10. The ergodic theory of hyperbolic dynamical systems

§10.1 Introduction

In Lecture 8 we studied thermodynamic formalism for shifts of finite type by defining a suitable transfer operator acting on a certain Banach space of functions and studying its spectral properties. In Lecture 9 we gave some applications of this methodology, mostly in the context of shifts of finite type. In this lecture we show how the use of thermodynamic formalism can be used to study a wide range of dynamical system that possesses some degree of 'hyperbolicity'.

§10.2 Anosov diffeomorphisms

Let M be a compact Riemannian manifold without boundary, with metric d on M derived from the Riemannian metric. For $x \in M$ let T_xM denote the tangent space at x and let TM denote the tangent bundle. Let $T : M \to M$ be a C^1 diffeomorphism of M and let $D_xT : T_xM \to T_{T(x)}M$ denote the derivative of T.

Definition. We say that $T : M \to M$ is an *Anosov diffeomorphism* if the tangent bundle TM has a continuous splitting into a Whitney sum of two DT-invariant sub-bundles $TM = E^s \oplus E^u$ such that there exist constant $C > 0$ and $\lambda \in (0,1)$ such that for all $n \geq 0$

$$\begin{aligned} \|D_xT^n v\| &\leq C\lambda^n \|v\|, \quad \text{for all } v \in E^s_x, \\ \|D_xT^{-n} v\| &\leq C\lambda^n \|v\|, \quad \text{for all } v \in E^u_x. \end{aligned}$$

We call E^s and E^u the stable and unstable sub-bundles, respectively.

Thus T is an Anosov diffeomorphism if each tangent plane splits into two complementary directions; one of which contracts exponentially fast under iteration and the other expands exponentially fast under iteration. (Note that contraction by DT^{-1} corresponds to expansion by DT; for technical reasons that we shall see below it is more convenient to describe expansion by T in terms of contraction by T^{-1}.)

The subspace E^s_x is a subspace of the tangent space T_xM at x. It is tangent to the stable manifold $W^s(x)$ through x. The stable manifold is an immersed submanifold in M and is characterised by:

$$W^s(x) = \{y \in M \mid d(T^n x, T^n y) \to 0 \text{ as } n \to \infty\}. \tag{10.1}$$

One can show that the convergence in (10.1) is necessary exponential, thus $W^s(x)$ is characterised by

$$W^s(x) = \{y \in M \mid d(T^n x, T^n y) \to 0 \text{ exponentially fast as } n \to \infty\}.$$

Similarly, there exist unstable manifolds. The unstable manifold $W^u(x)$ is an immersed manifold through x tangent to E^u_x and characterised by

$$W^s(x) = \{y \in M \mid d(T^{-n} x, T^{-n} y) \to 0 \text{ as } n \to \infty\}.$$

This illustrates why it makes sense to define the unstable directions in terms of contraction in backwards time. The manifold M is compact and the metric d gives M finite diameter; hence $d(T^n x, T^n y)$ cannot tend to infinity (although, if x and y are sufficiently close then for all sufficiently small n we do have that $d(T^n x, T^n y)$ increases exponentially fast).

§10.2.1 Example: the cat map

Recall that the cat map $T : \mathbb{R}^2/\mathbb{Z}^2 \to \mathbb{R}^2/\mathbb{Z}^2$ is defined by

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \bmod 1.$$

Let A denote this 2×2 matrix. Then A has two eigenvalues

$$\lambda_u = \frac{3+\sqrt{5}}{2} > 1, \ \lambda_s = \frac{3-\sqrt{5}}{2} \in (0,1),$$

with corresponding eigenvectors

$$v_u = \left(1, \frac{-1+\sqrt{5}}{2}\right), \ v_s = \left(1, \frac{-1-\sqrt{5}}{2}\right),$$

respectively.

Let $x \in M$. Then $T_x M$ can be decomposed as the direct sum of two subspaces, E^u_x and E^s_x, that are parallel to v_u and v_s, respectively. It is easy to see that $D_x T$ maps E^u_x to E^u_{Tx} and expands vectors by a factor of λ_u, and maps E^s_x to E^s_{Tx} and expands vectors by a factor of λ_s. This gives us the unstable and stable bundles.

Each subspace E^s_x is tangent to a one-dimensional submanifold of M in the direction v_s. This is the stable manifold through x and is the geodesic through x in direction v_s. As the two components of v_s are rationally independent, this geodesic is dense in M. Thus each stable manifold is dense in M. Similarly, we can construct the unstable manifolds.

§10.2.2 Which manifolds support Anosov diffeomorphisms?

It is an important open question to determine which manifolds M support an Anosov diffeomorphism, and, indeed, to classify all Anosov diffeomorphisms.

Recall that two continuous transformations $T_1 : X_1 \to X_1, T_2 : X_2 \to X_2$ are said to be *topologically conjugate* if there is a homeomorphism $h : X_1 \to X_2$ such that $hT_1 = T_2h$. If X_1, X_2 are manifolds then one can impose higher regularity conditions on the conjugacy h and speak of C^r-conjugate dynamical systems. In the context of hyperbolic dynamical systems it is natural to assume that a conjugacy h is $C^{1+\alpha}$, meaning that h is differentiable and the derivative is Hölder continuous.

By generalising the construction of the cat map and using a hyperbolic matrix in $SL(k, \mathbb{Z})$, it is clear that any k-dimensional torus (with $k \geq 2$) supports an Anosov diffeomorphism. More specifically, if $A \in SL(k, \mathbb{Z})$ has no eigenvalues of modulus 1 then A determines an Anosov diffeomorphism $T : \mathbb{R}^k/\mathbb{Z}^k \to \mathbb{R}^k/\mathbb{Z}^k$. We call such a map an *Anosov automorphism of a torus*. Note that the restriction $k \geq 2$ is necessary: there are no Anosov diffeomorphisms on a circle as the circle is 1-dimensional and the presence of stable and unstable bundles requires at least 2 dimensions.

One can view this construction in the following way: a matrix $A \in SL(k, \mathbb{Z})$ determines an automorphism of the abelian group $\mathbb{R}^k$. This automorphism leaves the discrete subgroup $\mathbb{Z}^k \subset \mathbb{R}^k$ invariant and so determines a well-defined map on the quotient space $\mathbb{R}^k/\mathbb{Z}^k$.

Recall that a *Lie group* G is a group that is also a smooth manifold (and the group operations are continuous). If G is a topological group then the commutator $[G, G]$ of G is the closed subgroup generated by elements of the form $ghg^{-1}h^{-1}$. We say that G is *nilpotent* if G^k is trivial for some k, where G^k is defined inductively by $G^0 = G$, $G^k = [G^{k-1}, G]$. Notice that abelian Lie groups, such as the k-dimensional torus, are nilpotent Lie groups.

Definition. Let N be a connected nilpotent Lie group and let $\Gamma \subset N$ be a discrete subgroup such that N/Γ is compact. (We do not assume that Γ is normal.) We call N/Γ a *nilmanifold*.

As an example of a nilpotent Lie group that is not abelian, consider the Heisenberg group, namely the group of matrices

$$N = \left\{ \left(\begin{array}{ccc} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{array} \right) \mid x, y, z \in \mathbb{R} \right\}.$$

The subgroup

$$\Gamma = \left\{ \left(\begin{array}{ccc} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{array} \right) \mid x, y, z \in \mathbb{Z} \right\}$$

is a cocompact lattice, and so N/Γ is a nilmanifold.

Let N be a connected Lie group and let $\Gamma \subset N$ be a cocompact lattice. Suppose that $A : N \to N$ is an automorphism of N such that $A(\Gamma) = \Gamma$. Then A induced a diffeomorphism $T : N/\Gamma \to N/\Gamma$ of the nilmanifold N/Γ defined by $T(gN) = A(g)N$. The derivative D_eA of A at the identity is a linear map defined on the tangent space T_eN of N at the identity (this is the Lie algebra of N). If the linear map $D_eA : T_eN \to T_eN$ has no eigenvalues of modulus 1 then $T : N/\Gamma \to N/\Gamma$ is an Anosov diffeomorphism. We call it an *Anosov automorphism of a nilmanifold*.

Of course, the above construction works for any automorphism of a connected Lie group G that leaves invariant a cocompact lattice Γ. However, if the quotient map on G/Γ is Anosov then the group G must be nilpotent. Not every nilmanifold admits an Anosov diffeomorphism; indeed, the smallest dimension in which one exists has $\dim N = 6$. (See the references.)

Essentially, Anosov automorphisms are the only known examples of Anosov diffeomorphisms.

Theorem 10.1
Let $T : M \to M$ be a $C^{1+\alpha}$ Anosov diffeomorphism. Then T is $C^{1+\alpha'}$ conjugate to an Anosov automorphism of a torus, nilmanifold, or infranilmanifold.

(An infranilmanifold is a manifold that has a nilmanifold as a finite cover.) The question for C^1 conjugacy is more subtle (see the references).

It is a major open problem to determine if these are all the Anosov diffeomorphisms. If a manifold admits an Anosov diffeomorphism then this requires very strong homological properties of the manifold; one can show that there are no Anosov diffeomorphisms on a Möbius band or Klein bottle, for example.

§10.3 Hyperbolic dynamical systems

Let $T : M \to M$ be a C^1 diffeomorphism of a smooth compact Riemannian manifold without boundary. Instead of requiring T_xM to have a hyperbolic splitting for all $x \in M$ (which, as we have seen, is a very strong assumption), we could instead require the spliting to exist only on a T-invariant compact subset $\Lambda \subset M$. We will not necessarily assume that Λ is a submanifold. Indeed, in many cases, Λ will be topologically very complicated; for example, Λ may be a Cantor set.

Definition. A compact T-invariant set $\Lambda \subset M$ is said to be a *locally maximal hyperbolic set* or a *basic set* if the following conditions hold:

(i) there exists a continuous DT-invariant splitting of the tangent bundle of M restricted to Λ:

$$T_xM = E^s_x \oplus E^u_x \text{ for all } x \in \Lambda$$

and constants $C > 0$, $\lambda \in (0,1)$ such that for all $n \geq 0$

$\|D_x T^n v\| \leq C\lambda^n \|v\|$, for all $v \in E^s_x$, $\|D_x T^{-n} v\| \leq C\lambda^n \|v\|$, for all $v \in E^u_x$;

(ii) $T : \Lambda \to \Lambda$ has a dense set of periodic points and a dense orbit;

(iii) there exists an open set $U \supset \Lambda$ such that

$$\bigcap_{n=-\infty}^{\infty} T^{-n} U = \Lambda.$$

In addition, we will normally assume that Λ is larger than a single orbit. For brevity, we shall often say that $T : \Lambda \to \Lambda$ is *hyperbolic*.

Note that condition (iii) says that orbits that start nearby to Λ will converge to Λ under either forwards or backwards iteration.

Definition. We say that a locally maximal hyperbolic set Λ is a *hyperbolic attractor* if all sufficiently nearby points converge to Λ under forwards iteration. That is, in (iii) above we have that

$$\bigcap_{n=0}^{\infty} T^n U = \Lambda.$$

It will be useful to impose an extra condition on the dynamics of T.

Definition. A continuous transformation T of a compact metric space X is said to be *topologically transitive* if for all non-empty open sets $U, V \subset X$, there exists $n \in \mathbb{N}$ such that $T^{-n}U \cap V \neq \emptyset$. It is straightforward to show that T is topologically transitive if and only if there is a dense orbit.

Moreover, we say that T is *topologically mixing* if for all non-empty open sets $U, V \subset X$, there exists $N \in \mathbb{N}$ such that for all $n \geq N$ we have $T^{-n}U \cap V \neq \emptyset$. (Compare this with the measure-theoretic notion of strong-mixing in Lecture 5.)

Thus topological mixing implies topological transitivity, but not conversely. If $\sigma : \Sigma \to \Sigma$ is a shift of finite type determined by a $0-1$ transition matrix A then σ is topologically transitive if and only if A is irreducible, and σ is topologically mixing if and only if A is aperiodic.

One can show that if $T : \Lambda \to \Lambda$ is hyperbolic, then Λ decomposes into a disjoint union $\Lambda = \bigcup_{j=0}^{k-1} \Lambda_j$ such that $T(\Lambda_j) = \Lambda_{j+1 \bmod k}$ and $T^k : \Lambda_j \to \Lambda_j$ is a topologically mixing hyperbolic map. Thus by replacing T by an iterate and decomposing Λ, we can, without loss of generality, assume that T is topologically mixing.

§10.4 Examples

§10.4.1 The Smale horseshoe

Let $\Delta \subset \mathbb{R}^2$ be a rectangle. Let $T : \Delta \to \mathbb{R}^2$ be a diffemorphism of Δ onto its image such that the intersection $\Delta \cap T(\Delta)$ consists of two horizontal 'bands' stretching across Δ. Moreover, we assume that T, when restricted to Δ is a hyperbolic affine map that contracts in the vertical direction and expands in the horizontal direction.

It is clear that $T^n(\Delta)$ consists of 2^n pairwise disjoint horizontal bands of exponentially shrinking height. Moreover, $T^{-n}(\Delta)$ consists of 2^n pairwise disjoint vertical bands of exponentially shrinking width.

The set $\Lambda = \cap_{n=-\infty}^{\infty} T^n \Delta \subset \Delta$ is a T-invariant subset of Δ. It is clear from the above discussion that Λ is the product of two Cantor sets. The dynamics of T restricted to Λ is conjugate to the full two-sided 2-shift, and hence has dense periodic orbits and a dense orbit. It is clear the Λ is a locally maximal hyperbolic set.

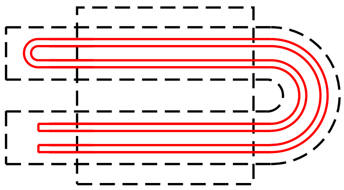

Figure 10.1: The Smale horseshoe and its first two iterates

§10.4.2 The solenoid

Let $M = \{(z, w) \in \mathbb{C}^2 \mid |z| = 1, |w| \leq 1\} = S^1 \times D$ denote the solid torus. Define $T : M \to M$ by

$$T(z, w) = \left(z^2, \frac{z}{2} + \frac{w}{4}\right).$$

Then each disc $\{z\} \times D$ is mapped onto the disc $\{z^2\} \times D$ with centre $z/2$ and radius $w/4$. Hence $T(M)$ intersects $\{z^2\} \times D$ in two discs of radius $1/4$, one with centre at $z/2$ and the other with centre at $-z/2$. Thus T acts by taking the solid torus M and wrapping it around twice, stretching it by a factor of two in one direction and contracting it by a factor of $1/4$ in the other.

Inductively, T^n wrap M around itself 2^n times. Thus $T^n(M)$ intersects each disc $\{z\} \times D$ in 2^n discs each of radiu $1/4^n$.

Let $\Lambda = \cap_{n=0}^{\infty} T^n(M)$. Then Λ has topological dimension 1. It intersects discs $\{z\} \times D$ in a Cantor set. Through each $x \in \Lambda$ there is a unique line L_x contained in Λ which wraps around the solid torus forever.

One can easily see that Λ is a hyperbolic attractor.

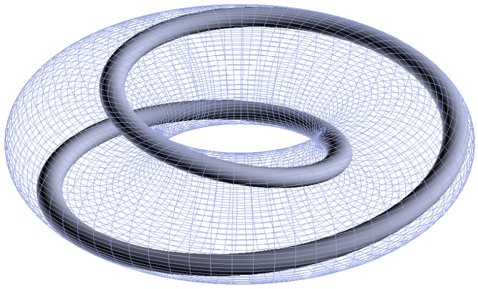

Figure 10.2: The solenoid

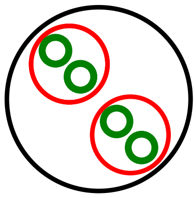

Figure 10.3: A cross-section of the solenoid after two iterations

§10.4.3 Relation to Axiom A

Locally maximal hyperbolic sets are related to Smale's Axiom A.

Let $T : M \to M$ be a C^1 diffeomorphism of a smooth compact Riemannian manifold M. A point $x \in M$ is said to be *non-wandering* if for each open neighbourhood U of x there exists $n > 0$ such that $T^{-n}U \cap U \neq \emptyset$. The *non-wandering set* Ω of T is defined to be the set of all non-wandering points. Then Ω is a compact T-invariant subset of M and is the largest subset of M on which the dynamics of T is 'interesting' (in terms of studying its recurrence properties).

Definition. The diffeomorphism T is said to satisfy *Axiom A* if

(i) the tangent bundle restricted to Ω has a hyperbolic splitting;

(ii) the periodic points are dense in Ω.

The following theorem then allows us to decompose the dynamics of an Axiom A diffeomorphism into locally maximal hyperbolic sets.

Theorem 10.2 (Smale's spectral decomposition)
Let T be an Axiom A diffeomorphism with non-wandering set Ω. Then Ω

can be written as a disjoint union $\Omega = \Omega_1 \cup \cdots \cup \Omega_n$ *where each* Ω_i *is a basic set.*

Thus in defining locally maximal hyperbolic sets we re just abstracting the properties of the basic sets that appear in Smale's spectral decomposition of an Axiom A diffeomorphism restricted to its non-wandering set.

§10.5 Markov partitions

Let Λ be a locally maximal hyperbolic set and let $T : \Lambda \to \Lambda$ be hyperbolic. We assume that T is topologically mixing. We want to code the dynamics of T by using an aperiodic shift of finite type.

Let $x \in \Lambda$. We have already seen that there will be a stable manifold $W^s(x)$ and unstable manifold $W^u(x)$ passing through x. Typically, $W^s(x)$ and $W^u(x)$ will be dense in Λ. Instead we want to work with a small region of $W^s(x)$ that contains x. We define the *local stable* and *local unstable manifolds* to be the following:

$$\begin{aligned} W^s_\varepsilon(x) &= \{y \in M : d(T^n x, T^n y) \leq \varepsilon \text{ for all } n \geq 0\} \\ W^u_\varepsilon(x) &= \{y \in M : d(T^{-n} x, T^{-n} y) \leq \varepsilon \text{ for all } n \geq 0\}. \end{aligned}$$

One can check that if $y \in W^s_\varepsilon(x)$ then $d(T^n x, T^n y) \to 0$ exponentially fast as $n \to \infty$ (and similarly for $W^u_\varepsilon(x)$). Thus $W^s_\varepsilon(x)$ is a neighbourhood of x in $W^s(x)$; in particular $W^s_\varepsilon(x)$ is tangent to E^s_x. The corresponding statements are also true for $W^u_\varepsilon(x)$. (Here $\varepsilon > 0$ is chosen to be small—it is not necessary (at least here) to state precisely how small.)

If $x, y \in \Lambda$ are sufficiently close then we define their *product* to be

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

Provided that ε is sufficiently small, this intersection is a single point. Note that $[x, y]$ has the same forward asymptotic dynamics as y and the same backward asymptotic dynamics as x.

A subset $R \subset \Lambda$ is called a *rectangle* if $x, y \in R$ implies $[x, y] \in R$. We say that a rectangle R is a *proper rectangle* if it is equal to the closure of its interior (as subsets of Λ).

If R is a proper rectangle and $x \in \operatorname{int} R$ then we define $W^s(x, R) = W^s_\varepsilon(x) \cap R$ and $W^u(x, R) = W^u_\varepsilon(x) \cap R$. Throughout we assume that $\operatorname{diam} R$ is small in comparison with ε.

Definition. A finite collection $\mathcal{R} = \{R_1, \ldots, R_k\}$ of proper rectangles is called a *Markov partition* of Λ if $\bigcup_{j=1}^k R_j = \Lambda$, the interiors of the R_js are pairwise disjoint, and

- if $x \in \operatorname{int} R_i$ and $Tx \in \operatorname{int} R_j$ then $T(W^s_\varepsilon(x, R_i)) \subset W^s(Tx, R_j)$,

- if $x \in \operatorname{int} R_i$ and $T^{-1}x \in \operatorname{int} R_j$ then $T^{-1}(W^u_\varepsilon(x, R_i)) \subset W^u(Tx, R_j)$.

Thus, if $x \in \operatorname{int} R_i$ then the possible rectangles that Tx can lie in is determined only by R_i (and not by the sequence of rectangles that contain $T^{-n}x$). This is directly analogous to the definition of a shift of finite type, where the rules that say when symbol i can be followed by the symbol j in a sequence are determined by the symbol i, and not on the preceding symbols.

Proposition 10.3
Let $T : \Lambda \to \Lambda$ be a hyperbolic map on a basix set. Then there exists a Markov partition with an arbitrarily small diameter.

Proof. See the references. □

If we have a Markov partition $\mathcal{R} = \{R_1, \ldots, R_k\}$ then we introduce a two-sided shift of finite type Σ on k symbols with transition matrix $A_{i,j} = 1$ if and only if $T(\operatorname{int} R_i) \cap \operatorname{int} R_j \neq \emptyset$. We can code each point $x \in \Lambda$ by recording the sequence of elements of $\mathcal{R}$ that the orbit of x visits. Thus the map

$$\pi : \Sigma \to \Lambda : (x_j)_{j=-\infty}^{\infty} \mapsto \bigcap_{j=-\infty}^{\infty} T^{-j} R_{x_j}$$

is a semiconjugacy from the shift map on Σ to T. (The hyperbolicity of T ensures that this intersection consists of only one point.) Note that, as the elements of the Markov partition may overlap on their boundaries, there is some ambiguity as to how to code points whose orbit hits the boundary of $\mathcal{R}$; however this is a small set in both a measure-theoretic and topological sense.

Proposition 10.4
The map $\pi : \Sigma \to \Lambda$ is

(i) Hölder continuous (for some $\theta \in (0,1)$) and surjective,

(ii) injective on a set of full-measure (for any ergodic T-invariant measure of full support) and on a dense residual set,

(iii) bounded-to-one,

(iv) and π conjugates the dynamics of the shift $\sigma : \Sigma \to \Sigma$ to $T : \Lambda \to \Lambda$, i.e. $T\pi = \pi\sigma$.

§10.6 Markov partitions for Anosov automorphisms of tori

For a given hyperbolic map, there is no canonical Markov partition. In some cases a Markov partition can be identified either by inspection or by simple

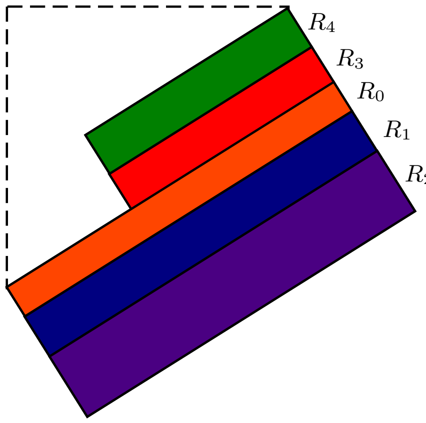

Figure 10.4: A Markov partition for the Cat map; the sides of the rectangles are parallel to the eigenvalues v_s, v_u

calculation. In the case of the cat map, a Markov partition can easily be found and is illustrated below.

However, for a general Anosov diffeomorphism, explicitly constructing a Markov partition is a non-trivial task. Indeed, even for Anosov automorphisms of tori, the rectangles that make up a Markov partition need not be rectangles in a geometrical sense. There are examples of Anosov automorphisms on the 3-dimensional torus for which the boundary of any Markov partition is nowehere differentiable.

§10.7 Ergodic theory of hyperbolic dynamical systems

We want to study some ergodic-theoretic properties of a hyperbolic map $T : \Lambda \to \Lambda$. To do this, we choose a Markov partition and code the dynamics as a shift of finite type. As T is invertible, this will be a two-sided shift of finite type. As thermodynamic formalism only works (at least in the way that we have presented it) for non-invertible maps, we need a method that allows us to pass from the two-sided shift to the one-sided.

We first consider invariant measures.

§10.7.1 Invariant measures and one- and two-sided shifts of finite type

Let A be an aperiodic $k \times k$ matrix with entries from $\{0, 1\}$. Then A determines a one-sided shift of finite type Σ^+ and a two-sided shift of finite type Σ. In both cases, let σ denote the shift map. Equip Σ^+ and Σ with their Borel σ-algebras, $\mathcal{B}^+, \mathcal{B}$, respectively.

Let μ be a σ-invariant probability measure on $(\Sigma^+, \mathcal{B}^+)$. We show how μ can be extended to give an invariant probability measure on $(\Sigma, \mathcal{B})$.

Let $C = [i_0, \ldots, i_{k-1}]_m = \{x = (x_j)_{j=-\infty}^{\infty} \mid x_{j+m} = i_j, 0 \leq j \leq k-1\}$ be a cylinder where we fix the symbols that appear in places m to $m+k-1$. It follows that $\sigma^{-m}C$ depends only on positive indices, and so can be regarded as a subset of Σ^+. Define $\mu(C) = \mu(\sigma^{-n}C)$ where n is chosen so that $\sigma^{-n}C$ depends only on future co-ordinates. As μ is σ-invariant, this is well-defined and independent of the choice of n.

Thus we have defined a measure on the algebra of cylinders in Σ. By the Kolmogorov Extension Theorem, this then extends to a measure defined on $\mathcal{B}$. As μ is σ-invariant on cylinders, it again follows from the Kolmogorov Extension Theorem that μ is a σ-invariant Borel probability measure defined on $(\Sigma, \mathcal{B})$.

§10.7.2 Functions of the future

Let $\theta \in (0, 1)$. We can define metrics on both Σ and Σ^+ as follows. For the one-sided shift of finite type Σ^+ we define $n(x, y) = \sup\{j \mid x_j = y_j, 0 \leq j \leq n-1\}$. For the two-sided shift of finite type Σ we define $n(x, y) = \sup\{j \mid x_j = y_j, -(n-1) \leq j \leq n-1\}$. (In both cases we define $n(x, y) = 0$ if $x = y$.) We then define a metric on the respective spaces by setting $d_\theta(x, y) = \theta^{n(x,y)}$. (With a small abuse of notation we use d_θ to denote a metric on both the one-sided and two-sided shifts.)

Let $F_\theta(\Sigma, \mathbb{R})$ and $F_\theta(\Sigma^+, \mathbb{R})$ denote the space of functions defined on Σ, Σ^+, respectively, that are θ-Holder with respect to d_θ. That is, $F_\theta(\Sigma, \mathbb{R})$ consists of functions $f : \Sigma \to \mathbb{R}$ such that

$$|f|_\theta = \sup_{x,y\in\Sigma, x\neq y} \frac{|f(x) - f(y)|}{d_\theta(x, y)} < \infty,$$

and similarly for $F_\theta(\Sigma^+, \mathbb{R})$.

Given a point $x = (x_j)_{j=-\infty}^{\infty} \in \Sigma$, we regard $(x_j)_{j=0}^{\infty}$ as 'the future' and $(x_j)_{j=-\infty}^{0}$ as 'the past'.

If $f \in F_\theta(\Sigma, \mathbb{R})$ then $f(x)$ will typically depend on both the future and the past of x. However, if $f(x)$ depends only on $x_0, x_1, x_2, \ldots$, i.e. f depends only on the future of x, then f can be regarded as a function defined on $\Sigma+$.

Recall that two functions f, g are said to be *cohomologous* if there exists a function u such that $f = g + u\sigma - u$.

The following proposition shows that any Hölder function on Σ is cohomologous to a function that depends only on future co-ordinates. Note that to achieve this we may have to increase the Hölder exponent from θ to $\theta^{1/2}$.

Proposition 10.5
Let $f \in F_\theta(\Sigma, \mathbb{R})$. Then there exists $u, g \in F_{\theta^{1/2}}(\Sigma^+, \mathbb{R})$ such that $f = g + u\sigma - u$ and g depends only on future co-ordinates.

Proof. For each symbol k choose an 'allowable past', that is, a sequence $(i_j^{(k)})_{j=-\infty}^0$ such that $i_0^{(k)} = k$ and $A_{i_j^{(k)}, i_{j+1}^{(k)}} = 1$ for all $j < 0$. Define a map $\phi : \Sigma \to \Sigma$ by

$$(\phi(x))_j = \begin{cases} x_j & \text{if } j \geq 0 \\ i_j^{(k)} & \text{if } j \leq 0 \text{ and } x_0 = k. \end{cases}$$

Thus $\phi(x)$ takes a sequence x and replaces its past by the past chosen above, determined by x_0.

Define

$$u(x) = \sum_{n=0}^{\infty} f(\sigma^n(x)) - f(\sigma^n(\phi(x))).$$

Note that $\sigma^n(x), \sigma^n(\phi(x))$ agree in at least places $-n \leq j < \infty$. Hence

$$|f(\sigma^n(x)) - f(\sigma^n(\phi(x)))| \leq |f|_\theta d_\theta(\sigma^n(x), \sigma^n(\phi(x))) \leq |f|_\theta \theta^n,$$

and it follows that $u(x)$ is well-defined.

Note that

$$\begin{aligned} & u(x) - u(\sigma(x)) \\ = \ & \sum_{n=0}^{\infty} f(\sigma^n(x)) - f(\sigma^n(\phi(x))) - \sum_{n=0}^{\infty} f(\sigma^{n+1}(x)) - f(\sigma^n \phi(\sigma(x))) \\ = \ & f(x) - \left(f(\phi(x)) + \sum_{n=0}^{\infty} f(\sigma^{n+1}(\phi(x))) - f(\sigma^n(\phi(\sigma(x)))) \right). \quad (10.2) \end{aligned}$$

Denote the bracketed term in (10.2) by $g(x)$ and notice that $g(x)$ depends only on the future co-ordinates.

It remains to show that u, and therefore g, belongs to $F_{\theta^{1/2}}(\Sigma, \mathbb{R})$. Let $x = (x_j)_{j=0}^{\infty}, y = (y_j)_{j=0}^{\infty} \in \Sigma$ be such that $x_j = y_j$ for $-2N \leq j \leq 2N$. Then for $0 \leq n \leq N$

$$|f(\sigma^n(x)) - f(\sigma^n(y))|, |f(\sigma^n(\phi(x))) - f(\sigma^n(\phi(y)))| \leq |f|_\theta \theta^{2N-n}.$$

Moreover, for all $n \geq 0$ we have

$$|f(\sigma^n(x)) - f(\sigma^n(\phi(x)))|, |f(\sigma^n(x)) - f(\sigma^n(\phi(x)))| \leq |f|_\theta \theta^n.$$

Hence

$$\begin{aligned} |u(x) - u(y)| & \leq 2|f|_\theta \sum_{n=0}^{N} \theta^{2N-n} + 2|f|_\theta \sum_{n=N+1}^{\infty} \theta^n \\ & = 2|f|_\theta \theta^{2N} \left(\frac{\theta^{-N-1} - 1}{\theta^{-1} - 1} \right) + 2|f|_\theta \frac{\theta^{N+1}}{1-\theta} \\ & \leq 4|f|_\theta \frac{\theta^N}{1-\theta}. \end{aligned}$$

It follows that $u \in F_{\theta^{1/2}}(\Sigma, \mathbb{R})$. $\square$

§10.8 Applications

We give some applications of how to use thermodynamic formalism to prove results about hyperbolic maps.

§10.8.1 Existence of equilibrium states

Let X be a compact metric space equipped with the Borel σ-algebra and let $T : X \to X$ be a continuous transformation.

Definition. Let $f : X \to \mathbb{R}$ be measurable. We say that a T-invariant probability measure μ_f is an *equilibrium state* for f if

$$h_{\mu_f}(T) + \int f\, d\mu_f = \sup h_\mu(T) + \int f\, d\mu \tag{10.3}$$

where the supremum is taken over all T-invariant Borel probability measures. We denote the supremum in (10.3) by $P(f)$ (or $P_T(f)$ if we wish to indicate the dependence on T) and call it the *pressure* of f.

This allows us to extend the definition of pressure, which was previously only defined for Hölder functions on shifts of finite type, to continuous functions on compact metric spaces.

It is natural to ask: (i) for which class of function f is there an equilibrium state, and (ii) if an equilibrium state exists, is it necessarily unique? We have already seen in Lecture 8 that, in the setting of a one-sided aperiodic shift of finite type, each Hölder continuous function has a unique equilibrium state. We can use symbolic dynamics to prove the same result for hyperbolic systems.

For convenience we first record the result for two-sided shifts of finite type.

Lemma 10.6
Let $\sigma : \Sigma \to \Sigma$ be an aperiodic shift of finite type and let $f \in F_\theta(\Sigma, \mathbb{R})$. Then f has a unique equilibrium state.

Proof. Choose $u \in F_{\theta^{1/2}}(\Sigma, \mathbb{R})$ and $g \in F_{\theta^{1/2}}(\Sigma^+, \mathbb{R})$ such that $f = u\sigma - u + g$. Then g has a unique equilibrium state; this is a σ-invariant probability measure on Σ^+ which we can extend to a σ-invariant probability measure μ_f on Σ. From §10.7.1 it follows that σ-invariant Borel probability measure on the two-sided shift of finite type Σ are in one-to-one correspondence with σ-invariant Borel probability measures on the one-sided shift of finite type Σ^+. Noting that $\int f\, d\mu = \int g\, d\mu$ for any σ-invariant Borel probability measure on Σ, it follows that

$$h_{\mu_f}(\sigma) + \int \hat{f}\, d\mu_f = \sup h_\mu(\sigma) + \int \hat{f}, d\mu$$

where the supremum is taken over all σ-invariant Borel probability measures on Σ, and μ_f is the only such measure that achieves this supremum. $\square$

We now prove the result for a hyperbolic map $T : \Lambda \to \Lambda$ on a hyperbolic basic set Λ. We assume, without loss of generality, that T is topologically mixing.

Proposition 10.7
Let $T : \Lambda \to \Lambda$ be a hyperbolic map restricted to a basic set Λ. Let $f : \Lambda \to \mathbb{R}$ be Hölder continuous. Then there exists a unique equilibrium state for f.

Proof. Choose a Markov partition and coding map $\pi : \Sigma \to \Lambda$, for an appropriate two-sided aperiodic shift of finite type Σ. Note that π is Hölder continuous. Hence $\hat{f} = f \circ \pi \in F_\theta(\Sigma, \mathbb{R})$ for an appropriate $\theta \in (0, 1)$. Let $\hat{\mu}_f$ denote the unique equilibrium state for $\hat{f}$.

Let R be an element of the Markov partition. Let ∂R denote the boundary of R as a subset of Λ and decompose $\partial R = \partial^s R \cup \partial^u R$ where

$$\begin{aligned} \partial^s R &= \{x \in R \mid x \notin \operatorname{int} W^u(x, R)\} \\ \partial^u R &= \{x \in R \mid x \notin \operatorname{int} W^s(x, R)\}. \end{aligned}$$

(The geometric intuition is that $\partial^s R$ are the 'edges' of R in the stable direction, and $\partial^u R$ are the 'edges' of R in the unstable direction.) Denote $\partial^s \mathcal{R} = \bigcup_{R \in \mathcal{R}} \partial^s R$ and $\partial^u \mathcal{R} = \bigcup_{R \in \mathcal{R}} \partial^u R$. Let $D_s = \pi^{-1}(\partial^s \mathcal{R})$ and $D_u = \pi^{-1}(\partial^u \mathcal{R})$. Then D_s, Du are non-empty strict closed subsets of Σ. Moreover, $\sigma D_s \subset D_s$ and $\sigma^{-1} D_u \subset D_u$. As $\hat{\mu}_f$ is σ-invariant, it follows that $\hat{\mu}_f(\sigma^n D_s) = \hat{\mu}_f(D_s)$. As $\sigma^{n+1} D_s \subset \sigma^n D_s$, it follows that

$$\hat{\mu}_f \left(\bigcap_{n=0}^{\infty} \sigma^n D_s \right) = \hat{\mu}_f(D_s).$$

As $\bigcap_{n=0}^{\infty} \sigma^n D_s$ is σ-invariant and $\hat{\mu}_f$ is ergodic, this intersection has measure either 0 or 1. Hence $\hat{\mu}_f(D_s) = 0$ or 1. As the complement of $D_s \subset \Sigma$ is a non-empty open subset, it follows that $\hat{\mu}_f(D_s)$ must be 0. Similarly, $\hat{\mu}_f(D_u) = 0$.

Define $\mu_f = \pi^* \hat{\mu}_f$ (so that $\mu_f(B) = \hat{\mu}_f(\pi^{-1} B)$). Then μ_f is a T-invariant probability measure defined on Λ. The above discussion shows that $\pi : \Sigma \to \Lambda$ is a measure-theoretic isomorphism between between σ (with respect to $\hat{\mu}_f$) and T (with respect to μ_f). We will show that μ_f is the unique equilibrium state for f.

As σ and T are measure-theoretically isomorphic (with respect to the measures $\hat{\mu}_f$ and μ_f, respectively), it follows that $h_{\hat{\mu}_f}(\sigma) = h_{\mu_f}(T)$. Hence

$$\begin{aligned} h_{\mu_f}(T) + \int f \, d\mu_f &= h_{\hat{\mu}_f}(\sigma) + \int \hat{f} \, d\hat{\mu}_f \\ &= P_\sigma(\hat{f}). \end{aligned}$$

The following result is easily proved:

Lemma 10.8
Let $T_j : X_j \to X_j$, $j = 1, 2$, be continuous transformations of compact metric spaces. Suppose that T_2 is a factor of T_1, i.e. there exists a continuous surjection $\pi : X_1 \to X_2$ such that $T_2\pi = \pi T_1$. Let $f : X_2 \to \mathbb{R}$ be continuous. Then $P_{T_1}(f\pi) \geq P_{T_2}(f)$.

From this lemma it follows that $P_\sigma(\hat{f}) \geq P_T(f)$. Hence

$$\begin{aligned} & h_{\mu_f}(T) + \int f \, d\mu_f \\ \geq \ & P_T(f) \\ = \ & \sup h_\mu(T) + \int f \, d\mu \end{aligned}$$

where the infimum is taken over all T-invariant Borel probability measures. Hence μ_f achieves this supremum and so is an equilibrium state for f.

Finally, we show that μ_f is the unique equilibrium state for f, i.e. μ_f is the only T-invariant probability measure that achieves the supremum in (10.3). We need the following lemma:

Lemma 10.9
Let $\mu \in M(\Lambda, T)$ be a T-invariant Borel probability measure. Then there exists a σ-invariant Borel probability measure $\nu \in M(\Sigma, \sigma)$ such that $\pi^*\nu = \mu$.

Let μ be any equilibrium state for f. Choose ν as in the lemma so that $\pi^*\nu = \mu$. Then $h_\nu(\sigma) \geq h_\mu(T)$. Hence

$$\begin{aligned} h_\nu(\sigma) + \int \hat{f} \, d\nu &\geq h_\mu(f) + \int f \, d\mu \\ &= P_T(f) \\ &= P_\sigma(\hat{f}). \end{aligned}$$

Hence ν is an equilibrium state for $\hat{f}$. As we know that Hölder functions defined on shifts of finite type have a unique equilibrium state, it follows that $\nu = \hat{\mu}_f$. Hence $\mu = \pi^*\nu = \mu_f$, and so f has a unique equilibrium state. □

§10.8.2 SRB measures

As we have seen, for a given dynamical system there may be many different ergodic measures. If the dynamical system is defined on a space with some additional structure, say a Riemannian manifold, then we may want to pick out ergodic measures that are related to this structure. We begin with the following observation.

Let X be a compact metric space equipped with the Borel σ-algebra $\mathcal{B}$ and let $T : X \to X$ be continuous. Let μ be an ergodic Borel probability measure for T. Recall that $C(X,\mathbb{R})$ is separable and choose a countable dense subset $\{f_i\} \subset C(X,\mathbb{R})$. Then for each i there exists a set $N_i \in \mathcal{B}$ with $\mu(N_i) = 0$ such that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} f_i(T^j x) \to \int f_i \, d\mu \tag{10.4}$$

for all $x \notin N_i$. Let $N = \bigcup_{i=1}^{\infty} N_i$ and note that $\mu(N) = 0$. Clearly (10.4) holds for each i, for all $x \in N$. By approximating an arbitrary continuous function $f \in C(X,\mathbb{R})$ by functions of the form f_i it is easy to see that

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

Thus, for continuous transformations of compact metric spaces and continuous observables, the set of measure zero for which Birkhoff's Ergodic Theorem fails can be chosen to be independent of the observation f.

Let M be a compact Riemannian manifold equipped with the Riemannian volume m. Let $T : M \to M$ be a diffeomorphism of M and let $f : M \to \mathbb{R}$ be continuous. We are interested in understanding the limit

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n-1} f(T^j x) \tag{10.5}$$

for m-a.e. $x \in M$. Typically the Riemannian volume will not be T-invariant; even if it is invariant then it need not be ergodic. Hence there is no reason to assume that (10.5) converges, or if it does, that it converges to $\int f \, dm$. However, the Riemannian volume is a distinguished measure in the sense that we view m-almost every point as being 'typical'.

Suppose that $T : M \to M$ contains a locally maximal (not necessarily hyperbolic) attractor. That is, there exists a T-invariant subset $\Lambda \subset M$ and an open set $U \supset \Lambda$ such that $\cap_{n=0}^{\infty} T^n U = \Lambda$. Given Λ, we call the largest such U the *basin of attraction* of Λ and denote it by $B(\Lambda)$. Thus the basin of attraction $B(\Lambda)$ of Λ is the set of all points whose orbits accumulate on Λ under forward iteration by T. As $B(\Lambda)$ is an open set, it has positive measure with respect to the Riemannian volume. Now consider ergodic averages of the form (10.5) for m-almost every point of $B(\Lambda)$; it is natural to expect (10.5) to converge to $\int f \, d\mu$ for some measure μ supported on the attractor Λ.

Definition. Let $T : M \to M$ be a diffeomorphism of a compact Riemannian manifold M with Riemannian volume m. Suppose Λ is an attractor for T with basin $B(\Lambda)$. We say that a probability measure μ (necessarily supported on Λ) is an *SRB measure*, or *Sinai-Ruelle-Bowen measure*, if

$$\frac{1}{n}\sum_{j=0}^{n-1} f(T^j x) \to \int f \, d\mu$$

for m-almost every point $x \in B(\Lambda)$ and for every continuous function $f : M \to \mathbb{R}$.

Example. Let $X \subset \mathbb{R}^2$ denote the circle of radius 1 centred at $(0,1)$ in $\mathbb{R}^2$. Call $(0,2)$ the North Pole (N) and $(0,0)$ the South Pole (S) of X. The Riemannian volume m is Lebesgue measure.

Recall that we defined the North-South map $T : X \to X$ as follows. Define $\phi : X \setminus \{N\} \to \mathbb{R} \times \{0\}$ by drawing a straight line through N and x and denoting by $\phi(x)$ the unique point on the x-axis that this line crosses (this is just stereographic projection of the circle). Now define $T : X \to X$ by

$$T(x) = \begin{cases} \phi^{-1}\left(\frac{1}{2}\phi(x)\right) & \text{if } x \in X \setminus \{N\}, \\ N & \text{if } x = N. \end{cases}$$

Hence $T(N) = N$, $T(S) = S$ and if $x \neq N$ then $T^n(x) \to S$ as $n \to \infty$. Thus $\Lambda = \{S\}$ is an attractor with basin $B(\Lambda) = X \setminus \{N\}$. Note that

Thus if $x \neq N$ and $f : X \to \mathbb{R}$ is continuous then we have

$$\frac{1}{n}\sum_{j=0}^{n-1} f(T^j x) \to f(S) = \int f \, d\delta_S.$$

Thus in this case, for m-almost every point x (10.5) converges to $\int f \, d\delta_S$ for every continuous function f. Hence the Dirac measure supported at the south pole is an SRB measure.

It is natural to ask when an SRB measure exists, and if it exists if it is unique.

Proposition 10.10
Let $T : M \to M$ be a $C^{1+\alpha}$ diffeomorphism of a compact Riemannian manifold M. Suppose that $\Lambda \subset M$ is a locally maximal hyperbolic attractor for T. Then there is a unique SRB measure supported on Λ and it is the equilibrium state of the Hölder continuous function $-\log DT|_{E^u}$.

Proof. See the references for the (lengthy) proof; the idea is as follows. Recall that for a shift of finite type σ and Hölder function f, the equilibrium state μ_f of f has an alternative characterisation as a Gibbs measure: there exist constants $C_1, C_2 > 0$ such that

$$C_1 \leq \frac{\mu f[x_0, \ldots, x_{n-1}]}{e^{\sum_{j=0}^{n-1} f(\sigma^j x) - nP(f)}} \leq C_2,$$

that is, the measure of a cylinder is approximated by the ergodic sum of f of a point in that cylinder (subject to a correction by the pressure $P(f)$ to eliminate the potential linear growth in this sum).

Let $T : \Lambda \to \Lambda$ be a hyperbolic map on a basic set Λ and choose a Markov partition $\mathcal{R} = \{R_1, \ldots, R_k\}$. The natural analogue of a cylinder is a set of the form

$$[x_0, \ldots, x_{n-1}] = R_{x_0} \cap T^{-1}R_{x_1} \cap \cdots \cap T^{-(n-1)}R_{x_{n-1}}.$$

Let m denote the Riemannian volume. By the change of variables formula, it seems reasonable to expect that

$$m[x_0, \ldots, x_{n-1}] \sim DT^n(x)^{-1}m(R_{x_{n-1}}) \sim Ce^{-\sum_{j=0}^{n-1} \log DT|_{E^u}(T^j x)}.$$

This indicates why it is natural to expect $-\log DT|_{E^u}$ to appear; the (many and lengthy) details are in the references. □

§10.9 References

Anosov diffeomorphisms were first discussed in

D. V. Anosov, *Geodesic Flows on Closed Riemannian Manifolds with Negative Curvature*, Proc. Steklov Inst., vol. 90, Amer. Math. Soc., Prov., Rhode Isl., 1969

as discrete time analogues of Anosov flows. Anosov flows were introduced as a generalisation of the geodesic flows on a compact Riemannian manifold with negative sectional curvatures.

The definition of Axiom A, and its abstraction to hyperbolic maps on basic sets, was first developed in

S. Smale, *Differentiable dynamical systems*, Bull. Amer. Math. Soc. **73** (1973), 747–817,

which also discusses in detail many of the examples given in this lecture. In particular, an Anosov automorphism of a nilmanifold (that is not a torus) is constructed. A particularly readable treatment that includes the examples discussed above can be found in

A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Encyclopædia of Math., vol. 54, C.U.P., Cambridge, 1995.

The construction of Markov partitions at this level of generality goes back to

R. Bowen, *Markov partitions for axiom A diffeomorphisms*, Amer. J. Math. **92**, 725–747

(although the idea for symbolically coding a dynamical system goes back to at least 1934 (Hedlund's proof that the geodesic flow on a surface of constant negative curvature is ergodic with respect to the Liouville measure)). The material on equilibrium states and SRB measures can be found in

R. Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Lecture Notes in Math., vol. 470, Springer, Berlin, 1975.

§10.10 Exercises

Exercise 10.1
Show that the solenoid $T : \Lambda \to \Lambda$ is topologically conjugate to an automorphism of a compact abelian group. (Hint: consider 2-adic numbers.)

Exercise 10.2
Let $T : \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ be a continuous map on the unit circle with $\sup |T'(x)| \geq r > 1$ for some r. We regard the circle as being equal to $[0, 1]$ with the end-points identified. Suppose there exist points $0 = a_0 \leq a_1 \leq \cdots a_{n-1} \leq a_n = 1$ such that (denoting $[a_j, a_{j+1}]$ by R_j) for each j, $T(R_j)$ is a union of sets of the form R_i. Then we call T a Markov map of the interval.

(The doubling map $Tx = 2x \bmod 1$ is an example: take $a_0 = 0, a_1 = 1/2, a_2 = 1$ and let $R_0 = [0, 1/2]$ and $R_1 = [1/2, 1]$. Then $T(R_0) = R_0 \cup R_1 = T(R_1)$.)

(i) *Define a $0-1$ matrix A by $A_{i,j} = 1$ if and only if $R_j \subset T(R_i)$ and let Σ denote the corresponding one-sided shift of finite type. Show that there exists a continuous surjective map $\pi : \Sigma \to \mathbb{R}/\mathbb{Z}$ defined by*

$$\pi(x_0, x_1, \cdots) = \bigcap_{j=0}^{\infty} T^{-j} R_{x_j}.$$

Show that $T\pi = \pi\sigma$. Show that π is injective except on a countable set.

(ii) *Suppose that the matrix A is aperiodic. Let $f : [0, 1] \to \mathbb{R}$ be Hölder. Show that f has a unique equilibrium state.*

Exercise 10.3
Prove Lemmas 10.8 and 10.9.

(Hint for Lemma 10.9: Let $\mu \in M(\Lambda, T)$ and define a continuous linear functional $w(f\pi) = \int f \, d\mu$ on the subspace of continuous functions in $C(\Sigma, \mathbb{R})$ of the form $f\pi$ for some $f \in C(\Lambda, T)$. Use the Hahn-Banach theorem to extend this to a functional on all of $C(\Sigma, \mathbb{R})$ and then use the Hahn-Banach theorem to find a measure ν on Σ such that $\pi^*\nu = \mu$. Consider weak-$*$ limits of $n^{-1} \sum_{j=0}^{n-1} \sigma_*^j \nu$ to find a suitable invariant measure.