.e.$$

for all $n, k \geq 1$.

Now we show how the existence of the limit in (39), almost everywhere, follows from the subadditive ergodic theorem.

Corollary 3.4. *Suppose $f : M \to M$ is a C^1-diffeomorphism of a compact Riemannian manifold, and $\mu \in \mathcal{M}_{\mathrm{erg}}(M, f)$. Then there exists $\lambda \in \mathbb{R}$ such that*

$$\lim_{n \to +\infty} \frac{1}{n} \log \|Df^n(x)\| \; = \; \lambda$$

for μ-almost all $x \in M$.

Proof. Let $F_n(x) := \log \|Df^n(x)\|$ and $T = f : M \to M$. Then by the sub-additive ergodic theorem

$$\lim_{n \to +\infty} \frac{1}{n} \log \|Df^n\| \; = \; \lim_{n \to +\infty} \frac{1}{n} F_n(x) \; = \; \lambda$$

μ-almost everywhere, for some $\lambda \in \mathbb{R} \cup \{-\infty\}$. It remains only to show that for this particular choice of F_n we have $\lambda > -\infty$. This follows using the chain rule:

$$\begin{aligned} 1 = \|Df^n(x) \circ Df^n(x)^{-1}\| \; &\leq \; \|Df^n(x)\| \|Df^n(x)^{-1}\| \\ &= \; \|Df^n(x)\| \|Df^{-n}(f^n(x))\| \\ &\leq \; \|Df^n(x)\| \|Df^{-n}\|_{C^0} \\ &\leq \; \|Df^n(x)\| \|Df^{-1}\|_{C^0}^n. \end{aligned}$$

Thus

$$\frac{1}{n} \log \|Df^n(x)\| \ \geq \ \log \|Df^{-1}\|_{C^0} \ > \ -\infty.$$

■

Remark 3.5. Corollary 3.4 does not depend on the choice of equivalent metric on M. Indeed, if $\| \cdot \|_x'$ is a norm on each tangent space varying continuously with $x \in M$ (i.e. a Finsler metric on M) equivalent to the given one, then there exist constants such that $c_1 \| \cdot \| \leq \| \cdot \|' \leq c_2 \| \cdot \|$ on all of M. Thus

$$\log c_1 \ + \ \log \|Df^n(x)\| \ \leq \ \log \|Df^n(x)\|' \ \leq \ \log c_2 \ + \ \log \|Df^n(x)\|$$

for all $x \in M$ and $n \geq 1$. Dividing by n and letting $n \to \infty$ we see that the c_1 and c_2 terms disapear.

To conclude this subsection: we have seen that the exponential growth rates of the *norms* of the linearized maps exist almost everywhere (corollary 3.4) as a consequence of the subadditive ergodic theorem (which we have yet to prove). The direction dependent asymptotic estimates in (36), which we are aiming for, require more work and will be the content of Oseledec's theorem whose statement we discuss in the next subsection.

3.2 Oseledec's theorem: statement and examples

Recall M is a smooth compact Riemannian manifold and $f : M \to M$ is a C^1-diffeomorphism. Let $x \in M$ and consider the sequence of linear maps

$$Df^n(x) : T_xM \to T_{f^n(x)}M.$$

Corollary 3.4 tells us something about the growth rate of the *norms* $\|Df^n(x)\|$, $n \geq 1$, of these maps. Oseledec's theorem gives us information about the growth rate *in each direction*; i.e. about the sequence $\|Df^n(x)v\|$ for each fixed $v \in T_xM$, as $n \to +\infty$.

We examine the statement first in dimension 2.

Theorem 3.6 (Oseledec for surfaces)**.** *Suppose $f : M \to M$ is a C^1-diffeomorphism*[*] *of a compact Riemannian surface. There exists an f-invariant set $\Lambda \subset M$ with $\mu(\Lambda) = 1$ for all $\mu \in \mathcal{M}_{\text{erg}}(M, f)$, such that one of the following two possibilities occurs: either*

1. *$\exists \lambda \in \mathbb{R}$ such that for all $x \in \Lambda$,*

   $$\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)v\| \; = \; \lambda$$

   for all non-zero $v \in T_xM$. Or,

2. *$\exists \lambda_1 > \lambda_2 \in \mathbb{R}$ and for all $x \in \Lambda$ a 1-dimensional subspace $E_x \subset T_xM$ (with the map $x \mapsto E_x$ measurable) that is f-invariant, meaning $Df(x)E_x = E_{f(x)}$ for all $x \in \Lambda$, and moreover for all non-zero $v \in T_xM$,*

   $$\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)v\| \; = \; \begin{cases} \lambda_1 & \text{if } v \notin E_x \\ \lambda_2 & \text{if } v \in E_x \end{cases}$$

   for all $x \in \Lambda$.

Definition 3.7 (Liapunov exponents)**.** The numbers λ, respectively λ_1, λ_2, occurring in theorem 3.6 are called the **Liapunov exponents** of $f : M \to M$ with respect to the measure $\mu \in \mathcal{M}_{\text{erg}}(M, f)$.

Remark 3.8. It is not necessary that the invariant measure μ in Oseledec's theorem be ergodic. Then the Liapunov exponents λ_1, λ_2 are no longer constants and have to be replaced by measurable functions

$$x \mapsto \lambda_1(x) \quad \text{and} \quad x \mapsto \lambda_2(x)$$

which are f invariant.

Remark 3.9. The following will come out of the proof of Oseledec's theorem. Fix $\mu \in \mathcal{M}_{\text{erg}}(M, f)$ and $x \in M$ with Liapunov exponents $\lambda_1 \geq \lambda_2$.

[*] See the footnote to Theorem 3.10 regarding the assumption of invertibility.

- λ_1 will correspond to the growth rate of the norms $\|Df^n(x)\|$ in the sense that

  $$\lambda_1 \;=\; \lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)\|,$$

  while λ_2 will be the growth rate of $1/\|Df^n(x)^{-1}\|$, that is

  $$\lambda_2 \;=\; -\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)^{-1}\|.$$

- Hence,

  $$\lim_{n\to+\infty} \frac{1}{n} \log |\det Df^n(x)| \;=\; \lambda_1 + \lambda_2.$$

  On the other hand, since $\det AB = \det A \det B$ the functions $F_n(x) = \log |\det Df^n(x)|$ are *additive* (earlier we considered $F_n(x) = \log \|Df^n(x)\|$ which is only sub additive because $\|AB\| \leq \|A\|\|B\|$), and therefore we can apply the Birkhoff ergodic theorem and deduce that if m is an ergodic invariant measure for $f : M \to M$ then

  $$\lim_{n\to\infty} \frac{1}{n} \log |\det Df^n(x)| \;=\; \int_M \log |\det Df(x)|\, dm(x)$$

  for m-almost all $x \in M$. It follows that

  $$\int_M \log |\det Df(x)|\, dm(x) \;=\; \lambda_1 + \lambda_2. \tag{45}$$

- Consider the special case that the Riemannian surface (M, g) is oriented and $f : M \to M$ is volume/area preserving (with respect to a volume form induced by g). Suppose also that the corresponding invariant probability measure μ_{vol} is ergodic. Then volume preserving means that $\det Df \equiv 1$. Thus the integral in (45) vanishes, and we conclude that the Liapunov exponents with respect to μ_{vol} satisfy

  $$\lambda_1 = -\lambda_2.$$

For completeness here is the statement of Oseledec's theorem in arbitrary dimensions.

Theorem 3.10 (Oseledec, for arbitrary dimensions)**.** *Suppose $f : M \to M$ is a C^1-diffeomorphism*[*] *of a compact d-dimensional Riemannian manifold. There exists $\Lambda \subset M$ with $\mu(\Lambda) = 1$ for all $\mu \in \mathcal{M}_{\mathrm{erg}}(M, f)$, such that for all $x \in \Lambda$ there exist:*

1. *real numbers $\lambda_1 > \cdots > \lambda_k$, $(k \leq d)$;*[*]
2. *positive integers $d_1, \cdots, d_k \in \mathbb{N}$ such that $d_1 + \cdots + d_k = d$;*
3. *nested linear subspaces (a "filtration")*

   $$E_x^k \;\subset\; \cdots \;\subset\; E_x^1 \;=\; T_xM$$

   with $\dim E_x^i = d_k + \cdots + d_i$, and $Df(x)E_x^i = E_{f(x)}^i$, such that for non-zero $v \in T_xM$ there holds

   $$\lim_{n \to +\infty} \frac{1}{n} \log \|Df^n(x)v\| \;=\; \lambda_i \tag{46}$$

   whenever $v \in E_x^i$, but $v \notin E_x^{i+1}$. Finally, the maps $x \mapsto E_x^i$ are measurable.

Let us look at a few examples.

I need to update the notation in the following examples !

Example 3.11. $M = T^2 = \mathbb{R}^2/\mathbb{Z}^2$ the 2-torus with the flat metric, and $f : T^2 \to T^2$ is

$$f(x, y) = (x + \alpha \mod 1,\; y + \beta \mod 1).$$

Recall that the Lebesgue measure μ_{leb} in an invariant measure for f, and is ergodic if α/β is irrational. With respect to the global trivialization $\{\partial_x, \partial_y\}$

[*] Actually f need not be invertible, and when it is invertible one can conclude a significantly stronger statement. Compare Theorems 2.1.1 and 2.1.2 in [18].

[*] When I get time I will probably reverse the order of the indices here, I think the other direction is easier to follow.

of $TM \to M$ we find that the differential $Df(p)$ at each point is represented by

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and in particular $Df^n(p)v = v$, for all $n \geq 1$ and $p \in M$ and $v \in T_pM$. Therefore

$$\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(p)v\| \;=\; \lim_{n\to+\infty} \frac{1}{n} \log \|v\| \;=\; 0$$

for all non-zero $v \in T_pM$, and so $\lambda_1 = \lambda_2 = 0$ and the filtration is trivial.

Example 3.12 (Arnold's "cat map", see book by Arnold-Avez)**.** Again $M = T^2 = \mathbb{R}^2/\mathbb{Z}^2$ with the flat metric, now $f : T^2 \to T^2$ is given by

$$f(x, y) = (2x + y \mod 1,\ x + y \mod 1).$$

Thus f is the projection of the linear map $\tilde{f} : \mathbb{R}^2 \to \mathbb{R}^2$ represented by the matrix

$$A \;:=\; \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$$

with respect to the $(1, 0)$ and $(0, 1)$. Hence $\det A = 1$ implies that f preserves the "Lebesgue" measure μ_{leb} on T^2. It turns out moreover that μ_{leb} is ergodic (see for example [9]). The eigenvalues for A are $3 \pm \sqrt{5}/2$, let v_1, v_2 be respective eigenvectors. Since all of these are for T_pM independent p, we easily see that the Liapunov exponents of f are

$$\lambda_1 \;=\; \log\left(\frac{3+\sqrt{5}}{2}\right) > 0, \qquad\qquad \lambda_2 \;=\; \log\left(\frac{3-\sqrt{5}}{2}\right) < 0$$

(notice that $\lambda_1 = -\lambda_2$ because f is volume preserving), and the splitting

$$E_p^1 \;\oplus\; E_p^2 \;=\; T_pM$$

is the translate of the splitting at the origin $p = (0, 0)$ given by the span of v_1 and v_2.

Example 3.13. Obviously if M is any surface and $x \in M$ we can find a diffeomorphism $f : M \to M$ such that $f(x) = x$ and so that $Df(x)$ is represented by

$$\begin{pmatrix} e^{\lambda_1} & 0 \\ 0 & e^{\lambda_2} \end{pmatrix}$$

for any choice $\lambda_1 > \lambda_2$. Then the point mass at x: $\mu = \delta_x$ is ergodic for f and the Liapunov exponents for f are $\lambda_1, \ \lambda_2$. We see that if there are multiple such fixed points for f then different (delta) measures lead to different Liapunov exponents.

Here is a more interesting example. We will work on $\mathbb{R}^2$ for simplicity, strictly speaking we should compactify at infinity and work in S^2 to be consistent with our framework.

Example 3.14 (Smale's horse-shoe)**.** $f : R^2 \to R^2$ is a C^1-diffeomorphism with the following property: let $Q \subset \mathbb{R}^2$ be a box, i.e. some kind of rectangle as in the figure. $f : Q \to \mathbb{R}^2$ is the map composed of a ”squash” map S and a “fold“ map B.

So some points don't get mapped into Q (escape), and some points do. We see that $f(Q) \cap Q$ has two components

which are horizontal strips H_0 and H_1. Tracing the map f backwards we find the points in Q which get mapped into H_0 and H_1, i.e. the points which don't escape under one forward iterate:

So the points in Q which get mapped back to Q, namely $f^{-1}Q \cap Q$, has two components, and forms two vertical strips we label V_0 and V_1.

Consider what happens under forwards iteration of H_0 and H_1 under f successively:

Thus f maps H_0 and H_1 again into H_0 and H_1 (ignoring escaping points)

$$f: \; Q \cap f(Q) \;\longrightarrow\; Q \cap f(Q) \cap f^2(Q)$$

and continuing successively applying f, we get a nested sequence of horizontal strips. Eventually

$$H \;:=\; \bigcap_{k=0}^{+\infty} f^k(Q)$$

which intuitively is equal to $[0,1] \times \mathcal{C}$ where $\mathcal{C}$ is the middle thirds Cantor set. Note also that

$$H \;=\; \Big\{\, x \in Q \;\Big|\; f^{-k}(x) \in Q \quad \forall k \geq 1 \Big\}.$$

Now consider what happens to Q under backwards iterates of f. We've seen that $f^{-1}(Q) \cap Q = V_0 \cup V_1$. Under the next iterate we get

$$f^{-2}(Q) \cap f^{-1}(Q) \cap Q \;=$$

Successively applying f^{-1} we get nested vertical strips, and in the limit

$$V \;:=\; \bigcap_{k=-1}^{-\infty} f^k(Q)$$

which intuitively is equal to $\mathcal{C} \times [0,1]$. Observe that

$$V \;=\; \Big\{\, x \in Q \;\Big|\; f^k(x) \in Q \quad \forall k \geq 0 \Big\}.$$

Thus we obtain a non-empty closed invariant set

$$\Omega \;:=\; H \cap V \;=\; \Big\{\, x \in Q \;\Big|\; f^k(x) \in Q \quad \forall k \in Z \,\Big\} \;\simeq\; \mathcal{C} \times \mathcal{C}$$

There is a natural bijection between the restriction of f to Ω and the so called Bernoulli process given by the shift operator $T : X \to X$ on the space of sequences in $\{0,1\}$,

$$\sigma : (f, \Omega) \to (T, X)$$

where

$$X \;:=\; \big\{(x_i)_{i \in \mathbb{Z}} \;\big|\; x_i \in \{0,1\}\big\} \qquad \text{and} \qquad T(x_i) = (x_{i+1}).$$

Give X the product topology from the discrete topology on $\{0,1\}$, then a basis for the resulting σ-algebra $\mathcal{B}$ are the so called cylinder sets $[x_0, \ldots, x_k]$ defined by

$$[x_{k_1}, \ldots, x_{k_r}] \;:=\; \Big\{\; (y_i)_{i \in \mathbb{Z}} \;\Big|\; y_{k_1} = x_{k_1}, \ldots, y_{k_r} = x_{k_r} \Big\}.$$

By the extension theorem the function $\mu : \{\text{Cylinder sets}\} \to [0, \infty)$ given by

$$\mu\,(C) \;=\; (1/2)^l$$

where $l \geq 1$ is the length of the cylinder C, extends to a unique Borel measure on $\mathcal{B}$. With respect to this so called Bernoulli measure T is invariant and ergodic. (This is not difficult to prove but we won't get sidetracked into this right now.)

fill in more here....

It turns out that if the "squash“ map S contracts in the vertical direction by a factor $0 < \alpha < 1$, and expands in the horizontal direction by a factor $1/\alpha > 1$, then there is a "natural“ ergodic invariant measure μ supported on Ω. For each point $p \in \Omega$, we have

$$Df(x) \;=\; \begin{pmatrix} \alpha & 0 \\ 0 & 1/\alpha \end{pmatrix}$$

and we find that the Liapunov exponents are

$$\lambda_1 = \log\left(\frac{1}{\alpha}\right) > 0, \qquad\qquad \lambda_2 = \log\left(\alpha\right) < 0$$

and that the splitting $T_x\mathbb{R}^2 = E^1_x \oplus E^2_x$ corresponds to the horizontal and vertical directions.

Example 3.15 (Non-uniform horse-shoes)**.** The previous example can be made more interesting if we arrange that the "squashing" map in the construction of $f : R^2 \to R^2$ has varying contraction and or expansion rates.

As before we obtain a closed invariant set $\Omega \subset Q$ homeomorphic to a product of Cantor sets $\mathcal{C} \times \mathcal{C}$. There is a dense set of periodic points in Ω, and in particular always two fixed points. We can arrange our squashing map so that the contraction/expansion rate degenerates at one of the two fixed points, call it p, so that $Df(p) = \mathrm{id}$. From the Oseledec theorem with respect to the Bernoulli measure on Ω, which is ergodic, it is not hard to believe that we pick up non-zero Liapunov exponents (for example using the relations to entropy via integrals that we consider in the next section). On the other hand due to the degenerate fixed point p it must be the case that the support of the Bernoulli measure is Ω minus a non-empty subset!

Complete!

3.3 Proof of Oseledec's theorem in 2D

Recall that $f : M \to M$ is a C^1-diffeomorphism of a smooth compact Riemannian 2-manifold (M, g).

Definition 3.16. Call $x \in M$ a **Lyapunov-regular** point for f if the limits

$$\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)\| \;=\; \lambda_1(x) \in \mathbb{R} \tag{47}$$

and

$$-\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)^{-1}\| \;=\; \lambda_2(x) \in \mathbb{R} \tag{48}$$

both exist (the minus sign in the second limit is not a typo). We denote the set of Lyapunov-regular points of f by $\Lambda = \Lambda_f$. The values $\lambda_1(x), \lambda_2(x)$ are called the **Lyapunov exponents** of x.

Remark 3.17. From the definition

$$\lambda_1(x) \;\geq\; \lambda_2(x)$$

because for any invertible matrix M we have $1 = \|MM^{-1}\| \leq \|M\|\|M^{-1}\|$.

Remark 3.18. Note that x regular for f does not imply that x is regular for f^{-1}. Indeed, many sources refer to forward and backward regular points, but we will not worry about the distinction here.

Lemma 3.19. *The set of Lyapunov-regular points, and the values of the Lyapunov exponents, are f-invariant. That is, $x \in \Lambda$ iff $f(x) \in \Lambda$, and $x \in \Lambda$ implies $\lambda_i(f(x)) = \lambda_i(x)$ for $i = 1, 2$.*

Proof. Let $x \in M$ and set $y := f(x)$. Then by the chain rule, for all $n \geq 1$,

$$Df^n(y)M = Df^{n+1}(x)$$

where $M = Df(x)$ is an invertible linear map. Thus there exist positive constants C_1, C_2 depending only on x such that

$$C_1\|Df^{n+1}(x)\| \leq \|Df^n(y)\| \leq C_2\|Df^{n+1}(x)\|$$

for all $n \geq 1$. Taking logarithms etc it follows that the limit (47) exists for x iff it exists for y, and then the limiting values are the same. Similarly for the limit in (48). Thus $x \in \Lambda$ if and only if $y \in \Lambda$, and $\lambda_i(x) = \lambda_i(y)$ for $i = 1, 2$. ■

Oseledec's theorem will follow by combining the next two propositions.

Proposition 3.20. *If $m \in \mathcal{M}_{\mathrm{erg}}(f, M)$ then m-almost every point $x \in M$ is Lyapunov regular.*

Proof. This follows immediately from the subadditive ergodic theorem as in the proof of corollary 3.4. Indeed, this corollary showed that the limit (47) exists for all points in some Borel set $\Lambda_0 \subset M$ with $m(\Lambda_0) = 1$. Essentially the same argument shows that the limit (48) exists on some $\Lambda_1 \subset M$ also with $m(\Lambda_1) = 1$. Then we can take $\Lambda = \Lambda_0 \cap \Lambda_1$. ■

Proposition 3.21. *Suppose $x \in \Lambda$, that is x is a regular point for $f : M \to M$. Let $\lambda_1 = \lambda_1(x)$ and $\lambda_2 = \lambda_2(x)$ denote the two Lyapunov exponents, in particular $\lambda_1 \geq \lambda_2$. Then:*

1. *If $\lambda_1 = \lambda_2$ then*

$$\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)v\| \;=\; \lambda_1$$

for all $v \in T_xM\backslash\{0\}$.

2. *If $\lambda_1 > \lambda_2$ then there exists a unique* 1*-dimensional subspace $E_x \subset T_xM$ such that for all non-zero $v \in T_xM$,*

$$\lim_{n\to+\infty} \frac{1}{n} \log \|Df^n(x)v\| \;=\; \begin{cases} \lambda_1 & \textit{if } v \notin E_x \\ \lambda_2 & \textit{if } v \in E_x \end{cases}$$

Remark 3.22. Observe that in general if $x \in \Lambda_f$, i.e. if λ_1 and λ_2 exist as in (47) and (48), then for any non-zero $v \in T_xM$,

$$\lambda_2 \;\le\; \underline{\lim}_{n\to\infty} \frac{1}{n} \log \|Df^n(x)v\| \;\le\; \overline{\lim}_{n\to\infty} \frac{1}{n} \log \|Df^n(x)v\| \;\le\; \lambda_1. \quad (49)$$

Indeed, in any dimension $d \ge 1$, take $M \in \mathrm{Gl}(\mathbb{R}, d)$, then

$$\frac{1}{\|M^{-1}\|} \;\le\; \|Mv\| \;\le\; \|M\|$$

for all $v \in \mathbb{R}^d$ with $\|v\| = 1$. (For the left hand inequality use $1 = \|M^{-1}Mv\| \le \|M^{-1}\|\|Mv\|$.) Applying this to the linear map $Df^n(x)$ and taking logarithms etc

$$-\frac{1}{n} \log \|Df^n(x)^{-1}\| \;\le\; \frac{1}{n} \log \|Df^n(x)v\| \;\le\; \frac{1}{n} \log \|Df^n(x)\|.$$

By assumption of x being a regular point, as $n \to +\infty$ the left and right sides converge to λ_2 and λ_1 respectively.

This remark proves case *1.* of the proposition. To prove case *2.* we will first use the following fact from linear algebra to make the linear maps “nicer”.

Lemma 3.23. *If $B \in \mathrm{Gl}(\mathbb{R}, d)$ then there exists a (unique) symmetric, positive definite, $A \in \mathrm{Gl}(\mathbb{R}, d)$ such that $A^2 = B^TB$. It follows that*

1. $\|Bv\| = \|Av\|$ *for all $v \in \mathbb{R}^d$,*

2. $\|B\| = \|A\|$

3. $\|B^{-1}\| = \|A^{-1}\|$.

Proof. $M := B^T B$ is symmetric, so can be diagonalized. That is, there exists $P \in \mathrm{Gl}(\mathbb{R}, d)$ so that $M = P^{-1}DP$ where D is diagonal. Taking the positive square-roots of the diagonal entries gives us a diagonal matrix $D^{1/2}$ whose square equals D. (Indeed, D has only positive entries on the diagonal because M is positive definite: $\langle Mv, v\rangle = \|Bv\|^2 > 0$ for all $v \neq 0$.) Then set $A = P^{-1}D^{1/2}P$. This gives us existence of A.

The relation in 1. is obvious, and 2. follows immediately. It remains to check 3. that $\|B^{-1}\| = \|A^{-1}\|$. Abbreviate $M^{-T} := (M^T)^{-1} = (M^{-1})^T$ for general invertible matrices. Then

$$(A^{-T})^2 = C^T C$$

where $C := B^{-T}$. Thus A^{-T} has the same relation to C that A has to B. Hence $\|C\| = \|A^{-T}\|$. But the norm is unchanged by the operation of transpose, so $\|B^{-1}\| = \|C\| = \|A^{-T}\| = \|A^{-1}\|$. ■

Proof. (Of proposition 3.21.) We are given a regular point $x \in M$, and a sequence of linear maps $Df^n(x) : T_xM \to T_{f^n(x)}M$. Picking an arbitrary orthonormal basis of each tangent space $T_{f^n(x)}M$ with respect to the given Riemannian metric g, we obtain a sequence of matrices $B_n \in \mathrm{Gl}(\mathbb{R}, 2)$, $n \geq 1$, and the inner product becomes the standard Euclidean one on $\mathbb{R}^2$. Our assumption that x is regular becomes the condition that

$$\begin{cases} \lim_{n\to+\infty} \frac{1}{n} \log \|B_n\| &= \lambda_1 \\ -\lim_{n\to+\infty} \frac{1}{n} \log \|B_n^{-1}\| &= \lambda_2 \end{cases}$$

where $\lambda_1 > \lambda_2 \in \mathbb{R}$ are the Lyapunov exponents of f at the point x. Using the lemma we find symmetric, positive definite 2×2 matrices A_n, for $n \geq 1$, such that $A_n^2 = B_n^T B_n$. From the lemma we have

$$\begin{cases} \lim_{n\to+\infty} \frac{1}{n} \log \|A_n\| &= \lambda_1 \\ -\lim_{n\to+\infty} \frac{1}{n} \log \|A_n^{-1}\| &= \lambda_2. \end{cases} \tag{50}$$

The advantage of using the A_n's is that they have real, strictly positive, eigenvalues and orthogonal eigenspaces.

The rest of the proof will go as follows.

Step 1 The eigenspaces of the matrices A_n converge! (This seems to me the most unobvious part). Denote the limiting orthogonal subspaces $E_1^\infty, E_2^\infty \subset \mathbb{R}^2$ (where E_1^∞ is the limit of the eigenspaces with greatest eigenvalue).

Step 2 That if $v_i \in E_i$ is non-zero, then

$$\lim_{n\to+\infty} \frac{1}{n} \log \|A_n v_i\| \quad = \quad \lambda_i$$

for $i = 1, 2$. This will be split into two substeps: **Step (2a)** to prove the limit for λ_1, and **Step (2b)** to prove the limit for λ_2.

This will complete the proof by setting $E_x := E_2^\infty$, since $\|Df^n(x)v\| = \|B_n v\| = \|A_n v\|$ for all v.

Before carrying out the above steps, let us give names to objects and choose convenient eigenvectors. Denote by $\mu_1^n \geq \mu_2^n > 0$ the eigenvalues of A_n for $n \geq 1$. Then since A_n is symmetric (and positive) its norm is achieved by the highest eigenvalue (similarly for A_n^{-1}), so

$$\mu_1^n = \|A_n\| \qquad \text{and} \qquad \mu_2^n = 1/\|A_n^{-1}\|.$$

Thus, in view of (50)

$$\begin{cases} \displaystyle\lim_{n\to+\infty} \frac{1}{n} \log \mu_1^n \;=\; \lambda_1 \\ \displaystyle\lim_{n\to+\infty} \frac{1}{n} \log \mu_2^n \;=\; \lambda_2. \end{cases}$$

Hence, for all $\varepsilon > 0$

$$e^{n(\lambda_i - \varepsilon)} \;\leq\; \mu_i^n \;\leq\; e^{n(\lambda_i+\varepsilon)} \tag{51}$$

for all n sufficiently large, for $i = 1, 2$. In particular for n large $\mu_1^n > \mu_2^n$ and A_n has orthogonal eigenspaces which we denote by E_1^n and E_2^n (corresponding to μ_1^n and μ_2^n respectively). To show that these spaces converge as $n \to +\infty$ it suffices to find unit eigenvectors $e_1^n \in E_1^n$ and $e_2^n \in E_2^n$ that converge. For each $n \geq 1$ choose e_1^n (the choice is just up to ± 1) so that

$$\langle e_1^{n+1}, e_1^n \rangle \;\geq\; 0$$

(this makes sense if we want to show the sequence e_1^n converges; e_1^{n+1} and e_1^n are far apart if $\langle e_1^{n+1}, e_1^n\rangle \;<\; 0$.) Then choose $e_2^n = ie_1^n$ where i here denotes complex multiplication, i.e. (e_1^n, e_2^n) is positively oriented.

Step 1 In this step we will show that the sequence of eigenvectors $(e_1^n)_n$ is (exponentially) Cauchy. More precisely, that there exists $C, \delta > 0$ so that

$$\|e_1^{n+k} - e_1^n\| \;\leq\; Ce^{-\delta n} \tag{52}$$

for all $n, k \geq 1$.

It suffices to show such an estimate for $k = 1$ due to the formula for the sum of a geometric series. Now observe that

$$\|e_1^{n+1} - e_1^n\| \;\leq\; \sqrt{2}|\langle e_1^{n+1}, e_2^n\rangle| \tag{53}$$

for each $n \geq 1$. Indeed, draw a picture and you'll see that for any unit vectors $v, w \in \mathbb{R}^2$ with $\langle v, w\rangle > 0$ we have $\|v - w\| \leq \sqrt{2}|\langle v, iw\rangle|$ where i denotes complex multiplication. Taking $v = e_1^{n+1}$ and $w = e_1^n$ gives us (53).

Thus to complete step 1 it suffices for us to show that there exists $C, \delta > 0$ so that

$$\big|\langle e_1^{n+1}, e_2^n\rangle\big| \;\leq\; Ce^{-\delta n} \tag{54}$$

for all $n \geq 1$. To prove this we argue as follows.

$$\begin{aligned}
\big|\langle e_1^{n+1}, e_2^n\rangle\big| &= \frac{1}{\mu_1^{n+1}}\big|\langle A_{n+1}e_1^{n+1}, e_2^n\rangle\big| \\
&= \frac{1}{\mu_1^{n+1}}\big|\langle e_1^{n+1}, A_{n+1}e_2^n\rangle\big| \\
&\leq \frac{1}{\mu_1^{n+1}}\|A_{n+1}e_2^n\| \\
&= \frac{1}{\mu_1^{n+1}}\|B_{n+1}e_2^n\| \\
&= \frac{1}{\mu_1^{n+1}}\|Df(f^n(x))B_ne_2^n\| \\
&\leq C_0\frac{1}{\mu_1^{n+1}}\|B_ne_2^n\|
\end{aligned}$$

where $C_0 = \|Df\|_{C^0(M)}$,

$$\begin{aligned} &= C_0 \frac{1}{\mu_1^{n+1}} \|A_n e_2^n\| \\ &= C_0 \frac{\mu_2^n}{\mu_1^{n+1}}. \end{aligned}$$

In view of (51), we therefore find that for all $\varepsilon > 0$

$$\left|\langle e_1^{n+1}, e_2^n \rangle\right| \ \leq \ C_0 \frac{e^{n(\lambda_2+\varepsilon)}}{e^{n(\lambda_1-\varepsilon)}}$$

if n is sufficiently large. In particular, taking $\varepsilon < (\lambda_1 - \lambda_2)/2$,

$$\left|\langle e_1^{n+1}, e_2^n \rangle\right| \ \leq \ C_0 e^{-n\delta}$$

(where $\delta = (\lambda_1 - \lambda_2)/2 \ - \varepsilon > 0$) for all n sufficiently large. This proves (54) and hence step 1.

Remark 3.24. For step 2b later we will need to observe that the argument in step 1 actually leads to more precise estimates, namely that δ can be chosen smaller than, but arbitrarily close to, the difference $\Delta := \lambda_1 - \lambda_2 > 0$ of the two limiting eigenvalues. That is, there exists $C > 0$ so that for all $0 < \varepsilon < \Delta$

$$\|e_1^{n+k} - e_1^n\| \ \leq \ Ce^{-(\Delta-\varepsilon)n} \tag{55}$$

for all $n, k \geq 1$.

Step 2a As a result of step 1 we know that the sequences of unit eigenvectors e_1^n and e_2^n converge to some orthogonal unit vectors in $\mathbb{R}^2$. (Note that this corresponds to a limit of certain orthonormal vectors in T_xM.) Let us denote the limits by

$$e_i^\infty := \lim_{n\to+\infty} e_i^n$$

and set

$$E_i^\infty \ := \ \operatorname{Span} e_i^\infty$$

for $i = 1, 2$. So E_1^∞ and E_2^∞ are orthogonal 1-dimensional subspaces of $\mathbb{R}^2$. In step 2a we will show that

$$\lim_{n\to+\infty} \frac{1}{n} \log \|A_n e_1^\infty\| \ = \ \lambda_1.$$

By remark (3.22) it suffices to show

$$\liminf_{n\to+\infty} \frac{1}{n} \log \|A_n e_1^\infty\| \;\geq\; \lambda_1. \tag{56}$$

To prove this, first observe that

$$\lim_{n\to+\infty} \frac{1}{n} \log \|A_n e_1^n\| \;=\; \lambda_1$$

by construction, because $\|A_n e_1^n\| = \mu_1^n$ for each $n \geq 1$. Thus it is natural to compare $\|A_n e_1^n\|$ with $\|A_n e_1^\infty\|$ as $n \to +\infty$, indeed it suffices to show that the difference grows at most exponentially at a rate strictly less than λ_1. We have,

$$\begin{aligned} \Big| \|A_n e_1^n\| \;-\; \|A_n e_1^\infty\| \Big| \;&\leq\; \|A_n e_1^n \;-\; A_n e_1^\infty\| \\ &\leq\; \|A_n\| \, \|e_1^n - e_1^\infty\|. \end{aligned} \tag{57}$$

From step 2, letting $k \to +\infty$ in (52), we have $\|e_1^n - e_1^\infty\| \leq Ce^{-n\delta}$ for some $C, \delta > 0$ and all $n \geq 1$. Moreover, for all $\varepsilon > 0$ we have $\|A_n\| \leq e^{n(\lambda_1+\varepsilon)}$ if n is sufficiently large. Thus if $\varepsilon < \delta/2$ then

$$\Big| \|A_n e_1^n\| \;-\; \|A_n e_1^\infty\| \Big| \;\leq\; Ce^{n(\lambda_1-\delta/2)}. \tag{58}$$

That is, there exists $C > 0$ so that (58) holds for all $n \geq 1$. In particular, for all $n \geq 1$,

$$\|A_n e_1^\infty\| \;\geq\; \|A_n e_1^n\| \;-\; Ce^{n(\lambda_1-\delta/2)}.$$

So, $\forall \varepsilon > 0, \exists N$ s.t. for all $n \geq N$

$$\|A_n e_1^\infty\| \;\geq\; e^{n(\lambda_1-\varepsilon)} \;-\; Ce^{n(\lambda_1-\delta/2)}. \tag{59}$$

Therefore for $0 < \varepsilon < \delta/2$,

$$\liminf_{n\to+\infty} \frac{1}{n} \log \|A_n e_1^\infty\| \;\geq\; \lambda_1 - \varepsilon$$

proving (56).

Step 2b In this step we will show that

$$\lim_{n\to+\infty} \frac{1}{n} \log \|A_n e_2^\infty\| \;=\; \lambda_2.$$

By remark (3.22) it suffices to show

$$\limsup_{n\to+\infty} \frac{1}{n} \log \|A_n e_2^\infty\| \;\le\; \lambda_2. \tag{60}$$

We begin with

$$\lim_{n\to+\infty} \frac{1}{n} \log \|A_n e_2^n\| \;=\; \lambda_2$$

because $\|A_n e_2^n\| = \mu_2^n$ for each $n \ge 1$. As in step 2a,

$$\Big| \|A_n e_2^n\| \;-\; \|A_n e_2^\infty\| \Big| \;\le\; \|A_n\| \, \|e_2^n - e_2^\infty\|. \tag{61}$$

Now we use the improved version of step 1, see Remark 3.24. Indeed, fix $\varepsilon \in (0, \Delta)$, where $\Delta = \lambda_1 - \lambda_2$. Letting $k \to +\infty$ in (55) we have $C > 0$ (independent of ε) such that $\|e_2^n - e_2^\infty\| \le Ce^{-n(\Delta-\varepsilon)}$ for all $n \ge 1$. Thus for n sufficiently large

$$\begin{aligned} \Big| \|A_n e_2^n\| \;-\; \|A_n e_2^\infty\| \Big| \;&\le\; e^{n(\lambda_1+\varepsilon)} \, Ce^{-n(\Delta-\varepsilon)} \\ &=\; C \, e^{n(\lambda_2+\varepsilon)}. \end{aligned} \tag{62}$$

Hence,

$$\begin{aligned} \|A_n e_2^\infty\| \;&\le\; \|A_n e_2^n\| \;+\; Ce^{n(\lambda_2+\varepsilon)} \\ &\le\; e^{n(\lambda_2+\varepsilon)} \;+\; Ce^{n(\lambda_2+\varepsilon)} \end{aligned} \tag{63}$$

if n is sufficiently large. Thus,

$$\limsup_{n\to+\infty} \frac{1}{n} \log \|A_n e_2^\infty\| \;\le\; \lambda_2 + \varepsilon.$$

This proves (60) and completes the proof of step 2b and therefore proposition 3.21. ■

Completing the proof of Oseledec: To complete the proof, given the last two propositions, it suffices to show that the map $\Lambda \ni x \mapsto E_1(x) \subset T_x M$ is measurable and that $Df(x)E_1(x) = E_1(f(x))$ for all $x \in \Lambda$.

3.4 Cocycles: an introduction

Fill in.

3.5 Proof of the subadditive ergodic theorem

Recall that the subadditive ergodic theorem is a generalization of the Birkhoff ergodic theorem that was crucial in our proof of Oseledec's theorem. More precisely we used it to show that the set of Liapunov-regular points of a C^1-diffeomorphism $f : M \to M$ on a compact surface M has full measure with respect to any f-invariant probability measure on M.

In this section we revert to an arbitrary measurable space $(X, \mathcal{B})$ with $T : X \to X$ a measurable transformation, and $\mu \in \mathcal{M}_{\text{inv}}(X, T)$.

For comparison, recall that our first version of the Birkhoff ergodic theorem stated the following. Consider a set $B \in \mathcal{B}$ fixed, and for $x \in X$ and $n \geq 1$ consider the total number of visits of the orbit $x, Tx, \ldots, T^n x$ to the set B:

$$S_n(x) \;:=\; \#\Big\{\; 0 \leq i \leq n-1 \;\Big|\; T^i x \in B \;\Big\}.$$

Then the Birkhoff theorem stated that the averages

$$A_n(x) \;:=\; \frac{1}{n} S_n(x)$$

converge for μ-almost all x.

Examining the proof carefully, we see that all we use is that the sequence of functions $(S_n)_{n \geq 1}$ is *additive*, that is, for almost all $x \in X$,

$$S_{n+k}(x) \;=\; S_n(x) \;+\; S_k(T^n x)$$

for all $n, k \geq 1$. We also used the property of the averages A_n that they are uniformly bounded from above (by 1) and are non-negative. These latter two assumptions can be reduced to each S_n being integrable, and this gave us the second version of the Birkhoff ergodic theorem. Thus we really proved a statement about additive sequences of integrable functions.

In contrast, the subadditive ergodic theorem draws the same conclusion for a *subadditive* sequence (S_n) of integrable functions, where subadditive means that the equalities above are merely the inequalities

$$S_{n+k}(x) \;\leq\; S_n(x) \;+\; S_k(T^n x)$$

for all $n, k \geq 1$. If one goes through the proof of the Birkhoff theorem carefully and tries to remove any use of addivity, i.e. only using subadditivity, then one arrives at the following proof that we explain now[*].

First let us recall the precise statement of the subadditive ergodic theorem using the suggestive notation that we used in theorem 2.26.

Theorem 3.25. *Let $S_n : (X, \mathcal{B}) \to \mathbb{R}$, for $n \geq 1$, be a sequence of measurable functions representing elements in $L^1(X, \mathcal{B}, \mu)$[*] and satisfying the subadditivity condition*

$$S_{n+k}(x) \;\leq\; S_n(x) \;+\; S_k(T^n x) \qquad \mu\text{-a.e.}$$

for all $n, k \geq 1$. Let $A_n := \frac{1}{n} S_n$ for each $n \geq 1$. Then the pointwise limit

$$A(x) \;:=\; \lim_{n\to\infty} A_n(x) \tag{64}$$

exists μ-almost everywhere, and defines a T-invariant function. That is, the full measure set of points Ω on which the above limit exists satisfies $T\Omega = \Omega = T^{-1}\Omega$, and $A(Tx) = A(x)$ for all $x \in \Omega$. Moreover,

$$\int_X A \, d\mu \;=\; \inf_{n\geq 1} \int_X A_n \, d\mu. \tag{65}$$

Proof. We concentrate on proving (64), since then (65) follows using the Lebesgue dominated convergence theorem plus the following observations: consider the sequence of extended real numbers $I_n = \int_X S_n \, d\mu \in \mathbb{R} \cup \{-\infty\}$. Integrating the subadditivity condition gives us $I_{n+k} = I_n + I_k$ for all $n, k \geq 1$, so by Lemma 3.2 the limit of the averages $\lim_{n\to\infty} I_n/n$ exists in $\mathbb{R} \cup \{-\infty\}$ and coincides with $\inf_n I_n/n$. It remains then to prove (64), and this is where the work begins.

[*] Having said this, we will also use the statement of the Birkhoff ergodic theorem (version II) at the end of our proof. For a proof of the subadditive ergodic theorem that doesn't use Birkhoff's theorem see Avila and Bochi's notes [2].

[*] Actually it suffices to assume that S_1^+ (the positive part of S_1) is in L^1. Then it follows from the subadditivity that S_n^+ is in L^1 for all $n \geq 1$. With this the proof goes through - I need to check this.

Set

$$\left\{\begin{array}{lcl} \overline{A}(x) & := & \limsup_{n\to\infty} A_n(x) \\ \underline{A}(x) & := & \liminf_{n\to\infty} A_n(x) \end{array}\right.$$

for each $x \in X$. Since $\overline{A} \geq \underline{A}$ the limit $A(x) = \lim_{n\to+\infty} A_n(x)$ will exist precisely at those $x \in X$ for which

$$\overline{A}(x) \;\leq\; \underline{A}(x). \tag{66}$$

Let us now show that $\overline{A}$ and $\underline{A}$ are T-invariant almost everywhere, this will be useful shortly. Since $\overline{A}\circ T$ and $\overline{A}$ integrated with respect to μ over X yield the same values it suffices to show an inequality such as $\overline{A}\circ T \geq \overline{A}$. Similarly it suffices to show $\underline{A}\circ T \geq \underline{A}$. Subadditivity gives us $S_{n+1}(x) \leq S_n(Tx) + S_1(x)$ and therefore

$$\frac{n+1}{n} A_{n+1}(x) \;\leq\; A_n(Tx) \;+\; \frac{1}{n} S_1(x) \qquad \mu - a.e. \tag{67}$$

for each $n \geq 1$. Fix x for which (67) holds for all $n \geq 1$. Choose $n_j \to +\infty$ so that $A_{n_j+1}(x) \to \overline{A}(x)$. Then using (67),

$$\overline{A}(x) \;\leq\; \lim_{j\to+\infty} A_{n_j}(Tx) \;\leq\; \overline{A}(Tx)$$

giving us one of the required inequalities for $\overline{A}$. For the other inequality (for $\underline{A}$) choose instead $n_j \to +\infty$ so that $A_{n_j}(Tx) \to \underline{A}(Tx)$. Then from (67),

$$\underline{A}(x) \;\leq\; \lim_{j\to+\infty} A_{n_j+1}(x) \;\leq\; \underline{A}(Tx).$$

Thus we have shown that

$$\overline{A}\circ T \;=\; \overline{A} \tag{68}$$

$$\underline{A}\circ T \;=\; \underline{A} \tag{69}$$

hold pointwise μ-almost everywhere. It follows of course that A is T-invariant.

Therefore, to prove the theorem, it suffices only to prove that (66) holds μ-a.e and the rest of the statement will follow. As in the proof of theorems 2.26 and 2.30 we split the proof into two cases, the first essentially being a “warm up” to the second.

Fix $\varepsilon > 0$. By definition of $\underline{A}$ we have that $\forall x \in X\ \exists n \geq 1$ such that

$$A_n(x) \ \leq \ \underline{A}(x) \ + \ \varepsilon.$$

Thus the measurable function $\tau : X \to \mathbb{N}$ defined by

$$\tau(x) \ := \ \min\left\{ \ n \geq 1 \ \middle| \ A_n(x) \ \leq \ \underline{A}(x) \ + \ \varepsilon \ \right\}$$

is everywhere finite. Consider first the following simpler case.

Case 1: Suppose $\exists N, M \in \mathbb{N}$ such that $\tau(x) \leq N$ and $S_1(x) \leq M$ for all $x \in X$. To prove (66) holds μ-a.e. we argue as follows.

Let $x \in X$ be in the full measure set of points on which $\underline{A}$ is T-invariant, and $n \geq 1$. Consider the length-n orbit $x, \ldots, T^n x$. Decompose the orbit into pieces with initial points $x_0 = x, x_1, x_2, \ldots, x_q$, so for $0 \leq j \leq q-1$ the orbit piece starting at x_j has length $\tau(x_j)$ and the last piece (which starts at x_q) has length $0 \leq l \leq \tau(x_q) < N$.

This means that on each piece, with the possible exception of the last one, the average number of visits to B is $\leq \underline{A}(x_j) + \varepsilon = \underline{A}(x) + \varepsilon$ μ-a.e. using (69). For the last piece we don't have this upper bound but we know that its length is $l \leq N$.

The subadditivity of the functions $(S_n)_{n\geq 1}$ allows us to mimick the crucial lines (23) to (24) in the proof of Birkhoff's ergodic theorem version I (compare also lines (32) to (33) in version II) and estimate $S_n(x)$ as follows,

$$\begin{aligned}
S_n(x) \ &= \ \sum_{j=0}^{q-1} S_{\tau(x_j)}(x_j) \ + \ S_l(x_q) && (70)\\
&\leq \ \sum_{j=0}^{q-1} \tau(x_j) A_{\tau(x_j)}(x_j) \ + \ lM && (71)\\
&\leq \ \sum_{j=0}^{q-1} \tau(x_j)\Big(\underline{A}(x) + \varepsilon\Big) \ + \ NM \\
&\leq \ n\Big(\underline{A}(x) + \varepsilon\Big) \ + \ NM.
\end{aligned}$$

To go from line (70) to (71) we used that $S_1 \leq M$ implies that $S_n \leq nM$ for all $n \geq 1$. Indeed this follows easily from the subadditivity condition by induction using $S_{n+1}(x) \leq S_n(x) + S_1(T^n x)$.

Now, dividing through by n yields

$$A_n(x) \ \leq \ \underline{A}(x) \ + \ \varepsilon \ + \ \frac{NM}{n} \tag{72}$$

for all $x \in X$ and $n \geq 1$. Taking lim sup of both sides we have

$$\overline{A}(x) \ \leq \ \underline{A}(x) + \varepsilon$$

for all $x \in X$ and all $\varepsilon > 0$. In particular (66) follows (in fact *everywhere*), which completes the proof of Case 1.

Case 2: In which τ and or S_1 are unbounded from above.

The "bad" points which prevent us from directly applying the argument in case 1 are those for which τ is large or for which S_1 is large.

Remark 3.26. Note the following general fact comes in useful here. For any subadditive sequence of functions $(F_n)_{n\geq 1}$ it suffices that F_1 is uniformly bounded from above by a constant C say, and it follows that all the averages also satisfy $\frac{1}{n}F_n \leq C$. Indeed this follows easily by induction since $F_{n+1}(x) \leq F_n(x) + F_1(T^n x)$. Thus it suffices to modify our sequence $(S_n)_{n\geq 1}$ to another subadditive sequence $(S_n')_{n\geq 1}$ for which S_1' is bounded from above, and it will follow automatically that the averages $A_n' := \frac{1}{n}S_n'$ are uniformly bounded from above for all n.

With these points in mind pick $N, M \geq 1$ sufficiently large that the set

$$E \ := \ \left\{ \ x \in X \ \middle| \ \ \tau(x) > N \ \text{ or } \ S_1(x) > M \ \right\}$$

is small in the sense that

$$\int_E S_1 \, d\mu \ < \ \varepsilon.$$

We can do this because $S_1 \in L^1(X, \mathcal{B}, \mu)$ and τ is almost everywhere finite. Now modify the subadditive sequence $(S_n)_{n\geq 1}$ to another subadditive sequence $(S'_n)_{n\geq 1}$ on the “bad“ set E as follows. Indeed, mimicking the proof of theorem 2.30 we modify S_1 to

$$S'_1(x) \;:=\; \begin{cases} S_1(x) & \text{if } x \notin E \\ \min\big\{\, S_1(x)\,,\underline{A}(x)\,, M\,\big\} & \text{if } x \in E. \end{cases} \tag{73}$$

Observe that $S'_1 \leq S_1$, and moreover S'_1 is uniformly bounded since on the "bad" set E it is now bounded (by M). Now how to modify S_n for $n > 1$? First rewrite the above as

$$S'_1 \;=\; S_1 \;-\; h$$

for some h. Indeed, we see that $h : (X, \mathcal{B}) \to [0, +\infty)$ is given by

$$h \;=\; \Big(S_1 \;-\; \min\big\{\, S_1,\, \underline{A},\, M\,\big\}\Big)\chi_E.$$

Now the sequence of functions generated by h

$$h_n(x) \;:=\; \sum_{j=0}^{n-1} h\big(T^j x\big)$$

is an additive sequence. The pointwise sum of an additive with a subadditive sequence gives a subadditive sequence, thus the functions

$$S'_n \;:=\; S_n \;-\; h_n \tag{74}$$

for $n \geq 1$ is a subadditive sequence of functions, and S'_1 coincides with our original definition in (73). We will see that S'_n has advantages over S_n allowing us to draw nice conclusions about S'_n. At the end of the day, we can then draw nice conclusions about S_n because the two sequences differ just by the sequence h_n for which we have already proved existence of a pointwise limit almost everywhere by the Birkhoff ergodic theorem.

Set $A'_n := \frac{1}{n} S'_n$ for $n \geq 1$. In view of remark (3.26) we have the uniform bound

$$A'_n \;\leq\; M$$

for all $n \geq 1$, since $S'_1 \leq M$. Thus the sequence $(S'_n)_{n\geq 1}$ has no "bad" points, in the sense that the averages are uniformly bounded from above. It remains to modify τ so that it too is uniformly bounded. Observe that

$$S'_n \;\leq\; S_n$$

for all $n \geq 1$, in view of (74) and the fact that $h \leq 0$. Thus the function $\tau' : X \to \mathbb{N}$ defined by

$$\tau'(x) \;:=\; \min\left\{\, n \geq 1 \;\middle|\; A'_n(x) \;\leq\; \underline{A}(x) \;+\; \varepsilon \,\right\}$$

is finite everywhere, and indeed $\tau' \leq \tau$. Moreover we see that τ' is uniformly bounded (by N) since on the "bad" set E we have $\tau' = 1$. Indeed, if $x \in E$, then

$$A'_1(x) \;=\; S'_1(x) \;\leq\; \underline{A}(x).$$

Thus for our modified data $(S'_n)_{n\geq 1}$ and τ' there are no "bad“ points, and we can follow the argument from case 1.

Indeed, decomposing an arbitrary orbit $x, Tx, \ldots, T^n x$ into pieces whose length is determined by the function τ' leads to an estimate analogous to (72). Indeed one obtains

$$A'_n(x) \;\leq\; \underline{A}(x) \;+\; \varepsilon \;+\; \frac{NM}{n}$$

for all $x \in X$ and $n \geq 1$. Therefore,

$$A_n(x) \;\leq\; \underline{A}(x) \;+\; \varepsilon \;+\; \frac{NM}{n} \;+\; \frac{1}{n}h_n(x)$$

for all $x \in X$ and $n \geq 1$. By the Birkhoff ergodic theorem, version II, we know that $\frac{1}{n}h_n \to h_*$ almost everywhere, for some integrable function h_* satisfying

$$\int_X h_* \, d\mu \;=\; \int_X h \, d\mu.$$

Therefore,

$$\overline{A}(x) \;\leq\; \underline{A}(x) \;+\; \varepsilon \;+\; h_*(x)$$

for all $x \in X$. Now, integrating gives

$$\int_X \overline{A} \, d\mu \;\leq\; \int_X \underline{A} \, d\mu \;+\; 2\varepsilon$$

because $\int h_* d\mu = \int h d\mu \leq \int_E S_1 d\mu < \varepsilon$ by construction. This holds for all $\varepsilon > 0$, so

$$\int_X \overline{A} - \underline{A} \, d\mu \;\leq\; 0.$$

Combining with $\overline{A} - \underline{A} \geq 0$ we must have $\overline{A} - \underline{A} = 0$ almost everywhere. This completes the proof of the subadditive ergodic theorem. ■

4 Metric and topological entropy

We are studying the dynamics of a transformation T on a measurable space (X, Ω) in terms of invariant probability measures $\mathcal{M}_{\text{inv}}(T)$. According to Katok [] the single most important real number one can associate to $\mu \in \mathcal{M}_{\text{inv}}(T)$ is the so called *measure theoretic entropy* $h_\mu(T) \in [0, +\infty]^*$. This number is often described as a measure of the orbit complexity of the dynamical system, as seen/weighted through the measure μ. Another useful notion we will discuss is the *topological entropy* $h_{\text{top}}(T) \in [0, +\infty]$ which does not require an invariant measure.

Before we dive into the definitions here are two important theorems in the subject. The first, known as the variational principle, relates the metric and topological entropies via

$$h_{\text{top}}(T) \;=\; \sup_{\mu} h_\mu(T)$$

where the supremum is taken over all invariant (Borel) probability measures for T. We will probably not prove the variational principle, although we may well make use of it. However we will prove the following famous theorem which is rather a statement in smooth ergodic theory, and relates the metric entropy of μ to its Lyapunov exponents. One version of this, known variously as the Ruelle or Margulis-Ruelle or the Pesin-Ruelle inequality, states that for a sufficiently smooth transformation ($C^{1,\varepsilon}$ suffices) $f : M \to M$ on a smooth compact manifold there holds

$$h_\mu(f) \;\leq\; \int_M \Sigma_{\lambda \geq 0} \lambda \; d\mu \tag{75}$$

where the integrand at each point (defined μ-almost everywhere) is the sum of the positive Lyapunov exponents for f (counted with multiplicity) with respect to μ. More difficult to prove is the result that equality holds when μ is a smooth volume form on M (and f is a diffeomorphism), this is known as Pesin's formula. This raises the following natural question: for which measures in general is (75) equality? This leads to the concept of Sinai-Ruelle-Bowen measures (SRB-measures).

[*] Also called the *metric entropy*. The word "metric" here has nothing to do with distance, it's just an adjective from the noun "measure".

4.1 Motivation through the information function

The definition of metric entropy is complicated and its' motivation far from apparent. To explain the definition roughly let $\alpha = A_1, \dots, A_r$ be a finite measurable partition of X. Then one first defines $h_\mu(T,\alpha) \in [0,+\infty]$ the "entropy with respect to the partition α" and then we set

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

over all partitions. It remains to give and motivate the definition of $h_\mu(T,\alpha)$. One frequently used motivation for the definition leads to the interpretion of $h_\mu(T,\alpha)$ as follows: suppose $x \in X$ is chosen at random. Then $h_\mu(T,\alpha)$ will be a measure of the asymptotic amount of information, per iterate of T, you would expect to gain about the location of the points

$$x, Tx, \dots, T^n x$$

as $n \to \infty$, if you know the α-address of each point on the orbit, that is you know the sequence of sets $A_{k_0}, \dots, A_{k_n} \in \alpha$ so that $T^j x \in A_{k_j}$ for each $0 \le j \le n$. We will explain this shortly.

The information function

Let us begin by discussing, purely for motivation, what the concept of "information" gain could mean. At this point there is no dynamical system.

Suppose that (X, Ω, μ) is a probability space (X =a non-empty set, Ω =a σ-algebra on X, $\mu : \Omega \to [0,1]$ is a probability measure.) An "experiment" or "process" on this space means that a point in X is selected at random. A subset $A \in \Omega$ can be identified with the "event" that this process picks a point in A, and $\mu(A)$ is the probability that this event actually occurrs when the experiment is run once. Here is an interesting informal question we can ask about a given fixed $A \in \Omega$:

> *Suppose x is a point randomly selected from X. How much information about its precise location do we gain on learning that $x \in A$?* (1)

For example, if $A = X$ then we don't learn anything new about x, we could say that the amount of information gain is zero. If A is a single point, this point must be x and we learn the precise location of x, so we could say that the amount of information gain is maximal. Question (1) will be especially interesting when we consider a partition $A_1, \dots, A_r$ of X and ask the following:

> *Suppose x will be randomly selected from X and we will be told which A_i it lies in. How much information about the precise location of x would we expect to gain? (i.e. on average)* (2)

On the one hand the smaller $\mu(A_i)$ is the more information we would gain from being told that $x \in A_i$, but on the other hand if $\mu(A_i)$ is small then it seems less likely that x should turn up in A_i.

Let's address the first question. We are looking for a function $I = I_\mu : \Omega \to [0, +\infty]$ so that informally

$$I(A) = \text{the amount of information gained on learning that } x \in A.$$

The following properties of I, with the possible exception of (5), seem reasonable:

1. $\mu(A) = \mu(B) \implies I(A) = I(B)$ (information depends only on the measure of a set).

2. $\mu(A) = 1 \implies I(A) = 0$ (no information is gained about the location of x).

3. $\mu(A) = 0 \implies I(A) = M$ where $M := \sup I \in [0, +\infty]$ (we have maximum information about location of x).

4. $\mu(A) < \mu(B) \implies I(A) > I(B)$ (the smaller A is the better we can pinpoint the location of x).

5. $\mu(A \cap B) = \mu(A)\mu(B)$ (i.e. A and B are independent) $\implies I(A \cap B) = I(A) + I(B)$.

Property (1) means that we are really looking for a function $\phi : [0, 1] \to [0, +\infty]$ so that

$$I(A) = \phi(\mu(A))$$

for all $A \in \Omega$. Let us boldly assume further that the function ϕ is universal in the sense that it is independent of the probability space (X, Ω, μ). Then we seek a function

$$\phi : [0,1] \to [0,+\infty]$$

that is strictly decreasing, with $\phi(1) = 0$ and achieving a unique maximum at 0. Moreover by property (5) ϕ satisfies

$$\phi(ab) = \phi(a) + \phi(b) \tag{76}$$

for all $a, b \in [0,1]$ since given any such a, b it is easy to cook up a probability space having two independent sets of measure a and b. It follows of course that ϕ is uniquely given by

$$\phi(x) = -\log x$$

up to a positive multiple (or base of logarithm).

Let us attempt to justify property (5). First recall how one defines conditional probability: namely given two elements $A, B \in \Omega$ what is the natural model for the probability that event A happens given that B has occurred, i.e. a reasonable definition of

$$\mu(A \,|\, B) \;=\; \text{the probability of } A \text{, given } B. \quad ?$$

Suppose a point $z \in X$ is selected at random and we are told that $z \in B$. What is the probability that $z \in A$ also? Since we know $z \in B$ we can regard B as a probability space whose measurable sets are of the form $E = F \cap B$ for $F \in \Omega$, and with probability measure

$$\mu_B(E) := \frac{\mu(E)}{\mu(B)}.$$

Then the probability that z lies in $A \cap B$ with respect to this measure is $\mu_B(A \cap B) = \mu(A \cap B)/\mu(B)$. Therefore it seems reasonable to set

$$\mu(A \,|\, B) \;:=\; \frac{\mu(A \cap B)}{\mu(B)}. \tag{77}$$

Similarly, it makes sense to define conditional information gain: given $A, B \in \Omega$ we may ask what is a reasonable definition of

$$I(A \,|\, B) \;=\; \text{the info gain on learning } z \in A \text{, given } z \in B \quad ?$$

where $z \in X$ is selected at random. Since we know $z \in B$ we can work on the probability space obtained by restricting μ to measurable subsets of B and renormalizing so as to have a probability measure. That is, we obtain a measure space (B, Ω_B, μ_B) with measurable sets $E = F \cap B$ for all $B \in \Omega$, and measure given by renormalizing μ to the probability measure $\mu_B(E) := \mu(E)/\mu(B)$ as above. Let I_B denote the information function associated to (B, Ω_B, μ_B), that is,

$$I_B(E) = \phi(\mu_B(E)).$$

Then the information gain on learning that $z \in A \cap B$ on this measure space is $I_B(A \cap B) = \phi(\mu_B(A \cap B)) = \phi(\mu(A \cap B)/\mu(B)) = \phi(\mu(A|B))$. It therefore seems reasonable to set

$$I(A \,|\, B) \;:=\; \phi(\mu(A \,|\, B)). \tag{78}$$

Now consider two measurable sets A, B, and we ask what is the information about a randomly selected point $x \in X$ we would gain on learning that $x \in A \cap B$? Can we express this in terms of $I(A)$ and $I(B)$? Intuitively,

$$\begin{matrix}\text{Info gained} \\ \text{learning } {\scriptstyle x\in A\cap B}\end{matrix} = \begin{matrix}\text{Info gained} \\ \text{learning } {\scriptstyle x\in B}\end{matrix} + \begin{matrix}\text{Info gained learning} \\ {\scriptstyle x\in A\cap B} \text{ given } {\scriptstyle x\in B}\end{matrix}$$

in other words,

$$I(A \cap B) = I(B) + I(A \,|\, B). \tag{79}$$

The left hand side is symmetric in A and B but it's not clear that the right hand side is, with our definition (78). A natural class of sets for which the right hand side is symmetric in A and B is when they are independent. Recall that $A, B \in \Omega$ are called *independent events* in (X, Ω, μ) if $\mu(A|B) = \mu(A)$, i.e. any information that the event B has occurred is irrelevant in determining the probability that A occurs. Now from our definition (77) A and B are independent iff $\mu(A \cap B) = \mu(A)\mu(B)$. In this case, (78) becomes

$$I(A \,|\, B) \;=\; \phi(\mu(A)) \;=\; I(A)$$

and therefore (79) becomes

$$I(A \cap B) \;=\; I(B) + I(A)$$

which is also symmetric in A and B. All of this discussion hopefully motivates property (5) of the information function I.

Now, any function $I : \Omega \to [0, +\infty]$ satisfying properties (1) to (5) above must take sets of measure x to $-c \log x$, where $c > 0$ is any fixed constant (or base for logarithm). Setting this constant equal to 1 motivates the following definition.

Definition 4.1. Let (X, Ω, μ) be a probability space. We call $I : \Omega \to [0, +\infty]$ given by

$$I(A) := -\log \mu(A)$$

the *information function* on (X, Ω, μ). We can think of $I(A)$ as the amount of information one gains about the precise location of a point x chosen at random in X, if told that $x \in A$.

4.2 Entropy of a partition

Throughout (X, Ω, μ) is a probability space, at this stage there is still no dynamical system. A (finite) *partition* α of X is a finite collection of measurable sets $A_1, \ldots, A_k \in \Omega$ that are disjoint and whose union equals X. The α-address of a point $x \in X$ is the unique A_i such that $x \in A_i$. Recall question (2) above,

> *Suppose a point x will be randomly selected from X and we will be told its α-address. How much information about the precise location of x would we expect to gain (on average)?* (3)

We are basically asking for the expectation of the random variable $I(\alpha) : X \to \mathbb{R}$ which takes $x \in A_i$ to

$$\begin{aligned} I(\alpha)(x) &:= \text{Prob}(x \in A_i) \cdot \big(\text{Info gained on learning } x \in A_i\big) \\ &= \mu(A_i) I(A_i). \end{aligned}$$

That is, $I(\alpha) := -\sum_i \log(\mu(A_i))\chi_{A_i}$. Recall that the expectation of a random variable $\mathbf{X} : X \to \mathbb{R}$ is defined to be $E(\mathbf{X}) := \int_X \mathbf{X} d\mu$. Thus the answer to the above question should be $\int_X I(\alpha) d\mu$, and we call this the entropy of the partition.

Definition 4.2. The *entropy* of a finite partition $\alpha = \{A_1, \ldots, A_k\}$ of the probability space (X, Ω, μ) is

$$H(\alpha) := -\sum_i \log(\mu(A_i))\mu(A_i).$$

We can interpret $H(\alpha)$ as the expected information gain on learning the α-address of a randomly selected point in X, i.e. an answer to Question 3.

It is sometimes convenient to think of H as a function of probability vectors (i.e. of $p = (p_1, \ldots, p_k)$ with $p_i \in [0, 1]$ and $\sum_i p_i = 1$) given by

$$H : \left\{\begin{matrix}\text{probability vectors} \\ \text{of all lengths}\end{matrix}\right\} \longrightarrow [0, +\infty). \qquad H(p) = -\sum_i p_i \log p_i$$

Lemma 4.3. *H satisfies the following four conditions:*

1. *H is continuous on vectors of fixed length, and $H(0, p_1, \ldots, p_k) = H(p_1, \ldots, p_k)$.*
2. *Symmetric under permutations of entries.*
3. *$H(1) = 0$ is the unique minimum.*
4. *$H(1/k, \ldots, 1/k) = \log k$ is the unique maximum over vectors of length k.*

Proof. Properties (1)-(3) are obvious. To prove (4) set $\phi : [0, 1] \to [0, +\infty)$ $x \mapsto -x \log x$. Then $\phi''(x) = -1/x < 0$ so ϕ is strictly convex downwards.

Now Jensen's inequality applied to ϕ gives

$$\frac{1}{k}\sum_i \phi(p_i) \;\le\; \phi\left(\frac{1}{k}\sum_i p_i\right) \;=\; \phi(1/k) = \frac{1}{k}\log k$$

as required. ■

Note that the fourth property in the above Lemma is equivalent to

$$H(\alpha) \;\leq\; \log \operatorname{card}(\alpha) \tag{80}$$

for each partition α of X.

4.3 Entropy of a transformation

Now we can define the entropy of a transformation T. Suppose $T : (X, \Omega) \to (X, \Omega)$ is a measurable transformation on a measurable space, and $\mu \in \mathcal{M}(T)$, i.e. is an invariant probability measure for T.

First, define the *join of* two partitions α and β to be the partition

$$\alpha \vee \beta := \left\{ \; A \cap B \;\; \middle| \; A \in \alpha, \;\; B \in \beta \; \right\}.$$

For a partition α of X define the pull-back partition

$$T^{-1}\alpha \;:=\; \left\{ \; T^{-1}A \; \middle| \; A \in \alpha \; \right\}.$$

Definition 4.4. For α a finite partition of X we define the **entropy of T with respect to μ and α** as the quantity

$$h_\mu(T;\alpha) := \lim_{n\to+\infty} \frac{1}{n} H\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha \right). \tag{81}$$

The **measure theoretic entropy of T with respect to μ** is

$$h_\mu(T) := \sup_\alpha h_\mu(T;\alpha) \;\in\; [0,+\infty].$$

We will justify shortly, see Corollary 4.18, that the limit in (81) exists in $[0, +\infty)$.

Definition 4.5 (Measure-theoretic isomorphism)**.** Suppose $T_i : (X_i, \Omega_i, \mu_i) \to (X_i, \Omega_i, \mu_i)$ are measure preserving transformations for $i = 1, 2$. We say T_1 and T_2 are **measure-theoretically isomorphic** if there exists an invertible measure preserving transformation between invariant subsets of full measure $\varphi : X_1' \to X_2'$ with $\varphi \circ T_1 = T_2 \circ \varphi$. In this case, by invariant is meant that $T_i X_i' \subset X_i'$.

Lemma 4.6. *The metric entropy is an invariant under measure theoretic isomorphism.*

Proof. From the definitions. ■

Exercise 4.7. Show that $h_\mu(T^k) = kh_\mu(T)$ for all $k \in \mathbb{N}$. This is actually very straightforward from the definition. If T is invertible show that $h_\mu(T^{-1}) = h_\mu(T)$.

Dynamical interpretation of entropy

Let's pause for a moment to think about what the value of $h_\mu(T;\alpha)$ (and hence also $h_\mu(T)$, is measuring, with a view to heuristically justifying the slogan that higher entropy of T implies more unpredictable dynamics.

First observe that knowing the $\vee_{i=0}^{n-1} T^{-i}\alpha$-address of x is equivalent to knowing the α-addresses of $x, Tx, \dots, T^{n-1}x$. Indeed,

$$x \;\in\; A_{i_0} \cap T^{-1}A_{i_1} \cap \ldots \cap T^{-(n-1)}A_{i_{n-1}} \;\in\; \bigvee_{i=0}^{n-1} T^{-i}\alpha$$

if and only if

$$x \in A_{i_0}, \quad Tx \in A_{i_1}, \quad \ldots \quad T^{(n-1)}x \in A_{i_{n-1}}. \tag{82}$$

Now we give two interpretations of the entropy of T.

Interpretation 1 From (82) we can interpret $H_\mu\left(\vee_{i=0}^{n-1} T^{-i}\alpha\right)$ as

> *the expected **information gain** regarding (the precise locations of) the points $x, Tx, \dots, T^{n-1}x$, for a randomly chosen $x \in X$, on learning of the α-addresses of the points $x, Tx, \dots, T^{n-1}x$.*

Thus $h_\mu(T;\alpha)$ is the asymptotic expected information gain per iterate for the partition α, and $h_\mu(T)$ the maximum information gain one could hope to obtain, per iterate, for large iterates.

Interpretation 2 Another way to interpret the entropy of T is as an amount of uncertainty with which one expects to be able to predict its orbits. Indeed, consider a partition $\alpha = \{A_1, \dots, A_k\}$ of a probability space (X, Ω, μ). One could ask the following question.

Suppose a point x is randomly selected from X. How much uncertainty is there predicting its α-address? (4)

I feel that the meaning of this question is less clear than questions (1),(2),(3) above so let me try to explain more, even though it is all only heuristic. Think of α as modelling the outcome of an experiment; the experiment is modelled by randomly picking a point $x \in X$. The outcome is the α-address of x, and the probability that the outcome is A_i is of course $\mu(A_i)$. If for example α is a single element then the outcome is certain and we can say that the uncertainty of α is zero. To see what the uncertainty could be more generally, we note that it is reasonable that the uncertainty in predicting the outcome of the experiment is the same as the amount of information one would expect to gain by learning the outcome of the experiment!

$$\begin{array}{c}\text{Uncertainty predicting}\\ \text{outcome of an experiment}\end{array} = \begin{array}{c}\text{Info gained}\\ \text{on learning outcome}\end{array} \tag{83}$$

(For a simple minded example, if we throw a die that has the number 5 on every side then the outcome can be predicted with complete certainty to be 5, that is the uncertainty in the outcome is 0. Clearly we also gain no information on learning the outcome is a 5, that is the information gained is also 0.) If we accept (83), then we can interpret the entropy of a partition α as the amount of uncertainty in predicting the α-address of a randomly selected point $x \in X$. It then follows that $H_\mu\left(\vee_{i=0}^{n-1} T^{-i}\alpha\right)$ is

*the **amount of uncertainty** in predicting the α-addresses of the points $x, Tx, \dots, T^{n-1}x$ for a randomly chosen $x \in X$.*

Thus we can interpret $h_\mu(T)$ as the asymptotic rate at which the unpredictability of the length-n orbits of T increases with n.

First examples

Example 4.8. $T = \mathrm{Rot} : S^1 \to S^1$, which preserves lebesgue measure μ. Let α be any partition of S^1. Then $\mathrm{card}(\alpha \vee T^{-1}\alpha) \leq \mathrm{card}(\alpha) + \mathrm{card}(T^{-1}\alpha) = 2\mathrm{card}(\alpha)$ (draw a picture).

Thus $\mathrm{card}(\alpha \vee T^{-1}\alpha \vee \cdots \vee T^{-(n-1)}\alpha) \leq n\,\mathrm{card}(\alpha)$, so by (80),

$$H_\mu\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right) \leq \log(n\,\mathrm{card}(\alpha)).$$

Thus $h_\mu(\mathrm{Rot};\alpha) = 0$ for all α, so $h_\mu(\mathrm{Rot}) = 0$.

Example 4.9 (Homeomorphisms of S^1)**.** If $T : S^1 \to S^1$ is a homeomorphism then it has a rotation number $\rho(T) \in \mathbb{R}/\mathbb{Z}$. If $\rho(T)$ is rational then $h_\mu(T) = 0$ because.... If $\rho(T) = \omega$ is irrational then T may not be topologically conjugate to the rotation $R_{2\pi\omega}$ but it will be uniquely ergodic, i.e. has a unique invariant probability measure μ, and it is *measure theoretically* conjugate to $R_{2\pi\omega}$ with lebesgue measure. Consequently by the previous example $h_\mu(T) = 0$.

Example 4.10 (Angle-doubling map)**.** Let $T : S^1 \to S^1$ be the non-invertible map $z \mapsto z^2$. This is the same as $T : [0,1] \to [0,1]$ $x \mapsto 2x \mod 1$ which we saw before preserves μ =lebesgue measure. Let us see why $h_\mu(T) > 0$.

Take $\alpha = \{[0,1/2),\, [1/2,1)\}$. Then

$$\alpha \vee T^{-1}\alpha = \{[0,1/4),\, [1/4,1/2),\, [1/2,3/4),\, [3/4,1)\}$$

and more generally $\alpha \vee T^{-1}\alpha \vee \cdots \vee T^{-(n-1)}\alpha$ is a partition of $[0,1]$ into 2^n intervals of equal length $1/2^n$.

Thus

$$H\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right) = \log \operatorname{card}\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right) = n\log 2$$

so that

$$h_\mu(T;\alpha) = \log 2.$$

Hence $h_\mu(T) \geq h_\mu(T;\alpha) = \log 2 > 0$. We will shortly show the reverse inequality also to find $h_\mu(T) = \log 2$.

Example 4.11 (Entropy of Bernoulli shift)**.** Let us show that the entropy of the $(1/2, 1/2)$-Bernoulli process is at least $\log 2$. Recall

- X is the set of sequences $\underline{x} = (x_1, x_2, \ldots)$ where $x_i \in \{0,1\}$,
- $T : X \to X$ takes $(x_1, x_2, x_3 \ldots)$ to $(x_2, x_3 \ldots)$,
- $m\big(A_{i_1,\ldots,i_r}\big) = 1/2^r$ where

  $$A_{i_1,\ldots,i_r} := \big\{\ \underline{x}\ \big|\ x_1 = i_1,\ x_2 = i_2 \ldots\ , x_r = i_r\big\} = 1/2^r$$

  for each $r \in \mathbb{N}$.

As our partition of X let us take

$$\alpha = \{A_0, A_1\}.$$

Then for example

$$\alpha \vee T^{-1}\alpha = \Big\{A_i \cap T^{-1}A_j\Big\}_{i,j} = \Big\{A_{i,j}\Big\}_{i,j}.$$

More generally

$$\alpha \vee T^{-1}\alpha \vee\ \cdots\ \vee T^{-(n-1)}\alpha\ =\ \Big\{A_{i_1,\ldots,i_n}\Big\}_{i_1,\ldots,i_n},$$

so

$$H\left(\bigvee_{k=0}^{n-1} T^{-k}\alpha\right)\ =\ n\log 2$$

giving the lower bound

$$h(T)\ \geq\ h(T;\alpha)\ \geq\ \log 2.$$

We will see shortly that $h(T) = \log 2$.

Example 4.12. More generally a similar calculation shows that the $(p_1, \ldots, p_s)$-Bernoulli process has entropy at least $\sum_i p_i \log p_i$, and again this can be shown to be equality when we have another lemma or two.

Lemma 4.13. *The angle-doubling map $S^1 \mapsto S^1$ $z \mapsto z^2$, where S^1 is the boundary of the unit disk and equipped with the Lebesgue measure, is measure-theoretically isomorphic to the Bernoulli-*$(1/2, 1/2)$ *process.*

Proof. Exercise. ■

4.4 Conditional entropy

We return to a further study of the definition of entropy for a partition and also that of a transformation, now that we have some motivation.

First let us extend the definition of entropy of a partition to that of conditional entropy. It's a useful definition. For example it allows us to compare the entropy of different partitions. It also allows us to rephrase the definition of entropy of a transformation in a convenient manner.

Throughout, (X, Ω, μ) is a probability space, and $\alpha = \{A_1, \ldots, A_k\}$ and $\beta = \{B_1, \ldots, B_l\}$ are finite partitions.

Definition 4.14. The *conditional entropy of β given α* is the quantity

$$H(\beta \,|\, \alpha) := -\sum_i \mu(A_i) \sum_j \mu(B_j|A_i) \log(\mu(B_j|A_i)).$$

One can arrive at this expression by interpreting $H(\beta \,|\, \alpha)$ as the expected information gain from the partition β that you didn't already know from α. Or more precisely; suppose that a point $x \in X$ is selected at random by someone else and you can't see which point it is. They then tell you of the α-address of x, and this gives you some information regarding its location. Then they tell you of the β-address of x, and this gives you possibly some additional information about its location. If you did this repeatedly then

$H(\beta|\alpha)$ is the amount of additional information you would expect to gain, on average, when told of the β-address.

For example, if $\beta = \alpha$ then we learn nothing new from the β-address so intuitively $H(\alpha\,|\,\alpha) = 0$ as is easily verified. More generally if α is a refinement of β meaning that for all $A_i \in \alpha$ there exists $B_j \in \beta$ with $A_i \subset B_j$, then we learn nothing more from β so that intuitively $H(\beta\,|\,\alpha) = 0$.

Definition 4.15. Let α, β, be two (finite) partitions of (X, Ω, μ). We write:

- $\alpha < \beta$ if for all $B \in \beta$ there exist $A \in \alpha$ with $B \subset A \mod 0$ (meaning $\mu(B \cap A^c) = 0$). We call β a **refinement** of α.
- $\alpha \perp \beta$ if $\mu(A \cap B) = \mu(A)\mu(B)$ for all $A \in \alpha$ and $B \in \beta$. We say α and β are **independent**.
- $\alpha \stackrel{\circ}{=} \beta$ if for all $A \in \alpha$ there exists $B \in \beta$ so that $\mu(A\Delta B) = 0$. We say α and β are **equivalent**.

Lemma 4.16. *For (finite) partitions α, β, γ,*

1. *$H(\alpha \vee \beta) = H(\alpha) + H(\beta\,|\,\alpha)$, and more generally*

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

2. *$H(\beta\,|\,\alpha) \leq H(\beta)$ with equality iff $\alpha \perp \beta$.*
3. *$\alpha < \beta$ implies $H(\alpha) \leq H(\beta)$ with equality iff $\alpha \stackrel{\circ}{=} \beta$.*
4. *More generally, $H(\beta\,|\,\alpha)$ is "increasing" in β and "decreasing" in α, in the sense that*

$$\begin{aligned} \beta < \beta' &\implies H(\beta\,|\,\alpha) \leq H(\beta'\,|\,\alpha) \\ \alpha < \alpha' &\implies H(\beta\,|\,\alpha) \geq H(\beta\,|\,\alpha'). \end{aligned}$$

Proof. We prove the first part of (1) and leave the others as exercises.

$$\begin{aligned}
H(\beta \,|\, \alpha) &= -\sum_i \mu(A_i) \sum_j \mu(B_j|A_i) \log(\mu(B_j|A_i)) \\
&= -\sum_{i,j} \mu(A_i \cap B_j) \log\left(\frac{\mu(A_i \cap B_j)}{\mu(A_i)}\right) \\
&= -\sum_{i,j} \mu(A_i \cap B_j) \log(\mu(A_i \cap B_j)) \;+\; \sum_{i,j} \mu(A_i \cap B_j) \log(\mu(A_i)) \\
&= H(\alpha \vee \beta) \;+\; \sum_i \sum_j \mu(A_i \cap B_j) \log(\mu(A_i)) \\
&= H(\alpha \vee \beta) \;+\; \sum_i \mu(A_i) \log(\mu(A_i))
\end{aligned}$$

because the sets $\{(A_i \cap B_j)\}_j$ form a partition of A_i,

$$= H(\alpha \vee \beta) \;-\; H(\alpha).$$

A couple of remarks on (2)-(4). The proof of (2) uses Jensen. Observe that the inequality makes intuitive sense, since both $H(\beta)$ and $H(\beta|\alpha)$ measure the expected information gain on learning the same fact, namely the β-address of a random point. However in $H(\beta|\alpha)$ we start off with additional information, namely the α-address. So the amount of information gained cannot be more than if we knew nothing before hand. The equality part of (2) is straightforward. Intuitively $\alpha \perp \beta$ means that any information from α is irrelevant in determining the uncertainty of the outcome of β, so $H(\beta \,|\, \alpha) = H(\beta)$.

Note (3) follows immediately from (1) using that $\alpha < \beta$ implies $\beta = \alpha \vee \beta$. ■

As promised, we can now prove that the limit in (81) exists, so that $h_\mu(T; \alpha)$ is well defined.

Lemma 4.17. *The sequence*

$$a_n \;:=\; H\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right)$$

is subadditive*, that is $a_{n+k} \leq a_n + a_k$ for all $n, k \in \mathbb{N}$.*

Proof.

$$\begin{aligned} a_{n+m} &= H\left(\vee_{i=0}^{n+m-1} T^{-i}\alpha\right) \\ &= H\left(\vee_{i=0}^{m-1} T^{-i}\alpha\right) \;+\; H\left(\vee_{i=m}^{n+m-1} T^{-i}\alpha \,\middle|\, \vee_{i=0}^{m-1} T^{-i}\alpha\right). \end{aligned}$$

The first term is a_m, and for the second term we have

$$\begin{aligned} H\left(\vee_{i=m}^{n+m-1} T^{-i}\alpha \,\middle|\, \vee_{i=0}^{m-1} T^{-i}\alpha\right) &\le H\left(\vee_{i=m}^{n+m-1} T^{-i}\alpha\right) \\ &= H\left(\vee_{i=0}^{n-1} T^{-i}\alpha\right) = a_n. \end{aligned}$$

■

Corollary 4.18. *$h_\mu(T;\alpha)$ is well defined and is equal to* $\inf_{n\ge 1} \frac{1}{n} H\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right)$.

Proof. The sequence $a_n = H\left(\vee_{i=0}^{n-1} T^{-i}\alpha\right)$ is subadditive and bounded from below (by zero), so the statement follows from Lemma 3.2. ■

We can now express the entropy of a transformation in terms of conditional entropy.

Proposition 4.19 (Equivalent definition of $h_\mu(T;\alpha)$)**.**

$$h_\mu(T;\alpha) \;=\; \lim_{n\to+\infty} H\left(\alpha \,\middle|\, \bigvee_{1}^{n} T^{-i}\alpha\right). \tag{84}$$

Moreover the sequence on the right hand side is monotonic decreasing.

Proof. Decreasing is easy: as n increases the partitions $\vee_1^n T^{-i}\alpha$ get finer, so by property (4) in Lemma 4.16 $H(\alpha | \vee_1^n T^{-i}\alpha)$ decreases.

Now to prove (84). Let $k \in \mathbb{N}$ be arbitrary.

$$\begin{aligned} H\big(\vee_0^k T^{-i}\alpha\big) &= H\Big(\alpha \vee \big(\vee_1^k T^{-i}\alpha\big)\Big) \\ &= H\Big(\alpha \,\Big|\, \vee_1^k T^{-i}\alpha\Big) \;+\; H\big(\vee_1^k T^{-i}\alpha\big). \end{aligned}$$

Thus,

$$
\begin{aligned}
H\Big(\alpha \Big| \vee_1^k T^{-i}\alpha\Big) &= H\big(\vee_0^k T^{-i}\alpha\big) - H\big(\vee_1^k T^{-i}\alpha\big) \\
&= H\big(\vee_0^k T^{-i}\alpha\big) - H\big(\vee_0^{k-1} T^{-i}\alpha\big).
\end{aligned}
$$

Summing both sides over k, the right hand side "telescopes" to give

$$
\sum_{k=1}^{n} H\Big(\alpha \Big| \vee_1^k T^{-i}\alpha\Big) = H\big(\vee_0^n T^{-i}\alpha\big) - H\big(\alpha\big)
$$

for all $n \in \mathbb{N}$. Dividing through by n and letting $n \to +\infty$ the left hand side converges in the Cesaro sense to $\lim_{n\to\infty} H(\alpha \mid \vee_1^n T^{-i}\alpha)$, while the right hand side converges to the definition of $h_\mu(T;\alpha)$. ■

How to interpret this alternative definition of entropy? For each $n \in \mathbb{N}$,

$$
H\left(\alpha \Big| \bigvee_1^n T^{-i}\alpha\right)
$$

is the expected additional information we would gain about the location of a random point $x \in X$ on learning its α-address, if we already knew in advance the α-addresses of the points $Tx, T^2x, \ldots, T^nx$.

In particular, we can interpret $h_\mu(T;\alpha) = 0$ as saying that the future, as given by T, determines the present. This suggests that zero entropy of T implies that T is somehow invertible? Indeed this is almost true provided the measurable space is not pathological, for example a Borel measure space on a reasonable metric space or a so called Lebesgue space. For the statement we say that a measure preserving transformation on a probability space is *invertible-* mod 0 if it restricts to an invertible transformation on some invariant subset of full measure.

Theorem 4.20. *$h_\mu(T) = 0$ implies that T is invertible-* mod 0 *if (X,Ω) is a complete, separable metric space with the Borel σ-algebra.*

Proof. See Corollary 4.14.3 in [19]. ■

Conversely, if T is not invertible- mod 0 then $h_\mu(T) > 0$.

For later let us record a couple of basic properties of the conditional entropy.

Lemma 4.21. *For (finite) partitions α, β there holds,*

1. $h(T, \alpha) \leq H(\alpha)$.
2. $\alpha < \beta$ *implies* $h(T, \alpha) \leq h(T, \beta)$.
3. $h(T, \alpha \vee \beta) \leq h(T, \alpha) + H(\beta \,|\, \alpha)$.
4. $h(T, \alpha) = h(T, \vee_{i=0}^{k} T^{-i}\alpha)$ *for any* $k \in \mathbb{N}$.

We suppressed the dependence of h and H on the invariant probability measure μ.

Proof. (1) follows easily from the equivalent definition of entropy using conditional entropy, Proposition 4.19, plus the fact that $H(\alpha \,|\, \beta) \leq H(\alpha)$ for all β.

(2) follows from the definition of $h(T, \alpha)$ and $h(T, \beta)$ and that $\alpha < \beta$ implies $\vee_0^n T^{-i}\alpha < \vee_0^n T^{-i}\beta$ for all n.

(3) is slightly tricky: from the definition $h(T, \alpha \vee \beta) = \lim_{n\to\infty} \frac{1}{n} H\big(\vee_0^{n-1} T^i(\alpha \vee \beta)\big)$. Examining the right hand side we have

$$\begin{aligned} H\Big(\vee_0^{n-1} T^i(\alpha \vee \beta)\Big) &= H\Big(\big(\vee_0^{n-1} T^i\alpha\big) \vee \big(\vee_0^{n-1} T^i\beta\big)\Big) \\ &= H\big(\vee_0^{n-1} T^i\alpha\big) + H\Big(\vee_0^{n-1} T^i\beta \,\Big|\, \vee_0^{n-1} T^i\alpha\Big). \end{aligned}$$

Since $\frac{1}{n}$ times the first term converges to $h(T, \alpha)$, it remains to show that $\frac{1}{n}$ times the second term is bounded above by $H(\beta \,|\, \alpha)$. From Lemma 4.16 (1) $H(\alpha_1 \vee \alpha_2 \,|\, \beta) \leq H(\alpha_1 \,|\, \beta) + H(\alpha_2 \,|\, \beta)$. Thus

$$H\Big(\vee_0^{n-1} T^i\beta \,\Big|\, \vee_0^{n-1} T^i\alpha\Big) \leq \sum_{j=0}^{n-1} H\Big(T^j\beta \,\Big|\, \vee_0^{n-1} T^i\alpha\Big)$$

for all i,

$$\begin{aligned} &\leq \sum_{j=0}^{n-1} H\Big(T^j\beta \,\Big|\, T^j\alpha\Big) \\ &= nH\big(\beta \,|\, \alpha\big). \end{aligned}$$

And (3) follows.

(4) follows easily from

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

Dividing by n and sending $n \to \infty$ the left hand side goes to $h(T, \vee_{i=0}^{k} T^{-i}\alpha)$. This is the same as dividing by $N := n + k - 1$ and sending $N \to \infty$, in which case the right hand side goes to $h(T, \alpha)$. ■

4.5 The Kolmogorov-Sinai theorem on generating σ-algebras

This is a useful tool to avoid taking a supremum over all partitions when calculating the entropy. Recall that any collection of subsets $\mathcal{S}$ of X generates a σ-algebra $\sigma(\mathcal{S})$ that is the smallest σ-algebra containing $\mathcal{S}$. In particular, if α is a finite partition of X then

$$\sigma(\alpha) := \text{the } \sigma\text{-algebra generated by } \alpha$$

is a finite σ-algebra on X. If $\alpha_1 < \alpha_2 < \cdots$ is a sequence of (finite) partitions of X, let us write

$$\bigvee_{i=0}^{+\infty} \alpha_i := \text{the smallest } \sigma\text{-algebra containing all the } \alpha_n\text{'s.}$$

Remark 4.22. Obviously this definition would be fine for any sequence of finite partitions $\{\alpha_n\}_{n\in\mathbb{N}}$, but if they are increasingly finer then $\vee_{i=m}^{+\infty}\alpha_i = \vee_{i=0}^{+\infty}\alpha_i$ for all $m \geq 0$ and so we can think of $\vee_{i=0}^{+\infty}\alpha_i$ as the *limit* of the σ-algebras $\sigma(\alpha_n)$.

Remark 4.23. Note that there is an obvious $1-1$ correspondence between finite partitions and finite sub-σ-algebras on (X, Ω):

$$\begin{matrix} \left\{\begin{matrix}\text{finite}\\ \text{partitions}\end{matrix}\right\} & \longleftrightarrow & \left\{\begin{matrix}\text{finite}\\ \sigma\text{−algebras}\end{matrix}\right\} \\ \alpha & \longleftrightarrow & \sigma(\alpha) \end{matrix}$$

(To go from right to left take all non-empty intersections from sets in $\sigma(\alpha)$ and the complement of these). However it's better to take the limit in the sense of σ-algebras; we risk losing information if we take a limit of partitions because while each σ-algebra gives rise to a unique partition, an infinite partition can come from many different σ-algebras.

As before, throughout T is a measurable transformation on the probability space (X, Ω, μ).

Proposition 4.24 (Kolmorgorov-Sinai). *Let $\alpha_1 < \alpha_2 < \cdots$ be a sequence of partitions of (X, Ω, μ) with "limit" $\vee_{i=0}^{\infty} \alpha_i \doteq \Omega$. Then*

$$h_\mu(T) = \lim_{n \to \infty} h_\mu(T; \alpha_n). \tag{85}$$

Note that since the sequence in (85) is monotonic we also have $h_\mu(T) = \sup_{n \geq 1} h_\mu(T; \alpha_n)$. To prove this Proposition we will use:

Lemma 4.25 (Approximation Lemma). *Fix $r \in \mathbb{N}$. Then for all $\varepsilon > 0$ there exists $\delta = \delta(r) > 0$ so that if $\alpha = \{A_1, \ldots, A_r\}$ and $\beta = \{B_1, \ldots, B_r\}$ are partitions of X with $\mu(A_i \,\Delta\, B_i) < \delta$ for all $1 \leq i \leq r$, then $H(\beta \,|\, \alpha) < \varepsilon$.*

The conclusion implies that $H(\alpha)$ and $H(\beta)$ are close because $H(\alpha) + H(\beta \,|\, \alpha) = H(\alpha \vee \beta) = H(\beta) + H(\alpha \,|\, \beta)$). But smallness of $H(\beta \,|\, \alpha)$ is stronger than closeness of $H(\alpha)$ and $H(\beta)$, as it says that α and β are close as partitions.

Proof. The trick is to observe that there is a partition γ so that $H(\beta \,|\, \alpha) = H(\gamma \,|\, \alpha)$ and where $H(\gamma) < \varepsilon$. Of course γ need not have cardinality r. From this the Lemma will follow as $H(\gamma \,|\, \alpha) \leq H(\gamma)$.

Take γ to be the following partition of X:

$$\gamma := \{\, C \,\} \cup \{\, A_i \cap B_j \,\big|\, i \neq j \,\}$$

where

$$C := \bigcup_{i=1}^{r} A_i \cap B_i.$$

That is, γ is the partition obtained from $\alpha \vee \beta$ by combining all the $A_i \cap B_j$ sets for which $i = j$ into one set. Now clearly $\alpha \vee \beta \;=\; \alpha \vee \gamma$, and therefore $H(\beta \,|\, \alpha) = H(\gamma \,|\, \alpha)$ because

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

It remains then to see why γ has small entropy; this is because it consists of a single set C of almost full measure, and all the other sets have very small measure, and the function $x \mapsto -x \log x$ is small near both 0 and 1. More precisely, since $\mu(A_i \cap B_j) < \delta$ for all $i \neq j$, and $\mu(C) > 1 - r\delta$,

$$\begin{aligned} H(\gamma) \;&\leq\; -\mu(C) \log \mu(C) \quad - \quad \sum_{i \neq j} \mu(A_i \cap B_j) \log \mu(A_i \cap B_j) && (86) \\ &\leq\; -(1 - r\delta) \log(1 - r\delta) \quad - \quad r(r-1)\delta \log \delta. && (87) \end{aligned}$$

For fixed r the right hand side goes to zero as $\delta \to 0$. ■

We can now prove the theorem of Kolmogorov and Sinai:

Proof. (Of Proposition 4.24) Fix $\varepsilon > 0$. Pick a finite partition β so that

$$h_\mu(T, \beta) \;\geq\; h_\mu(T) \;-\; \varepsilon.$$

Set $r := \operatorname{card}(\beta)$ and let $\delta = \delta(r) > 0$ be as in the approximation lemma. Claim: it follows from $\vee_{i=0}^{\infty} \alpha_i \doteq \Omega$ that there exists $n \in \mathbb{N}$ and $\alpha < \alpha_n$ that is a δ-good approximation of β, meaning that $\operatorname{card}(\alpha) = r$ and

$$\mu(A_i \,\Delta\, B_i) \;<\; \delta$$

for all $1 \leq i \leq r$, where $\alpha = \{A_1, \ldots, A_r\}$. (Think about this). It follows by the approximation lemma that $H(\beta \,|\, \alpha) < \varepsilon$. Thus,

$$h_\mu(T, \beta) \leq h_\mu(T, \alpha \vee \beta) \leq h_\mu(T, \alpha) + H(\beta \,|\, \alpha) \leq h_\mu(T, \alpha_n) + \varepsilon.$$

The second from last inequality here used Lemma 4.21. So

$$h_\mu(T, \alpha_n) \geq h_\mu(T) - 2\varepsilon.$$

This establishes (85). ■

We now have the following useful corollary to the Kolmogorov-Sinai theorem, which sometimes allows to avoid taking a supremum over all partitions when calculating the entropy of a transformation. First we need:

Definition 4.26 (Generating partition)**.** Say that a (finite) partition α of X is a **generator** for T if

$$\bigvee_{n=0}^{+\infty} \alpha_n \ \overset{\circ}{=} \ \Omega$$

where $\alpha_n := \vee_{i=0}^{n} T^{-i}\alpha$.

Corollary 4.27. *If α is a (finite) generating partition for T then*

$$h_\mu(T) = h_\mu(T;\alpha). \tag{88}$$

Proof. For each $n \in \mathbb{N}$ set $\alpha_n := \vee_0^n T^{-i}\alpha$. Then by Proposition 4.24

$$h(T) \ = \ \lim_{n\to+\infty} h(T,\alpha_n). \tag{89}$$

However, for each fixed $n \geq 1$ we have

$$h(T,\alpha_n) \ = \ h\left(T,\, \vee_{i=0}^{n} T^{-i}\alpha\right) \ = \ h(T,\alpha)$$

by (4) of Lemma 4.21. So the sequence in (89) is constant and equal to $h(T,\alpha)$. ■

Example 4.28. Let $T : [0,1] \to [0,1]$ be $x \mapsto 2x \mod 1$ which we saw before preserves μ =lebesgue measure. Then we claim that

$$\alpha = \big\{[0,1/2),\, [1/2,1)\big\}$$

is a generating partition. Indeed, the ambient σ-algebra is the Borel-σ-algebra on $[0,1]$. Thus it suffices to generate every open interval $(a,b) \subset [0,1]$.

For each $n \in \mathbb{N}$, we saw before that

$$\alpha_n := \bigvee_{i=0}^{n} T^{-i}\alpha = \Big\{ \ [0,1/2^n),\ [1/2^n,2/2^n),\ \ \ldots\ \ \Big\}$$

is the partition of $[0,1]$ into 2^n intervals of length $1/2^n$. Thus if $(a,b) \subset [0,1]$ is an arbitrary open interval we easily find intervals of the form

$$(a_n, b_n) \in \sigma(\alpha_n)$$

with a_n decreasing to a and b_n increasing to b as $n \to \infty$. Thus $(a,b) = \cup_n^\infty (a_n, b_n)$ lies in $\vee_{i=0}^\infty T^{-i}\alpha$.

Let us use a generating partition to calculate the entropy of Bernoulli processes precisely, showing that the lower bound we got in Example 4.11 is infact equality.

Example 4.29 (Entropy of Bernoulli shift again)**.** Let us now show that the entropy of the $(p_1, \dots, p_s)$-Bernoulli process equals $\sum_i p_i \log p_i$, in particular the $(1/2, 1/2)$-shift has entropy $\log 2$.

Recall that X is the space of half infinite sequences $\underline{x} = (x_1, x_2, \dots)$ with values in $\{1, \dots, s\}$. Take $\alpha = \{A_1, \dots, A_s\}$ to be the partition of X determined by the 1-st coordinate:

$$A_r := \left\{ \ \underline{x} = (x_1, x_2, \dots) \ \middle| \ x_1 = r \ \right\}.$$

Arguing along the lines of Example 4.28 one can show that α is a generator, and so

$$h_\mu(T) = h_\mu(T;\alpha) = \lim_{n \to +\infty} \frac{1}{n} H\left(\bigvee_{k=0}^{n-1} T^{-k}\alpha \right).$$

Now one can calculate directly that for each n, $H\left(\vee_{k=0}^{n-1} T^{-k}\alpha\right) = n \sum_i p_i \log p_i$. Alternatively we can use the definition of entropy in terms of conditional entropy

$$h_\mu(T;\alpha) = \lim_{n \to +\infty} H\big(\alpha \,\big|\, \vee_1^n T^{-k}\alpha\big).$$

The partitions α and $\vee_1^n T^{-k}\alpha$ are independent for each n, and so

$$H\big(\alpha \,\big|\, \vee_1^n T^{-k}\alpha\big) = H(\alpha) = \sum_i p_i \log p_i.$$

4.6 Ruelle's inequality

In preparation for proving inequality (75) we prove two lemmas.

Lemma 4.30. *Suppose $\alpha = \{A_1, \ldots, A_k\}$ and $\beta = \{B_1, \ldots, B_l\}$ are partitions of (X, Ω, μ). Then*

$$H(\beta \,|\, \alpha) \;\leq\; \sum_{i=1}^{k} \mu(A_i) \log r_\beta(A_i)$$

where $r_\beta(A_i)$ is the cardinality of the set $\big\{B \in \beta \,\big|\, B \cap A_i \neq \emptyset\big\}$.

Proof. From the definition of $H(\beta \,|\, \alpha)$ it suffices to show that for each $A_i \in \alpha$ having nonzero measure,

$$-\sum_j \mu(B_j \,|\, A_i) \log(\mu(B_j \,|\, A_i)) \;\leq\; \log r_\beta(A_i).$$

Observe that the left hand side is nothing other than the entropy of the partition γ of A_i given by all non-empty intersections with elements of β, when we view A_i as a probability space with measurable sets $E \cap A_i$ for all $E \in \Omega$, and measure $\tilde{\mu}(E \cap A_i) := \mu(E \cap A_i)/\mu(A_i)$. Then the lemma follows from

$$H(\gamma) \;\leq\; \log \operatorname{card}(\gamma) \;=\; \log r_\beta(A_i).$$

■

By a $(v_1, \ldots, v_d)$-**parallelogram** in $\mathbb{R}^d$ we mean the image of a Euclidean isometry of the set $R(v_1, \ldots, v_d)$ spanned by these vectors;

$$R(v_1, \ldots, v_d) := \Big\{ (s_1 v_1, \ldots, s_d v_d) \in \mathbb{R}^d \,\Big|\, 0 \leq s_i \leq 1, \text{ for all } i = 1 \Big\}.$$

Clearly the d-dimensional volume $\operatorname{vol}(R)$ is bounded by the product of the lengths $|v_1| \cdots |v_d|$. For any $r \geq 0$, consider the r-neighborhood R_r of $R = R(v_1, \ldots, v_d)$ in $\mathbb{R}^d$. Draw a picture and you'll see that

$$\operatorname{vol}(R_r) \;\leq\; \sum_{k=0}^{d} \sum_{\substack{\sigma \in R \\ \dim \sigma = k}} (2r)^{d-k} \operatorname{vol}(\sigma) \tag{90}$$

where on the right hand side σ runs over all the k-dimensional parallelograms consisting of k-tuples formed from the vectors $v_1, \ldots, v_d$, and $\mathrm{vol}(\sigma)$ means the k-dimensional volume.

For $\varepsilon > 0$ let $\mathcal{P}_\varepsilon = \mathcal{P}_\varepsilon(\mathbb{R}^d)$ be the **partition of $\mathbb{R}^d$ into ε-cubes**, that is $\mathcal{P}_\varepsilon$ consists of all sets of the form $\left\{ k_i\varepsilon \le x_i < (k_i+1)\varepsilon \,\middle|\, i = 1, \ldots, d \right\}$, over all d-tuples of integers $(k_1, \ldots, k_d)$. Note that each element of $\mathcal{P}_\varepsilon$ has Euclidean diameter $\varepsilon\sqrt{d}$.

Lemma 4.31. *Let $d \in \mathbb{N}$, and R be a $(v_1, \ldots, v_d)$-parallelogram in $\mathbb{R}^d$. Then for all $\varepsilon > 0$,*

$$\mathrm{card}\left\{ P \in P_\varepsilon \,\middle|\, P \cap R \neq \emptyset \right\} \;\le\; \sum_{k=0}^{d} \sum_{(i_1,\ldots,i_k)} (2\sqrt{d})^{d-k} \frac{|v_{i_1}|}{\varepsilon} \cdots \frac{|v_{i_k}|}{\varepsilon}$$

where the inner sum is over all k-tuples of non-repeated indices in $\{1, \ldots, d\}$.

Proof. The diameter of an ε-cube in $\mathbb{R}^d$ is $\varepsilon\sqrt{d}$. Thus if $P \in \mathcal{P}_\varepsilon$ and $P \cap R \neq \emptyset$ then P lies in $R' =$ the $\varepsilon\sqrt{d}$-neighborhood (in the Euclidean metric) of R. Hence

$$\begin{aligned} \mathrm{card}\left\{ P \in P_\varepsilon \,\middle|\, P \cap R \neq \emptyset \right\} \;&\le\; \mathrm{card}\left\{ P \in P_\varepsilon \,\middle|\, P \subset R' \neq \emptyset \right\} \\ &\le\; \frac{1}{\varepsilon^d} \cdot \text{Euclidean volume of } R'. \end{aligned}$$

Now we estimate the volume of R' using (90), with $r = \varepsilon\sqrt{d}$, and then bound the volume of each face σ of R by the product of the lengths of the vectors determining σ. ∎

Now we can prove the inequality. There is nothing to be gained by working in dimension 2, so the dimension $d \ge 1$ of M is assumed to be arbitrary.

Theorem 4.32 (Ruelle's inequality)**.** *Let $f : M \to M$ be a C^1-diffeomorphism*[*] *of a compact smooth manifold. Then for any f-invariant Borel probability*

[*] f need not be invertible and there is nothing to be gained by assuming this. However, to dispence with this assumption we need to use the non-invertible version of Oseledec.

measure μ there holds

$$h_\mu(f) \ \leq \ \int_M \lambda^+ \, d\mu$$

where $\lambda^+ : M \to [0, \infty)$ is the measurable function, defined μ-almost everywhere, taking $x \in M$ to the sum of the non-negative Lyapunov exponents (counted with multiplicity) of (f, μ) at x.

We will follow the original proof in ([16]) and the outline in ([18]). If α is a partition of M into small sets then for $A \in \alpha$ the set $f^j A$ has small measure but could still intersect a large number of elements in α. If however we can bound the "shape“ of $f^j A$, by bounding its diameters in different "directions“ then we can hope to bound the number of intersections with elements in α. The proof of the inequality is along the following lines: for a fine partition α of M and j large, the Oseledec theorem applies to most elements $A \in \alpha$ and tells us that $f^{-j} A$ is distorted in different directions according to the Lyapunov exponents. This allows to control the number of elements in α which intersect $f^{-j} A$ in terms of the exponents, leading to a bound on the entropy.

In the proof, let $\Lambda \subset M$ be the full measure set of Lyapunov regular points given by Oseledec, and let $\lambda_1(x) > \cdots > \lambda_k(x)$ be the Lyapunov exponents at $x \in \Lambda$, $(k \leq d)$, and let $E^k_x \ \subset \ \cdots \ \subset \ E^1_x \ = \ T_x M$ denote the filtration at x.

Proof. Every smooth manifold, compact or not, admits a smooth triangulation. We fix a finite triangulation for M and denote it by T and its cells by Δ. Let Y denote the "skeleton“ of T, that is,

$$Y := \bigcup_{\Delta \in T} \partial \Delta.$$

Without loss of generality $\mu(Y) = 0$ by perturbing Y. By definition each $\Delta \in T$ is the image of a smooth diffeomorphism from the standard d-simplex in $\mathbb{R}^d$:

$$\Delta^d := \Big\{ (x_1, \ldots, x_d) \Big| \, 0 \leq x_i \leq 1, \ \sum_i x_i \leq 1 \Big\}.$$

For each $\varepsilon > 0$ let a_ε denote the partition of the standard simplex Δ^d into ε-"boxes" (plus some triangles) obtained by intersecting elements in $\mathcal{P}_\varepsilon$ with

Δ^d. This determines a partition of each cell in the triangulation and therefore a partition of the whole manifold M, which we denote by α_ε. Clearly for any sequence $\varepsilon_1 > \varepsilon_2 > \ldots$ decreasing to zero, the partitions α_{ε_n} get finer as $n \to 0$ and the limit σ-algebra $\vee_n \alpha_{\varepsilon_n}$ coincides with the Borel σ-algebra on M.

Fix numbers $\delta_1, \delta_2, \delta_3, \delta_4 > 0$. Now:

1. Let $B \subset M$ be an open neighborhood of the skeleton Y, sufficiently small that $\mu(B) < \delta_1$.
2. Choose $N \in \mathbb{N}$ sufficiently large that the set G of "good points" x for which $Df^N(x)$ is δ_2-close to its Lyapunov representation, satisfies $\mu(G) > 1 - \delta_3$. Meaning that if $x \in G$ then

   $$\|Df^N(x)v\| \leq \|v\|e^{(\lambda_i+\delta_2)N}$$

   for all $v \in E_x^i$. We also assume that N is so large that every $\lambda_i > 0$ satisfies $e^{\lambda_i N} > 2\sqrt{d}$, and every $\lambda_j \leq 0$ satisfies $e^{\lambda_j N} < 2\sqrt{d}$.
3. Choose $\varepsilon > 0$ sufficiently small that all of the following hold:
   (a) $h(f;\alpha_\varepsilon) \geq h(f^N) - \delta_4$. (Can do by the Kolmogorov-Sinai theorem.)
   (b) For each $A \in \alpha_\varepsilon$, either $f^N(A) \subset B$ or there exists $\Delta \in T$ so that $f^N(A) \subset \Delta$. (To avoid dealing with $f^N(A)$ being spread over different cells in the triangulation.) Moreover, in the latter case, for each $x \in A \cap G$

   $$Df^N(x)\big(B_\varepsilon(x)\big) \ \subset \ R$$

   for some $R = R_x \subset \Delta$ corresponding to a parallelogram with side lengths $\varepsilon e^{(\lambda_1+\delta_2)N}, \ldots, \varepsilon e^{(\lambda_d+\delta_2)N}$. To show this we use that the pull back of the Riemannian metric g on each cell in the triangulation to the standard simplex in $\mathbb{R}^d$ is equivalent to the Euclidean metric.

Now in view of

$$h(f) \ = \ \frac{1}{N}h(f^N) \ \leq \ \frac{1}{N}h\big(f^N;\alpha_\varepsilon\big) + \frac{\delta_4}{N},$$

it suffices for us to show that

$$\lim_{N\to+\infty}\lim_{\varepsilon\to 0}\ \frac{1}{N}h\big(f^N;\alpha_\varepsilon\big)\ \le\ \int_M \lambda^+\,d\mu. \tag{91}$$

We have

$$\begin{aligned} h\big(f^N;\alpha_\varepsilon\big) &= \inf_{n\ge 1} H\left(\alpha_\varepsilon\,\Big|\,\bigvee_{j=1}^{n}(f^N)^{-j}\alpha_\varepsilon\right)\\ &\le H\big(\alpha_\varepsilon\big|f^{-N}\alpha_\varepsilon\big)\\ &\le \sum_{A\in\alpha_\varepsilon}\mu(A)\log r(A) \end{aligned} \tag{92}$$

by lemma 4.30, where

$$r(A) := \mathrm{card}\big\{\,B\in\alpha_\varepsilon\,\big|\,f^N A\cap B\neq\emptyset\,\big\}.$$

For "most" $A\in\alpha_\varepsilon$ we have $f^N(A)$ lies in a single simplex Δ and satisfies $A\cap G\neq\emptyset$. For such A, by (3)b) we can estimate $r(A)$ by the number of elements in a_ε which meet an $\varepsilon e^{(\lambda_1+\delta_2)N}\times\cdots\times\varepsilon e^{(\lambda_d+\delta_2)N}$-parallelogram. So by lemma 4.31,

$$r(A)\le\sum_{k=0}^{d}\sum_{(i_1,\dots,i_k)}(2\sqrt{d})^{d-k}e^{(\lambda_{i_1}+\delta_2)N}\cdots e^{(\lambda_{i_k}+\delta_2)N}.$$

This sum is bounded by some constant multiple C, depending only on d, of the maximal term. The maximal term clearly comes from the unique tuple that picks out those λ_i for which $e^{(\lambda_i+\delta_2)N}>2\sqrt{d}$. We arranged N is sufficiently large that this corresponds precisely to the strictly positive λ_i's. Thus we get an estimate of the form

$$\frac{1}{N}\log r(A)\ \le\ \sum_{\lambda_i>0}(\lambda_i+\delta_2)\ +\ \frac{C_0}{N} \tag{93}$$

for some large constant $C_0>0$ depending only on d.

The remaining $A\in\alpha_\varepsilon$ satisfy either $f^N(A)\subset B$ or $A\cap G=\emptyset$. For such sets we use a cruder estimate on $r(A)$. Fix an auxilliary Riemannian metric g on M. Let $\mathrm{Diam}_g(E)$ denote the diameter of $E\subset M$ with respect to g, and if

E lies in a single cell in the triangulation write $\mathrm{Diam}(E)$ for the Euclidean diameter of the corresponding subset of the standard simplex Δ^d. Then, $\mathrm{Diam}_g(f^N A) \leq \|Df\|^N_{C^0(M)} \mathrm{Diam}_g(A)$, and therefore there is some constant $C_1 > 0$ so that $\mathrm{Diam}\big(f^N(A) \cap T\big) \leq C_1^N \mathrm{Diam}(A)$ for each face $\Delta \in T$. Hence for all $A \in \alpha_\varepsilon$ we have

$$\frac{1}{N} \log r(A) \ \leq \ C_2 \tag{94}$$

for some constant $C_2 > 0$. Now we will use the better estimate (93) whenever A permits, and otherwise use (94), but the latter will only be for sets $A \in \alpha_\varepsilon$ taking up very little measure in M.

More precisely, combining (92), (93), and (94) we have

$$\begin{aligned} \frac{1}{N} h\big(f^N; \alpha_\varepsilon\big) \ &\leq \ \sum_{A \in \alpha_\varepsilon} \mu(A) \frac{1}{N} \log r(A) \\ &\leq \ \sum_{A \in \alpha_\varepsilon} \mu(A) \left[\sum_{\lambda_i \geq 0} (\lambda_i + \delta_2) \ + \ \frac{C_0}{N} \right] \\ &\qquad\qquad + \ \sum_{\substack{A \in \alpha_\varepsilon \\ A \subset f^{-N} B}} \mu(A) C_2 \ + \ \sum_{\substack{A \in \alpha_\varepsilon \\ A \subset M \setminus G}} \mu(A) C_2. \end{aligned}$$

Letting $\varepsilon \to 0$ gives

$$\begin{aligned} \lim_{\varepsilon \to 0} \frac{1}{N} h\big(f^N; \alpha_\varepsilon\big) \ &\leq \ \int_M \lambda^+ \, d\mu \ + \ d\delta_2 \ + \ \frac{C_0}{N} \\ &\qquad\qquad + \ C_2 \mu(B) + C_2 \mu(M \setminus G). \end{aligned}$$

So

$$\lim_{N \to \infty} \lim_{\varepsilon \to 0} \frac{1}{N} h\big(f^N; \alpha_\varepsilon\big) \ \leq \ \int_M \lambda^+ \, d\mu \ + \ d\delta_2 \ + \ C_2 \delta_1 + C_2 \delta_3.$$

This proves (91) and therefore theorem 4.32. ■

For comparison, let us state a couple of results without proof. First Pesin's formula [15], see also [11], or the discussion in [18]:

Theorem 4.33 (Pesin)**.** *Let $f : M \to M$ be a C^2-diffeomorphism of a compact smooth manifold, and μ an f-invariant Borel probability measure equivalent to the volume (meaning). Then*

$$h_\mu(f) \;=\; \int_M \lambda^+ \, d\mu.$$

So what is the bridge between these two statements? Of course if μ is ergodic then the right hand side is simply $\int_M \lambda^+ \, d\mu = \sum_{\lambda_i > 0} \lambda_i m_i$ where $m_i \in \mathbb{N}$ is the multiplicity of λ_i. In the case where $\lambda_i = \lambda > 0$ for all i, this becomes $\lambda \cdot m$ where $m = \dim M$. For a measure μ define its dimension $\dim(\lambda)$ to be the infimum of the Hausdorff dimensions of all supports of μ. Then the following is a theorem of Ledrappier and Young [10] from 1985.

Theorem 4.34 (Ledrappier-Young)**.** *Let $f : M \to M$ be a C^2-mapping of a compact smooth manifold, and μ an f-invariant ergodic Borel probability measure with all Lyapunov exponents $\lambda_i = \lambda$ for all i. Then*

$$h_\mu(f) \;=\; \lambda \cdot \dim(\mu).$$

It is not necessary that μ be f-ergodic, but makes it easier to state. See [18] for an outline of the proof.

References

[1] D. V. Anosov, A. B. Katok, New examples in smooth ergodic theory. Ergodic diffeomorphisms. Trans. Moscow Math. Soc. 23 (1970), 1–35

[2] A. Avila, J. Bochi, On the subadditive ergodic theorem. www.mat.uc.cl/∼jairo.bochi/

[3] J. Conway, A course in functional analysis, Graduate texts in mathematics, Springer (2010).

[4] J. Franks, Generalizations of the Poincaré-Birkhoff theorem. Ann. of Math. (2) 128 (1988), no. 1, 139–151.

[5] J. Franks, Geodesics on S^2 and periodic points of annulus homeomorphisms. Invent. Math. 108 (1992), no. 2, 403–418.

[6] M. Herman, Sur la conjugaison différentiable des difféomorphismes du cercle á des rotations. Inst. Hautes Études Sci. Publ. Math. No. 49 (1979), 5–233.

[7] A. Katok, Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. Inst. Hautes Études Sci. Publ. Math. No. 51 (1980), 137–173.

[8] A. Katok, Open problems in elliptic dynamics, www.math.psu.edu/katok_a/problems.html

[9] A. Katok, B. Hasselblatt, Introduction to the modern theory of dynamical systems. With a supplementary chapter by Katok and Leonardo Mendoza. Encyclopedia of Mathematics and its Applications, 54. Cambridge University Press, Cambridge, 1995. xviii+802 pp.

[10] F. Ledrappier, L.-S. Young, The metric entropy of diffeomorphisms, Annals of Math. 122 (1985) 509–574.

[11] R. Mañe, A proof of Pesin's formula, Ergod. Th. Dynam. Sys. 1 (1981) 95–102.

[12] R. Mañe, Ergodic theory and differentiable dynamics, Springer (1983).

[13] Parthasarathy, Introduction to probability and measure, Macmillan press, (1977).

[14] R. Phelps, Lectures on Choquet's theorem, Lecture notes in mathematics, Springer (2001).

[15] Ya. B. Pesin, Characteristic Lyapunov exponents and smooth ergodic theory, Russ. Math. Surveys 32 (1977) 55–114.

[16] D. Ruelle, An inequality for the entropy of differentiable maps, Bol. Soc. Bras. Mat. 9 (1978), 83–87.

[17] O. Sarig, Lecture notes on ergodic theory, available online, (Version Dec 2009).

[18] L-S. Young, Ergodic theory of differentiable dynamical systems, survey/slash lecture notes available at www.cims.nyu.edu/∼lsy/

[19] P. Walters, An introduction to ergodic theory, graduate texts in mathematics, Springer (1982)