Dynamical systems, Spring 2020

Péter Bálint

May 19, 2020

Abstract

On March 12 2020, the Technical University of Budapest has decided to continue teaching in the form of distance education, effective March 23. These lecture notes contain the material for the Dynamical systems course (BMETE93MM02).

Contents

1 Introduction to ergodic theory

1.1 Recap of the definition of ergodicity.

Ergodicity concerns the *relation of a dynamical system* $T : X \to X$ and *an invariant measure* μ. that is, we are given an endomorphism, that is a quadruple $(X, \mathcal{B}, T, \mu)$, where X is the phase space, $\mathcal{B}$ is a sigma algebra, the map $T : X \to X$ is measurable with respect to $\mathcal{B}$, and μ is an invariant probability measure, that is, $\mu(A) = \mu(T^{-1}A)$ for every $A \in \mathcal{B}$. As usual, L^1_μ denotes the set of (with respect to μ) integrable functions $f : X \to \mathbb{R}$.

Invariant sets and functions. A set $A \in \mathcal{B}$ is invariant if $A = T^{-1}A$ μ-almost surely, that is, $\mu(A \Delta T^{-1}A) = 0$. A function $f \in L^1_\mu$ is invariant if $f = f \circ T$ almost surely, that is, $f(x) = f(Tx)$ for μ almost every $x \in X$. Note that if $\mu(A) = 0$ or $\mu(A) = 1$, then it is automatically invariant. Also, if f is (almost surely) constant (i.e. there exists $c \in \mathbb{R}$ such that $f(x) = c$ for almost every $x \in X$), then it is automatically invariant. We will call these trivial invariant sets/functions.

Definition 1.1 (Ergodicity)**.** *The endomorphism $(X, \mathcal{B}, T, \mu)$ is ergodic (or, the invariant measure μ is ergodic with respect to the map $T : X \to X$) if it has only trivial invariant sets (i.e. if A is invariant then $\mu(A) = 0$ or $\mu(A) = 1$). Equivalently, $(X, \mathcal{B}, T, \mu)$ is ergodic if it has only trivial invariant functions.*

We have seen (at the class on March 10) that the two definitions of ergodicity are equivalent. Also, we have discussed that ergodicity means that the phase space cannot be split into two (from the viewpoint of the measure) non-trivial invariant components.

1.2 Birkhoff's ergodic theorem

Consider an endomorphism $(X, \mathcal{B}, T, \mu)$ and an arbitrary integrable function $f \in L^1_\mu$. We introduce the following notations. For $x \in X$ and $n \geq 1$, let

$$S_n f(x) = f(x) + f(Tx) + ... + f(T^{n-1}x), \qquad A_n f(x) = \frac{S_n f(x)}{n}.$$

Here S stands for sum, and A stands for average.

Theorem 1.2 (Birkhoff's ergodic theorem for ergodic endomorphisms)**.** *Let $(X, \mathcal{B}, T, \mu)$ be ergodic. Then, for any $f \in L^1_\mu$:*

$$A_n f(x) \to \int f d\mu, \qquad \text{as n} \to \infty, \text{ for } \mu\text{– a. e. x} \in \text{X}. \tag{1.1}$$

The proof of this theorem is postponed to a later section. For now, we make the following comments.

- There is also a version for nonergodic endomorphisms, in which case it states that the sequence converges for μ-almost every $x \in X$, but not (necessarily) to a constant value.

- *Probabilistic interpretation.* $(X, \mathcal{B}, \mu)$ is a probability space, and $f \in L^1_\mu$ can be regarded as *a random variable*, the expected value of which is $\int f d\mu$. $f, f \circ T, f \circ T^2, ...$ is a sequence of random variables, and, by the invariance of the measure, this is an identically distributed sequence. Then Theorem 1.2 is an analogue of the *Strong Law of Large Numbers.* Note, however, that this sequence is by far not independent. In fact, the only source of randomness is the choice of the initial point $x \in X$, which then determines deterministically all later points $T^k x$, $(k \geq 1)$.

- *Physics interpretation.* If X is the phase space of a physical system, which is evolved by the map T, then functions $f : X \to \mathbb{R}$ can be regarded as an observable, and the values $f(x), f(Tx), ...$ are the consecutive measurements. In this terminology Theorem 1.2 is often formulated as "time averages converge to the ensemble average" for almost every initial condition $x \in X$.

1.3 Linear operators associated to a dynamical system

Given a map $T : X \to X$, and an appropriately chosen linear space $\mathcal{L}$ of functions $f : X \to \mathbb{R}$, we may consider the operator $\widehat{T} : \mathcal{L} \to \mathcal{L}$ defined by

$$\widehat{T}f = f \circ T \qquad \text{that is } (\widehat{\mathrm{T}}\mathrm{f})(\mathrm{x}) = \mathrm{f}(\mathrm{Tx}) \text{ for } \mathrm{x} \in \mathrm{X}.$$

The properties of this operator depend on the choice of $\mathcal{L}$, yet, as long as $\mathcal{L}$ is preserved, it is apparently a linear operator: $\widehat{T}(\lambda_1 f_1 + \lambda_2 f_2) = \lambda_1 \widehat{T}(f_1) + \lambda_2 \widehat{T}(f_2)$ for every $f_1, f_2 \in \mathcal{L}$ and $\lambda_1, \lambda_2 \in \mathbb{R}$.

1.3.1 Action of $\widehat{T}$ on L^p_μ.

Let $(X, \mathcal{B}, T, \mu)$ be an endomorphism, that is, the probability measure μ is preserved. In particular, $(X, \mathcal{B}, \mu)$ is a probability space, and for any $p \in [1, \infty]$ we may consider the space

$$L^p_\mu = (L^p_\mu(X) =)\{f : X \to \mathbb{R} \mid \int_X |f(x)|^p d\mu(x) < \infty\},$$

with the norm

$$\|f\|_p = (\int |f|^p d\mu)^{1/p}.$$

With this norm, these spaces are Banach spaces and for the particular case of $p = 2$, L^2_μ is a Hilbert space with the inner product:

$$< f, g >_2 = \int f(x)g(x)d\mu(x).$$

Let us emphasized that *real* L^p_μ spaces are considered.

As the measure μ is invariant, $\widehat{T}$ not only preserves the space L^p_μ, it is, actually, an L^p_μ*-isometry*:

$$\|\hat{T}f\|_p = \|f\|_p \quad \forall f, g \in L^p_\mu.$$

Of special interest to us is the action of $\widehat{T}$ on the Hilbert space L^2_μ. A Lemma is included here for future reference. Recall that the *adjoint* of $\widehat{T}$ is the operator $\widehat{T}^* : L^2_\mu \to L^2_\mu$ such that for any $f, g \in L^2_\mu$ we have $< f, \widehat{T}^* g >_2 = < \widehat{T} f, g >_2$. This is not to be confused with the adjoint of $\widehat{T}$ acting on the space of continuous functions, to be discussed in the next subsection.

Lemma 1.3. *If $f \in L^2_\mu$ is such that $\widehat{T}^* f = f$, then $\widehat{T} f = f$.*

Comment: This is true for any isometry of a *real* Hilbert space.

Proof. Assume $\widehat{T}^* f = f$. Then $(\widehat{T} f, f) = (f, \widehat{T} f)_2 = (\widehat{T}^* f, f)_2 = (f, f)_2 = \|f\|_2$. Also $\|\widehat{T} f\|_2^2 = \|f\|^2$ as $\widehat{T}$ is an isometry. Hence

$$\|f - \widehat{T} f\|_2^2 = \|f\|_2^2 - (f, \widehat{T} f)_2 - (\widehat{T} f, f)_2 + \|\widehat{T} f\|_2^2 = 0,$$

which implies $f = \widehat{T} f$. □

1.3.2 Action of $\widehat{T}$ on the space of continuous functions.

In the remainder of the section we consider maps $T : X \to X$ without any particular invariant measure fixed. Actually, as it will be discussed later, we are interested in questions like: does T have any invariant (probability) measure at all? If yes, how can we describe the invariant measures that it has?

Setting: Let X be a *compact metric space* and let $T : X \to X$ be *continuous*. Then, unless specified differently, we make the natural choice that $\mathcal{B}$ is the Borel σ-algebra. Terminology: this setting is often called a *topological dynamical system*.

First we discuss some spaces associated to the compact metric space X. Let $\mathcal{C}(= \mathcal{C}(X))$ denote the linear space of *continuous functions* $f : X \to \mathbb{R}$. This is a Banach space with the supremum norm

$$\|f\| = \sup_{x \in X} |f(x)| < \infty \qquad f \in \mathcal{C}.$$

As usual, $\mathcal{C}^*(X)$ denotes the *dual space*, that is, the space of bounded linear functionals acting on $\mathcal{C}(X)$. That is, if $\alpha \in \mathcal{C}^*$, then $\alpha : \mathcal{C} \to \mathbb{R}$, and the dependence of $\alpha(f)$ on f is linear and continuous.

Finally, let $\mathcal{M}(= \mathcal{M}(X))$ denote the collection of *Borel probability measures* on X. For $\mu \in \mathcal{M}$ and $f \in \mathcal{C}$, let $\mu(f) = \int f d\mu$. This way μ acts on $\mathcal{C}$, and this action has the following properties.

- the action is apparently linear,
- the action is bounded:

$$|\mu(f)| = \left| \int_X f(x) d\mu(x) \right| \leq \sup_{x \in X} |f(x)| \cdot \mu(X) = \|f\|.$$

- the above computation not only shows that μ can be regarded as an element of $\mathcal{C}^*$, it is actually of *unit norm*, and as such it is an element of the unit ball of $\mathcal{C}^*$. (To see that the norm is not just less but actually equal to 1, note that for the special case of $f_0 \equiv 1$ we have $\mu(f_0) = 1 = \|f_0\|$.)
- μ is a *positive* functional in the sense that if $f(x) \geq 0$ for every $x \in X$, then $\mu(f) \geq 0$.

By the *Riesz representation theorem* (more precisely, by one of the several Riesz representation theorems) any element $\alpha \in \mathcal{C}^*$ which is positive and has unit norm, can be identified with an element $\mu \in \mathcal{M}$.[1]

Note that this way $\mathcal{M}$ is a closed subset of the unit ball in $\mathcal{C}^*$.

The *weak-$*$ topology* on $\mathcal{C}^*$ (and thus, on $\mathcal{M}$) is defined as follows. Given a sequence $\mu_n \in \mathcal{M}$, we have $\mu_n \xrightarrow{*} \mu \in \mathcal{M}$ in the weak-$*$ sense if for any continuous function $f \in \mathcal{C}$ we have that $\mu_n(f)$ (as a numerical sequence) converges to $\mu(f)$. By the *Banach-Alaoglu theorem* the (closed) unit ball is compact in the weak-$*$ topology.[2] Hence, in particular, $\mathcal{M}$, which is a closed subset of the unit ball in $\mathcal{C}^*$, is compact in the weak-$*$ topology, too.

Now we get back to dynamical systems. Given a continuous map $T : X \to X$, for any continuous $f : X \to \mathbb{R}$ we have that $f \circ T$ is continuous, too. Hence we may consider the linear operator

$$\widehat{T} : \mathcal{C} \to \mathcal{C}, \qquad (\widehat{T}f)(x) = f(Tx).$$

Note that for the supremum norm $\|\widehat{T}f\| \leq \|f\|$ (actually, $\|\widehat{T}f\| = \|f\|$ if $T : X \to X$ is onto), hence $\widehat{T} : \mathcal{C} \to \mathcal{C}$ is a bounded linear operator.

Let us denote by $\widehat{T}_*$ the adjoint of $\widehat{T}$, which acts on $\mathcal{C}$, and thus, can be restricted to $\mathcal{M}$:

$$\widehat{T}_* : \mathcal{M} \to \mathcal{M}, \qquad (\widehat{T}_*\mu)(f) = \mu(\widehat{T}f) = \mu(f \circ T).$$

Now $\widehat{T}_*$ defined this way is an operator we have already considered, the push-forward operator on measures. [3]

This implies in particular that *a probability measure $\mu \in \mathcal{M}$ is invariant for $T : X \to X$ if and only if it is a fixed point of the push-forward operator, that is, if $T_*\mu = \mu$.* We introduce the following notation for the collection of invariant measures for $T : X \to X$:

$$\mathcal{M}_{\text{inv}} = (\mathcal{M}_{\text{inv}}(T) =)\{\mu \in \mathcal{M} \,|\, T_*\mu = \mu\}.$$

Finally, we show that the operator T_* is continuous in the weak-$*$ topology. Let the sequence $m_n \in \mathcal{M}$ converge to $m \in \mathcal{M}$, that is, $m_n \xrightarrow{*} m$. This means that $m_n(f) \to m(f)$ for every $f \in \mathcal{C}$. In particular as $f \circ T$ is continuous, we have that $m_n(f \circ T) \to m(f \circ T)$ form every $f \in \mathcal{C}$, which actually means $T_*(m_n) \xrightarrow{*} T_*m$ in the weak-$*$ sense.

[1] This particular theorem appears on wikipedia as the Riesz-Markov-Kakutani representation theorem, and as the Riesz representation theorem on wofram mathworld.

[2] Note that, for nontrivial examples of X, $\mathcal{C}$ (and hence $\mathcal{C}^*$) is infinite dimensional and thus the (closed) unit ball is *not compact in the norm topology.*

[3] This can be seen by approximating the indicator function χ_A of a $A \in \mathcal{B}$ by a sequence of continuous functions, and noting that $\chi_A \circ T = \chi_{T^{-1}A}$.

1.4 Krylov-Bogolyubov theorem

Theorem 1.4 (Krylov-Bogolyubov theorem)**.** *Consider a topological dynamical system, that is $T : X \to X$, with X a compact metric space and T continuous. Then there exists at least one invariant probability measure, that is, $\mathcal{M}_{\text{inv}}(T) \neq \emptyset$.*

Proof. Fix an arbitrary $x \in X$. (The invariant measure to be constructed may depend on the choice of x.) Let

$$\mu_n = \frac{\delta_x + \delta_{Tx} + \dots + \delta_{T^{n-1}x}}{n}$$

where δ_y denotes the Dirac delta measure at the point $y \in X$.[4] As $\mu_n \in \mathcal{M}$ which is compact in the weak star topology, there exists a subsequence μ_{n_j} and a limit measure $\mu \in \mathcal{M}$ such that $\mu_{n_j} \xrightarrow{*} \mu$ as $j \to \infty$. . It is claimed that this measure is invariant for $T : X \to X$. To verify this claim, it has to be shown that $T_*\mu = \mu$.

By the weak-$*$ continuity of T_*, we have $T_*\mu_{n_j} \xrightarrow{*} T_*\mu$. On the other hand, for any $f \in \mathcal{C}$:

$$\begin{aligned} |T_*(\mu_{n_j})(f) - \mu_{n_j}(f)| &= \left| \frac{f(x) + f(Tx) \dots + f(T^{n_j}x)}{n_j} - \frac{f(Tx) + f(T^2x) \dots + f(T^{n_j+1}x)}{n_j} \right| = \\ &= \left| \frac{f(x) + f(T^{n_j+1}x)}{n_j} \right| \leq \frac{2\|f\|}{n_j} \to 0, \end{aligned}$$

as $j \to \infty$, hence the weak-$*$ limit points of the sequences μ_{n_j} and $T_*\mu_{n_j}$ coincide, which means $\mu = T_*\mu$. $\square$

Alternative argument for the proof of the Krylov-Bolgolyubov theorem, which is of some interest for what follows. We have seen that $\mathcal{M}$ is a compact set in the weak-$*$ topology. Another important property is that $\mathcal{M}$ is *convex*: for any $m_1, m_2 \in \mathcal{M}$ and $\lambda \in [0, 1]$ we have that $\lambda m_1 + (1 - \lambda) m_2 \in \mathcal{M}$. Also, we have seen that $T_* : \mathcal{M} \to \mathcal{M}$ is continuous. Now, by the Schauder fixed point theorem given any continuous mapping of a compact, convex set in a Hausdorff topological vector space into itself has a (not necessarily unique) fixed point. If $\mu \in \mathcal{M}$ is a fixed point for T_*, then $\mu = T_*\mu$, that is, $\mu \in \mathcal{M}_{\text{inv}}$.

Example 1.5. *Consider $T : [0, 1] \to [0, 1]$, $Tx = x/2$. 0 is a fixed point of T, hence δ_0 is an invariant measure. We claim that this is the only invariant measure for T, that is, $\mathcal{M} = \{\delta_0\}$.*

This is related to the fact that the fixed point 0 attracts the whole phase space. For a formal argument, note that $T[0, 1] = [0, 1/2]$, and, moreover $T[0, 2^{-k}] = [0, 2^{-k-1}]$. Hence, for any $\mu \in \mathcal{M}_{\text{inv}}$:

$$1 = \mu([0, 1]) = \mu(T^{-1}[0, 1/2]) = \mu([0, 1/2]) = \dots = \mu([0, 2^{-n}]) = \dots$$

[4] If the orbit of x was periodic with period p, then μ_p would be an invariant measure. However, this is not the general case. It may happen that $T : X \to X$ does not have any periodic points.

then by the continuity of the measure (which is equivalent to σ-additivity):

$$1 = \lim_{n\to\infty} \mu([0, 2^{-n}]) = \mu\left(\bigcap_{n=1}^{\infty} [0, 2^{-n}]\right) = \mu(\{0\}).$$

Thus the full mass is carried by the point 0, which means $\mu = \delta_0$.

Further comments:

- As mentioned earlier, topological dynamical systems for which $\mathcal{M}_{\text{inv}}$ consist of a single measure are called *uniquely ergodic* (some explanation for this terminology is given in the next section). Another uniquely ergodic example is the rotation of the circle by an irrational angle. In this case the only invariant measure is Lebesgue measure.

- For "chaotic" examples the set $\mathcal{M}_{\text{inv}}$ is typically very large. For example, consider the full shift $\sigma : \Sigma^+ \to \Sigma^+$ (or the two-sided –invertible – $\sigma : \Sigma \to \Sigma$). We have seen that there are many periodic points, which all give rise to invariant measures. another class of invariant measures are Bernoulli measures. There are also further examples, for example Markov measures, to be discussed later. By (almost)-conjugacy, this rich structure of $\mathcal{M}_{\text{inv}}$ is characteristic to several other examples: the doubling map, the logistic map with parameter greater than 4, or Smale's horseshoe.

1.5 Convexity and ergodicity

Recall that $\mathcal{M}$ is convex and compact in the weak-$*$ topology. By the continuity oh T_*, $\mathcal{M}_{\text{inv}}$ is a closed subset, hence it is compact. It can be checked by direct inspection that convex: if $\mu_1, \mu_2 \in \mathcal{M}_{\text{inv}}$ and $\lambda \in [0, 1]$, then $\lambda\mu_1 + (1-\lambda)\mu_2 \in \mathcal{M}_{\text{inv}}$.

Some terminology from convex geometry: if K is a convex set, then a point $x \in K$ is *extreme* if $x = \lambda y_1 + (1-\lambda)y_2$ can hold for some $y_1, y_2 \in K$ and $\lambda \in [0, 1]$ only in the trivial cases of $\lambda = 1$ (when $y_1 = x$) or $\lambda = 0$ (when $y_2 = x$). Examples: the extreme points of a segment are its two endpoints; the extreme points of a triangle are its three vertices.

Given $T : X \to X$, let us denote by $\mathcal{M}_{\text{erg}}(= \mathcal{M}_{\text{erg}}(T)) \subset \mathcal{M}_{\text{inv}}(= \mathcal{M}_{\text{inv}}(T))$ the collection of those invariant measures that are *ergodic* for $T : X \to X$. We have the following Lemma.

Lemma 1.6. *For $\mu \in \mathcal{M}_{\text{inv}}$ the following properties are equivalent:*

(i) $\mu \in \mathcal{M}_{\text{erg}}$.

(ii) μ *is an extreme point of* $\mathcal{M}_{\text{inv}}$.

Proof. (ii) $\Rightarrow$ (i) will be a homework. You can start thinking about it.

To prove (i) $\Rightarrow$ (ii), let us fix some $\mu \in \mathcal{M}_{\text{erg}}$ and assume that $\mu = t\mu_1 + (1-t)\mu_2$ for $\mu_1, \mu_2 \in \mathcal{M}_{\text{inv}}$, $t \in [0, 1]$. Then if for some measurable set we have $\mu(A) = 0$, it has to hold that $\mu_1(A) = 0$. This means that $\mu_1 \ll \mu$ (μ_1 is absolutely continuous w.r. to μ). Hence, by the Radon-Nikodym theorem, there exist some density function $\rho \in L^1_\mu$, $\rho = \frac{d\mu_1}{d\mu}$. In fact, we can say more: $\rho(x) \leq 1/t$ for μ-almost every $x \in X$. To see this, assume, by contradiction, that there exists some measurable set $B \subset X$ with $\mu(B) > 0$ such that $\rho(x) > 1/t$ for $x \in B$. Then

$$\mu(B) = t\mu_1(B) + (1-t)\mu_2(B) \geq t\mu_1(B) = t\int_B \rho(x)d\mu(x) > t \cdot (1/t)\mu(B) = \mu(B),$$

which is a contradiction. Hence $\rho(x)$ is in L^∞_μ, and thus in L^2_μ. Now, for any $f \in L^2_\mu$:

$$(f,\rho)_2 = \int f(x)\rho(x)d\mu(x) = \int f d\mu_1 = \in (f \circ T)d\mu_1 = \int (f\circ T)(x)\rho(x)d\mu(x) = (\widehat{T}f,\rho)_2,$$

where we have used that $\mu_1 \in \mathcal{M}_{\mathrm{inv}}$. But this means that $\widehat{T}^*\rho = \rho$, which, by Lemma 1.3, implies that $\widehat{T}\rho = \rho$. Hence ρ is an invariant function, which, as μ is ergodic, implies that ρ is almost surely constant. Now as ρ is a probability density, we have that $\rho(x) = 1$ for μ-almost every $x \in X$. but this implies that $\mu_1 = \mu$, so we have $t = 1$, and μ is an extreme point. □

As a byproduct of this argument, the following statement has been proved:

Lemma 1.7. *If $\mu \in \mathcal{M}_{\mathrm{erg}}$ and $\mu_1 \in \mathcal{M}_{\mathrm{inv}}$ such that $\mu_1 \ll \mu$, we have $\mu = \mu_1$.*

Comment. By the Krein-Milman theorem, any compact and convex set $K \subset X$ in a locally convex topological vector space arises as the convex hull of its extreme points. (Check wikipedia.) This implies in particular *ergodic decomposition*: any invariant measure can be decomposed as a (potentially, uncountable) convex combination of ergodic measures.

One more Lemma is included to give further insight into the structure of $\mathcal{M}_{\mathrm{inv}}$ and $\mathcal{M}_{\mathrm{erg}}$. Recall that two measures $m_1, m_2 \in \mathcal{M}$ are mutually singular (denoted $m_1 \perp m_2$)if there exists some measurable set A such that $m_1(A) = 1$ and $m_2(A) = 0$.

Lemma 1.8. *If $m, \mu \in \mathcal{M}_{\mathrm{erg}}$, $m \neq \mu$, then $m \perp \mu$.*

Proof. Consider $\mu, m \in \mathcal{M}_{\mathrm{erg}}$. By the Lebesgue decomposition theorem, there exists $t \in [0,1]$ and $\mu_1, \mu_2 \in \mathcal{M}$ such that $\mu_1 \ll m$, $\mu_2 \perp m$

$$\mu = t\mu_1 + (1-t)\mu_2. \tag{1.2}$$

We *claim* that $\mu_1, \mu_2 \in \mathcal{M}_{\mathrm{erg}}$. This implies, by Lemma 1.6, that either $t = 1$ (and $m = \mu$) or $t = 0$ (and $\mu \perp m$). Hence, in the rest of the argument, we verify the claim, which completes the proof of the Lemma.

On the one hand, $\mu \in \mathcal{M}_{\mathrm{inv}}$, hence $\mu = T_*\mu$. On the other hand, apply T_* to (1.2):

$$\mu = T_*\mu = tT_*\mu_1 + (1-t)T_*\mu_2 \tag{1.3}$$

if we prove (a) $T_*\mu_1 \ll m$ and (b) $T_*\mu_2 \perp m$ then, by the uniqueness of Lebesgue decomposition, $\mu_1 = T_*\mu_1$ and $\mu_2 = T_*\mu_2$. Hence (a), (b) would complete the proof of the claim and thus, of the Lemma.

To prove (a), consider A with $m(A) = 0$. Then $m(T^{-1}A) = 0$ as $m \in \mathcal{M}_{\text{inv}}$. By (1.2), $\mu_1(T^{-1}A) = 0$. So $T_*\mu_1(A) = 0$, and thus $T_*\mu_1 \ll m$.

To prove (b), by $\mu_2 \perp m$, there exists B such the $\mu_2(B) = 1$ and $m(B) = 0$. Then by $m \in \mathcal{M}_{\text{inv}}$ we have $T_*m(B) = 0$. Also, by (a) $T_*\mu_1(B) = 0$. By (1.2) and (1.3),

$$(1-t) = \mu(B) = T_*\mu(B) = (1-t)T_*\mu_2(B)$$

Hence $T_*\mu_2(B) = 1$, which completes the proof of (b). $\square$

1.6 Proof of the Birkhoff ergodic theorem

In this subsection we give the proof of Theorem 1.2. This is a proof by Katznelson and Weiss, not the original argument of Birkhoff.

Preliminary observations:

- Recall that for any $n \geq 1$ we have $S_nf(x) = f(x) + f(Tx) + \cdots + f(T^{n-1}x)$. By the invariance of the measure μ:

$$\int S_nfd\mu = \int fd\mu + \int f \circ Td\mu + \cdots + \int f \circ T^{n-1}d\mu = n \cdot \int fd\mu. \tag{1.4}$$

- We may assume that $f(x) \geq 0$ for μ-a.e. $x \in X$. Otherwise, split f into a positive and negative part, and use linearity.

- Introduce

$$f^+(x) = \limsup_{n\to\infty} \frac{S_nf(x)}{n}, \qquad f^-(x) = \liminf_{n\to\infty} \frac{S_nf(x)}{n}.$$

  We claim that these are invariant functions. To see this:

$$\frac{S_{n+1}f(x)}{n+1} = \frac{n}{n+1}f(x) + \left(\frac{n}{n+1}\right) \cdot \frac{S_nf(Tx)}{n}$$

  Taking lim sup, the left had side tends to $f^+(x)$, the first term on the right hand side tends to 0, $\frac{n}{n+1} \to 1$, while $\limsup_{n\to\infty} \frac{S_nf(Tx)}{n} = f^+(Tx)$. This gives $f^+(x) = f^+(Tx)$, and similarly for $f^-(x)$.

- As we are in the ergodic case, these invariant functions are μ-almost surely equal to some constants, which we may denote by f^+ and f^-. Then our aim is to show $f^+ \leq \int fd\mu$ and $\int fd\mu \leq f^-$, which by the obvious $f^- \leq f^+$ implies that the three quantities are equal, and the limit μ-almost surely exists and is equal to $\int fd\mu$.

We focus on $f^+ \leq \int fd\mu$. It is enough to show that for an arbitrary small $\varepsilon > 0$ we have $\int fd\mu \geq f^+ - 3\varepsilon$. So let us fix $\varepsilon > 0$. Let

$$n(x) := \min\{n \geq 1 \,|\, \frac{S_nf(x)}{n} \geq f^+ - \varepsilon\}.$$

We have that

$$S_{n(x)}f(x) \geq n(x)(f^+ - \varepsilon). \tag{1.5}$$

Note that $n(x)$ is finite for almost every $x \in X$, however, it may not be uniformly bounded. *Let us complete first the argument assuming that* $n(x) \leq N$, almost surely, for some $N \in \mathbb{N}$. Then we will see how to fix the argument when this does not hold. Introduce the following stopping times:

$$\begin{aligned} n_1(x) &= n(x) \\ n_2(x) &= n_1(x) + n(T^{n_1(x)}x), \quad \text{etc.} \\ n_k(x) &= n_{k-1}(x) + n(T^{n_{k-1}(x)}x). \end{aligned}$$

Also define

$$B_j(x) = S_{n_j(x)-n_{j-1}(x)}f(T^{n_{j-1}(x)}x), \qquad j = 1, 2, ... \ (n_0(x) = 0).$$

For brevity, let us introduce $x_j = T^{n_j(x)}x$. Note that by (1.5):

$$B_j(x) = S_{n(x_{j-1})}f(x_{j-1}) \geq n(x_{j-1})(f^+ - \varepsilon) = (n_j(x) - n_{j-1}(x))(f^+ - \varepsilon). \tag{1.6}$$

Also, by our assumption, B_j consists almost surely of at most N terms. Let, furthermore, L be so large that $\frac{Nf^+}{L} < \varepsilon$. Introduce one more stopping time:

$$k(x) = \sup\{k \geq 1 | n_k(x) \leq L - 1\}.$$

Then

$$L - N \leq n_{k(x)}(x) \leq L. \tag{1.7}$$

Now

$$\begin{aligned} S_L f(x) &= S_{n_1(x)}f(x) + S_{n_2(x)-n_1(x)}f(T^{n_1(x)}x) + ... = \\ &= B_1(x) + B_2(x) + ... + B_{k(x)}(x) + \text{remainder} \geq \\ &\geq (n_1(x) + n_2(x) - n_1(x) + ... + n_{k(x)}(x) - n_{k(x)-1}(x))(f^+ - \varepsilon) \geq \\ &\geq = n_{k(x)}(x)(f^+ - \varepsilon) \geq (L - N)(f^+ - \varepsilon) \end{aligned}$$

where, we have used that, μ almost surely (i) by $f(x) \geq 0$, the remainder is nonnegative, (ii) (1.6), (iii) (1.7). Now, integration, (1.4) and then division by L gives:

$$L \int f d\mu \geq (L - N)(f^+ - \varepsilon) \quad \Longrightarrow \quad \int f d\mu \geq f^+ - \frac{Nf^+}{L} - \varepsilon \geq f^+ - 2\varepsilon.$$

Now *let us see how to fix the argument when* $n(x)$ *is not uniformly bounded.* Even if $n8x)$ is not bounded, it is almost surely finite, hence, for N sufficiently large, the measure of set

$$A = \{x \in X | n(x) > N\}$$

can be made arbitrarily small. In particular choose N so large that $\mu(A) \leq \frac{\varepsilon}{f^+}$. Now define

$$\hat{f}(x) = \begin{cases} f(x) & \text{if } x \notin A, \\ \max(f(x), f^+) & \text{if } x \in A; \end{cases} \qquad \hat{n}(x) = \begin{cases} n(x) & \text{if } x \notin A, \\ 1 & \text{if } x \in A. \end{cases}$$

Then

- $$\int \hat{f} d\mu \leq \int_{X \setminus A} f d\mu + \int_A (f + f^+) d\mu \leq \int f d\mu + f^+ \cdot \mu(A) \leq \int f d\mu + \varepsilon$$
- $\hat{n}(x) \leq N$,
- Both for $x \in A$ and for $x \in X \setminus A$:
$$S_{\hat{n}(x)} \hat{f}(x) \geq \hat{n}(x)(f^+ - \varepsilon)$$
almost surely.

Hence the previous argument (introduce stopping times etc.) applies to $\hat{f}(x)$ (with $\hat{n}(x)$ instead of $n(x)$). We arrive at:

$$f^+ \leq \int \hat{f} d\mu + 2\varepsilon \leq \int f d\mu + 3\varepsilon.$$

Let us see how to adapt this argument to prove $f^- \geq \int f d\mu - 4\varepsilon$ for all $\varepsilon > 0$:

- truncation: choose $C > 0$ large enough such that, for $f_C(x) = \min(f(x), C)$ we have $\int f d\mu \leq \int f_C d\mu - \varepsilon$.
- let $n(x) = \min\{n \geq 1 \,|\, \frac{S_n(f_C)(x)}{n} \leq f^- + \varepsilon$.
- assuming temporarily that there exists some N for which $n(x) \leq N$ almost surely, proceed as in the case of f^+: construct stopping times etc. In particular, if L is large enough to ensure $\frac{NC}{L} < \varepsilon$, then
$$S_L(f_C)(x) \leq L(f^- + \varepsilon) + NC \quad \Longrightarrow \quad \int f_C d\mu \leq f^+ + 2\varepsilon.$$
- to account for unbounded $n(x)$, let N be large enough that for $A = \{x \in X \,|\, n(x) > N\}$, $\mu(A) \leq \frac{\varepsilon}{C}$. Then let
$$\hat{f}(x) = \begin{cases} f_C(x) & \text{if } x \notin A, \\ 0 & \text{if } x \in A; \end{cases} \qquad \hat{n}(x) = \begin{cases} n(x) & \text{if } x \notin A, \\ 1 & \text{if } x \in A. \end{cases}$$
The argument with the stopping times applies to $\hat{f}$ and $\hat{n}$. So
$$f^- \geq \int \hat{f} d\mu - 2\varepsilon$$
- Finally, use that
$$\int \hat{f} d\mu \geq \int f_c d\mu - \varepsilon \geq \int f d\mu - 2\varepsilon.$$

1.7 von Neumann's ergodic theorem and further characterizations of ergodicity

Theorem 1.9 (von Neumann's ergodic theorem)**.** *Let $(X, \mathcal{B}, T, \mu)$ be ergodic, and let $f : X \to \mathbb{R}$ be square integrable, i.e. $f \in L^2_\mu$. Then $A_n f \to \int f d\mu$ in L^2_μ, that is*

$$\int_X \left| \frac{S_n f(x)}{n} - \int f d\mu \right|^2 d\mu(x) \longrightarrow 0 \quad \text{as} \quad \mathrm{n} \to \infty. \tag{1.8}$$

Generally, convergence almost surely and in L^2 are independent, none of them implies the other.

Question 1.10. *Can you construct a sequence of random variables – say on $[0, 1]$ with the Lebesgue measure – that (i) converges in L^2, but does not converge almost surely; (ii) that converges almost surely, but does not converge in L^2?*

Proof. von Neumann's ergodic theorem can be proved directly, actually, with a very nice argument, but we do not include it here. Instead, we show that Birkhoff's ergodic theorem implies von Neumann's ergodic theorem.

Step 1. Let us consider first $f \in L^\infty_\mu$, that is, assume that there exists $C > 0$ such that $|f(x)| \leq C$ for μ-a.e. $x \in X$. Then

$$|A_n f(x)| = \left| \frac{f(x) + f(Tx) + \dots f(T^{n-1}x)}{n} \right| \leq C$$

almost surely and $|\int f d\mu| \leq C$. Hence the sequence of functions $|A_n f(x) - \int f d\mu|^2$ is uniformly bounded, and, by Birkhoff's ergodic theorem converges to 0, μ-almost surely. (1.8) follows by the dominated convergence theorem.

Step 2. L^∞_μ is dense in L^2_μ. That is, for any $f \in L^2_\mu$ and any $\varepsilon > 0$, there exists $g \in L^\infty_\mu$ with $\|f - g\|_{L^2_\mu} < \varepsilon$. Then, as $\widehat{T}$ is an L^2_μ-isometry, $\|A_n f - A_n g\| < \varepsilon$ for any $n \geq 1$, also $|\int f d\mu - \int g d\mu|$ can be controlled in terms of ε, and extension form L^∞_μ to L^2_μ follows by a standard approximation argument. ☐

Proposition 1.11. *Consider an endomorphism $(X, \mathcal{B}, T, \mu)$. The following properties are equivalent:*

(i) *μ is ergodic with respect to T.*

(ii) *$\forall f, g \in L^2_\mu$ we have* $\lim_{n\to\infty} \left(\frac{1}{n} \sum_{k=0}^{n-1} \int f(x) g(T^k x) d\mu(x) \right) = \int f d\mu \cdot \int g d\mu.$

(iii) *$\forall A, B \subset X$ measurable we have* $\lim_{n\to\infty} \left(\frac{1}{n} \sum_{k=0}^{n-1} \mu(T^{-k} A \cap B) \right) = \mu(A) \cdot \mu(B).$

Proof. To prove (ii) $\Rightarrow$ (iii), let $f = \chi_B$ and $g = \chi_A$ where χ stands for the indicator function of a set, which is bounded, hence in L^2_μ. Then $f(x)g(T^k x) = \chi_B(x)\chi_A(T^k x) = \chi_{B\cap T^{-k}A}(x)$, thus

$$\int f(x)g(T^k)d\mu(x) = \mu(T^{-k}A \cap B)$$

and we see that (iii) reduces to (ii).

To prove (iii) $\Rightarrow$ (i), let A be an invariant set, and let $B = A$. Then $T^{-k}A \cap A = A$ almost surely, hence $\frac{1}{n}\sum_{k=0}^{n-1}\mu(T^{-k}A\cap B) = \mu(A)$ for every $n \geq 1$. Hence (iii) gives $\mu(A) = \mu(A)^2$ in this case, which implies that either $\mu(A) = 0$ or $\mu(A) = 1$. Since this holds for any invariant set, $(X, \mathcal{B}, T, \mu)$ is ergodic.

Finally, we prove (i) $\Rightarrow$ (ii). Let us introduce $E_f = \int f d\mu$ for brevity. If $(X, \mathcal{B}, T, \mu)$ is ergodic and $f \in L^2_\mu$, $A_n f(x) \to E_f$ in L^2_μ, by von Neumann's ergodic theorem. Then, for any $g \in L^2_\mu$,

$$|(A_n f - E_f, g)_{L^2_\mu}| \leq \|A_n f - E_f\|_{L^2_\mu}\|g\|_{L^2_\mu} \to 0$$

by the Cauchy-Schwartz inequality. Now

$$\begin{aligned}(A_n f - E_f, g)_{L^2_\mu} &= \left(\frac{1}{n}\sum_{k=0}^{n-1}\int (f(x) - E_f)g(T^k x)d\mu(x)\right) = \\ &= \left(\frac{1}{n}\sum_{k=0}^{n-1}\int f(x)g(T^k x)d\mu(x)\right) - \int f d\mu \cdot \int g d\mu,\end{aligned}$$

so we obtain precisely (ii). □

Definition 1.12. *An endomorphism $(X, \mathcal{B}, T, \mu)$ is mixing if for any $A, B \in \mathcal{B}$*

$$\lim_{n\to\infty}\mu(T^{-n}A \cap B) = \mu(A)\cdot\mu(B). \tag{1.9}$$

Equivalently, it is mixing if for any $f, g \in L^2_\mu$

$$\lim_{n\to\infty}\left(\int f(x)g(T^n x)d\mu(x)\right) = \int f d\mu \cdot \int g d\mu \tag{1.10}$$

Comments:

- Recall that for a numerical sequence a_n:

$$\lim_{n\to\infty} a_n = 0 \implies \lim_{n\to\infty}\frac{1}{n}\left(\sum_{k=1}^{n} a_k\right) = 0$$

but the converse is, in general, false. (Think of a counterexample.) Hence *mixing implies ergodicity.* In the next section we will see a system which is ergodic, but not mixing.

- To verify mixing, by standard approximation arguments, it is enough to check either (1.9) for a collection of sets which generates the sigma algebra $\mathcal{B}$, or (1.10) for a collection of functions which is dense in L^2_μ.
- As discussed earlier, $f, g \in L^2_\mu$ can be regarded as *random variables*, actually, with finite second moment (or, equivalently, finite variance). Hence (1.10) is just the statement that the covariance of the random variables f and $g \circ T^n$ tends to 0 as $n \to \infty$. In other words, mixing is equivalent to the *decay of correlations* as the time gap of the measurements made increases.
- Here we give another aspect of mixing/decay of correlations. Let μ be mixing for $T : X \to X$, and let $m \ll \mu$. Then we know from Lemma 1.7 that (except the trivial case $m = \mu$) the measure m cannot be invariant for T. Now let $m_n = T^n_* m$, the push-forward of μ by T^n, that is, for any A measurable $m_n(A) = m(T^{-n}A)$. Now let $\rho = \frac{dm}{d\mu}$ the Radon-Nikodym derivative, that is, $\rho(x)$ is the density of m with respect to μ. Assume $\rho \in L^2_\mu$. We have

$$\begin{aligned} m_n(A) &= \mu(T^{-n}A) = \int \chi_{T^{-n}A}(x)\rho(x)d\mu(x) \\ &= \int \chi_A(T^n x)\rho(x)d\mu(x) \longrightarrow \int \chi_A d\mu \cdot \int \rho d\mu = \mu(A) \end{aligned}$$

  as $n \to \infty$. This property can be interpreted as "*relaxation to equilibrium*". For decay of correlations/relaxation to equilibrium, the rate – how fast is the convergence in terms of n – is a question of interest, which will be revisited below for specific examples.

2 Ergodic properties of some examples

In what follows we discuss the question of ergodicity and mixing for some examples.

2.1 Irrational rotations

In this subsection $T : \mathbb{S}^1 \to \mathbb{S}^1$ is a rotation with some irrational angle α, and μ denotes the Lebesgue measure on $\mathbb{S}^1$, which is preserved by T.

Proposition 2.1. *The Lebesgue measure is ergodic for the irrational rotation.*

Proof. We have discussed earlier that the irrational rotation is uniquely ergodic, that is, $\mathcal{M}_{\text{inv}} = \{\mu\}$, hence μ is extremal, thus ergodic. Nonetheless, here we argue directly: we show that if $f \in L^2_\mu$ is invariant, then it is constant. It is somewhat easier to construct complex valued functions, by Fourier expansion:

$$f(x) = \sum_{n=-\infty}^{\infty} a_n e^{i2\pi nx},$$

then, as $Tx = (x+\alpha)(\mod 1)$, we have

$$f(Tx) = \sum_{n=-\infty}^{\infty} a_n e^{i2\pi nTx} = \sum_{n=-\infty}^{\infty} a_n e^{i2\pi n(x+\alpha)} = \sum_{n=-\infty}^{\infty} a_n e^{i2\pi n\alpha} e^{i2\pi nx}.$$

Let us introduce $b_n = a_n e^{i2\pi n\alpha}$, which are the Fourier coefficients of $f(Tx)$. If $f(x) = f(Tx)$ almost surely, then, by uniqueness of Fourier coefficients,

$$a_n = b_n, \qquad \forall n \in \mathbb{Z} \qquad \Longrightarrow \qquad \text{either } \mathrm{e^{i2\pi n\alpha}} = 1 \text{ or } \mathrm{a_n} = 0.$$

For $n \neq 0$, the first possibility cannot occur as α is irrational. Hence a_0 is arbitrary, but all other Fourier coefficients vanish, which means that $f(x)$ is (almost surely) constant. □

Proposition 2.2 (Weyl's theorem)**.** *Let $T : \mathbb{S}^1 \to \mathbb{S}^1$ be an irrational rotation, and let $I \subset \mathbb{S}^1$ be an interval. Then, for every $x \in \mathbb{S}^1$*

$$\frac{\#\{k = 0, \ldots, n-1 \,|\, T^k x \in I\}}{n} \longrightarrow \mu(I), \quad \text{as } n \to \infty, \tag{2.1}$$

where $\mu(I)$ is the Lebesgue measure (i.e. the length) of the interval.

Proof. Note that

$$\#\{k = 0, \ldots, n-1 \,|\, T^k x \in I\} = \sum_{k=0}^{n-1} \chi_I(T^k x)$$

so (2.1) is nothing else but (1.1), for the function $f = \chi_I$. We know that by Birkhoff's ergodic theorem this holds for μ-a.e. $x \in \mathbb{S}^1$, however, Proposition 2.2 states this for *every* $x \in \mathbb{S}^1$.

First step: we show that if T is an irrational rotation, and $f : \mathbb{S}^1 \to \mathbb{R}$ is *continuous*, then (1.1) holds for *every* $x \in \mathbb{S}^1$. Let $\varepsilon > 0$ be arbitrary small.

- By compactness of $\mathbb{S}^1$, f is uniformly continuous, hence there exists $\delta > 0$ such that whenever $d(x,y) \leq \delta$, we have $|f(x) - f(y)| \leq \varepsilon$.
- (This is specific to rotations!) $d(T^k x, T^k y) = d(x,y)$ for every $k \geq 1$. So if $d(x,y) \leq \delta$, then $|f(T^k x) - f(T^k y)| \leq \varepsilon$, and by the triangular inequality $|A_n f(x) - A_n f(y)| \leq \varepsilon$.
- Let $A \subset \mathbb{S}^1$ denote the set of points for which the convergence (1.1) is guaranteed by Theorem 1.2. For every $x \in \mathbb{S}^1$ there exists $y \in A$ such that $d(x,y)\delta$ (otherwise, $\mathbb{S}^1 \setminus A$, which is of zero measure, would contain an interval).
- Putting all this together, we see that for every $x \in \mathbb{S}^1$, $|A_n f(x) - \int f d\mu| \leq 2\varepsilon$ for n sufficiently large.

Second step. There exist two continuous functions $f_1 : \mathbb{S}^1 \to \mathbb{R}$ and $f_2 : \mathbb{S}^1 \to \mathbb{R}$ such that $f_1(x) \leq \chi_I(x) \leq f_2(x)$, for every $x \in \mathbb{S}^1$ and $\int (f_2 - f_1) d\mu \leq \varepsilon$. Hence, for every $x \in \mathbb{S}^1$:

$$A_n(f_1)(x) \leq A_n(\chi_I)(x) \leq A_n(f_2)(x)$$

and, as by the first step, $A_n(f_i)(x) \to \int f_i d\mu$ $(i = 1, 2)$, we have

$$\mu(I) - \varepsilon \leq \int f_1 d\mu \leq \liminf_{n\to\infty} A_n \chi_I(x) \leq \limsup_{n\to\infty} A_n \chi_I(x) \leq \int f_2 d\mu \leq \mu(I) + \varepsilon,$$

for every $\varepsilon > 0$. □

Comment: (2.1) can be interpreted as the *equidistribution* of the sequence $x, Tx, T^2x, ...$ on $\mathbb{S}^1$. That is, for any interval $I \subset \mathbb{S}^1$, the frequency of points in I converges to the length of the interval I.

Question 2.3 (Arnold's problem)**.** *Let $a_n \in \{1, \dots, 9\}^{\mathbb{N}}$ be the sequence obtained by considering the first digits in the decimal expansion of the numbers $1, 2, 4, \dots, 2^n, \dots$. Does 8 occur in this sequence? Does 7 occur? Which one is more frequent?*

Question 2.4. *Is the irrational rotation mixing?*

2.2 Bernoulli shifts

Recall that, for $K \geq 2$, $\Sigma_K = \{0, 1, \dots, K-1\}^{\mathbb{Z}}$ and $\Sigma_K^+ = \{0, 1, \dots, K-1\}^{\mathbb{N}}$ denote the set of bi-infinite and semi-infinite sequences with K symbols, respectively, while $\sigma : \Sigma_K \to \Sigma_K$ and $\sigma : \Sigma_K^+ \to \Sigma_K^+$ denote the associated two-sided and one-sided *full* shift maps, respectively. These are compact metric spaces with the symbolic metric, and σ is continuous, hence Theorem 1.4 applies. As mentioned earlier, the set $\mathcal{M}_{\text{inv}}$ is enormous. Here we consider a very important class of invariant measures, the *Bernoulli measures.* The associated endomorphisms/automorphisms are called two-sided/one-sided Bernoulli shifts. Here we focus on the non-invertible $\sigma : \Sigma_K^+ \to \Sigma_K^+$, the invertible $\sigma : \Sigma_K \to \Sigma_K$ can be discussed analogously.

First we fix some terminology.

- Let

$$\Delta_K = \left\{ p = (p_0, p_1, \dots, p_{K-1}) \in \mathbb{R}^K \,|\, p_j \geq 0, (j = 0, \dots, K-1);\ \sum_{j=0}^{K-1} p_j = 1 \right\}$$

  denote the K *dimensional symplex.* An element $p \in \Delta_K$ is just a probability distribution on the finite set $\{0, \dots, K-1\}$.

- Recall that $\underline{x} \in \Sigma_K^+$ is an infinite sequence $\underline{x} = (x_1, x_2, \dots)$. We introduce the notation $\underline{a} = (a_1, \dots, a_\ell) \in \{0, \dots, K-1\}^\ell$ for words of (finite) length $\ell \geq 1$. The *cylinder set* associated to $\underline{a}$ is

$$C_{\underline{a}} = \{\underline{x} \in \Sigma_K^+ \,|\, (x_1, \dots, x_\ell) = \underline{a}\}.$$

  Cylinder sets associated to distinct words of the same length are disjoint.

- Let $\mathcal{F}_\ell$ denote the (finite) sigma algebra generated by cylinder sets of length $\ell \geq 1$. Then $\mathcal{F}_\ell \subset \mathcal{F}_{\ell+1}$, and the collection of cylinder sets of arbitrary length $\bigcup_{\ell=1}^\infty \mathcal{F}_\ell$, generates the Borel sigma algebra $\mathcal{B}$. This implies, in particular, that for any Borel measure $m \in \mathcal{M}$, any $\varepsilon > 0$ and any $A \in \mathcal{B}$ there exists $A_\ell \in \mathcal{F}_\ell$ such that $m(A \Delta A_\ell) < \varepsilon$, where Δ is the symmetric difference. Note also that $A_\ell = \cup_s A_{\ell,s}$, a finite, disjoint union of cylinder sets of length ℓ.

Definition 2.5. *The Bernoulli measure $\mu \in \mathcal{M}_{\text{inv}}(\sigma)$ is defined by*

$$\mu(C_{\underline{a}}) = \prod_{i=1}^{\ell} p_{a_i}$$

for any cylinder set associated to some finite word $\underline{a} = (a_1, \ldots, a_\ell)$. This then extends from $\bigcup_{\ell=1}^\infty \mathcal{F}_\ell$ to $\mathcal{B}$ by sigma additivity.

In probabilistic terms, this means that the letters are i.i.d. and distributed according to the distribution p on $\{0, \ldots, K-1\}$. The measure μ is preserved by σ:

$$\sigma^{-1}(C_{\underline{a}}) = \bigcup_{j=0}^{K-1} C_{j,\underline{a}} \implies \mu(\sigma^{-1}(C_{\underline{a}})) = \sum_{j=0}^{K-1} \left(p_j \cdot \prod_{i=1}^{\ell} p_{a_i} \right) = \prod_{i=1}^{\ell} p_{a_i} = \mu(C_{\underline{a}}).$$

The endomorphism $(\Sigma_K^+, \mathcal{B}, \sigma, \mu)$ is called the *(one-sided) Bernoulli shift* associated to $p \in \Delta_K$.

Proposition 2.6. *Bernoulli shifts are mixing.*

Proof. Fix $A, B \in \mathcal{B}$, and then some $\varepsilon > 0$. We have to show that there exists $N_\varepsilon \geq 1$ such that

$$|\mu(\sigma^{-n} A \cap B) - \mu(A)\mu(B)| < \varepsilon$$

whenever $n \geq N_\varepsilon$. There exist $\ell \geq 1$ and $A_\ell, B_\ell \in \mathcal{F}_\ell$ such that $\mu(A \Delta A_\ell) < \varepsilon/10$ and $\mu(B \Delta B_\ell) < \varepsilon/10$. We claim that $N_\varepsilon = \ell$ works. We have

$$\begin{aligned} |\mu(A) - \mu(A_\ell)| < \varepsilon/10,\ |\mu(B) - \mu(B_\ell)| < \varepsilon/10 \quad &\Rightarrow \quad |\mu(A)\mu(B) - \mu(A_\ell)\mu(B_\ell)| < \frac{2\varepsilon}{10} \\ \mu(\sigma^{-n} A \Delta \sigma^{-n} A_\ell) = \mu(\sigma^{-n}(A \Delta A_\ell)) < \varepsilon/10 \quad &\Rightarrow \quad |\mu(\sigma^{-n} A \cap B) - \mu(\sigma^{-n} A_\ell \cap B_\ell)| < \varepsilon/10, \end{aligned}$$

hence

$$\begin{aligned} |\mu(\sigma^{-n} A \cap B) - \mu(A)\mu(B)| \leq \quad & |\mu(\sigma^{-n} A_\ell \cap B_\ell) - \mu(A_\ell)\mu(B_\ell)| + \\ + \quad & |\mu(\sigma^{-n} A \cap B) - \mu(\sigma^{-n} A_\ell \cap B_\ell)| + |\mu(A)\mu(B) - \mu(A_\ell)\mu(B_\ell)| \end{aligned}$$

where we have already controlled the second and the third term. Here we show that the first term vanishes if $n \geq \ell$. Note that $A_\ell = \cup_i A_{\ell,i}$ and $B_\ell = \cup_{i'} B_{\ell,i'}$, finite disjoint unions of cylinder sets of length ℓ. This means that for any i, i' fixed there exist $\underline{a}, \underline{b} \in \{0, \ldots, K-1\}^\ell$ such that

$$\begin{aligned} \underline{x} \in B_{\ell,i'} &\iff (x_1, \ldots, x_\ell) = \underline{b}; \\ \underline{x} \in \sigma^{-n} A_{\ell,i} &\iff (x_{n+1}, \ldots, x_{n+\ell}) = \underline{a}; \\ \underline{x} \in \sigma^{-n} A_{\ell,i} \cap B_{\ell,i'} &\iff (x_1, \ldots, x_\ell) = \underline{b} \text{ and } (x_{n+1}, \ldots, x_{n+\ell}) = \underline{a}. \end{aligned} \tag{2.2}$$

Hence if $n \geq \ell$, the conditions for $\sigma^{-n} A_{\ell,i} \cap B_{\ell,i'}$ fix the letters at positions $1, \ldots, \ell$ and at $n+1, \ldots, \ell$, and the letters at the positions in between $(\ell, \ldots, n)$ are arbitrary. Thus

$$\mu(\sigma^{-n} A_{\ell,i} \cap B_{\ell,i'}) = \prod_{k=1}^{\ell} p_{a_k} \prod_{k'=1}^{\ell} p_{b_{k'}} = \mu(A_{\ell,i}) \cdot \mu(B_{\ell,i'}),$$

which, when summed on i and i', implies $\mu(\sigma^{-n} A_\ell \cap B_\ell) = \mu(A_\ell)\mu(B_\ell)$. □

2.3 Topological Markov chains and Markov shifts

2.3.1 Preliminaries – recap of finite Markov chains from stochastic processes

A matrix $\pi = \pi_{ij}$ $i, j = 0, ..., (K-1)$ is a *stochastic matrix* (or transition matrix) if

- $\pi_{ij} \geq 0$ $\forall i, j$ and
- $\sum_{j=0}^{K-1} \pi_{ij} = 1$ $\forall i$.

The second bullet says that 1 is an eigenvalue of the matrix with right eigenvector $\mathbf{1} = (1, ..., 1)$.

Interpretation: associated to π there exists a finite state space, discrete time Markov chain; a stationary sequence of random variables $X_1, X_2, ..., X_n, ...$ taking values in the finite set $0, ..., K-1$ such that:

$$\begin{aligned} &P(X_n = i) = P(X_1 = i) = p_i; \\ &P(X_{n+1} = j | X_1 = i_1, ..., X_n = i) = P(X_{n+1} = j | X_n = i) = \pi_{ij}; \end{aligned}$$

Let π^n denote the nth power π.
The matrix π is

- *irreducible* if for any pair i, j there exists n such that $\pi^n_{ij} > 0$.
- *primitive* (or irreducible and aperiodic) if there exists $n_0 \geq 1$ such that for any pair i, j we have $\pi^{n_0}_{ij} > 0$. (This implies that for any $n \geq n_0$ $\pi^n_{i,j} > 0$ $\forall i, j$.)

From now on, we *restrict to the primitive case.*

Theorem 2.7 (Perron's theorem). *Let π be a primitive stochastic matrix. Then*

(a) $\lambda = 1$ is a simple eigenvalue, (with $\mathbf{1} = (1, ..., 1)$ the only right eigenvector up to multiplicity).

(b) for the associated left eigenvector $p = (p_0, \ldots, p_{K-1})$ we have $p_i > 0$ for all i. This is unique up to multiplicity, so we can normalize $\sum_{i=0}^{K-1} p_i = 0$.

(c) all other eigenvalues of π have absolute value strictly less than 1.

(d) $\pi^n_{i,j}$ converges to p_j exponentially. That is, there exists $\beta < 1$ such that $\forall i, j$ we have $\pi^n_{i,j} = p_j + O(\beta^n)$ as $n \to \infty$.

Comment: The theorem generalizes to arbitrary matrices with nonnegative entries $B_{i,j} \geq 0$. In this case primitivity means that there exists n_0 such that $B^{n_0}_{i,j} > 0$, $\forall i, j$. Perron's theorem states that there is a simple eigenvalue λ with an eigenvector that has positive components, and that there is a spectral gap (the rest of the spectrum is contained in a disc of radius strictly smaller than λ).

2.3.2 Definition of topological Markov chains and Markov shifts

A matrix $A = A_{ij}$ is an *(adjacency) matrix*, if all its entries are 0 or 1. Given a stochastic matrix π, the associated adjacency matrix is

$$A_{i,j} = \begin{cases} 0 & \text{if } \pi_{i,j} = 0, \\ 1 & \text{if } \pi_{i,j} > 0. \end{cases}$$

We may think of $A_{i,j}$ as a directed graph.

Recall the definition of $\sigma : \Sigma^+_K \to \Sigma^+_K$. Let

$$\Sigma^+_A = \{(x_1, x_2 ...) = \underline{x} \in \Sigma^+_K | A_{x_k x_{k+1}} = 1; \forall k = 1, 2, ...\},$$

that is, A_{ij} defines the "permitted transitions". In this case $\Sigma^+_A \subset \Sigma^+$ is closed (hence compact) and σ-invariant. Hence we may consider the topological dynamical system $\sigma : \Sigma^+_A \to \Sigma^+_A$, a *topological Markov chain.* (A natural generalization is a subshift of finite type, where the closed and σ-invariant subset of Σ^+_K is defined by forbidding a collection of finite words.)

Comment: Analogously, we may consider two-sided (invertible) topological Markov chains $\sigma : \Sigma_A \to \Sigma_A$. Here we focus on the one-sided case $\sigma : \Sigma^+_A \to \Sigma^+_A$, the two-sided case can be treated similarly.

Note that several stochastic matrices π may correspond to the same adjacency matrix A. Nevertheless, primitivity of π is equivalent to the primitivity of A. For a given stochastic matrix π, we construct the *Markov measure* μ , which is invariant for $\sigma\Sigma^+_A \to \Sigma^+_A$. As in the case of Bernoulli measures, it is enough to specify μ on cylinder sets. For $C_{\underline{b}} \in \mathcal{F}_\ell$ associated to $\underline{b} = (b_1, \ldots, b_\ell)$:

$$\mu(C_{\underline{b}}) = p_{b_1} \pi_{b_1 b_2} \pi_{b_2 b_3} \ldots \pi_{b_{\ell-1} b_\ell},$$

where p_i, $i = 0, \ldots, (K-1)$ is the (unique) stationary distribution corresponding to π_{ij}.

Note that $\mu(C_{\underline{b}})$ is precisely the probability that the associated (stationary) Markov chain assigns to the finite word $\underline{b} = (b_1, \ldots, b_\ell)$.

The endomorphism $(\Sigma_A^+, \mathcal{B}, \sigma, \mu)$ is called the Markov shift (associated to the primitive stochastic matrix π).

2.3.3 Markov shifts corresponding to primitive π are mixing

Proposition 2.8. *Let π_{ij} be a primitive stochastic matrix. The associated Markov shift $(\Sigma_A^+, \mathcal{B}, \sigma, \mu)$ is mixing.*

Proof. As in the proof of Proposition 2.6, we may approximate arbitrary Borel sets B, D with finite disjoint unions of cylinder sets of length ℓ. Hence it remains to show that, if $B_{\ell,i'} = C_{\underline{b}}$ and $D_{\ell,i} = C_{\underline{d}}$ are cylinder sets associated to the words $\underline{b} = (b_1, \ldots, b_\ell)$ and $\underline{d} = (d_1, \ldots, d_\ell)$, respectively, then

$$\mu(\sigma^{-n} D_{\ell,i} \cap B_{\ell,i'}) \longrightarrow \mu(D_{\ell,i}) \cdot \mu(B_{\ell,i'}); \qquad \text{as } n \to \infty.$$

As in the proof of Proposition 2.6,

$$\underline{x} \in \sigma^{-n} D_{\ell,i} \cap B_{\ell,i'} \iff (x_1, \ldots, x_\ell) = \underline{b} \text{ and } (x_{n+1}, \ldots, x_{n+\ell}) = \underline{d}.$$

If $n \geq \ell$:

$$\begin{aligned}
\mu(\sigma^{-n} D_{\ell,i} \cap B_{\ell,i'}) &= \sum_{x_{\ell+1}, \ldots x_n = 0}^{K-1} p_{b_1} \pi_{b_1,b_2} \ldots \pi_{b_\ell x_{\ell+1}} \pi_{x_{\ell+1} x_{\ell+2}} \ldots \pi_{x_n d_1} \pi_{d_1 d_2} \ldots \pi_{d_{\ell-1} d_\ell} \\
&= p_{b_1} \pi_{b_1,b_2} \ldots \pi_{b_{\ell-1} b_\ell} \cdot \pi^{n-\ell+1}_{b_\ell d_1} \cdot \pi_{d_1 d_2} \ldots \pi_{d_{\ell-1} d_\ell} \to \\
&\to p_{b_1} \pi_{b_1,b_2} \ldots \pi_{b_{\ell-1} b_\ell} \cdot p_{d_1} \cdot \pi_{d_1 d_2} \ldots \pi_{d_{\ell-1} d_\ell} = \mu(B_{\ell,i'}) \mu(D_{\ell,i}),
\end{aligned}$$

where we have used Perron's theorem. For future reference, it is worth noting that (again by Perron's theorem) the rate of convergence is exponential, in fact:

$$\mu(\sigma^{-n}(C_{\underline{d}}) \cap C_{\underline{b}}) = \mu(C_{\underline{b}}) \mu(C_{\underline{d}}) \cdot (1 + O(\beta^{n-\ell})). \tag{2.3}$$

□

2.3.4 Exponential decay of correlations

Let (M, d) be a compact metric space. For $\alpha \in (0, 1]$, the function $f : M \to \mathbb{R}$ is *Hölder continuous with exponent α* if there exist $C > 0$ such that $|f(x) - f(y)| \leq C(d(x, y))^\alpha$ for every $x, y \in M$.[5] The smallest possible $C > 0$ for which this holds is called the *Hölder constant* of f, denoted by $C(f, \alpha)$. Hölder continuous functions are continuous, and thus belong to the Banach space $\mathcal{C}$ – let $\|f\|_0$ denote the corresponding supremum norm. However, we may also define:

$$\|f\|_\alpha = \|f\|_0 + C(f, \alpha)$$

the *Hölder norm*, with which Hölder continuous functions make another Banach space, $\mathcal{C}^\alpha = \mathcal{C}^\alpha(M)$.

Let $T : M \to M$ be a topological dynamical system (compact metric space, continuous map), with mixing invariant measure μ. Then for any $f, g \in L^2_\mu$ we have that the time correlations vanish asymptotically, that is:

$$Corr(f, g; n) = \left| \int f(T^n x) g(x) d\mu(x) - \int f d\mu \int g d\mu \right| \longrightarrow 0, \qquad \text{as } n \to \infty.$$

For brevity, we will sometimes use $E(f) = \int f d\mu$, where E stands for expectation.

Definition 2.9. *The dynamical system $T : M \to M$ has exponential decay of correlations with respect to the measure μ if for any $\alpha \in (0, 1]$ there exists $\gamma = \gamma(\alpha) \in (0, 1)$ such that, for any $f, g \in \mathcal{C}^\alpha$ there exists $C(f, g) > 0$ such that*

$$Corr(f, g; n) \leq C(f, g)\gamma^n.$$

Comments

- Even for the "nicest" (or "most chaotic") systems, such as the doubling map, it is possible to construct $f, g \in L^2_\mu$ such that $Corr(f, g; n)$ decays arbitrarily slowly. That is why it is necessary to put some (reasonable) smoothness requirement on the class of functions for which the rate of decay is to be investigated.

- If there is exponential decay of correlations, we often have $C(f, g) = C(M) \cdot \|f\|_\alpha \cdot \|g\|_\alpha$, where the constant $C(M) > 0$ depends only on the map, and not on the pair of functions f, g (or the same type of expression with some other "nice norm").

- To prove exponential decay of correlations, it is enough to consider functions f, g with $E(f) = E(g) = 0$. Indeed, otherwise let $f^0 = f - E(f)$, $g^0 = g - E(g)$, and $Corr(f, g; n) = Corr(f^0, g^0; n)$.

Proposition 2.10. *Consider a mixing Markov shift (that is, corresponding to a primitive π). Correlations decay exponentially.*

Proof. For $\ell \geq 1$ fixed, we will refer to functions measurable with respect to $\mathcal{F}_\ell$ as *(ℓ-)step functions*, which form a function space to be denoted by $\mathcal{S}_\ell$. Functions in $\mathcal{S}_\ell$ take

[5] In the special case of $\alpha = 1$, f is called Lipschitz continuous.

constant values on each of the (disjoint) cylinder sets of length ℓ. Accordingly, $\mathcal{S}_\ell$ is a finite dimensional linear space. For $\widehat{f}, \widehat{g} \in \mathcal{S}_\ell$, by (2.3):

$$Corr(\widehat{f}, \widehat{g}; n) \leq C \cdot \|\widehat{f}\|_0 \|\widehat{g}\|_0 \beta^{n-\ell} \tag{2.4}$$

(recall that $\|f\|_0$ is the supremum norm).

Now let $f, g \in \mathcal{C}^\alpha$, and for n fixed, let us estimate $Corr(f, g; n)$. Let us fix some $\ell \geq 1$, which we will optimize in terms of n later. Let us introduce

$$\widehat{f} = E(f|\mathcal{F}_\ell), \quad \widetilde{f} = f - \widehat{f}, \quad f^n = f \circ T^n, \quad \widehat{f^n} = \widehat{f} \circ T^n, \quad \widetilde{f^n} = \widetilde{f} \circ T^n,$$

and similar notations for g. (Slightly abusing notation, we drop denoting the dependence on ℓ.) Note $E(f) = E(\widehat{f})$ by the tower rule, and apparently $\|\widehat{f}\|_0 \leq \|f\|_0$. Then

$$Corr(f, g; n) = Corr(\widehat{f}, \widehat{g}; n) + E(\widehat{f^n}\widetilde{g}) + E(\widetilde{f^n}\widehat{g}) + E(\widetilde{f^n}\widetilde{g}) \tag{2.5}$$

For the first term, (2.4) applies. On the other hand, cylinder sets of length ℓ have diameter $2^{-\ell}$. As the values of $\widetilde{f}$ are the fluctuations of f on such sets, using Hölder continuity,

$$|\widetilde{f}(\underline{x})| \leq C(f, \alpha) 2^{-\alpha\ell}.$$

A similar bound applies to $\widetilde{g}$. Hence, bounding the integral simply by the supremum, the second term in (2.5) can be estimated by $\|f\|_0 C(g, \alpha) 2^{-\alpha\ell}$. The third and the fourth terms in (2.5) can be treated analogously. We arrive at

$$\begin{aligned} Corr(f, g; n) &\leq C\|f\|_0\|g\|_0\beta^{n-\ell} + (\|f\|_0 C(g, \alpha) + C(f, \alpha)\|g\|_0 + C(f, \alpha)C(g, \alpha))\, 2^{-\alpha\ell} \\ &\leq C\|f\|_\alpha \|g\|_\alpha \gamma^n, \end{aligned}$$

where $\gamma = \gamma(\alpha) = \max(2^{-\alpha}, \beta)^{1/2}$, if $\ell = n/2$.

□

2.4 Markov partitions

In this section we consider *hyperbolic toral automorphisms*, $(T_A) = T : \mathbb{T}^2 \to \mathbb{T}^2$, where $Tx = Ax(\mod \mathbb{Z}^2)$, and the spectrum of $A \in SL(2, \mathbb{Z})$ is disjoint from the unit circle. For simplicity, we will discuss the CAT map associated to

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

Recall that

- *T preserves the Lebesgue measure m.* The aim of the section is to prove that m is mixing for T, in fact, has exponential decay of correlations.

- A has two eigenvalues: $\lambda(= (3+\sqrt{5})/2) > 1$ and λ^{-1}. The associated (unit) eigenvectors are e_u (unstable eigenvector) and e_s (stable eigenvector), respectively.
- Accordingly, for any $x \in \mathbb{T}^2$ define the (global) *stable and unstable manifolds* of x as

  $$W^s(x) = \{x + t\cdot e_s \,|\, t \in \mathbb{R}\}; \qquad W^u(x) = \{x + t\cdot e_u \,|\, t \in \mathbb{R}\},$$

  understood $\mod \mathbb{Z}^2$. As the slope of these vectors is irrational, these manifolds wrap around the torus and cover it densely.
- We have $TW^s(x) = W^s(Tx)$ and $TW^u(x) = W^u(Tx)$, with the distances contracted and expanded by T along the curves, respectively. The origin O is a fixed point, hence $TW^u(O) = W^u(O)$ and $TW^s(O) = W^s(O)$. Intersections in $W^u(O) \cap W^s(O)$ are homoclinic points to the origin.

The directions e_u and e_s define a *local product structure* on $\mathbb{T}^2$ in the following sense. For some fixed small $\delta > 0$, let us define the *local stable and unstable manifolds* as follows:

$$W^s_\delta(x) = \{x + t\cdot e_s \,|\, |t| \le \delta\}; \qquad W^u_\delta(x) = \{x + t\cdot e_u \,|\, |t| \le \delta\}.$$

Then, for any $x, y \in \mathbb{T}^2$, let

$$[x, y] = W^u_\delta(x) \cap W^s_\delta(y)$$

which consists of at most one point, and precisely one point if $d(x, y) < \delta_1 = \delta_1(\delta)$.

Definition 2.11. *$R \subset \mathbb{T}^2$ is a* ***rectangle****, if $R = \overline{int\, R}$ and $\forall x, y \in R$ we have $[x, y] \in R$.*

Immediate properties:

- if R is a rectangle, then so is TR;
- if R and R' are rectangles, then so is $R \cap R'$.

For the particular case of the CAT map, we may consider R-s which are indeed rectangles with sides parallel to e_u and e_s – we may think of them this way for the remainder of the section. In such a case it makes sense to talk about its boundary $\partial R = \partial^u R \cup \partial^s R$, where $\partial^u R$ and $\partial^s R$ are the sides parallel to e^u and e^s, respectively.

Some further properties and terminology.

- Let

  $$W^u(x, R) = \cup_{y\in R}[x, y]; \qquad W^s(x, R) = \cup_{y\in R}[y, x].$$

  If for two rectangles $R_1 \subset R_2$ we have that

  $$\forall x \in R_1 : \quad W^u(x, R_1) = W^u(x, R_2)$$

  then R_1 is a *u-subrectangle* of R_2. This means that R_1 "stretches along R_2 in the u-direction".(For the case of the CAT map this is equivalent to $\partial^s R_1 \subset \partial^s R_2$.) s-subrectangles can be defined analogously.

  It is easy to see that R_1 is a u-subrectangle of R_2 if and only if TR_1 is a u-subrectangle of TR_2, and analogously for s-subrectangles.

- Given two rectangles R and R', TR *properly intersects* R' if $TR \cap R'$ is a u-subrectangle in R', or equivalently, $TR \cap R'$ is an s-subrectangle in TR, or equivalently, $R \cap T^{-1}R'$ is an s-subrectangle in R. In this case we automatically have that whenever $x \in R \cap T^{-1}R'$ (i.e. $x \in R$ and $Tx \in R'$)

$$W^u(Tx, R') \subset T(W^u(x, R)); \qquad \text{and} \qquad W^s(x, R) \subset T^{-1}(W^s(Tx, R')).$$

Definition 2.12. *The collection of finite rectangles $R_0, R_1, ..., R_{K-1}$ is a* ***Markov partition*** *if*

- $\bigcup_{i=0}^{K-1} R_i = \mathbb{T}^2$*;*
- *whenever $i \neq j$, $\mathrm{int}\, R_i \cap \mathrm{int}\, R_j = \emptyset$, (rectangles can only intersect at their boundaries). That is, up to m-measure 0, this is indeed a partition.*
- *if for some pair i, j (possibly for $i = j$) $m(TR_i \cap R_j) \neq 0$, then this intersection is proper and connected.*

Consider a Markov partition, and let s_i and u_i denote the unstable and stable dimensions of rectangle R_i, respectively.

Let us consider now indices i, j, k such that $m(T^2R_i \cap TR_j \cap R_k) \neq 0$. Then $T^2R_i \cap TR_j \cap R_k$ is a u-subrectangle in R_k, or equivalently, $R_i \cap T^{-1}R_j \cap T^{-2}R_k$ is an s-subrectangle in R_i. The essence of the Markov property (proper intersection) is that the unstable and stable dimensions of such a rectangle are determined by R_k, and R_i, respectively, irrespective of the intermediate R_j. In particular:

$$m(T^2R_i \cap TR_j \cap R_k) = \lambda^{-2} u_k s_i.$$

This property extends of intersections of arbitrary multiplicity.

Let us define the adjacency matrix

$$A_{ij} = \begin{cases} 1 & \text{if } m(TR_i \cap R_j) \neq 0, \\ 0 & \text{if } m(TR_i \cap R_j) = 0 \end{cases}$$

If $A_{ij} \neq 0$, then $m(TR_i \cap R_j) = \frac{s_i u_j}{\lambda}$. Let, furthermore

$$\pi_{ij} = \frac{m(TR_i \cap R_j)}{m(R_i)} = \begin{cases} \frac{u_j}{\lambda u_i} & \text{if } A_{ij} = 1, \\ 0 & \text{if } A_{ij} = 0. \end{cases}$$

Then π_{ij} is a stochastic matrix, and

$$p_i = m(R_i) = u_i s_i$$

is a left eigenvector for the eigenvalue 1 – that is, a stationary distribution.

Our aim is to relate the CAT map with the invariant measure m to the (double-sided) Markov shift $(\Sigma_A, \mathcal{F}, \mu_\pi, \sigma)$. Let $(..., i_{-1}, i_0, i_1, ...) = \underline{i} \in \Sigma_A$. Then, as we have seen,

$$m(TR_{i_0} \cap R_{i_1}) = m(R_{i_0} \cap T^{-1}R_{i_1}) = \frac{s_{i_0}u_{i_1}}{\lambda} = p_{i_0}\pi_{i_0 i_1},$$

and

$$m(T^2(R_{i_{-1}}) \cap TR_{i_0} \cap R_{i_1}) = \lambda^2 s_{i_{-1}}u_{i_1} = p_{i_{-1}}\pi_{i_{-1}i_0}\pi_{i_0 i_1}.$$

Furthermore, for any $m, n \in \mathbb{Z}$, $m < n$ let

$$R_{m,n}(i_m, ..., i_n) = T^{-m}R_{i_m} \cap T^{-m-1}R_{i_{m+1}} \cap \cdots \cap T^{-n}R_{i_n}.$$

By the proper intersection property, the measure of this later rectangle can be easily computed, too:

$$m(R_{m,n}(i_m, ..., i_n)) = \frac{s_{i_m}u_{i_n}}{\lambda^{n-m}} = p_{i_m}\pi_{i_m i_{m+1}}...\pi_{i_{n-1}i_n}.$$

Define $\Phi : \Sigma_A \to \mathbb{T}^2$ as follows:

$$\Phi(\underline{i}) = \bigcap_{n=0}^{\infty} R_{-n,n}(i_{-n}, i_{-n+1}, ..., i_{n-1}, i_n).$$

The property that the intersections are connected (whenever $A_{ij} = 1$) is needed for Φ to be well defined. This guarantees that the rectangle $R_{-n,n}(i_{-n}, i_{-n+1}, ..., i_{n-1}, i_n)$ is connected, and as n tends to infinity, its size (both in the s and in the u direction) tends to 0 – actually, exponentially, with rate λ^{-1}. Hence the intersection defining $\Phi(\underline{i})$ consists of a single point.

Φ has the following important properties:

- Φ is an isomorphism of the CAT map and the Markov shift. Indeed, m-almost every point of $\mathbb{T}^2$ arises as the image $\Phi(\underline{i})$ for a unique $\underline{i} \in \Sigma_A$ (a point $x \in \mathbb{T}^2$ is exceptional if there exists $k \in \mathbb{Z}$ and $i \in \{0, ..., K-1\}$ such that $T^k x \in \partial R_i$). Also, for cylinder sets, it has been just checked above that $\Phi_*\mu_\pi = m$. $\Phi \circ \sigma = T \circ \Phi$ holds by construction.

- Φ is a continuous, what is more, a Hölder continuous map. To see this, note that

$$d(\underline{i}, \underline{j}) < 2^{-n} \implies \Phi(\underline{i}), \Phi(\underline{j}) \in R_{-n,n}(i_{-n}, i_{-n+1}, ..., i_{n-1}, i_n)$$

  and thus $d(\Phi(\underline{i}), \Phi(\underline{j})) < \lambda^{-n}$. Hence, for any Hölder continuous function $f : \mathbb{T}^2 \to \mathbb{R}$ we have that $\Phi^* f : \Sigma_A \to \mathbb{R}$ is Hölder continuous, too (where $\Phi^* f(\underline{i}) = f(\Phi(\underline{i}))$).

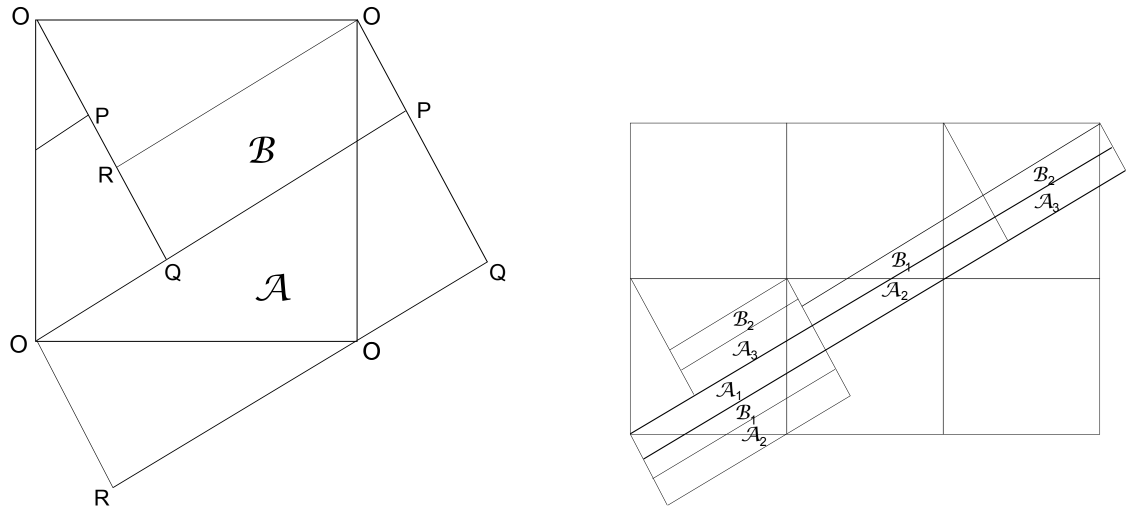
This later property has a special significance as it ensures

$$Corr(n, f, g) = Corr(n, \Phi^* f, \Phi^* g),$$

where $Corr$ denotes, for both systems, the corresponding correlation. Hence, if it can be shown that π_{ij} is primitive, the Markov partition ensures that **the CAT map has exponential decay of correlations**.

Construction of the Markov partition for the CAT map

Let us describe now how to construct a Markov partition for the CAT map (according to Adler and Weiss). As before, let O denote the origin, and let $W^s(O)$ and $W^u(O)$ denote the stable and the unstable manifolds of O, respectively. Let us consider the partition formed by two rectangles $\mathcal{A}$ and $\mathcal{B}$, the stable and unstable sides of which are formed by finite segments about the origin of $W^s(O)$ and $W^u(O)$, respectively. See figure 1, left. In particular the four vertices (of both $\mathcal{A}$ and $\mathcal{B}$) are the origin O and the three homoclinic points P, Q and R.



(a) The partition

(b) Image of the partition

Figure 1: Markov partition for the CAT map

To see that the intersections of $T\mathcal{A}$ and $T\mathcal{B}$ are proper both with $\mathcal{A}$ and with $\mathcal{B}$, we make the following observations.

- The stable sides $\partial^s\mathcal{A}$ and $\partial^s\mathcal{B}$ are formed by the segment L_s of $W^s(O)$ between O and Q. The image of this segment TL_s is a shorter segment of $W^s(O)$, that is $TL_s \subset L_s$. This means in particular that $(\partial^s T\mathcal{A} \cup \partial^s T\mathcal{B}) \subset (\partial^s\mathcal{A} \cup \partial^s\mathcal{B})$. If TR enters R' at a stable side, it has to go all the way through.

- The unstable sides $\partial^u\mathcal{A}$ and $\partial^u\mathcal{B}$ are formed by the segment L_u of $W^u(O)$ between R and P. The image of this segment TL_u is a longer segment of $W^u(O)$, that is $L_u \subset TL_u$. This means in particular that $(\partial^u T\mathcal{A} \cup \partial^u T\mathcal{B}) \supset (\partial^u\mathcal{A} \cup \partial^u\mathcal{B})$. Hence no unstable side of a rectangle R' can be found in the interior of some TR.

The images of the rectangles $\mathcal{A}$ and $\mathcal{B}$ are depicted on Figure 1, right. We have $T\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3$: going rightwards along $W^u(O)$ from O, $T\mathcal{A}$ first intersects $\mathcal{A}$ in $\mathcal{A}_1$, then again $\mathcal{A}$ in $\mathcal{A}_2$, and finally $\mathcal{B}$ in $\mathcal{A}_3$. Also, $T\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$: Going leftwards along $W^u(O)$ from O, $T\mathcal{A}$ first intersects $\mathcal{B}$ in $\mathcal{B}_2$, and then $\mathcal{A}$ in $\mathcal{B}_1$.

There is only one issue why the obtained partition does not fulfill entirely the requirements of Definition 2.12: the intersection $T\mathcal{A} \cap \mathcal{A}$ is not connected. However, if we consider instead the iterated rectangles $\mathcal{A}_1, \mathcal{B}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{B}_2$ as elements of our partition, all requirements of Definition 2.12 are fulfilled. The adjacency matrix is:

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

3 Hyperbolic dynamical systems

In this section we again focus on topological properties. Our aim is to give an introduction to hyperbolic dynamics.

3.1 Attractors

Let M be a complete metric space, and $T : M \to M$ continuous, with continuous inverse T^{-1}.

Definition 3.1. *An open set $U \subset M$ is a trapping region if for its closure $N = \overline{U}$ we have $TN \subset U$. In this case $\Lambda = \bigcap_{n=0}^{\infty} T^n N$ is the corresponding attractor.*

In this case

- the attractor Λ is a closed and T-invariant set;
- $\forall x \in U$ we have $T^n x \to \Lambda$ as $n \to \infty$.

By the first bullet, we may consider the restriction of the dynamics to the attractor, obtaining this way $T : \Lambda \to \Lambda$, a topological dynamical system. The attractor is called *transitive* is $T : \Lambda \to \Lambda$ is topologically transitive.

Simple examples are an attracting fixed point ($M = \mathbb{R}$) or a sink ($M = \mathbb{R}^2$). Similarly, a source is an attractor for T^{-1}. In the next subsection we describe a more interesting example.

3.1.1 The Solenoid

Let

$$M = \mathbb{S}^1 \times \mathbb{B}^2 \qquad \text{where} \qquad \mathbb{B}^2 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\},$$

That is, $\mathbb{B}^2$ is the unit disc in the plane, which we consider as a subset of $\mathbb{C}$ (i.e. introduce $z = x + iy$). We think of M as the *solid torus*. On M we will use the coordinates $(t, z) \in M$, $t \in \mathbb{S}^1$ and $z \in \mathbb{B}^2$.

Now let us define $T : M \to M$ as

$$T(t,z) = \left(g(t), \frac{z}{10} + \frac{1}{2}e^{2\pi i t} \right); \qquad \text{where } g : \mathbb{S}^1 \to \mathbb{S}^1,\ g(t) = 2t \pmod 1.$$

That is, $g(t)$ is the doubling map. To describe TM, introduce the notation, given any $t \in \mathbb{S}^1$

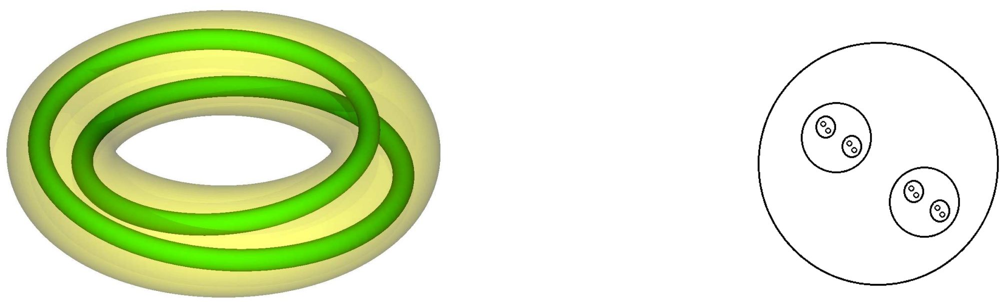
$$D_t = \{t\} \times \mathbb{B}^2 \subset M$$

for the cross-section of M at t.

Let us consider some $t_0 \in [0, 1/2) \subset \mathbb{S}^1$, and let $t_1 = g(t_0) = g(t_0 + 1/2)$. Accordingly:

$$D_{t_1} \cap TM = T(D_{t_0}) \cup T(D_{t_0+1/2})$$

$T(D_{t_0})$ and $T(D_{t_0+1/2})$ are disjoint discs of radius $\frac{1}{10}$ each, and centers distance $1/2$ apart. Accordingly, TM is a tube of radius $1/10$, which wraps around the solid torus M twice. See Figure 2.



(a) The first iterate

(b) Cross section

Figure 2: Solenoid

To proceed, note that by $M \subset TM$ we have

$$M \subset TM \subset T^2M \subset \cdots \subset T^nM \subset \ldots$$

T^nM has the following properties:

- for any $t \in \mathbb{S}^1$, $D_t \cap T^nM$ consists of 2^n disjoint discs of radius 10^{-n} each;
- T^nM is a tube of radius 10^{-n}, which wraps around the torus 2^n times.
- its volume tends to 0 exponentially as $n \to \infty$.

Now consider the attractor $\Lambda = \bigcap_{n=0}^{\infty} T^nM$.

- Λ has zero Lebesgue measure.

- For any $t \in \mathbb{S}^1$, $\Lambda_t = D_t \cap \Lambda$ is a Cantor set.
- Λ_t varies continuously when changing t.

Now we prove that Λ is a transitive attractor, more precisely, that $T : \Lambda \to \Lambda$ is topologically conjugate to the two-sided full shift $\sigma : \Sigma_2 \to \Sigma_2$. Consider the Markov partition:

$$V_0 = [0, 1/2] \times \mathbb{B}^2; \qquad V_1 = [1/2, 1] \times \mathbb{B}^2.$$

For any $(i_0, i_1, \ldots, i_{m-1}) \in \{0, 1\}^m$, let

$$V_{i_0,i_1,\ldots,i_{m-1}} = \bigcap_{k=0}^{m-1} T^{-k} V_{i_k},$$

which, by standard properties of the doubling map, is a short tube of radius 1 and length 2^{-n}. Then, for any $(j_n, \ldots, j_1) \in \{0, 1\}^n$, let

$$H_{j_n,\ldots,j_1} = T^n V_{j_n,\ldots,j_1}$$

which is a tube of radius 10^{-n} and length 1 (it goes around the torus once, but does not close). Finally, for $(i_{-n}, \ldots, i_{-1}, i_0, i_1, \ldots, i_n) \in \{0, 1\}^{2n+1}$, let

$$S_{i_{-n},\ldots,i_{-1},i_0,i_1,\ldots,i_n} = H_{i_{-n},\ldots,i_{-1}} \cap V_{i_0,i_1,\ldots,i_n} = \bigcap_{k=-n}^{n} T^{-k} V_{i_k}$$

As $n \to \infty$, the diameter of $(i_{-n}, \ldots, i_{-1}, i_0, i_1, \ldots, i_n)$ tends to 0. Now, for

$$\underline{i} = (\ldots, i_{-n}, \ldots, i_{-1}, i_0, i_1, \ldots, i_n, \ldots) \in \Sigma_2,$$

let

$$\pi : \Sigma_2 \to \Lambda; \qquad \pi(\underline{i}) = \bigcap_{n=0}^{\infty} S_{i_{-n},\ldots,i_{-1},i_0,i_1,\ldots,i_n}$$

which is well defined, as the intersection consists of a single point. π topologically conjugates $T : \Lambda \to \Lambda$ with $\sigma : \Sigma_2 \to \Sigma_2$.

3.2 Unstable manifold theorem

In what follows we study $T : M \to M$ where M is a compact Riemannian manifold. However, a strong background in differential geometry is *not* needed. As most of our considerations are local, we may just think of M as being locally Euclidean, and T as map of (an open ball of) $\mathbb{R}^d$. Most of our examples are two dimensional, so M is just a closed surface, like the sphere or the torus. The differential geometric notions that show up are as follows.

- The *tangent space* $\mathcal{T}_x M$ at a point $x \in M$. It is a d dimensional vector space. M is Riemannian, which means that the spaces $\mathcal{T}_x M$ have a natural Euclidean structure. This generates a distance on M.

- For $T : M \to M$, consider a point $x \in M$ and its image $Tx \in M$. This then generated a *tangent map* (essentially, the derivative) $DT_x : \mathcal{T}_x M \to \mathcal{T}_{Tx} M$. DT_x is an *invertible* linear mapping between the d dimensional tangent spaces. Invertibility of DT_x follows as it is assumed that $T : M \to M$ is a diffeomorphism.

3.2.1 Unstable manifold theorem for a hyperbolic fixed point

Assume, for simplicity, that M is two dimensional. Let us consider a fixed point $x_0 = Tx_0$ for $T : M \to M$. Then DT_{x_0} is just an invertible linear map of $\mathcal{T}_{x_0} M$ (which we may identify with $\mathbb{R}^2$) onto itself. That is, DT_{x_0} is a 2×2 invertible matrix. We restrict to the case of *hyperbolic* fixed points, which means that the spectrum of DT_{x_0} is disjoint from the unit circle. Specifically, we assume that x_0 is a *saddle*,and thus DT_{x_0} has two eigenvalues $0 < \lambda < 1 < \mu$. The unstable manifold theorem formulates the phenomena that in leading order the dynamical behavior near x_0 is determined by the matrix DT_{x_0}.

The theorem naturally generalizes to periodic points of period $n_0 \geq 1$. In this case, we consider the tangent map of the iterate, $DT^{n_0}_{x_0}$. Denoting $x_k = T^k x_0$, $k = 1, \ldots, (n_0 - 1)$ we have that

$$DT^{n_0}_{x_k} = DT^k_{k_k} DT^{n_0}_{x_0} (DT^k_{k_k})^{-1}$$

and thus the spectrum of $DT^{n_0}_{x_k}$ is the same for all points along the periodic orbit.

The unstable manifold manifold theorem below is a local statement, so we may consider $T : \mathbb{R}^2 \to \mathbb{R}^2$ instead of $T : M \to M$.

Theorem 3.2 (Unstable manifold theorem for a saddle)**.** *Let us consider $T : \mathbb{R}^2 \to \mathbb{R}^2$ continuously differentiable, and $p = T(p)$ a saddle point for T. Then there exists $\varepsilon > 0$ and a continuously differentiable curve $\gamma^u_{loc} : (-\varepsilon, \varepsilon) \to \mathbb{R}^2$ such that*

- $\gamma^u_{loc}(0) = p$,

- *The tangent vector of γ^u_{loc} at p satisfies $(\gamma^u_{loc})'(0) \neq 0$, and it is , actually, an unstable eigenvector of DT_p,*

- *γ^u_{loc} is invariant under T^{-1},*

- *for any $t \in (-\varepsilon, \varepsilon)$, $T^{-n}(\gamma^u_{loc}(t)) \to p$ as $n \to \infty$,*

- *if $|T^{-n}(q) - p| < \varepsilon$ for all $n \geq 0$, then $q = \gamma^u_{loc}(t)$ for some $t \in (-\varepsilon, \varepsilon)$.*

The curve γ^u_{loc} is called the local unstable manifold of p.

Proof. Preparations. We may assume that $p = 0$ (the origin) and that

$$DT_p = D = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}.$$

This is just an affine change of the coordinates on $\mathbb{R}^2$ (moving the origin to p and then conjugating the matrix to its canonical form). Also, we may assume that $\lambda > 2$ and $\mu < \frac{1}{2}$ If this does not hold for T, we may switch to a higher iterate T^{n_0}, for which p remains a saddle fixed point.

For tangent vectors $\begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in \mathbb{R}^2 (= \mathcal{T}_0 M)$, we introduce the *stable and unstable cones* as

$$S^s(0) = \left\{ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in \mathbb{R}^2 \;\middle|\; |\alpha| < \frac{1}{2}|\beta| \right\} \qquad S^u(0) = \left\{ \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in \mathbb{R}^2 \;\middle|\; |\alpha| > 2|\beta| \right\}.$$

Also, for

$$\begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix} \in \mathbb{R}^2, \quad \text{let} \quad \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix} = D \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \alpha_{-1} \\ \beta_{-1} \end{pmatrix} = D^{-1} \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}.$$

We have

$$\text{If } \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix} \in S^u(0), \text{ then } \begin{pmatrix} \alpha_1 \\ \beta_1 \end{pmatrix} \in S^u(0), \text{ thus } |\alpha_1| > 2|\beta_1| \text{ and } |\alpha_1| > 2|\alpha_0|. \tag{3.1}$$

Also

$$\text{If } \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix} \in S^s(0), \text{ then } \begin{pmatrix} \alpha_{-1} \\ \beta_{-1} \end{pmatrix} \in S^s(0), \text{ thus } |\beta_{-1}| > 2|\alpha_{-1}| \text{ and } |\beta_{-1}| > 2|\beta_0|. \tag{3.2}$$

Choice of ε. Conditions (3.1) and (3.2) are *open*: if they hold for a matrix D, they remain true for any other matrix $\hat{D}$ all elements of which are sufficiently close to the elements of D. For some small $\varepsilon > 0$ to be determined, let

$$U = \{(x, y) \in \mathbb{R}^2 \mid |x| < \varepsilon, |y| < \varepsilon\}.$$

By choosing ε small enough, as $T : M \to M$ is C^1, it can be ensured that (i) for $q \in U$, the matrices $\hat{D} = DT_q$ are arbitrarily close to $D (= DT_0)$; (ii) second order effects can be made arbitrary small. In particular, by choosing ε small, the following properties can be ensured.

- Consider $p_1 = (x_0, \varphi_1(x_0)) \in U$ and $p_2 = (x_0, \varphi_2(x_0)) \in U$, that is, two points that have the same horizontal coordinate. Let $T^{-1} p_1 = q_1 = (x_1, y_1)$ and $T^{-1} p_2 = q_2 = (x_2, y_2)$ denote the preimages of these two points. Then

$$|y_2 - y_1| > 2|\varphi_2(x_0) - \varphi_1(x_0)|, \qquad |y_2 - y_1| > 2|x_2 - x_1|, \tag{3.3}$$

  by an extension of (3.2) from the tangent plane $\mathcal{T}_0 M$ to the small neighborhood U.

- Another requirement on the smallness of ε is that the *horizontal curves are preserved* by T (see Definition 3.3 below).

Definition 3.3 (Horizontal curves)**.** *A curve* $\{(x, h(x)) \,|\, |x| \leq \varepsilon\}$ *is horizontal if*

- $h(0) = 0$,
- *h is Lipschitz continuous with Lipschitz constant $\frac{1}{2}$; that is, for any x, x' we have* $|h(x) - h(x')| \leq \frac{|x-x'|}{2}$.

Let $\mathcal{H}$ denote the space of horizontal curves.

If we apply T to each point of a horizontal curve, we obtain a new curve. By choosing ε sufficiently small, it can be ensured that this image curve is horizontal, too. This follows as 0 is a fixed point and by the extension of (3.1) from the tangent plane $\mathcal{T}_0 M$ to the small neighborhood U. This way we obtain a map $\Phi : \mathcal{H} \to \mathcal{H}$ called a *the graph transform.*

Let us introduce the following *metric* on $\mathcal{H}$:

$$d(\gamma_1, \gamma_2) = \sup_{-\varepsilon \leq x \leq \varepsilon} |h_1(x) - h_2(x)| \qquad \mathcal{H} \ni \gamma_i = (x, h_i(x)); \; i = 1, 2.$$

With this metric, $\mathcal{H}$ is a complete metric space.

Lemma 3.4. *There exists some $\kappa < 1$ such that $d(\Phi\gamma_1, \Phi\gamma_2) \leq \kappa d(\gamma_1, \gamma_2)$.*

Hence we have that $\Phi : \mathcal{H} \to \mathcal{H}$ is a contraction of the complete metric space $\mathcal{H}$. Then by the *Banach fixed point theorem* there is a unique fixed point $\gamma^u_{loc} \in \mathcal{H}$, $\Phi\gamma^u_{loc} = \gamma^u_{loc}$. This horizontal curve which is preserved by the graph transform will be the local unstable manifold.

Indeed, if $\gamma^u_{loc} \ni q = (x_0, h(x_0))$, then $T^{-1}q = (x_{-1}, h(x_{-1})) \in \gamma^u_{loc}$, and $|x_{-1}| < \frac{1}{2}|x_0|$ by (3.1). Hence $T^{-n}q \to 0$ as $n \to \infty$.

On the other hand, if $q \in U \setminus \gamma^u_{loc}$, then there is a vertical segment connecting q to γ^u_{loc}. Applying repeatedly (3.3), it is obtained that the vertical distance of $T^{-n}q$ from γ^u_{loc} keeps growing, and sooner or later $T^{-n}q$ has to leave the neighborhood U.

It remains to prove Lemma 3.4 – for instance, with $\kappa = \frac{5}{6}$. We argue by contradiction. Let us assume that there exist $\gamma_1, \gamma_2 \in \mathcal{H}$ such that

$$d(\Phi\gamma_1, \Phi\gamma_2) > \kappa d(\gamma_1, \gamma_2). \tag{3.4}$$

Let us introduce the notation

$$\gamma_i = \{(z, h_i(z)) \,|\, |z| \leq \varepsilon\} \quad \text{and} \quad \Phi\gamma_i = \{(x, \varphi_i(x)) \,|\, |x| \leq \varepsilon\},$$

$i = 1, 2$. (3.4) implies that there exists some x_0 for which

$$|\varphi_1(x_0) - \varphi_2(x_0)| > \kappa \sup_{|z| \leq \varepsilon} |h_1(z) - h_2(z)|. \tag{3.5}$$

Let $p_i = (x_0, \varphi_i(x_0))$, $i = 1, 2$. the preimages of these two point are $T^{-1}p_i = q_i = (z_i, h_i(z_i)) \in \gamma_i$, $i = 1, 2$. (3.3) applies:

$$\begin{aligned} |h_2(z_2) - h_1(z_1)| &> 2|\varphi_2(x_0) - \varphi_1(x_0)| \text{ and also} && (3.6)\\ |h_2(z_2) - h_1(z_1)| &> 2|z_2 - z_1| && (3.7) \end{aligned}$$

(3.5) and (3.6) imply:

$$|h_2(z_2) - h_1(z_1)| > 2\kappa |h_1(z_1) - h_2(z_1)|.$$

Then, by the triangular inequality:

$$|h_2(z_2) - h_2(z_1)| \geq |h_2(z_2) - h_1(z_1)| - |h_1(z_1) - h_2(z_1)| > (1 - (2\kappa)^{-1})|h_2(z_2) - h_1(z_1)|,$$

which, by (3.7), results in

$$|h_2(z_2) - h_2(z_1)| > 2 \cdot (1 - (2\kappa)^{-1}) \cdot |z_2 - z_1| = 0.8|z_2 - z_1|$$

if $\kappa = \frac{5}{6}$. But this contradicts that γ_2 is a horizontal curve. Hence Lemma 3.4 is proved.

We are almost done, just need to recall that this way we have constructed γ^u_{loc}, a local unstable manifold for a *fixed higher iterate* T^{n_0}, and we still need to argue that γ^u_{loc} is a local unstable manifold for the initial map T as well. The only way this can fail is that $T^{-1}\gamma^u_{loc} \neq \gamma^u_{loc}$ but another horizontal curve. But then $T^{-1}\gamma^u_{loc}$ would be another local unstable manifold for T^{n_0}. But this would contradict the uniqueness of the local unstable manifold for T^{n_0}, expressed in the last bullet of the characterization of Theorem 3.2. ☐

Analogously, the local stable manifold of the hyperbolic fixed point p, $\gamma^s_{loc}(p)$ could be constructed. If p is not a saddle but a sink/source, than the entire small neighborhood U is $\gamma^s_{loc}(p)/\gamma^u_{loc}(p)$, respectively. Higher dimensional cases could be discussed analogously, just instead of curves we may have higher dimensional submanifolds.

The *global stable and unstable manifolds* of p can be defined as

$$\begin{aligned} \gamma^s(p) &= \bigcup_{m\geq 1} T^{-m}\gamma^s_{loc}(p) = \{q \in M \,|\, T^n q \to p\}, \quad \text{as } n \to \infty, \\ \gamma^u(p) &= \bigcup_{m\geq 1} T^{m}\gamma^u_{loc}(p) = \{q \in M \,|\, T^{-n} q \to p\}, \quad \text{as } n \to \infty. \end{aligned}$$

3.3 Hyperbolic sets

For $T : M \to M$, $\Lambda \subset M$ is a *hyperbolic set* if:

- $T\Lambda = \Lambda$ and Λ is compact. (So we may consider the restriction $T : \Lambda \to \Lambda$ which is a topological dynamical system.)

- For every $x \in \Lambda$ the tangent plane $\mathcal{T}_x M$ can be represented as the direct sum

$$\mathcal{T}_x M = E^u_x \oplus E^s_x, \tag{3.8}$$

where E^u_x and E^s_x are called the *unstable and the stable subspace* of x, respectively.

  - The subspaces depend on x continuously;
  - For the tangent maps we have $DT_x E^u_x = E^u_{Tx}$ and $DT_x E^s_x = E^s_{Tx}$, accordingly, (3.8) is called an *invariant splitting*.
  - There exist constants $C > 0$ and $0 < \lambda < 1$ such that

$$\begin{aligned} \forall x \in \Lambda, \forall v \in E^u_x, \forall n \geq 1 : \quad & |DT^{-n}_x v| \leq C\lambda^n |v|; \\ \forall x \in \Lambda, \forall v \in E^s_x, \forall n \geq 1 : \quad & |DT^{n}_x v| \leq C\lambda^n |v|. \end{aligned}$$

The simplest example is a hyperbolic fixed point $\Lambda = \{p\}$. Further examples are hyperbolic toral automorphisms, when $M = \Lambda$, or the solenoid map, when Λ is the attractor.

Given a hyperbolic set, with techniques similar to the ones used in the proof of Theorem 3.2 (but with more work), it can be proved that there exists some $\varepsilon > 0$ and for any $x \in \Lambda$ a local stable manifold

$$W^s_{loc}(x) = \{y \in M \,|\, d(T^n x, T^n y) \leq \varepsilon, \forall n \geq 0\}.$$

$W^s_{loc}(x)$ depends on x continuously, satisfies $\mathcal{T}_x W^s_{loc}(x) = E^s_x$, and for $y \in W^s_{loc}(x)$ we have $d(T^n x, T^n y) \leq C\lambda^n d(x, y)$. The global stable manifold of x is defined as

$$W^s(x) = \bigcup_{m \geq 1} T^{-m} W^s_{loc}(x) = \{y \in M \,|\, d(T^n x, T^n y) \to 0\}, \quad \text{as } n \to \infty.$$

Local and global unstable manifolds of $x \in \Lambda$ can be defined analogously.
Here we define two major classes of smooth, uniformly hyperbolic dynamical systems.

Definition 3.5 (Anosov maps). *$T : M \to M$ is an Anosov diffeomorphism if the entire phase space, M is a hyperbolic set.*

The main example is the CAT map (or any other hyperbolic toral automorphism). To define the other major class, we need a little more terminology. Given a topological dynamical system $T : M \to M$

- For $x \in M$ the *ω-limit* points (and, for invertible maps, the α-limit points) of x are

$$\omega(x) = \bigcap_{n \in \mathbb{Z}^+} \overline{\bigcup_{j \geq n} T^j x}; \qquad \left(\alpha(x) = \bigcap_{n \in \mathbb{Z}^+} \overline{\bigcup_{j \geq n} T^{-j} x} \right)$$

that is $y \in \omega(x)$ ($y \in \alpha(x)$) if and only if there is a subsequence $n_k \to \infty$ such that $T^{n_k} x \to y$ ($T^{-n_k} x \to y$).

- The *recurrent points* of T are:

$$\mathcal{R}(T) = \{x \in M \mid x \in \omega(x)(\cap\alpha(x))\}.$$

- $x \in M$ is a *non-wandering* point for T if given any open neighborhood $U \ni x$ there exists $n \geq 1$ such that $T^n U \cap U \neq \emptyset$. The collection of non-wandering points is denoted by $\Omega(T)$.

Let, furthermore $Per(T) = \{x \in M \mid \exists n_0 : T^{n_0}x = x\}$ denote the periodic points of $T : M \to M$. Then

$$Per(T) \subset \mathcal{R}(T) \subset \Omega(T),$$

moreover, $Per(T)$ and $\mathcal{R}(T)$ are not necessarily closed sets, but $\Omega(T)$ is always closed.

Definition 3.6 (Axiom A attractors)**.** *$T : M \to M$ is an Axiom A diffeomorphism if*

- *$\Omega(T)$ is a hyperbolic set,*
- *$\Omega(T) = \overline{Per(T)}$.*

The typical example is the solenoid, when $\Omega(T)$ is the attractor Λ.
We will see later that Anosov maps satisfy Axiom A, that is, for Anosov maps (hyperbolic) periodic points are dense in the phase space M.

3.4 Shadowing

3.4.1 The shadowing property

Let $T : M \to M$ be a topological dynamical system.

- A sequence of points $x_0, x_1, \ldots, x_n \in M$ is a *δ-pseudo orbit* if for any $i = 1, \ldots n$ we have $d(x_i, Tx_{i-1}) < \delta$. Analogously, infinite δ-pseudo orbits – and, for invertible T, bi-infinite δ-pseudo orbits – can be defined.

- A δ-pseudo orbit is *ε-shadowed* by a true orbit if there exists $y \in M$ such that $d(T^k y, x_k) < \varepsilon$, for any $k = 0, \ldots, n$.

- The dynamical system $T : M \to M$ has the *shadowing property* if for any $\varepsilon > 0$ there exists $\delta > 0$ such that any (infinite, in the invertible case bi-infinite) δ-pseudo orbit is ε-shadowed by some true orbit.

The significance of the shadowing property is that it expresses the stability of the phase portrait with respect to small perturbations. As it is demonstrated below, the shadowing property is characteristic to hyperbolic systems. These can be thought of as chaotic, as (typical) small perturbations grow rapidly (exponentially) with time. This instability applies, however, only to the individual orbits. The phase portrait as a whole is stable for small enough perturbations.

As a first example, let us show that – rational or irrational – *rotations do not have the shadowing property.* To see this, let $T : \mathbb{S}^1 \to \mathbb{S}^1$ be a rotation, and construct a δ-pseudo orbit in such a way that we always make a small perturbation in the clockwise direction. That is, $x_i = Tx_{i-1} + \delta$, and thus, $x_n = T^n x_0 + n\delta$ (addition is always understood $\mod 1$). In particular, $d(x_n, T^n x_0) = n\delta$. Now, let us assume that there exits some $y \in \mathbb{S}^1$ the orbit of which ε-shadows this pseudo orbit. Then $d(x_0, y) < \varepsilon$. As the rotation is an isometry, this implies $d(T^n x_0, T^n y) < \varepsilon$. But this then implies $d(x_n, T^n y) \geq n\delta - \varepsilon$ which can be made macroscopic for large enough n.

Lemma 3.7. *The doubling map has the shadowing property.*

Proof. Let $T : \mathbb{S}^1 \to \mathbb{S}^1$ be now the doubling map. We will prove the shadowing property with $\delta = \varepsilon$. Let us consider first a finite δ-pseudo orbit $x_0, x_1, \ldots, x_n$. We construct the shadowing point $y(= y^{(n)})$ in such a way that we *fix first the endpoint* of the orbit by letting $T^n y = x_n$. Note that as the doubling map is not invertible, this does not determine y uniquely. Not even $T^{n-1}y$ is uniquely determined, as the point $T^n y$ has two pre-images under T. We will choose first $T^{n-1}y$, then $T^{n-2}y$, and so on, iteratively, we pick $T^k y$ for decreasing k.

As we have a pseudo orbit, $d(Tx_{n-1}, T^n y) = d(Tx_{n-1}, x_n) < \delta$. Now Tx_{n-1} has two preimages under T, one in $[0, 1/2)$ and another in $[1/2, 1)$, one of which is x_{n-1}. Similarly, $T^n y$ has two preimages. If we obtain $T^{n-1}y$ by applying the inverse of the branch that maps x_{n-1} to Tx_{n-1}, then we have, by the contraction of the inverse branch, $d(x_{n-1}, T^{n-1}y) < \frac{\delta}{2}$. As we have a pseudo orbit, this implies, in turn:

$$d(Tx_{n-2}, T^{n-1}y) \leq d(Tx_{n-2}, x_{n-1}) + d(x_{n-1}, T^{n-1}y) < \frac{3\delta}{2} < 2\delta.$$

Now we proceed inductively. Let us assume that $T^k y$ has already been chosen such that $d(T^k y, x_k) < \delta$. Then, as we have a δ-pseudo orbit, by the triangular inequality $d(T^k y, Tx_{k-1}) < 2\delta$. Choosing the preimage $T^{k-1}y$ appropriately, we can arrange $d(T^{k-1}y, x_{k-1}) < \delta$.

To extend to infinite pseudo orbit, note that this way we have a sequence of points $y^{(n)}$ that shadow longer and longer sections of the pseudo orbit. We claim that this is a Cauchy sequence. To see this, note that $T^n y^{(n)} = x_n$, and thus $d(T^n y^{(n)}, T^n y^{(n+1)}) < \delta$. As we follow the same inverse branches for n iterations, this implies $d(y^{(n)}, y^{(n+1)}) < \frac{\delta}{2^n}$. The limit point $y^{(\infty)}$ shadows the infinite pseudo orbit. □

Lemma 3.8. *The CAT map has the shadowing property.*

Proof. The main idea in the proof of Lemma 3.7 was to use the non-invertibility (and the expansion) of the dynamics. Here the map is invertible, but there are the stable and unstable manifolds. The proof, actually, works for open neighborhoods of hyperbolic sets in general.

Let $T : M \to M$ denote the CAT map. Recall the following notations: $\lambda > 1$ is the unstable eigenvalue, and λ^{-1} is the stable eigenvalue of the matrix. Also, for $x, y \in M$:

$$[x, y] = W^u_\delta(x) \cap W^s_\delta(y)$$

which is a unique point for x and y sufficiently close.

Let us consider a δ-pseudo orbit $x_0, x_1, \ldots, x_n \in M$. We will find, recursively, points y_k that shadow the pseudo orbit up to time k, that is, $d(T^i y_k, x_i) < \varepsilon$ (where ε is a fixed multiple of δ) for $i = 0, 1, \ldots, k$.
To start, let $y_0 = x_0$.
Now, to find y_1, we construct Ty_1, which determines y_1 uniquely (T is invertible). Let

$$Ty_1 = [Ty_0, x_1].$$

Note that $y_0 = x_0$, so $d(Ty_0, x_1) < \delta$ as we have a pseudo orbit. This implies

$$d(Ty_0, Ty_1) < \delta,\ Ty_1 \in W^u_{loc}(Ty_0) \implies d(y_0, y_1) < \delta\lambda^{-1}$$

as unstable manifolds are contracted by the inverse map T^{-1}. Also,

$$d(x_1, Ty_1) < \delta,\ Ty_1 \in W^s_{loc}(x_1) \implies d(T^2 y_1, Tx_1) < \delta\lambda^{-1}.$$

This, on the one hand, shows that the orbit y_1, Ty_1 shadows the two-element pseudo orbit x_0, x_1. On the other hand, prepares the next step of the iterative construction. The distance of $T^2 y_1$ and x_2 is controlled by the triangular inequality:

$$d(T^2 y_1, x_2) \leq d(T^2 y_1, Tx_1) + d(Tx_1, x_2) \leq \delta(1 + \lambda^{-1}).$$

So we may define y_2 via $T^2 y_2$ as

$$T^2 y_2 = [T^2 y_1, x_2].$$

This way

$$d(T^2 y_2, T^2 y_1) < \delta(1 + \lambda^{-1}),\ T^2 y_2 \in W^u_{loc}(T^2 y_1) \implies d(y_1, y_2) < \delta(\lambda^{-2} + \lambda^{-3}).$$

As distances along unstable manifolds contract in backward time, the orbit if y_2 inherits the shadowing from y_1 for the past orbit. Also

$$d(x_2, T^2 y_2) < \delta(1 + \lambda^{-1}),\ T^2 y_2 \in W^s_{loc}(x_2) \implies d(T^3 y_2, Tx_2) < \delta(\lambda^{-1} + \lambda^{-2}),$$

which ensures that the construction can be iterated by letting $T^3 y_3 = [T^3 y_2, x_3]$.

In the inductive step, we assume that y_{k-1} shadows the pseudo orbit up to time $k-1$, in such a way that

$$T^{k-1} y_{k-1} \in W^s_{loc}(x_{k-1}) \implies d(T^k y_{k-1}, x_k) \text{ controlled.}$$

Now let

$$T^k y_k = [T^k y_{k-1}, x_k].$$

As $T^k y_k$ and $T^k y_{k-1}$ are on the same unstable manifold, shadowing of the pseudo-orbit up to time $k-1$ is inherited by y_k. Also, it extends to time k, and as $T^k y_k$ and x_k are on the same stable manifold, the construction can be iterated. Distances are controlled as we deal with the consecutive terms of a geometric series.

To include bi-infinite pseudo obits, when constructing $T^k y_k$, instead of letting $y_0 := x_0$, rather have $T^{-k} y_{-k} := x_{-k}$ as the first step, and follow the above iterative procedure for $2k+1$ steps. $\square$

3.4.2 Consequences of shadowing

Definition 3.9. *An invertible topological dynamical system $T : M \to M$ is expansive if there exists $\varepsilon > 0$ such that if for $x, y \in M$ we have $d(T^j x, T^j y) < \varepsilon$ for every $j \in \mathbb{Z}$ then $x = y$.*

Comments:

- If $T : M \to M$ is not invertible, then the requirement for expansivity is that $d(T^j x, T^j y) < \varepsilon$ for every $j \geq 0$ implies $x = y$.
- It can be checked by direct inspection that the (full) shift is expansive.
- The main examples for expansive systems are hyperbolic systems. Indeed if $d(T^j x, T^j y) < \delta$ for every $j \in \mathbb{Z}$ then $y \in W^u_{loc}(x) \cap W^s_{loc}(x)$ which can happen only if $x = y$.
- Thus hyperbolic systems are expansive and have the shadowing property. This implies that the *point shadowing a bi-infinite pseudo-orbit is unique.* Indeed, it x and y shadows the same bi-infinite pseudo-orbit then $d(T^j x, T^j y) < 2\varepsilon$ for any $j \in \mathbb{Z}$ which implies, for ε sufficiently small, $x = y$ by expansivity.

Lemma 3.10. *In a hyperbolic dynamical system, periodic points are dense in $\Omega(T)$ (the set of non-wandering points).*

Proof. Choose ε so small that we have 2ε-expansivity. By the shadowing property, there exists $\delta > 0$ such that any bi-infinite δ-pseudo orbit is ε-shadowed by a unique true orbit. Let us fix $x \in \Omega(T)$, and let $U = B_{\delta/2}(x)$, the open ball of radius $\delta/2$ around x. As x is non-wandering, there exists $n \geq 1$ such that $T^n U \cap U \neq \emptyset$. Hence there exists some point $\widetilde{x} \in U$ such that $T^n \widetilde{x} \in U$ as well. In particular,

$$d(\widetilde{x}, x) < \delta, \qquad d(\widetilde{x}, T^n \widetilde{x}) < \delta. \tag{3.9}$$

Hence

$$\dots, T^{n-1}\widetilde{x}, \widetilde{x}, T\widetilde{x}, \dots, T^{n-1}\widetilde{x}, \widetilde{x}, \dots$$

is a bi-infinite δ-pseudo orbit. Hence there exists a *unique* $y \in M$ ε-shadowing it. It is claimed that y is periodic, more precisely $T^n y = y$. To see this, we argue by contradiction: assume $z = T^n y \neq y$. Then the (bi-)infinite orbit of z is just the T^n-shifted (bi-)infinite orbit of y, hence it is shadowing the pseudo-orbit (3.9). Hence, by uniqueness, $z = y$. So y is periodic and

$$d(x, y) \leq d(x, \widetilde{x}) + d(\widetilde{x}, y) < \delta + \varepsilon$$

which can be made arbitrary small by shrinking ε and δ (which may require, naturally, a larger period n). □

Finally, we mention another consequence of shadowing and expansivity: hyperbolic systems are C^1-*structurally stable*. Recall that structural stability of $T : M \to M$ means that there exists some $\varepsilon > 0$ such that any $\widehat{T} : M \to M$ that is ε-close to $T : M \to M$ in the C^1-metric is topologically conjugate to T.

The idea behind structural stability is that for any $y \in M$ we may regard the $\widehat{T}$-obit:

$$\ldots, \widehat{T}^{-1}y, y, \widehat{T}y, \ldots, \widehat{T}^k y, \ldots$$

as a pseudo-orbit for T. Hence, by the shadowing property, there exists a unique point $z \in M$ the T-orbit of which is shadowing this pseudo-orbit. We define $\Phi : M \to M$ by letting $\Phi(y) := z$. Then it can be shown that Φ topologically conjugates T with $\widehat{T}$.

4 Entropy

There are several constructions in dynamical systems called some type of entropy. For all of them, the aim is to measure the rate of the growth of complexity in some sense. To consider growth rates, the following lemma on numerical sequences is very useful.

Lemma 4.1 (Subadditive convergence lemma, or Fekete lemma)**.** *Let $a_n \in \mathbb{R}$ be a numerical sequence with the following subadditive property:*

$$a_{n+m} \le a_n + a_m, \quad \forall n, m \in \mathbb{N}. \tag{4.1}$$

Then the sequence $\dfrac{a_n}{n}$ converges, in fact

$$\lim_{n\to\infty} \frac{a_n}{n} = \inf_{n\ge 1} \left\{ \frac{a_n}{n} \right\}.$$

Proof. Let us introduce

$$\alpha = \inf_{n\ge 1} \left\{ \frac{a_n}{n} \right\}.$$

Then, apparently, $\liminf\limits_{n\to\infty} \dfrac{a_n}{n} \ge \alpha$. Hence, it is enough to prove that

$$\limsup_{n\to\infty} \frac{a_n}{n} \le \alpha + \varepsilon \quad \text{for any } \varepsilon > 0.$$

So let us fix $\varepsilon > 0$. By definition, there exists some $m \ge 1$ such that $\dfrac{a_m}{m} < \alpha + \varepsilon$. Now, by the subadditivity property (4.1), for any $n \ge 1$:

$$a_n = a_{\ell m + r} \le \ell a_m + a_r; \qquad \text{where } 0 \le r \le m-1; \ \frac{n-1}{m} \le \ell \le \frac{n}{m}.$$

Let $K = \max(|a_1|, \ldots, |a_{m-1}|)$, Now, dividing by n, and taking the $\limsup$:

$$\limsup_{n\to\infty} \frac{a_n}{n} \le a_m \limsup_{n\to\infty} \left(\frac{\ell}{n} \right) + \limsup_{n\to\infty} \frac{K}{n} = \frac{a_m}{m} < \alpha + \varepsilon.$$

$\square$

4.1 Topological entropy

Let M be a *compact* metric space, and let $\varepsilon > 0$.

- A *finite ε-cover* of M is a finite collection of open sets U_i, $i = 1, \dots, I$ such that $\bigcup_{i=1}^{I} U_i \supset M$ and $diam(U_i) < \varepsilon$ for every i. By compactness, there exist finite ε-covers for every $\varepsilon > 0$. So we may define

$$C(\varepsilon) = \text{The minimum cardinality of } \varepsilon\text{-covers.}$$

- A finite *ε-net* is a set $\{x_1, \dots, x_K\} \subset M$ such that, for any $y \in M$ there exists some $k \in \{1, \dots, K\}$ for which $d(y, x_k) < \varepsilon$. Again by compactness, for any $\varepsilon > 0$ there exist finite ε-nets. Accordingly, we may define

$$N(\varepsilon) = \text{The minimum cardinality of } \varepsilon\text{-nets.}$$

- $S \subset M$ is an *ε-separated set* if $\forall x, y \in S$ we have $d(x, y) \geq \varepsilon$. Again by compactness, for any $\varepsilon > 0$ an ε-separated set has to have finite cardinality. Accordingly, we may define

$$S(\varepsilon) = \text{The maximum cardinality of } \varepsilon\text{-separated sets.}$$

It will be a homework to prove that:

$$C(2\varepsilon) \leq N(\varepsilon) \leq S(\varepsilon) \leq C(\varepsilon) \tag{4.2}$$

for any $\varepsilon > 0$.

Now let us consider a topological dynamical system $T : M \to M$, that is, M is a compact metric space and T is continuous. For any $n \geq 1$, let us define the *Bowen metric* as

$$d_n(x, y) = \max_{i=0,1,\dots,n-1} d(T^i x, T^i y).$$

$d_n(x, y) < \varepsilon$ means that the orbits of x and y "stay together" (i.e. ε-close) for n iterations. With growing n, the $d_n - \varepsilon$ neighborhoods of a point x shrink. However, by continuity of T, for any $n \geq 1$ d_n is a metric that generates the same topology as the original metric d. Accordingly, as the space M is compact, the above defined quantities make sense for the metric d_n. We will denote them $C(n, \varepsilon, T)$, $N(n, \varepsilon, T)$ and $S(n, \varepsilon, T)$, respectively.

Claim 4.2. *For any $n, m \geq 1$, we have*

$$C(n + m, \varepsilon, T) \leq C(n, \varepsilon, T) \cdot C(m, \varepsilon, T).$$

This holds because if U_i ($i = 1, \dots, I$) is an (n, ε)-cover, and V_j ($j = 1, \dots, J$) is an (m, ε)-cover, then $U_i \cap T^{-n} V_j$ is an $(n + m, \varepsilon)$-cover of cardinality $I \cdot J$. Accordingly, the sequence $\log(C(n, \varepsilon, T))$ is subadditive. Hence the limit

$$h_\varepsilon(T) = \lim_{n \to \infty} \frac{\log(C(n, \varepsilon, T))}{n}$$

exists. Also, as $C(n, \varepsilon, T)$ is an integer, $h_\varepsilon(T) \geq 0$. Moreover, as for any $n \geq 1$ we have that $C(n, \varepsilon, T)$ increases as $\varepsilon \searrow 0$, so does $h_\varepsilon(T)$. Hence we may define

Definition 4.3. *The topological entropy of* $T : M \to M$ *is defined as*

$$(h_{top}(T) =)h(T) = \lim_{\varepsilon \to 0+} h_\varepsilon(T).$$

Comments:

- $h(T)$ may be 0, a finite positive number or $+\infty$. The base of the logarithm may effect the value of $h(T)$, but it does not effect which one of these three cases occurs. In the theory of entropy, typically logarithm of base 2 is used.
- By (4.2), $N(n, \varepsilon, T)$ or $S(n, \varepsilon, T)$ may be used instead of $C(n, \varepsilon, T)$ in the definition of $h(T)$. Also, instead of lim we may as well use lim inf or lim sup.

Finally, we compute the topological entropy for two examples.

Let $T : \mathbb{S}^1 \to \mathbb{S}^1$ be a *rotation*. For the space $\mathbb{S}^1$, we have

$$N(\varepsilon) \le \left\lceil \frac{2}{\varepsilon} \right\rceil + 1,$$

as simply putting points $\frac{\varepsilon}{2}$-apart on the circle, an ε-net is obtained. But this is at the same time an (n, ε)-net for any $n \ge 1$, because T is an isometry, hence $d(x, y) < \varepsilon$ implies $d(T^k x, T^k y) < \varepsilon$ for any $k \ge 1$. Hence

$$h_\varepsilon(T) = \lim_{n \to \infty} \frac{1}{n} \log(N(n, \varepsilon, T)) = 0$$

and thus $h(T) = 0$ for rotations.

Let $\sigma : \Sigma^+ \to \Sigma^+$ be the (one-sided) full shift with two symbols. Then,

$$2^{-m-1} < \varepsilon \le 2^{-m} \implies C(\varepsilon) = 2^m$$

as we need words – cylinder sets – of length m to cover the space. For an (n, ε) cover, we need words of length $(n + m)$, so

$$2^{-m-1} < \varepsilon \le 2^{-m} \implies C(n, \varepsilon, \sigma) = 2^{n+m}.$$

This implies

$$h_\varepsilon(\sigma) = \lim_{n \to \infty} \frac{1}{n} \log(C(n, \varepsilon, \sigma)) = \lim_{n \to \infty} \left(\frac{n+m}{n} \cdot \log 2 \right) = \log 2$$

independently of ε, hence the topological entropy of the full shift with two symbols is $\log 2$.

Similarly, the one-sided and the two-sided shifts of K symbols, $\sigma : \Sigma_K^+ \to \Sigma_K^+$ and $\sigma : \Sigma_K \to \Sigma_K$, both have topological entropy is $\log K$.

4.2 Entropy for finite partitions

Some terminology

Let $(M, \mathcal{B}, \mu)$ be a probability space. To avoid technical complications, throughout, *Lebesgue spaces* are considered. This means that the probability space is isomorphic to $[0, a]$ for some $a \in [0, 1]$ with Lebesgue measure and the union of countably many atoms. As any complete separable metric space with a completed Borel probability measure is a Lebesgue space, this is definitely general enough for the examples considered in these notes.

A *finite partition* is a finite collection of measurable sets $\alpha = \{A_1, \ldots, A_I\}$ that are pairwise disjoint and cover M, up to μ- measure 0. Finite partitions generate finite σ-algebras.

Given two partitions $\alpha = \{A_1, \ldots, A_I\}$ and $\beta = \{B_1, \ldots, B_J\}$ the following notations are introduced.

- The common refinement or *join* is

$$\alpha \vee \beta = \{A_i \cap B_j \,|\, i = 1, \ldots, I; j = 1, \ldots, J\}.$$

- β is a *refinement* of α, denoted $\alpha \leq \beta$, if for any j there exists i such that $B_j \subset A_i$. (equivalently, if $\alpha \vee \beta = \beta$).
- $\mathcal{N}$ denotes the trivial partition $\{M\}$.
- By including some zero measure sets, we may assume $I = J$. Accordingly, let $\mathcal{P}_I$ denote the collection of partitions of M with I elements. Then we may define the *distance of the two partitions*, a metric on $\mathcal{P}_I$ as

$$d(\alpha, \beta) = \min_{\sigma \in S_I} \sum_{i=1}^{I} \mu(A_i \Delta B_{\sigma(i)}) \tag{4.3}$$

  where S_I denotes all possible permutations of the indices $\{1, \ldots, I\}$, and Δ is the symmetric difference. We have

$$(\alpha \leq \beta \text{ and } \beta \leq \alpha) \iff d(\alpha, \beta) = 0 \iff \alpha = \beta \quad \text{mod } \mu \text{ measure } 0.$$

- $\alpha \perp \beta$ means that the two partitions are independent in the sense that $\mu(A_i \cap B_j) = \mu(A_i) \cdot \mu(B_j)$ for any pair i, j.

Entropy of a finite partition

The "information content" of an event A is defined as some $I(p)$ where $p = \mathbb{P}(A)$, and

(i) $I(1) = 0$, (ii) $I(p) \geq 0,\ \forall p \in (0, 1)$, (iii) $I(pq) = I(p) + I(q),\ \forall p, q \in (0, 1)$.

This fixes $I(p) = -\log p$. Note that to any partition with K elements, $\alpha = \{A_1, \dots, A_K\}$ we can associate a point in the K dimensional simplex Δ_K (namely $(\mu(A_1), \dots, \mu(A_K)) = (p_1, \dots, p_K)$). Letting

$$\Phi(x) = \begin{cases} 0 & \text{if } x = 0, \\ -x \log x & \text{if } x \in (0, 1]; \end{cases} \tag{4.4}$$

define $H : \Delta_K \to \mathbb{R}^+$ by

$$H(\beta) = H(\underline{p}) = -\sum_{i=1}^{K} p_i \log p_i = \sum_{i=1}^{K} \Phi(p_i),$$

the *entropy of the partition*. This is the "average information content" of a partition element as $H(\alpha) = \int_M I(x) d\mu(x)$.

The function (4.4) is strictly *concave down* on the interval $[0, 1]$, hence for every $\lambda \in (0, 1)$

$$\Phi(\lambda x + (1 - \lambda)y) \geq \lambda \Phi(x) + (1 - \lambda)\Phi(y)$$

with equality if an only if $x = y$. Also

$$\Phi\left(\sum_{i=1}^{K} \lambda_i x_i\right) \geq \sum_{i=1}^{K} \lambda_i \Phi(x_i) \qquad \left(\sum_{i=1}^{K} \lambda_i = 1\right), \tag{4.5}$$

again with equality iff all the x_i coincide. In particular, choosing $x_i = p_i$ and $\lambda_i = \frac{1}{K}$ $(i = 1, \dots, K)$, we have

$$H(\underline{p}) \leq \log K, \qquad \forall \underline{p} \in \Delta_K \tag{4.6}$$

with equality if and only if $p_i = \frac{1}{K}$ $(\forall i = 1, \dots, K)$. That is, the uniform distribution maximizes entropy.

Conditional entropy

Let α and β be two partitions. For j fixed, that is, $B_j \in \beta$ fixed, we may consider the entropy of the probability distribution obtained by conditioning the partition α on B_j:

$$-\sum_{i=1}^{I} \mu(A_i|B_j) \log(\mu(A_i|B_j)); \qquad \left(\mu(A_i|B_j) = \frac{\mu(A_i \cap B_j)}{\mu(B_j)}\right).$$

Averaging this on $B_j \in \beta$ the conditional entropy of α with respect to β is obtained:

$$H(\alpha|\beta) = -\sum_{j=1}^{J} \sum_{i=1}^{I} \mu(A_i \cap B_j) \log\left(\frac{\mu(A_i \cap B_j)}{\mu(B_j)}\right).$$

Some immediate properties:

- $H(\alpha|\mathcal{N}) = H(\alpha)$,
- $\alpha = \gamma \implies H(\alpha|\beta) = H(\gamma|\beta)$,
- $\alpha = \gamma \implies H(\beta|\alpha) = H(\beta|\gamma)$.

Properties of entropy and conditional entropy

Throughout α, β and γ denote finite partitions $\alpha = \{A_1, \ldots, A_I\}$, $\beta = \{B_1, \ldots, B_J\}$ and $\gamma = \{C_1, \ldots, C_K\}$.

(i) $H(\alpha \vee \gamma|\beta) = H(\alpha|\beta) + H(\gamma|\alpha \vee \beta)$.

Proof.

$$\begin{aligned} -H(\alpha \vee \gamma|\beta) &= \sum_{i,j,k} \mu(A_i \cap B_j \cap C_k) \log \frac{\mu(A_i \cap B_j \cap C_k)}{\mu(B_j)} = \\ &= \sum_{i,j,k} \mu(A_i \cap B_j \cap C_k) \log \frac{\mu(A_i \cap B_j \cap C_k)}{\mu(A_i \cap B_j)} + \sum_{i,j} \mu(A_i \cap B_j) \log \frac{\mu(A_i \cap B_j)}{\mu(B_j)} \\ &= -H(\gamma|\alpha \vee \beta) - H(\alpha|\beta). \end{aligned}$$

□

The proof of properties (ii), (iii) and (iv) will be part of your homework.

(ii) $H(\alpha \vee \gamma) = H(\alpha) + H(\gamma|\alpha)$.

(iii) If $\alpha \leq \gamma$, then $H(\alpha|\beta) \leq H(\gamma|\beta)$.

(iv) If $\alpha \leq \gamma$, then $H(\alpha) \leq H(\gamma)$.

(v) If $\alpha \leq \gamma$, then $H(\beta|\alpha) \geq H(\beta|\gamma)$.

Proof. Note that here we have the *opposite inequality* as compared to property (iii). Recall that the function Φ defined by (4.4) is strictly concave down. We have that the LHS is

$$H(\beta|\alpha) = \sum_{i,j} \mu(A_i)\Phi(\mu(B_j|A_i)).$$

Now, as $\alpha \leq \gamma$, for any pair i, k we have either $\mu(A_i \cap C_k) = 0$ or $\mu(A_i \cap C_k) = \mu(C_k)$. Hence, using the concavity of Φ with $\lambda_k = \frac{\mu(A_i \cap C_k)}{\mu(A_i)}$ and $x_k = \frac{\mu(B_j \cap C_k)}{\mu(C_k)}$

$$\begin{aligned} \Phi(\mu(B_j|A_i)) &= \Phi\left(\frac{\mu(B_j \cap A_i)}{\mu(A_i)}\right) = \\ &= \Phi\left(\sum_k \frac{\mu(A_i \cap C_k)}{\mu(C_k)} \cdot \frac{\mu(B_j \cap C_k)}{\mu(A_i)}\right) = \\ &= \Phi\left(\sum_k \frac{\mu(A_i \cap C_k)}{\mu(A_i)} \cdot \frac{\mu(B_j \cap C_k)}{\mu(C_k)}\right) \geq \\ &\geq \sum_k \frac{\mu(A_i \cap C_k)}{\mu(A_i)} \cdot \Phi\left(\frac{\mu(B_j \cap C_k)}{\mu(C_k)}\right). \end{aligned}$$

Now, if we multiply this relation by $\mu(A_i)$ and sum on i, j we get

$$\begin{aligned} H(\beta|\alpha) \;\geq\; & \sum_{i,j,k} \mu(A_i \cap C_k) \cdot \Phi\left(\frac{\mu(B_j \cap C_k)}{\mu(C_k)}\right) = \\ & \sum_{j,k} \mu(C_k)\Phi\left(\frac{\mu(B_j \cap C_k)}{\mu(C_k)}\right) = \\ & H(\beta|\gamma). \end{aligned}$$

□

The proof of properties (vi), (vii) and (viii) will be part of your homework.

(vi) $H(\alpha) \geq H(\alpha|\gamma)$.

(vii) $H(\alpha \vee \gamma|\beta) \leq H(\alpha|\beta) + H(\gamma|\beta)$.

(viii) $H(\alpha \vee \gamma) \leq H(\alpha) + H(\gamma)$.
Let $T : M \to M$ preserve the measure μ. Then, for any partition $\alpha = \{A_1, \ldots, A_I\}$, $T^{-1}\alpha = \{T^{-1}A_1, \ldots, T^{-1}A_I\}$ is another partition. Properties (ix) and (x) immediately follow from the invariance of the measure μ.

(ix) $H(T^{-1}\alpha|T^{-1}\beta) = H(\alpha|\beta)$.

(x) $H(T^{-1}\alpha) = H(\alpha)$.

(xi) $H(\alpha|\gamma) = 0 \iff H(\alpha \vee \gamma) = H(\gamma) \iff \alpha \leq \gamma$.

Proof. The first equivalence immediately follows from property (ii). For the second equivalence, $\Longleftarrow$ is immediate, too. For $\Longrightarrow$, note that $H(\alpha|\gamma) = 0$ means that

$$\mu(A_i \cap C_k) \log(\mu(A_i|C_k)) = 0$$

for any pair i, k. But this implies that, for any pair i, k, there are two possibilities: either $\mu(A_i \cap C_k) = 0$, or $\mu(A_i|C_k) = 1$, which means that $C_k \subset A_i$ (up to μ-measure 0). So for any k there exists i such that $C_k \subset A_i$, which precisely means that $\alpha \leq \gamma$. □

(xii) $H(\alpha|\gamma) = H(\alpha) \iff H(\alpha \vee \gamma) = H(\alpha) + H(\gamma) \iff \alpha \perp \gamma$.

Proof. The first equivalence immediately follows from property (ii). For the second equivalence, recall that

$$H(\alpha) = \sum_i \Phi(\mu(A_i)).$$

We have a similar expression for $H(\alpha|\gamma)$, in which we can use, for any fixed i, the concavity of Φ with $\lambda_k = \mu(C_k)$ and $x_k = \mu(A_i|C_k)$:

$$H(\alpha|\gamma) = \sum_i \left(\sum_k \mu(C_k)\Phi(\mu(A_i|C_k)) \right) \leq H(\alpha)$$

with equality if an only if for any i fixed, the value of $\mu(A_i|C_k) = \rho_i$ for some ρ_i independent of k. But the $\rho_i = \mu(A_i)$, hence $\mu(A_i \cap C_k) = \mu(A_i) \cdot \mu(C_k)$ for any pair i, k, which means precisely $\alpha \perp \gamma$. $\square$

Rokhlin metric on $\mathcal{P}_K$

Recall (4.3). Here we define another distance, the Rokhlin metric on $\mathcal{P}_K$ by

$$\rho(\alpha, \beta) = H(\alpha|\beta) + H(\beta|\alpha).$$

Symmetry and $\rho(\alpha, \beta) \geq 0$ is immediate. Also, by (xi) above, $\rho(\alpha, \beta) = 0$ implies $\alpha \leq \beta$ and $\beta \leq \alpha$ simultaneously, which means $\alpha = \beta$ (up to μ-measure 0). Now, applying (iii), (i) and then (v):

$$\begin{aligned} H(\alpha|\gamma) &\leq H(\alpha \vee \beta|\gamma) = H(\beta|\gamma) + H(\alpha|\beta \vee \gamma) \leq H(\beta|\gamma) + H(\alpha|\beta) \\ H(\gamma|\alpha) &\leq H(\gamma \vee \beta|\alpha) = H(\beta|\alpha) + H(\gamma|\beta \vee \alpha) \leq H(\beta|\alpha) + H(\gamma|\beta); \end{aligned}$$

and adding the two relations the triangular inequality is obtained for the Rokhlin metric.

Lemma 4.4. *The Rokhlin metric is uniformly continuous in the metric* (4.3). *That is, for any $\varepsilon > 0$ there exists $\delta(= \delta(\varepsilon, K))$ such that for any $\alpha, \beta \in \mathcal{P}_K$ such that $d(\alpha, \beta) < \delta$ we have $\rho(\alpha, \beta) < \varepsilon$.*

Proof. If $d(\alpha, \beta) < \delta$, we may assume that $\alpha = \{A_1, \ldots, A_K\}$ and $\beta = \{B_1, \ldots, B_K\}$ such that $\mu(A_i \Delta B_i) < \delta$, $i = 1, \ldots, K$. Now let us introduce another partition

$$\gamma = \{A_i \cap B_j | i \neq j\} \bigcup \{\cup_{k=1}^K (A_k \cap B_k)\}.$$

γ is a partition of $K^2 - K + 1$ elements, 1 large element and $K(K-1)$ tiny elements. Accordingly:

$$H(\gamma) \leq K(K-1)\Phi(\delta) + \Phi(1-\delta)$$

which can be made smaller than $\varepsilon/2$ by choosing δ small enough. Now

$$\alpha \vee \beta = \alpha \vee \gamma = \beta \vee \gamma.$$

This implies:

$$H(\alpha) + H(\alpha|\beta) = H(\alpha \vee \beta) = H(\alpha \vee \gamma) \leq H(\alpha) + H(\gamma) \leq H(\alpha) + \varepsilon/2$$

and thus $H(\alpha|\beta) \leq \varepsilon/2$. Similarly, $H(\beta|\alpha) \leq \varepsilon/2$ so $\rho(\alpha, \beta) \leq \varepsilon$. $\square$

Conditional entropy with respect to a sigma algebra

If $\beta = (B_1, \ldots, B_K)$ is a finite partition, then there is an associated finite σ-algebra, which, by slight abuse of notation, will be denoted by β as well. If A is a measurable set, then

$$\mathbb{E}(\chi_A|\beta)(x) = \sum_{i=1}^{K} \frac{\mu(A \cap B_k)}{\mu(B_k)} \cdot \chi_{B_k}(x)$$

for the conditional expectation of the indicator function. Then let

$$\Phi(\mathbb{E}(\chi_A|\beta)(x)) = -\mathbb{E}(\chi_A|\beta)(x) \cdot \log(\mathbb{E}(\chi_A|\beta)(x)) = \sum_{k=1}^{K} \Phi(\mu(A|B_k)) \cdot \chi_{B_k}(x)$$

and by integration

$$\int \Phi(\mathbb{E}(\chi_A|\beta)(x))d\mu(x) = \sum_{k=1}^{K} \mu(B_k)\Phi(\mu(A|B_k)).$$

Now if $\alpha = (A_1, \ldots, A_I)$ is a finite partition, than applying this formula for each of the A_i, and then summing on i gives:

$$H(\alpha|\beta) = \int \sum_{i=1}^{I} \Phi(\mathbb{E}(\chi_{A_i}|\beta)(x))d\mu(x).$$

This formula for the conditional entropy has the advantage that it can be generalized to the case when instead of β we condition on an arbitrary σ-algebra $\mathcal{F}$:

$$H(\alpha|\mathcal{F}) = \int \sum_{i=1}^{I} \Phi(\mathbb{E}(\chi_{A_i}|\mathcal{F})(x))d\mu(x).$$

Claim 4.5. *Let $\alpha = (A_1, \ldots, A_I)$ be a finite partition, and $\mathcal{F}_n$ a filtration (a refining sequence of σ-algebras) for example, corresponding to a refining sequence of finite partitions, and let*

$$\mathcal{F} = \bigvee_{n=1}^{\infty} \mathcal{F}_n$$

the sigma algebra generated by the sequence $\mathcal{F}_n$. Then

$$H(\alpha|\mathcal{F}_n) \to H(\alpha|\mathcal{F}) \qquad \text{as } n \to \infty.$$

Proof. For any measurable set A, by the martingale convergence theorem

$$\mathbb{E}(\chi_A|\mathcal{F}_n)(x) \to \mathbb{E}(\chi_A|\mathcal{F})(x)$$

almost surely, and thus

$$\Phi(\mathbb{E}(\chi_{A_i}|\mathcal{F}_n)(x)) \to \Phi(\mathbb{E}(\chi_{A_i}|\mathcal{F})(x))$$

almost surely, for $i = 1, \dots, I$. Then as

$$\max_{t\in[0,1]} \Phi(t) = \frac{1}{e} \implies \sum_{i=1}^{I} \Phi(\mathbb{E}(\chi_{A_i}|\mathcal{F}_n)(x)) \leq I \cdot \frac{1}{e}$$

where I is the cardinality of the fixed partition α, the claim follows by the dominated convergence theorem. □

4.3 Kolmogorov-Sinai entropy

Entropy of a dynamical system w.r.t. a partition

Let $T : M \to M$ preserve the probability measure μ. Recall that for a finite partition $\alpha = (A_1, \dots, A_I)$, $T^{-1}\alpha = (T^{-1}A_1, \dots, T^{-1}A_I)$ is another partition. Similarly we can define $T^{-n}\alpha$ for $n \geq 1$. Let, furthermore

$$\alpha_n = \alpha \vee T^{-1}\alpha \vee \dots T^{-n+1}\alpha = \bigvee_{i=0}^{n-1} T^{-i}\alpha$$

which is a refining sequence of partitions.

Definition 4.6. *The entropy of the measure preserving transformation $T : M \to M$ with respect to the partition α is defined by*

$$h(T, \alpha) = \lim_{n\to\infty} \frac{1}{n} H(\alpha_n)$$

It is still to be verified that the limit in Definition 4.6 exists. We give two arguments.

First argument: we show that $a_n = H(\alpha_n)$ is a subadditive sequence, and then Lemma 4.1 applies. To see this, we apply properties (viii) and (x) as follows:

$$a_{n+m} = H\left(\bigvee_{i=0}^{n+m-1} T^{-i}\alpha\right) \leq \underbrace{H\left(\bigvee_{i=0}^{n-1} T^{-i}\alpha\right)}_{a_n} + \underbrace{H\left(T^{-n}\left(\bigvee_{i=0}^{m-1} T^{-i}\alpha\right)\right)}_{a_m} =$$

Second argument: we will prove that the sequence $\frac{1}{n}\, a_n$ is nonincreasing. Then, as it is a positive sequence, the limit exists. The proof relies on the following Lemma, which is of independent interest.

Lemma 4.7.

$$H(\alpha_n) = H(\alpha) + \sum_{j=1}^{n-1} H\left(\alpha \;\middle|\; \bigvee_{i=1}^{j} T^{-i}\alpha\right).$$

Proof. This is proved by induction on n, the base case is immediate. Then, using properties (ii), (x) and the inductive assumption:

$$\begin{aligned}
\alpha_{n+1} &= H\left(\bigvee_{i=0}^{n} T^{-i}\alpha\right) = H\left(\left(\bigvee_{i=1}^{n} T^{-i}\alpha\right) \vee \alpha\right) = \\
&= H(\alpha_n) + H\left(\alpha \;\middle|\; \bigvee_{i=1}^{n} T^{-i}\alpha\right) = H(\alpha) + \sum_{j=1}^{n} H\left(\alpha \;\middle|\; \bigvee_{i=1}^{j} T^{-i}\alpha\right).
\end{aligned}$$

$\square$

A direct consequence of Lemma 4.7 and property (vi) is that

$$H(\alpha_n) \geq n \cdot H\left(\alpha \;\middle|\; \bigvee_{i=1}^{n} T^{-i}\alpha\right).$$

Now

$$\begin{aligned}
n \cdot a_{n+1} &= n \cdot H\left(\bigvee_{i=0}^{n} T^{-i}\alpha\right) = n \cdot \left(H(\alpha_n) + H\left(\alpha \;\middle|\; \bigvee_{i=1}^{n} T^{-i}\alpha\right)\right) \\
&\leq n \cdot H(\alpha_n) + H(\alpha_n) = (n+1) \cdot a_n
\end{aligned}$$

which readily implies $\frac{a_{n+1}}{n+1} \leq \frac{a_n}{n}$.

The Kolmogorov-Sinai entropy

Recall that $\mathcal{P}_K (= \mathcal{P}_K(M))$ denotes the collections of all K-element partitions of the probability space M. Let $(\mathcal{P}(M) =)\mathcal{P} = \bigcup_{k=1}^{\infty} \mathcal{P}_K$, the collection of all finite partitions.

Definition 4.8. *The Kolmogorov-Sinai entropy of the endomorphism (or automorphism) $(M, \mathcal{B}, T, \mu)$ is defined as*

$$h(T) = \sup_{\alpha \in \mathcal{P}} h(T, \alpha).$$

Lemma 4.9. *If $(\widehat{M}, \widehat{\mathcal{B}}, \widehat{T}, \widehat{\mu})$ is a factor of $(M, \mathcal{B}, T, \mu)$, then $h(\widehat{T}) \leq h(T)$.*

Proof. Recall that if $(\widehat{M}, \widehat{cB}, \widehat{T}, \widehat{\mu})$ is a factor of $(M, \mathcal{B}, T, \mu)$, this means that there exists some (not necessarily invertible) $\pi : M \to \widehat{M}$ such that $\pi_* \mu = \widehat{\mu}$ and $\pi \circ \widehat{T} = T \circ \pi$. Then, for any $\widehat{\alpha} \in \mathcal{P}(\widehat{M})$ let $\alpha = \pi^{-1}\widehat{\alpha}$, then

$$H(\alpha) = H(\widehat{\alpha}); \quad H(\alpha_n) = H(\widehat{\alpha}_n),\ \forall n \geq 1; \quad \Longrightarrow \quad h(T, \alpha) = h(\widehat{T}, \widehat{\alpha}).$$

Taking a supremum on $\widehat{\alpha} \in \mathcal{P}(\widehat{M})$ then implies

$$h(\widehat{T}) = \sup_{\widehat{\alpha} \in \mathcal{P}(\widehat{M})} h(\widehat{T}, \widehat{\alpha}) \leq \sup_{\alpha \in \mathcal{P}(M)} h(T, \alpha) = h(T).$$

$\square$

Corollary 4.10. *The Kolmogorov-Sinai entropy is an isomorphism invariant. That is, if* $(\widehat{M}, \widehat{\mathcal{B}}, \widehat{T}, \widehat{\mu})$ *and* $(M, \mathcal{B}, T, \mu)$ *are isomorphic, then* $h(\widehat{T}) = h(T)$.

A natural question is whether entropy is a *complete* isomorphism invariant, that is, does $h(\widehat{T}) = h(T)$ imply that $(\widehat{M}, \widehat{\mathcal{B}}, \widehat{T}, \widehat{\mu})$ and $(M, \mathcal{B}, T, \mu)$ are isomorphic? It is a famous theorem of Ornstein that the answer is yes in the category of (two-sided) Bernoulli-shifts.

Properties of $h(T, \alpha)$ and $h(T)$

The proof of properties (1), (2), (3) and (5), (6), (7) will be part of your homework.

(1) $h(T, \alpha) \leq H(\alpha)$.

(2) $h(T, \alpha \vee \beta) \leq h(T, \alpha) + h(T, \beta)$.

(3) If $\alpha \leq \beta$, then $h(T, \alpha) \leq h(T, \beta)$.

(4) $h(T, \beta) \leq h(T, \alpha) + H(\beta|\alpha)$.

Proof. Using properties (iv), (ii), (vii), (v) and (ix) we get:

$$\begin{aligned} H(\beta_n) &\leq H(\alpha_n \vee \beta_n) = H(\alpha_n) + H(\beta_n|\alpha_n) \leq H(\alpha_n) + \sum_{i=1}^{n} H(T^{-i}\beta|\alpha_n) \leq \\ &\leq H(\alpha_n) + \sum_{i=1}^{n} H(T^{-i}\beta|T^{-i}\alpha) = H(\alpha_n) + nH(\beta|\alpha). \end{aligned}$$

Then dividing by n and taking the limit results in (4). □

(5) $h(T, \alpha) = h(T, T^{-1}\alpha)$.

(6) $h(T, \alpha) = h(T, \bigvee_{i=0}^{k-1} T^{-i}\alpha)$, $\forall k \in \mathbb{Z}^+$.

(7) If T is invertible (that is, if $(M, \mathcal{B}, T, \mu)$ is an automorphism), then $h(T, \alpha) = h(T, \bigvee_{i=-k}^{k} T^{i}\alpha)$, $\forall k \in \mathbb{Z}^+$.

(8) $h(T^k) = kh(T)$, $\forall k \in \mathbb{Z}^+$.

Proof. For an arbitrary $\alpha \in \mathcal{P}$ and $k \geq 1$ fixed, we may apply property (6) to get

$$\begin{aligned} h(T^k, \alpha) &= h(T^k, \alpha_k) = \lim_{n\to\infty} \frac{1}{n} \cdot H\left(\bigvee_{j=0}^{n-1} T^{-j}(\alpha_k) \right) \\ &= k \cdot \lim_{n\to\infty} \frac{1}{n \cdot k} H(\alpha_{n\cdot k}) = k \cdot h(T, \alpha), \end{aligned}$$

and then take supremum on $\alpha \in \mathcal{P}$ to complete the argument. □

(9) If T is invertible (that is, if $(M, \mathcal{B}, T, \mu)$ is an automorphism), then $h(T^k) = |k| h(T)$, $\forall k \in \mathbb{Z}$.

Proof. In view of property (8), it is enough to prove $h(T^{-1}) = h(T)$, which follows as by (x) we have

$$H(\alpha \vee T\alpha \vee \cdots \vee T^{n-2}\alpha \vee T^{n-1}\alpha) = H(T^{-n+1}\alpha \vee T^{-n+2}\alpha \vee \cdots \vee T^{-1}\alpha \vee \alpha) = H(\alpha_n).$$

□

Some further useful properties:

- It follows directly from property (4) that

$$|h(T, \alpha) - h(T, \beta)| \leq \rho(\alpha, \beta). \tag{4.7}$$

- Using the notation of Claim 4.5, we have

$$h(T, \alpha) = \lim_{n \to \infty} H(\alpha | T^{-1}\alpha_n) = H(\alpha | T^{-1}\mathcal{F}) \tag{4.8}$$

  where

$$\mathcal{F} = \bigvee_{n=0}^{\infty} \alpha_n$$

  the sigma algebra generated by the refining sequence of finite partitions. Here property (v) ensures that the sequence is non-increasing, hence the limit exists, and, using Lemma 4.7:

$$\begin{aligned} h(T, \alpha) &= \lim_{n \to \infty} \frac{1}{n} H(\alpha_n) = \lim_{n \to \infty} \frac{1}{n} \left(H(\alpha) + \sum_{j=1}^{n-1} H(\alpha | T^{-1}\alpha_j) \right) = \\ &= \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n-1} H(\alpha | T^{-1}\alpha_j) = \lim_{n \to \infty} H(\alpha | T^{-1}\alpha_n) \end{aligned}$$

  as if the limit exists, it is equal to the Cesaro limit.

The Kolmogorov-Sinai theorem

Given an endomorphism $(M, \mathcal{B}, T, \mu)$ and a finite partition $\alpha = (A_1, \ldots, A_I)$, consider α_n, the a refining sequence of partitions, with associated finite sigma algebras $\mathcal{F}_n$. Let

$$\mathcal{A} = \bigcup_{n=1}^{\infty} \alpha_n \qquad \text{and} \qquad \mathcal{F} = \sigma(\mathcal{A}) = \bigvee_{n=1}^{\infty} \alpha_n = \bigvee_{n=1}^{\infty} \mathcal{F}_n$$

That is, $\mathcal{A}$ is the *algebra* (collection of subsets of M closed under *finite* intersections and complements) obtained as the union of all the finite σ algebras $\mathcal{F}_n$, while $\mathcal{F}$ is the generated *sigma algebra.*

Definition 4.11. *Given an endomorphism $(M, \mathcal{B}, T, \mu)$, the finite partition α is a (one-sided) generator if*

$$\mathcal{B} = \mathcal{F}\left(= \sigma(\mathcal{A}) = \bigvee_{n=1}^{\infty} \alpha_n = \bigvee_{n=1}^{\infty} \mathcal{F}_n\right).$$

Given an automorphism $(M, \mathcal{B}, T, \mu)$ (that is, $T : M \to M$ is invertible), the finite partition α is a two-sided generator if

$$\mathcal{B} = \bigvee_{n=1}^{\infty} \alpha_{-n}^{n} = \bigvee_{n=1}^{\infty}\left(\bigvee_{i=-n}^{n} T^i\alpha\right).$$

Proposition 4.12. *Let α be a generator. Then for any $\varepsilon > 0$ and any finite partition $\beta = (B_1, \ldots, B_K)$, there exists another finite partition $\gamma = (C_1, \ldots, C_K)$ with $C_i \in \mathcal{A}$ $(i = 1, \ldots, K)$ and $\rho(\beta, \gamma) < \varepsilon$.*

Proof. By Lemma 4.4, it is enough to ensure that $d(\beta, \gamma) < \delta$ for δ chosen appropriately. As $\sigma(\mathcal{A}) = \mathcal{F}$, there exist $D_1, \ldots, D_K \in \mathcal{A}$ such that $d(B_i \Delta D_i) < \frac{\delta}{K^4}$ for $i = 1, \ldots, K$, where, as usual, Δ denotes the symmetric difference. The issue is that the sets D_i are not necessarily disjoint, hence may not form a partition. However,

$$D_i \cap D_j \subset (B_i \Delta D_i) \cap (B_j \Delta D_j), \qquad \forall i \neq j.$$

So let

$$\begin{aligned} C_i &= D_i \setminus \left(\bigcup_{k\neq\ell}(D_k \cap D_\ell)\right), \qquad (i = 1, \ldots, K-1),\\ C_K &= D_K \bigcup \left(\bigcup_{k\neq\ell}(D_k \cap D_\ell)\right). \end{aligned}$$

Then $\gamma = (C_1, \ldots, C_K)$ is a partition such that $C_i \in \mathcal{A}$ for $i = 1, \ldots, K$, and

$$\begin{aligned} \mu(C_i \Delta B_i) &\leq K^2 \cdot 2\frac{\delta}{K^4} = \frac{2\delta}{K^2}, \qquad (i = 1, \ldots, K-1),\\ \mu(C_K \Delta B_K) &\leq K^2 \cdot 2\frac{\delta}{K^4} = \frac{2\delta}{K^2}, \end{aligned}$$

which implies $d(\beta, \gamma) < \delta$. □

Corollary 4.13. *If α is a generator, then, for any finite partition β, $H(\beta|\alpha_n) \to 0$ as $n \to \infty$. Similarly, if α is a two-sided generator for an automorphism, then $H(\beta|\alpha_{-n}^{n}) \to 0$.*

Proof. By Proposition 4.12, for any ε there exists some partition $\gamma = (C_1, \ldots, C_K) \in \mathcal{A}$ such that $H(\beta|\gamma) < \varepsilon$. As $C_i \in \mathcal{A}$ there exists some $n_i \geq 1$ such that $C_i \in \mathcal{F}_{n_i}$; $(i = 1, \ldots, K)$. Let $n_0 = \max(n_1, \ldots, n_K)$. Then $\gamma \leq \alpha_{n_0}$, and for any $n \geq n_0$ we have

$$H(\beta|\alpha_n) \leq H(\beta|\alpha_{n_0}) \leq H(\beta|\gamma) < \varepsilon.$$

□

Theorem 4.14 (Kolmogorov-Sinai)**.** *Let the partition α be a one-sided generator for an endomorphism $(M, \mathcal{B}, T, \mu)$. Then*

$$h(T) = h(T, \alpha).$$

Analogously, if α is a two-sided generator for an automorphism $(M, \mathcal{B}, T, \mu)$, then $h(T) = h(T, \alpha)$.

Proof. $h(T) \geq h(T, \alpha)$ follows from the definition. Hence it is enough to prove that for any $\varepsilon > 0$ and any finite partition β we have

$$h(T, \beta) \leq h(T, \alpha) + \varepsilon.$$

By Corollary 4.13, there exists $n_0 \geq 0$ such that $H(\beta|\alpha_{n_0}) < \varepsilon$. Using properties (4) and (6) we have:

$$h(T, \beta) \leq h(T, \alpha_{n_0}) + H(\beta|\alpha_{n_0}) = h(T, \alpha) + H(\beta|\alpha_{n_0}) \leq h(T, \alpha) + \varepsilon.$$

The proof for a two-sided generator of an automorphism is analogous. □

Corollary 4.15. *If an automorphism $(M, \mathcal{B}, T, \mu)$ has a one-sided generator α, then $h(T) = 0$.*

Proof. As $T : M \to M$ is invertible and T^{-1} is measurable, we have $T^{-1}\mathcal{B} = \mathcal{B}$. Let α denote the one-sided generator. Then

$$h(T) = h(T, \alpha) = \lim_{n \to \infty} H(\alpha|T^{-1}\alpha_n) = H(\alpha|T^{-1}\mathcal{B}) = H(\alpha|\mathcal{B}) = 0.$$

□

4.4 Examples

Rotations

Rational case. We have $h(Id) = 0$, as for the identity $\alpha_n = \alpha$ for any finite partition α and any $n \geq 0$. Now if T is a rational rotation, then there exists some $k \geq 1$ such that $T^k = Id$. Hence, using property (8),

$$h(T) = \frac{1}{k} \cdot h(T^k) = \frac{1}{k} \cdot h(Id) = 0.$$

Irrational case. Let $\alpha = \{[0, 1/2); [1/2, 1)\}$, a partition of $\mathbb{S}^1$. As the orbit of any point is dense, α is a one-sided generator. $h(T) = 0$ follows from Corollary 4.15.

Bernoulli shifts

Let $\sigma : \Sigma_K^+ \to \Sigma_K^+$ denote the one-sided full shift with K symbols. For $\underline{p} = (p_0, \ldots, p_{K-1}) \in \Delta_K$, let $\mu(= \mu_{\underline{p}})$ denote the associated Bernoulli measure on Σ_K^+, and let us consider the endomorphism $(\Sigma_K^+, \mathcal{B}, \sigma, \mu)$, the one-sided Bernoulli shift. Recall that the points in this space are sequences $\Sigma_K^+ \ni \underline{x} = (x_0, x_1, \ldots..)$, and let us consider the partition:

$$\alpha = (A_0, \ldots, A_{K-1}); \qquad A_j = \{\underline{x} \in \Sigma_K^+ \,|\, x_0 = j\};\ j = 0, \ldots, K-1. \tag{4.9}$$

Then α_n is exactly the partition into cylinder sets of length n, and thus α is a (one-sided) generator.

Also, for $n \geq 1$

$$\sigma^{-n}\alpha = (\sigma^{-n}A_0, \ldots, \sigma^{-n}A_{K-1}); \qquad \sigma^{-n}A_j = \{\underline{x} \in \Sigma_K^+ \,|\, x_n = j\};\ j = 0, \ldots, K-1,$$

and as the letters at the different positions are independent for a Bernoulli measure, we have, by property (xii),

$$\alpha_n \perp \sigma^{-n}\alpha, \forall n \geq 1, \implies H(\alpha_n) = nH(\alpha) = nH(\underline{p}),$$

and thus, for the metric entropy of the one-sided full shift

$$h_\mu(\sigma) = \lim_{n\to\infty} \frac{1}{n} H(\alpha_n) = H(\underline{p}) = -\sum_{i=0}^{K-1} p_i \log p_i.$$

The same Formula holds for the metric entropy of the automorphism $(\Sigma_K, \mathcal{B}, \sigma, \mu)$, the two-sided Bernoulli shift.

Comments:

- Recall that for the topological entropy we have already seen $h_{TOP}(\sigma) = \log K$, and thus by (4.6)
$$h_\mu(\sigma) = H(\underline{p}) \leq \log K = h_{TOP}(\sigma)$$
for any Bernoulli measure μ. This is a special case of the *variational principle*, to be discussed in further detail later.

- We have already mentioned Ornstein's famous theorem on the *isomorphism of two-sided Bernoulli shifts with the same metric entropy.* That is, if for the Bernoulli automorphisms $(\Sigma_{K_1}, \mathcal{B}, \sigma, \mu_{\underline{p}_1})$ and $(\Sigma_{K_2}, \mathcal{B}, \sigma, \mu_{\underline{p}_2})$ we have $h_{\mu_{\underline{p}_2}}(\sigma) = H(\underline{p}_2) = H(\underline{p}_1) = h_{\mu_{\underline{p}_2}}(\sigma)$, then the two automorphisms are isomorphic.

- For a partition $\beta = (B_0, \ldots, B_J)$, let us introduce the notation
$$\beta(x) = B_j \text{ and } I_\beta(x) = -\log(\mu(B_j)) \text{ if } x \in B_j$$

for the "information content of the randomly chosen point $x \in M$". Then β_n is a refining sequence of partitions, and thus $I_{\beta_n}(x)$ keeps growing with increasing n. Specifically, for a Benoulli shift and the generator α defined in (4.9), by independence:

$$I_{\alpha_n}(\underline{x}) = \sum_{i=0}^{n-1} I_\alpha(\sigma^i \underline{x})$$

and thus, by Birkhoff's ergodic theorem, for μ-a.e. $\underline{x} \in \Sigma_K$ (or Σ_K^+):

$$\lim_{n\to\infty} \frac{1}{n} I_{\alpha_n}(\underline{x}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} I_\alpha(\sigma^i \underline{x}) = \int I_\alpha(\underline{x}) d\mu(\underline{x}) = H(\underline{p}) = h_\mu(\sigma).$$

According to the *Shannon-McMillan-Breiman* theorem, this turns out to be true in a much wider generality: for any ergodic endomorphism $(M, \mathcal{B}, T, \nu)$ and any finite partition β we have

$$\lim_{n\to\infty} \frac{1}{n} I_{\beta_n}(x) = h(T, \beta) \qquad \text{for } \nu-\text{a.e. } x \in M.$$

Markov shifts

Let us compute first the *topological entropy* for topological Markov chains $\sigma : \Sigma_A^+ \to \Sigma_A^+$, where the adjacency matrix A_{ij} is primitive (irreducible and aperiodic). For $n \geq 1$, let $W(n, A)$ denote the set of admissible words of length n, that is:

$$W(n, A) = \{(a_0, \ldots, a_{n-1}) \in \{0, \ldots, K-1\}^n \,|\, A_{a_{j-1}, a_j} = 1; j = 1, \ldots, n-1\},$$

and let $\#W(n, A)$ denote the cardinality of this set. Then, along the argument for the topological entropy of the full shift at the end of subsection 4.1, we have

$$h_{TOP}(\sigma, \Sigma_A^+) = \lim_{n\to\infty} \frac{1}{n} \log\left(\#W(n, A)\right),$$

so it is $\#W(n, A)$ that we have to compute. Note that for any n, A^n is a matrix with integer entries.

Claim 4.16. *$(A^{n-1})_{ij}$, the ij entry of the matrix A^{n-1} is the number of admissible words of length n that start with symbol i and end with symbol j.*

Proof. This can be proved by induction on n. The base case $n = 2$ follows from the definition of A. Then the inductive step follows as

$$(A^n)_{ij} = \sum_{k=0}^{K-1} (A^{n-1})_{ik} A_{kj}$$

and the words of length $n + 1$ from symbol i to symbol j arise precisely as words of length n from symbol i to some symbol k such that the transition from symbol k to symbol j is allowed. □

As a corollary

$$\#W(n,A) = \sum_{i=0}^{K-1}\sum_{j=0}^{K-1}(A^n)_{ij} = \|A^{n-1}\|_1$$

where $\|B\|_1$ denotes the L^1 norm of the matrix B. On a finite dimensional vector space all norms are equivalent, that is, there exists a $C > 0$ such that

$$C^{-1}\|B\| \le \|B\|_1 \le C\|B\|$$

for any $K \times K$ matrix B, where $\|B\|$ denotes the usual Euclidean norm. Hence

$$h_{TOP}(\sigma, \Sigma_A^+) = \lim_{n\to\infty}\frac{1}{n}\log\|A^n\| = \log\left(\lim_{n\to\infty}\sqrt[n]{\|A^n\|}\right) = \log\lambda$$

where λ is the spectral radius of the matrix A. In fact, $\lambda > 1$ is the largest eigenvalue of A, which is simple by the Perron-Frobenius theorem.

To compute the *metric entropy*, let π_{ij} be a transition matrix that corresponds to the adjacency matrix A_{ij}. By assumption, π is primitive, so there exists a unique stationary distribution p_i. Let μ denote the corresponding Markov measure on Σ_A^+; our aim is to compute the metric entropy for the Markov shift $(\Sigma_A^+, \mathcal{B}, \sigma, \mu)$.

(4.9) remains a generator (to be denoted as $\beta = (B_0, \dots, B_{K-1})$ now). We may use Formula (4.8):

$$h_\mu(\sigma) = h_\mu(\sigma, \beta) = \lim_{n\to\infty} H(\beta|\sigma^{-1}\beta_n)$$

Now elements of β_n are cylinder sets of the form

$$C = \{\underline{x} \in \Sigma_A^+ | x_0 \dots x_{n-1} = j_0 \dots j_{n-1}\} \text{ for some } j_0 \dots j_{n-1} \in W(n,A)$$

while elements of $\sigma^{-1}\beta_n$ are of the form

$$\sigma^{-1}C = \{\underline{x} \in \Sigma_A^+ | x_1 \dots x_n = j_1 \dots j_n\} \text{ for some } j_1 \dots j_n \in W(n,A).$$

Then

$$\begin{aligned} H(\beta|\sigma^{-1}\beta_n) &= -\sum_{B\in\beta, C\in\beta_n} \mu(B\cap\sigma^{-1}C)\log\frac{\mu(B\cap\sigma^{-1}C)}{\mu(C)} = \\ &= -\sum_{j_0,j_1,\dots,j_n} p_{j_0}\prod_{k=0}^{n-1}(\pi_{j_kj_{k+1}})\log\frac{p_{j_0}\prod_{k=0}^{n-1}(\pi_{j_kj_{k+1}})}{p_{j_1}\prod_{k=1}^{n-1}(\pi_{j_kj_{k+1}})} = \\ &= -\sum_{j_0,j_1,\dots,j_n} p_{j_0}\prod_{k=0}^{n-1}(\pi_{j_kj_{k+1}})\log\frac{p_{j_0}\pi_{j_0j_1}}{p_{j_1}} \end{aligned}$$

Now as π_{ij} is a stochastic matrix, we have $\sum\limits_{j_{k+1}=1}^{K-1}\pi_{j_kj_{k+1}} = 1$ for $k = 1, \dots, n-1$, so

$$-H(\beta|\sigma^{-1}\beta_n) = \sum_{j_0,j_1} p_{j_0}\pi_{j_0j_1}(\log p_{j_0} + \log\pi_{j_0j_1} - \log p_{j_1})$$

which is constant in n. Moreover, using $\sum_{j_1} \pi_{j_0 j_1} = 1$ in the first and $\sum_{j_0} p_{j_0}\pi_{j_0 j_1} = p_{j_1}$ in the third term, we see that these two cancel and

$$h_\mu(\sigma) = -\sum_{j_0,j_1} p_{j_0}\pi_{j_0 j_1} \log \pi_{j_0 j_1}.$$

Parry measure. As in the context of Bernoulli shifts, there are many possible ways to assign an invariant (Markov) measure μ to the same topological system (subshift), that is, a transition matrix π_{ij} to an adjacency matrix A_{ij}. Here is a construction for which $h_\mu(\sigma) = h_{TOP}(\sigma, \Sigma_A^+)$. As it is assumed that A_{ij} is primitive, by the Perron-Frobenius theorem the maximal eigenvalue $\lambda > 1$ is simple. Let s_k and u_k denote the associated left and right eigenvectors, respectively, normalized so that $\sum_{k=0}^{K-1} s_k u_k = 1$. It will be a homework to verify that $\pi_{kl} = \lambda^{-1} u_k^{-1} A_{kl} u_l$ is the transition matrix of an irreducible aperiodic Markov chain, and that the corresponding stationary distribution is $p_k = s_k u_k$. Also, to check that for the associated Markov shift the metric entropy is equal to the topological entropy ($h_\mu(\sigma) = \log\lambda$). Hence, this Parry measure is a measure of maximal entropy.

5 Thermodynamic formalism

5.1 Topological pressure

We continue working with $\sigma : \Sigma_A^+ \to \Sigma_A^+$ (and occasionally with the two-sided $\sigma : \Sigma_A \to \Sigma_A$) for some primitive adjacency matrix A. Some terminology:

- Points $\Sigma_A^+ \ni \underline{x} = (x_0 x_1 \dots)$ are (semi-)infinite sequences such that $A_{x_i x_{i+1}} = 1$, $\forall i \geq 0$.
- For $m \geq 1$, $W(m, A)$ is the set of words of length m, i.e. $W(m, A) \ni \underline{a} = (a_0, \dots, a_{m-1})$ with $A_{a_i a_{i+1}} = 1$, $\forall i = 0, \dots, m-2$.
- For $\underline{a} \in W(m, A)$, let

  $$C(\underline{a}) = \{\underline{x} \in \Sigma_A^+ \,|\, x_0, \dots, x_{m-1} = a_0, \dots, a_{m-1}\},$$

  the associated cylinder set.
- $\Phi : \Sigma_A^+ \to \mathbb{R}$ is a *Hölder continuous potential* if there exist $C > 0$ and $\alpha \in (0, 1|$ such that

  $$|\Phi(\underline{x}) - \Phi(\underline{y})| \leq C d(\underline{x}, \underline{y})^\alpha = C K^{-\alpha s(\underline{x},\underline{y})}$$

  or equivalently, there exists $\beta < 1$ such that

  $$\forall m \geq 1; \forall \underline{a} \in W(m, A); \forall \underline{x}, \underline{y} \in C(\underline{a}) : \quad |\Phi(\underline{x}) - \Phi(\underline{y})| \leq C\beta^m.$$

  $\mathcal{C}(\Sigma_A^+)$ will denote the space of continuous functions $f : \Sigma_A^+ \to \mathbb{R}$, while $\mathcal{H}(\Sigma_A^+)$ will denote the space of Hölder continuous functions (potentials) $\Phi : \Sigma_A^+ \to \mathbb{R}$.

- For $\Phi \in \mathcal{H}(\Sigma_A^+)$ and $n \geq 1$, let

$$S_n\Phi(\underline{x}) = \sum_{k=0}^{n-1} \Phi(\sigma^k \underline{x})$$

that is, the Birkhoff sum of Φ at $\underline{x} \in \Sigma_A^+$, up to time n. Let, furthermore, for $m \geq 1$ and $\underline{a} \in W(m, A)$:

$$\mathcal{S}_m(\Phi, \underline{a}) = \sup_{\underline{x} \in C(\underline{a})} S_m\Phi(\underline{x})$$

and finally

$$Z_m\Phi = \sum_{\underline{a} \in W(m,A)} \exp\left(\mathcal{S}_m(\Phi, \underline{a})\right).$$

In particular, for $\Phi \equiv 0$, $Z_m\Phi = \#W(m, A)$.

Lemma 5.1. *For and $\Phi \in \mathcal{H}(\Sigma_A^+)$ and $m, n \geq 1$, we have $Z_{m+n}\Phi \leq Z_m\Phi \cdot Z_n\Phi$.*

Proof. Fix some $m \geq 1$ and $n \geq 1$. For $W(m + n, A) \ni \underline{a} = (a_0 \ldots a_{m+n-1})$, let

$$W(m, A) \ni \underline{a}^{(1)} = (a_0 \ldots a_{m-1}); \qquad W(n, A) \ni \underline{a}^{(2)} = (a_m \ldots a_{m+n-1}).$$

For $\underline{x} \in C(\underline{a})$, we have $\underline{x} \in C(\underline{a}^{(1)})$ and $\sigma^m \underline{x} \in C(\underline{a}^{(2)})$. Also,

$$S_{m+n}(\Phi)(\underline{x}) = S_m(\Phi)(\underline{x}) + S_n(\Phi)(\sigma^m \underline{x}).$$

Taking a supremum on $\underline{x} \in C(\underline{a})$ we have

$$\mathcal{S}_{m+n}(\Phi, \underline{a}) \leq \mathcal{S}_m(\Phi, \underline{a}^{(1)}) + \mathcal{S}_n(\Phi, \underline{a}^{(2)}),$$

where we have $\leq$ as the supremum is subject to more restrictions in the LHS than in the two terms of the RHS. Taking the exponential and then summing on $\underline{a} \in W(m + n, A)$ results in

$$\begin{aligned} Z_{m+n}\Phi &= \sum_{\underline{a} \in W(m+n,A)} \exp\left(\mathcal{S}_{m+n}(\Phi, \underline{a})\right) \leq \sum_{\underline{a} \in W(m+n,A)} \exp\left(\mathcal{S}_m(\Phi, \underline{a}^{(1)})\right) \cdot \exp\left(\mathcal{S}_n(\Phi, \underline{a}^{(2)})\right) \\ &\leq \left(\sum_{\underline{a}_1 \in W(m,A)} \exp\left(\mathcal{S}_m(\Phi, \underline{a}_1)\right)\right) \cdot \left(\sum_{\underline{a}_2 \in W(n,A)} \exp\left(\mathcal{S}_n(\Phi, \underline{a}_2)\right)\right) = Z_m\Phi \cdot Z_n\Phi, \end{aligned}$$

where again there is an inequality as – given the condition $A_{a_{m-1}a_m} = 1$ – it may happen that not all $\underline{a}_1 \in W(m, A)$ and $\underline{a}_2 \in W(n, A)$ arises as $\underline{a}^{(1)}$ and $\underline{a}^{(2)}$ for some $\underline{a} \in W(m+n, A)$. $\square$

Definition 5.2. *For $\Phi \in \mathcal{H}(\Sigma_A^+)$, let*

$$P(\Phi) = \lim_{m \to \infty} \frac{1}{m} \log Z_m\Phi,$$

the topological pressure associated to the Hölder continuous potential Φ. The limit exists by Lemma 5.1.

Comment: For $\Phi \equiv 0$, $P(\Phi)$ is the topological entropy.

5.2 Variational principle

Recall that $\mathcal{M}_{\text{inv}}(\Sigma_A^+)$ denotes the space of shift invariant Borel probability measures. Let $h_\mu(\sigma)$ denote the Kolmogorov-Sinai entropy of $(\Sigma_A^+, \mathcal{B}, \sigma, \mu)$ for $\mu \in \mathcal{M}_{\text{inv}}(\Sigma_A^+)$.

Lemma 5.3 (Variational principle). *Let $\mu \in \mathcal{M}_{\text{inv}}(\Sigma_A^+)$ be an arbitrary invariant measure, and $\Phi \in \mathcal{H}(\Sigma_A^+)$ an arbitrary Hölder continuous potential. We have*

$$h_\mu(\sigma) + \int \Phi d\mu \leq P(\Phi) \tag{5.1}$$

Proof. Lemma 5.3 relies on the following calculus lemma, the proof of which is part of your homework.

Lemma 5.4. *Let us fix the parameters $d_1, ..., d_r \in \mathbb{R}$, and introduce the notation $Z = \sum_{i=1}^{r} e^{d_i}$.*

1. *Consider the simplex*

$$\Delta = \{ \ \underline{p} = (p_1, ..., p_r) \in \mathbb{R}^r \mid p_i \geq 0, \ \sum_{i=1}^{r} p_i = 1 \ \},$$

and the function $F : \Delta \to \mathbb{R}$, $F(\underline{p}) = -\sum_{i=1}^{r} p_i \log p_i + \sum_{i=1}^{r} d_i \cdot p_i$. Show that the maximum of $F(\underline{p})$ on Δ is $\log Z$, attained at the unique point $p_j = \frac{e^{d_j}}{Z}$, $j = 1, ..., r$.

2. *Let us introduce furthermore*

$$\Delta_s = \{ \ \underline{p} = (p_1, ..., p_r) \in \mathbb{R}^r \mid p_i \geq 0, \ \sum_{i=1}^{r} p_i = s \ \},$$

for $0 < s \leq 1$. Show that the maximum of F on Δ_s is $s(\log Z - \log s)$, taken at the point $p_j = \frac{se^{d_j}}{Z}$, $j = 1, ..., r$.

By the invariance of μ:

$$\frac{1}{m} \int S_m \Phi d\mu = \int \Phi d\mu$$

Using at the consecutive inequalities that the partition into letters is a generator; Lemma 5.4 and the definition of Z_m, respectively:

$$\begin{aligned}
h_\mu(\sigma) + \int \Phi d\mu &= \lim_{m\to\infty} \frac{1}{m} \left(- \sum_{\underline{a} \in W(m,A)} \mu(C(\underline{a})) \log \mu(C(\underline{a})) + \int S_m \Phi d\mu \right) \leq \\
&\leq \lim_{m\to\infty} \frac{1}{m} \left(- \sum_{\underline{a} \in W(m,A)} \mu(C(\underline{a})) \log \mu(C(\underline{a})) \ + \sum_{\underline{a} \in W(m,A)} \mathcal{S}_m(\Phi, \underline{a}) \mu(C(\underline{a})) \right) \leq \\
&\leq \lim_{m\to\infty} \frac{1}{m} \log Z_m \Phi = P(\Phi).
\end{aligned}$$

□

Definition 5.5. *$\mu \in \mathcal{M}_{\text{inv}}(\Sigma_A^+)$ is an equilibrium measure for the potential $\Phi \in \mathcal{H}(\Sigma_A^+)$ if the supremum of the LHS is attained in Lemma 5.3, that is,* (5.1) *holds for μ with an equality.*

Comment: For $\Phi \equiv 0$, an equilibrium measure is a measure of maximal entropy.

Definition 5.6. *$\mu_\Phi \in \mathcal{M}_{\text{inv}}(\Sigma_A^+)$ is a Gibbs measure for the potential $\Phi \in \mathcal{H}(\Sigma_A^+)$ if there exists $P_\mu \in \mathbb{R}$ and $0 < c_1 < c_2$ such that, for any $m \geq 1$ an $\underline{a} \in W(m, A)$ we have:*

$$c_1 \leq \frac{\mu_\Phi(C(\underline{a}))}{\exp(-P_\mu m + S_m\Phi(\underline{x}))} \leq c_2 \qquad \forall \underline{x} \in C(\underline{a}).$$

Proposition 5.7. *If μ_Φ is a Gibbs measure, then it is an equilibrium measure for $\Phi \in \mathcal{H}(\Sigma_A^+)$.*

Proof. Step 1: $P_\mu = P(\Phi)$. As Φ and σ are continuous, $S_m\Phi$ is continuous, too. As $C(\underline{a})$ is compact, there exists $\widehat{\underline{x}} \in C(\underline{a})$ such that $S_m\Phi(\widehat{\underline{x}}) = \mathcal{S}_m(\Phi, \underline{a})$. Hence

$$c_1 \leq \frac{\mu_\Phi(C(\underline{a}))}{\exp(-P_\mu m + \mathcal{S}_m(\Phi, \underline{a}))} \leq c_2$$

for every $m \geq 1$ and every $\underline{a} \in W(m, A)$. Summation on $\underline{a} \in W(m, A)$ gives

$$c_1 \leq \frac{\exp(P_\mu m)}{Z_m(\Phi)} \leq c_2.$$

Taking the logarithm, division by m and then taking the limit as $m \to \infty$ gives

$$0 = \lim_{m\to\infty} \frac{\log c_1}{m} \leq P_\mu - \lim_{m\to\infty} \frac{\log Z_m(\Phi)}{m} \leq \lim_{m\to\infty} \frac{\log c_2}{m} = 0,$$

hence $P_\mu = P(\Phi)$.

Step 2: Hölder properties of $S_m(\Phi)$. For $(a_0, \ldots, a_{m-1}) = \underline{a} \in W(m, A)$ and $k = 0, \ldots, m-1$, let $\underline{a}_k = (a_k, \ldots, a_{m-1}) \in W(m-k, A)$. Then for any $\underline{x}, \underline{y} \in C(\underline{a})$ we have $\sigma^k\underline{x}, \sigma^k\underline{y} \in C(\underline{a}_k)$, hence, by Hölder continuity of Φ

$$|\Phi(\sigma^k\underline{x}) - \Phi(\sigma^k\underline{y})| \leq C\beta^{m-k}.$$

This implies

$$|S_m\Phi(\underline{x}) - S_m\Phi(\underline{y})| \leq C(\beta^m + \beta^{m-1} + \cdots + \beta + 1) \leq C \cdot \frac{1}{1-\beta} = D, \tag{5.2}$$

where the constant $D > 0$ is uniform in m.

Step 3. By the variational principle $h_\mu(\sigma) + \int \Phi d\mu \leq P(\Phi)$. To prove $h_\mu(\sigma) + \int \Phi d\mu \geq P(\Phi)$, recall that the partition into letters α is a generator, and that for $m \geq 1$, α_m is the partition into cylinder sets $C(\underline{a})$ for $\underline{a} \in W(m, A)$. For any $\underline{a} \in W(m, A)$, using Step 2, and then the Gibbs property:

$$\begin{aligned}
-\mu(C(\underline{a})) \log \mu(C(\underline{a})) + \int\limits_{C(\underline{a})} (S_m \Phi) d\mu &\geq -\mu(C(\underline{a})) \log \mu(C(\underline{a})) \ + \mu(C(\underline{a}))(\mathcal{S}_m(\Phi, \underline{a}) - D) \geq \\
&\geq \mu(C(\underline{a})) \left(-(-P(\Phi)m + \mathcal{S}_m(\Phi, \underline{a}))\right) - \log(c_2) + \mathcal{S}_m(\Phi, \underline{a}) - D) \geq \\
&\geq \mu(C(\underline{a})) \left(P(\Phi)m - \widetilde{D}\right)
\end{aligned}$$

for some $\widetilde{D} > 0$. Now summation on $\underline{a} \in W(m, A)$ gives:

$$H(\alpha_m) + m \int\limits_{\Sigma_A^+} \Phi d\mu = H(\alpha_m) + \int\limits_{\Sigma_A^+} S_m \Phi d\mu \geq P(\Phi)m - \widetilde{D}$$

then division by m and then taking $\lim\limits_{m\to\infty}$ gives

$$h_\mu(\sigma) + \int \Phi d\mu \geq P(\Phi)$$

which completes the proof of the Proposition. $\square$

5.3 Homologous potentials

In this section two-sided (invertible) shifts $\sigma : \Sigma_A \to \Sigma_A$ are considered. The aim is to reduce the two-sided case to the one-sided case.

Definition 5.8. *Two Hölder continuous potentials $\varphi, \psi \in \mathcal{H}(\Sigma_A)$ are homologous – notation: $\psi \sim \varphi$ – if there exists $u \in \mathcal{H}(\Sigma_A)$ such that*

$$\psi = \varphi + u \circ \sigma - u, \quad \textit{that is,} \quad \psi(\underline{x}) = \varphi(\underline{x}) + u(\sigma \underline{x}) - u(\underline{x});\ \forall \underline{x} \in \Sigma_A.$$

Lemma 5.9. *If $\varphi \sim \psi$, then $P(\varphi) = P(\psi)$ and $\mu_\psi = \mu_\varphi$ – more precisely, if $\mu_\varphi \in \mathcal{M}_{\text{inv}}(\Sigma_A)$ is a Gibbs masure for φ, then it is a Gibbs measure for ψ, too.*

Proof. We have $\psi = \varphi + u \circ \sigma - u$. Let $\|u\|$ denote the supremum norm of u, which is finite as $u \in \mathcal{H}(\Sigma_A)$, hence continuous. We have, for any $m \geq 1$ and any $\underline{x} \in \Sigma_A$:

$$\begin{aligned}
|S_m \psi(\underline{x}) - S_m \varphi(\underline{x})| &= \left| \sum_{k=0}^{m-1} (\psi(\sigma^k \underline{x}) - \varphi(\sigma^k \underline{x})) \right| \leq \left| \sum_{k=0}^{m-1} (u(\sigma^{k+1} \underline{x}) - u(\sigma^k \underline{x})) \right| \leq \\
&\leq |u(\sigma^{m+1} \underline{x}) - u(\underline{x})| \leq 2\|u\|
\end{aligned}$$

independently of $m \geq 1$. This implies that there exist $d_2 > d_1 > 0$ such that

$$d_1 < \frac{\exp(S_m \varphi(\underline{x}))}{\exp(S_m \psi(\underline{x}))} < d_2 \quad \Longrightarrow \quad d_1 < \frac{Z_m \varphi}{Z_m \psi} < d_2.$$

The usual procedure of taking log, dividing by m and $\lim_{m\to\infty}$ then shows $P(\varphi) = P(\psi)$. A further consequence is that, if $\mu = \mu_\psi$ is a Gibbs measure for ψ with $c_2 > c_1 > 0$, then

$$c_1 \cdot d_1 < \frac{\mu(C(\underline{a}))}{\exp(-P \cdot m + S_m\varphi(\underline{x}))} < c_2 \cdot d_2; \qquad \forall m \geq 1,\ \forall \underline{a} \in W(m, A),\ \forall \underline{x} \in C(\underline{a})$$

where $P = P(\varphi) = P(\psi)$. This means that μ is a Gibbs measure for $\varphi \in \mathcal{H}(\Sigma_A)$, too. $\square$

Definition 5.10. *$\psi \in \mathcal{H}(\Sigma_A)$ depends only on the future – notation: $\psi \in \widetilde{\mathcal{H}}(\Sigma_A)$ – if $\psi(\underline{x}) = \psi(\underline{y})$ whenever $\underline{x} = (\dots x_{-1}x_0x_1\dots)$ and $\underline{y} = (\dots y_{-1}y_0y_1\dots)$ are such that $x_k = y_k$ for $k \geq 0$.*

Lemma 5.11. *For any $\varphi \in \mathcal{H}(\Sigma_A)$ there exists $\psi \in \widetilde{\mathcal{H}}(\Sigma_A)$ such that $\psi \sim \varphi$.*

Proof. For any $t \in \{0, \dots, K-1\}$ there exists

$$(\dots z^t_{-1}z^t_0z^t_1\dots) = \underline{z}^t \in \Sigma_A \quad \text{such that} \quad z^t_0 = t.$$

There is at least one such possibility; if there are several, let us fix one of them. Now let us define $r : \Sigma_A \to \Sigma_A$ by

$$r(\underline{x}) = r(\dots x_{-2}x_{-1}x_0x_1x_2\dots) = (\dots z^{x_0}_{-2}z^{x_0}_{-1}x_0x_1x_2\dots),$$

that is,

$$(r(\underline{x}))_k = \begin{cases} x_k & \text{if } k \geq 0; \\ z^{x_0}_k & \text{if } k < 0. \end{cases}$$

By Hölder continuity of φ

$$|\varphi(\sigma^k\underline{x}) - \varphi(\sigma^k r(\underline{x}))| \leq C\beta^k \tag{5.3}$$

as $(\sigma^k\underline{x})_\ell = (\sigma^k r(\underline{x}))_\ell$ for $\ell \geq -k$. ($\beta < 1$ will be referred to as the Hölder exponent of φ.) Now let

$$u(\underline{x}) = \sum_{k=0}^{\infty}(\varphi(\sigma^k\underline{x}) - \varphi(\sigma^k r(\underline{x}))).$$

Claim 5.12. *The function $u : \Sigma_A \to \mathbb{R}$ defined above is Hölder continuous with exponent $\sqrt{\beta}(< 1)$.*

Proof. of the Claim: let $\underline{a} \in W(m, A)$ and $\underline{x}, \underline{y} \in C(\underline{a})$, then

$$\begin{aligned} |u(\underline{x}) - u(\underline{x})| &= \left|\sum_{j=0}^{\infty}(\varphi(\sigma^j\underline{x}) - \varphi(\sigma^j r(\underline{x})) - \varphi(\sigma^j\underline{y}) + \varphi(\sigma^j r(\underline{y})))\right| \\ &\leq \sum_{j=0}^{\lceil \frac{m}{2} \rceil}\left(|\varphi(\sigma^j\underline{x}) - \varphi(\sigma^j\underline{y})| + |\varphi(\sigma^j r(\underline{x})) - \varphi(\sigma^j r(\underline{y}))|\right) + \\ &\quad + \sum_{j=\lceil \frac{m}{2} \rceil}^{\infty}\left(|\varphi(\sigma^j\underline{x}) - \varphi(\sigma^j r(\underline{x}))| + |\varphi(\sigma^j\underline{y}) - \varphi(\sigma^j r(\underline{y}))|\right) \\ &= \quad I \quad + \quad II. \end{aligned}$$

To bound I, we use that $s(\underline{x},\underline{y}) \geq m$ and thus $s(r(\underline{x}),r(\underline{y})) \geq m$, while the shift σ can decrease separation at most by one, so $s(\sigma^j\underline{x},\sigma^j\underline{y}) \geq m-j$ and $s(\sigma^j r(\underline{x}),\sigma^j r(\underline{y})) \geq m-j$. Hence:

$$I \;\leq\; \sum_{j=0}^{\lceil \frac{m}{2} \rceil} 2C\beta^{m-j} \leq \frac{2C\beta^{\lceil \frac{m}{2} \rceil}}{1-\beta}.$$

To bound II we use (5.3):

$$II \;\leq\; \sum_{j=\lceil \frac{m}{2} \rceil}^{\infty} 2C\beta^j \leq \frac{2C\beta^{\lceil \frac{m}{2} \rceil}}{1-\beta}$$

Altogether, we find that there exists some $D > 0$ such that

$$|u(\underline{x}) - u(\underline{y})| \leq D(\sqrt{\beta})^m \qquad \forall m \geq 1,\, \forall \underline{a} \in W(m,A),\, \forall \underline{x},\underline{y} \in C(\underline{a})$$

which completes the proof of the claim. $\square$

Let $\psi = \varphi + u \circ \sigma - u$. As $\varphi \in \mathcal{H}(\Sigma_A)$ by assumption and $u \in \mathcal{H}(\Sigma_A)$ by the claim, we have $\psi \in \mathcal{H}(\Sigma_A)$. Also

$$\begin{aligned} \psi(\underline{x}) \;&=\; \varphi(\underline{x}) + u(\sigma\underline{x}) - u(\underline{x}) = \varphi(\underline{x}) + \sum_{j=0}^{\infty} \varphi(\sigma^{j+1}\underline{x}) - \sum_{j=0}^{\infty} \varphi(\sigma^j r(\sigma\underline{x})) - \\ &\quad - \sum_{j=0}^{\infty} \varphi(\sigma^j\underline{x}) + \sum_{j=0}^{\infty} \varphi(\sigma^j r(\underline{x})) = \sum_{j=0}^{\infty} \left(\varphi(\sigma^j r(\underline{x})) - \varphi(\sigma^j r(\sigma\underline{x}))\right). \end{aligned}$$

Now for $\underline{x} = (\dots x_{-1}x_0x_1\dots)$, we have that $r(\underline{x})$ and $r(\sigma\underline{x})$ depend only on the future coordinates $x_0, x_1, \dots$ of $\underline{x}$, so it follows that $\psi \in \widetilde{\mathcal{H}}(\Sigma_A)$. $\square$

Now the construction of a Gibbs measure for a potential $\varphi \in \mathcal{H}(\Sigma_A)$ proceeds as follows. By the above considerations, there exists $\psi \sim \varphi$ such that $\psi \in \widetilde{\mathcal{H}}(\Sigma_A)$. This ψ can be identified with some $\widehat{\psi} \in \mathcal{H}(\Sigma_A^+)$. The construction of a Gibbs measure $\mu_{\widehat{\psi}} \in \mathcal{M}_{\text{inv}}(\Sigma_A^+)$ for such a one-sided case will be discussed in the next section. Then we "pull back" $\mu_{\widehat{\psi}}$ to Σ_A as follows. To $f \in \mathcal{C}(\Sigma_A)$ assign $f^* \in \mathcal{C}(\Sigma_A^+)$ by

$$\text{for} \quad \underline{x} \in \Sigma_A^+ \quad \text{let} \quad f^*(\underline{x}) = \min\{f(\underline{y}) \,|\, (\dots y_{-1}y_0y_1\dots) = \underline{y} \in \Sigma_A, (y_0y_1\dots) = (x_0x_1\dots)\}.$$

Then define

$$\mu_\psi(f) = \lim_{n\to\infty} \mu_{\widehat{\psi}}((f \circ \sigma^n)^*).$$

It can be checked that μ_ψ is a Gibbs measure for ψ, and hence for φ.

5.4 Gibbs measures

The Ruelle-Perron-Frobenius theorem

The construction of the Gibbs measure $\mu_\Phi \in \mathcal{M}_{\text{inv}}(\Sigma_A^+)$ for $\Phi \in \mathcal{H}(\Sigma_A^+)$ relies on the Ruelle-Perron-Frobenius theorem stated below and proved in the next section. In a sense, this is a functional analytic generalization of the *Perron-Frobenius theorem* from linear algebra, which we recall here.

If B_{ij} is some nonnegative $K \times K$ matrix (i.e. $B_{ij} \geq 0$ for any $(i,j) \in \{0,\dots,K-1\}^2$), and B is primitive (there exists some $N \geq 1$ such that the Nth power of B is positive, i.e. $(B^N)_{ij} > 0$ for any $(i,j) \in \{0,\dots,K-1\}^2$), then

- B_{ij} has a maximal simple eigenvalue $\lambda > 0$,
- both the corresponding left eigenvector $u_i > 0$ and the corresponding right eigenvector $s_i > 0$ for any $i \in \{0,\dots,K-1\}$, and can be normalized so that $\sum\limits_{i=0}^{K-1} u_i s_i = 1$,
- we have $\lim\limits_{n\to\infty} \lambda^{-n}(B^n)_{ij} = s_i u_j$.

To formulate the Ruelle-Perron-Frobenius theorem, recall that $\mathcal{C}(\Sigma_A^+)$ denotes the space of continuous functions $f : \Sigma_A^+ \to \mathbb{R}$, and that we are given a Hölder continuous potential $\Phi \in \mathcal{H}(\Sigma_A^+) \subset \mathcal{C}(\Sigma_A^+)$. For the non-invertible $\sigma : \Sigma_A^+ \to \Sigma_A^+$ and to the potential $\Phi \in \mathcal{H}(\Sigma_A^+)$, assign the *Ruelle-Perron-Frobenius operator*

$$\mathcal{L}_\Phi : \mathcal{C}(\Sigma_A^+) \to \mathcal{C}(\Sigma_A^+); \qquad (\mathcal{L}_\Phi f)(\underline{x}) = \sum_{\sigma \underline{y} = \underline{x}} e^{\Phi(\underline{y})} f(\underline{y}).$$

$\mathcal{L} = \mathcal{L}_\Phi$ is a linear operator acting on $\mathcal{C}(\Sigma_A^+)$, and thus its adjoint $\mathcal{L}_\Phi^*$ is a linear operator that is acting on $\mathcal{M}(\Sigma_A^+)$, the space of Borel probability measures on Σ_A^+. The *Ruelle-Perron-Frobenius theorem* states that

- $\mathcal{L}_\Phi$ has a maximal eigenvalue $\lambda > 0$;
- there exists $h \in \mathcal{C}(\Sigma_A^+)$ such that $\mathcal{L}h = \lambda h$ and $h(\underline{x}) > 0$, $\forall \underline{x} \in \Sigma_A^+$;
- there exists $\nu \in \mathcal{M}(\Sigma_A^+)$ such that $\mathcal{L}^*\nu = \lambda\nu$, moreover, $\nu(h) = 1$ (so h is a probability density for ν);
- for every $g \in \mathcal{C}(\Sigma_A^+)$, we have

  $$\lim_{n\to\infty} \|\lambda^{-n}\mathcal{L}^n g - \nu(g)h\| = 0, \tag{5.4}$$

  where $\|.\|$ denotes the supremum norm.

Using the result and the notations of the Ruelle-Perron-Frobenius theorem, let us define $(\mu_\Phi =)\mu \in \mathcal{M}(\Sigma_A^+)$ as

$$\mu = h\nu, \qquad \text{that is } \mu(g) = \int_{\Sigma_A^+} g(\underline{x}) \cdot h(\underline{x}) d\nu(\underline{x});\ \forall g \in \mathcal{C}(\Sigma_A^+).$$

We will prove that μ defines this way is invariant, mixing, and that it is the unique Gibbs measure for $\Phi \in \mathcal{H}(\Sigma_A^+)$. First we prove three useful formulae.

$$(\mathcal{L}f) \cdot g = \mathcal{L}(f \cdot g \circ \sigma); \qquad \forall f, g \in \mathcal{C}(\Sigma_A^+). \tag{5.5}$$

To see this, note that:

$$((\mathcal{L}f) \cdot g)(\underline{x}) = \sum_{\underline{y}:\sigma\underline{y}=\underline{x}} e^{\Phi(\underline{y})} f(\underline{y}) \cdot g(\underline{x}) = \sum_{\underline{y}:\sigma\underline{y}=\underline{x}} e^{\Phi(\underline{y})} f(\underline{y}) \cdot g(\sigma\underline{y}).$$

The second formula is:

$$(\mathcal{L}^m f)(\underline{x}) = \sum_{\underline{y}:\sigma^m\underline{y}=\underline{x}} e^{S_m\Phi(\underline{y})} f(\underline{y}); \qquad \forall m \geq 1. \tag{5.6}$$

This can be proved by induction on m:

$$\begin{aligned}
(\mathcal{L}^{m+1} f)(\underline{x}) &= (\mathcal{L}(\mathcal{L}^m f))(\underline{x}) = \sum_{\underline{z}:\sigma\underline{z}=\underline{x}} e^{\Phi(\underline{z})} (\mathcal{L}^m f)(\underline{z}) = \\
&= \sum_{\underline{z}:\sigma\underline{z}=\underline{x}} e^{\Phi(\underline{z})} \left(\sum_{\underline{y}:\sigma^m\underline{y}=\underline{z}} e^{S_m\Phi(\underline{y})} f(\underline{y}) \right) = \\
&= \sum_{\underline{y}:\sigma^{m+1}\underline{y}=\underline{x}} \exp(\Phi(\sigma^m\underline{y}) + S_m\Phi(\underline{y})) \cdot f(\underline{y}) = \sum_{\underline{y}:\sigma^{m+1}\underline{y}=\underline{x}} e^{S_{m+1}\Phi(\underline{y})} \cdot f(\underline{y}).
\end{aligned}$$

Finally:, for any $m \geq 1$:

$$(\mathcal{L}^m f) \cdot g = \mathcal{L}^m(f \cdot g \circ \sigma^m); \qquad \forall f, g \in \mathcal{C}(\Sigma_A^+), \tag{5.7}$$

which can be derived form (5.6) exactly the same way as (5.5) is derived from the definition of $\mathcal{L}$.

Invariance of μ. Using at the consecutive steps that $\mathcal{L}h = \lambda h$, Formula (5.5) and that $\mathcal{L}^*\nu = \lambda\nu$, we have

$$\begin{aligned}
\mu(f) &= \nu(h \cdot f) = \nu(\lambda^{-1}\mathcal{L}(h) \cdot f) = \nu(\lambda^{-1}\mathcal{L}(h \cdot f \circ \sigma)) = \\
&= \lambda^{-1}(\mathcal{L}^*\nu)(h \cdot f \circ \sigma) = \nu(h \cdot f \circ \sigma) = \mu(f \circ \sigma)
\end{aligned}$$

for every $f \in \mathcal{C}(\Sigma_A^+)$, which proves that $\mu \in \mathcal{M}_{\text{inv}}(\Sigma_A^+)$.

Mixing of μ. Let E, F be cylinder sets, and let χ_E and χ_F denote the corresponding indicator functions. By the topology of Σ_A^+, we have $\chi_E, \chi_F \in \mathcal{C}(\Sigma_A^+)$. Now

$$\begin{aligned}\mu(E \cap \sigma^{-n} F) &= \nu(h \cdot \chi_E(\chi_F \circ \sigma^n)) = (\lambda^{-n}(\mathcal{L}^*)^n \nu)(h\chi_E\chi_F \circ \sigma^n) = \\ &= \lambda^{-n}\nu(\mathcal{L}^n(h\chi_E\chi_F \circ \sigma^n)) = \lambda^{-n}\nu(\mathcal{L}^n(h\chi_E)\chi_F) = \nu(\lambda^{-n}\mathcal{L}^n(h\chi_E)\chi_F))\end{aligned}$$

while

$$\mu(E) \cdot \mu(F) = \nu(\chi_E h) \cdot \nu(\chi_F h) = \nu(\nu(\chi_E h) \cdot h \cdot \chi_F).$$

Hence

$$\begin{aligned}\left|\mu(E \cap \sigma^{-n} F) - \mu(E) \cdot \mu(F)\right| &= \left|\nu\left(\left(\lambda^{-n}\mathcal{L}^n(h\chi_E) - \nu(h\chi_E)h\right)\chi_F\right)\right| \leq \\ &\leq \|\lambda^{-n}\mathcal{L}^n(h\chi_E) - \nu(h\chi_E)h\|\nu(F) \to 0\end{aligned}$$

by (5.4) of the Ruelle-Perron-Frobenius theorem.

Gibbs property of μ. For this, we will represent $\lambda > 0$ $\lambda = e^P$ for some $P \in \mathbb{R}$. We will see that μ is a Gibbs measure with $P = P_\mu$, and hence, by Lemma 5.7, $P = P(\Phi)$. Fix $m \geq 1$, $\underline{a} \in W(m, A)$ and let $E = C(\underline{a})$. We have to prove an upper and a lower bound on $\mu(E)$. For the *upper bound*, note that, as E is a cylinder set of length m, for any $\underline{z} \in \Sigma_A^+$ there exists *at most* one $\underline{x}' \in E$ such that $\sigma^m \underline{x}' = \underline{z}$ (namely, $\underline{x}' = (\underline{az})$ if $A_{a_{m-1}z_0} = 1$). Hence, for any $\underline{z} \in \Sigma_A^+$ the sum below (which arises from (5.6)) consists of at most one term:

$$\mathcal{L}^m(h\chi_E)(\underline{z}) = \sum_{\underline{y}:\sigma^m\underline{y}=\underline{z}} e^{S_m\Phi(\underline{y})}h(\underline{y})\chi_E(\underline{y}) \leq e^{S_m\Phi(\underline{x}')}h(\underline{x}') \leq e^{\mathcal{S}_m(\Phi,\underline{a})}\|h\|.$$

This implies

$$\mu(E) = \nu(h\chi_E) = \lambda^{-m}(\mathcal{L}^*)^m(\nu)(h\chi_E) = \lambda^{-m}\nu(\mathcal{L}^m(h\chi_E)) \leq c_2 e^{-P(\Phi)m + \mathcal{S}_m(\Phi,\underline{a})}$$

where $c_2 = \|h\|$.

For the *lower bound*, it can be exploited that the adjacency matrix A is primitive: there exists $M \geq 1$ such that $A_{ij}^M > 0$ for any pair $i, j \in \{0, \ldots, K-1\}^2$. Hence, for any $\underline{z} \in \Sigma_A^+$ there exists *at least* one $\underline{x}' \in E$ such that $\sigma^{m+M}\underline{x}' = \underline{z}$. Hence

$$\begin{aligned}\mathcal{L}^{m+M}(h\chi_E)(\underline{z}) &= \sum_{\underline{y}:\sigma^{m+M}\underline{y}=\underline{z}} e^{S_{m+M}\Phi(\underline{y})}h(\underline{y})\chi_E(\underline{y}) \geq \\ &\geq e^{S_{m+M}\Phi(\underline{x}')}h(\underline{x}') \geq e^{-M\|\Phi\|} \cdot e^{-D} \cdot e^{\mathcal{S}_m(\Phi,\underline{a})}\inf(h),\end{aligned}$$

where we have used (5.2). Also, by continuity and $h > 0$, we have $\inf(h) > 0$. This implies

$$\begin{aligned}\mu(E) = \lambda^{-(m+M)}\nu(\mathcal{L}^{m+M}(h\chi_E)) &\geq e^{-P(\Phi)m}\lambda^{-M} \cdot e^{-M\|\Phi\|} \cdot e^{-D} \cdot e^{\mathcal{S}_m(\Phi,\underline{a})}\inf(h) = \\ &\geq c_1 e^{-P(\Phi)m + \mathcal{S}_m(\Phi,\underline{a})},\end{aligned}$$

where $c_1 = \lambda^{-M} \cdot e^{-M\|\Phi\|} \cdot e^{-D} \cdot \inf(h)$.

Uniqueness. Note as μ is mixing it is also ergodic. Assume there exists another Gibbs measure μ' for the potential $\Phi \in \mathcal{H}(\Sigma_A^+)$. As both μ and μ' are Gibbs, there exists some $C > 0$ such that $\mu'(E) \leq C\mu(E)$ for any cylinder set E. But this implies $\mu' \ll \mu$, and thus by Lemma 1.7, $\mu' = \mu$.

5.5 Proof of the Ruelle-Perron-Frobenius theorem

Step 1: construction of ν and λ. Let us define the following (nonlinear) operator on $\mathcal{M} = \mathcal{M}(\Sigma_A^+)$:

$$G(m) = \frac{1}{(\mathcal{L}^* m)(1)} \mathcal{L}^* m \qquad m \in \mathcal{M},$$

where $1 \in \mathcal{C}(\Sigma_A^+)$ denotes the function $1(\underline{x}) \equiv 1$. It is claimed that $G(m) \in \mathcal{M}$:

- $\mathcal{L}$ is positive: if $\mathcal{C}(\Sigma_A^+) \ni f \geq 0$, then $\mathcal{L} f \geq 0$. This implies that for $f \geq 0$

$$(\mathcal{L}^* m)(f) = m(\mathcal{L} f) \geq 0 \implies (Gm)(f) \geq 0.$$

- by rescaling with $(\mathcal{L}^* m)(1)$, we have $(Gm)(1) = 1$.

Now as

- $\mathcal{M}$ is convex and compact in the weak-$*$ topology,
- $G : \mathcal{M} \to \mathcal{M}$ is continuous in the weak-$*$ topology,

it follows by Schauder's fixed point theorem that there exists some fixed point $\nu \in \mathcal{M}$ such that

$$G(\nu) = \nu \iff \mathcal{L}^* \nu = \lambda \nu, \qquad \text{where } \lambda = (\mathcal{L}^* \nu)(1).$$

Step 2. Definition and invariance of $\Lambda \subset \mathcal{C}(\Sigma_A^+)$. We would like to mimic the previous step to construct h as a fixed point, using Schauder's theorem, however, $\mathcal{C}(\Sigma_A^+)$ is not compact in the supremum topology, and the restriction $\nu(f) = 1$ would not result in a compact subset either, as the space is infinite dimensional. Recall the *Arzela-Ascoli theorem* which states that $N \subset \mathcal{C}(\Sigma_A^+)$ is compact if and only if two conditions are met:

- *uniformly bounded*: there exists $K > 0$ such that $|f(\underline{x})| \leq K$, for every $f \in N$ and every $\underline{x} \in \Sigma_A^+$.
- *uniformly equicontinuous*: for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $f \in N$, whenever $d(\underline{x}, \underline{y}) < \delta$ we have $|f(\underline{x}) - f(\underline{y})| < \varepsilon$.

Here we construct a subset $\Lambda \subset \mathcal{C}(\Sigma_A^+)$ that meets these criteria.

As $\Phi \in \mathcal{H}(\Sigma_A^+)$, there exist $b > 0$ and $\beta < 1$ such that, for any

$$|\Phi(\underline{x}) - \Phi(\underline{y})| \leq b \beta^m, \qquad \forall m \geq 1,\ \underline{a} \in W(m, A),\ \underline{x}, \underline{y} \in C(\underline{a}).$$

Let

$$\Lambda = \left\{ f \in \mathcal{C}(\Sigma_A^+) \,\middle|\, f \geq 0,\ \nu(f) = 1,\ |\log f(\underline{x}) - \log f(\underline{y})| \leq 2b \frac{\beta^{m+1}}{1 - \beta} \right\}$$

where the condition is again on $\forall m \geq 1$, $\underline{a} \in W(m, A)$, $\underline{x}, \underline{y} \in C(\underline{a})$. Equivalently

$$\frac{f(\underline{x})}{f(\underline{y})} \leq B_m; \qquad \text{where } B_m = \exp\left(2b \frac{\beta^{m+1}}{1 - \beta}\right); \qquad \forall m \geq 1,\ \underline{a} \in W(m, A),\ \underline{x}, \underline{y} \in C(\underline{a}).$$

We claim that Λ is preserved by the action of $\lambda^{-1} \mathcal{L}$. To see this, consider $f \in \Lambda$. Then

- $\lambda^{-1}\mathcal{L}f \geq 0$ is obvious.

- $\nu(\lambda^{-1}\mathcal{L}f) = \lambda^{-1}\nu(\mathcal{L}f) = (\lambda^{-1}\mathcal{L}^*\nu)(f) = G(\nu)(f) = \nu(f) = 1$

- let $\underline{x}, \underline{y} \in C(\underline{a})$ for $\underline{a} \in W(m, A)$, $m \geq 1$. This means $x_0 \dots x_{m-1} = y_0 \dots y_{m-1} = \underline{a}$, and, in particular, $x_0 = y_0 = a_0$. Let us denote by $\mathcal{A}_{a_0}$, the collection of $j \in \{0, \dots, K-1\}$ for which $A_{ja_0} = 1$ (or equivalently, for which $A_{jx_0} = 1$, or equivalently, for which $A_{jy_0} = 1$). We have

$$(\mathcal{L}f)(\underline{x}) = \sum_{j \in \mathcal{A}_{a_0}} e^{\Phi(j\underline{x})} f(j\underline{x}); \qquad (\mathcal{L}f)(\underline{y}) = \sum_{j \in \mathcal{A}_{a_0}} e^{\Phi(j\underline{y})} f(j\underline{y})$$

and these two expressions can be compared term by term. Indeed, for any $j \in \mathcal{A}_{a_0}$, we have $j\underline{x}, j\underline{y} \in C(j\underline{a})$, where $j\underline{a} \in W(m+1, A)$, hence

$$\frac{e^{\Phi(j\underline{x})}}{e^{\Phi(j\underline{y})}} \leq e^{b\beta^{m+1}} \qquad \text{by Hölder continuity of } \Phi,$$
$$\frac{f(j\underline{x})}{f(j\underline{y})} \leq B_{m+1} \qquad \text{as } f \in \Lambda.$$

Now

$$e^{b\beta^{m+1}} B_{m+1} = \exp\left(b\beta^{m+1} + 2b\frac{\beta^{m+2}}{1-\beta}\right) \leq \exp\left(2b\frac{\beta^{m+1}(1-\beta) + \beta^{m+2}}{1-\beta}\right) = B_m.$$

So

$$e^{\Phi(j\underline{x})} f(j\underline{x}) \leq B_m \cdot e^{\Phi(j\underline{y})} f(j\underline{y}), \qquad \forall j \in \mathcal{A}_{a_0},$$

and summing up on j we get

$$(\mathcal{L}f)(\underline{x}) \leq B_m \cdot (\mathcal{L}f)(\underline{y}) \qquad \text{hence} \quad (\lambda^{-1}\mathcal{L}f) \in \Lambda.$$

Comment. Note that we have proved a little more, namely

$$(\mathcal{L}f)(\underline{x}) \leq e^{b\beta^{m+1}} B_{m+1} \cdot (\mathcal{L}f)(\underline{y}) \tag{5.8}$$

which will be useful for later purposes.

Step 3: compactness of Λ and construction of h. We prove that $\Lambda \subset \mathcal{C}(\Sigma_A^+)$ is compact by verifying the Arzela-Ascoli conditions. First let us show that there exists $K > 0$ such that $\forall f \in \Lambda$ we have $\|f\| \leq K$ for the supremum norm. Recall that A is primitive: there exists $M \geq 1$ such that $A_{ij}^M > 0$ for every pair i, j. Hence, for any two points $\underline{x}, \underline{z} \in \Sigma_A^+$ there exists at least one $\underline{y} \in \Sigma_A^+$ such that $\sigma^M \underline{y} = \underline{x}$ and $y_0 = z_0$. Hence, using (5.6):

$$(\mathcal{L}^M f)(\underline{x}) \geq e^{-M\|\Phi\|} f(\underline{y}) \geq e^{-M\|\Phi\|} B_0^{-1} f(\underline{z})$$

and thus

$$(\lambda^{-M}\mathcal{L}^M f)(\underline{x}) \geq K^{-1} f(\underline{z}), \qquad \forall \underline{x}, \underline{z} \in \Sigma_A^+$$

with $K = \lambda^M e^{M\|\Phi\|} B_0$, Now as $\nu(\lambda^{-M}\mathcal{L}^M f) = 1$, there is at least one $\underline{x} \in \Sigma_A^+$ where $(\lambda^{-M}\mathcal{L}^M f)(\underline{x}) \geq 1$, hence $f(\underline{z}) \leq K$, $\forall \underline{z} \in \Sigma_A^+$ – that is, by $f \geq 0$, $\|f\| \leq K$ follows.

Comment: By a similar argument, using $\nu(f) = 1$, it also follows that

$$\inf(\lambda^{-M}\mathcal{L}^M f) \geq K^{-1}; \qquad \forall f \in \Lambda. \tag{5.9}$$

This also implies that for any $f \in \Lambda$

$$|f(\underline{x}) - f(\underline{x}')| \leq K \left| \frac{f(\underline{x})}{f(\underline{x}')} - 1 \right| \leq K|B_m - 1|$$

whenever $\underline{x}, \underline{x}' \in C(\underline{a})$ for some $\underline{a} \in W(m, A)$, which implies uniform equicontinuity as $B_m \to 1$ when $m \to \infty$.

Now as $\lambda^{-1}\mathcal{L}$ preserves Λ, which is a convex, compact set in the supremum norm, by Schauder fixpoint theorem there exists $h \in \Lambda$ such that $\mathcal{L}h = \lambda h$, and as $h \in \Lambda$ we have $\nu(h) = 1$ and $h > 0$. (Actually, as $\lambda^{-M}\mathcal{L}^M h = h \in \Lambda$, the above considerations also imply $\inf h \geq K^{-1}$.)

Step 4. Coupling.

Lemma 5.13. *There exists $\eta > 0$ such that for every $f \in \Lambda$ there exists $f' \in \Lambda$ with*

$$\lambda^{-M}\mathcal{L}^M f = \eta h + (1 - \eta) f'$$

Proof. We have to find η so small that

$$g = \lambda^{-M}\mathcal{L}^M f - \eta h$$

is positive and satisfies the regularity properties included in the definition of Λ. Then, letting $f' = (1-\eta)^{-1} g$, $\nu(f') = 1$ is automatic as

$$\nu(g) = \nu(\lambda^{-M}\mathcal{L}^M f) - \eta\nu(h) = 1 - \eta.$$

First note that by (5.9) $\inf(\lambda^{-M}\mathcal{L}^M f) \geq K^{-1}$ and as $\|h\| \leq K$, choosing $\eta < K^{-2}$ we have $g \geq K^{-1} - \eta K > 0$. We still need to ensure that for any $m \geq 1$, $\underline{a} \in W(m, A)$ and $\underline{x}, \underline{x}' \in C(\underline{a})$ we have

$$g(\underline{x}) \leq B_m g(\underline{x}') \quad \text{that is} \quad \lambda^{-M}\mathcal{L}^M f(\underline{x}) - \eta h(\underline{x}) \leq B_m(\lambda^{-M}\mathcal{L}^M f(\underline{x}') - \eta h(\underline{x}'))$$

for every $f \in \Lambda$. Equivalently, by choosing η sufficiently small, it can be ensured that

$$\eta(B_m h(\underline{x}') - h(\underline{x})) \leq B_m \lambda^{-M}\mathcal{L}^M f(\underline{x}') - \lambda^{-M}\mathcal{L}^M f(\underline{x}). \tag{5.10}$$

By (5.8):

$$\lambda^{-M}\mathcal{L}^M f(\underline{x}) \leq e^{b\beta^{m+1}} B_{m+1} \lambda^{-M}\mathcal{L}^M f(\underline{x}')$$

and thus

$$\text{RHS of (5.10)} \ \geq (B_m - e^{b\beta^{m+1}} B_{m+1})\lambda^{-M}\mathcal{L}^M f(\underline{x}') \geq (B_m - e^{b\beta^{m+1}} B_{m+1})K^{-1}$$

by (5.9). Yet, as $h \in \Lambda$, we have $h(\underline{x}) \geq B_m^{-1} h(\underline{x}')$, and thus

$$\text{LHS of (5.10)} \ \leq \eta(B_m - B_m^{-1})h(\underline{x}) \leq \eta(B_m - B_m^{-1})K.$$

So it is enough to prove that, with a suitably small η it can be ensured that

$$\eta(B_m - B_m^{-1}) \leq (B_m - e^{b\beta^{m+1}} B_{m+1})K^{-2}, \qquad \forall m \geq 1. \tag{5.11}$$

Now as $B_m \to 1$ as $m \to \infty$, there exists some $L > 0$ such that

$$\log B_m, \log B_m^{-1}, \log(e^{b\beta^{m+1}} B_{m+1}) \in [L^{-1}, L]; \qquad \forall m \geq 1.$$

Also, there exist $0 < u_1 < u_2$ such that

$$u_1|x - y| \leq |e^x - e^y| \leq u_2|x - y|; \qquad \forall x, y \in [L^{-1}, L].$$

Hence, to prove (5.11), it is enough to ensure

$$\eta u_2 |\log B_m - \log B_m^{-1}| \leq K^{-2} u_1 |\log B_m - \log(e^{b\beta^{m+1}} B_{m+1})|.$$

Now

$$\begin{aligned} |\log B_m - \log B_m^{-1}| &= 4b\frac{\beta^{m+1}}{1-\beta}, \\ |\log B_m - \log(e^{b\beta^{m+1}} B_{m+1})| &= b\beta^{m+1}, \end{aligned}$$

so

$$\eta < \frac{u_1(1-\beta)}{4u_2K^2}$$

will work. ◻

Step 4. Convergence for $f \in \Lambda$.

Lemma 5.14. *There exist $A > 0$ and $\alpha < 1$ such that, for any $f \in \Lambda$ and $n \geq 1$*

$$\|\lambda^{-n}\mathcal{L}^n f - h\| \leq A\alpha^n; \ \forall n \geq 1.$$

Proof. Let us write $n = Mq + r$ where $q \geq 0$ and $0 \leq r < M$. Applying Lemma 5.13 repeatedly, we arrive at

$$\begin{aligned} \lambda^{-qM}\mathcal{L}^{qM} f &= (\eta + \eta(1-\eta) + \cdots + \eta(1-\eta)^{q-1})h + (1-\eta)^q f_q = \\ &= (1 - (1-\eta)^q)h + (1-\eta)^q f_q \qquad \text{for some } f_q \in \Lambda. \end{aligned}$$

Using $\|f_q\| \leq K$ and $\|h\| \leq K$, this implies

$$\|\lambda^{-qM}\mathcal{L}^{qM}f - h\| \leq 2K(1-\eta)^q$$

and thus

$$\|\lambda^{-n}\mathcal{L}^n f - h\| = \|\lambda^{-r}\mathcal{L}^r(\lambda^{-qM}\mathcal{L}^{qM}f - h)\| \leq A\alpha^n$$

with

$$\begin{aligned} A &= 2K(1-\eta)^{-1}\max_{0\leq r<M}\|\lambda^{-r}\mathcal{L}^r\|; \quad \text{and} \\ \alpha &= (1-\eta)^{1/M}. \end{aligned}$$

Step 5. Approximation. To prove (5.4), the convergence of Lemma 5.14 is extended form Λ to $\mathcal{C}(\Sigma_A^+)$ as follows.For $r \geq 1$, let $\mathcal{C}_r \subset \mathcal{C}(\Sigma_A^+)$ denote the set of step functions that are constant on cylinder sets of length r.

Claim 5.15. *For $f \in \mathcal{C}_r$ we have*

$$\|\lambda^{-n-r}\mathcal{L}^{n+r}f - \nu(f)h\| \leq A\nu(|f|)\alpha^n;\ \forall n \geq 1.$$

Proof. We may assume that (i) $\nu(f) = 1$ (otherwise rescale by a constant factor and use linearity) and that (ii) $f \geq 0$ (otherwise represent as $f_+ - f_-$ and again use linearity). Now the claim follows from Lemma 5.14 if we show that for $f \in \mathcal{C}_r$ with $f \geq 0$ and $\nu(f) = 1$, $\lambda^{-r}\mathcal{L}^r f \in \Lambda$. To get the required regularity, let $m \geq 1$ $\underline{x}, \underline{x}' \in C(\underline{a})$ for $\underline{a} \in W(m, A)$. Let $\mathcal{A}^r_{a_0}$ denote the set of $\underline{j} = (j_0 \dots j_{r-1}) \in W(r, A)$ for which $A_{j_{r-1}a_0} = 1$. We have

$$(\mathcal{L}^r f)(\underline{x}) = \sum_{\underline{j}\in\mathcal{A}^r_{a_0}} e^{S_r\Phi(\underline{j}\underline{x})}f(\underline{j}\underline{x}); \qquad (\mathcal{L}^r f)(\underline{x}') = \sum_{\underline{j}\in\mathcal{A}^r_{a_0}} e^{S_r\Phi(\underline{j}\underline{x}')}f(\underline{j}\underline{x}')$$

and the two expressions can be compared term by term: as f is constant on cylinder sets of length r, $f(\underline{j}\underline{x}) = f(\underline{j}\underline{x}')$, while, arguing as in the proof of (5.2),

$$|S_r\Phi(\underline{j}\underline{x}) - S_r\Phi(\underline{j}\underline{x}')| \leq b(\beta^{r+m} + \cdots + \beta^m) \implies e^{S_r\Phi(\underline{j}\underline{x})} \leq B_m e^{S_r\Phi(\underline{j}\underline{x}')}.$$

□

To extend to arbitrary $g \in \mathcal{C}(\Sigma_A^+)$ use that for any ε there exists $r \geq 1$ and $f_1, f_2 \in \mathcal{C}_r$ such that $f_1 \leq g \leq f_2$ and $0 \leq f_2 - f_1 \leq \varepsilon$, which implies $|\nu(g) - \nu(f_i)| \leq \varepsilon$ and thus $|\lambda^{-m}\mathcal{L}^m f_i - \nu(g)h| \leq \varepsilon(1 + \|h\|)$ for m large enough and $i = 1, 2$. Moreover, as $\mathcal{L}$ is a positive operator, $\lambda^{-m}\mathcal{L}^m f_1 \leq \lambda^{-m}\mathcal{L}^m g \leq \lambda^{-m}\mathcal{L}^m f_2$. □