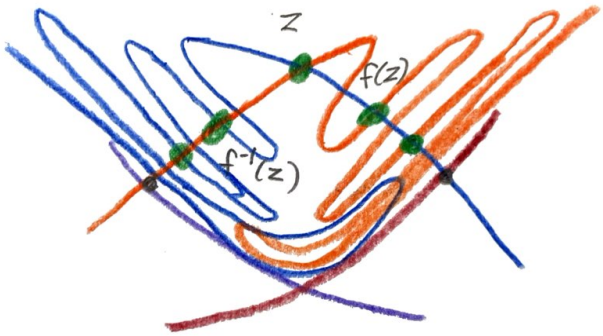
Ergodic Theory and Hyperbolic Dynamics

July 24, 2025 – version 0.1

Math 7326 course notes, Univ. of Houston, Spring 2025



Vaughn Climenhaga

This material is based upon work supported by the National Science Foundation under Award No. DMS-1554794 and DMS-2154378.

This work was supported by a grant from the Simons Foundation (915869, Climenhaga).

Contents

Preface

This document represents course notes from Math 7326, Dynamical Systems, in Spring 2025 at the University of Houston. I plan to eventually develop these notes into a book, which will include topics beyond those presented here: thus this document should be treated very much as a work in progress, rather than as a final static product. Since this is the first version that is developed enough to post on my website, I am labeling it as version 0.1. Future versions will be incremented to help distinguish them. All theorem and equation numbers are subject to change in future versions, so exercise due caution if you decide to cite anything from this document.

I welcome comments, corrections, and questions: this text undoubtedly contains errors, omissions, and unclear explanations that could be improved.

The exercises are of varying degrees of difficulty, which I have made some (rather hasty) attempt to assess and to indicate by adding one marking if the exercise is straightforward, two if it requires some more thought or a bit more calculation but should not pose undue difficulties, and three if it requires a more substantial amount of strategizing to see your way through.

Eventually, this preface will include some discussion of how this book compares with existing books on hyperbolic dynamics and ergodic theory, and why I have begun the process of adding another book to this rich body of literature. For now, I simply list some (rather large but not systematically chosen) subset of the extant literature, beginning with some fairly comprehensive introductions to the overall theory that cover both hyperbolic dynamics and ergodic theory, although they do not go as deeply into the theory of thermodynamic formalism that lies at the interface of these topics.

- Anatole Katok and Boris Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*. Cambridge University Press, 1995.
- Todd Fisher and Boris Hasselblatt, *Hyperbolic Flows*. European Mathematical Society, 2019.

Shorter general introductions to both dynamical systems and ergodic theory at the graduate level include:

- Michael Brin and Garrett Stuck, *Introduction to Dynamical Systems*. Cambridge University Press, 2002.
- Yves Coudène, *Ergodic Theory and Dynamical Systems*. Springer, 2013.

There are many texts in ergodic theory that do not dwell on hyperbolic dynamics. These include:

- Peter Walters, *An Introduction to Ergodic Theory*. Springer, 1982.
- Karl Petersen, *Ergodic Theory*. Cambridge University Press, 1983.
- Manfred Einsiedler and Thomas Ward, *Ergodic Theory: with a view towards Number Theory*. Springer, 2011.
- Eli Glasner, *Ergodic Theory via Joinings*. American Mathematical Society, 2003.

The following treat in more depth the interface between ergodic theory and hyperbolic dynamics, including some of the key ideas in thermodynamic formalism.

- Rufus Bowen, *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*. Springer, 1975. (2nd revised edition, 2008.)
- David Ruelle, *Thermodynamic Formalism*. 1978. (2nd edition, Cambridge University Press, 2004.)
- Manfred Denker, Christian Grillenberger, and Karl Sigmund, *Ergodic Theory on Compact Spaces*. Springer, 1976.
- Ricardo Mañé, *Ergodic Theory and Differentiable Dynamics*. Springer–Verlag, 1983.
- William Parry and Mark Pollicott, *Zeta Functions and the Periodic Orbit Structure of Hyperbolic Dynamics*. Société mathématique de France, Astérisque, 1990.
- Gerhard Keller, *Equilibrium States in Ergodic Theory*. Cambridge University Press, 1998.
- Mark Pollicott and Michiko Yuri, *Dynamical Systems and Ergodic Theory*. Cambridge University Press, 1998.
- Viviane Baladi, *Positive Transfer Operators and Decay of Correlations*. World Scientific, 2000.
- Marcelo Viana and Krerley Oliveira, *Foundations of Ergodic Theory*. Cambridge University Press, 2016.
- Mariusz Urbański, Mario Roy, and Sara Munday. *Non-invertible Dynamical Systems, Volume 1: Ergodic Theory – Finite and Infinite, Thermodynamic Formalism, Symbolic Dynamics and Distance Expanding Maps*. De Gruyter, 2022.

In this same vein the following book should be mentioned – it includes an analysis of the Chirikov–Taylor standard map.

- Yakov G. Sinai, *Topics in Ergodic Theory*. Princeton University Press, 1994.

Some introductory reading suitable for an undergraduate audience include the following: these are less comprehensive and in particular do not include any ergodic theory, with the exception of Silva's book, which introduces the necessary measure theory from scratch.

- Edward Ott, *Chaos in Dynamical Systems*. Cambridge University Press, 1993.
- Steven H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Westview Press, 1994.
- Boris Hasselblatt and Anatole Katok, *A First Course in Dynamics: With a Panorama of Recent Developments*. Cambridge University Press, 2003.

- Yakov Pesin and Vaughn Climenhaga, *Lectures on Fractal Geometry and Dynamical Systems*. American Mathematical Society, Student Mathematical Library **52**, 2009.
- Diana Davis, Bryce Weaver, Roland K.W. Roeder, and Pablo Lessa, *Dynamics Done with Your Bare Hands*. European Mathematical Society, 2016.
- Cesar E. Silva, *Invitation to Ergodic Theory*. American Mathematical Society, Student Mathematical Library **42**, 2018.

A detailed study of the Lorenz flow can be found in:

- Colin Sparrow, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*. Springer–Verlag, 1982.

The following books focus specifically on symbolic dynamics, including ergodic theory in that setting.

- Douglas Lind and Brian Marcus, *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, 1996.
- Bruce P. Kitchens, *Symbolic Dynamics: One-sided, Two-sided and Countable State Markov Shifts*. Springer, 1998.
- Henk Bruin, *Topological and Ergodic Theory of Symbolic Dynamics*. American Mathematical Society, Graduate Studies in Mathematics **228**, 2022.

For the general theory of nonuniform hyperbolicity ("Pesin theory"), see:

- Luis Barreira and Yakov Pesin, *Nonuniform Hyperbolicity: Dynamics of Systems with Nonzero Lyapunov Exponents*. Cambridge University Press, 2007.
- Anatole Katok and Jean-Marie Strelcyn. *Invariant Manifolds, Entropy and Billiards; Smooth Maps with Singularities*. Springer-Verlag, Lecture Notes in Mathematics **1222**, 1986.
- Luis Barreira and Yakov Pesin, *Introduction to Smooth Ergodic Theory*. American Mathematical Society, Graduate Studies in Mathematics **148**, 2013.
- Mark Pollicott, *Lectures on Ergodic Theory and Pesin Theory on Compact Manifolds*. London Mathematical Society, Lecture Note Series **180**, 1993.

References for fractal geometry include:

- Kenneth Falconer, *The Geometry of Fractal Sets*. Cambridge University Press, 1985.
- Kenneth Falconer, *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, 1989.

For more in-depth introductions to the qualitative study of dynamical systems, including homoclinic and heteroclinic points and their consequences, see:

- Jacob Palis and Welington de Melo, *Geometric Theory of Dynamical Systems: An Introduction*. Springer–Verlag, 1982.
- Michael Shub, *Global Stability of Dynamical Systems*. Springer–Verlag, 1987.
- Jacob Palis and Floris Takens, *Hyperbolicity and sensitive chaotic dynamics at homoclinic bifurcations*. Cambridge University Press, 1993.

- Gerald Teschl, *Ordinary Differential Equations and Dynamical Systems*. American Mathematical Society, 2012.

There is a vast literature treating the qualitative study of dynamical systems from a more applied point of view; here I mention only a few examples.

- John Guckenheimer and Philip Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer–Verlag, 1983.
- Christian Beck and Friedrich Schlögl, *Thermodynamics of Chaotic Systems*. Cambridge University Press, 1993.
- Robert Gilmore and Marc Lefranc, *The Topology of Chaos: Alice in Stretch and Squeezeland*. Wiley, 2002.
- Henk Broer and Floris Takens, *Dynamical Systems and Chaos*. Springer, 2011.

Standard references for the use of Ornstein's $\bar{d}$-metric in ergodic theory include:

- Donald S. Ornstein, *Ergodic Theory, Randomness, and Dynamical Systems*. Yale Mathematical Monographs, 1974.
- Daniel J. Rudolph, *Fundamentals of Measurable Dynamics: Ergodic Theory on Lebesgue Spaces*. Oxford University Press, 1990.

The following classical references for ergodic theory can provide more in-depth details; they vary in their perspective and motivation.

- Paul R. Halmos, *Lectures on Ergodic Theory*. American Mathematical Society, 1956.
- V.I. Arnold and A. Avez, *Ergodic Problems of Classical Mechanics*. W.A. Benjamin, Inc., 1968.
- Yakov G. Sinai, *Introduction to Ergodic Theory*. Princeton University Press, 1977.
- William Parry, *Topics in Ergodic Theory*. Cambridge University Press, 1981.
- I.P. Cornfeld, S.V. Fomin, and Yakov Sinai, *Ergodic Theory*. Springer-Verlag, 1982.

The following texts develop ergodic theory and related ideas in particular directions that we will not fully explore in this book.

- Yakov B. Pesin, *Dimension Theory in Dynamical Systems: Contemporary Views and Applications*. University of Chicago Press, 1997.
- Nikolai Chernov and Roberto Markarian, *Chaotic Billiards*. American Mathematical Society, 2006.
- Tanja Eisner, Bálint Farkas, Markus Haase, and Rainer Nagel, *Operator Theoretic Aspects of Ergodic Theory*. Springer, 2015.
- David Damanik and Jake Fillman, *One-Dimensional Ergodic Schrödinger Operators: I. General Theory*. American Mathematical Society, Graduate Studies in Mathematics **221**, 2022.

The theory of one-dimensional dynamical systems is described in:

- Welington de Melo and Sebastian van Strien, *One-Dimensional Dynamics*. Springer, 1993.

- Edson de Faria and Welington de Melo, *Mathematical Tools for One-Dimensional Dynamics*. Cambridge University Press, 2008.
- Lluís Alsedà, Jaume Llibre, and MichałMisiurewicz, *Combinatorial Dynamics and Entropy in Dimension One*. World Scientific, 2000.
- Pierre Collet and Jean-Pierre Eckmann, *Iterated Maps on the Interval as Dynamical Systems*. Birkhäuser, 1980.

Finally, there are a number of books aimed at a wider audience, which can provide a broad perspective on the mathematics, its applications, and the historical context.

- Benoit B. Mandelbrot, *The Fractal Geometry of Nature*. W.H. Freeman and Company, 1977.
- James Gleick, *Chaos: Making a New Science*. Penguin Books, 1987.
- Edward N. Lorenz, *The Essence of Chaos*. University of Washington Press, 1993.
- Florin Diacu and Philip Holmes, *Celestial Encounters: The Origins of Chaos and Stability*. Princeton University Press, 1996.
- June Barrow-Green, *Poincaré and the Three Body Problem*. American Mathematical Society, 1997.
- H. Scott Dumas, *The KAM Story: A Friendly Introduction to the Content, History, and Significance of Classical Kolmogorov–Arnold–Moser Theory*. World Scientific, 2014.

CHAPTER 1

Setting the stage

1.1. Determinism and randomness

Lec 1
Mon, Jan 13

Imagine rolling a pair of dice. The outcome is random: this is accepted in board games, in casinos, in probability classes, and in our everyday speech. But how can this be? The dice fly and bounce according to the deterministic laws of physics; where does the randomness come from? If the final outcome depends only on the initial conditions and nothing else, why can we not predict how the dice will land?

The motion of the dice has two parts: flying in the air and bouncing on the table. Let us consider each in turn.

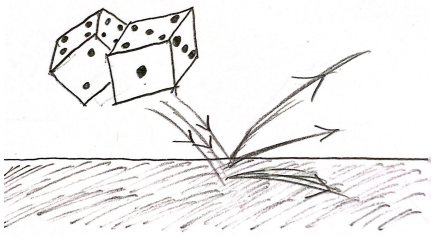


FIGURE 1.1. Rolling the dice.

While the dice fly through the air, we can accurately predict where they will hit the table, just as if we were watching a ball, and how they will be oriented. Even if our information about the initial position, velocity, and rotation is not perfect – which will always be the case – that measurement error only leads to a small error in our prediction of where and how the flying dice first hit the table.

Things change when the dice bounce. Because of the sharp angles at the edges and corners, the direction of the bounce is highly sensitive to the exact angle at which the dice hit the table, as is the rotation of the dice afterwards. Even a small uncertainty in our initial information is enough to render the future trajectory unpredictable.

This example illustrates the following fundamental fact lying at the heart of what is sometimes called "chaos theory":

> *A system governed by deterministic rules can be unpredictable in the long run, and behave in ways that appear random.*

In other words, even if we have perfect knowledge of how a system's state changes from one time to another, we may not be able to make effective forecasts of the system's behavior far into the future. Indeed, in many systems – but not all! – our ability to make concrete predictions over long time scales is not noticeably

better than if the system was behaving randomly. Thus we are led to the following questions.

(1) How do we describe and identify systems in which we lose predictability and perceive randomness?
(2) How can we give a precise mathematical description of random behavior in a deterministic system?

Roughly speaking, the first question is answered by "hyperbolic dynamics", and the second by "ergodic theory". As the title of this book suggests, we will explore the confluence of these related but distinct theories.

In this chapter, we will focus on two fundamental examples, motivated by real-world problems, which exhibit hyperbolic behavior:

- the "standard map" of Taylor and Chirikov (§1.2), and
- the "strange attractor" of Lorenz (§1.12).

These examples turn out to be quite hard to analyze rigorously in as much detail as we would like, because their hyperbolicity is *nonuniform*. Consequently, in much of the book we will study more tractable families of *uniformly hyperbolic* systems, introduced in Chapter **??**.

This chapter also introduces three fundamental theorems that will play a central role in the remainder of the book:

- the Hadamard–Perron theorem on stable and unstable manifolds (§1.5),
- the Birkhoff ergodic theorem on asymptotic averages (§1.11), and
- the Perron–Frobenius theorem on eigendata of positive matrices (§1.13).

The proofs of these theorems illustrate techniques we will use frequently, including contractions in auxiliary dynamics induced by phase space expansion. At the end of the chapter (§1.13), we will discuss Bernoulli and Markov measures, which begin to illustrate the bridge between hyperbolic dynamics and stochastic processes that will be explored in the rest of the book.

1.2. The basics, by example

A dynamical system has two ingredients.

- First, a *phase space* X: the set of all possible states the system can be in.
- Second, a *time evolution rule* governing how each future state of the system is determined by its present state.

We will mostly study examples where the phase space X is not merely a set, but also carries the structure of a metric space, so that it makes sense to describe two states $x, y \in X$ as "close together" or "far apart". When we discuss ergodic theory, X will be a measure space as well. Often X will be a smooth manifold, so we can talk about "direction" as well as proximity. For now, it is fine to think of X as being Euclidean space $\mathbb{R}^N$, so that the state of the system is determined by N real numbers, and the metric is given by $d(x, y) = \|x - y\|$, where $\|\cdot\|$ is the Euclidean norm.

1.2.1. The rotator. To make things more concrete, consider a *rotator*: A point mass connected by a rigid, massless rod to a pivot point, around which it rotates without friction, gravity, or any external forces.

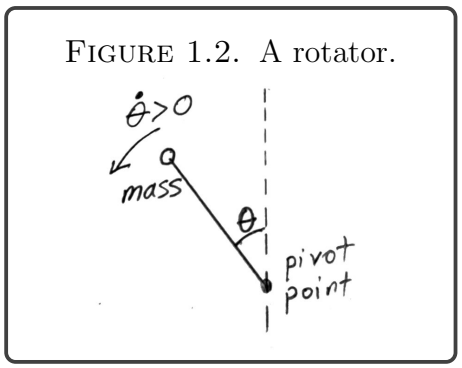


FIGURE 1.2. A rotator.

Let θ denote the angle the rod makes with a fixed reference direction; then the state of the system is determined by the rod's position θ and velocity $\dot\theta$, so our phase space is $\mathbb{R}^2$, the Euclidean plane.[1] We use coordinates (x_1, x_2) on $\mathbb{R}^2$, where $x_1 = \theta/2\pi$ and $x_2 = \dot x_1 = \dot\theta/2\pi$; then in the absence of any external forces, the time evolution rule of the system is given by the differential equations[2]

$$\dot x_1 = x_2, \qquad \dot x_2 = 0. \tag{1.1}$$

This is easy enough to solve: if the system begins in state $x = (x_1, x_2)$, then after a time $t \in \mathbb{R}$ has elapsed, it is in state

$$f_t(x_1, x_2) := (x_1 + tx_2, x_2). \tag{1.2}$$

The map $f_t \colon X \to X$ is called the *time-t map* of the system. Because the right-hand side of (1.1) has no explicit dependence on t (the ODEs are *autonomous*), the time evolution rule does not depend on what time we start, and f_t describes not just the evolution from time 0 to time t, but also the evolution from time s to time $s+t$ for every $s \in \mathbb{R}$. That is, we have

$$f_{t+s} = f_t \circ f_s \text{ for all } s, t \in \mathbb{R}. \tag{1.3}$$

We refer to a one-parameter family of maps $\{f_t\}_{t\in\mathbb{R}}$ satisfying (1.3) as a *flow*, and call the system $(X, \{f_t\})$ a *continuous-time system*.

▶ EXERCISE 1.1. Prove that if $\{f_t\}$ is a flow, then each map $f_t \colon X \to X$ is a bijection. A *continuous flow* is one in which the map $(t, x) \mapsto f_t(x)$ is continuous; prove that in this case each f_t is a homeomorphism.

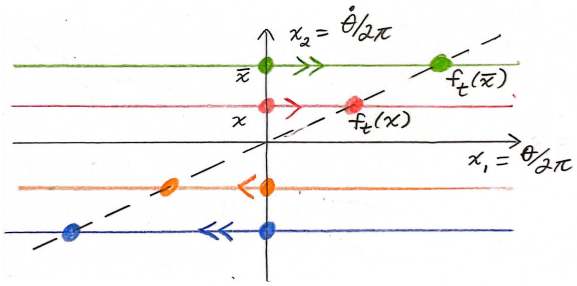
DEFINITION 1.1. Given a flow $\{f_t\}$ on a space X and an initial condition $x \in X$, the *orbit* (or *trajectory*) of x under the flow can refer either to the map $t \mapsto f_t(x)$, or to the set $\{f_t(x) : t \in \mathbb{R}\}$. Restricting to $t \in [0, \infty)$ gives the *forward orbit*, and $t \in (-\infty, 0]$ gives the *backward orbit*.

Since the flow (1.2) of the rotator has such a simple expression, it is easy to answer the question of how quickly predictability can be lost.

[1] Since θ and $\theta + 2\pi$ both describe the same state of the system, it would be more accurate to consider the phase space to be the cylinder $S^1 \times \mathbb{R}$, but for now we will stick with the plane.

[2] We will always use a dot to denote a derivative with respect to time.

FIGURE 1.3. Phase space and orbits of the rotator. The dashed line is the image of the vertical axis under the time-t map f_t.



▶ EXERCISE 1.2. Consider two orbits of the rotator with initial conditions x and y, separated by an initial "measurement error" of $\Delta_0 = (v_1, v_2)$; the idea is that one orbit is the true trajectory, and the other is the predicted trajectory. Let Δ_t be the separation between the orbits at time $t \geq 0$.

(1) Prove that $\Delta_t = (v_1 + tv_2, v_2)$, so that $\|\Delta_t\| \leq (1+t)\|\Delta_0\|$.
(2) Fix $\epsilon > \delta > 0$, and suppose that the initial measurement is accurate to within δ. For how long is the forecast based on (1.2) guaranteed to be accurate to within ϵ?

The orbit separation in Exercise 1.2 is illustrated in Figure 1.3. Although nearby orbits of this system can separate, they do so rather slowly, and if we want to double the length of time for which our prediction is valid, we can accomplish this by doubling the accuracy of our initial measurement. As we will soon see, for many systems the situation is much, much worse.

1.2.2. The pendulum. Now orient the rotator in Figure 1.2 in the vertical plane, and subject it to the force of gravity, so that it becomes a pendulum. If the gravitational force is mg, then the component perpendicular to the rod is $mg\sin\theta$; see Figure 1.4.[3] By Newton's second law this is equal to $mr\ddot\theta$, where r is the length of the rod, so $r\ddot\theta = g\sin\theta$.

As before, we can describe the phase space as the plane $\mathbb{R}^2$ (or the cylinder, if you prefer). Using coordinates $x_1 = \theta/2\pi$ and $x_2 = \dot x_1 = \dot\theta/2\pi$, (1.1) is replaced by

FIGURE 1.4. A pendulum.

$$\dot x_1 = x_2, \qquad \dot x_2 = \ddot x_1 = \frac{g}{2\pi r}\sin(2\pi x_1). \tag{1.4}$$

Lec 2
Wed, Jan 15

[3]Here we are using the *upward* pointing vertical as the reference direction, so that the more familiar downward-pointing configuration corresponds to $\theta = \pm\pi$.

▶ EXERCISE 1.3. Prove that there is $\beta > 0$ such that writing $\tau = \beta t$, we have $\frac{d^2}{d\tau^2}x_1 = \sin(2\pi x_1)$, so that writing $y_1 = x_1$ and $y_2 = \frac{d}{d\tau}y_1$, (1.4) becomes

$$\frac{d}{d\tau}y_1 = y_2, \qquad \frac{d}{d\tau}y_2 = \sin(2\pi y_1).$$

The result of Exercise 1.3 is usually expressed more succinctly by keeping the same variable names and saying that a rescaling of time converts the equations of the pendulum to the dimensionless form

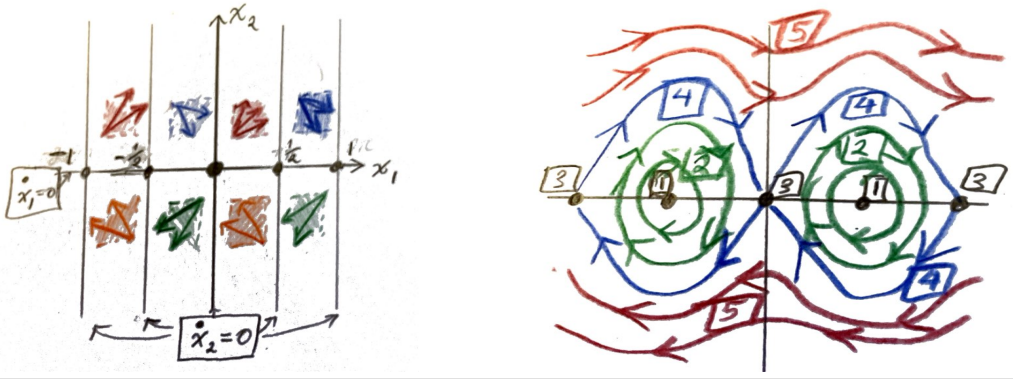
$$\dot{x}_1 = x_2, \qquad \dot{x}_2 = \sin(2\pi x_1). \tag{1.5}$$

From now on we will work with (1.5) for convenience. This system of ODEs induces a flow f_t on $\mathbb{R}^2$. We cannot write a simple equation for the flow as we did in (1.2), but can describe its qualitative behavior by:

(1) finding its fixed points;
(2) determining the regions in which the flow moves left, right, up, or down;
(3) exhibiting a conserved quantity whose level sets contain the orbits.

This is illustrated in Figure 1.5. The first picture shows the regions in which the vector field points into each of the four quadrants, using (1.5) to determine whether $\dot{x}_1$ and $\dot{x}_2$ are positive, negative, or zero at each point in the plane; this also yields the fixed points at $(\frac{n}{2}, 0)$ for every $n \in \mathbb{Z}$.

FIGURE 1.5. Drawing the phase portrait of the pendulum.



For the third step, start by recalling the rotator, where the value of x_2 is not changed by the flow, so every orbit of the flow is a horizontal line. In other words, horizontal lines and the function $H(x_1, x_2) = x_2$ are invariant, in the following sense.

DEFINITION 1.2. Given a flow f_t on a phase space X, a function $H\colon X \to \mathbb{R}$ is *invariant*[4] if $H \circ f_t = H$ for every $t \in \mathbb{R}$; that is, if $H(f_t(x)) = H(x)$ for every $x \in X$ and $t \in \mathbb{R}$. An *invariant set* is $A \subset X$ such that $f_t(A) = A$ for every $t \in \mathbb{R}$.

[4]The term *flow-invariant* is also used. Such a function is sometimes called a *first integral* of the flow, a *constant of motion*, or a *conserved quantity*.

For the pendulum (1.5), an invariant function is provided by conservation of energy: writing

$$H(x) = \pi x_2^2 + \cos(2\pi x_1), \tag{1.6}$$

where the first term represents kinetic energy and the second term potential energy, we have (using the chain rule and (1.5))

$$\frac{d}{dt} H(f_t(x)) = 2\pi x_2 \dot{x}_2 - 2\pi \sin(2\pi x_1)\dot{x}_1 = 0.$$

Thus the level sets of H are invariant, and every orbit of the flow induced by (1.5) must lie on one of these level sets, which are illustrated in the second picture in Figure 1.5. Here the information from the first picture is used to determine the fixed points and the direction of the flow on each level set. We can identify five types of orbits.

(1) *Elliptic fixed point*: At $x = (\pm\frac{1}{2}, 0)$, and more generally $(\frac{n}{2}, 0)$ for any odd integer n, we have $\dot{x} = 0$, so this is a fixed point for the flow. Here $H(x) = -1$, and this is a global minimum for H. A short computation reveals that the Hessian matrix of H is positive definite here, so up to quadratic order, the nearby level sets are ellipses surrounding x. Orbits beginning close to x remain close to x, so it is a *stable* fixed point.

Physically, these fixed points correspond to the configuration where the pendulum is hanging motionless from the pivot point.

(2) *Periodic oscillation*: If $|x_2| \leq \sqrt{\cos(2\pi x_1)/\pi}$, then $-1 < H(x) < 1$, and the corresponding level set is a closed curve, so the orbit is periodic.

Physically, this corresponds to the most familiar behavior of a pendulum, when it oscillates periodically around the downward-pointing configuration. If $H(x) > 0$, then the pendulum has enough energy to move above the horizontal configuration before reversing direction and swinging the other way.

(3) *Hyperbolic fixed point*: At $x = \mathbf{0}$, and more generally $(n, 0)$ for any $n \in \mathbb{Z}$, we have $\dot{x} = 0$, so this is again a fixed point. Here $H(x) = 1$ and the Hessian matrix has eigenvalues of both signs, so x is a saddle for H, and up to quadratic order, the nearby level sets are hyperbolas. Nearby orbits do not remain within a neighborhood of x, so it is an *unstable* fixed point.

Physically, this corresponds to a pendulum balancing vertically, pointing straight up from the pivot point.

(4) *Heteroclinic connection*: if $H(x) = 1$ and x is not a fixed point, then the orbit of x is a curve connecting $(n, 0)$ and $(n + 1, 0)$ for some $n \in \mathbb{Z}$; $\lim_{t\to-\infty} f_t(x)$ is one of these points, and $\lim_{t\to\infty} f_t(x)$ is the other. If we consider the phase space to be $\mathbb{R}^2$, then these are distinct fixed points, and the orbit of x is called a *heteroclinic connection*; if we consider the phase space to be the cylinder $S^1 \times \mathbb{R}$, then they are the same fixed point, and the orbit of x is called a *homoclinic connection*.

Physically, the pendulum approaches the vertical direction in both the future and the past, never quite reaching it.

(5) *Rotation*: if $H(x) > 1$, then the trajectory in $\mathbb{R}^2$ either always moves to the right (if $x_2 > 0$) or always moves to the left (if $x_2 < 0$). There is $T > 0$ such that $f_T(x) = x + (1, 0)$, so the orbit is periodic if we consider the phase space to be the cylinder.

Physically, the pendulum has enough energy to pass the vertical direction and complete a full rotation before starting its next period.

Using the periodicity of the oscillating and rotating solutions, it can be shown that for initial conditions away from the heteroclinic connections and the unstable fixed points, we have an analogue of the result in Exercise 1.2: the displacement between two trajectories can grow at most linearly.

►► EXERCISE 1.4. Prove that for every $\epsilon > 0$, there exists $C > 0$ such that given any $x, y \in \mathbb{R}^2$ satisfying $H(x) \le 1 - \epsilon$ and $H(y) \le 1 - \epsilon$, we have

$$\|f_t x - f_t y\| \le (1 + Ct)\|x - y\| \text{ for all } t \ge 0. \tag{1.7}$$

Similarly, prove that $C = C(\epsilon)$ can be chosen such that (1.7) holds for all $x, y \in \mathbb{R}^2$ satisfying $H(x) \ge 1 + \epsilon$ and $H(y) \ge 1 + \epsilon$. Can you extend this result to the case when $H(x) \le 1 - \epsilon$ and $H(y) \ge 1 + \epsilon$?

1.2.3. The Chirikov–Taylor standard map. Suppose we decide to use a computer to approximate some orbits of the pendulum. Although the pendulum is a continuous-time system where t can take any real value, the simplest way to perform a numerical approximation is to fix a time step $\tau > 0$, find a map $g\colon \mathbb{R}^2 \to \mathbb{R}^2$ that approximates the time-τ map f_τ, and iterate g. If $x \in \mathbb{R}^2$ represents the initial state of the system, then $g(x)$ represents the state after one time step (at time τ), $g^2(x) = g(g(x))$ represents the state after two time steps (at time 2τ), and in general $g^n(x) = g(g^{n-1}(x))$ represents the state after n time steps (at time $n\tau$).

Another way of saying all of this is that we replace the *continuous-time* flow with the *discrete-time* system given by the time-τ map, and then compute (an approximation to) this map numerically.

REMARK 1.3. In a discrete-time system, the map g could be either invertible or non-invertible. If g is a bijection then we write g^{-n} for the g-fold iteration of g^{-1}, so that g^n makes sense for all $n \in \mathbb{Z}$, and an analogue of (1.3) holds. If g is non-invertible then g^n is only defined for $n \ge 0$. Orbits are defined as for flows, except now an orbit is a sequence rather than a path.

One way of producing $g \approx f_\tau$ is to define $\bar{x} = g(x)$ by using the linear approximation resulting from (1.5) to first update x_2, and then to update x_1 using the new value of $\bar{x}_2$. This gives

$$\bar{x}_2 = x_2 + \tau \sin(2\pi x_1), \qquad \bar{x}_1 = x_1 + \tau \bar{x}_2. \tag{1.8}$$

Thus the map $g\colon \mathbb{R}^2 \to \mathbb{R}^2$ approximating f_τ is defined by

$$g(x_1, x_2) = \big(x_1 + \tau x_2 + \tau^2 \sin(2\pi x_1),\ x_2 + \tau \sin(2\pi x_1)\big). \tag{1.9}$$

REMARK 1.4. One could also update x_1 first, and then x_2, to get

$$\bar{x}_1 = x_1 + \tau x_2, \qquad \bar{x}_2 = x_2 + \tau \sin(2\pi \bar{x}_1),$$

so that instead of g we approximate f_τ with the map

$$\tilde{g}(x_1, x_2) = \big(x_1 + \tau x_2,\ x_2 + \tau \sin(2\pi(x_1 + \tau x_2))\big). \tag{1.10}$$

One could also update both x_1 and x_2 at the same time, obtaining the map

$$\tilde{f}(x_1, x_2) = \big(x_1 + \tau x_2,\ x_2 + \tau \sin(2\pi x_1)\big). \tag{1.11}$$

This represents one step of Euler's method for approximating the solution of the ODEs in (1.5).

►► EXERCISE 1.5. Prove that g and $\tilde{g}$ are homeomorphisms, and that each one can be obtained from the other in the following sense: there exists a homeomorphism $h\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that $g \circ h = h \circ \tilde{g}$.

►► EXERCISE 1.6. Prove that $\tilde{f}$ is a homeomorphism if and only if $\tau^2 < \frac{1}{2\pi}$.

Now we compare the dynamics of iterates of g to the pendulum flow. We consider g as acting on the cylinder $S^1 \times \mathbb{R}$, and draw this domain as $[0,1] \times \mathbb{R}$, where the lines $x_1 = 0$ and $x_1 = 1$ are identified. Figure 1.6 shows three orbits of g for to the time step $\tau = 0.4$, computed to 300 iterates.

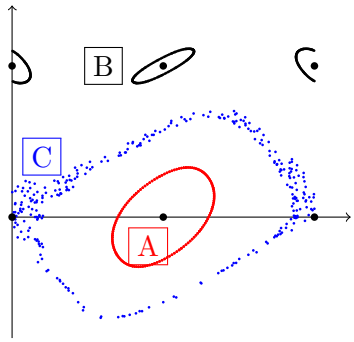


FIGURE 1.6. Some numerics.

If these were actually orbits of f_τ, then each one would lie on one of the invariant curves sketched in Figure 1.5. The orbit labeled A does appear to lie on a curve surrounding the fixed point at $(\frac{1}{2}, 0)$. However, while the orbit labeled B appears to lie on a curve, this curve has two distinct pieces, which surround the points $(0, \frac{5}{4})$ and $(\frac{1}{2}, \frac{5}{4})$. These points form a period-2 orbit of both g and f_τ, but Figure 1.5 shows no flow-invariant curves surrounding them. Finally, the orbit labeled C does not appear to lie on a curve at all; although the iterates mostly cluster near the homoclinic connections associated to the unstable fixed point, they spread out over a broader region of phase space. The dynamics of g appear qualitatively different than those of the time-τ map f_τ; what is happening?

Lec 3
Fri, Jan 17

Before analyzing g further, let us make a change of coordinates. Given $x = (x_1, x_2)$ and $\bar{x} = g(x)$, let $y = (y_1, y_2) = (x_1, \tau x_2)$, and similarly define $\bar{y} = (\bar{x}_1, \tau \bar{x}_2)$. Recalling (1.8) and (1.9), we see that y and $\bar{y}$ are related by

$$\bar{y}_1 = \bar{x}_1 = x_1 + \tau x_2 + \tau^2 \sin(2\pi x_1) = y_1 + y_2 + \tau^2 \sin(2\pi y_1),$$
$$\bar{y}_2 = \tau \bar{x}_2 = \tau x_2 + \tau^2 \sin(2\pi x_1) = y_2 + \tau^2 \sin(2\pi y_1).$$

Thus $\bar{y} = f(y)$, where $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ is given by

$$f(y_1, y_2) = \big(y_1 + y_2 + \tau^2 \sin(2\pi y_1),\ y_2 + \tau^2 \sin(2\pi y_1)\big). \tag{1.12}$$

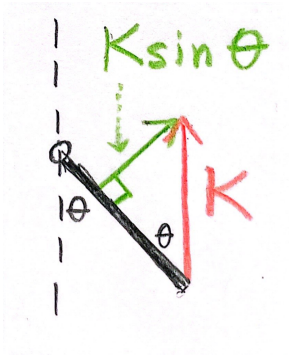
DEFINITION 1.5. Given a topological space X, two continuous maps $f\colon X \to X$ and $g\colon X \to X$ are said to be *topologically conjugate* if there exists a homeomorphism $h\colon X \to X$ such that $g \circ h = h \circ f$, or equivalently, $f = h^{-1} \circ g \circ h$.

The discussion above shows that the maps f and g in (1.9) and (1.12) are topologically conjugate via the homeomorphism $h(y_1, y_2) = (y_1, y_2/\tau)$. Exercise 1.5 shows that the map $\tilde{g}$ in (1.10) is topologically conjugate to both of these. Exercise 1.6 shows that the map $\tilde{f}$ in (1.11) coming from Euler's method cannot be topologically conjugate to these maps when $\tau^2 \geq \frac{1}{2\pi}$, because it is not a homeomorphism.

The map f in (1.12) also arises directly in physical models. The simplest of these is given by returning to the rotator from §1.2.1 and adding an external force that acts not continuously, as with gravity, but impulsively – that is, at specific moments in time, the value of $\dot{\theta}$ is changed instantaneously.

To make this more concrete, suppose that at periodic intervals, the rotator receives a "kick" in a fixed direction from some external source, which has the effect of increasing $\dot{\theta}$ if it is rotating in the direction of the kick, and decreasing $\dot{\theta}$ if it is rotating against the kick. We model the resulting change in $x_2 = \dot{\theta}/2\pi$ by $K \sin\theta = K \sin 2\pi x_1$, where $K > 0$ is the strength of the kick; see Figure 1.7. For simplicity, assume that the kicks occur at each integer time; then the evolution of the system through one unit of time is given by composing the two maps

FIGURE 1.7. Kicking the rotator.



$$(x_1, x_2) \mapsto (x_1, x_2 + K \sin 2\pi x_1) =: (\bar{x}_1, \bar{x}_2) \qquad \text{(kick)}$$
$$(\bar{x}_1, \bar{x}_2) \mapsto (\bar{x}_1 + \bar{x}_2, \bar{x}_2) \qquad \text{(flow)}$$

to obtain the following map $\mathbb{R}^2 \to \mathbb{R}^2$, which is illustrated in Figure 1.8:

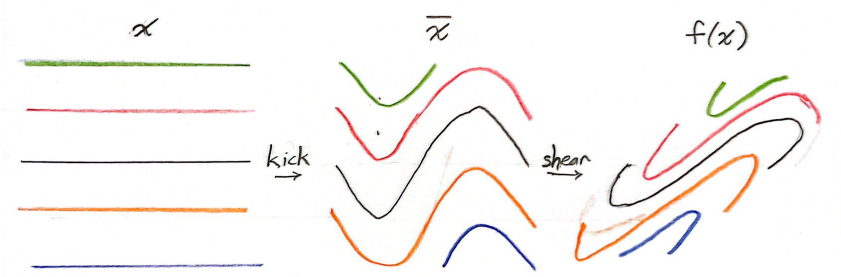
$$f(x_1, x_2) = (x_1 + x_2 + K \sin 2\pi x_1,\ x_2 + K \sin 2\pi x_1). \tag{1.13}$$

Observe that writing $K = \tau^2$, this is exactly the map we produced in (1.12). As discussed there, we consider the kicked rotator as a *discrete-time* dynamical system, for which time moves in discrete increments and only takes integer values; it is not the time-1 map of a flow on $\mathbb{R}^2$.

REMARK 1.6. The map (1.13) is often referred to as the *Chirikov–Taylor standard map*, or just as the *standard map*. The case $K = 0$, corresponding to an unkicked rotator, gives the map $f(x_1, x_2) = (x_1 + x_2, x_2)$, which is an example of a

twist map. Perturbations of twist maps, and the standard map in particular, occur in a broad range of physical applications. From the conjugacy to the map g in (1.9), we see that the standard map can also be viewed as a perturbation of the pendulum.

FIGURE 1.8. A geometric interpretation of the standard map.



For the unkicked rotator given in (1.2), it was easy to write a formula for the future state, even at large time, and you can solve Exercise 1.2 this way. It is folly to pursue this approach with the standard map of (1.13), and so another kind of argument will be needed if we want to understand the relationship between measurement error and forecast error here.

With the pendulum, Exercise 1.4 controlled the forecast error using the periodicity of the oscillating and rotating orbits, which we deduced from the existence of the invariant function H in (1.6). The solution of this exercise relies on having uniform bounds for the period, which is why it only applies to initial conditions away from the heteroclinic curves connecting the hyperbolic fixed points. The importance of the periodic orbits, the invariant curves, and the fixed points in this discussion suggests the following.

QUESTION 1.7. *Consider the standard map f from* (1.13).

(a) *What are the periodic orbits of f?*
(b) *Is there a continuous function $H\colon \mathbb{R}^2 \to \mathbb{R}$ that is f-invariant?*
(c) *Are there f-invariant curves in the plane?*
(d) *If so, do any of these connect hyperbolic fixed points?*
(e) *Do any invariant curves surround elliptic fixed points?*
(f) *Do any invariant curves connect $\{0\}\times\mathbb{R}$ to $\{1\}\times\mathbb{R}$, so that they "wrap around the cylinder"?*
(g) *If Δ_n represents the distance between $f^n(y)$ and $f^n(x)$ for two initial conditions x, y, how does Δ_n change with time?*
(h) *How can we describe the orbits that are illustrated in Figure 1.6?*
(i) *How do the answers to the questions depend on the parameter K?*

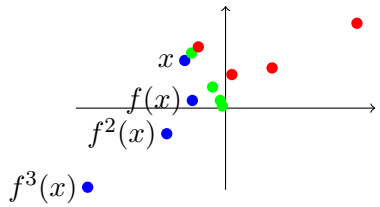
For now, we focus on the behavior near a hyperbolic fixed point, and in particular on the question of existence of invariant curves there. One can quickly check that

f has fixed points at $(\frac{n}{2}, 0)$ for all $n \in \mathbb{Z}$, and that these are the only fixed points. We consider the fixed point at $\mathbf{0}$.

Figure 1.9 shows several orbits of f for $K = 1/6$. Recalling Figure 1.5, these look similar to orbits of the time-τ map of the pendulum near $\mathbf{0}$. One orbit seems to be approaching $\mathbf{0}$ (perhaps along an invariant curve?), and the other two could be moving along hyperbola-like curves.

Indeed, it turns out that near $\mathbf{0}$, the time-τ map of the pendulum and of f behave similarly, and the discussion from here through the end of §1.5 applies to both of these systems, although we focus on the standard map f.

FIGURE 1.9. Orbits near $\mathbf{0}$.



Given $x \approx \mathbf{0}$, linear approximation suggests that $f(x) \approx Df(\mathbf{0})x$. Iterating, if $f^k(x) \approx \mathbf{0}$ for all $0 \le k < n$, then we expect to have $f^n(x) \approx (Df(\mathbf{0}))^n x$. Thus we are led to study the powers of the matrix

$$L = Df(\mathbf{0}) = \begin{pmatrix} 1 + 2\pi K & 1 \\ 2\pi K & 1 \end{pmatrix}. \tag{1.14}$$

In the next two sections we study the linear dynamics of $v \mapsto Lv$, before returning in §1.5 to the question of the connection between the linear and nonlinear dynamics of (1.13) near the origin.

1.3. Hyperbolicity in linear dynamics

The linearized dynamics of the kicked rotator near the fixed point at the origin are given by the matrix powers L^n. For concreteness, we start with the case $K = \frac{1}{2\pi}$, when (1.14) becomes

$$L = Df(\mathbf{0}) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}. \tag{1.15}$$

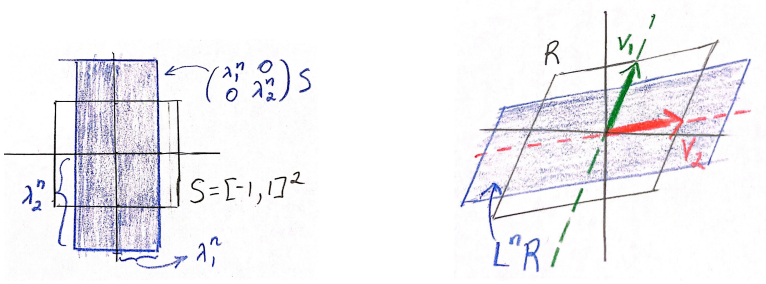
We can study L^n in terms of the eigendata of L. Following the approach we learned in linear algebra class,[5] we compute the characteristic polynomial $\det(L - \lambda I) = \lambda^2 - 3\lambda + 1$, and then use the quadratic formula to find its roots, yielding the eigenvalues $\lambda_1 = \frac{1}{2}(3 - \sqrt{5}) \in (0, 1)$ and $\lambda_2 = \frac{1}{2}(3 + \sqrt{5}) > 1$.

▶ EXERCISE 1.7. Use row reduction of $L - \lambda I$ (or another method) to find the eigenvectors v_1, v_2 of $L = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, observe that they are orthogonal, and remind yourself of why this follows from the fact that L is symmetric.

The fact that $0 < \lambda_1 < 1 < \lambda_2$ already tells us a lot about the dynamics of L^n, so from now on we let L be *any* 2×2 matrix whose eigenvalues satisfy these

[5] Which we will soon jettison in favor of a more dynamical approach...

FIGURE 1.10. Powers of matrices with eigenvalues $0 < \lambda_1 < 1 < \lambda_2$.



(a) (b)

inequalities. The simplest case is a diagonal matrix:

$$L = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \quad \Rightarrow \quad L^n = \begin{pmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{pmatrix}. \tag{1.16}$$

The effect of the matrix on the unit square is illustrated in Figure 1.10(a). More generally, if v_1, v_2 are eigenvectors for λ_1, λ_2, then the matrix C whose columns are v_1, v_2 diagonalizes L, and we have

$$L = C \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} C^{-1} \quad \Rightarrow \quad L^n = C \begin{pmatrix} \lambda_1^n & 0 \\ 0 & \lambda_2^n \end{pmatrix} C^{-1}. \tag{1.17}$$

In this case the iterates L^n act as shown in Figure 1.10(b).

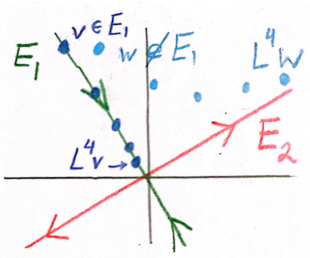
Lec 4
Fri, Jan 24

To study the orbit of an arbitrary $v \in \mathbb{R}^2$, we write $v = c_1 v_1 + c_2 v_2$ for some $c_1, c_2 \in \mathbb{R}$ so that we have

$$L^n v = c_1 L^n v_1 + c_2 L^n v_2 = c_1 \lambda_1^n v_1 + c_2 \lambda_2^n v_2. \tag{1.18}$$

Observe that the two terms of (1.18) are the components of $L^n v$ in the eigenspaces E_1, E_2, and that as $n \to \infty$ we have $\lambda_2^n \to \infty$ and $\lambda_1^n \to 0$. Thus the E_1 component of $L^n v$ always goes to 0 exponentially fast, while the E_2 component grows exponentially, unless it was 0 to begin with. We deduce the following result, which Figure 1.11 illustrates for the specific matrix $L = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$.[6]

FIGURE 1.11. Linear iteration.



LEMMA 1.8. *Let L be a 2×2 matrix with eigenvalues $0 < \lambda_1 < 1 < \lambda_2$, eigenvectors v_i, and eigenspaces E_i. Given $v = c_1 v_1 + c_2 v_2$, we have the following dichotomy for $\|L^n v\|$ as $n \to \infty$.*

[6]Compare this picture to the nonlinear trajectories in Figure 1.9.

- *If $v \in E_1$, then $c_2 = 0$ and $L^n v \in E_1$ for all n. In this case $L^n v \to 0$ exponentially fast:* $\|L^n v\| = \|c_1 v_1\| \lambda_1^n$.
- *If $v \notin E_1$, then $c_2 \neq 0$ and $L^n v \to E_2$ as $n \to \infty$. In this case $L^n v \to \infty$ exponentially fast:* $\lim_{n\to\infty} \|L^n v\|/\lambda_2^n = \|c_2 v_2\|$.

FIGURE 1.12. Evolution of a "cloud" of initial conditions.

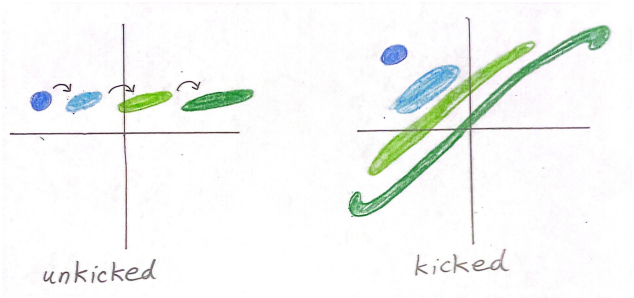


Figure 1.12 shows the growth of an initial measurement error v for the unkicked and kicked rotators. We argued in (1.14) that the powers L^n determine the behavior of orbits of the kicked rotator near the fixed point $(0, 0)$: if the orbit remains near this fixed point through time n, then the error in the forecast at time n should be roughly $L^n v$.

For the unkicked rotator, the error grew linearly in n (Exercise 1.2), and doubling our measurement accuracy also doubled the duration of our forecast's validity. The situation here is much worse: if our measurement error has an E_2 component of size δ, then after time n we expect our forecast error to be of magnitude $\delta\lambda_2^n$, so that the forecast is only valid to within ϵ if $\lambda_2^n \leq \epsilon/\delta$, or equivalently, $n \leq \frac{1}{\log \lambda_2}(\log \epsilon - \log \delta)$. What happens if we double our measurement accuracy now? We still increase our forecast's duration, but not by a multiplicative factor; rather, we only add a fixed amount of time (in this case, $\log 2/\log \lambda_2$). This is the basic mechanism driving the appearance of randomness in deterministic systems:

Exponential separation of orbits makes it prohibitively difficult to control the error term in long-term forecasts.

REMARK 1.9. Not all measurement errors are amplified; if the initial error is in the direction of E_1 then the linearization suggests that the orbits will draw closer and the error will decrease. In this case, however, the orbits separate exponentially fast as we go *backward* in time, creating difficulties in recovering the initial conditions given the final state.

The mechanism described above is not the whole story. Forecast error is driven not only by amplification of measurement error as orbits separate, but also by numerical error (the computed trajectory is only an approximation of the true trajectory)

and by model error (the model is only an approximation of reality). These considerations are largely beyond the scope of this book, and require a more complete discussion of numerical analysis and scientific modeling. However, they do suggest two important ideas:

- we should also study *pseudo-orbits* $x_0, x_1, x_2, \dots$ of a system, for which we merely have $x_{n+1} \approx f(x_n)$ instead of $x_{n+1} = f(x_n)$;
- it is important to understand whether a nearby system $g \approx f$ must have similar dynamics to f, or whether it can behave differently.

This second idea reminds us that have yet to justify why the behavior of the linear system $v \mapsto Lv$ also describes the nonlinear system $x \mapsto f(x)$ in a neighborhood of the origin, where $f \approx L$. A formal theorem doing this will come in §1.5. Here we clarify what such a result ought to say, and then in §1.4 we develop a dynamical approach to eigendata that will prove more adaptable to the nonlinear case than the algebraic approach above.

From Lemma 1.8 above, and the corresponding result for $n \to -\infty$ with the roles of E_1 and E_2 reversed, we see that the eigenspaces can be characterized dynamically:

$$\begin{aligned} E_1 &= \{v \in \mathbb{R}^2 : L^n v \to 0 \text{ as } n \to \infty\}, \\ E_2 &= \{v \in \mathbb{R}^2 : L^{-n} v \to 0 \text{ as } n \to \infty\}. \end{aligned} \tag{1.19}$$

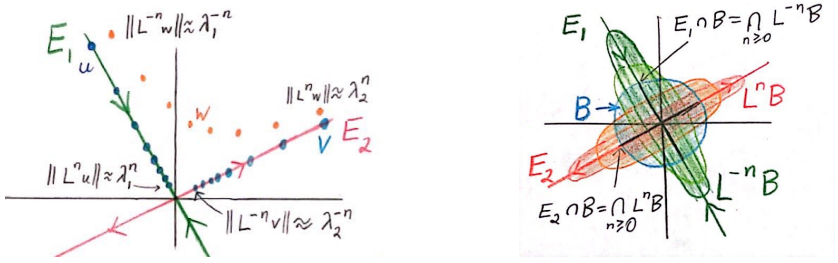
This is illustrated in Figure 1.13(a). We can also write the eigenspaces as

$$\begin{aligned} E_1 &= \{v \in \mathbb{R}^2 : \{L^n v : n \geq 0\} \text{ is bounded}\}, \\ E_2 &= \{v \in \mathbb{R}^2 : \{L^{-n} v : n \geq 0\} \text{ is bounded}\}. \end{aligned} \tag{1.20}$$

Figure 1.13(b) illustrates the following equivalent description: if $B \subset \mathbb{R}^2$ is a ball centered at the origin, then

$$E_1 \cap B = \bigcap_{n \geq 0} L^{-n} B, \qquad E_2 \cap B = \bigcap_{n \geq 0} L^n B. \tag{1.21}$$

FIGURE 1.13. Eigenspaces via convergence and intersections.



These dynamical descriptions suggest that in the nonlinear case, we might profitably study the sets obtained from (1.19)–(1.21) by replacing L with f. That is,

we can study the sets

$$\{x \in \mathbb{R}^2 : f^n(x) \to 0 \text{ as } n \to \infty\} \quad \text{and} \quad \{x \in \mathbb{R}^2 : f^{-n}(x) \to 0 \text{ as } n \to \infty\},$$

or the sets

$$\bigcap_{n\geq 0} f^{-n}(B) \quad \text{and} \quad \bigcap_{n\geq 0} f^n(B),$$

and ask whether these are "close to" the lines E_1 and E_2. In §1.5 we will see that indeed these sets are curves that are close to these lines. To prove this, we need a completely dynamical approach to the eigendata of L: the arguments in this section relied on knowing in advance that L has eigenvalues $0 < \lambda_1 < 1 < \lambda_2$, which we deduced (for the matrix $L = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$) via the quadratic formula. In the nonlinear setting, we have no such algebraic tool.

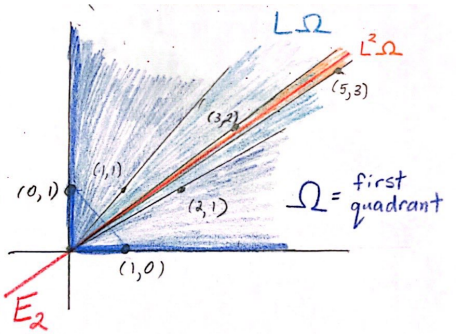
1.4. Eigenspaces via fixed point theorems

The algebraic approach to eigendata requires us to first find the eigenvalues as roots of the characteristic polynomial, and then use our knowledge of the eigenvalues to find the eigenvectors. In this section we describe a dynamical approach that proceeds in the opposite direction, finding the eigenspaces first. This idea will play an important role in the Hadamard–Perron theorem in §1.5, in the Perron–Frobenius theorem in §1.13, and in various other settings where the algebraic techniques of the previous section are insufficient.

1.4.1. Positive matrices and invariant cones. Start with the matrix $L = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$. We saw in Lemma 1.8 that every $v \notin E_1$ has $L^n v \to E_2$ as $n \to \infty$; in particular, this is true of every v in the first quadrant.

Motivated by this, we look at the images of the first quadrant Ω under the iterates L^n, which are illustrated in Figure 1.14. As n increases, these appear to be cones that are getting narrower and converging to the eigenspace E_2.[7] Indeed, one can prove using Lemma 1.8 that E_2 is an attracting fixed point for the action of L on the space of lines through the origin in $\mathbb{R}^2$.

FIGURE 1.14. Narrowing images.



Let us make this more precise, and then explore how to bypass the use of Lemma 1.8. Given $d \in \mathbb{N}$, consider the *projective space* $\mathbb{RP}^{d-1}$ of one-dimensional subspaces (lines through the origin) in $\mathbb{R}^d$:

$$\mathbb{RP}^{d-1} = (\mathbb{R}^d \setminus 0)/\sim, \text{ where } v \sim cv \text{ for all } v \in \mathbb{R}^d \setminus 0 \text{ and } c \in \mathbb{R} \setminus 0. \tag{1.22}$$

[7] If you like, you can interpret this convergence in terms of the limiting ratio between successive Fibonacci numbers.

Write $[v]$ for the equivalence class of $v \in \mathbb{R}^d \setminus 0$. A $d \times d$ matrix L acts on $\mathbb{RP}^{d-1}$ by $L[v] = [Lv]$, and v is an eigenvector if and only if $L[v] = [v]$, so we are led to search for fixed points of $L\colon \mathbb{RP}^{d-1} \to \mathbb{RP}^{d-1}$.

FIGURE 1.15. The action of L on projective space $\mathbb{RP}^1$.

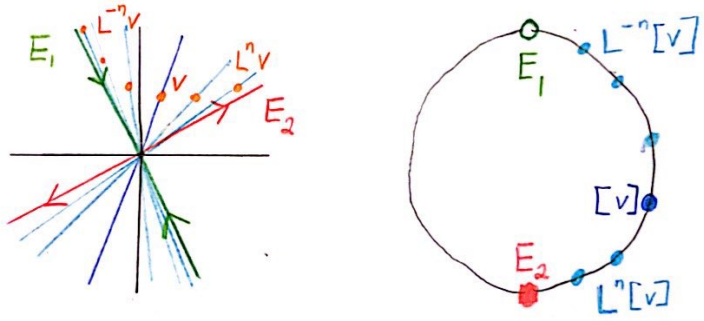


Figure 1.15 illustrates the following consequence of Lemma 1.8.

COROLLARY 1.10. *When L is a 2×2 matrix with eigenvalues $0 < \lambda_1 < 1 < \lambda_2$, the following are true.*

- *The projective map $L\colon \mathbb{RP}^1 \to \mathbb{RP}^1$ has exactly two fixed points, corresponding to the eigenspaces E_1 and E_2.*
- *The fixed point at E_2 is* asymptotically stable*: for every $[v] \in \mathbb{RP}^1 \setminus \{E_1\}$, we have $L^n[v] \to E_2$ in the quotient topology that $\mathbb{RP}^1$ inherits from $\mathbb{R}^2$.*
- *The unstable fixed point at E_1 becomes stable if we reverse time: for every $[v] \in \mathbb{RP}^1 \setminus \{E_2\}$, we have $L^n[v] \to E_1$ as $n \to -\infty$.*

A map on $\mathbb{RP}^1$ (which is topologically a circle) with the structure described in Corollary 1.10 is sometimes called a *north-south map*. Note that from the point of view of $\|L^n v\|$, the unstable subspace was E_2 and the stable was E_1. This illustrates a general principle:

Expansion in phase space leads to contraction in projective space.

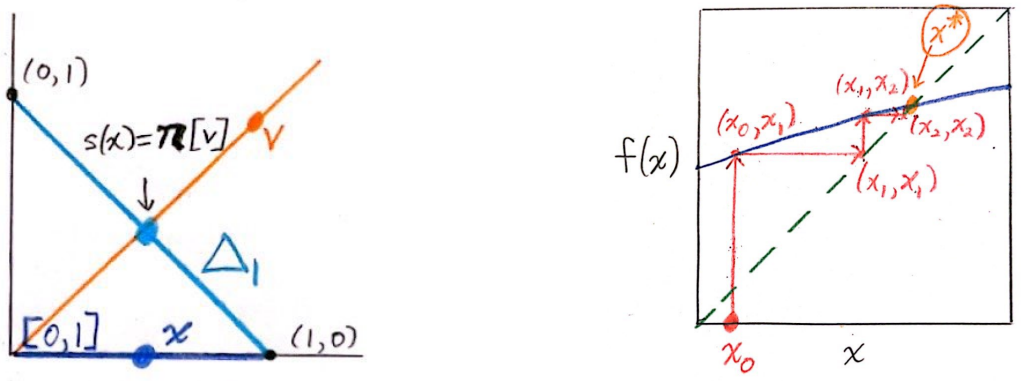
As we will see in §1.5 and elsewhere, this extends beyond "projectivization" itself, to a broader class of auxiliary dynamics. For the moment we observe one consequence of the stability of E_2: if we want to avoid doing much algebra and are content to have an approximation to the eigenspace, then we can simply take any v in the first quadrant, fix a large value of n, and use $L^n v$ as an approximate eigenvector. Of course this is most useful if we can also prove some error bounds; we will soon see how to do this.

Now comes a crucial change in strategy: we can bypass the algebra and prove Corollary 1.10 without relying on Lemma 1.8. As we saw in Figure 1.14, $L = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$

maps the first quadrant $\Omega = \{(x, y) \in \mathbb{R}^2 : x \geq 0, y \geq 0\}$ into itself. With $\sim$ as in (1.22), the quotient space $\mathbb{P}\Omega = (\Omega \setminus 0)/\sim$ is the projectivization of the first quadrant, and can be represented as a compact interval:

▶ EXERCISE 1.8. As shown in Figure 1.16, let $\Delta_1 = \{(x, y) \in \Omega : x + y = 1\}$, and define the normalization map $\mathfrak{n}\colon \mathbb{P}\Omega \to \Delta_1$ by $\mathfrak{n}\colon [v] \mapsto v/\|v\|_1$, where $\|v\|_1 = \sum_i |v_i|$. Prove that $\mathfrak{n}$ is a homeomorphism between $\mathbb{P}\Omega$ and Δ_1.

FIGURE 1.16. Parametrizing $\mathbb{RP}^1$ and obtaining a cobweb diagram.



The line segment Δ_1 can be naturally parametrized with the homeomorphism $s\colon [0, 1] \to \Delta_1$ given by $s(x) = (x, 1 - x)$, so we can write $L\colon \mathbb{P}\Omega \to \mathbb{P}\Omega$ in terms of a map $f\colon [0, 1] \to [0, 1]$ such that the following diagram commutes (in particular, f is topologically conjugate to $L\colon \mathbb{P}\Omega \to \mathbb{P}\omega$).

$$\begin{array}{ccccc} \mathbb{P}\Omega & \xrightarrow{\mathfrak{n}} & \Delta_1 & \xleftarrow{s} & [0,1] \\ \downarrow{\scriptstyle L} & & \downarrow & & \downarrow{\scriptstyle f} \\ \mathbb{P}\Omega & \xrightarrow{\mathfrak{n}} & \Delta_1 & \xleftarrow{s} & [0,1] \end{array} \tag{1.23}$$

▶ EXERCISE 1.9. Prove that when $L = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, the map f in (1.23) is given by $f(x) = \frac{x+1}{x+2}$; in particular, it is continuous (hence $L\colon \mathbb{P}\Omega \to \mathbb{P}\Omega$ is as well). Use the intermediate value theorem (or Brouwer's fixed point theorem) to deduce that it has a fixed point,[8] so that L has an eigenvector in the first quadrant. Then prove that f admits an analogous formula and conclusion when $L = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ for any $a, b, c, d > 0$.

This approach has the benefit that it applies to a broad class of matrices – all positive 2×2 matrices! – without the need to carry out the algebraic computations of Exercise 1.7 for each one. However, those algebraic considerations do provide extra information on uniqueness and stability: for $L = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, $\mathbb{P}\Omega$ contains only one

Lec 5
Mon, Jan 27

[8] Of course you could also do this quite easily by solving the equation $x = \frac{x+1}{x+2}$, but the whole point of this section is to avoid the algebraic approach!

fixed point E_2, which is stable in the sense of Lemma 1.8. This is illustrated in Figure 1.16, which shows the sequence $x_n = f^n(x_0)$ converging to the fixed point x^*. To obtain these extra properties from a fixed point theorem, we must go beyond the purely topological approach in Exercise 1.9.

THEOREM 1.11 (Banach Fixed Point Theorem). *Let X be a complete metric space and $f\colon X \to X$ a* contraction, *meaning that there is $\gamma \in (0,1)$ such that $d(fx, fy) \le \gamma d(x,y)$ for all $x, y \in X$. Then f has a unique fixed point $x^* \in X$, and this fixed point is exponentially stable: for every $x \in X$ we have $d(f^n x, x^*) \le d(x, x^*)\gamma^n$.*

▶ EXERCISE 1.10. Use the formula in Exercise 1.9 to prove that for $L = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$, the interval map $f\colon [0,1] \to [0,1]$ from (1.23) is differentiable and satisfies $f'(x) \in (0, \frac{1}{4}]$ for all $x \in [0,1]$, so that for any $0 \le x < y \le 1$ we have

$$f(y) - f(x) = \int_x^y f'(t)\,dt \le \frac{1}{4}(y - x). \tag{1.24}$$

Then apply the Banach fixed point theorem to conclude that L has a *unique* (up to a scalar) eigenvector $v \in \Omega$, and that given any $w \in \Omega$, the angle between $L^n w$ and v goes to 0 exponentially fast. Use a similar argument with L^{-1} and the second quadrant to prove Corollary 1.10.

Unlike Exercise 1.9, this result does not automatically generalize quite as broadly as we might expect.

▶▶ EXERCISE 1.11. Given $L = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with $a, b, c, d > 0$, use the formula for f from Exercise 1.9 to repeat the computation in Exercise 1.10, and prove that the corresponding map $f\colon [0,1] \to [0,1]$ is a contraction if and only if $\min(a+c, b+d)^2 > ad - bc$.

The important feature of the identification between $\mathbb{P}\Omega$ and $[0,1]$ used here is not the parametrization itself, but the metric it induces on $\mathbb{P}\Omega$.

▶ EXERCISE 1.12. Let $X \subset \Omega$ be any continuous curve with the property that every $[v] \in \mathbb{P}\Omega$ intersects X in exactly one point. Denote this point by $\mathfrak{n}v$, and prove that $d(v,w) := \|\mathfrak{n}v - \mathfrak{n}w\|$ defines a complete metric on $\mathbb{P}\Omega$.

The limitations revealed in Exercise 1.11 are a weakness of the specific choice of cross-sectional curve used in (1.23), rather than an intrinsic property of positive matrices. In §1.13 we will see that by choosing a different curve, one can guarantee that the map f is a contraction whenever $a, b, c, d > 0$; this leads to the proof of the Perron–Frobenius theorem.

1.4.2. Transverse cones and perturbations of matrices. In the meantime, we pursue a slightly different use of fixed point theorems to study eigenspaces. Suppose we have some 2×2 matrix L that we know has positive real eigenvalues $\lambda_1 < \lambda_2$ and associated eigenspaces E_1, E_2, and another 2×2 matrix $\tilde{L}$ that is close to L. Can we conclude that $\tilde{L}$ has positive real eigenvalues $\tilde{\lambda}_i \approx \lambda_i$, and associated

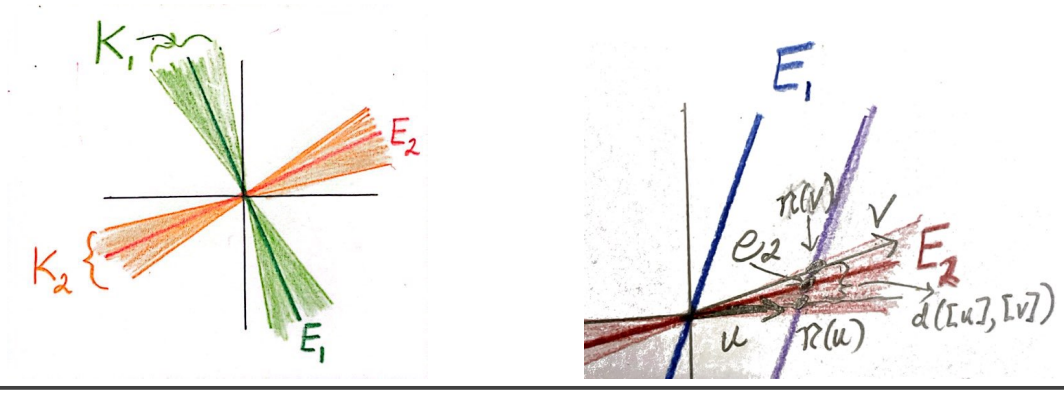
eigenspaces $\tilde{E}_i \approx E_i$? This is a warm-up question for the problem of studying *nonlinear* maps close to L in §1.5.

Corollary 1.10 described E_1 and E_2 as stable fixed points for the projective action of L^{-1} and L, respectively. We will strengthen this: there is a neighborhood of E_1 in $\mathbb{RP}^1$ on which L^{-1} acts as a contraction, and similarly for L near E_2. We will show that $\tilde{L}^{\pm 1}$ also act as contractions on these neighborhoods, which implies the existence of eigenspaces $\tilde{E}_i \approx E_i$. One can also get quantitative control using the following exercise.

▶▶ EXERCISE 1.13. Let X be a complete metric space and $f\colon X \to X$ a contraction by some ratio $\gamma \in (0,1)$. Let $x_f^* \in X$ be the unique fixed point guaranteed by the Banach fixed point theorem.

(1) Prove that for any $x \in X$ we have $d(x, x_f^*) \le \frac{1}{1-\gamma} d(x, fx)$.
(2) Fix $\delta > 0$ and let $g\colon X \to X$ be a contraction such that $d(fx, gx) \le \delta$ for all $x \in X$. Prove that the fixed points x_f^* and x_g^* satisfy $d(x_f^*, x_g^*) \le \frac{\delta}{1-\gamma}$.

FIGURE 1.17. Cones in $\mathbb{R}^2$, and a projective metric on a cone.



To describe projective neighborhoods on which $L^{\pm 1}$ and nearby matrices act as contractions, fix a small parameter $\alpha > 0$ and consider the sets

$$\begin{aligned} K_1 &:= \{v_1 + v_2 : v_i \in E_i \text{ for } i = 1,2 \text{ and } \|v_2\| \le \alpha \|v_1\|\}, \\ K_2 &:= \{v_1 + v_2 : v_i \in E_i \text{ for } i = 1,2 \text{ and } \|v_1\| \le \alpha \|v_2\|\}, \end{aligned} \tag{1.25}$$

which are shown in Figure 1.17. These are the *cones* of width α associated to the decomposition $\mathbb{R}^2 = E_1 \oplus E_2$. We study the action of $\tilde{L}$ on K_2; the situation with $\tilde{L}^{-1}$ and K_1 is analogous.

Consider the projectivization $\mathbb{P}K_2 = (K_2 \setminus 0)/\!\sim$, where as in (1.22) we write $v \sim cv$ for all $c \in \mathbb{R} \setminus 0$, and $[v]$ for the equivalence class of v. To define a metric d on $\mathbb{P}K_2$ in which every $\tilde{L} \approx L$ acts as a contraction, fix a unit vector $e_2 \in K_2$ and let

$$X = e_2 + E_1 = \{e_2 + v : v \in E_1\}.$$

Then for any $u \in K_2 \setminus 0$, the equivalence class $[u]$ intersects X in exactly one point, which we denote $\mathfrak{n}(u)$; see Figure 1.17. Observe that if u is in the same half of the cone as e_2, then

$$\mathfrak{n}(u) = \frac{u}{\|u_2\|}. \tag{1.26}$$

Recalling Exercise 1.12, we can define a complete metric on $\mathbb{P}K_2$ by[9]

$$d([u],[v]) = \|\mathfrak{n}u - \mathfrak{n}v\|. \tag{1.27}$$

We will prove the following general result.

PROPOSITION 1.12. *Let $E_1, E_2 \subset \mathbb{R}^2$ be subspaces such that $\mathbb{R}^2 = E_1 \oplus E_2$. Fix $\alpha \in (0,1)$, and let K_1, K_2 be the cones defined in (1.25). Fix $\chi \in (0,1)$ and let $L\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a linear map such that the following are true.*
(a) *$L(K_2) \subset K_2$, and $L^{-1}(K_1) \subset K_1$.*
(b) *$\|L(v)\| \geq \chi^{-1}\|v\|$ for all $v \in K_2$, and $\|L(v)\| \leq \chi\|v\|$ for all $v \in K_1$.*
Then for every $u, v \in K_2 \setminus 0$, we have

$$d([Lu],[Lv]) \leq \frac{\chi^2(1+\alpha)}{(1-\alpha)^2} d([u],[v]). \tag{1.28}$$

Before proving Proposition 1.12, we note the following consequence: if $\alpha > 0$ is small enough that $\frac{\chi^2(1+\alpha)}{(1-\alpha)^2} < 1$, then L acts as a contraction on the complete metric space $(\mathbb{P}K_2, d)$. We can summarize this as follows.

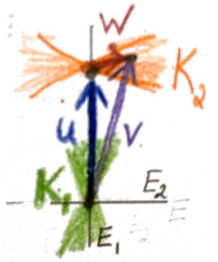
> *If a linear map L carries a narrow cone K into itself, expanding vectors as it does so, and if L^{-1} does the same to a cone transverse to K, then L acts as a contraction on $\mathbb{P}K$.*

We will return to this idea in §1.13, and will see that it is enough for L to carry K into itself, as was the case in §1.4.1 when $K = \Omega$ and L was any positive matrix; even with no information on vector expansion or on a transverse cone, one still has a contraction in the *Hilbert projective metric*.

Lec 6
Wed, Jan 29

For now, we prove Proposition 1.12. An important tool will be the observation that if $v \in K_i$, then $\|v\| \approx \|v_i\|$, and more generally $\|v\| \approx \|u\|$ whenever v and u are as shown in Figure 1.18. This is made precise by the following lemmas.

FIGURE 1.18



LEMMA 1.13. *Let $E_1, E_2, K_1, K_2 \subset \mathbb{R}^2$ and $\alpha \in (0,1)$ be as in Proposition 1.12. Given $v \in \mathbb{R}^2$, write $v = v_1 + v_2$, where $v_i \in E_i$.*
(a) *If $v \in K_i$, then $(1-\alpha)\|v_i\| \leq \|v\| \leq (1+\alpha)\|v_i\|$.*

[9] The map $\mathfrak{n}\colon \mathbb{P}K_2 \to X \cap K_2$ is a homeomorphism, and we could also proceed by defining a map $f\colon X \cap K_2 \to X \cap K_2$ such that $f \circ \mathfrak{n} = \mathfrak{n} \circ L$.

(b) *If* $v, w \in K_i$, *then* $\frac{\|v_i\|}{\|w_i\|} \leq \frac{1+\alpha}{1-\alpha}\frac{\|v\|}{\|w\|}$.

PROOF. Consider the case $i = 1$. For (a), observe that

$$\|v_1\| - \|v_2\| \leq \|v_1 + v_2\| \leq \|v_1\| + \|v_2\|$$

and use the fact that $\|v_2\| \leq \alpha\|v_1\|$. Part (b) follows immediately from (a), and the case $i = 2$ is similar. □

LEMMA 1.14. *Let* $E_1, E_2, K_1, K_2 \subset \mathbb{R}^2$ *and* $\alpha \in (0, 1)$ *be as in Proposition 1.12. Suppose that* $u \in E_1$, *and that* $v \in K_1$ *is such that* $w := v - u \in K_2$, *as in Figure 1.18. Then* $\|v\| \leq \frac{1}{1-\alpha}\|u\|$.

PROOF. Observe that $w_2 = v_2$ since $u \in E_1$, so since $w \in K_2$ and $v \in K_1$, we have

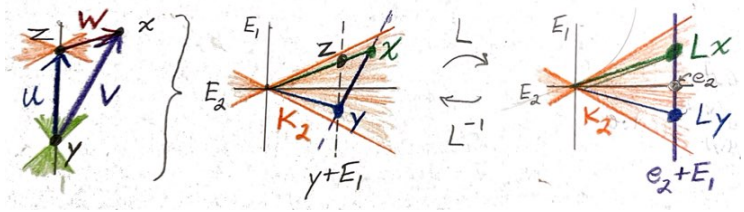
$$\|w_1\| \leq \alpha\|w_2\| = \alpha\|v_2\| \leq \alpha^2\|v_1\|$$

Now we have

$$\|u\| = \|u_1\| \geq \|v_1\| - \|w_1\| \geq (1 - \alpha^2)\|v_1\| \geq (1 - \alpha)\|v\|,$$

where the last inequality uses Lemma 1.13(a). □

FIGURE 1.19. Proving Proposition 1.12.



PROOF OF PROPOSITION 1.12. Fix $x, y \in K_2 \setminus 0$, and assume without loss of generality that $Lx, Ly \in e_2 + E_1$ (see Figure 1.19). Since $Lx - Ly \in E_1 \subset K_1$ and $L^{-1}(K_1) \subset K_1$, we have $v := x - y \in K_1$, so

$$d([Lx], [Ly]) = \|Lx - Ly\| = \|Lv\| \leq \chi\|v\|. \tag{1.29}$$

Writing $z := \frac{\|y_2\|}{\|x_2\|}x$ for the point where $y+E_1$ intersects $[x]$, observe that $u := z-y \in E_1$ and $w = v - u = x - z \in K_2$, so we can apply Lemma 1.14 to turn (1.29) into

$$d([Lx], [Ly]) \leq \chi\|v\| \leq \frac{\chi}{1-\alpha}\|u\|. \tag{1.30}$$

Moreover, from (1.26) we get

$$d([x], [y]) = \|\mathfrak{n}x - \mathfrak{n}y\| = \left\|\frac{z}{\|y_2\|} - \frac{y}{\|y_2\|}\right\| = \frac{\|z - y\|}{\|y_2\|} = \frac{\|u\|}{\|y_2\|}. \tag{1.31}$$

Since $\frac{\|y\|}{\|Ly\|} \le \chi$, Lemma 1.13(b) gives

$$\|y_2\| \le \frac{\chi(1+\alpha)}{1-\alpha}\|(Ly)_2\|, \tag{1.32}$$

and since $\|(Ly)_2\| = 1$, we can combine (1.30), (1.31), and (1.32) to get

$$d([Lx],[Ly]) \le \frac{\chi}{1-\alpha}\|y_2\|d([x],[y]) \le \frac{\chi}{1-\alpha}\cdot\frac{\chi(1+\alpha)}{1-\alpha}d([x],[y]),$$

which completes the proof of Proposition 1.12. □

►►► EXERCISE 1.14. Let $L = \left(\begin{smallmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{smallmatrix}\right)$ for some $0 < \lambda_1 < \lambda_2$. Use Exercise 1.13 and Proposition 1.12 to find $\epsilon_0 > 0$ and a function $\theta\colon (0,\epsilon_0] \to (0,\frac{\pi}{4}]$ such that the following is true:

- $\lim_{\epsilon\to 0}\theta(\epsilon) = 0$; and
- given any 2×2 real matrix A with $|A_{ij} - L_{ij}| < \epsilon$ for all $i,j \in \{1,2\}$, the matrix A has eigenspaces making an angle $\le \theta(\epsilon)$ with the coordinate axes.

1.5. The Hadamard–Perron theorem

Lec 7
Fri, Jan 31

Now we return to our motivating question from §1.2.3 of understanding the dynamics of the standard map (1.13) near the origin. Lemma 1.8 described the linear dynamics of $L = Df(\mathbf{0}) = \left(\begin{smallmatrix}2 & 1\\ 1 & 1\end{smallmatrix}\right)$ in terms of its eigenspaces. In §§1.3–1.4, we saw that

- the iterates of the cone K^u under the linear map L converge to E^u, and
- if B is a small ball around the origin, then the line segment $E^u \cap B$ can be characterized as the sets of points whose backward L-orbit converges to 0, with all other backward L-orbits leaving B.

Armed with these insights, we can now prove that similar constructions work for the nonlinear map f, as illustrated in Figure 1.20. The following is the first major theorem of the book, and we present it here in the simplest setting. Later on, we will see more sophisticated versions.

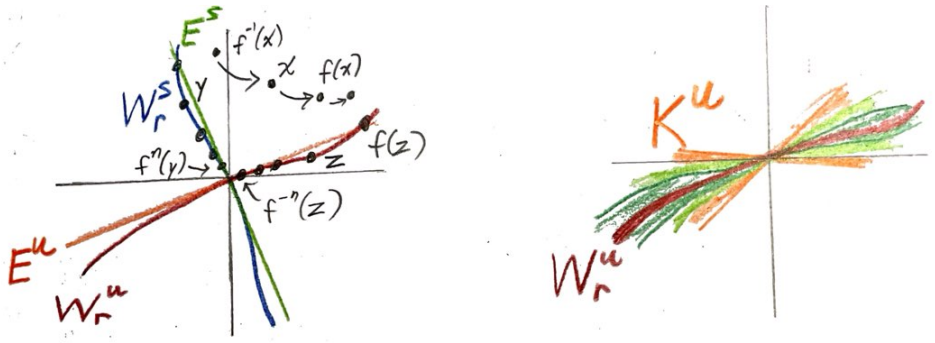
THEOREM 1.15 (Hadamard–Perron). *Let $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a C^1 diffeomorphism fixing the origin, and suppose that $L = Df(\mathbf{0})$ has eigenvalues λ_s, λ_u satisfying $0 < |\lambda_s| < 1 < |\lambda_u|$, with corresponding eigenspaces E^s, E^u. Then for every sufficiently small $r > 0$, the sets*

$$\begin{aligned} W^s_r &:= \{x \in \mathbb{R}^2 : \|f^n x\| \le r \text{ for all } n \ge 0\},\\ W^u_r &:= \{x \in \mathbb{R}^2 : \|f^{-n} x\| \le r \text{ for all } n \ge 0\} \end{aligned} \tag{1.33}$$

are C^1 curves tangent at the origin to E^s and E^u, respectively. Moreover, for every $x \in W^s_r$ we have $f^n(x) \to \mathbf{0}$ as $n \to \infty$, and for every $x \in W^u_r$ we have $f^{-n}(x) \to \mathbf{0}$ as $n \to \infty$.

The curve W^s_r is called the *local stable manifold* of the fixed point $\mathbf{0}$, and W^u_r is the *local unstable manifold*. Observe that these curves only depend on $f|_{B(\mathbf{0},r)}$. It

FIGURE 1.20. Nonlinear dynamics near a hyperbolic fixed point.



follows immediately from (1.33) that $f(W^s_r) \subset W^s_r$ and $f^{-1}(W^u_r) \subset W^u_r$, so one can think of these curves as nonlinear analogues of the eigenspaces of $Df(\mathbf{0})$. With this in mind, compare Figure 1.20 to Figures 1.13(b) and 1.14. The idea of the proof will be to mimic the procedure from §1.4.2 as much as possible.

The conclusion $\lim_{n\to\infty} f^n(x) = \mathbf{0}$ holds not just for points in W^s_r, but also for points in the (larger) set $W^s := \bigcup_{n=0}^{\infty} f^{-n}(W^s_r)$, which is called the *global stable manifold* of $\mathbf{0}$. We will see later that unlike W^s_r, which is quite close to linear, the global stable manifold W^s can be immersed in $\mathbb{R}^2$ in very complicated ways, and can return to $B(\mathbf{0}, r)$. Similar statements apply to the unstable manifold.

Before proving Theorem 1.15, we obtain estimates on how the difference between two orbits evolves in time when that difference is close to E^u or E^s, in the sense of cones as in (1.25): every $v \in \mathbb{R}^2$ can be written uniquely as $v = v_u + v_s$ where $v_u \in E^u$ and $v_s \in E^s$, and fixing $\alpha > 0$, we write

$$K^u_\alpha = \{v \in \mathbb{R}^2 : \|v_s\| \le \alpha \|v_u\|\}, \qquad K^s_\alpha = \{v \in \mathbb{R}^2 : \|v_u\| \le \alpha \|v_s\|\}.$$

LEMMA 1.16. *Let f be as in Theorem 1.15. For every $\alpha > 0$ there exists $r_0 = r_0(\alpha) > 0$ such that the cones $K^{s,u} = K^{s,u}_\alpha$ satisfy:*

(a) *for every $x \in B(\mathbf{0}, r_0)$ we have $Df(x)K^u \subset K^u$ and $Df(x)^{-1}K^s \subset K^s$.*

Moreover, for every χ such that $\max(|\lambda_s|, |\lambda_u|^{-1}) < \chi < 1$, there exist $\alpha_1, r_1 > 0$ such that if $\alpha \in (0, \alpha_1]$ and $r_2 \in (0, \min(r_0(\alpha), r_1)]$, then for every $x \in B(\mathbf{0}, r_2)$, we have:

(b) *if $v^{s,u} \in K^{s,u}$, then $\|Df(x)v^s\| \le \chi\|v^s\|$ and $\|Df^{-1}(x)v^u\| \le \chi\|v^u\|$.*

PROOF. Both (a) and (b) follow from continuity of $(x, v) \mapsto Df(x)v$ along with the fact that $L = Df(\mathbf{0})$ maps K^u strictly inside itself and contracts vectors in E^s by a factor of $|\lambda_s|$, while L^{-1} does the same with s and u swapped. □

The choices of constants in Lemma 1.16 are illustrated in Figure 1.21. The sequence of events here is very typical in hyperbolic dynamics:

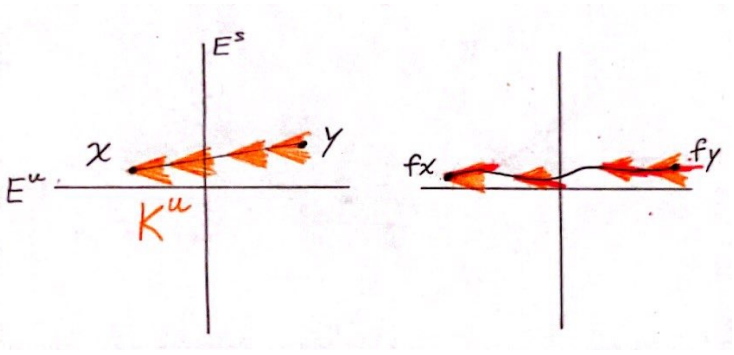
(1) we start with hyperbolicity rates $|\lambda_s|$ and $|\lambda_u|$ in the stable and unstable subspaces at $\mathbf{0}$;
(2) we weaken these rates slightly to get χ and χ^{-1};
(3) these weakened rates apply not just to $E^{s,u}$, but to the cones $K^{s,u}$ (provided α is small enough);
(4) they also apply not just at the reference point $\mathbf{0}$, but in a small neighborhood (provided r_2 is small enough).

FIGURE 1.21. Constants in Lemmas 1.16 and 1.17.

LEMMA 1.17. *Fix $f, r_0, \chi, \alpha_1, r_1, \alpha, r_2$ as in Theorem 1.15 and Lemma 1.16.*
(a) *Given $z, y \in B(\mathbf{0}, r_0)$ with $z - y \in K^u$, we have $f(z) - f(y) \in K^u$.*
(b) *Given $z, y \in B(\mathbf{0}, r_0)$ with $z - y \in K^s$, we have $f^{-1}(z) - f^{-1}(y) \in K^s$.*
(c) *Given $z, y \in B(\mathbf{0}, r_2)$ with $z - y \in K^u$, we have $\|f^{-1}(z) - f^{-1}(y)\| \le \chi \|z - y\|$.*
(d) *Given $z, y \in B(\mathbf{0}, r_2)$ with $z - y \in K^s$, we have $\|f(z) - f(y)\| \le \chi \|z - y\|$.*

PROOF. Part (a) follows from Lemma 1.16(a), as illustrated in Figure 1.22: writing $c\colon [0,1] \to B(\mathbf{0}, r_0)$ for a parametrization of the straight line between y and z, we see that $f \circ c$ parametrizes a C^1 curve from $f(y)$ to $f(z)$, all of whose tangent vectors lie in K^u, and thus $f(z) - f(y) \in K^u$. Part (b) is proved similarly.

FIGURE 1.22. Points separated by directions in K^u.



For part (d), observe that if c is any curve whose tangent vectors all lie in K^s, then Lemma 1.16(b) gives $\ell(f \circ c) \le \chi \ell(c)$, where ℓ denotes the length of a curve.

Taking c to be the straight line segment from y to z gives the result. Part (c) is proved similarly. □

PROOF OF THEOREM 1.15. We will prove the claims concerning W^u_r; the result for W^s_r proceeds analogously with f^{-1} replacing f. As in the discussion following (1.25), E^u is the stable fixed point of the linear map L acting as a contraction on the space of lines through the origin that lie in K^u. Each such line is the graph of a linear map $E^u \to E^s$. To obtain W^u_δ for the nonlinear map f, we will consider nonlinear maps $E^u \to E^s$.

First we need to determine our constants (recall Figure 1.21). Fixing a hyperbolicity rate $\chi \in \big(\max(|\lambda_s|, |\lambda_u|^{-1}), 1\big)$, we then choose a cone width $\alpha > 0$ and a scale $r_2 > 0$ such that Lemmas 1.16 and 1.17 hold. We also require α to be sufficiently small that

$$\frac{\chi^2(1+\alpha)}{(1-\alpha)^2} < 1; \tag{1.34}$$

then we put

$$\delta := r_2/(1+\alpha) \quad \text{and} \quad r := \delta \sin\theta, \text{ where } \theta = \measuredangle(E^u, E^s). \tag{1.35}$$

REMARK 1.18. Observe that the smaller the angle θ between E^u and E^s becomes, the smaller we must take r to be; this means we must zoom in even closer to $\mathbf{0}$ in order to control the nonlinear behavior.

FIGURE 1.23. Relationship between r, δ, and r_2.

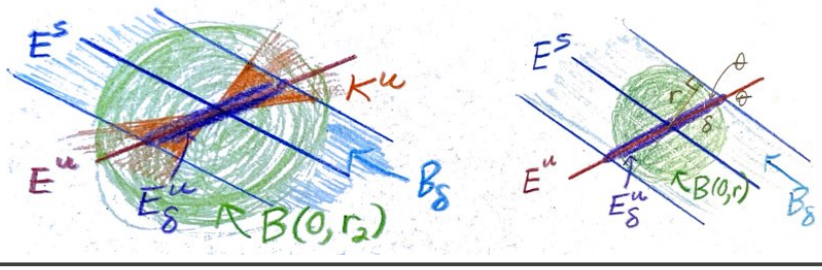


Figure 1.23 illustrates the consequences of (1.35): writing

$$E^u_\delta = \{x \in E^u : \|x\| \le \delta\} \quad \text{and} \quad B_\delta := E^u_\delta + E^s, \tag{1.36}$$

we have

$$B_\delta \cap K^u \subset B(\mathbf{0}, r_2) \quad \text{and} \quad B(\mathbf{0}, r) \subset B_\delta. \tag{1.37}$$

Now consider the set of α-Lipschitz maps

$$\mathrm{Lip}_\alpha = \{c\colon E^u_\delta \to E^s : c(\mathbf{0}) = \mathbf{0} \text{ and } \|c(x) - c(y)\| \le \alpha\|x - y\| \text{ for all } x, y \in E^u_\delta\},$$

Lec 8
Mon, Feb 3

equipped with the uniform distance

$$\rho(c_1, c_2) = \sup_{x \in E^u_\delta} \|c_1(x) - c_2(x)\|,$$

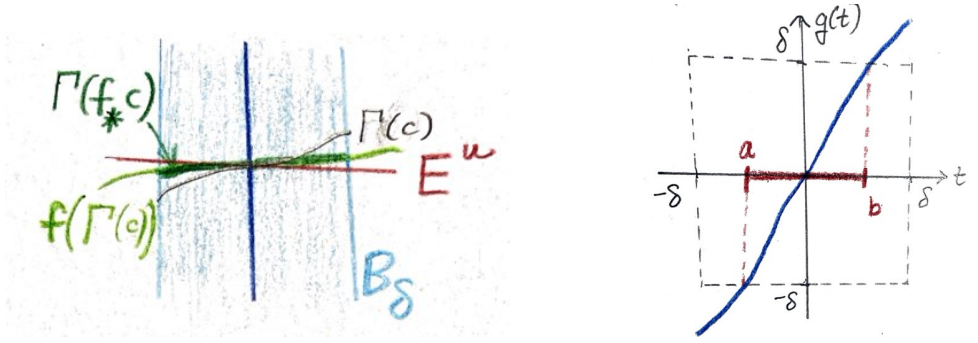
under which it is a complete metric space by the Arzelà–Ascoli theorem. To establish Theorem 1.15, we will prove the following.

(1) f induces a map from Lip_α to itself, called the *graph transform*.
(2) The graph transform is a contraction in the metric ρ, and thus has a unique fixed point $c^* \in \mathrm{Lip}_\alpha$.
(3) The graph of c^* intersected with $B(\mathbf{0}, r)$ is the set W^u_r defined in (1.33). In other words, given any $x \in B(\mathbf{0}, r)$, we have $\|x_u\| \le \delta$, so $c^*(x_u)$ is defined, and exactly one of the following two cases happens:
   - $x_s = c^*(x_u)$, and $\|f^{-n}(x)\| \le \chi^n \|x\|$ for all $n \ge 0$, or
   - $x_s \ne c^*(x_u)$, and $f^{-n}(x)$ eventually leaves $B(\mathbf{0}, r)$.
(4) The function c^* is C^1, not just Lipschitz, and $(Dc^*)(0) = 0$, so its graph is tangent to E^u at the origin.

STEP 1: *Defining the graph transform.*

Given $c \in \mathrm{Lip}_\alpha$, denote its graph by $\Gamma(c) := \{x_u + c(x_u) : x_u \in E^u_\delta\} \subset B_\delta$. Observe that $\Gamma(c) \subset B(\mathbf{0}, r_2)$ by (1.37). We will prove the following, which is illustrated in Figure 1.24.

FIGURE 1.24. The graph transform



PROPOSITION 1.19 (Graph transform). *Given f, χ, α, r_2 as in Lemma 1.17 and δ, r as in (1.35), there exists a unique map $f_*\colon \mathrm{Lip}_\alpha \to \mathrm{Lip}_\alpha$ such that $\Gamma(f_* c) = f(\Gamma(c)) \cap B_\delta$ for every $c \in \mathrm{Lip}_\alpha$.*

In the proof, we will use the following characterization of graphs of functions in Lip_α, which is immediate from the definitions.

LEMMA 1.20. *A set $Z \subset B_\delta$ can be written as $Z = \Gamma(c)$ for some $c \in \mathrm{Lip}_\alpha$ if and only if*

- $y - z \in K^u$ *for every* $y, z \in Z$, *and*
- $\pi(Z) = E^u_\delta$, *where* $\pi\colon \mathbb{R}^2 \to E^u$ *is projection along* E^s.

PROOF OF PROPOSITION 1.19. It suffices to show that given any $c \in \mathrm{Lip}_\alpha$, the set $Z := f(\Gamma(c)) \cap B(\mathbf{0}, r)$ satisfies the two criteria in Lemma 1.20. First, given $y, z \in Z$, there exist $p, q \in \Gamma(c) \subset B(\mathbf{0}, r_0)$ such that $f(p) = y$ and $f(q) = z$. Then $p - q \in K^u$, so by Lemma 1.17(a), we have $y - z = f(p) - f(q) \in K^u$. Thus the first criterion is satisfied.

To check the second criterion in Lemma 1.20, we will let $e_u \in E^u$ be a unit vector, and consider the continuous map $g\colon [-\delta, \delta] \to \mathbb{R}$ defined by

$$g(t) = \langle e_u,\, \pi(f(te_u + c(te_u)))\rangle,$$

which is illustrated in Figure 1.24. Observe that $g(0) = 0$; we must prove that $g([-\delta, \delta]) \supset [-\delta, \delta]$. Given any $s, t \in [-\delta, \delta]$, let

$$p := se_u + c(se_u) \quad \text{and} \quad q := te_u + c(te_u)$$

be the corresponding points on $\Gamma(c)$, so $p - q \in K^u$ and $f(p) - f(q) \in K^u$. Lemma 1.17(c) gives $\|p - q\| \le \chi \|f(p) - f(q)\|$, so by Lemma 1.13(b),

$$\|\pi(p) - \pi(q)\| \le \frac{\chi(1+\alpha)}{1-\alpha} \|\pi(f(p)) - \pi(f(q))\|.$$

Writing $\omega := \frac{1-\alpha}{\chi(1+\alpha)}$, this is equivalent to

$$\|se_u - te_u\| \le \omega^{-1} \|g(s)e_u - g(t)e_u\|,$$

and we conclude that

$$|g(s) - g(t)| \ge \omega |s - t|, \tag{1.38}$$

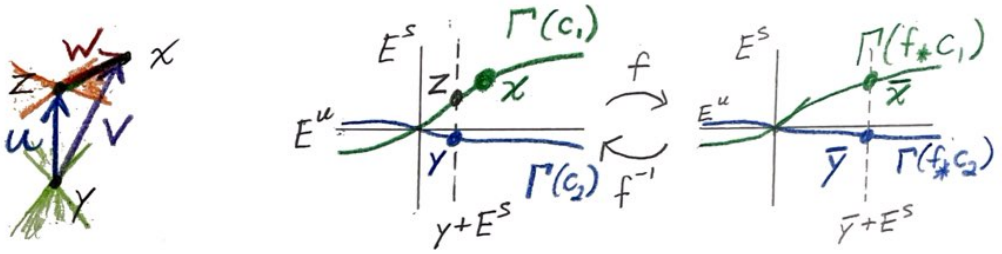
where $\omega > 1$ by (1.34). We say that g is *uniformly expanding*; in particular, it is injective. Since $g\colon [-\delta, \delta] \to \mathbb{R}$ is continuous and $g(0) = 0$, (1.38) yields $|g(\delta)| \ge \omega\delta$, and by the intermediate value theorem, there exists $b \in (0, \delta)$ such that $|g(b)| = \delta$; see Figure 1.24. Similarly, there exists $a \in (-\delta, 0)$ such that $|g(a)| = \delta$. Since $a \ne b$, we have $g(a) \ne g(b)$, so $\{g(a), g(b)\} = \{-\delta, \delta\}$, and one more application of the intermediate value theorem shows that $g\colon [a, b] \to [-\delta, \delta]$ is a bijection. This verifies the second criterion in Lemma 1.20, and completes the proof of Proposition 1.19. □

STEP 2: *The graph transform is a contraction.*

PROPOSITION 1.21. *Given* f, χ, α, r_2 *as in Lemma 1.17 and* δ, r *as in* (1.35), *the graph transform* $f_*\colon \mathrm{Lip}_\alpha \to \mathrm{Lip}_\alpha$ *has the property that*

$$\rho(f_*c_1, f_*c_2) \le \frac{\chi}{1-\alpha} \rho(c_1, c_2) \text{ for every } c_1, c_2 \in \mathrm{Lip}_\alpha\,. \tag{1.39}$$

FIGURE 1.25. Contraction under the graph transform.



PROOF. The argument is similar to the proof of Proposition 1.12; it is illustrated in Figure 1.25, which should be compared to Figure 1.19.

Given $c_1, c_2 \in \text{Lip}_\alpha$, consider the graph transforms f_*c_1 and f_*c_2, and choose any points $\bar{x} \in \Gamma(f_*c_1)$ and $\bar{y} \in \Gamma(f_*c_2)$ such that $\bar{x}_u = \bar{y}_u$. Let $x = f^{-1}(\bar{x}) \in \Gamma(c_1)$ and $y = f^{-1}(\bar{y}) \in \Gamma(c_2)$. Let $z = y_u + c_1(y_u)$ be the unique point in $\Gamma(c_1)$ with $z_u = y_u$, so that

$$\|z - y\| = \|z_s - y_s\| = \|c_1(y_u) - c_2(y_u)\| \le \rho(c_1, c_2). \tag{1.40}$$

Then $x - z \in K^u$ since $c_1 \in \text{Lip}_\alpha$, and since $\bar{x} - \bar{y} \in E^s \subset K^s$, Lemma 1.17(b) gives $x - y \in K^s$. Using Lemma 1.17(d), Lemma 1.14, and (1.40), we get

$$\|\bar{x} - \bar{y}\| \le \chi\|x - y\| \le \frac{\chi}{1-\alpha}\|z - y\| \le \frac{\chi}{1-\alpha}\rho(c_1, c_2),$$

and taking a supremum over $\bar{x}$ and $\bar{y}$ gives the result. □

We assumed in (1.34) that $\frac{\chi^2(1+\alpha)}{(1-\alpha)^2} < 1$, which implies that $\frac{\chi}{1-\alpha} < 1$, so Proposition 1.21 shows that f_* is a contraction on the complete metric space Lip_α. Thus the Banach Fixed Point Theorem 1.11 gives a unique $c^* \in \text{Lip}_\alpha$ such that $f_*c^* = c^*$.

STEP 3: *The fixed point of the graph transform is W^u_r.*

Let $c^* \in \text{Lip}_\alpha$ be the unique fixed point of the graph transform f_*. Recall from (1.33) that W^u_r is defined to be the set of $x \in \mathbb{R}^2$ such that $f^{-n}(x) \in B(\mathbf{0}, r)$ for all $n \ge 0$. We claim that

$$W^u_r = \Gamma(c^*) \cap B(\mathbf{0}, r), \tag{1.41}$$

and that every $x \in W^u_r$ has the property that

$$\|f^{-n}(x)\| \le \chi^n\|x\| \text{ for all } n \ge 0. \tag{1.42}$$

LEMMA 1.22. *For every $x \in \Gamma(c^*)$ and $n \ge 0$, we have $f^{-n}(x) \in \Gamma(c^*)$ and $\|f^{-n}(x)\| \le \chi^n\|x\|$.*

PROOF. For the first conclusion: the definition of the graph transform f_* in Proposition 1.19 gives $\Gamma(c^*) = f(\Gamma(c^*)) \cap B_\delta$, so $f^{-1}(\Gamma(c^*)) \subset \Gamma(c^*)$.

For the second conclusion: since $c^* \in \mathrm{Lip}_\alpha$, for every $x \in \Gamma(c^*)$ we have $x = x - \mathbf{0} \in K^u$, so (1.37) gives $x \in B(\mathbf{0}, r_2)$, and Lemma 1.17(c) gives

$$\|f^{-1}(x)\| = \|f^{-1}(x) - f^{-1}(\mathbf{0})\| \le \chi\|x - \mathbf{0}\| = \chi\|x\|. \tag{1.43}$$

Iterating (1.43) proves the lemma. □

Lec 9
Wed, Feb 5

It follows from Lemma 1.22 that $\Gamma(c^*) \cap B(\mathbf{0}, r) \subset W^u_r$. For the other inclusion, suppose that $x \in \bigcap_{k=0}^n f^k(B(\mathbf{0}, r))$, and write $x = x_u + x_s$, where $x_u \in E^u_\delta$ and $x_s \in E^s$. Let $y = x_u + c^*(x_u)$; then

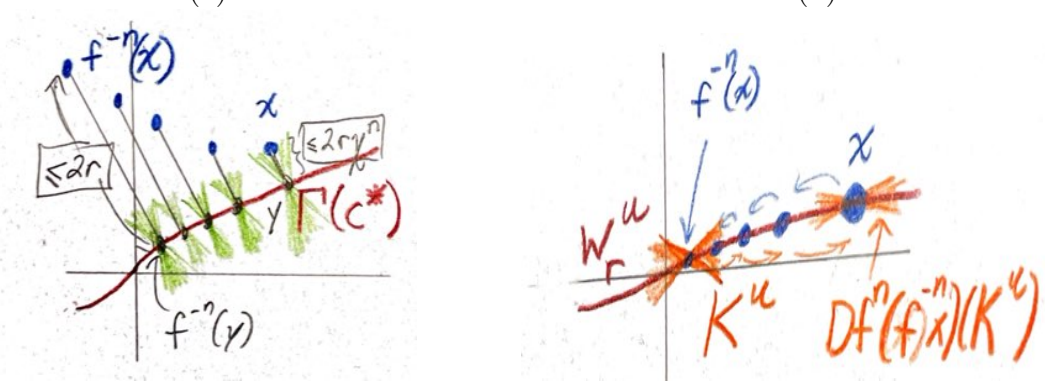
$$x - y = x_s - c^*(x_u) \in E^s \subset K^s.$$

For every $k = 0, 1, \dots, n$, $B(\mathbf{0}, r)$ contains $f^{-k}(x)$ by assumption and $f^{-k}(y)$ by Lemma 1.22. Thus $f^{-k}(x) - f^{-k}(y) \in K^s$ by Lemma 1.17(b), and applying Lemma 1.17(d) n times gives

$$\|x - y\| \le \chi^n \|f^{-n}(x) - f^{-n}(y)\| \le \chi^n \cdot 2r, \tag{1.44}$$

as shown in Figure 1.26(a). If $x \in W^u_r = \bigcap_{k=0}^\infty f^k(B(\mathbf{0}, r))$, then (1.44) holds for all $n \ge 0$, so $x = y \in \Gamma(c^*)$. Having proved (1.41), Lemma 1.22 shows that every $x \in W^u_r$ satisfies (1.42).

FIGURE 1.26. Information at x by pushing forward from $f^{-n}(x)$.



STEP 4: *The local unstable manifold is C^1, not just Lipschitz.*

Now we complete the proof of Theorem 1.15 by showing that $c^*\colon E^u_\delta \to E^s$ is continuously differentiable. To do this, we must show that for every $x \in E^u_\delta$ and $v \in E^u$, the limit

$$Dc^*(x)(v) = \lim_{t \to 0} \frac{1}{t}(c^*(x + tv) - c^*(x)) \in E^s \tag{1.45}$$

exists, and that it depends linearly on v and continuously on x.

The main idea is to use invariance of W^u_r to show that the set of limit points in (1.45) corresponds to a subset of $Df^n(f^{-n}x)(K^u)$, as shown in Figure 1.26(b), and that since Df contracts K^u uniformly, this set of limit points must be a single point. This will basically be a nonstationary version of the Banach Fixed Point Theorem 1.11.

DEFINITION 1.23. Given $x \in W^u_r$, say that $v \in \mathbb{R}^2$ is *tangent* to W^u_r at x if the line $x + \mathbb{R}v$ is a limit of secant lines in the following sense: there exist sequences $y_n, z_n \in W^u_r$ with $y_n \neq z_n$ and $a_n > 0$ such that

$$\lim_{n\to\infty} y_n = \lim_{n\to\infty} z_n = x \quad \text{and} \quad v = \lim_{n\to\infty} a_n(y_n - z_n). \tag{1.46}$$

LEMMA 1.24. *The* tangent set

$$T_xW^u_r = \{v \in \mathbb{R}^2 : v \text{ is tangent to } W^u_r \text{ at } x\} \tag{1.47}$$

has the following properties.

(a) *Homogeneity: if $v \in T_xW^u_r$, then $av \in T_xW^u_r$ for every $a \in \mathbb{R}$.*
(b) *For every $x \in W^u_r$, we have $\{\mathbf{0}\} \subsetneq T_xW^u_r \subset K^u$.*
(c) *$c^*\colon E^u_\delta \to E^s$ is differentiable at $y \in E^u_\delta$ if and only if for $x = y + c^*(y)$, the tangent set $T_xW^u_r$ is a one-dimensional subspace of $\mathbb{R}^2$.*
(d) *If $x \in f^{-1}(W^u_r)$, then $Df(x)(T_xW^u_r) = T_{f(x)}W^u_r$.*

PROOF. Conclusions (a) and (c) are immediate from the definition. For (b), observe that since $c^* \in \mathrm{Lip}_\alpha$, we have $y_n - z_n \in K^u$ for every $y_n, z_n \in W^u_r$, which implies that $T_xW^u_r \subset K^u$ since the cone is closed. The fact that $T_xW^u_r$ is nontrivial follows from compactness of $\mathbb{P}K^u$. Finally, (d) follows from the observation that by continuity of f, we have $y_n, z_n \to x$ if and only if $f(y_n), f(z_n) \to f(x)$, and in this case we have

$$\lim_{n\to\infty} a_n(f(y_n) - f(z_n)) = Df(x) \lim_{n\to\infty} a_n(y_n - z_n). \tag{1.48}$$

□

Now we can complete the proof of Theorem 1.15 by arguing that:

- for every $x \in W^u_r$, the $T_xW^u_r$ is a one-dimensional subspace of $\mathbb{R}^2$;
- the map $x \mapsto T_xW^u_r$ is continuous.

The key tool is Proposition 1.12, which guarantees that for every $y \in W^u_r$, the map $Df(y)\colon K^u \to K^u$ is a contraction in the metric defined in (1.27). Figure 1.26(b) illustrates the idea: applying Lemma 1.24(d) to $f^{-k}(x)$ for each $1 \le k \le n$, and using the fact that $T_{f^{-n}x}W^u_r \subset K^u$, we see that

$$T_xW^u_r \subset Df(f^{-1}x)Df(f^{-2}x)\cdots Df(f^{-n}x)K^u. \tag{1.49}$$

By Proposition 1.12, each of the maps $Df(f^{-k}x)$ contracts $\mathbb{P}K^u$ by a factor of $\gamma := \frac{\chi^2(1+\alpha)}{(1-\alpha)^2}$, so (1.49) gives

$$\operatorname{diam} T_xW^u_r \le \gamma^n \cdot 2\alpha.$$

This holds for all $n \ge 0$, so $\operatorname{diam} T_xW^u_r = 0$, and we conclude that $T_xW^u_r$ is a single line, establishing differentiability of c^*.

To establish continuity of Dc^*, we use an analogue of the solution to Exercise 1.13. Since f is C^1, for every $\delta > 0$ there exists $\beta > 0$ such that if $y, z \in W^u_r$ have $\|y - z\| < \beta$, then for every $v \in K^u$, we have $d(Df(y)[v], Df(z)[v]) \leq \delta$. Now suppose that $x, y \in W^u_r$ have $\|x - y\| < \beta$: we claim that

$$d(Df^n(f^{-n}x)E^u, Df^n(f^{-n}y)E^u) \leq \frac{\delta}{1-\gamma} \text{ for all } n \geq 0. \tag{1.50}$$

Sending $n \to \infty$, this will imply that $d(T_xW^u_r, T_yW^u_r) \leq \frac{\delta}{1-\gamma}$, completing the proof of Theorem 1.15.

To prove (1.50), fix $e^u \in E^u \setminus 0$, and for each $0 \leq k \leq n$, let

$$v_k := Df^k(f^{-k}(y))Df^{n-k}(f^{-n}(x))e^u.$$

By our choice of β and by the contraction property of Df on $\mathbb{P}K^u$, we have $d([v_k], [v_{k+1}]) \leq \delta\gamma^k$, so the triangle inequality gives

$$d([v_0], [v_n]) \leq \sum_{k=0}^{n-1} d([v_k], [v_{k+1}]) \leq \sum_{k=0}^{\infty} \delta\gamma^k.$$

Since $v_0 = Df^n(f^{-n}(x))e^u$ and $v_n = Df^n(f^{-n}(y))e^u$, this proves (1.50). □

Remark 1.25. The last step of the proof of Theorem 1.15 introduced an important idea: the mechanism behind the Banach Fixed Point Theorem continues to apply when we iterate a *sequence* of contractions, even if they are not all the same map. Later, we will use this to prove a more general version of the Hadamard–Perron Theorem.

Remark 1.26. Our proof of Theorem 1.15 follows *Hadamard's method*, built around the graph transform. There is also *Perron's method*, in which one considers the space of sequences in $B(\mathbf{0}, r)$ that converge to $\mathbf{0}$, and uses an implicit function theorem to identify those that are orbits of the system. One way of comparing the approaches is this: Hadamard's method works with true orbits that might leave the neighborhood, and searches for those that do not leave, while Perron's method works with sequences that stay in the neighborhood but might not be true orbits, and searches for those that are.

1.6. A transverse homoclinic intersection

Recalling Question 1.7, we have now constructed invariant curves – the stable and unstable manifolds – near hyperbolic fixed points. The Hadamard–Perron theorem itself only describes the behavior near the fixed point, but as was pointed out in §1.5, once we have the local stable and unstable manifolds $W^{s,u}_r$, we can define the

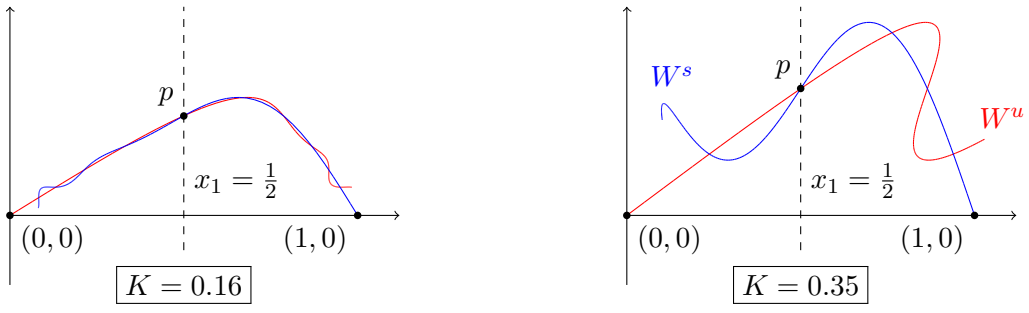
global stable and unstable manifolds

$$W^s := \bigcup_{n=0}^{\infty} f^{-n}(W^s_r) = \Big\{x \in \mathbb{R}^2 : \lim_{n\to\infty} f^n(x) = \mathbf{0}\Big\},$$
$$W^u := \bigcup_{n=0}^{\infty} f^{n}(W^u_r) = \Big\{x \in \mathbb{R}^2 : \lim_{n\to\infty} f^{-n}(x) = \mathbf{0}\Big\}. \tag{1.51}$$

For the pendulum, we saw in Figure 1.5 that these curves actually coincide, once one identifies the fixed points at $(n, 0)$ for all $n \in \mathbb{Z}$. Now we would like to understand whether this is true for the standard map (1.13) as well, or whether something different happens.

Our proof of the Hadamard–Perron theorem via the graph transform suggests that if V is any curve that approximates W^u_r, then $f^n(V)$ should be a good approximation of a bounded part of the global unstable manifold W^u. Figure 1.27 shows the result of this procedure, and its analogue for W^s, for the standard map (1.13) with parameter values $K = 0.16$ and $K = 0.35$. (Observe that the first of these values corresponds to the time step $\tau = 0.4$ in the discrete-time approximation to the pendulum from §1.2.3, which produced Figure 1.6; moreover, $0.16 \approx \frac{1}{2\pi}$, which was the value of K used in §1.3.)

FIGURE 1.27. W^u and W^s appear to cross, with angle increasing in K.



Looking at the picture, it appears that W^u and W^s do *not* coincide for the standard map, and that instead they cross each other repeatedly, including at a point p lying on the line $x_1 = \frac{1}{2}$. In the remainder of this section, we will prove that this is indeed the case. In the next section, we explore some consequences of this *transverse homoclinic intersection*.

PROPOSITION 1.27. *For every $K \geq 0.35$, the standard map (1.13) has the property that there exists a point p on the line $x_1 = \frac{1}{2}$ at which W^u and W^s intersect transversely: that is, $p \in W^u \cap W^s$, and $\mathbb{R}^2 = T_pW^u \oplus T_pW^s$.*

REMARK 1.28. In fact, Proposition 1.27 holds for *all* $K > 0$, but I am not aware of a simple argument that gives the result in this generality. By restricting to the case $K \geq 0.35$, we can give a fairly short proof. See Exercise 1.15 for an outline of

how to extend the argument at least a little bit further. Regarding small values of K, it turns out that the angle between W^u and W^s at p is very small as a function of K: there is a constant $C > 0$ such that this angle is less than $e^{-C/\sqrt{K}}$, so in particular it shrinks faster than any power of K.[10]

REMARK 1.29. Proposition 1.27 implies that f is not the time-t map of a flow. Indeed, the orbit of p under such a flow would be an invariant curve connection $\mathbf{0}$ to $(1,0)$, implying that $W^s = W^u$, which the proposition shows is not the case.

The rest of this section will be devoted to the proof of Proposition 1.27. We begin by observing a time reversal symmetry of the standard map, which shows that f and f^{-1} are topologically conjugate by an involution that exchanges $(0,0)$ and $(1,0)$, and fixes the line $x_1 = \frac{1}{2}$.

LEMMA 1.30. *Define* $h\colon \mathbb{R}^2 \to \mathbb{R}^2$ *by* $h(x_1, x_2) = (1 - x_1, x_2 + K \sin 2\pi x_1)$. *Then* h^2 *is the identity, and* $f^{-1} = h \circ f \circ h$.

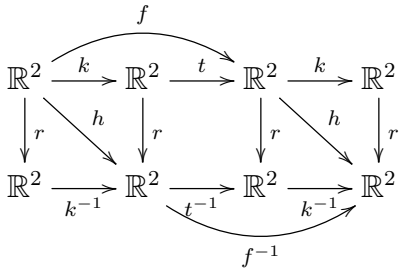
PROOF. Recall from the discussion before (1.13) that $f = t \circ k$, where

- $k(x_1, x_2) = (x_1, x_2 + K \sin 2\pi x_1)$ is the *kick map*, and
- $t(x_1, x_2) = (x_1 + x_2, x_1)$ is the *twist map*.

Let $r(x_1, x_2) = (1 - x_1, x_2)$ be reflection in the line $x_1 = \frac{1}{2}$; then $r^2 = \mathrm{Id}$ and $h = r \circ k$. Conjugating k and t by r gives their inverses:

$$\begin{aligned} r \circ t \circ r(x) &= r(t(1 - x_1, x_2)) = r(1 - x_1 + x_2, x_2) = (x_1 - x_2, x_2) = t^{-1}(x), \\ r \circ k \circ r(x) &= r(k(1 - x_1, x_2)) = r(1 - x_1, x_2 + K \sin 2\pi(1 - x_1)) \\ &= (x_1, x_2 - K \sin 2\pi x_1) = k^{-1}(x). \end{aligned}$$

From this we see that $h^2 = (r \circ k)^2 = \mathrm{Id}$, and that the following diagram commutes.



In particular, this implies that $f^{-1} \circ h = h \circ f$, and thus $f^{-1} = h \circ f \circ h$. □

Given $x \in \mathbb{R}^2$, it follows from Lemma 1.30 that $\lim_{n\to\infty} f^{-n}(x) = \mathbf{0}$ if and only if $\lim_{n\to\infty} f^n(h(x)) = (1,0)$, and similarly with $f^{\pm n}$ reversed. Using the characterization of the global stable and unstable manifolds in (1.51), we deduce that

$$h(W^u) = W^s \quad \text{and} \quad h(W^s) = W^u. \tag{1.52}$$

[10] See Lecture 14 of Sinai's "Topics in Ergodic Theory" (Princeton, 1994).

This has the following consequence.

(1.53) If $p \in W^u$ has $p_1 = \frac{1}{2}$, then $h(p) = p$, so $p \in W^u \cap W^s$.

Lec 10
Fri, Feb 7

(A similar result holds if we assume $p \in W^s$.) After demonstrating the existence of a point p satisfying (1.53), the proof of Proposition 1.27 will reduce to showing that the one-dimensional subspaces T_pW^u and T_pW^s do not coincide, which implies that $T_pW^u \oplus T_pW^s = \mathbb{R}^2$. By (1.52), we have $Dh(p)T_pW^u = T_pW^s$, so we can complete the proof by showing that T_pW^u is not an eigenspace of $Dh(p)$. Direct computation gives

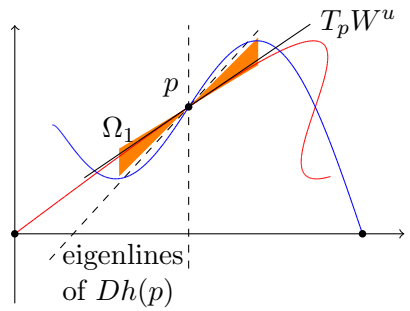
$$Dh(p) = \begin{pmatrix} -1 & 0 \\ -2\pi K & 1 \end{pmatrix},$$

which has eigenvalues ± 1 and corresponding eigenvectors $\binom{0}{1}$ and $\binom{1}{\pi K}$, so it will suffice to prove that

(1.54) T_pW^u does not contain either $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ or $\begin{pmatrix} 1 \\ \pi K \end{pmatrix}$.

So our task is to find a point $p \in W^u$ satisfying both (1.53) and (1.54). Figure 1.28 illustrates our strategy: we will exhibit a cone Ω_1 that does not contain either eigenline of $Dh(p)$, but which does contain all tangent lines T_xW^u when x lies on the initial part of W^u with $0 \le x_1 \le \frac{1}{2}$. Recall that we have fixed $K \ge 0.35$; this lower bound is not quite optimal for our proof, but will keep our computations simpler. Consider the cone

FIGURE 1.28. Proving transversality.



(1.55) $$\Omega_1 := \{\mathbf{0}\} \cup \left\{ v \in \mathbb{R}^2 \setminus 0 : 0.6 \le \frac{v_2}{v_1} \le 1 \right\}.$$

LEMMA 1.31. *Given any $q \in [0, \frac{1}{6}] \times \mathbb{R}$, we have $Df(q)\Omega_1 \subset \Omega_1$.*

PROOF. It suffices to prove that given $s \in [0.6, 1]$, the vector $\binom{v_1}{v_2} = Df(q)\binom{1}{s}$ has the property that $0.6 \le \frac{v_2}{v_1} \le 1$. Observe that

(1.56) $$Df(q) = \begin{pmatrix} 1 + 2\pi K \cos 2\pi q_1 & 1 \\ 2\pi K \cos 2\pi q_1 & 1 \end{pmatrix} = \begin{pmatrix} 1 + a & 1 \\ a & 1 \end{pmatrix},$$

where we write

$$a := 2\pi K \cos 2\pi q_1 \ge 2\pi K \cos \frac{\pi}{3} = \pi K \ge 1,$$

using the fact that cosine is decreasing on $[0, \pi]$ and $K \ge 0.35$. Thus

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = Df(q) \begin{pmatrix} 1 \\ s \end{pmatrix} = \begin{pmatrix} 1 + a & 1 \\ a & 1 \end{pmatrix} \begin{pmatrix} 1 \\ s \end{pmatrix} = \begin{pmatrix} a + s \\ 1 + a + s \end{pmatrix},$$

and since a and s are positive, we have

$$\frac{v_2}{v_1} = \frac{a+s}{1+a+s} = 1 - \frac{1}{1+a+s} < 1.$$

Moreover, $1+a+s \geq 1+1+0.6 > 2.5 = \frac{5}{2}$, so

$$\frac{v_2}{v_1} > 1 - \frac{1}{2.5} = 1 - \frac{2}{5} = \frac{3}{5} = 0.6,$$

which proves the lemma. □

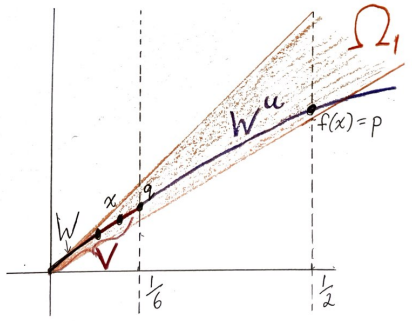
LEMMA 1.32. *Given any point $q \in \Omega_1$ with $q_1 \in [0, \frac{1}{4}]$, the point $x = f(q)$ has $x_1 \geq 3q_1$.*

PROOF. The function $r \mapsto \sin 2\pi r$ is concave and goes through the points $(0,0)$ and $(\frac{1}{4}, 1)$, so we have $\sin 2\pi r \geq 4r$ for all $r \in [0, \frac{1}{4}]$. Thus

$$x_1 = q_1 + q_2 + K \sin 2\pi q_1 \geq q_1 + 0.6q_1 + 4Kq_1 \geq q_1(1 + 0.6 + 1.4) = 3q_1. \quad \square$$

Now we prove Proposition 1.27 using Lemmas 1.31 and 1.32. Figure 1.29 shows the procedure. Let W_r^u be the local unstable manifold provided by Theorem 1.15, and let W be the part of W_r^u lying in $[0, \frac{1}{6}] \times \mathbb{R}$. (Note that although we did not get explicit bounds on r, one should not expect W_r^u to extend past the line $x_1 = \frac{1}{6}$.)

FIGURE 1.29. Extending W_r^u.



Since the cone Ω_1 is $Df(q)$ invariant for every $q \in W$, we have $T_qW \subset \Omega_1$ for all $q \in W$. This in turn implies that $W \subset \Omega_1$, so Lemma 1.32 guarantees that the x_1-footprint of W is expanded by a factor of at least 3 when we apply f. If $f(W)$ lies in $[0, \frac{1}{6}] \times \mathbb{R}$, we can repeat this argument, iterating until we obtain a curve $V \subset W^u$ whose endpoints are $\mathbf{0}$ and q with $q_1 = \frac{1}{6}$, and such that $T_xV \subset \Omega_1$ for all $x \in V$.

By Lemma 1.32, $f(q)$ lies to the right of the line $x_1 = \frac{1}{2}$, so by the intermediate value theorem, there exists $x \in V$ such that $f(x)$ lies on this line. Write $p = f(x)$. Lemma 1.31 implies that $T_pW^u = Df(x)T_xV \subset \Omega_1$. Observe that Ω_1 does not contain either $\binom{0}{1}$ or $\binom{1}{\pi K}$ (since $\pi K > 1$), so (1.54) holds, and p is the transverse homoclinic intersection that we sought. This completes the proof of Proposition 1.27.

►►► EXERCISE 1.15. Prove that the transverse homoclinic intersection in the left half of Figure 1.27 does indeed occur by using the following strategy to extend Proposition 1.27. What range of K can you prove the result for?

(1) Find a cone similar to Ω_1 that is invariant on $[0,t] \times \mathbb{R}$ for an appropriate value of t; use this to produce a curve $V \subset W^u$ tangent to K^u, whose endpoints are $\mathbf{0}$ and q with $q_1 = t$.

(2) Choose t and the cone such that $f^2(q)$ lies ot the right of the line $x_1 = \frac{1}{2}$, and thus there exists $x \in V$ for which $p := f^2(x)$ lies on this line.
(3) Estimate the slope of $T_pW^u = Df(y)Df(x)T_xV$, where $y = f(x)$. In particular, prove that this slope is $< \pi K$, so (1.54) holds.

1.7. Consequences of transversality

Remark 1.29 pointed out one consequence of the existence of a transverse homoclinic intersection: the standard map is not the time-t map of a flow. There are important consequences beyond this one, and we explore some of these now.

1.7.1. A homoclinic tangle. The curves W^u and W^s are invariant, so the set of points at which they intersect transversely is invariant as well:

LEMMA 1.33. *The curves W^u and W^s intersect transversely at $p \in \mathbb{R}^2$ if and only if they intersect transversely at $f(p)$.*

PROOF. Since $f(W^u \cap W^s) = f(W^u) \cap f(W^s) = W^u \cap W^s$, the curves intersect at p if and only if they intersect at $f(p)$. Moreover,

$$T_{f(p)}W^u \oplus T_{f(p)}W^s = Df(p)T_pW^u \oplus Df(p)T_pW^u = Df(p)(T_pW^u \oplus T_pW^s),$$

and since $Df(p)$ is invertible, we conclude that p is a transverse intersection if and only if $f(p)$ is. □

FIGURE 1.30. A homoclinic tangle.

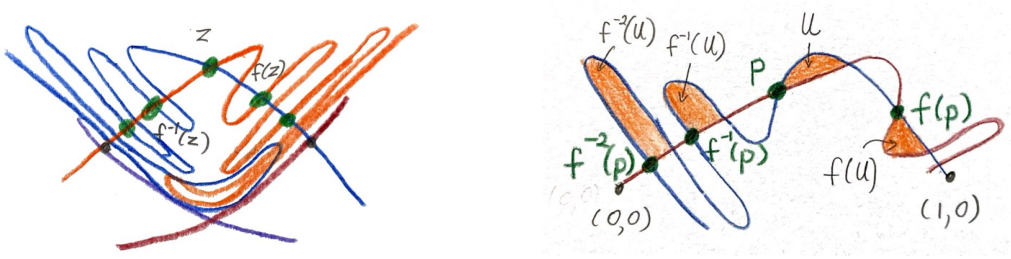


Figure 1.27 showed the initial consequences of this fact: moving along W^u away from $\mathbf{0}$, after passing through p the curve W^u starts winding back and forth so that it can pass through W^s at $f(p)$. The continuation of this process is shown in Figure 1.30. We can make the following observations regarding the resulting *homoclinic tangle*.

(1) By Lemma 1.33, the existence of a single transverse homoclinic intersection implies the existence of infinitely many such points.
(2) As the "lobes" $f^n(U)$ between successive intersections are mapped towards the fixed point, they appear to stretch farther and farther out along W^u (as $n \to \infty$) or W^s (as $n \to -\infty$).

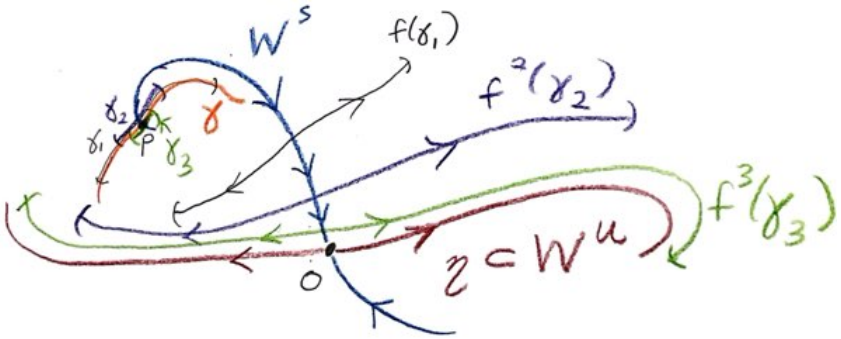
(3) Because $Df(x)$ has determinant 1 for every $x \in \mathbb{R}^2$ (recall the formula in (1.56)), the map f preserves area, and in particular, each of the lobes $f^n(U)$ has the same area.
(4) As the lobes stretch further out, we see the creation of "secondary" transverse homoclinic intersections.

Figure 1.30 only displays the beginnings of the complexity that arises as a result of a transverse homoclinic intersection. There are further levels of detail beyond the secondary intersections hinted at here. We will explore some of these in due course, while others lie beyond the scope of this book.

▶ EXERCISE 1.16. Use the fact that the eigenvalues of $Df(\mathbf{0})$ are both positive to prove that as we move along W^u from p to $f(p)$, we must encounter another transverse homoclinic intersection that is not on the orbit of p.

1.7.2. The Inclination Lemma. Let us examine the second of the four observations in the previous section: the lobes $f^n(U)$ appear to stretch out along W^u as $n \to \infty$, and along W^s as $n \to -\infty$. This idea is made precise by the following result, commonly called the *Inclination Lemma* or *Lambda Lemma*, which is illustrated in Figure 1.31.

FIGURE 1.31. The Inclination Lemma.



LEMMA 1.34. *Let $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a C^1-diffeomorphism satisfying the conditions of the Hadamard–Perron Theorem 1.15.*

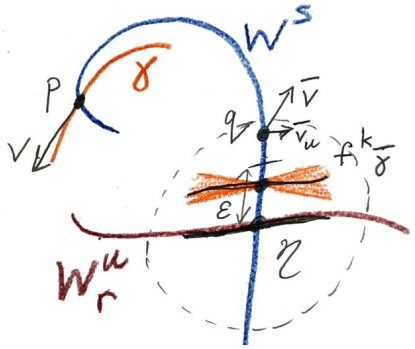
(1) For every $\ell \in \mathbb{N}$ and $\theta > 0$, there exists $n \geq \ell$ such that given any $p \in f^{-\ell}W^s_r$ and any $v \in \mathbb{R}^2$ with $\measuredangle(v, T_pW^s) \geq \theta$, we have $Df^n(p)v \in K^u$.
(2) Given any C^1 curve $\gamma \subset \mathbb{R}^2$ that intersects W^s transversely at a point $p \in \gamma \cap W^s$, and any bounded curve $\eta \subset W^u$, there exists a sequence of curves $\gamma_k \subset \gamma$ such that $f^k\gamma_k \to \eta$ in the C^1 topology.

Lec 11
Mon, Feb 10

▶▶ EXERCISE 1.17. The second part of Lemma 1.34 appears to validate the comment in §1.6 about the procedure for generating Figure 1.27: start with a curve that approximates W^u_r, then iterate it forward to get a good approximation of part of W^u.

However, some caution is warranted: explain how it could happen that even if $\gamma \approx W^u_r$, the curves $f^n\gamma$ might have parts that do not approximate anything in W^u.

FIGURE 1.32. Proving the Inclination Lemma.



PROOF OF LEMMA 1.34. For the first statement, we start by observing that there exists an angle $\zeta > 0$ such that if $v \in \mathbb{R}^2$ satisfies $\measuredangle(v, E^u) \le \zeta$, then $v \in K^u$. Moreover, there exists $r_0 > 0$ such that given any $q \in W^s_{r_0}$, we have $\measuredangle(Df^m(q)E^u, E^u) < \zeta/2$ for all $m \ge 0$. Let $j \in \mathbb{N}$ be sufficiently large that $r\chi^j < r_0$.

Now we claim that there exists $\bar\theta = \bar\theta(\ell, \theta) > 0$ such that given any p, v as in the lemma, if we write $q := f^{\ell+j}(p) \in W^s_{r_0}$ and $\bar v := Df^{\ell+j}(p)v$, then we have $\measuredangle(\bar v, T_qW^s) \ge \bar\theta$. This claim follows from a compactness argument: on the compact set

$$X = \{(p, v, w) \in f^{-\ell}(W^s_r) \times S^1 \times S^1 : \measuredangle(v, w) \ge \theta\},$$

we define the function $(p, v, w) \mapsto \measuredangle(Df^{\ell+j}(p)v, Df^{\ell+j}(p)w)$, which is continuous (since f is C^1) and positive (since f is invertible), thus it admits a positive lower bound $\bar\theta > 0$.

Now we use the decomposition $\mathbb{R}^2 = T_qW^s \oplus E^u$ to write $\bar v = \bar v_s + \bar v_u$, where $\bar v_s \in T_qW^s$ and $\bar v_u \in E^u$. As shown in Figure 1.32, we have $\bar v_u \neq 0$ because $\bar v \notin T_qW^s_r$; in fact, since $\measuredangle(\bar v, T_qW^s) \ge \bar\theta$, there exists $\beta = \beta(\theta, \ell)$ such that $\|\bar v_u\| \ge \beta\|\bar v_s\|$.

Pushing $\bar v$ forward under $Df^m(q)$, we see that

$$\underbrace{Df^{m+\ell+j}(p)v}_{w} = \underbrace{Df^m(q)\bar v_s}_{w^1} + \underbrace{Df^m(q)\bar v_u}_{w^2}.$$

With the notation indicated, we have $w^1 \in K^s$ and $w^2 \in K^u$; moreover,

$$\|w^1\| \le \chi^m\|\bar v_s\| \le \chi^m\beta^{-1}\|\bar v_u\| \le \chi^{2m}\beta^{-1}\|w^2\|,$$

which implies that

$$\measuredangle(w, w^2) \le \arctan(\chi^{2m}\beta^{-1}).$$

Choosing m sufficiently large that the right-hand side is $< \zeta/2$, and recalling that by our choice of r_0 we have $\measuredangle(w^2, E^u) < \zeta/2$, we obtain $\measuredangle(w, E^u) < \zeta$, so $w \in K^u$. Observe that m only depends on χ and β.

For the second part of Lemma 1.34, it suffices to prove the result when $\eta = W^u_r$. Indeed, if $\gamma_n \subset \gamma$ are such that $f^n\gamma_n \to W^u_r$, then given any bounded curve $\eta \subset W^u$, there exists $\ell \in \mathbb{N}$ such that $f^{-\ell}\eta \subset W^u$. Thus there are curves $\bar\gamma_n \subset \gamma_n$ such that $f^n\bar\gamma_n \to f^{-\ell}\eta$, and since f is C^1, this gives $f^{n+\ell}\bar\gamma_n \to \eta$.

This vector is tangent to $f^n(\gamma)$ at $f^n(p)$, and since the curve is C^1, there is a subcurve $\bar\gamma \subset \gamma$ such that $f^n(\bar\gamma)$ is "u-admissible" in the sense that all of its tangent vectors lie in K^u. Pushing this curve forward and restricting to an appropriate subcurve, we can approximate W^u_r arbitrarily well in the C^1 topology. □

►► EXERCISE 1.18. Fill in the details of the last step in the proof of the Inclination Lemma 1.34, by using the techniques developed in the proof of the Hadamard–Perron Theorem 1.15 to prove the following.

(1) Given any u-admissible curve $\zeta \subset B(\mathbf{0}, r)$, there exists $k \in \mathbb{N}$ such that $f^n(\zeta)$ completely crosses the strip B_δ from (1.36): in particular, $f^k(\zeta) \cap B_\delta$ is the graph of an α-Lipschitz function $c\colon E^u_\delta \to E^s$.
(2) The definition of the graph transform f_* can be extended from Lip_α to the space of all α-Lipschitz functions $c\colon E^u_\delta \to E^s$ for which $c(0)$ is sufficiently small, and it is a contraction on this space.
(3) The sequence $f^n_* c$ converges to the fixed point c^* not only in the uniform metric, but also in the C^1 topology (mimic Step 4 of the proof).

1.7.3. A horseshoe. Returning now to the specific example of the Chirikov–Taylor standard map, the numerically-generated orbit labeled "C" in Figure 1.6 suggested that something interesting is happening to orbits that pass through the region lying just above the fixed point. We are now in a position to give a better explanation of this.

Recall that one of our original motivating questions concerned the magnitude of the displacement $\Delta_n := f^n(x) - f^n(y)$ between two orbits of f, and in particular, how to control $\|\Delta_n\|$ in terms of $\|\Delta_0\|$. Figure 1.9 and the discussion in §§1.2.3–1.4, especially Lemma 1.8, suggested that this error term might grow exponentially fast when the orbits remain near the hyperbolic fixed point at $\mathbf{0}$. This was verified rigorously in §1.5, and we now formulate a precise statement clarifying this, after first setting up some notation.

Let $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a C^1-diffeomorphism satisfying the conditions of the Hadamard–Perron Theorem 1.15. Let $r_0, \chi, \alpha_1, r_1, \alpha, r_2 > 0$ be as in Lemma 1.16, and let $\delta = \frac{1}{2}r_2$. With E^u_δ as in (1.36), let

$$R = R_\delta := E^u_\delta + E^s_\delta = \{x_u + x_s : x_u \in E^u, x_s \in E^s, \|x_u\| \le \delta, \|x_s\| \le \delta\}.$$

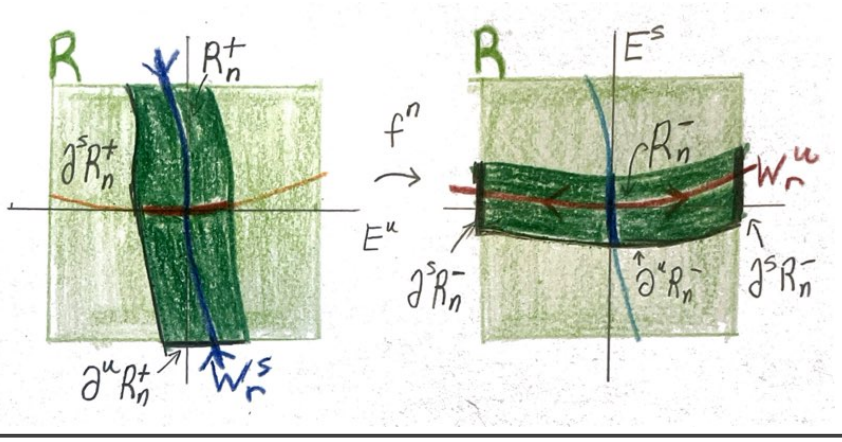
It is customary to refer to R as a "rectangle", as shown in Figure 1.33, although if E^u and E^s are not orthogonal then geometrically it is only a parallelogram. (We will give a more careful definition of dynamical rectangles in the next section.) Two of the sides of R are line segments parallel to E^u; we refer to their union as *unstable boundary* of R, and denote it by

$$\partial^u R := \{x_u + x_s : x_u \in E^u, x_s \in E^s, \|x_u\| \le \delta, \|x_s\| = \delta\}.$$

Interchanging the roles of s and u gives the *stable boundary* $\partial^s R$.

DEFINITION 1.35. In the setting of the Hadamard–Perron Theorem 1.15, we say that a continuous curve $\gamma \subset B(\mathbf{0}, r)$ is *u-admissible* if $x - y \in K^u$ for every $x, y \in \gamma$, and *s-admissible* if $x - y \in K^s$ for every $x, y \in \gamma$.

FIGURE 1.33. Orbits remaining near $\mathbf{0}$ for n iterates.



LEMMA 1.36. *Let $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a C^1-diffeomorphism satisfying the conditions of the Hadamard–Perron Theorem 1.15, and let R be as above. For each $n \in \mathbb{N}$, consider the sets*

$$\begin{aligned} R_n^+ &:= \{x \in R : f^k x \in R \text{ for all } 0 \le k \le n\} = \bigcap_{k=0}^{n} f^{-k} R, \\ R_n^- &:= \{x \in R : f^{-k} x \in R \text{ for all } 0 \le k \le n\} = \bigcap_{k=0}^{n} f^{k} R, \end{aligned} \tag{1.57}$$

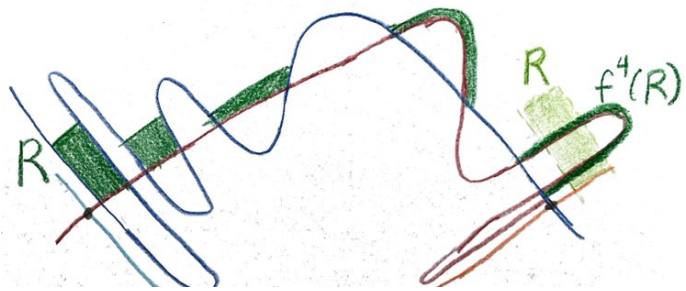
for which f maps $R_n^+ \to R_n^-$ bijectively. Let $\partial^u R_n^+ = R_n^+ \cap \partial^u R$ and $\partial^s R_n^- = R_n^- \cap \partial^s R$. Let $\partial^u R_n^- := f^n \partial^u R_n^+$ and $\partial^s R_n^+ := f^{-n} \partial^s R_n^-$. Then the following are true.

(1) Each of $\partial^u R_n^+$ and $\partial^s R_n^-$ is the union of two disjoint line segments.
(2) The set $\partial^u R_n^-$ is the union of two disjoint u-admissible curves whose endpoints lie on $\partial^s R_n^-$, and the set $\partial^s R_n^+$ is the union of two disjoint s-admissible curves whose endpoints lie on $\partial^u R_n^+$.
(3) Each of $R_n^\pm$ is a compact set that is the closure of its interior, and whose boundary is $\partial R_n^\pm = \partial^u R_n^\pm \cup \partial^s R_n^\pm$.
(4) If $x, y \in R_n^+$ have $x - y \in K^u$, then $\|x - y\| \le \chi^n \|f^n x - f^n y\| \le 4\delta\chi^n$.
(5) If $x, y \in R_n^-$ have $x - y \in K^s$, then $\|x - y\| \le \chi^n \|f^{-n} x - f^{-n} y\| \le 4\delta\chi^n$.

►► EXERCISE 1.19. Prove Lemma 1.36 using the ideas from the proof of the Hadamard–Perron Theorem 1.15 in §1.5.

Lemma 1.36 rigorously justifies the comments preceding Remark 1.9: for a given measurement error, the forecast error grows exponentially fast in n. However, there is an important caveat: this only applies as long as we remain near $\mathbf{0}$, and in particular, the only orbit near which Lemma 1.36 guarantees this exponential divergence for all time (both forward and backward) is $\mathbf{0}$ itself! This raises the natural question of whether we see a similar divergence near any other orbits. It turns out that we do; we will use the Inclination Lemma to prove an analogue of Lemma 1.36 for orbits remaining close to the orbit of a transverse homoclinic point.

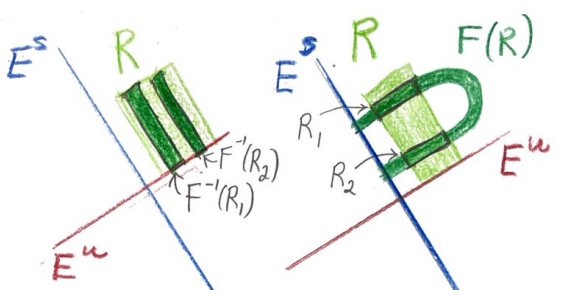
FIGURE 1.34. An admissible rectangle and a horseshoe.



As a first illustration of how this occurs, consider the standard map and let R be the "rectangular" region illustrated in Figure 1.34. Its sides are not quite lines: two of them are pieces of W^s in the region where it has developed rather large back-and-forth oscillations, and these pieces are close to W^s_r by the Inclination Lemma; the third side is a piece of the local unstable manifold W^u_r; and the fourth is a short curve roughly parallel to W^u_r, on the opposite sides to one of the lobes.

As we iterate R, Figure 1.34 shows how these two "unstable" curves are expanded and come closer together. After passing p, they start to bend, and ultimately $f^4(R)$ is a kind of "horseshoe" shape. Because the phase space is periodic in the horizontal direction, $f^4(R)$ intersects R as subsets of the cylinder, crossing it twice.

FIGURE 1.35. A "standard" horseshoe.



The essential aspects of this picture will appear anytime there is a transverse homoclinic intersection, producing what we will call a horseshoe. Roughly speaking, this will include a region R ("rectangle") bounded by curves close to the stable and unstable directions, together with an iterate $F = f^\ell$ such that each of $F^{-1}(R) \cap R$ and $R \cap F(R)$ is the disjoint

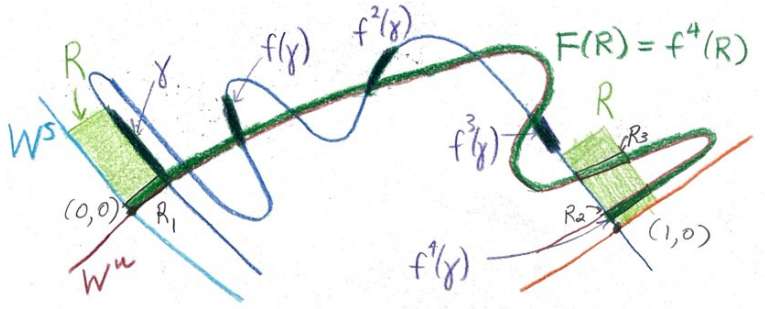
Lec 12
Wed, Feb 12

union of narrower rectangles ("branches"), as in Figure 1.35, on which the map F exhibits hyperbolic behavior similar to what is described in Lemma 1.36.

While the choice of rectangle here has the effect of making $F(R)$ look like an actual horseshoe, providing some justification for the terminology, other choices of rectangle lead to horseshoes that would be less practical in equine applications, such as the one in Figure 1.36. Observe that this horseshoe has three branches: one appears near $\mathbf{0}$ and is essentially the branch provided by Lemma 1.36, while the other two appear in the copy of R near $(1, 0)$.

▶ EXERCISE 1.20. Figure 1.36 gives the impression that R_1 and R_2 could possibly overlap as subsets of the cylinder. Prove that this is impossible.

FIGURE 1.36. Another horseshoe in the standard map.



In the next section, we will define the words "rectangle", "horseshoe", and "branches" more carefully. For now, we use the Inclination Lemma 1.34 to prove an analogue of Lemma 1.36 for orbits that remain near the orbit of a transverse homoclinic intersection.

LEMMA 1.37. *Let f, R be as in Lemma 1.36, and suppose that W^u and W^s intersect transversely at a point p. Let $R_n^{\pm}(\mathbf{0})$ denote the sets constructed in* (1.57). *For given values of $k, n \in \mathbb{N}$ and $\epsilon > 0$, consider the sets*

$$R_n^+(p) := R_n^+(\mathbf{0}) \cap f^{-(n+k)}(B(p, \epsilon)) \cap f^{-2(n+k)} R_n^-(\mathbf{0}),$$
$$R_n^-(p) := f^{2(n+k)} R_n^+(p),$$

which are shown in Figure 1.37. Then there are $k \in \mathbb{N}$ and $\epsilon > 0$, $C > 0$ such that for every sufficiently large $n \in \mathbb{N}$, the map $F := f^{2(n+k)} \colon R_n^+(p) \to R_n^-(p)$ has the property that

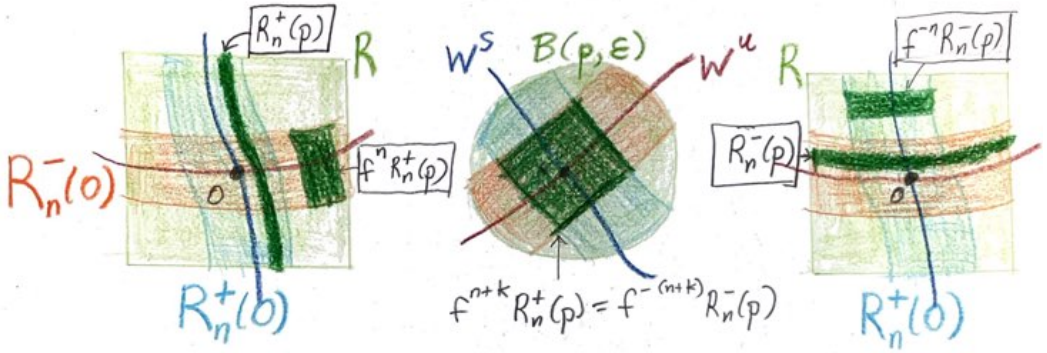
$$\begin{aligned} DF(x)(K^u) &\subset K^u \text{ for every } x \in R_n^+(p), \\ DF^{-1}(y)(K^s) &\subset K^s \text{ for every } y \in R_n^-(p), \end{aligned} \tag{1.58}$$

and moreover, the sets $R_n^{\pm}(p)$ satisfy the conclusions of Lemma 1.36, if we define $\partial^{u,s} R_n^{\pm}(p)$ as in that lemma, and replace the inequalities in the last two statements

of that lemma with:

$$\|x-y\| \le C\chi^{2n}\|Fx-Fy\| \le 4Cr\chi^{2n} \text{ when } x,y \in R_n^+(p),\ x-y \in K^u,$$
$$\|x-y\| \le C\chi^{2n}\|F^{-1}x-F^{-1}y\| \le 4Cr\chi^{2n} \text{ when } x,y \in R_n^-(p),\ x-y \in K^s.$$

FIGURE 1.37. A branch of $R \cap f^{-N}(R)$ that follows the orbit of p.



PROOF. Let $\ell \in \mathbb{N}$ be sufficiently large that $f^{-\ell}(p) \in W_r^u \cap R$ and $f^\ell(p) \in W_r^s \cap R$. By the Inclination Lemma 1.34, there exists $k \ge \ell$ such that

$$Df^k(p)T_pW^u \subset K^u_{\alpha/4} \quad \text{and} \quad Df^{-k}(p)T_pW^s \subset K^s_{\alpha/4}. \tag{1.59}$$

Writing $\bar{x} = f^{-k}(p) \in W_r^u \cap R$ and $\bar{y} = f^k(p) \in W_r^s \cap R$, we obtain

$$Df^{2k}(\bar{x})T_{\bar{x}}W^u \subset K^u_{\alpha/4} \quad \text{and} \quad Df^{-2k}(\bar{y})T_{\bar{y}}W^s \subset K^s_{\alpha/4}. \tag{1.60}$$

Fix $\theta > 0$ sufficiently small that

$$\begin{aligned} &\text{if } v \in \mathbb{R}^2 \text{ has } \measuredangle(v, T_{\bar{x}}W^u) < \theta, \text{ then } Df^{2k}(\bar{x})v \in K^u_{\alpha/2}, \\ &\text{if } v \in \mathbb{R}^2 \text{ has } \measuredangle(v, T_{\bar{y}}W^s) < \theta, \text{ then } Df^{-2k}(\bar{y})v \in K^s_{\alpha/2}. \end{aligned} \tag{1.61}$$

Then use continuity of $f^{\pm k}$ and $Df^{\pm 2k}$ to fix $\epsilon > 0$ sufficiently small that given any $q \in B(p,\epsilon)$, the points $x = f^{-k}(q)$ and $y = f^k(q)$ have the property that

$$\begin{aligned} &\text{if } v \in \mathbb{R}^2 \text{ has } \measuredangle(v, T_{\bar{x}}W^u) < \theta, \text{ then } Df^{2k}(x)v \in K^u_\alpha = K^u, \\ &\text{if } v \in \mathbb{R}^2 \text{ has } \measuredangle(v, T_{\bar{y}}W^s) < \theta, \text{ then } Df^{-2k}(y)v \in K^s_\alpha = K^s. \end{aligned} \tag{1.62}$$

Decreasing $\epsilon > 0$ if necessary, we may assume that given any $x \in W_r^u \cap f^{-k}(B(p,\epsilon))$, we have $\measuredangle(T_xW^u, T_{\bar{x}}W^u) < \theta/2$, and similarly for the stables.

Now observe that as $n \to \infty$, the sets $R_n^-(\mathbf{0})$ shrink in the stable direction, and their intersection is $W_r^u \cap R$; similarly, the sets $R_n^+(\mathbf{0})$ shrink in the unstable direction, and their intersection is $W_r^s \cap R$. In particular $\bigcap_{n=0}^\infty f^k(R_n^-(\mathbf{0})) = f^k(W_r^u \cap R)$, while $\bigcap_{n=0}^\infty f^{-k}(R_n^+(\mathbf{0})) = f^{-k}(W_r^s \cap R)$. The intersection of these two sets is the single point p. Fix $n \in \mathbb{N}$ sufficiently large that

$$\big(f^k(R_n^-(\mathbf{0})) \cap f^{-k}(R_n^+(\mathbf{0}))\big) \subset B(p,\epsilon), \tag{1.63}$$

and such that we also have:

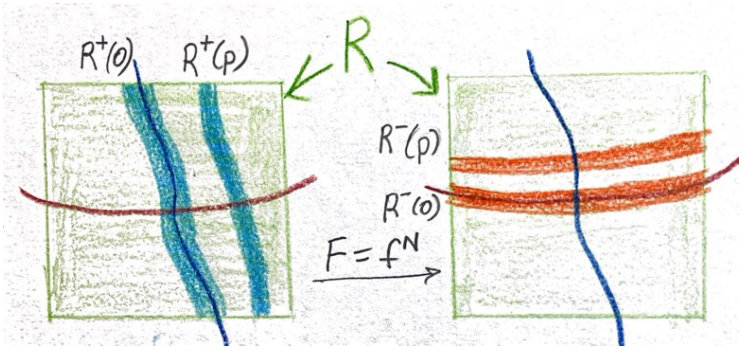
$$\text{if } x \in R_n^+(\mathbf{0}) \cap f^{-(n+k)}B(p,\epsilon) \text{ and } v \in K^u, \text{ then } \measuredangle(Df^n(x)v, T_{\bar{x}}W^u) < \theta,$$
$$\text{if } y \in R_n^-(\mathbf{0}) \cap f^{n+k}B(p,\epsilon) \text{ and } v \in K^s, \text{ then } \measuredangle(Df^{-n}(x)v, T_{\bar{y}}W^s) < \theta.$$

Combining these with (1.62), we obtain the cone invariance claimed in (1.58). The remaining conclusions from Lemma 1.36, appropriately modified, are left as an exercise. □

►► EXERCISE 1.21. Complete the proof of Lemma 1.37.

With k, n as in Lemma 1.37 and writing $N = 2(k+n)$, $F = f^N$, Figure 1.38 illustrates the relationship between the sets $R^\pm(\mathbf{0}) := R_N^\pm(\mathbf{0})$ from Lemma 1.36 and $R^\pm(p) := R_n^\pm(p)$ from Lemma 1.37. In the next section, we will explore the consequences of this picture.

FIGURE 1.38. Two branches of $R \cap f^{-N}(R)$.



1.8. Linear and nonlinear horseshoes

Lec 13
Fri, Feb 14

1.8.1. Describing a linear horseshoe. The basic local picture of hyperbolic dynamics is a hyperbolic fixed point, and §§1.3–1.5 were devoted to the following tasks:

(1) study the linear model at a hyperbolic fixed point and identify the main dynamical phenomena;
(2) use the cones $K^{u,s}$ to prove that analogues of these phenomena persist in the original nonlinear system.

In §1.5, the main phenomenon we studied was the existence of local stable and unstable manifolds. Lemma 1.36 in the previous section reframed the discussion in terms of the dynamics of orbits that remain near the fixed point. In this case the linear model is of a "vertical" rectangle being mapped to a "horizontal" one as shown in Figure 1.39, by a linear map that expands horizontal vectors by a factor of $\lambda_u > 1$, and contracts vertical vectors by a factor of $\lambda_s \in (0,1)$. Lemma 1.36 gave a nonlinear version of this model.

Lemma 1.37 showed that this same model appears for orbits that remain near the orbit of a transverse homoclinic intersection, so that as Figure 1.38 illustrates, we in fact have two "hyperbolic branches" crossing R: one corresponding to orbits that stay near $\mathbf{0}$, and one to orbits that follow the orbit of p. This leads us to the basic *global* picture of hyperbolic dynamics, which is called a *horseshoe* after the shape illustrated in Figure 1.35.

FIGURE 1.39. The basic linear models of hyperbolic dynamics.

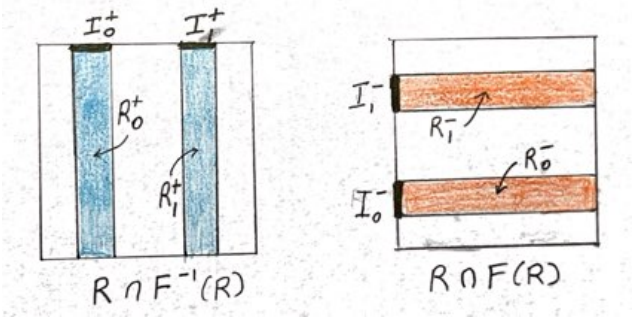
Local: hyperbolic fixed point Global: horseshoe

In this section and §1.8.2, we will study the *linear* model of a horseshoe shown in Figure 1.39 to see what kind of dynamical behavior it implies; then in §1.8.3, we will prove that this behavior also appears for the nonlinear system.

We start by describing a linear horseshoe more precisely. Let $R = [0,1]^2$ be the unit square, and let $F\colon R \to \mathbb{R}^2$ be a C^1 diffeomorphism onto its image, with the following properties, illustrated in Figures 1.39 and 1.40.

(1) $F(R) \cap R$ is the disjoint union of rectangles $R_0^- = [0,1] \times I_0^-$ and $R_1^- = [0,1] \times I_1^-$, where $I_0^-, I_1^- \subset [0,1]$ are disjoint closed intervals.
(2) For each $j \in \{0,1\}$, the set $R_j^+ := F^{-1}R_j^- \subset R$ is a rectangle of the form $I_j^+ \times [0,1]$ for some closed interval $I_j^+ \subset [0,1]$.
(3) On each R_j^+, we have $F(x_1, x_2) = (g_j x_1, h_j x_2)$, where each $g_j\colon I_j^+ \to [0,1]$ is an affine homeomorphism, and similarly for each $h_j\colon [0,1] \to I_j^-$.

FIGURE 1.40. The fundamental branches of a linear horseshoe.

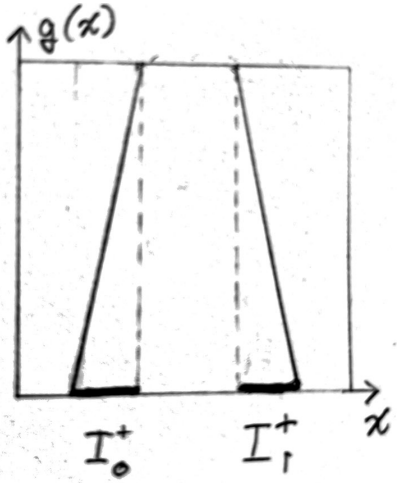


REMARK 1.38. The linear horseshoe illustrated in Figure 1.39 is often called a *Smale horseshoe* after Stephen Smale; see §1.9 for more on the history. One could also consider horseshoes with more than two branches, or where the part of $F(R)$ outside of R behaves differently.

▶ EXERCISE 1.22. Suppose we have $I_0^+ = I_0^- = [\frac{1}{6}, \frac{1}{3}]$ and $I_1^+ = I_1^- = [\frac{2}{3}, \frac{5}{6}]$. Find all fixed points of F, and all points of period 2.

We can define a map $g\colon I_0^+ \cup I_1^+ \to [0,1]$ by $g|_{I_j^+} = g_j$ for $j \in \{0,1\}$, as shown in Figure 1.41. Observe that because each basic interval $I_j^\pm$ has length < 1, the maps g_j and h_j^{-1} expand distances, while g_j^{-1} and h_j contract distances. It is common to assume that there are constants $\lambda_u > 2$ and $\lambda_s \in (0, \frac{1}{2})$ such that each g_j has slope $\pm\lambda_u$ and each h_j has slope $\pm\lambda_s$, and most of our pictures will be drawn as if this is the case.

FIGURE 1.41. Basic intervals.



Some points in R are mapped outside of R by F, and thus cannot be iterated more than once. Similarly, not every point in R has a preimage, so not all points have backwards orbits. We will study the set of points Λ that have infinite forward and backwards orbits. To this end, first consider the sets

$$\begin{aligned} \Lambda_n^+ &:= \{x \in R : F^k(x) \in R \text{ for all } 0 \le k \le n\} = \bigcap_{k=0}^{n} F^{-k}(R), \\ \Lambda_n^- &:= \{x \in R : F^{-k}(x) \in R \text{ for all } 0 \le k \le n\} = \bigcap_{k=0}^{n} F^{k}(R). \end{aligned} \tag{1.64}$$

(Compare this to the definition of $R_n^\pm$ in (1.57).) That is, Λ_n^+ consists of all points that have n forwards iterates, and Λ_n^- consists of all points that have n backwards iterates. Then define

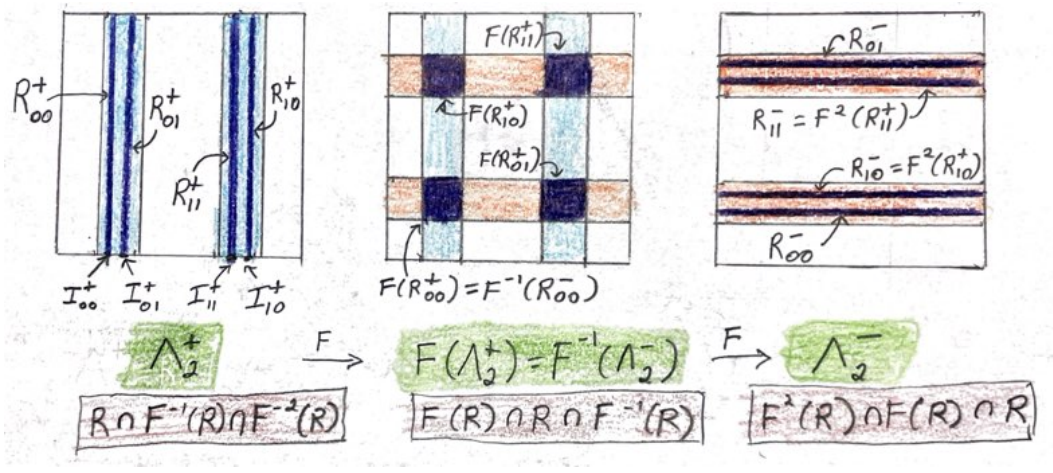
$$\Lambda^+ := \bigcap_{n\in\mathbb{N}} \Lambda_n^+, \qquad \Lambda^- := \bigcap_{n\in\mathbb{N}} \Lambda_n^-, \qquad \Lambda := \Lambda^+ \cap \Lambda^- , \tag{1.65}$$

so Λ^+ consists of all points with infinite forward orbits, Λ^- consists of all points with infinite backward orbits, and Λ consists of all points with bi-infinite orbits. To understand the structure of these sets, we will first examine the sets $\Lambda_2^\pm$, which are shown in Figure 1.42.

The middle picture in Figure 1.42 shows the set

$$F(\Lambda_2^+) = F^{-1}(\Lambda_2^-) = F(R) \cap R \cap F^{-1}(R) = \Lambda_1^- \cap \Lambda_1^+. \tag{1.66}$$

FIGURE 1.42. Two iterations of a linear horseshoe.



From the definition of a linear horseshoe, we have $\Lambda_1^\pm = R_0^\pm \cup R_1^\pm$, so

$$\begin{aligned} \Lambda_1^- \cap \Lambda_1^+ &= (R_0^- \cup R_1^-) \cap (R_0^+ \cup R_1^+) \\ &= (R_0^- \cap R_0^+) \cup (R_0^- \cap R_1^+) \cup (R_1^- \cap R_0^+) \cup (R_1^- \cap R_1^+). \end{aligned} \tag{1.67}$$

Each of these intersections represents one of the four darker squares in the middle picture: observe that for each $i, j \in \{0, 1\}$, we have

$$R_i^- \cap R_j^+ = ([0,1] \times I_i^-) \cap (I_j^+ \times [0,1]) = I_j^+ \times I_i^-. \tag{1.68}$$

REMARK 1.39. It may look innocuous, but (1.68) captures one of the core ideas of hyperbolic dynamics: *any future can be connected to any past.* Later, we will explore this idea more, as well as its nonlinear generalizations.

Given $(i, j) \in \{0, 1\}^2$, consider the sets

$$\begin{aligned} R_{ij}^+ &:= F^{-1}(R_i^- \cap R_j^+) = R_i^+ \cap F^{-1}(R_j^+), \\ R_{ij}^- &:= F(R_i^- \cap R_j^+) = F(R_i^-) \cap R_j^-, \end{aligned} \tag{1.69}$$

so that by (1.66) and (1.67), we have

$$\Lambda_2^+ = \bigcup_{(i,j)\in\{0,1\}^2} R_{ij}^+ \quad \text{and} \quad \Lambda_2^- = \bigcup_{(i,j)\in\{0,1\}^2} R_{ij}^-.$$

Let us say that a single step $x \mapsto F(x)$ of an orbit is "coded by $a \in \{0, 1\}$" if $x \in R_a^+$ and $F(x) \in R_a^-$. Then R_{ij}^+ consists of those points for which the first two forward steps are coded by the symbols i and j (in that order), and R_{ij}^- has a similar description going backwards. Moreover, given any $x \in R_{ij}^+$, we have $F(x) = (g_i x_1, h_i x_2)$, so (1.68) gives

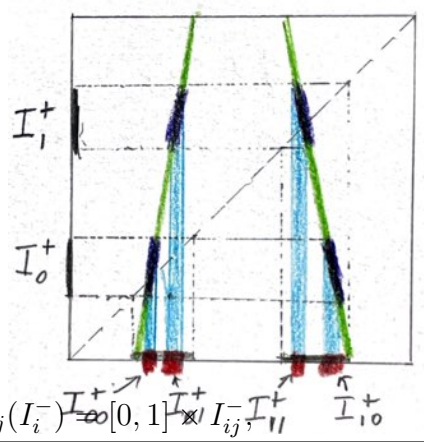
$$R_{ij}^+ = F^{-1}(I_j^+ \times I_i^-) = g_i^{-1}(I_j^+) \times h_i^{-1}(I_i^-) = I_{ij}^+ \times [0,1],$$

where we write

$$I_{ij}^+ := I_i^+ \cap g^{-1}I_j^+ = g_i^{-1}I_j^+ = g_i^{-1}g_j^{-1}([0,1]). \tag{1.70}$$

Figure 1.43 shows the four intervals I_{ij}^+. The corresponding sets R_{ij}^+ are the four narrow vertical strips in the first picture in Figure 1.42, whose union is Λ_2^+. Writing $C_2^+ := \bigcup_{(i,j)\in\{0,1\}^2} I_{ij}^+$, we see that $C_2^+ := g^{-2}([0,1])$, and the set Λ_2^+ is a direct product: $\Lambda_2^+ = C_2^+ \times [0,1]$.

FIGURE 1.43. More intervals.



Similarly, the sets R_{ij}^- making up Λ_2^- are the four narrow horizontal strips in the third picture of Figure 1.43. Each of these sets, and Λ_2^- itself, can again be described as a direct product:

$$R_{ij}^- = F(I_j^+ \times I_i^-) = g_j(I_j^+) \times h_j(I_i^-) = [0,1] \times I_{ij}^-,$$

where we write

$$I_{ij}^- := h_j(I_i^-) = h_j h_i([0,1]), \tag{1.71}$$

so that writing $C_2^- = \bigcup_{(i,j)\in\{0,1\}^2} I_{ij}^-$, we have $\Lambda_2^- = [0,1] \times C_2^-$.

With this discussion to guide us, we now give a general description of $\Lambda_n^\pm$. This will rely on coding points in $\Lambda_n^\pm$ in terms of their orbits. The symbols used for the coding will come from the set $A = \{0,1\}$, which we refer to as the *alphabet*. Given $n \in \mathbb{N}$ and $w \in A^n$, we will refer to w as a *word of length* n, and will write $w = w_1 w_2 \cdots w_n$; in particular, note that juxtaposition denotes concatenation rather than multiplication.

PROPOSITION 1.40. *Given $w \in A^n$, the sets*

$$R_w^+ := \bigcap_{k=0}^{n-1} F^{-k}(R_{w_k}^+) \quad \text{and} \quad R_w^- := \bigcap_{k=0}^{n-1} F^k(R_{w_k}^-) \tag{1.72}$$

have the property that $F^n(R_w^+) = R_w^-$, and can be written as direct products

$$R_w^+ = I_w^+ \times [0,1] \quad \text{and} \quad R_w^- = [0,1] \times I_w^-, \tag{1.73}$$

where $I_w^\pm$ are nonempty intervals given by

$$\begin{aligned} I_w^+ &:= g_{w_1}^{-1} \circ g_{w_2}^{-1} \circ \cdots \circ g_{w_n}^{-1}([0,1]), \\ I_w^- &:= h_{w_n} \circ h_{w_{n-1}} \circ \cdots \circ h_{w_1}([0,1]), \end{aligned} \tag{1.74}$$

each with length $\le \chi^n$, where $\chi \in (0,1)$ is the maximum of the lengths of the intervals $I_a^\pm$ with $a \in \{0,1\}$. Writing $C_n^\pm = \bigcup_{w\in A^n} I_w^\pm$, we have

$$\Lambda_n^+ = \bigcup_{w\in A^n} R_w^+ = C_n^+ \times [0,1] \quad \text{and} \quad \Lambda_n^- = \bigcup_{w\in A^n} R_w^- = [0,1] \times C_n^-. \tag{1.75}$$

Moreover, the sets $C^{\pm}_{\infty} := \bigcap_{n=1}^{\infty} C^{\pm}_n \subset [0,1]$ *are Cantor sets,*[11] *and*

$$\Lambda^+ = C^+_\infty \times [0,1], \qquad \Lambda^- = [0,1] \times C^-_\infty, \qquad \Lambda = C^+_\infty \times C^-_\infty. \tag{1.76}$$

PROOF. The fact that $F^n(R^+_w) = R^-_w$ is immediate from the definition (1.72). The main thing to prove is (1.73); once we prove this, the rest of the proposition will follow quickly from the definitions.

We prove (1.73) by induction, observing that the case $n = 1$ is in the definition of a linear horseshoe, and the discussion preceding Proposition 1.40 showed the first inductive step, going from $n = 1$ to $n = 2$. For the general inductive step, suppose (1.73) holds for all $w \in A^n$; then for any $a \in A$, the definitions in (1.72) give

$$\begin{aligned} R^+_{aw} &= R^+_a \cap F^{-1}(R^+_w) = F^{-1}(R^-_a \cap R^+_w) = F^{-1}\big(([0,1] \times I^-_a) \cap (I^+_w \times [0,1])\big) \\ &= F^{-1}(I^+_w \times I^-_a) = g_a^{-1}(I^+_w) \times h_a^{-1}(I^-_a) = I^+_{aw} \times [0,1], \end{aligned}$$

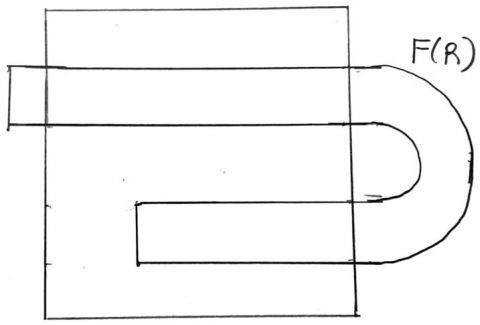
and similarly

$$\begin{aligned} R^-_{wa} &= F(R^-_w) \cap R^-_a = F(R^-_w \cap R^+_a) = F\big(([0,1] \times I^-_w) \cap (I^+_a \times [0,1])\big) \\ &= F(I^+_a \times I^-_w) = g_a(I^+_a) \times h_a(I^-_w) = [0,1] \times I^-_{wa}. \end{aligned}$$

This proves (1.73) for all n by induction. The descriptions in (1.75) and (1.76) follow from the definitions and from (1.73). The claim about the lengths of the intervals follows because each g_a^{-1} and h_a contracts distances by a factor of $|I^{\pm}_a| \le \chi$. Thus each of $C^{\pm}_n$ is the union of 2^n closed intervals with length $\le \chi^n$. Since $C^{\pm}_{n+1} \subset C^{\pm}_n$, with each basic interval of $C^{\pm}_n$ containing two from $C^{\pm}_{n+1}$, we conclude that $C^{\pm}_\infty$ are Cantor sets. □

►► EXERCISE 1.23. What parts of the preceding discussion would fail if the image $F(R)$ had the shape shown in Figure 1.44?

FIGURE 1.44. A partial branch.



Lec 14
Mon, Feb 17

1.8.2. Coding with bi-infinite sequences. The use of finite words $w \in A^n$ to describe the component rectangles of $\Lambda_n^{\pm}$ can be extended to infinite sequences, which describe individual points of Λ. To this end, consider the set

$$A^{\mathbb{Z}} := \{\bar{x} = (\bar{x}_n)_{n\in\mathbb{Z}} : \bar{x}_n \in A = \{0,1\} \text{ for all } n \in \mathbb{Z}\} \tag{1.77}$$

of bi-infinite sequences of symbols in A, equipped with the metric

$$d(\bar{x},\bar{y}) = 2^{-s(\bar{x},\bar{y})}, \qquad s(\bar{x},\bar{y}) = \min\{|n| : n \in \mathbb{Z}, \bar{x}_n \neq \bar{y}_n\}. \tag{1.78}$$

Given $\bar{x} \in A^{\mathbb{Z}}$ and $i, j \in \mathbb{Z}$ with $i \le j$, we will refer to $[i,j] \cap \mathbb{Z}$ as an interval of integers, and will write

$$\bar{x}_{[i,j]} = \bar{x}_i \bar{x}_{i+1} \cdots \bar{x}_{j-1} \bar{x}_j, \tag{1.79}$$

We will also sometimes use the notation

$$\bar{x}_{(i,j]} = \bar{x}_{[i+1,j]}, \qquad \bar{x}_{[i,j)} = \bar{x}_{[i,j-1]}, \qquad \bar{x}_{(i,j)} = \bar{x}_{[i+1,j-1]}. \tag{1.80}$$

Now given $\bar{x} \in A^{\mathbb{Z}}$ and $n \in \mathbb{N}$, let

$$R_n^+(\bar{x}) := R^+_{\bar{x}_{[0,n)}} \quad \text{and} \quad R_n^-(\bar{x}) := R^-_{\bar{x}_{[-n,0)}} = F^n(R^+_{\bar{x}_{[-n,0)}}); \tag{1.81}$$

that is, $R_n^+(\bar{x})$ consists of all points $z \in R$ such that the iterates $F^k(z)$ are defined for every $k \in [0,n)$, and have the property that $F^k(z) \in R^+_{\bar{x}_k}$, and $R_n^-(\bar{x})$ has a similar description replacing the interval $[0,n)$ with the interval $[-n,0)$. By Proposition 1.40, we have

$$R_n^+(\bar{x}) \cap R_n^-(\bar{x}) = I^+_{\bar{x}_{[0,n)}} \times I^-_{\bar{x}_{[-n,0)}},$$

so this set has diameter $\le 2\chi^n$. We conclude that $\bigcap_{n=1}^{\infty} R_n^+(\bar{x}) \cap R_n^-(\bar{x})$ is a single point, which we denote $\pi(\bar{x})$, obtaining a map $\pi\colon A^{\mathbb{Z}} \to \Lambda$.

▶ Exercise 1.24. Prove that the map π is a homeomorphism. Moreover, show that if $\sigma\colon A^{\mathbb{Z}} \to A^{\mathbb{Z}}$ denotes the shift map defined by $\sigma(\bar{x})_n = \bar{x}_{n+1}$, then $F \circ \pi = \pi \circ \sigma$, so that $F|_\Lambda$ and σ are topologically conjugate.

We conclude that the set Λ is uncountably infinite: in particular, R contains uncountably many points y for which the entire orbit $\{F^n(y) : n \in \mathbb{Z}\}$ remains in R. Moreover, despite the fact that we are dealing with a deterministic system, there is a certain past–future independence if we "blur our vision" and only consider approximate information about the current state of the system:

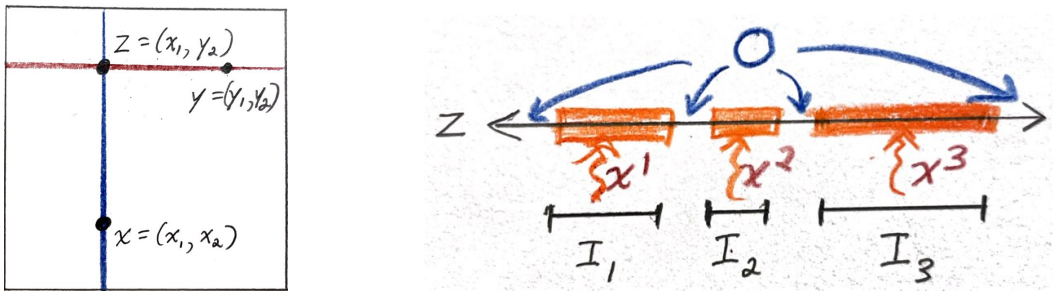
Proposition 1.41. *Let $F\colon R \to R$ be a linear horseshoe, and let $\Lambda = \bigcap_{n\in\mathbb{Z}} F^{-n}(R)$ be its maximal F-invariant set. Define a map $b\colon \Lambda \to \{0,1\}$ that records which branch of the horseshoe a point is in by writing $b(x) = a$ for all $x \in R_a^+$. Write $x \approx y$ if $b(x) = b(y)$.*

(1) Given any $x, y \in \Lambda$, there is a unique $z \in \Lambda$ such that $F^n(z) \approx F^n(x)$ for all $n \ge 0$, and $F^n(z) \approx F^n(y)$ for all $n < 0$.

[11] Topologically, this means that they are compact, nonempty, do not contain any intervals, and do not have any isolated points.

(2) *Given any points $x^1, \dots, x^k \in \Lambda$ and any pairwise disjoint intervals of integers $I_1, \dots, I_k \subset \mathbb{Z}$, there exists $z \in \Lambda$ such that $F^n(z) \approx F^n(x^j)$ for all $j \in \{1, \dots, k\}$ and all $n \in I_j$.*

FIGURE 1.45. Product structure, and specifying an orbit.



(a) (b)

PROOF. We prove the first statement twice: once geometrically, and once symbolically.

Here is the geometric proof: given $x, y \in \Lambda$, let $z = (x_1, y_2)$ be the point where the vertical line through x intersects the horizontal line through y, as shown in Figure 1.45(a). Recall from (1.73) in Proposition 1.40 that $R_n^+(x) = I^+_{x_{[0,n)}} \times [0,1]$ is a union of vertical lines for all $n \geq 0$, so $R_n^+(z) = R_n^+(x)$ for all such n, and in particular, $F^n(z) \approx F^n(y)$. For $n < 0$, observe similarly that $R^-_{|n|}(y) = [0,1] \times I_{y_{[n,0)}}$ is a union of horizontal lines, so $F^n(z) \approx F^n(y)$.

Here is the symbolic proof of the first statement: let $\bar{x} = \pi^{-1}(x)$ and $\bar{y} = \pi^{-1}(y)$ be the elements of $A^{\mathbb{Z}}$ that code the points x and y. Define $\bar{z} \in A^{\mathbb{Z}}$ by

$$\bar{z}_n = \begin{cases} y_n & \text{if } n < 0, \\ x_n & \text{if } n \geq 0. \end{cases}$$

Then $z = \pi(\bar{z})$ has the desired property, by definition of the coding map π.

This symbolic proof extends easily to cover the second statement: for each $j \in \{1, \dots, k\}$, let $\bar{x}^j = \pi^{-1}(x^j)$, and define $\bar{z} \in A^{\mathbb{Z}}$ by

$$\bar{z}_n = \begin{cases} x^j_n & \text{if } n \in I_j \text{ for some } j \in \{1, \dots, k\}, \\ 0 & \text{otherwise.} \end{cases}$$

Then $z = \pi(\bar{z}) \in \Lambda$ has the desired property. □

Informally, the two parts of Proposition 1.41 have the following descriptions.

(1) The first part says that any past can be joined to any future, so that the set of all orbits in the horseshoe has a *product structure*:

(all possible orbits) = (all possible futures) × (all possible pasts).

(2) The second part says that any sequence of orbit segments can be connected by a single orbit that "shadows" each of them in turn, so that on each interval the orbit of z is *specified* up to a small error by the orbits of the points x^j.

Both of these properties – product structure and specification – occupy a central place in hyperbolic dynamics, and we will return to them throughout the book.

1.8.3. Back to the nonlinear case. Now comes the payoff of the results developed in §1.7.3, §1.8.1, and §1.8.2: everything we said about linear horseshoes remains true (with appropriate modifications) for nonlinear horseshoes, and in particular, applies to every system with a transverse homoclinic intersection, such as the standard map. In this section, we will formulate these ideas more precisely, and will explain several senses in which this means that all such systems exhibit "chaotic behavior".

Let us start with some general definitions, motivated by the arguments we used in the Hadamard–Perron Theorem 1.15 and in 1.7.3.

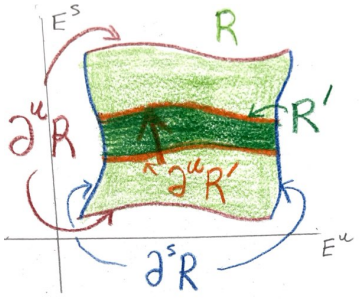
DEFINITION 1.42. Let E^u and E^s be transverse one-dimensional subspaces of $\mathbb{R}^2$, fix $\alpha > 0$, and let $K^{u,s}$ be as in (1.25). We call $K^{u,s}$ a *cone decomposition*.

In the following definitions, we fix a cone decomposition $K^{u,s}$. The next definition extends Definition 1.35.

DEFINITION 1.43. A continuous curve $\gamma \subset \mathbb{R}^2$ is *u-admissible* if $x - y \in K^u$ for every $x, y \in \gamma$. Similarly, γ is *s-admissible* if $x - y \in K^s$ for every $x, y \in \gamma$.

Observe that when γ is C^1, admissibility is equivalent to the condition that all tangent lines of γ lie in K^u or K^s. The next definition extends the ideas introduced in the paragraphs preceding Definition 1.35.

FIGURE 1.46. Rectangles.



DEFINITION 1.44. A compact set $R \subset \mathbb{R}^2$ is an *admissible rectangle* if it is the closure of its interior and if its boundary can be written as $\partial R = \partial^s R \cup \partial^u R$, where $\partial^s R$ is the disjoint union of two s-admissible curves, and $\partial^u R$ is the disjoint union of two u-admissible curves, as shown in Figure 1.46.

One of the main conclusions of Lemmas 1.36 and 1.37 was that the sets $R_n^{\pm}(\mathbf{0})$ and $R_n^{\pm}(p)$ are admissible rectangles. Those lemmas also established that

(1) the rectangles R_n^+ "cross R completely in the stable direction", and the rectangles R_n^- "cross R completely in the unstable direction";
(2) the maps carrying $R_n^+ \to R_n^-$ have derivatives treating the cones $K^{u,s}$ according to the hypotheses of Proposition 1.12.

The following definitions capture these ideas.

DEFINITION 1.45. Given an admissible rectangle R, say that a u-admissible curve $\gamma \subset \mathbb{R}^2$ *crosses* R if $\gamma \cap R$ is a connected curve with positive length whose endpoints lie on $\partial^s R$. Similarly, an s-admissible curve γ *crosses* R if $\gamma \cap R$ is a connected curve with positive length whose endpoints lie on $\partial^u R$.

The following exercise provides some further justification for the term "rectangle".

▶▶ EXERCISE 1.25. Let R be an admissible rectangle. Prove that if γ is a u-admissible curve that crosses R, and η is an s-admissible curve that crosses R, then $\gamma \cap \eta$ is a single point.

▶▶ EXERCISE 1.26. Let R be an admissible rectangle. Suppose that $\eta \subset R$ is a set with the following property: every u-admissible curve γ that crosses R intersects η in exactly one point. Prove that η is an s-admissible curve that crosses R.

Lec 15
Wed, Feb 19

LEMMA 1.46. *Let R and R' be admissible rectangles with $R' \subset R$. The following are equivalent (see Figure 1.46):*

(1) $\partial^s R' \subset \partial^s R$;
(2) every u-admissible curve that crosses R' also crosses R.

PROOF. ($\Rightarrow$): Every u-admissible curve γ crossing R' has $\gamma \cap R = \gamma \cap R'$ with endpoints in $\partial^s R' \subset \partial^s R$.
($\Leftarrow$): Every point $x \in \partial^s R'$ is the endpoint of a u-admissible curve crossing R', which thus crosses R, so $x \in \partial^s R$. □

DEFINITION 1.47. Let R and R' be admissible rectangles with $R' \subset R$. If one (and hence both) of the conditions in Lemma 1.46 holds, then we say that R' is a *u-subrectangle* of R. Similarly, R' is an s-subrectangle if $\partial^u R' \subset \partial^u R$, or (equivalently) every s-admissible curve crossing R' also crosses R.

Now we can reword the first parts of Lemmas 1.36 and 1.37 by saying that the sets $R_n^+(\mathbf{0})$ and $R_n^+(p)$ are s-subrectangles of R, and the sets $R_n^-(\mathbf{0})$ and $R_n^-(p)$ are u-subrectangles of R; see Figure 1.38.

The preceding definitions were entirely geometric. Now we bring some dynamics into the picture.

DEFINITION 1.48. Let $K^{u,s}$ be a cone decomposition in $\mathbb{R}^2$, and let $\operatorname{int} K^{u,s} \subset K^{u,s}$ be obtained by replacing $\leq$ with $<$ in (1.25). Let $L\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a linear map.

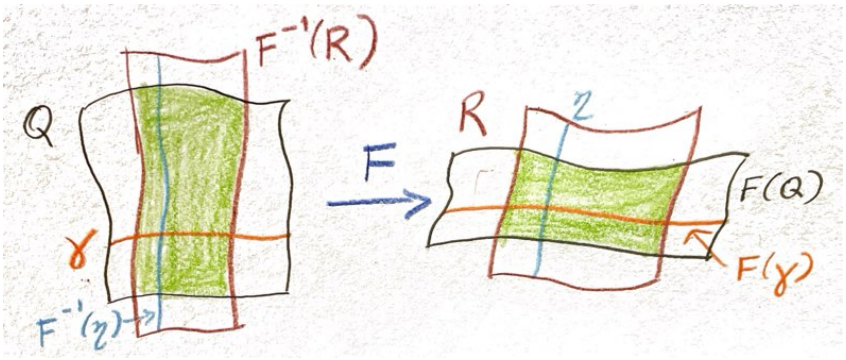
(1) If $L(K^u) \subset K^u$ and $L^{-1}K^s \subset K^s$, then we say that L *preserves* the cone decomposition $K^{u,s}$.
(2) If $L(K^u) \subset \operatorname{int} K^u$ and $L^{-1}K^s \subset \operatorname{int} K^s$, then we say that L *stably preserves* the cone decomposition $K^{u,s}$.
(3) If L stably preserves the cones $K^{u,s}$ and if in addition $\|Lv\| \geq \chi^{-1}\|v\|$ for all $v \in K^u$, and $\|Lv\| \leq \chi\|v\|$ for all $v \in K^s$, then we say that L is *χ-hyperbolic* with respect to the cone decomposition $K^{u,s}$.

REMARK 1.49. The first part of the proof of the Hadamard–Perron Theorem 1.15 amounts to fixing a neighborhood of $\mathbf{0}$ on which $Df(x)$ is χ-hyperbolic, so that Proposition 1.12 can be applied.

DEFINITION 1.50. Let $U, V \subset \mathbb{R}^2$ be open and let $F\colon U \to V$ be a C^1 diffeomorphism. We say that F is *χ-hyperbolic* with respect to a cone decomposition $K^{u,s}$ if $DF(x)$ is χ-hyperbolic for every $x \in U$.

REMARK 1.51. Let $F\colon U \to V$ be χ-hyperbolic. Following the argument in Lemma 1.17, we see immediately that if $\gamma \subset U$ is u-admissible, then $F(\gamma) \subset V$ is u-admissible; similarly, if $\gamma \subset V$ is s-admissible, then $F^{-1}(\gamma) \subset U$ is s-admissible. Moreover, if $x, y \in U$ satisfy $x - y \in K^u$, then the argument there gives $\|F(x) - F(y)\| \geq \chi^{-1}\|x - y\|$; similarly, if $x, y \in V$ satisfy $x - y \in K^s$, then $\|F^{-1}(x) - F^{-1}(y)\| \leq \chi\|x - y\|$.

FIGURE 1.47. Mapping one rectangle across another.



DEFINITION 1.52. Let $F\colon U \to V$ be χ-hyperbolic. Given admissible rectangles $Q \subset U$ and $R \subset V$, we say that F *maps Q across R*, and write $Q \overset{F}{\rightsquigarrow} R$, if the following are both true (see Figure 1.47):

(1) if $\gamma \subset U$ is a u-admissible curve crossing Q, then $F(\gamma)$ crosses R;
(2) if $\eta \subset V$ is a s-admissible curve crossing R, then $F^{-1}(\eta)$ crosses Q.

LEMMA 1.53. *Given admissible rectangles Q and R, a χ-hyperbolic map F maps Q across R if and only if $P := Q \cap F^{-1}(R)$ is an s-subrectangle of Q and $F(P) = F(Q) \cap R$ is a u-subrectangle of R.*

▶ EXERCISE 1.27. Prove Lemma 1.53.

REMARK 1.54. It follows immediately from Definitions 1.47 and 1.52 that if P is an s-subrectangle of Q and $P \overset{F}{\rightsquigarrow} R$, then $Q \overset{F}{\rightsquigarrow} R$. Similarly, if P is a u-subrectangle of R and $Q \overset{F}{\rightsquigarrow} P$, then $Q \overset{F}{\rightsquigarrow} R$.

REMARK 1.55. Our arguments regarding the graph transform in Proposition 1.19 can be interpreted as follows. If f is as in in the Hadamard–Perron Theorem 1.15, then f maps the following rectangle across itself:

$$R := \{v_u + v_s : v_u \in E^u, v_s \in E^s, \|v_u\| \le \delta, \|v_s\| \le \delta\}.$$

Now we can finally give a precise definition of a "nonlinear horseshoe".

DEFINITION 1.56. A *horseshoe* with respect to a cone decomposition $K^{u,s}$ consists of a finite set of disjoint admissible rectangles $R_1, \dots, R_d$, with $d \ge 2$, and a map F defined on a neighborhood of each R_i such that:

(1) F is χ-hyperbolic on each R_i, and
(2) given any $i, j \in \{1, \dots, d\}$, F maps R_i across R_j.

With these definitions in hand, Lemmas 1.36 and 1.37 together imply the following.

PROPOSITION 1.57. *Let f be as in the Hadamard–Perron Theorem 1.15, and suppose that W^s and W^u intersect transversely at some point $p \neq \mathbf{0}$. Then there are disjoint admissible rectangles $R_0, R_1 \subset B(\mathbf{0}, r)$ and an integer $\ell \in \mathbb{N}$ such that the map $F := f^\ell \colon R_0 \sqcup R_1 \to \mathbb{R}^2$ is a horseshoe.*

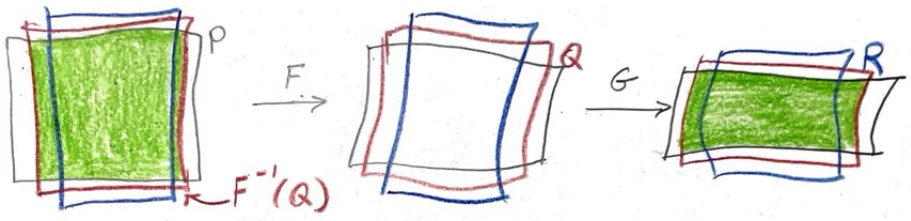
REMARK 1.58. As in §1.8.1, it is natural to consider the maximal invariant set $\Lambda := \bigcap_{n\in\mathbb{Z}} F^{-n}(R_0 \sqcup R_1)$, and it is often this set that is referred to as the "horseshoe".

Now we are in a position to extend the "symbolic coding" results of §§1.8.1–1.8.2 to nonlinear horseshoes.

PROPOSITION 1.59. *Let $F \colon \bigsqcup_{i=1}^d R_i \to \mathbb{R}^2$ be a horseshoe with respect to a cone decomposition $K^{u,s}$, and let Λ be as in Remark 1.58. Let $A = \{1, \dots, d\}$, and equip $A^{\mathbb{Z}}$ with the metric (1.78). Given $\bar{x} \in A^{\mathbb{Z}}$, there exists a unique $x \in \Lambda$ such that $F^n(x) \in R_{\bar{x}_n}$ for all $n \in \mathbb{Z}$. Moreover, the map $\bar{x} \mapsto x$ defines a topological conjugacy between the shift space $(A^{\mathbb{Z}}, \sigma)$ and the horseshoe (Λ, F).*

The key to the proof of Proposition 1.59 is an analogue of (1.68) for nonlinear rectangles (recall Remark 1.39 regarding the centrality of this fact), which relies on the *mapping across* property in Definition 1.52, and is illustrated in Figure 1.48.

FIGURE 1.48. The concatenation of two rectangle crossings.



LEMMA 1.60. *Let P, Q, R be admissible rectangles and F, G be χ-hyperbolic maps such that $P \overset{F}{\rightsquigarrow} Q$ and $Q \overset{G}{\rightsquigarrow} R$. Then $(P \cap F^{-1}Q) \overset{G\circ F}{\rightsquigarrow} R$.*

PROOF. If γ is a u-admissible curve crossing $P \cap F^{-1}Q$, then $F(\gamma)$ is a u-admissible curve crossing Q since $P \overset{F}{\rightsquigarrow} Q$, and thus $G \circ F(\gamma)$ is a u-admissible curve crossing R since $Q \overset{G}{\rightsquigarrow} R$. Similarly, if γ is an s-admissible curve crossing R, we conclude that $G^{-1}(\gamma)$ is an s-admissible curve crossing Q, and thus $F^{-1} \circ G^{-1}(\gamma)$ crosses $P \cap F^{-1}Q$. □

PROOF OF PROPOSITION 1.59. Given $w \in A^n$, let $R_w := \bigcap_{k=0}^{n-1} F^{-k}(R_{w_{k+1}})$ be the set of points $x \in R_{w_1}$ with the property that $F(x) \in R_{w_2}$, $F^2(x) \in R_{w_3}$, and so on. We claim that for every $n \in \mathbb{N}$, every $w \in A^n$, and every $a \in A$, we have:

(1) R_w is an s-subrectangle of R_{w_1};
(2) $R_w \overset{F^n}{\rightsquigarrow} R_a$.

By Definition 1.56, we have $R_i \overset{F}{\rightsquigarrow} R_j$ for all $i, j \in A$, which proves the claim when $n = 1$. If the claim is true for some n, then for every $w \in A^n$ and $a, b \in A$, we see that R_{wa} is an s-subrectangle of R_w since $R_w \overset{F^n}{\rightsquigarrow} R_a$, and moreover, by Lemma 1.60 we have

$$R_w \overset{F^n}{\rightsquigarrow} R_a \text{ and } R_a \overset{F}{\rightsquigarrow} R_b \quad \Rightarrow \quad R_{wa} \overset{F^{n+1}}{\rightsquigarrow} R_b.$$

This proves the claim for $n+1$, so by induction the claim is true for all n. Given $\bar{x} \in A^{\mathbb{Z}}$ and $n \in \mathbb{N}$, consider the following analogues of (1.81):

$$R_n^+(\bar{x}) := R_{\bar{x}_{[0,n)}} \quad \text{and} \quad R_n^-(\bar{x}) := F^n(R_{\bar{x}_{[-n,0)}}).$$

Let $C := \max_{a\in A} \operatorname{diam}(R_a)$. By the χ-hyperbolicity property, every u-admissible curve contained in $R_n^+(\bar{x})$ has length $\le C\chi^n$, and similarly for every s-admissible curve contained in $R_n^-(\bar{x})$. With this in mind, consider the sets

$$V^s(\bar{x}) := \bigcap_{n=0}^{\infty} R_n^+(\bar{x}) \quad \text{and} \quad V^u(\bar{x}) := \bigcap_{n=0}^{\infty} R_n^-(\bar{x}).$$

The set $V^s(\bar{x})$ is the set of all points that have infinite forward orbits coded by the positive half of $\bar{x}$. Given any u-admissible curve γ that crosses $R_{\bar{x}_0}$, observe that

$$\gamma \cap V^s(\bar{x}) = \bigcap_{n\in\mathbb{N}} \big(\gamma \cap R_n^+(\bar{x})\big),$$

where each u-admissible curve $\gamma \cap R_n^+(\bar{x})$ has length $\le C\chi^n$ by the discussion above, so $\gamma \cap V^s(\bar{x})$ is a single point. By Exercise 1.26, it follows that $V^s(\bar{x})$ is an s-admissible curve that crosses $R_{\bar{x}_0}$.

A similar argument shows that $V^u(\bar{x})$ is a u-admissible curve that crosses $F(R_{\bar{x}_{-1}})$, and thus it crosses $R_{\bar{x}_0}$ (since $R_{\bar{x}_{-1}} \overset{F}{\rightsquigarrow} R_{\bar{x}_0}$). By Exercise 1.25, the intersection $V^s(\bar{x}) \cap V^u(\bar{x})$ is a single point, which we denote x. Now the proof of Proposition 1.59 is completed by the following exercise. □

▶▶ EXERCISE 1.28. Prove that the map $\pi\colon \bar{x} \mapsto x$ defines a topological conjugacy from $(A^{\mathbb{Z}}, \sigma)$ to (Λ, F).

Lec 16
Fri, Feb 21

Now we can finally provide a more definitive explanation for the "cloud-like" appearance of the orbit labeled C in Figure 1.6: the standard map has a transverse homoclinic intersection, which leads to a horseshoe Λ by Proposition 1.57. By Proposition 1.59, Λ contains uncountably many distinct orbits, and since the proof of Proposition 1.41 only relied on the topological conjugacy to the full shift $(A^{\mathbb{Z}}, \sigma)$, the "product structure" and "specification" properties in that proposition continue to hold for $f^\ell\colon \Lambda \to \Lambda$ here.

In later chapters, we will explore further properties of horseshoes and their generalizations: locally maximal hyperbolic sets. We conclude this section with the observation that every orbit in a horseshoe experiences exponential separation of nearby trajectories: if $x, y \in \Lambda$ have the property that $x - y \in K^u$ and that $\|f^k(x) - f^k(y)\| \le r$ for all $0 \le k \le n$, then we have

$$\|x - y\| \le C\chi^{n/\ell}\|f^n(x) - f^n(y)\|,$$

where $\ell \in \mathbb{N}$ is as in Proposition 1.57. This leads to the same conclusion as in the discussion preceding Remark 1.9: if we want our forecast to be accurate to within some tolerance $\epsilon \in (0, r]$, and our initial measurement is accurate to within δ, then our forecast is only guaranteed to be accurate to within ϵ as long as n is small enough that

$$\delta \le C\chi^{n/\ell}\epsilon \Leftrightarrow n \le \ell\Big(\frac{\log\delta - \log(C\epsilon)}{|\log\chi|}\Big).$$

Since the forecast error increases exponentially fast, doubling the measurement accuracy only adds a fixed amount of time to the duration for which the forecast is valid to within a specified error. We have now shown that this is true for uncountably many orbits, whereas the original discussion is really only applicable when one of the orbits is the hyperbolic fixed point. This property of exponential growth of displacements will return later when we discuss *Lyapunov exponents*.

1.9. Poincaré and the three-body problem

The exposition so far has centered around the Chirikov–Taylor standard map, in order to demonstrate how an apparently simple setting can produce complicated dynamical phenomena. In the next chapter, we will begin exploring *classes* of dynamical systems more systematically, rather than focusing on specific examples. The remainder of this chapter describes two more concrete examples that have played an important historical role: the three-body problem, and the Lorenz attractor.[12]

REMARK 1.61. The present section is primarily historical, and the three-body problem described here will not play a role in the remainder of the book, so this section could be skipped by the reader who only wishes to see all the mathematical

[12]Of course many other examples have played important historical roles as well, and we will mention some of these in due course.

concepts as quickly as possible.[13] The remaining sections in this chapter will introduce important concepts from ergodic theory that will be developed in the rest of the book.

Celestial mechanics has played an important role in the historical development of dynamical systems, and indeed of mathematics more broadly. In 1609, Johannes Kepler formulated three laws describing the motion of the planets around the sun; these were based on empirical observations. In 1689, Isaac Newton's *Principia* gave Kepler's laws a theoretical foundation, via laws of motion and of universal gravitation. Kepler's first law states that each planet orbits the sun in an ellipse, with the sun at one focus. Newton considered the *two-body problem*, in which we consider only the sun and a single planet, disregarding the influence of all other planets, moons, etc., and proved that the two bodies travel in ellipses with a focus at their center of mass. Since the sun is several orders of magnitude more massive than the planets, the center of mass of the sun-planet system lies so close to the center of the sun that Kepler's law is a reasonable description.

Things become more complicated when we include the effect of multiple planets, or of other objects in the solar system. The first step is the *three-body problem*; consider, for example, the sun, Jupiter, and Earth. Henri Poincaré's work on this problem in the late 19th century can be reasonably considered the beginning of the modern theory of hyperbolic dynamical systems; in particular, he was the first to observe the existence of a homoclinic tangle, as we will now see.

Remark 1.62. This section only gives a brief survey of the history and of the mathematical ideas in Poincaré's work that are most relevant for our present discussion. For a more detailed description of both the history and the mathematics, see June Barrow-Green's 1997 book "Poincaré and the three-body problem". Other helpful accounts include "Poincaré's Discovery of Homoclinic Points" by K.G. Andersson (*Arch. Hist. Exact Sci.* **48**, 1994, no. 2, pp. 133–147), and "Poincaré, Celestial Mechanics, Dynamical Systems Theory and 'Chaos'" by Philip Holmes (*Physics Reports* **193**, No. 3, 1990, pp. 137–163).[14]

By the 1880s, several solutions of the three-body problem (or versions of it) had been given in terms of infinite series, but it was not known whether or not these series actually converged, and natural questions such as the stability of the solar system remained open. In the mid-1880s, the mathematician Gösta Mittag-Leffler, founder of the journal *Acta Mathematica*, organized a competition sponsored by King Oscar II of Norway and Sweden. The announcement of the competition included four questions, one of which concerned the stability of the solar system and a solution of

[13] But where is the fun in that?

[14] For broader accounts of the historical development of the theory of celestial mechanics, including not just Poincaré's work but also more recent developments, see the books "The KAM Story" by H. Scott Dumas (World Scientific, 2014), and "Celestial Encounters: The Origins of Chaos and Stability" by Diacu and Holmes (Princeton, 1996).

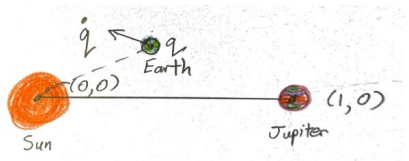
the n-body problem via convergent series. The prize was to be awarded in 1889, on the king's sixtieth birthday.

Poincaré would eventually win the prize for an entry studying the *restricted circular planar three-body problem* (which we will abbreviate RCP3BP); this problem makes the following assumptions.

- *Restricted:* The first body (the sun) has mass much larger than the second body (Jupiter), and the third body (Earth) has negligible mass compared to the other two. In particular, we neglect the gravitational pull of the third body on the other two, so they reduce to a two-body problem.
- *Circular:* The first two bodies orbit their center of mass in perfect circles.
- *Planar:* All three bodies move in a common plane $\mathbb{R}^2$.

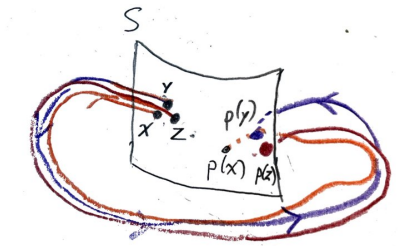
Use a rotating reference system with the center of mass at the origin and Jupiter (the second body) at a fixed position, say $(1, 0)$, as shown in Figure 1.49. Let $q \in \mathbb{R}^2$ denote the position of the Earth in this reference system; then the current state and future evolution of the system is determined by the pair $(q, \dot{q}) \in \mathbb{R}^4$. By conservation of energy,[15] the system remains on a three-dimensional manifold $X \subset \mathbb{R}^4$, which is our phase space.

FIGURE 1.49. A rotating frame.



Poincaré introduced a technique for studying this three-dimensional flow $\{f_t\}_{t\in\mathbb{R}}$ via a two-dimensional discrete-time system; taking $S \subset X$ to be a two-dimensional surface transverse to the flow direction, consider the *first return time function* $\tau\colon S \to (0, \infty]$ defined by $\tau(x) = \inf\{t > 0 : f_t(x) \in S\}$, and the *first return map* $P(x) = f_{\tau(x)}(x)$.

FIGURE 1.50. First return map.

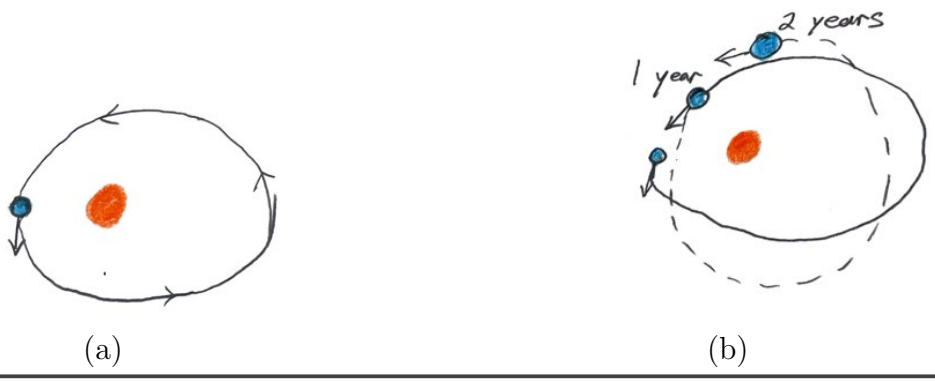


One possible choice of surface S can be described by first considering the limiting case in which the mass of Jupiter is so small (relative to the mass of the sun) that we disregard its gravitational attraction on the Earth, which therefore orbits the sun in an ellipse. Let S be the surface consisting of all points $x = (q, \dot{q}) \in X$ such that this ellipse is non-circular and for which x represents the point at which the Earth is closest to the sun (the *perihelion*).

If we worked in a nonrotating reference frame, then the Earth's elliptical orbit would appear closed, as in Figure 1.51(a). In our rotating reference frame, however, it has the appearance shown in Figure 1.51(b); after one (Earth) year, Jupiter has moved partway around the sun, so the Earth is in a different position relative

[15] Since we work in the limit when the third body is massless, one must actually consider "energy per unit mass" to obtain a quantity that remains useful in the limit.

FIGURE 1.51. An elliptical orbit in nonrotating and rotating frames.



to Jupiter. If we represent points on S with coordinates (θ, T), where $\theta \in S^1$ is the angular coordinate of the Earth at perihelion (with respect to the Sun–Jupiter axis), and $T \in (0, \infty)$ is the Earth's orbital period relative to Jupiter's, then the Poincaré return map becomes the twist map $P(\theta, T) = (\theta + T, T)$. We have of course encountered this twist map already, as the time-1 map of the unkicked rotator in (1.2).

One may now consider the RCP3BP as a perturbation of this twist map by reintroducing the (small) gravitational effect that Jupiter has on the earth: writing $\epsilon > 0$ for the ratio of the masses of Jupiter and the sun, the unperturbed case above corresponds to the limit as $\epsilon \to 0$, we we want to understand perturbations corresponding to small values of $\epsilon > 0$. Recall that we have already studied two such perturbations: the time-1 map of the pendulum, and the Chirikov–Taylor standard map. As we saw, these two systems have very different behaviors: both have a homoclinic orbit associated to a hyperbolic fixed point, but for the pendulum this orbit is part of a homoclinic connection, while for the standard map it is a transverse intersection that leads to a homoclinic tangle and a horseshoe.

Poincaré's analysis led to the following conclusions.[16]

- For a suitable truncation of the RCP3BP, in which certain higher-order terms (with respect to ϵ) are omitted, we obtain a picture analogous to the pendulum: the corresponding perturbation of the twist map has a hyperbolic fixed point with a homoclinic connection.
- Once these higher-order terms are restored, the homoclinic connection splits and becomes a transverse homoclinic intersection, as we observed for the standard map. In particular, the RCP3BP has a homoclinic tangle and a horseshoe.

The realization that this complicated dynamical behavior was present in the three-body problem represented a profound shift away from the idea that with better

[16] As we describe below, these conclusions were not immediately realized, and Poincaré's original work contained a significant error.

mathematics would always come better predictions, a vision that had been articulated clearly by Pierre-Simon Laplace, who wrote in 1814, in a work entitled "A Philosophical Essay on Probabilities":[17]

> We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situations of the beings who compose it – an intelligence sufficiently vast to submit these data to analysis – it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes.

These words of Laplace are often quoted and set in contrast to the unpredictability revealed by Poincaré's work. Recall our own initial motivations in §1.1 of understanding how randomness arises in deterministic systems. With this in mind, and given that Laplace's essay concerned the foundations of probability theory, it is worth highlighting two further passages written there, the first of which immediately follows the preceding one:

> The human mind offers, in the perfection which it has been able to give to astronomy, a feeble idea of this intelligence ... All these efforts in the search for truth tend to lead it back continually to the vast intelligence which we have just mentioned, but from which it will always remain infinitely removed.

Two pages later, having described how developments in celestial mechanics transformed comets from a supernatural phenomenon into a predictable occurrence, Laplace goes on to say:

> The regularity which astronomy shows us in the movements of comets doubtless exists also in all phenomena.
>
> The curve described by a simple molecule of air or vapor is regulated in a manner just as certain as the planetary orbits; the only difference between them is that which comes from our ignorance.
>
> Probability is relative, in part to this ignorance, in part to our knowledge.

Laplace connects randomness to our ignorance, to the limitations of our knowledge. One might read these lines and hope for a future in which the "perfection" and "regularity" of our astronomy have been extended so far that we can even predict the motion of a single molecule of a gas. Poincaré's work shows us that in fact, the unpredictability of the gas molecule's motion appears also in the motions of the solar system, and that our ignorance – at least as far as making forecasts is concerned – is greater than we might have thought.

[17] The quote here is taken from the 1902 translation by F.W. Truscott and F.W. Emory, where it appears on page 4.

REMARK 1.63. We will later see that the mathematics of statistical mechanics used to study large collections of gas molecules can in fact be very productively applied to study hyperbolic dynamical systems and to gain insights into the nature of the randomness they produce. One aspect of this theory is the close relationship between "entropy" and information, which echoes Laplace's remarks about probability and ignorance.

The development of the ideas above – a transverse homoclinic intersection, a homoclinic tangle, and a horseshoe – was not so straightforward as the description so far might suggest.[18] In fact, Poincaré's prize-winning submission mistakenly asserted that the "pendulum picture" of a homoclinic connection not only describes the truncated RCP3BP, but remains true once the higher-order terms are restored, leading to a stability result for the true RCP3BP.

In January 1889, the prize was awarded to Poincaré, and publication of his winning memoir in *Acta Mathematica* was planned for October 1889. In preparation for this, Edvard Phragmén, an editor of *Acta*, asked a number of questions seeking clarification of certain points in Poincaré's manuscript. Responding to these questions led Poincaré to write lengthy appendices to his memoir. In the course of this work, Poincaré eventually realized that one part of his argument contained a substantial error: in the process of correcting it, he proved that the restoration of the higher-order terms does *not* preserve the homoclinic connection, but instead leads to a transverse homoclinic intersection, and in fact to infinitely many of them, due to the invariance that we observed in §1.7.1. Later, Poincaré would write:[19]

> When one tries to depict the figure formed by these two curves[20] and their infinity of intersections, each corresponding to a doubly asymptotic solution,[21] these intersections form a sort of trellis, web, or infinitely tight mesh; neither of the two curves can ever intersect itself, but must fold back on itself in a very complex way in order to intersect all the links of the mesh infinitely many times.
>
> One is struck by the complexity of this figure that I shall not even attempt to draw.[22] Nothing is better suited to give us an idea of the complexity of the three body problem and all of the problems of dynamics in general where there is no uniform integral and Bohlin's series diverge.

[18] The next few paragraphs follow the account in Barrow-Green's book, pages 66–67.

[19] The quote here is reproduced from page 50 of "The KAM Story" by Dumas; the original quote is taken from Volume 3 of Poincaré's "Les Méthodes nouvelles de la mécanique céleste", 1899: see page 389 of the 1993 English translation (D. Goroff, American Institute of Physics).

[20] "These two curves" refers to the global stable and unstable manifolds W^s and W^u.

[21] A "doubly asymptotic solution" refers to a homoclinic point.

[22] See Figure 1.30. The reader can judge whether we have been wise in treading where Poincaré feared to draw.

By the time of Poincaré's realization in late 1889, pre-publication copies of the memoir had already been sent to a number of mathematicians. Mittag-Leffler retrieved these and arranged to have them destroyed, with Poincaré agreeing to pay the associated printing costs, an amount which exceeded the prize money. The corrected version was published in *Acta* in 1890. It was not until 1985 that some surviving copies of the original version were discovered, revealing more details of the development of Poincaré's ideas.[23]

In §§1.7–1.8, we described various consequences of a transverse homoclinic intersection, beyond the tangle itself. These were initially studied not by Poincaré, but by later mathematicians, most notably George David Birkhoff[24] and Stephen Smale.

Birkhoff, in his 1927 book "Dynamical Systems", observed that the presence of a transverse homoclinic intersection implied the coexistence of countably many periodic orbits. Chapter 4 of his 1935 book "Nouvelle recherches sur les systèmes dynamiques" studied further consequences.[25]

Later, in the 1960s, Stephen Smale introduced the linear horseshoe shown in Figure 1.39 and proved that a transverse homoclinic intersection implies the existence of a nonlinear horseshoe and thus of a compact invariant set on which the dynamics is topologically conjugate to the full shift map, as described in Proposition 1.59.[26] As described on page 59 of Diacu and Holmes, Smale's motivation was not the three-body problem but the *van der Pol equation* that modeled an oscillatory electrical circuit with periodic forcing and had been studied by Mary Cartwright and John Littlewood as part of their work on radar during World War II. They proved the existence of infinitely many periodic solutions (a la Birkhoff); further work on the system was done by Levinson,[27] who suggested that Smale consider this system as a possible counterexample to a conjecture that he (Smale) had formulated concerning finiteness of the number of periodic points in a bounded region for a "typical" system.

With this history in mind, it is reasonable to refer to the combination of Propositions 1.57 and 1.59 – which together state that a transverse homoclinic intersection leads to a nonlinear horseshoe on which (an iterate of) the dynamics is topologically conjugate to a full shift – as the *Poincaré–Birkhoff–Smale Theorem*.

1.10. Stochastic processes, invariant measures, and recurrence

Lec 17
Mon, Feb 24

The results of the preceding sections reveal an inherent unpredictability present in at least some deterministic systems. Confronted by this phenomenon, we must

[23]See pages 48–49 of the book by Diacu and Holmes for an account of Richard McGehee's discovery of the original, supposedly destroyed, version.

[24]Not to be confused with his son Garrett Birkhoff, whose work on convex cones we will soon encounter in §1.13.

[25]See page 74 of Diacu and Holmes.

[26]See "Diffeomorphisms with many periodic points", *Differential and Combinatorial Topology (A Sympsoium in Honor of Marston Morse)*, 1965, pages 63–80, MR0182020.

[27]"A second order differential equation with singular solutions", *Ann. of Math.* **50** (1949), pages 127–153, MR0030079.

adapt our mindset to a setting in which we cannot make definitive predictions ("this specific trajectory will do the following"). Instead, we must incorporate a certain amount of randomness into our forecasts ("the following will happen with high probability"). Making this precise leads us into *ergodic theory*.

Consider a dynamical system $f\colon X \to X$. We can interpret a function $\varphi\colon X \to \mathbb{R}$ as a measurement, or observation, of the system's current state; if the system is in state $x \in X$, then the measurement takes value $\varphi(x) \in \mathbb{R}$. A measurement that is made after one iteration of the system is represented by $\varphi \circ f$; if the system is in state $x \in X$ now, then tomorrow it is in state $f(x) \in X$, and the measurement takes value $\varphi(f(x))$. Similarly, a measurement made after n iterations is represented by the function $\varphi \circ f^n$.

With this notation, making a definitive prediction would consist of gathering information about the current state x via some initial measurements, and then using this information to make a statement about the value of $\varphi(f^n(x))$ for some $n > 0$. If the initial measurement tells us that $x \in U$, where $U \subset X$ is some small open set, then our prediction would need to be a statement about $\varphi(y)$ that is true for every $y \in f^n(U)$. When $f^n(U)$ spreads out and has large diameter, as illustrated in Figure 1.12, this may quickly become impossible to do in any meaningful way; even if φ is very regular (Lipschitz, smooth, etc.), the function $\varphi \circ f^n$ becomes very irregular and highly sensitive to small changes in x.

Even when $f^n(U)$ spreads out, we can still make probabilistic statements by equipping the phase space X with a probability measure μ, and regarding each $\varphi \circ f^n$ as a random variable with respect to μ. Then the sequence $(\varphi \circ f^n)_n$ becomes a stochastic process, whose behavior we can study with the language of probability theory. Now we must keep the following two issues in mind.

(1) Given a compact metric space X, there are many different Borel probability measures on X. Which one should we choose?
(2) The most complete results in probability theory are available for stochastic processes that are *independent* and *identically distributed* (see below). Do either of these properties hold for $(X, \mu, (\varphi \circ f^n)_n)$? If not, how do we proceed?

Regarding the first question, it is often the case that X carries extra structure that can inform our choice of μ; for example, in §1.2, the standard map was defined on $\mathbb{R}^2$, where it is natural to consider Lebesgue measure. (Although Lebesgue measure on $\mathbb{R}^2$ is infinite, we can obtain a probability measure by considering the system on the torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$.) In §1.9, the first return map is defined for a certain surface $S \subset \mathbb{R}^4$, and it is natural to consider (normalized) area on S.

The question of how to select an appropriate measure will play a major role in our story later on, beginning in Chapter **??**; it lies at the heart of the theory of *thermodynamic formalism*. For now we consider only the most elementary examples, such as Lebesgue measure.

For the second question, concerning properties of $(X, \mu, (\varphi \circ f^n)_n)$, we need to recall some definitions from measure theory and probability theory.

1.10.1. Probability spaces.

DEFINITION 1.64. A *measurable space* is a pair $(X, \mathcal{A})$, where X is a set and $\mathcal{A}$ is a σ-algebra of subsets of X.[28]

EXAMPLE 1.65. If X is a metric space, then we will usually consider the Borel σ-algebra $\mathcal{B}$, so that $(X, \mathcal{B})$ is a measurable space. This includes $\mathbb{R}^d$ with the Euclidean metric, as well as $S^{\mathbb{Z}}$ and $S^{\mathbb{N}}$, where S is a finite set and the metric is given by (1.78).

DEFINITION 1.66. A measure μ on a measurable space $(X, \mathcal{A})$ is a *probability measure* if $\mu(X) = 1$. We refer to the triple $(X, \mathcal{A}, \mu)$ as a *probability space*. Given a measurable space $(X, \mathcal{A})$, we will denote the set of all probability measures on $(X, \mathcal{A})$ by $\mathcal{M} = \mathcal{M}(X, \mathcal{A})$.

We often want to describe a measure without needing to explicitly identify its value on every set in the σ-algebra, and instead working with a smaller collection $\mathcal{C} \subset \mathcal{A}$.

DEFINITION 1.67. A *π-system* on X is a non-empty collection $\mathcal{C}$ of subsets of X that is closed under pairwise intersection: for every $A, B \in \mathcal{C}$, we have $A \cap B \in \mathcal{C}$. We say that $\mathcal{C} \subset \mathcal{A}$ *generates* $\mathcal{A}$ if $\mathcal{A}$ is the smallest σ-algebra containing $\mathcal{C}$.

EXAMPLE 1.68. When $X = \mathbb{R}$, the collection $\mathcal{C} = \{(-\infty, t] : t \in \mathbb{R}\}$ is a π-system that generates the Borel σ-algebra.

LEMMA 1.69. *If two probability measures $\mu, \nu \in \mathcal{M}(X, \mathcal{A})$ agree on a π-system $\mathcal{C} \subset \mathcal{A}$ that generates $\mathcal{A}$, then $\mu = \nu$.*

PROOF. See Lemma 1.6 in "Probability with Martingales" by David Williams, Cambridge, 1991. □

If we want to *construct* a measure, and not just check equality of two measures that have already been constructed, then we need a little more.

DEFINITION 1.70. A *semialgebra*[29] on X is a π-system $\mathcal{C}$ on X that contains $\emptyset$ and has the additional property that for every $A \in \mathcal{C}$, the set $X \setminus A$ is a finite disjoint union of elements of $\mathcal{C}$.

EXAMPLE 1.71. The π-system in Example 1.68 is not a semialgebra. On the real line, one possible choice of semialgebra that generates the Borel σ-algebra is $\mathcal{C}_\ell = \{(a, b] : -\infty \le a \le b \le \infty\}$.

►► EXERCISE 1.29. Fix a finite set S (the *symbols*). On $S^{\mathbb{N}}$, consider for every $n \in \mathbb{N}$ and every *word* $w \in S^n$ the *cylinder set*

$$[w] := \{x \in S^{\mathbb{N}} : x_i = w_i \text{ for all } 1 \le i \le n\}, \tag{1.82}$$

[28] Recall that this means that $\emptyset \in \mathcal{A}$, $X \in \mathcal{A}$, and $\mathcal{A}$ is closed under complements, countable intersections, and countable unions.

[29] Here we follow the terminology in Walters, "An Introduction to Ergodic Theory", p. 3, and in Royden, "Real Analysis", §12.2. The term *elementary family* is used in Folland, "Real Analysis", §1.2.

consisting of all infinite sequences that start with the word w. Let $S^* := \bigcup_{n=1}^{\infty} S^n$ be the set of all finite words. Then

$$\mathcal{C} = \{\emptyset\} \cup \{[w] : w \in S^*\}$$

is a semialgebra on $S^{\mathbb{N}}$ that generates the Borel σ-algebra.

DEFINITION 1.72. Given a semialgebra $\mathcal{C}$ on X, a set function $\ell\colon \mathcal{C} \to [0,\infty)$ is *countably additive* if it has the following property: whenever $A_1, A_2, \dots \in \mathcal{C}$ is a countable collection[30] of pairwise disjoint sets in $\mathcal{C}$ whose union $A := \bigcup_{n=1}^{\infty} A_n$ is also in $\mathcal{C}$, we have $\ell(A) = \sum_{n=1}^{\infty} \ell(A_n)$.

THEOREM 1.73 (Carathéodory Extension Theorem). *If $\mathcal{C}$ is a semialgebra on X that generates the σ-algebra $\mathcal{A}$, and $\ell\colon \mathcal{C} \to [0,\infty)$ is countably additive and finite-valued, then there exists a unique finite measure μ on $(X,\mathcal{A})$ such that $\mu|_{\mathcal{C}} = \ell$.*

EXAMPLE 1.74. Applying Theorem 1.73 to the set function $\ell((a,b]) = b-a$ leads to Lebesgue measure on $[0,1]$. (And more generally, on $\mathbb{R}$, although one needs to be a bit more careful because this measure is infinite.)

EXAMPLE 1.75. Let $S = \{1,\dots,d\}$ and let $\mathcal{C}$ be as in Exercise 1.29. Fix a probability vector $p \in \mathbb{R}^d$ (that is, $p_i \geq 0$ and $\sum_{i=1}^d p_i = 1$), and define a set function $\ell\colon \mathcal{C} \to [0,1]$ by

$$\ell(w_1 \cdots w_n) = p_{w_1} p_{w_2} \cdots p_{w_n}. \tag{1.83}$$

One can prove that this set function is countably additive, so by Theorem 1.73 it extends to a unique measure $\mu = \mu_p$ on $S^{\mathbb{N}}$, which is called a *Bernoulli measure* and models a sequence of independent experiments that each have finitely many outcomes, with probabilities given by p.

►► EXERCISE 1.30. Complete Example 1.75 by showing that the set function ℓ in (1.83) is indeed countably additive on the semialgebra $\mathcal{C}$.

1.10.2. Measurable maps and functions. Treating a measurable space $(X,\mathcal{A})$ as the phase space of a dynamical system, we will be interested in functions with domain X from three different points of view.

(1) A transformation $f\colon X \to X$ can represent the dynamics, evolving the system through one unit of time.
(2) A function $\varphi\colon X \to \mathbb{R}$ can represent a measurement or observation made of the system's current state; this is what we hope to treat as a random variable.
(3) A map $h\colon X \to Y$, where $(Y,\mathcal{B})$ is a measurable space (which may or may not be the same as $(X,\mathcal{A})$), can represent a "change of coordinates" such as the topological conjugacies we encountered earlier.

We will give several definitions in the setting of a map between two measurable spaces $(X,\mathcal{A})$ and $(Y,\mathcal{B})$, with the understanding that this covers all three points of view above.

[30] This includes the case of finite collections by taking $A_n = \emptyset$ for all n sufficiently large.

DEFINITION 1.76. Given two measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$, a map $f\colon X \to Y$ is *measurable* if for every $B \in \mathcal{B}$, we have $f^{-1}(B) \in \mathcal{A}$. We will refer to a measurable map f as a *measurable transformation*. In the specific case when $Y = \mathbb{R}$ and $\mathcal{B}$ is the Borel σ-algebra, we refer to a measurable map $\varphi\colon X \to \mathbb{R}$ as a *random variable*.

LEMMA 1.77. *If $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ are measurable spaces, $\mathcal{C} \subset \mathcal{B}$ generates the σ-algebra $\mathcal{B}$, and $f\colon X \to Y$ has the property that for every $C \in \mathcal{C}$, we have $f^{-1}(C) \in \mathcal{A}$, then f is measurable.*

PROOF. Let $\mathcal{B}' = \{B \in \mathcal{B} : f^{-1}(B) \in \mathcal{A}\}$. Then $\mathcal{B}' \supset \mathcal{C}$, and $\mathcal{B}'$ is closed under countable unions since for any $B_1, B_2, \dots \in \mathcal{B}'$ and $B = \bigcup_{n=1}^{\infty} B_n$, we have $f^{-1}(B) = \bigcup_{n=1}^{\infty} f^{-1}(B_n) \in \mathcal{A}$. Similarly, $\mathcal{B}'$ is closed under complements and countable intersections, so it is a σ-algebra that contains $\mathcal{C}$, and thus $\mathcal{B}' = \mathcal{B}$ since $\mathcal{C}$ generates $\mathcal{B}$. □

EXAMPLE 1.78. If X, Y are metric spaces and $\mathcal{A}, \mathcal{B}$ are their Borel σ-algebras, then every continuous map $f\colon X \to Y$ is measurable: indeed, given any open $U \subset Y$, we have $f^{-1}(U) \subset X$ open by continuity, hence $f^{-1}(U) \in \mathcal{A}$. Since open sets generate the Borel σ-algebra, Lemma 1.77 proves that f is measurable.

REMARK 1.79. Upon encountering Definition 1.76 for the first time, one may naturally wonder why it is formulated in terms of preimages (f^{-1}) rather than images (f). Several answers may be given.

- This is analogous to the definition of continuous map via preimages of open sets, which is what makes Example 1.78 work, so that continuous maps are measurable.
- In the case of a random variable $\varphi\colon X \to \mathbb{R}$ and a probability measure μ on $(X, \mathcal{A})$, we want to be able to make sense of "the probability that φ takes a value in B", where $B \subset \mathbb{R}$ is Borel measurable, and this requires that the set $\varphi^{-1}(B)$ be measurable.
- Thinking of a probability measure μ on $(X, \mathcal{A})$ as describing the distribution of some amount of mass across the space X, we may imagine that this mass is moved by the map f and is thereafter distributed across the space Y, and ask what measure on $(Y, \mathcal{B})$ describes this distribution. As we will see in the next definition, this requires the measurability property formulated in Definition 1.76.

Lec 18
Wed, Feb 26

DEFINITION 1.80. Given two measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$, a measurable map $f\colon X \to Y$, and a measure μ on $(X, \mathcal{A})$, the *pushforward of μ under f* is the measure $f_*\mu = \mu \circ f^{-1}$ on $(Y, \mathcal{B})$ defined by $(f_*\mu)(B) = \mu(f^{-1}(B))$ for all $B \in \mathcal{B}$:

$$X \xrightarrow{f} Y \qquad \Rightarrow \qquad [0,1] \xleftarrow{\mu} \mathcal{A} \xleftarrow{f^{-1}} \mathcal{B}.$$

DEFINITION 1.81. If $(X, \mathcal{A}, \mu)$ is a probability space and $\varphi\colon X \to \mathbb{R}$ is a random variable, the *distribution* (or *law*) of φ is the Borel probability measure $\varphi_*\mu$ on $\mathbb{R}$.

Given a measurable transformation $f\colon (X, \mathcal{A}) \to (Y, \mathcal{B})$, the pushforward from Definition 1.80 gives a map $f_*\colon \mathcal{M}(X, \mathcal{A}) \to \mathcal{M}(Y, \mathcal{B})$. In particular, when $(Y, \mathcal{B}) =$

$(X, \mathcal{A})$ so that f can be interpreted as representing the time evolution of a dynamical system, we obtain a map f_* from $\mathcal{M}(X, \mathcal{A})$ to itself, which is another example of *auxiliary dynamics* associated to a dynamical system. We will explore this later in more depth as the *transfer operator* or *Ruelle–Perron–Frobenius operator*,[31] and we will see that when restricted to a suitable space, we encounter once again the phenomenon observed in §1.4 for projectivization of linear maps and in §1.6 for the graph transform: when expansion in phase space is present, it leads to contraction in various auxiliary dynamics.

We can also consider the auxiliary dynamics induced on the space of random variables, using the following elementary fact.

LEMMA 1.82. *If $f\colon (X, \mathcal{A}) \to (Y, \mathcal{B})$ and $g\colon (Y, \mathcal{B}) \to (Z, \mathcal{C})$ are measurable, then $g \circ f\colon (X, \mathcal{A}) \to (Z, \mathcal{C})$ is measurable.*

PROOF. For every $C \in \mathcal{C}$, we have $g^{-1}(C) \in \mathcal{B}$ since g is measurable, so $(g \circ f)^{-1})(C) = f^{-1}(g^{-1}(C)) \in \mathcal{A}$ since f is measurable. □

As a consequence of Lemma 1.82, given a measurable transformation $f\colon (X, \mathcal{A}) \to (Y, \mathcal{B})$ and a random variable $\varphi\colon Y \to \mathbb{R}$, the composition $\varphi \circ f$ is a random variable on X.[32] Again, when the spaces are the same, the map $\varphi \mapsto \varphi \circ f$ becomes an operator on the space of measurable functions, called the *Koopman operator*.

PROPOSITION 1.83. *Given measurable spaces $(X, \mathcal{A})$ and $(Y, \mathcal{B})$, a measurable transformation $f\colon X \to Y$, a probability measure μ on $(X, \mathcal{A})$, and a random variable $\varphi\colon Y \to \mathbb{R}$ that is either nonnegative or integrable with respect to $f_*\mu$, we have $\int_X \varphi \circ f \, d\mu = \int_Y \varphi \, d(f_*\mu)$.*

We will prove Proposition 1.83 momentarily, via an application of what David Williams calls the "standard machine":[33] start with characteristic functions, deduce the result for simple functions (finite linear combinations of characteristic functions), and then apply the Monotone Convergence Theorem to establish it for all nonnegative measurable functions, which will also prove it for all integrable functions via decomposition into positive and negative parts.

REMARK 1.84. Informally, Proposition 1.83 says that if we have a measurement φ to make tomorrow and we know the distribution μ of states of the system today, we can either work with $(\varphi \circ f, \mu)$ as a delayed measurement sampled from the present distribution, or with $(\varphi, f_*\mu)$ as an immediate measurement sampled from the future distribution; both approaches give the same result. More formally, it says that the transfer operator ("evolve the distribution forward in time") and the Koopman operator ("wait to make the measurement") are adjoints.

[31] The reader is cautioned that in the literature these terms are usually reserved for an operator acting on a space of functions, to which certain measures can eventually be associated, so the discussion here should not be taken as a precise definition.

[32] This is the *pullback* of φ; note that it moves in the opposite direction to the pushforward.

[33] See §5.12 in "Probability with Martingales", *Cambridge*, 1991.

To prove Proposition 1.83 for characteristic functions, we will use the following elementary observation.

LEMMA 1.85. *Given sets X, Y, a map $f\colon X \to Y$, and $E \subset Y$, we have*

$$\mathbf{1}_E \circ f = \mathbf{1}_{f^{-1}(E)}. \tag{1.84}$$

PROOF. It suffices to observe that

$$\mathbf{1}_E \circ f(x) = \begin{cases} 1 & \text{if } f(x) \in E \\ 0 & \text{otherwise} \end{cases} = \begin{cases} 1 & \text{if } x \in f^{-1}(E) \\ 0 & \text{otherwise} \end{cases} = \mathbf{1}_{f^{-1}(E)}(x). \qquad \square$$

PROOF OF PROPOSITION 1.83. By Lemma 1.85, we see that if $\varphi = \mathbf{1}_B$ for some $B \in \mathcal{B}$, then

$$\int_X \varphi \circ f \, d\mu = \int_X \mathbf{1}_{f^{-1}(B)} \, d\mu = \mu(f^{-1}(B)) = (f_*\mu)(B) = \int_Y \varphi \, d(f_*\mu).$$

This proves the proposition for characteristic functions. Both sides are linear so the result holds for simple functions as well, and since every nonnegative measurable function $\varphi\colon Y \to [0, \infty)$ is an increasing pointwise limit of simple functions φ_n, we have (observing that $\varphi_n \circ f \nearrow \varphi \circ f$ pointwise), by the Monotone Convergence Theorem:

$$\int_X \varphi \circ f \, d\mu = \lim_{n\to\infty} \int_X \varphi_n \circ f \, d\mu = \lim_{n\to\infty} \int_Y \varphi_n \, d(f_*\mu) = \int_Y \varphi \, d(f_*\mu).$$

Finally, every integrable $\varphi \in L^1(f_*\mu)$ can be written as $\varphi = \varphi_+ - \varphi_-$, where $\varphi_\pm$ are nonnegative integrable functions, so

$$\int_X \varphi \circ f \, d\mu = \int_X \varphi_+ \circ f \, d\mu - \int_X \varphi_- \circ f \, d\mu = \int_Y \varphi_* \, d(f_*\mu) - \int_Y \varphi_* \, d(f_*\mu),$$

which proves the proposition. $\square$

1.10.3. Independence and invariance.

DEFINITION 1.86. Given a probability space $(X, \mathcal{A}, \mu)$, two events $A_1, A_2 \in \mathcal{A}$ are *independent* if $\mu(A_1 \cap A_2) = \mu(A_1)\mu(A_2)$. Two random variables $\varphi, \psi\colon X \to \mathbb{R}$ are *independent* if for every pair of Borel sets $B_1, B_2 \subset \mathbb{R}$, the events "$\varphi$ takes a value in B_1" and "ψ takes a value in B_2" are independent:

$$\mu(\{x \in X : \varphi(x) \in B_1 \text{ and } \psi(x) \in B_2\}) = \mu(\varphi^{-1}(B_1))\mu(\psi^{-1}(B_2)). \tag{1.85}$$

More generally, a set of random variables $\{\varphi_i\}_{i\in I}$ is independent if for every finite $J \subset I$ and all Borel sets $B_i \subset \mathbb{R}$, $i \in I$, we have

$$\mu\Big(\bigcap_{i=J} \varphi_i^{-1}(B_i)\Big) = \prod_{i\in J} \mu(\varphi_i^{-1}(B_i)). \tag{1.86}$$

REMARK 1.87. It suffices to check (1.85) and (1.86) in the case when $B_i = (-\infty, t_i]$ for some $t_i \in \mathbb{R}$. This fact can be proved using Lemma 1.69 and the fact that the family $\{(-\infty, t] : t \in \mathbb{R}\}$ is a π-system that generates $\mathcal{B}$; see §4.2 in Williams' book.

REMARK 1.88. Clearly, the question of whether φ and ψ are independent involves properties of the measure μ, not just the functions $\varphi, \psi \colon X \to \mathbb{R}$. In the previous section, random variables were described as measurable functions, without reference to a particular probability measure, but whenever we wish to discuss anything involving the distribution of a random variable, we must fix a choice of probability measure, even if it is not explicitly stated.

REMARK 1.89. The definition of independence of random variables can be reformulated using pushforwards: $\varphi, \psi \colon X \to \mathbb{R}$ are independent if the joint random variable $\varphi \times \psi \colon X \to \mathbb{R}^2$ defined by $(\varphi \times \psi)(x) = (\varphi(x), \psi(x))$ has the property that $(\varphi \times \psi)_* \mu = (\varphi_* \mu) \times (\psi_* \mu)$. Thus there is a strong connection between independence of random variables and the question of whether a certain measure has a product structure. This will reappear later.

DEFINITION 1.90. Two random variables φ, ψ carried by a probability space $(X, \mathcal{A}, \mu)$ are *identically distributed* if $\varphi_* \mu = \psi_* \mu$. By Lemma 1.69, this is equivalent to the condition that

$$\mu(\{x \in X : \varphi(x) \le t\}) = \mu(\{x \in X : \psi(x) \le t\}) \quad \text{for all } t \in \mathbb{R}. \tag{1.87}$$

A set of random variables is identically distributed if any two are identically distributed.

Now we bring dynamics into the discussion. Fix a probability space $(X, \mathcal{A}, \mu)$ and a measurable transformation $f \colon X \to X$. Given a measurable function $\varphi \colon X \to \mathbb{R}$, Lemma 1.82 shows that $\varphi \circ f^n$ is measurable for every $n \ge 0$.[34] If φ represents a measurement of the system, the sequence $\varphi \circ f^n$ represents measurements made at different times, and we want to study this sequence as a sequence of random variables with respect to the measure μ. We should not expect these random variables to be independent in general, since one of $f^n(x)$ and $f^k(x)$ determines the other.[35]

▶ EXERCISE 1.31. Let f be the standard map on the torus $\mathbb{T}^2$ equipped with Lebesgue measure, and give an example of a function $\varphi \colon \mathbb{T}^2 \to \mathbb{R}$ such that φ and $\varphi \circ f$ are not independent.

▶ EXERCISE 1.32. Let $S = \{1, \dots, d\}$, let $\sigma \colon S^{\mathbb{N}} \to S^{\mathbb{N}}$ be the shift map defined by $\sigma(x)_n = x_{n+1}$, and let μ the Bernoulli measure associated to a probability vector p as in Example 1.75.

(1) Define a function $\varphi \colon S^{\mathbb{N}} \to \mathbb{R}$ by $\varphi(x) = x_1$. Prove that the sequence of random variables $\varphi \circ \sigma^n$ are independent and identically distributed.
(2) Define a function $\psi \colon S^{\mathbb{N}} \to \mathbb{R}$ by $\psi(x) = x_1 + x_2$. Prove that the random variables $\psi \circ \sigma^n$ are not independent, but are identically distributed.

[34] If f is invertible and f^{-1} is measurable then we can also work with negative values of n, but for now we do not assume this.

[35] There are exceptions, as we will see, and we will eventually study situations in which the *correlation* between $\varphi \circ f^n$ and $\varphi \circ f^k$ decreases as $|n - k|$ becomes large, but for now we will proceed without any assumption of independence.

With independence unavailable in general, we may at least hope to have the random variables $\varphi \circ f^n$ be identically distributed with respect to μ. The following condition on (μ, f) will guarantee this.

DEFINITION 1.91. Given a measurable space $(X, \mathcal{A})$, a probability measure $\mu \in \mathcal{M}(X, \mathcal{A})$, and a measurable transformation $f\colon X \to X$, we say that the measure μ is *f-invariant* if $f_*\mu = \mu$. In this case we also say that f *preserves* μ, and refer to f as a *measure-preserving transformation*.

REMARK 1.92. There are some instances in which we have a given transformation f, and may wish to consider different measures μ; in this case it is more common to use the terminology "f-invariant" to single out those that are fixed by f_*. There are also instances in which we have a given reference measure, such as Lebesgue, and may wish to consider different transformations f; in this case it is more common to use the terminology "measure-preserving" to single out those transformations whose pushforwards fix μ.

The invariance/measure-preserving property can be checked (and used) via a number of equivalent conditions.

PROPOSITION 1.93. *Given a measure space $(X, \mathcal{A}, \mu)$ and a measurable transformation $f\colon X \to X$, the following are equivalent.*

(1) $f_*\mu = \mu$.
(2) $\mu(f^{-1}A) = \mu(A)$ *for every* $A \in \mathcal{A}$.
(3) For every measurable $\varphi\colon X \to \mathbb{R}$ *that is either nonnegative or μ-integrable, we have* $\int \varphi \circ f \, d\mu = \int \varphi \, d\mu$.
(4) For every measurable $\varphi\colon X \to \mathbb{R}$, *the sequence* $(\varphi \circ f^n)_n$ *is identically distributed.*
(5) There is a π-system $\mathcal{C} \subset \mathcal{A}$ *that generates* $\mathcal{A}$ *with the property that* $\mu(f^{-1}A) = \mu(A)$ *for every* $A \in \mathcal{C}$.

PROOF. 1⇔2 This is immediate from the definition $f_*\mu = \mu \circ f^{-1}$.
1⇒3 This follows from Proposition 1.83, which gave $\int \varphi \circ f \, d\mu = \int \varphi \, d(f_*\mu)$.
3⇒2 This follows by taking $\varphi = \mathbf{1}_A$ and using Lemma 1.85 to get

$$\mu(f^{-1}A) = \int \mathbf{1}_{f^{-1}A} \, d\mu = \int \mathbf{1}_A \circ f \, d\mu = \int \mathbf{1}_A \, d\mu = \mu(A).$$

1⇒4 If $f_*\mu = \mu$, then $(\varphi \circ f)_*\mu = \varphi_* f_* \mu = \varphi_* \mu$, and by induction we have $(\varphi \circ f^n)_*\mu = \varphi_*\mu$ for all $n \in \mathbb{N}$.
4⇒3 The probability measure $\varphi_*\mu$ on $\mathbb{R}$ determines $\int \varphi \, d\mu$, so if $(\varphi \circ f)_*\mu = \varphi_*\mu$, then $\int \varphi \circ f \, d\mu = \int \varphi \, d\mu$.
2⇒5 This is immediate.
5⇒1 If the measures μ and $f_*\mu = \mu \circ f^{-1}$ agree on $\mathcal{C}$, then by Lemma 1.69, they agree on $\mathcal{A}$. □

EXAMPLE 1.94. Let $X = \mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$, and let $f\colon X \to X$ be the standard map defined in (1.13). The Jacobian determinant of f is 1 at every point, so given any open set $U \subset \mathbb{R}^2$, the usual change-of-variables formula for diffeomorphisms shows that U and $f^{-1}(U)$ have the same two-dimensional Lebesgue measure. Since open sets in $\mathbb{R}^2$ form a π-system that generates the Borel σ-algebra, it follows from Proposition 1.93 that f preserves two-dimensional Lebesgue measure.

▶ EXERCISE 1.33. Prove that the Bernoulli measure from Example 1.75 is σ-invariant.

Lec 19
Fri, Feb 28

1.10.4. Recurrence. Although we will not prove it here, the flow associated to the 3-body problem preserves Lebesgue measure. Poincaré used this fact to deduce that despite the presence of the homoclinic tangle described in §1.9, we can still make precise statements about the behavior of typical trajectories; he proved the following results.

THEOREM 1.95 (Poincaré Recurrence Theorem). *Let $(X, \mathcal{A}, \mu)$ be a measure space with $\mu(X) < \infty$, and $f\colon X \to X$ a measure-preserving transformation. Given any $A \in \mathcal{A}$, μ-a.e. $x \in A$ has an orbit that returns to A infinitely often; that is, there exists $A' \in \mathcal{A}$ with $\mu(A \setminus A') = 0$ such that for every $x \in A'$, the set $\{n \in \mathbb{N} : f^n(x) \in A\}$ is infinite.*

PROOF. Let $A_n = \bigcup_{k \geq n} f^{-k}A$ be the set of points in X whose orbit enters A sometime at or after time n, and let $A' = \bigcap_{n=0}^{\infty} A_n$. Then A' has the desired recurrence property, and it remains to show that $\mu(A \setminus A') = 0$.

Observe that $A_n = A_{n+1} \cup f^{-n}A \supset A_{n+1}$, so we have

$$A_0 \supset A_1 \supset A_2 \supset \cdots \supset A' = \bigcap_{n=0}^{\infty} A_n \quad \Rightarrow \quad A_0 \setminus A' = \bigcup_{n=0}^{\infty} A_n \setminus A_{n+1},$$

and since $A \subset A_0$, monotonicity and countable subadditivity of μ give

$$\mu(A \setminus A') \leq \mu(A_0 \setminus A') \leq \sum_{n=0}^{\infty} \mu(A_n \setminus A_{n+1}).$$

Moreover, we have

$$A_{n+1} = \bigcup_{k \geq n} f^{-(k+1)A} = f^{-1}\Big(\bigcup_{k \geq n} f^{-k}A\Big) = f^{-1}(A_n),$$

so invariance gives $\mu(A_{n+1}) = \mu(A_n)$, and since μ is finite, we deduce that $\mu(A_n \setminus A_{n+1}) = 0$ for every n, which completes the proof. □

The word "recurrence" in Theorem 1.95 refers to returns to a given set. Another notion of recurrence involves returns to smaller and smaller sets, which requires us to work in a metric space where the word "smaller" makes sense.[36]

[36] In fact a topological space would be enough. But we will stick to the metric setting.

DEFINITION 1.96. Let X be a metric space and consider a map $f\colon X \to X$. Say that a point $x \in X$ is *recurrent* if for every $\epsilon > 0$, there exists $n \geq 1$ such that $f^n(x) \in B(x, \epsilon)$.

THEOREM 1.97. *Let X be a separable metric space, μ a finite Borel measure on X, and $f\colon X \to X$ a measure-preserving transformation. Then μ-a.e. point is recurrent in the sense of Definition 1.96.*

PROOF. Since X is a separable metric space, there exists a countable collection $\{U_n\}_{n=1}^{\infty}$ of open sets such that every open set in X contains at least one of the sets U_n. By Theorem 1.95, for each n there exists a set $Z_n \subset U_n$ such that

- $\mu(Z_n) = 0$, and
- every $x \in U_n \setminus Z_n$ has an orbit that returns to U_n infinitely often.

Let $Z = \bigcup_{n=1}^{\infty} Z_n$. Then $\mu(Z) = 0$, and given any $x \in X \setminus Z$, we see that the orbit of x returns infinitely often to every U_n that contains x. In particular, x is recurrent. □

Observe that homoclinic orbits are *not* recurrent under either f or f^{-1}. Thus the standard map has infinitely many points with this strong nonrecurrence property. In fact, it has uncountably many:

▶ EXERCISE 1.34. Let S be a finite set. Prove that $x \in S^{\mathbb{Z}}$ is recurrent under the shift map σ if and only if for every $n \in \mathbb{N}$, there exists $k \geq 1$ such that $x_{-n} \cdots x_0 x_1 \cdots x_n = x_{k-n} \cdots x_k x_{k+1} \cdots x_{k+n}$.

▶▶ EXERCISE 1.35. Prove that the standard map contains uncountably many points that are not recurrent for either f or f^{-1}. (Hint: first prove it for $(S^{\mathbb{Z}}, \sigma)$, then use the Poincaré–Birkhoff–Smale Theorem.)

We conclude this section by pointing out that although the horseshoe provided by the Poincaré–Birkhoff–Smale Theorem exhibits very rich dynamical behavior that we can describe quite precisely, this horseshoe is in fact only a small part of the standard map's dynamics: it is a Lebesgue-null set, and thus in order to understand the dynamics of Lebesgue-typical points, one must look beyond the horseshoe.

1.11. Birkhoff and the ergodic theorem

The Poincaré Recurrence Theorem 1.95 provides a *qualitative* recurrence result. It is natural to ask whether there is also a *quantitative* version that provides information about the frequency of returns to the set A. For example, we would intuitively expect that a typical orbit would return more frequently when $\mu(A)$ is large, and less frequently when $\mu(A)$ is small.

To make this more precise, let $R_A(x, n)$ denote the number of times $k \in \{0, 1, \ldots, n-1\}$ such that $f^k(x) \in A$, and consider the following questions.

- Does the limit $\lim_{n\to\infty} \frac{1}{n} R_A(x, n)$ exist for every, or at least almost every, point x? This would represent the average frequency of returns to A along the orbit of x; let us denote this by $r_A(x)$, when it exists.

- If $r_A(x)$ exists, does it depend on the choice of x? Or can different orbits return with different frequencies?
- How is the value of $r_A(x)$ related to $\mu(A)$?

We address these in reverse order, since the third turns out to be the easiest, while the first is the hardest.

Observe that by Lemma 1.85, we have

$$R_A(x,n) = \sum_{k=0}^{n-1} \mathbf{1}_A(f^k x) = \sum_{k=0}^{n-1} \mathbf{1}_{f^{-k}A}(x), \tag{1.88}$$

so for every $n \in \mathbb{N}$, invariance of μ gives

$$\int_X \frac{1}{n} R_A(x,n)\, d\mu(x) = \frac{1}{n} \sum_{k=0}^{n-1} \mu(f^{-k}A) = \mu(A). \tag{1.89}$$

Since $\frac{1}{n} R_A(x,n) \in [0,1]$ for every $x \in X$ and $n \in \mathbb{N}$, the Dominated Convergence Theorem implies that if $r_A(x)$ exists for μ-a.e. x, then we have

$$\int r_A(x)\, d\mu(x) = \mu(A). \tag{1.90}$$

Moreover, if it turns out that r_A is constant, or at least constant μ-a.e., then that constant value must be $\mu(A)$, consistent with our earlier intuition.

This answers the third question, contingent on answering the first two. Turning our attention to the second question, we observe that there is one simple mechanism that might lead $r_A(x)$ to take different values for different choices of x: it could be that some orbits start outside A and never enter A at all. (Note that the Poincaré Recurrence Theorem 1.95 gives no information about orbits starting outside A.) For such points x we would have $r_A(x) = 0$, while (1.90) implies that r_A must take positive values somewhere else.

To see if there is another way that r_A could vary, let us consider systems in which this phenomenon does not occur, in the following sense.

DEFINITION 1.98. Given a measure-preserving transformation f on a probability space $(X, \mathcal{A}, \mu)$, let us say that a set $A \in \mathcal{A}$ is a *sweep-out set*[37] if $\mu(\bigcup_{n=0}^{\infty} f^{-n}(A)) = 1$; equivalently, if μ-a.e. orbit enters A. The transformation is *ergodic* if every set $A \in \mathcal{A}$ with $\mu(A) > 0$ is a sweep-out set.

As with invariance, ergodicity is a property of the pair (f, μ); in situations where the transformation is regarded as given, it is common to speak of ergodic *measures*. The following exercise provides the simplest examples of ergodic measures.

▶ EXERCISE 1.36. Let $(X, \mathcal{A})$ be a measurable space such that $\{x\} \in \mathcal{A}$ for every $x \in X$, and let $f\colon X \to X$ be a measurable transformation. Suppose that $p \in X$ has the property that $f^n(p) = p$ for some $n \in \mathbb{N}$. Prove that the probability measure $\mu := \frac{1}{n} \sum_{k=0}^{n-1} \delta_{f^k p}$ is invariant and ergodic.

[37] This terminology is not entirely standard, but will be useful. Be warned that occasionally the same term is used to refer to a set A such that $\bigcup_{n=0}^{\infty} f^n(A)$ has full measure.

Definition 1.98 says that ergodic transformations are precisely those for which the Poincaré Recurrence Theorem 1.95 can be strengthened to say that recurrence occurs for μ-a.e. point in X, not just in A: that is, given any $A \in \mathcal{A}$ with $\mu(A) > 0$, there is a set $X_A \in \mathcal{A}$ with $\mu(X \setminus X_A) = 0$ such that for every $x \in X_A$, the set $\{n \in \mathbb{N} : f^n(x) \in A\}$ is infinite.

REMARK 1.99. Following the argument in the proof of Theorem 1.97, we see that if X is a separable metric space, μ a Borel probability measure on X that is *fully supported* (meaning that $\mu(U) > 0$ for every open $U \subset X$), and $f\colon X \to X$ an *ergodic* measure-preserving transformation, then μ-a.e. $x \in X$ has a dense orbit.

One often sees ergodicity introduced via one of several equivalent conditions instead of Definition 1.98. We will formulate some of these in Proposition 1.102 below, but first we need the following notions of invariance.

DEFINITION 1.100. Given a measure-preserving transformation f on a probability space $(X, \mathcal{A}, \mu)$, a set $A \in \mathcal{A}$ is *invariant* if $f^{-1}(A) = A$, and *invariant mod zero* if $\mu(f^{-1}(A) \triangle A) = 0$, where we recall that $A \triangle B = (B \setminus A) \cup (A \setminus B)$. A measurable function $\varphi\colon X \to \mathbb{R}$ is *invariant* if $\varphi \circ f = \varphi$, and *invariant mod zero* if $\varphi \circ f(x) = \varphi(x)$ for μ-a.e. x.

The following technical lemma allows us to pass from "invariant mod zero" to "invariant" by considering the set of points whose orbits return infinitely often. It uses a construction similar to the proof of the Poincaré Recurrence Theorem 1.95.

LEMMA 1.101. *Let f be a measure-preserving transformation on a probability space $(X, \mathcal{A}, \mu)$. If $A \in \mathcal{A}$ is invariant mod zero, then the set*

$$A_\infty := \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} f^{-k}(A) \tag{1.91}$$

is invariant and satisfies $\mu(A_\infty \triangle A) = 0$.

PROOF. Let $A_n := \bigcup_{k=n}^{\infty} f^{-k}(A)$ and observe that $A_1 \supset A_2 \supset A_3 \supset \cdots$. Moreover, we have

$$f^{-1}(A_n) = \bigcup_{k=n}^{\infty} f^{-1}(f^{-k}(A)) = \bigcup_{\ell=n+1}^{\infty} f^{-\ell}(A) = A_{n+1},$$

and thus since the sets are nested decreasing,

$$f^{-1}(A_\infty) = \bigcap_{n=1}^{\infty} f^{-1}(A_n) = \bigcap_{n=1}^{\infty} A_{n+1} = A_\infty.$$

This proves invariance. To show that $\mu(A_\infty \triangle A) = 0$, first observe that for every n, we have

$$A_n \triangle A = \Big(\bigcup_{k=n}^{\infty} f^{-k}(A) \Big) \triangle A \subset \bigcup_{k=n}^{\infty} \big(f^{-k}(A) \triangle A \big). \tag{1.92}$$

For any sets $U, V, W \subset X$, we have $(U \triangle V) \subset (U \triangle W) \cup (W \triangle V)$, and thus

$$f^{-k}(A) \triangle A \subset \bigcup_{\ell=0}^{k-1} \big(f^{-(\ell+1)}(A) \triangle f^{-\ell}(A)\big) = \bigcup_{\ell=0}^{k-1} f^{-\ell}\big(f^{-1}(A) \triangle A\big). \tag{1.93}$$

Since $\mu(f^{-1}(A) \triangle A) = 0$ and μ is invariant, (1.93) gives $\mu(f^{-k}(A) \triangle A) = 0$ for every $k \in \mathbb{N}$, and by (1.92) we conclude that

$$\mu(A_n \triangle A) = 0 \text{ for every } n \in \mathbb{N}. \tag{1.94}$$

Finally, since $A \setminus A_\infty \subset A \setminus \bigcup_{n=1}^\infty A_n$ and $A_\infty \setminus A \subset A_n \setminus A$ for every n, we have $A_\infty \triangle A \subset \bigcup_{n=1}^\infty A_n \triangle A$, so (1.94) gives $\mu(A_\infty \triangle A) = 0$, proving the lemma. □

PROPOSITION 1.102. *Let f be a measure-preserving transformation on a probability space $(X, \mathcal{A}, \mu)$. The following are all equivalent to ergodicity.*

(1) For every $A \in \mathcal{A}$ with $\mu(A) > 0$, we have $\mu(\bigcup_{n=0}^\infty f^{-n}(A)) = 1$.
(2) For every $A, B \in \mathcal{A}$ with $\mu(A) > 0$ and $\mu(B) > 0$, there exists $n \in \mathbb{N}$ such that $\mu(f^{-n}(A) \cap B) > 0$.
(3) If $A \in \mathcal{A}$ is invariant, then $\mu(A) = 0$ or 1.
(4) If $A \in \mathcal{A}$ is invariant mod zero, then $\mu(A) = 0$ or 1.
(5) If $\varphi\colon X \to \mathbb{R}$ is invariant, then φ is constant μ-a.e.,[38] meaning that there exists $c \in \mathbb{R}$ such that $\varphi(x) = c$ for μ-a.e. x.
(6) If $\varphi\colon X \to \mathbb{R}$ is invariant mod zero, then φ is constant μ-a.e.
(7) If $\nu \in \mathcal{M}(X, \mathcal{A})$ is an f-invariant probability with $\nu \ll \mu$, then $\nu = \mu$.
(8) If $\mu = p_1\nu_1 + p_2\nu_2$ for some f-invariant probabilities $\nu_1, \nu_2 \in \mathcal{M}(X, \mathcal{A})$ and some $p_1, p_2 \in (0, 1)$ such that $p_1 + p_2 = 1$, then $\nu_1 = \nu_2 = \mu$.

REMARK 1.103. The property of being ergodic was originally called *metrically transitive*; Condition 2 reflects this way of thinking. The term *ergodic* originates from the ergodic hypothesis in statistical mechanics, under which "the time average is equal to the space average". See Corollary 1.106 below for a result that captures this principle in the setting of dynamical systems, and which will allow for a strengthening of Condition 2.

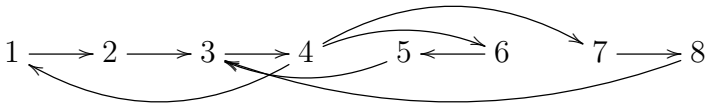
Condition 8 is an irreducibility property, saying that ergodic measures are precisely those which cannot be written as a convex combination of other invariant probability measures. Geometrically, ergodic measures are the extreme points of the space of invariant measures, a fact to which we shall return later on.

The characterization of ergodicity most relevant to our discussion of recurrence frequencies at the beginning of this section is Condition 5 via invariant functions; see Exercise 1.38 below and the discussion following it.

Lec 20
Mon, Mar 3

[38] The term *essentially constant* is sometimes used.

PROOF OF PROPOSITION 1.102. Condition 1 is the definition of ergodicity given above. To prove that it is equivalent to Conditions 2–8, we will prove the implications shown:



1⇒2 If $A, B \in \mathcal{A}$ have $\mu(A) > 0$ and $\mu(B) > 0$, then Condition (1) gives $1 = \mu(\bigcup_{n=0}^{\infty} f^{-n}(A))$, hence

$$\mu(B) = \mu\Big(\Big(\bigcup_{n=0}^{\infty} f^{-n}A\Big) \cap B\Big) = \mu\Big(\bigcup_{n=0}^{\infty} \big(f^{-n}(A) \cap B\big)\Big) \le \sum_{n=0}^{\infty} \mu\big(f^{-n}(A) \cap B\big).$$

Since $\mu(B) > 0$, some term in the sum must be positive.

2⇒3 If $A \in \mathcal{A}$ is invariant, then for every $n \ge 0$, we have $f^{-n}(A) = A$, and thus $f^{-n}(A) \cap A^c = \emptyset$. It follows from the contrapositive of Condition 2 that either $\mu(A) = 0$ or $\mu(A^c) = 0$.

3⇒4 If $A \in \mathcal{A}$ is invariant mod zero, then by Lemma 1.101 there is an invariant set $A_\infty \in \mathcal{A}$ such that $\mu(A_\infty \triangle A) = 0$. By Condition 3 we have $\mu(A_\infty) = 0$ or 1, so the same is true of A.

4⇒1 Given $A \in \mathcal{A}$ with $\mu(A) > 0$, let $B = \bigcup_{n=0}^{\infty} f^{-n}(A)$, and observe that $f^{-1}(B) = \bigcup_{n=1}^{\infty} f^{-n}(A)$. In particular,

$$B \triangle f^{-1}(B) = \{x \in A : f^n(x) \notin A \text{ for all } n \ge 1\},$$

which has μ-measure 0 by the Poincaré Recurrence Theorem 1.95. By Condition 4, we have $\mu(B) = 0$ or $\mu(B) = 1$. Since $B \supset A$ and $\mu(A) > 0$, we conclude that $\mu(B) = 1$.

4⇒6 Let $\varphi\colon X \to \mathbb{R}$ be measurable and invariant mod zero. Given $t \in \mathbb{R}$, let

$$A_t := \varphi^{-1}((-\infty, t]) = \{x \in X : \varphi(x) \le t\},$$

and observe that $f^{-1}(A_t) = \{x \in X : \varphi \circ f(x) \le t\}$, so if $x \in A_t \triangle f^{-1}(A_t)$ then we either have $\varphi(x) \le t < \varphi(f(x))$, or $\varphi(f(x)) \le t < \varphi(x)$. Since φ is invariant mod zero, we conclude that $\mu(A_t \triangle f^{-1}(A_t)) = 0$, and by Condition 4, this implies that $\mu(A_t) = 0$ or 1.

Observing that the sets A_t are nested, we see that $t \mapsto \mu(A_t)$ is nondecreasing in t, and that $\lim_{t\to-\infty} \mu(A_t) = \mu(\emptyset) = 0$ while $\lim_{t\to\infty} \mu(A_t) = \mu(X) = 1$, so $c := \inf\{t \in \mathbb{R} : \mu(A_t) = 1\} \in \mathbb{R}$, and we have $\varphi = c$ μ-a.e.

6⇒5 This implication is immediate.

5⇒3 If $A \in \mathcal{A}$ is invariant, then $\varphi = \mathbf{1}_A$ is invariant by Lemma 1.85: $\varphi \circ f = \mathbf{1}_A \circ f = \mathbf{1}_{f^{-1}(A)} = \mathbf{1}_A = \varphi$. Condition 5 implies that $\mathbf{1}_A$ is constant μ-a.e., so $\mu(A) = 0$ or 1.

4⇒7 Given $\nu \ll \mu$ as in Condition 7, let $h = \frac{d\nu}{d\mu}$ be the Radon–Nikodym derivative. Let $A = \{x \in X : h(x) < 1\}$ and $B = \{x \in X : h(x) > 1\}$. Observe that exactly one of following two cases occurs:

- $\mu(A) = \mu(B) = 0$ and thus $h = 1$ μ-a.e., so $\nu = \mu$.
- $\mu(A) > 0$, in which case we must also have $\mu(B) > 0$ since $\int_X h\, d\mu = \nu(X) = 1$. Thus in this case, we have $0 < \mu(A) < 1$.

In particular, to prove that $\nu = \mu$, it will suffice to show that A is invariant mod zero and then apply Condition 4, which rules out the second case above.

We start with the following general observation: given any f-invariant $m \in \mathcal{M}(X, \mathcal{A})$ and any $E \in \mathcal{A}$, we have

$$E = (E \cap f^{-1}E) \sqcup (E \setminus f^{-1}E),$$
$$f^{-1}(E) = (E \cap f^{-1}E) \sqcup (f^{-1}E \setminus E).$$

Since $m(E) = m(f^{-1}(E))$, this implies that

$$m(E \setminus f^{-1}E) = m(f^{-1}E \setminus E). \tag{1.95}$$

We will apply (1.95) both with $m = \nu$ and $m = \mu$, and with $E = A$. We have

$$\begin{aligned} \int_{A \setminus f^{-1}A} h\, d\mu = \nu(A \setminus f^{-1}A) = \nu(f^{-1}A \setminus A) = \int_{f^{-1}A \setminus A} h\, d\mu \\ \geq \mu(f^{-1}A \setminus A) = \mu(A \setminus f^{-1}A), \end{aligned} \tag{1.96}$$

where the first and third equalities come from the definition of the Radon–Nikodym derivative, the second and final equalities come from (1.95), and the inequality uses the fact that $h \geq 1$ on A^c. Since $h < 1$ on A, (1.96) implies that $\mu(A \setminus f^{-1}A) = 0$. By (1.95), we conclude that $\mu(f^{-1}A \setminus A) = 0$ as well, so $\mu(A \triangle f^{-1}A) = 0$. As discussed above, this suffices.

7⇒8 If $\mu = p_1\nu_1 + p_2\nu_2$, where $\nu_i \in \mathcal{M}(X, \mathcal{A})$ are both f-invariant, and $p_i \in (0, 1)$ satisfy $p_1 + p_2 = 1$, then given any $A \in \mathcal{A}$ with $\mu(A) = 0$, we have $p_1\nu_1(A) + p_2\nu_2(A) = 0$, so $\nu_i(A) = 0$. Thus $\nu_i \ll \mu$, and by Condition 7, this implies that $\nu_1 = \nu_2 = \mu$.

8⇒3 We prove the contrapositive. Suppose there exists $A \in \mathcal{A}$ with $f^{-1}(A) = A$ and $0 < \mu(A) < 1$. Let $p_1 = \mu(A)$ and $p_2 = 1 - \mu(A)$. Define measures ν_1 and ν_2 by

$$\nu_1(E) = \frac{1}{p_1}\mu(A \cap E) \quad \text{and} \quad \nu_2(E) = \frac{1}{p_2}\mu(A^c \cap E).$$

Then $p_1\nu_1 + p_2\nu_2 = \mu$. Moreover, given any $E \in \mathcal{A}$, we have

$$A \cap f^{-1}(E) = f^{-1}(A) \cap f^{-1}(E) = f^{-1}(A \cap E)$$

since A is invariant, so invariance of μ gives

$$\nu_1(f^{-1}(E)) = \frac{1}{p_1}\mu(A \cap f^{-1}(E)) = \frac{1}{p_1}\mu(f^{-1}(A \cap E)) = \frac{1}{p_1}\mu(A \cap E) = \nu_1(E).$$

We conclude that ν_1 is invariant, and a similar argument proves that ν_2 is as well. □

▶ EXERCISE 1.37. Prove that n in Condition 2 can be taken to be arbitrarily large: if (f, μ) is ergodic, then for every $A, B \in \mathcal{A}$ with $\mu(A) > 0$ and $\mu(B) > 0$, and every $k \in \mathbb{N}$, there exists $n \geq k$ such that $\mu(f^{-n}(A) \cap B) > 0$.

REMARK 1.104. Despite our long list of conditions equivalent to ergodicity, we do not yet have any *examples* of ergodic measures apart from the periodic orbit measures in Exercise 1.36. Let us mention the following facts now, deferring proofs until later: the Bernoulli measures of Example 1.75 are ergodic for the shift map on $S^{\mathbb{Z}}$, while Lebesgue measure on $\mathbb{T}^2$ is not ergodic for the standard map.[39]

Lec 21
Wed, Mar 5

Now recall the motivating questions from the beginning of this section, which concerned the frequency of returns to some set $A \in \mathcal{A}$.

▶ EXERCISE 1.38. Let f be a measure-preserving transformation on a probability space $(X, \mathcal{A}, \mu)$. Fix $A \in \mathcal{A}$ and define a function $\overline{r}_A \colon X \to [0,1]$ by

$$\overline{r}_A(x) := \overline{\lim_{n\to\infty}} \frac{1}{n} R_A(x,n).$$

Prove that $\overline{r}_A$ is measurable and invariant.

It follows from Exercise 1.38 that if f is ergodic, then the function $\overline{r}_A$ is constant μ-a.e. If in addition we know that the limit $r_A(x)$ exists μ-a.e., then we conclude that $r_A(x)$ is constant μ-a.e., and (1.90) implies that $r_A(x) = \mu(A)$ for μ-a.e. $x \in X$.

It only remains to address the first question from the beginning of this section: do the functions $x \mapsto \frac{1}{n} R_A(x,n)$ converge pointwise as $n \to \infty$? From (1.88) these functions can be rewritten as $\frac{1}{n}\sum_{k=0}^{n-1} \mathbf{1}_A \circ f^k$. Thus the problem can be framed in terms of one of our motivations from §1.10 for introducing the language of measure-preserving transformations: to take a measurable function $\varphi \colon X \to \mathbb{R}$ and treat the sequence of functions $\varphi \circ f^n$ as a sequence of random variables. One important problem in probability theory is to understand the limiting behavior as $n \to \infty$ of the average of the first n random variables in such a sequence: for example, in the IID setting, these averages converge almost everywhere[40] to the mean of their common distribution. This is the *strong law of large numbers*.

We can now prove the Birkhoff[41] Ergodic Theorem, which provides the analogous result in the framework of ergodic theory. To state the result, we need to consider for each $\varphi \in L^1(\mu)$, $n \in \mathbb{N}$, and $x \in X$ the *ergodic sum*

$$\mathsf{S}_n\varphi(x) = \sum_{k=0}^{n-1} \varphi(f^k x) = \varphi(x) + \varphi(fx) + \cdots + \varphi(f^{n-1}x), \tag{1.97}$$

[39] More precisely, there is a very large set of parameters K for which it has been proved that ergodicity fails, and there are no parameters for which it has been proved that ergodicity holds.

[40] Or *almost surely* to use the terminology more common in probability theory.

[41] Named for George David Birkhoff, who proved it in 1931. We also encountered him at the end of §1.9 through his work on homoclinic tangles. As described in §1.2 of Ulrich Krengel's "Ergodic Theorems" (De Gruyter, 1985), Birkhoff's paper proved the result in the case when φ is a characteristic function, and the case of more general functions appeared in a 1933 paper by Khinchin in *Mathematische Annalen*, following Birkhoff's ideas. Because of this, the result is sometimes called the *Birkhoff–Khinchin Ergodic Theorem*.

as well as the *ergodic average*[42]

$$\mathsf{A}_n\varphi(x) = \frac{1}{n}\mathsf{S}_n\varphi(x) = \frac{1}{n}\sum_{k=0}^{n-1}\varphi(f^k x). \tag{1.98}$$

THEOREM 1.105 (Birkhoff Ergodic Theorem). *Let f be a measure-preserving transformation on a probability space $(X, \mathcal{A}, \mu)$. Given $\varphi \in L^1(\mu)$, the limit*

$$\tilde{\varphi}(x) := \lim_{n\to\infty} \mathsf{A}_n\varphi(x) \tag{1.99}$$

exists for μ-a.e. x, and $\mathsf{A}_n\varphi \to \tilde{\varphi}$ in L^1. The function $\tilde{\varphi}$ is f-invariant and $\int \tilde{\varphi}\, d\mu = \int \varphi\, d\mu$.

Given an ergodic system, the f-invariant function $\tilde{\varphi}$ in (1.99) must be equal to $\int \tilde{\varphi}\, d\mu = \int \varphi\, d\mu$ at μ-a.e. point, and so Theorem 1.105 has the following consequence, which generalizes[43] the strong law of large numbers.

COROLLARY 1.106 (Birkhoff Ergodic theorem – ergodic case). *Let f be an ergodic measure-preserving transformation on a probability space $(X, \mathcal{A}, \mu)$. Given any $\varphi \in L^1(\mu)$, we have $\mathsf{A}_n\varphi(x) \to \int \varphi\, d\mu$ for μ-a.e. x.*

PROOF OF THEOREM 1.105. There are various proofs of the Birkhoff Ergodic Theorem. I learned the argument here from Omri Sarig's lecture notes on ergodic theory,[44] although I have modified the exposition. He in turn follows Keane,[45] who attributes it to Kamae.[46]

We will prove the theorem for nonnegative functions, and leave the extension to integrable functions as an exercise. Assume that $\varphi\colon X \to [0, \infty)$ is integrable, and consider the functions

$$\overline{A}(x) = \overline{\lim_{n\to\infty}}\, \mathsf{A}_n\varphi(x) \quad \text{and} \quad \underline{A}(x) = \underline{\lim_{n\to\infty}}\, \mathsf{A}_n\varphi(x). \tag{1.100}$$

Observe that $\overline{A}$ and $\underline{A}$ are measurable and f-invariant (as in Exercise 1.38). Our goal is to prove that $\overline{A}(x) = \underline{A}(x)$ for μ-a.e. x.

[42] These are sometimes also referred to as the *Birkhoff sum* and *Birkhoff average*. The notation $\mathsf{S}_n\varphi$ is standard; the notation $\mathsf{A}_n\varphi$, which appears for example in Einsiedler and Ward's book "Ergodic Theory with a view towards Number Theory", is less common but is helpful.

[43] To show that this is a generalization, we still need to provide an argument that independence implies ergodicity of an appropriate transformation, such as the Bernoulli measures, as suggested in Remark 1.104.

[44] Currently available at `https://www.weizmann.ac.il/math/sarigo/sites/math.sarigo/files/uploads/ergodicnotes.pdf` – see Theorem 2.2 for the proof, and see the notes at the end of Chapter 2 (page 88) for some history.

[45] Trieste conference, 1989, Oxford Press, pages 35–70.

[46] *Israel Journal of Mathematics*, 1982. It seems Kamae found the proof using some ideas from nonstandard analysis. See also a paper by Katznelson and Weiss that year in the same journal. This proof also appears in §3.3.2 of "Foundations of Ergodic Theory" by Marcelo Viana and Krerley Oliveira, Cambridge, 2016.

Since $\overline{A} - \underline{A} \geq 0$, this will follow if we show that $\int (\overline{A} - \underline{A})\, d\mu \leq 0$, or equivalently, $\int \overline{A}\, d\mu \leq \int \underline{A}\, d\mu$. We will accomplish this by proving that

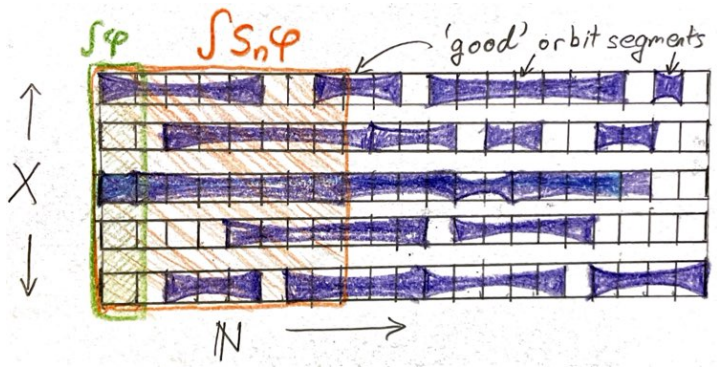
$$\int \overline{A}\, d\mu \leq \int \varphi\, d\mu \leq \int \underline{A}\, d\mu. \tag{1.101}$$

In doing so, we will need the following consequence of invariance of μ:

$$\int \mathsf{S}_n \varphi\, d\mu = \sum_{k=0}^{n-1} \int \varphi \circ f^k\, d\mu = n \int \varphi\, d\mu. \tag{1.102}$$

Figure 1.52 illustrates the overall strategy: treating $[i, j) \times \{x\} \subset \mathbb{N} \times X$ as a finite segment of the orbit of x,[47] we will "pack" each $[0, N) \times \{x\}$ with orbit segments on which we have good control of $\mathsf{A}_n \varphi$. We will do this twice, once for each of the inequalities in (1.101).

FIGURE 1.52. Controlling $\int \varphi$ by packing $[0, N) \times X$ with orbit segments.



STEP 1: $\int \varphi \leq \int \underline{A}$ *when φ is nonnegative and bounded.*

We start by considering the case when φ is bounded, so there exists $L > 0$ such that $0 \leq \varphi \leq L$. This implies that $0 \leq \underline{A} \leq L$ as well.

Fix $\epsilon > 0$. Let us call a set $[i, j) \times \{x\}$ a "good orbit segment" if

$$\mathsf{A}_{j-i} \varphi(f^i x) \leq \underline{A}(x) + \epsilon. \tag{1.103}$$

A disjoint set of good orbit segments in $[0, N) \times X$ will be called a "good packing".

Suppose we have chosen a good packing. For each $x \in X$, let $b(x)$ denote the number of elements $k \in [0, N)$ that are *not* contained in any of the good orbit segments we have chosen; we refer to these as "bad times". Then the length of the chosen orbit segments in $[0, N) \times \{x\}$ add up to $N - b(x)$. From (1.103) we see that on each good orbit segment $[i, j) \times \{x\}$, we have

$$\mathsf{S}_{j-i} \varphi(f^i x) \leq (j - i)(\underline{A}(x) + \epsilon).$$

[47] We abuse notation by using $[i, j)$ to refer to an interval of integers, namely $\{i, i+1, \ldots, j-1\}$.

Summing over the chosen orbit segments gives

$$\mathsf{S}_N\varphi(x) \le (N - b(x))(\underline{A}(x) + \epsilon) + b(x)L \le N(\underline{A}(x) + \epsilon) + b(x)L. \tag{1.104}$$

With (1.104) in mind, our goal is to produce a good packing for which $b(x) \in [0, N)$ is as small as possible, at least on average.[48] The general result providing this is Lemma 1.109 below, but instead of stating this result immediately, we first describe how good packings can be produced.

If we were packing $\mathbb{N} \times X$, then we could do use the definition of $\underline{A}$ and its f-invariance to argue as follows:

- there exists $k_1 \in \mathbb{N}$ such that $\mathsf{A}_{k_1}\varphi(x) \le \underline{A}(x) + \epsilon$;
- there exists $k_2 \in \mathbb{N}$ such that $\mathsf{A}_{k_2}\varphi(f^{k_1}x) \le \underline{A}(f^{k_1}x) + \epsilon = \underline{A}(x) + \epsilon$;
- similarly, there exist $k_j \in \mathbb{N}$ such that writing $n_j = k_1 + k_2 + \cdots + k_{j-1}$, each $[n_j, n_{j+1}) \times \{x\}$ is a good orbit segment.

The orbit segments obtained this way provide a good packing with *no* bad times! However, this does not produce a good packing of $[0, N) \times X$, because we might choose an orbit segment that is too long: if we have chosen $k_1, \dots, k_{j-1}$, and there are no values of k_j such that $n_j + k_j \le N$, then we must treat n_j as a bad time, and try again starting at $n_j + 1$.

Lec 22
Fri, Mar 7

Describing this process a little more carefully will give our algorithm for producing a good packing of $[0, N) \times X$. First we frame things in a more flexible way, so that we can reuse these ideas later on. Let us represent the "good orbit segments" by

$$\mathcal{G} := \{(x, n) \in X \times \mathbb{N} : \mathsf{A}_n(x) \le \underline{A}(x) + \epsilon\}. \tag{1.105}$$

By the definition of $\underline{A}$, this collection of orbit segments has the following property of being "eventually good":

$$\text{for every } x \in X, \text{ there exists } n \in \mathbb{N} \text{ such that } (x, n) \in \mathcal{G}. \tag{1.106}$$

In what follows, we will not use the specific definition of $\mathcal{G}$ from (1.105), but only the fact that it is a subset of $X \times \mathbb{N}$ that satisfies (1.106).

Given $x \in X$, a *good packing* of $[0, N)$ will mean a collection P_x of disjoint subintervals $[i, j) \subset [0, N)$ with the property that $(f^i x, j - i) \in \mathcal{G}$ for each $[i, j) \in P_x$. A *good packing* of $[0, N) \times X$ will mean a function $P\colon x \mapsto P_x$, where each P_x is a good packing of $[0, N)$.

Given a good packing P of $[0, N) \times X$, let $b_P(x)$ denote the number of elements of $[0, N)$ that are not contained in any of the intervals $[i, j)$ from P_x; these are the "bad times". Now we describe an algorithm that produces a packing of $[0, N) \times X$ for which $b_P(x)$ is small on average.

[48] We cannot hope to make $b(x)$ small at *every* point x: for a given value of N, there could be points x such that $\varphi(f^k x) = L$ for all $k \in [0, N)$, but $\varphi(f^k x) = 0$ for all $k \ge N$, so $\overline{A}(x) = 0$, but $\mathsf{A}_{j-i}(f^i x) = L$ for all $i, j \in [0, N)$.

ALGORITHM 1.107. *Note that by* (1.106)*, the following quantity is finite for every* $x \in X$*, but may be arbitrarily large:*

$$\tau(x) = \min\{n \geq 1 : (x, n) \in \mathcal{G}\}. \tag{1.107}$$

Given $x \in X$*, produce a disjoint set* P_x *of good intervals in* $[0, N)$ *as follows.*

(1) Set our "present location in time" to be 0*.*

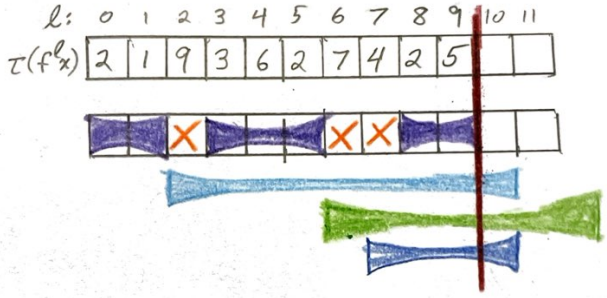
(2) Writing $\ell \in [0, N)$ *for our present location in time, check whether or not* $k := \ell + \tau(f^\ell x) \leq N$*. This determines whether or not there is a* $\mathcal{G}$*-orbit segment beginning at* $f^\ell x$ *that finishes by time* N*.*

(a) If $k \leq N$*, then add* $[\ell, k-1)$ *to* P_x*, and let* $\ell' := k$*.*

(b) If $k > N$*, then declare* ℓ *to be a "bad time" for* x*, and let* $\ell' = \ell + 1$*.*

(3) If $\ell' = N$*, then terminate the algorithm, otherwise declare our present location in time to be* ℓ'*, and return to Step 2.*

FIGURE 1.53. Producing a good packing using Algorithm 1.107.



EXAMPLE 1.108. Suppose $N = 10$ and the sequence $\tau(f^\ell x)$ is as shown in Figure 1.53. Then the first time through the algorithm produces the good interval $[0, 2)$, and updates the present location to $\ell = 2$. Because the shortest good orbit segment starting here is too long (length 9, so that $k = 2 + 9 > 10 = N$), we declare $\ell = 2$ to be a "bad time", and move to $\ell = 3$, where we obtain the good interval $[3, 6)$. The times $\ell = 6$ and $\ell = 7$ are then marked as bad, because τ is too large, and at $\ell = 8$ we obtain the good interval $[8, 10)$. Thus $P_x = \{[0, 2), [3, 6), [8, 10)\}$, and $b_P(x) = 3$.

Algorithm 1.107 produces a good packing P with the following property: if $\ell \in [0, N)$ is a bad time for x, then $\tau(f^\ell x) > N - \ell$. (Note that the converse need not hold.) In particular, we have

$$b_P(x) \leq \sum_{\ell=0}^{N-1} \mathbf{1}_{Z_{N-\ell}}(f^\ell x), \qquad \text{where } Z_k := \{z \in X : \tau(z) > k\}.$$

Using invariance of μ, this gives

$$\int b_P(x)\, d\mu(x) \leq \sum_{\ell=0}^{N-1} \int \mathbf{1}_{Z_{N-\ell}}\, d\mu = \sum_{\ell=0}^{N-1} \mu(Z_{N-\ell}) = \sum_{k=1}^{N} \mu(Z_k), \tag{1.108}$$

and we are led to the following conclusion.

LEMMA 1.109. *Let $\mathcal{G} \subset X \times \mathbb{N}$ be* any *collection of orbit segments satisfying the "eventually good" property* (1.106). *Then for every $\delta > 0$, there exist $N \in \mathbb{N}$ and a good packing P of $[0, N) \times X$ such that*

$$\frac{1}{N} \int b_P(x)\, d\mu(x) \le \delta. \tag{1.109}$$

PROOF. Define $\tau \colon X \to \mathbb{N}$ by (1.107), and observe that $\tau < \infty$ everywhere by (1.106), so $\mu(Z_k) \to 0$ as $k \to \infty$.

Given $N \in \mathbb{N}$, let $P(N)$ be the good packing of $[0, N) \times X$ produced by the algorithm described above, so that (1.108) gives

$$\lim_{N\to\infty} \frac{1}{N} \int b_{P(N)}(x)\, d\mu(x) \le \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mu(Z_k) = 0,$$

where the last equality follows since $\lim_{k\to\infty} \mu(Z_k) = 0$. □

Fixing $\epsilon > 0$ and returning to the specific choice of $\mathcal{G}$ from (1.105), Lemma 1.109 proves that for every $\delta > 0$, there are $N \in \mathbb{N}$ and a good packing of $[0, N) \times X$ such that (1.109) holds, and then (1.102) and (1.104) together give

$$\int \varphi\, d\mu = \frac{1}{N} \int \mathsf{S}_N \varphi(x)\, d\mu(x) \le \underline{A}(x) + \epsilon + \frac{L}{N} \int b_P\, d\mu \le \underline{A}(x) + \epsilon + L\delta.$$

Since $\epsilon, \delta > 0$ were arbitrary, this shows that

$$\int \varphi\, d\mu \le \int \underline{A}\, d\mu. \tag{1.110}$$

STEP 2: $\int \varphi \le \int \underline{A}$ *for any nonnegative* φ.

Given $L > 0$, let $\varphi_L := \min(\varphi, L)$, so that φ_L takes values in $[0, L]$. The result of Step 1 applies to φ_L, giving

$$\int \varphi_L\, d\mu \le \int \varliminf_{n\to\infty} \mathsf{A}_n \varphi_L(x)\, d\mu \le \int \varliminf_{n\to\infty} \mathsf{A}_n \varphi(x)\, d\mu = \int \underline{A}\, d\mu. \tag{1.111}$$

where the first inequality uses (1.110), and the second uses the fact that $\varphi_L \le \varphi$. Note that here $\underline{A}$ refers to the lower limit associated to φ, not φ_L. In particular, the right-hand side of (1.111) is independent of L, and since $\varphi_L \nearrow \varphi$ pointwise, the Monotone Convergence Theorem gives

$$\int \varphi\, d\mu = \lim_{L\to\infty} \int \varphi_L\, d\mu \le \int \underline{A}\, d\mu.$$

STEP 3: $\int \varphi \ge \int \overline{A}$ *for any nonnegative* φ.

This step follows the same idea as in Step 1, replacing $\underline{A}$ with $\overline{A}$: take $\mathcal{G}$ to be the set of orbit segments (x,n) for which $\mathsf{A}_n(x)$ is close to $\overline{A}(x)$, and then apply Lemma 1.109. There is one small subtlety: it would be natural to interpret "close to $\overline{A}(x)$" as meaning "$\geq \overline{A}(x)-\epsilon$", but if we proceed this way, then in the analogue of (1.104) we would obtain

$$\mathsf{S}_N\varphi(x) \geq (N-b(x))(\overline{A}(x)-\epsilon) = N(\overline{A}(x)-\epsilon) - b(x)(\overline{A}(x)-\epsilon).$$

This will create problems because $b(x)$ is multiplied by a factor that could be arbitrarily large, instead of the constant factor L in (1.104). Thus instead of using $\overline{A}(x)-\epsilon$ as the threshold, we fix $k \in \mathbb{N}$ and use the threshold function $\overline{A}_k(x) := \min(k, \overline{A}(x) - \frac{1}{k})$, defining

$$\mathcal{G} := \{(x,n) \in X \times \mathbb{N} : \mathsf{A}_n\varphi(x) \geq \overline{A}_k(x)\}.$$

This satisfies (1.106), and given any good packing P of $[0,N)\times X$, we have

$$\mathsf{S}_N\varphi(x) \geq (N - b_P(x))\overline{A}_k(x) \geq N\overline{A}_k(x) - b_P(x)k.$$

Fixing $\delta := \frac{1}{k^2} > 0$ and taking N, P to be given by Lemma 1.109, we obtain

$$\int \varphi\,d\mu = \frac{1}{N}\int \mathsf{S}_N\varphi\,d\mu \geq \int \overline{A}_k(x)\,d\mu(x) - \delta k = \int \overline{A}_k\,d\mu - \frac{1}{k}.$$

As in Step 2, we observe that $\overline{A}_k \nearrow \overline{A}$ pointwise as $k\to\infty$, and thus by the Monotone Convergence Theorem,

$$\int \varphi\,d\mu \geq \lim_{k\to\infty}\Big(\int \overline{A}_k\,d\mu - \frac{1}{k}\Big) = \int \overline{A}\,d\mu. \tag{1.112}$$

Combining this with the result of Step 2 proves (1.101), and recalling the discussion there, we conclude that $\tilde{\varphi}(x) := \lim_{n\to\infty}\mathsf{A}_n\varphi(x)$ exists μ-a.e. when $\varphi \geq 0$, and that $\tilde{\varphi}$ is f-invariant. Moreover, (1.101) gives $\int\tilde{\varphi}\,d\mu = \int\varphi\,d\mu$, so it only remains to prove that $\mathsf{A}_n\varphi \to \tilde{\varphi}$ in L^1, and to extend the results to arbitrary integrable functions φ.

Step 4: $\mathsf{A}_n\varphi \to \tilde{\varphi}$ *in* L^1.

To prove that the convergence to $\tilde{\varphi}$ occurs in L^1, not just pointwise, start by fixing $\epsilon > 0$ and choosing a bounded $\psi\colon X\to\mathbb{R}$ such that $\int|\varphi-\psi|\,d\mu < \epsilon$. Then we have

$$\int|\mathsf{A}_n\varphi - \tilde{\varphi}|\,d\mu \leq \overbrace{\int|\mathsf{A}_n\varphi - \mathsf{A}_n\psi|\,d\mu}^{\text{I}} + \overbrace{\int|\mathsf{A}_n\psi - \tilde{\psi}|\,d\mu}^{\text{II}} + \overbrace{\int|\tilde{\psi}-\tilde{\varphi}|\,d\mu}^{\text{III}},$$

and we estimate each term separately.

I We have $|\mathsf{A}_n\varphi - \mathsf{A}_n\psi| = |\mathsf{A}_n(\varphi-\psi)| \leq \mathsf{A}_n|\varphi-\psi|$, so (1.102) gives

$$\int|\mathsf{A}_n\varphi - \mathsf{A}_n\psi|\,d\mu \leq \int \mathsf{A}_n|\varphi-\psi|\,d\mu = \int|\varphi-\psi|\,d\mu < \epsilon. \tag{1.113}$$

II Applying the results so far to ψ, we have $\mathsf{A}_n\psi \to \tilde{\psi}$ pointwise μ-a.e., so by the Bounded Convergence Theorem, $\int |\mathsf{A}_n\psi - \tilde{\psi}|\, d\mu \to 0$. Thus there exists n_0 such that for every $n \geq n_0$, we have $\int |\mathsf{A}_n\psi - \tilde{\psi}|\, d\mu < \epsilon$.

III We have $|\tilde{\psi} - \tilde{\varphi}| = |\lim_{n\to\infty}(\mathsf{A}_n\psi - \mathsf{A}_n\varphi)| = \lim_{n\to\infty} |\mathsf{A}_n\psi - \mathsf{A}_n\varphi|$, so applying Fatou's Lemma and recalling (1.113) gives

$$\int |\tilde{\psi} - \tilde{\varphi}|\, d\mu \leq \varliminf_{n\to\infty} \int |\mathsf{A}_n\psi - \mathsf{A}_n\varphi|\, d\mu \leq \int |\varphi - \psi|\, d\mu < \epsilon. \tag{1.114}$$

Combining these, we see that for every $n \geq n_0$, we have $\int |A_n\varphi - \tilde{\varphi}|\, d\mu < 3\epsilon$. Since $\epsilon > 0$ was arbitrary, this proves that $\mathsf{A}_n\varphi \to \tilde{\varphi}$ in $L^1(\mu)$. □

▶ EXERCISE 1.39. Complete the proof of Theorem 1.105 by decomposing an arbitrary $\varphi \in L^1(\mu)$ as $\varphi = \varphi^+ - \varphi^-$, where $\varphi^\pm \colon X \to [0, \infty)$, then applying the result we proved to each of $\varphi^\pm$ individually.

REMARK 1.110. As mentioned at the beginning of the proof, the overall argument followed here can be found in various other places in the literature, but the exposition here is somewhat unique: in particular, the emphasis on *orbit segments* in (1.105) and Lemma 1.109 differs from the usual descriptions in terms of "colorings". This point of view is influenced by the author's work with Dan Thompson on thermodynamic formalism for non-uniformly hyperbolic systems, where related ideas of decompositions of orbit segments play a central role; this will appear in a later chapter.

The function $\tilde{\varphi}(x) := \lim_{n\to\infty} \mathsf{A}_n\varphi(x)$ in the Birkhoff Ergodic Theorem 1.105 admits another description, which is worth mentioning.

THEOREM 1.111. *Let f be a measure-preserving transformation on a probability space $(X, \mathcal{A}, \mu)$, and let*

$$\mathcal{I} := \{A \in \mathcal{A} : \mu(A \triangle f^{-1}A) = 0\}$$

be the σ-algebra of measurable sets that are invariant mod zero. Given $\varphi \in L^1(X, \mathcal{A}, \mu)$, define a measure ν on $(X, \mathcal{A})$ by $\nu(A) = \int_A \varphi\, d\mu$, and let $\mu_{\mathcal{I}}$, $\nu_{\mathcal{I}}$ denote the restrictions of μ and ν to $\mathcal{I}$. Then the function $\tilde{\varphi}$ from the Birkhoff Ergodic Theorem 1.105 is equal μ-a.e. to the Radon–Nikodym derivative $\frac{d\nu_{\mathcal{I}}}{d\mu_{\mathcal{I}}}$.

PROOF. By uniqueness of the Radon–Nikodym derivative, it suffices to prove that $\tilde{\varphi}$ is $\mathcal{I}$-measurable, and that for every $E \in \mathcal{I}$, we have $\int_E \varphi\, d\mu = \int_E \tilde{\varphi}\, d\mu$. The measurability follows since $\tilde{\varphi}$ is f-invariant, so $f^{-1}((-\infty, t]) \in \mathcal{I}$ for every $t \in \mathbb{R}$. For equality of the integrals, we observe that given $E \in \mathcal{I}$, the function $\varphi\mathbf{1}_E$ has $\mathsf{A}_n(\varphi\mathbf{1}_E) = (\mathsf{A}_n\varphi)\mathbf{1}_E \to \tilde{\varphi}\mathbf{1}_E$ since E is invariant, so applying the ergodic theorem to $\varphi\mathbf{1}_E$ gives $\int \varphi\mathbf{1}_E\, d\mu = \int \tilde{\varphi}\mathbf{1}_E\, d\mu$. □

REMARK 1.112. The function $\varphi \in L^1(X, \mathcal{A}, \mu)$ is the Radon–Nikodym derivative of ν with respect to μ on the measure space $(X, \mathcal{A})$; Theorem 1.111 highlights the important role played by the choice of σ-algebra in the Radon–Nikodym Theorem.

More generally, given a sub-σ-algebra $\mathcal{B} \subset \mathcal{A}$, the Radon–Nikodym derivative $\frac{d\nu_{\mathcal{B}}}{d\mu_{\mathcal{B}}}$ is called the *conditional expectation* of φ with respect to $\mathcal{B}$. This will reappear later on.

Lec 23
Mon, Mar 17

Returning to the questions posed at the beginning of this section, we see that the Birkhoff Ergodic Theorem 1.105 answers the first one: given a set $A \in \mathcal{A}$, the asymptotic frequency of returns $r_A(x) := \lim_{n\to\infty} \frac{1}{n} R_A(x,n)$ exists for μ-a.e. $x \in X$. As we saw earlier, $\int r_A \, d\mu = \mu(A)$, and if μ is ergodic, then $r_A(x) = \mu(A)$ for μ-a.e. x.

One consequence of this fact is the following strengthening of Condition 2 from Proposition 1.102, which characterizes ergodicity as a sort of "asymptotic averaged independence".

PROPOSITION 1.113. *A measure-preserving transformation f on a probability space $(X, \mathcal{A}, \mu)$ is ergodic if and only if for every $A, B \in \mathcal{A}$, we have*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu(f^{-k}(A) \cap B) = \mu(A)\mu(B). \tag{1.115}$$

PROOF. $\Leftarrow$ This direction is straightforward: if $\mu(A) > 0$ and $\mu(B) > 0$, then (1.115) implies that $\mu(f^{-k}(A) \cap \mu(B))$ for some k, verifying ergodicity by Condition 2 from Proposition 1.102.

$\Rightarrow$ By the Birkhoff Ergodic Theorem, the following holds for μ-a.e. x:

$$\frac{1}{n} \mathsf{S}_n \mathbf{1}_A(x) \to \mu(A) \qquad \Rightarrow \qquad \frac{1}{n} \big(\mathsf{S}_n \mathbf{1}_A(x)\big) \mathbf{1}_B(x) \to \mu(A) \mathbf{1}_B(x).$$

By the Bounded Convergence Theorem,

$$\frac{1}{n} \int \big(\mathsf{S}_n \mathbf{1}_A(x)\big) \mathbf{1}_B(x) \, d\mu(x) \to \int \mu(A) \mathbf{1}_B \, d\mu = \mu(A)\mu(B).$$

The left-hand side is equal to

$$\frac{1}{n} \int \sum_{k=0}^{n-1} (\mathbf{1}_A \circ f^k) \mathbf{1}_B \, d\mu = \frac{1}{n} \sum_{k=0}^{n-1} \int \mathbf{1}_{f^{-k}(A)} \mathbf{1}_B \, d\mu = \frac{1}{n} \sum_{k=0}^{n-1} \mu(f^{-k}(A) \cap B),$$

so this proves (1.115). □

Now we address the question of ergodicity for the examples of invariant measures we have studied so far: Bernoulli measures on $S^{\mathbb{N}}$, and Lebesgue measure for the standard map. As with invariance, it will be helpful if we can avoid checking the ergodicity criteria for *every* measurable set, and restrict our attention to a more tractable class.

DEFINITION 1.114. Given a probability space $(X, \mathcal{A}, \mu)$, a collection $\mathcal{Z} \subset \mathcal{A}$ of measurable sets is *μ-dense in $\mathcal{A}$* if for every $A \in \mathcal{A}$ and $\epsilon > 0$, there exists $Z \in \mathcal{Z}$ such that $\mu(Z \triangle A) < \epsilon$.

▶ EXERCISE 1.40. Prove that $\rho_\mu(A,B) := \mu(A \triangle B)$ defines a pseudometric on $\mathcal{A}$. This motivates the terminology "μ-dense".

LEMMA 1.115. *Given a probability space $(X, \mathcal{A}, \mu)$, let $\mathcal{Z} \subset \mathcal{A}$ be an algebra on X that generates the σ-algebra $\mathcal{A}$. Then $\mathcal{Z}$ is μ-dense in $\mathcal{A}$.*

PROOF. Let $\mathcal{B} \subset \mathcal{A}$ be the collection of sets B such that for every $\epsilon > 0$, there exists $Z \in \mathcal{Z}$ satisfying $\mu(Z \triangle B) < \epsilon$. It will suffice to prove that

(1) if $B \in \mathcal{B}$, then $B^c \in \mathcal{B}$, and
(2) if $B_n \in \mathcal{B}$ and $B_1 \subset B_2 \subset \cdots$, then $\bigcup_{n\in\mathbb{N}} B_n \in \mathcal{B}$.

Indeed, these imply that $\mathcal{B}$ is also closed under countable decreasing intersections, so it is a monotone class containing $\mathcal{Z}$, and must therefore contain $\mathcal{A}$ by the Monotone Class Theorem.

Given $B \in \mathcal{B}$, if $Z \in \mathcal{Z}$ satisfies $\mu(Z \triangle B) < \epsilon$, then $\mu(Z^c \triangle B^c) < \epsilon$, and $Z^c \in \mathcal{Z}$ since $\mathcal{Z}$ is an algebra, so $B^c \in \mathcal{B}$.

If $B_1 \subset B_2 \subset \cdots$ and $B_n \in \mathcal{B}$, let $B = \bigcup_{n\in\mathbb{N}} B_n$. Given $\epsilon > 0$, there exists $n \in \mathbb{N}$ such that $\mu(B \setminus B_n) < \epsilon/2$, and $Z \in \mathcal{Z}$ such that $\mu(Z \triangle B_n) < \epsilon/2$, so $\mu(B \triangle Z) < \epsilon$. Thus $B \in \mathcal{B}$, completing the proof. □

PROPOSITION 1.116. *Let f be a measure-preserving transformation on a probability space $(X, \mathcal{A}, \mu)$. Suppose $\mathcal{Z} \subset \mathcal{A}$ is μ-dense, and that there exists $c > 0$ with the following property: for every $A, B \in \mathcal{Z}$,*

$$\text{there exists } n \in \mathbb{N} \text{ such that } \mu(f^{-n}(A) \cap B) \geq c\mu(A)\mu(B). \tag{1.116}$$

Then (f, μ) is ergodic.

REMARK 1.117. The property in (1.116) is a sort of "partial independence". If $\mu(A) = 0$ or $\mu(B) = 0$, then it is automatically true, so we are really only concerned with what happens when both sets have positive measure. Observe that (1.116) is stronger than simply asking that Condition 2 from Proposition 1.102 hold for every $A, B \in \mathcal{Z}$ with positive measure, because we require that the same constant $c > 0$ work for every A, B.

At the same time, (1.116) is weaker than requiring (1.115) as in Proposition 1.113; indeed, if (1.115) holds for some A, B, then (1.116) must hold for every $c \in (0, 1)$.

PROOF OF PROPOSITION 1.116. Given $A, B \in \mathcal{A}$ with $\mu(A) > 0$ and $\mu(B) > 0$, let $\epsilon > 0$ be arbitrary, and use the μ-density property to obtain $Y, Z \in \mathcal{Z}$ such that

$$\mu(Y \triangle A) < \epsilon \quad \text{and} \quad \mu(Z \triangle B) < \epsilon. \tag{1.117}$$

By the hypothesis, there exists $n = n(\epsilon) \in \mathbb{N}$ such that

$$\mu(f^{-n}(Y) \cap Z) \geq c\mu(Y)\mu(Z) \geq c(\mu(A) - \epsilon)(\mu(B) - \epsilon). \tag{1.118}$$

Observe that

$$\big(f^{-n}(A) \cap B\big) \triangle \big(f^{-n}(Y) \cap Z\big) \subset \big(f^{-n}(A) \triangle f^{-n}(Y)\big) \cup \big(B \triangle Z\big)$$

and that $f^{-n}(A)\triangle f^{-n}(Y) = f^{-n}(A\triangle Y)$, so that by invariance,

$$\mu\big(f^{-n}(A)\cap B\big)\triangle\big(f^{-n}(Y)\cap Z\big) \le \mu(A\triangle Y) + \mu(B\triangle Z) < 2\epsilon.$$

Combining this with (1.118) gives

$$\mu(f^{-n}(A)\cap B) \ge c(\mu(A)-\epsilon)(\mu(B)-\epsilon) - 2\epsilon.$$

Since $\epsilon > 0$ was arbitrary, we conclude that

$$\sup_{n\in\mathbb{N}} \mu(f^{-n}(A)\cap B) \ge c\mu(A)\mu(B).$$

By Condition 2 of Proposition 1.102, it follows that (f,μ) is ergodic. □

The following exercise provides a more general version of Proposition 1.116, which will be useful later on.

►► EXERCISE 1.41. Suppose there is a nested family $\mathcal{Z}_1 \subset \mathcal{Z}_2 \subset \cdots \subset \mathcal{A}$ with the following properties.

(1) For every $\epsilon > 0$, there exists $M = M(\epsilon) \in \mathbb{N}$ such that $\mathcal{Z}_M$ is (μ,ϵ)-dense in $\mathcal{A}$, meaning that for every $A \in \mathcal{A}$ there exists $Z \in \mathcal{Z}_M$ such that $\mu(Z\triangle A) < \epsilon$.
(2) For every $M \in \mathbb{N}$, there exists $c = c(M) > 0$ such that for every $Y, Z \in \mathcal{Z}_M$, there exists $n \in \mathbb{N}$ such that $\mu(f^{-n}(Y)\cap Z) \ge c\mu(Y)\mu(Z)$.

Prove that (f,μ) is ergodic.

For now, we will benefit from the following immediate consequence of Lemma 1.115 and Proposition 1.116.

PROPOSITION 1.118. *Let f be a measure-preserving transformation on a probability space $(X,\mathcal{A},\mu)$. Suppose that the σ-algebra $\mathcal{A}$ is generated by an algebra $\mathcal{Z}\subset\mathcal{A}$ with the following property: there exists $c>0$ such that for every $A,B\in\mathcal{Z}$,*

$$\text{there exists } n\in\mathbb{N} \text{ such that } \mu(f^{-n}(A)\cap B) \ge c\mu(A)\mu(B). \tag{1.119}$$

Then (f,μ) is ergodic.

To formulate this criterion for a semialgebra, we need to strengthen (1.119) slightly.

PROPOSITION 1.119. *Let f be a measure-preserving transformation on a probability space $(X,\mathcal{A},\mu)$. Suppose that the σ-algebra $\mathcal{A}$ is generated by a semialgebra $\mathcal{S}\subset\mathcal{A}$ with the following property: there exists $c>0$ such that for every $A,B\in\mathcal{S}$,*

$$\varliminf_{n\to\infty} \mu(f^{-n}(A)\cap B) \ge c\mu(A)\mu(B). \tag{1.120}$$

Then (f,μ) is ergodic.

PROOF. Let $\mathcal{Z}$ be the algebra generated by $\mathcal{S}$, so every $Z\in\mathcal{Z}$ is a finite disjoint union of elements of $\mathcal{S}$. Then given any $A,B\in\mathcal{Z}$, we have

$$A = \bigsqcup_{i=1}^{k} Y_i \quad\text{and}\quad B = \bigsqcup_{j=1}^{\ell} Z_j \qquad \text{for some } Y_i, Z_j \in \mathcal{S}.$$

It follows that for every $n \in \mathbb{N}$, we have

$$f^{-n}(A) \cap B = \bigsqcup_{i,j} f^{-n}(Y_i) \cap Z_j. \tag{1.121}$$

For each i, j, by (1.120) there exists n_{ij} such that for every $n \geq n_{ij}$, we have

$$\mu(f^{-n}(Y_i) \cap Z_j) \geq \frac{c}{2}\mu(Y_i)\mu(Z_j).$$

Taking $n = \max_{i,j} n_{ij}$, we use this together with (1.121) to obtain

$$\mu(f^{-n}(A) \cap B) = \sum_{i,j} \mu(f^{-n}(Y_i) \cap Z_j) \geq \sum_{i,j} \frac{c}{2}\mu(Y_i)\mu(Z_j) = \frac{c}{2}\mu(A)\mu(B).$$

Now ergodicity follows from Proposition 1.118. □

▶ EXERCISE 1.42. Prove that Proposition 1.119 remains true if we replace the condition (1.120) with

$$\varliminf_{k\to\infty} \mu(f^{-n_k}(A) \cap B) \geq c\mu(A)\mu(B), \tag{1.122}$$

where $n_k \to \infty$ is a subsequence that can depend on $\mathcal{S}$ and c but not on A, B.

EXAMPLE 1.120 (Bernoulli measures are ergodic). Let $S = \{1, \dots, d\}$ and consider $X = S^{\mathbb{N}}$ with the Borel σ-algebra $\mathcal{B}$. Let $p = (p_1, \dots, p_d)$ be a probability vector and let μ be the associated Bernoulli measure on $(X, \mathcal{B})$, as in Example 1.75. Recall from Exercise 1.33 that μ is σ-invariant, where $\sigma\colon X \to X$ is the shift map, and from Exercise 1.29 that the collection $\mathcal{C}$ of cylinders is a semialgebra that generates $\mathcal{B}$.

Given any words $v \in S^k$ and $w \in S^\ell$, for every $n \geq \ell$ the corresponding cylinders satisfy

$$\sigma^{-n}([v]) \cap [w] = \bigsqcup_{u \in S^{\ell-n}} [wuv],$$

and by the definition of the Bernoulli measure μ we have[49]

$$\mu[wuv] = \mu[w] \cdot \mu[u] \cdot \mu[v],$$

from which we conclude that

$$\mu(\sigma^{-n}([v]) \cap [w]) = \sum_{u \in S^{\ell-n}} \mu[w] \cdot \mu[u] \cdot \mu[v] = \mu[w] \cdot \mu[v].$$

By Proposition 1.119, it follows that μ is ergodic. This provides a concrete sense in which the Birkhoff Ergodic Theorem (in the form of Corollary 1.106) generalizes the Strong Law of Large Numbers.

Lec 24
Wed, Mar 19

EXAMPLE 1.121 (Lebesgue measure for the twist map). Let $f\colon \mathbb{T}^2 \to \mathbb{T}^2$ be the standard map for some parameter value K. We saw in Example 1.94 that two-dimensional Lebesgue measure m on $\mathbb{T}^2$ is f-invariant. Remark 1.104 mentioned that

[49] We abuse notation slightly and write $\mu[w]$ in place of $\mu([w])$, in order to avoid a deluge of nested brackets in expressions such as "$\mu([w])\mu([u])\mu([v])$".

(f, m) is not ergodic in general. We will not prove this now – it would lead us into KAM theory[50] and would require too great a deviation from the primary thrust of our narrative – but we can examine the *unkicked* twist map $f(x, y) = (x+y, y) \bmod \mathbb{Z}^2$, which corresponds to $K = 0$. Observe that for every $y \in S^1 = \mathbb{R}/\mathbb{Z}$, the circle $S^1 \times \{y\} = (\mathbb{R}/\mathbb{Z}) \times \{y\}$ is f-invariant, and thus any union of these circles is as well. In particular, $S^1 \times [0, \frac{1}{2})$ is an f-invariant set with Lebesgue measure $\frac{1}{2}$, so m is not ergodic. Indeed, in the spirit of Condition 8 from Proposition 1.102, we see that using Fubini's theorem, m can be written as a (continuous) convex combination of one-dimensional Lebesgue measures on the circles $S^1 \times \{y\}$: for any m-integrable $\varphi\colon \mathbb{T}^2 \to \mathbb{R}$, we have

$$\int_{\mathbb{T}^2} \varphi(x, y)\, dm(x, y) = \int_{S^1} \int_{S^1 \times \{y\}} \varphi(x, y)\, dm_y(x)\, dy, \tag{1.123}$$

where m_y is one-dimensional Lebesgue measure on $S^1 \times \{y\}$, and dy means we integrate with respect to one-dimensional Lebesgue measure on S^1.

On the circle $S^1 \times \{y\}$, the twist map f restricts to a rotation by y. (Here the rotation angle is measured in fractions of a full rotation, rather than in degrees or radians.) This rotation preserves Lebesgue measure of intervals, so by Proposition 1.93, each m_y is invariant.

The decomposition in (1.123) is a step towards the *ergodic decomposition*, which we will study later. For now we look at whether or not the measures m_y are ergodic. Since the twist map restricts to rotation by y on each $S^1 \times \{y\}$, we study the following maps: given $\theta \in \mathbb{R}/\mathbb{Z}$, let $R_\theta\colon S^1 \to S^1$ be defined by $R_\theta(x) = x + \theta \pmod 1$. This is a *circle rotation*.

►► EXERCISE 1.43. Show that in the rational case $\theta = p/q$, where $\gcd(p, q) = 1$, every point $x \in S^1$ is R_θ-periodic with minimal period q, and the ergodic measures for R_θ are precisely the periodic orbit measures on these orbits, as in Exercise 1.36.

THEOREM 1.122. *If $\theta \in \mathbb{R}/\mathbb{Z}$ is irrational, then:*

(1) given any $x \in S^1$, the orbit $\{R_\theta^n(x) : n \geq 0\}$ is dense in S^1; and
(2) R_θ is ergodic with respect to Lebesgue measure m on S^1.

PROOF. Density For density, start by observing that $R_\theta^n(x) = R_x R_\theta^n(0)$, so it suffices to prove that the orbit of 0 is dense. Given $x \in \mathbb{R}/\mathbb{Z}$, let $|x| \in [0, \frac{1}{2}]$ denote the distance from x to the nearest integer. Let $\theta_0 = \theta$ and choose $n \in \mathbb{N}$ such that 1 lies in the interval between $n|\theta_0|$ and $(n+1)|\theta_0|$, dividing it into two subintervals of positive length (because $\theta \notin \mathbb{Q}$). The whole interval has length $|\theta_0|$, so one of the subintervals must have length $\leq \frac{1}{2}\theta_0$; that is, we can take $n_0 \in \{n, n+1\}$ with the property that $|R_{\theta_0}^{n_0}(0)| = |n_0\theta_0| \leq \frac{1}{2}|\theta_0|$.

Let $\theta_1 = n_0\theta_0$ and then follow the above procedure to choose $n_1 \in \mathbb{N}$ such that $|n_1\theta_1| \leq \frac{1}{2}|\theta_1| \leq \frac{1}{4}|\theta_0|$. Putting $\theta_2 = n_1\theta_1 = n_1 n_0 \theta$ and iterating this procedure, we obtain $n_k \in \mathbb{N}$ and $\theta_k \in S^1$ for which $\theta_{k+1} = n_k\theta_k$ and $|\theta_k| \leq 2^{-k}|\theta|$.

[50] Named after Andrey Kolmogorov, Vladimir Arnold, and Jürgen Moser.

Taking $N_k := \prod_{j=0}^{k-1} n_j$, we have $R_{\theta_k} = R_\theta^{N_k}$. Given any $\epsilon > 0$, we can take $k \in \mathbb{N}$ sufficiently large that $2^{-k}|\theta| < \epsilon$, so $R_\theta^{N_k}$ is rotation by a nonzero number smaller than ϵ. Its iterates therefore enter every ϵ-neighborhood in S^1. Since $\epsilon > 0$ was arbitrary, this proves the density result.

Ergodicity Now we prove that (R_θ, m) is ergodic. Suppose $A \subset S^1$ is measurable and R_θ-invariant, with $m(A) > 0$. By the Lebesgue Density Theorem, there exists $x \in S^1$ and $\epsilon > 0$ such that for every interval I centered at x with length $< \epsilon$, we have $m(A \cap I) > \frac{4}{5} m(I)$.

We claim that given any interval $J \subset S^1$ with length $< \epsilon$, we have $m(A \cap J) > \frac{3}{5} m(J)$. To prove this, let I be the interval centered at x with length $\frac{3}{4}|J|$. Since the orbit of x is dense, there exists $n \in \mathbb{N}$ such that $R_\theta^n(I) \subset J$. Invariance of A gives $A \cap J \supset A \cap R_\theta^n(I) = R_\theta^n(A \cap I)$, and since m is invariant, we have

$$m(A \cap J) \geq m(A \cap I) > \frac{4}{5} m(I) = \frac{4}{5} \cdot \frac{3}{4} m(J) = \frac{3}{5} m(J).$$

It follows that A^c has no Lebesgue density points, so $m(A^c) = 0$. □

▶ EXERCISE 1.44. Modify the above proof of ergodicity to avoid the use of the Lebesgue Density Theorem, as follows.

(1) Prove that if $A \subset \mathbb{R}$ and $\gamma \in (0,1)$ are such that $m(A \cap I) \leq \gamma m(I)$ for every interval $I \subset \mathbb{R}$ then $m(A) = 0$.
(2) Prove that given any measurable $A \subset \mathbb{R}$ with $m(A) > 0$, and any $\gamma \in (0,1)$, there is a sequence of intervals $I_n \subset \mathbb{R}$ such that $m(A \cap I_n) > \gamma m(I_n)$ for every n, and $|I_{n+1}| = \frac{1}{2}|I_n|$.
(3) Use the previous part to prove ergodicity, following the ideas in the proof of Theorem 1.122.

We now present an alternative proof of Theorem 1.122, which leads to an even stronger conclusion. Given $k \in \mathbb{Z}$, let $e_k \colon \mathbb{R}/\mathbb{Z} \to \mathbb{C}$ be given by $e_k(x) = e^{2\pi i k x}$. Observe that if $k = 0$, we have $e_0 \equiv 1$, so $\mathsf{A}_n e_0 \equiv 1$ for every $n \in \mathbb{N}$. Otherwise, for every $k \neq 0$, we can use the formula for the sum of a geometric series with ratio $z = e^{2\pi i k \theta}$ to obtain

$$\mathsf{A}_n e_k(x) = \frac{1}{n} \sum_{j=0}^{n-1} e^{2\pi i k (x + j\theta)} = \frac{1}{n} e^{2\pi i k x} \sum_{j=0}^{n-1} z^j = \frac{1}{n} e^{2\pi i k x} \cdot \frac{z^n - 1}{z - 1}.$$

From this we deduce that $\|\mathsf{A}_n e_k\| \leq \frac{2}{n}$, where $\|\cdot\|$ is the uniform norm, so $\mathsf{A}_n e_k \to 0$ uniformly in x. Writing $\mathcal{T} \subset C(S^1)$ for the linear span of the functions $\{e_k : k \in \mathbb{Z}\}$, we conclude that for every $\psi \in \mathcal{T}$, we have $\mathsf{A}_n \psi \to \int \psi \, dm$ uniformly.

By the Stone–Weierstrass Theorem, $\mathcal{T}$ is dense in $C(S^1)$ (with respect to the uniform norm), so given any $\varphi \in C(S^1)$ and $\epsilon > 0$, there exists $\psi \in \mathcal{T}$ with $\|\varphi - \psi\| < \epsilon/3$. This implies that $\|\mathsf{A}_n \varphi - \mathsf{A}_n \psi\| < \epsilon/3$, and taking n sufficiently large that $\|\mathsf{A}_n \psi - \int \psi \, dm\| < \epsilon/3$, we conclude that $\|\mathsf{A}_n \varphi - \int \varphi \, dm\| < \epsilon$. In other

words, we have proved that:

$$\text{given any } \varphi \in C(S^1), \text{ we have } \frac{1}{n}\sum_{k=0}^{n-1} \varphi \circ R_\theta^k \to \int \varphi \, dm \text{ uniformly.} \tag{1.124}$$

PROPOSITION 1.123. *If $\theta \in \mathbb{R}/\mathbb{Z}$ is irrational, then Lebesgue measure m is the only R_θ-invariant Borel probability measure on S^1.*

PROOF. Given $\varphi \in C(S^1)$, it follows from (1.124) that the function $\tilde{\varphi}$ from the Birkhoff Ergodic Theorem is constant everywhere and takes the value $\int \varphi \, dm$. Thus if μ is any R_θ-invariant Borel probability measure on S^1, we have $\int \varphi \, d\mu = \int \tilde{\varphi} \, d\mu = \int \varphi \, dm$. This holds for every continuous φ, so $\mu = m$. □

Proposition 1.123 (which we proved without relying on Theorem 1.122) implies that m is ergodic, as a consequence of Condition 8 from Proposition 1.102. In fact, the property in Proposition 1.123 is called *unique ergodicity*.

We have now studied two classes of ergodic measures, in addition to the trivial examples of periodic orbit measures: Bernoulli measures on $S^{\mathbb{N}}$, and Lebesgue measure for irrational circle rotations. It is worth highlighting the following facts.

- The shift map arose naturally in our study of dynamical systems with hyperbolic behavior, and it supports many ergodic measures; we have Bernoulli measures associated to every probability vector, and in fact we will later encounter an even broader array of ergodic measures for this system.
- An irrational circle rotation does not display any hyperbolic behavior at all: it is an isometry. Instead of large families of ergodic measures, as for the full shift, there is only a single invariant measure.

Later, we will explore in more detail the connection between hyperbolic behavior and the existence of a rich and large space of invariant measures. For the moment, we conclude this section by describing one consequence of ergodicity of Bernoulli measures.

Lec 25
Fri, Mar 21

DEFINITION 1.124. Given an integer $b \geq 2$, let $S_b := \{0, 1, \ldots, b-1\}$. A real number $x \in [0,1]$ is *normal in base b* if its base-b representation $\bar{x}^b \in S_b^{\mathbb{N}}$ contains every finite sequence of digits with the same asymptotic frequency. More precisely, given any $k \in \mathbb{N}$ and $w \in S_b^k$, for each $n \geq k$ we write

$$R_x(w,n) := \#\{j \in [0, n-k) : \bar{x}^b_{[j,j+k)} = w\}$$

for the number of times the word w is repeated in $x_{[0,n)}$, and say that x is normal in base b if for every such k, w, we have

$$\lim_{n\to\infty} \frac{1}{n} R_x(w,n) = b^{-k}. \tag{1.125}$$

THEOREM 1.125. *Lebesgue-a.e. $x \in [0,1]$ is normal in every integer base.*

PROOF. Given $b \geq 2$, let μ_b denote the Bernoulli measure on $S_b^{\mathbb{N}}$ associated to the probability vector $(\frac{1}{b}, \ldots, \frac{1}{b})$, so that $\mu_b([w]) = b^{-k}$ for every $k \in \mathbb{N}$ and $w \in S_b^k$.

Since μ_b is ergodic for the shift map, the Birkhoff Ergodic Theorem tells us that there is a μ_b-null set $Z_b \subset S_b^{\mathbb{N}}$ such that for every $z \in Z_b^c$, we have

$$\lim_{n\to\infty} \frac{1}{n} \mathsf{S}_n \mathbf{1}_{[w]}(z) = \mu_b([w]) = b^{-k}. \tag{1.126}$$

Defining $\pi^b \colon S_b^{\mathbb{N}} \to [0,1]$ by $\pi^b(z) = \sum_{n=1}^{\infty} z_k b^{-k}$, it follows from (1.126) that every $x \in \pi^b(Z_b)^c = \pi^b(Z_b^c)$ is normal in base b. Moreover, the pushforward measure $\pi^b_* \mu_b$ on $[0,1]$ has the property that every interval of the form $[jb^{-k}, (j+1)b^{-k}]$ has weight b^{-k}, since its preimage under π^b is $[w] \subset S_b^{\mathbb{N}}$ for some $w \in S_b^k$. It follows that $\pi^b_* \mu_b$ is Lebesgue measure m on $[0,1]$, and thus $m(\pi^b(Z_b)) = \mu_b(Z_b) = 0$. Taking $Z = \bigcup_{b\geq 2} \pi^b(Z_b) \subset [0,1]$, we see that $m(Z) = 0$ and that every $x \in Z^c$ is normal in every integer base. □

REMARK 1.126. The original 1909 proof of Theorem 1.125 by Émile Borel[51] was more direct, involving an application of the Borel–Cantelli Lemma to the sets of "unfavorable" numbers where the frequencies of repetitions in $\bar{x}^b_{[0,n)}$ deviate from the expected value by more than an appropriately chosen threshold.

1.12. Lorenz and a strange attractor

Our main physical examples so far – the pendulum, the standard map, and the 3-body problem – have all had the property that they preserve Lebesgue measure. Such systems are often referred to as *conservative*.

We saw in §1.10 and §1.11 that the measure-preserving nature of these systems leads to strong results, including existence of the asymptotic ergodic averages $\lim_{n\to\infty} \mathsf{A}_n \varphi(x)$ for any integrable φ and Lebesgue-a.e. x. What are we to do, however, if we are confronted with a system that does *not* preserve Lebesgue measure?

Many such nonconservative systems arise quite naturally: for example, if we return to the example of the pendulum but now incorporate the effect of friction, it is no longer conservative, and apart from trajectories lying on the stable manifold of the hyperbolic fixed point (the configuration where the pendulum points directly upwards), every trajectory converges as $t \to \infty$ to the stable fixed point where the pendulum hangs straight down. In this example, then, it does not matter that we have lost the invariance of Lebesgue measure: its *dissipative* nature allows us to say exactly what the asymptotic behavior of the system is. Here we use "dissipative" to mean that time evolution decreases Lebesgue measure: if m is Lebesgue measure and A, A' are two positive measure sets such that A is mapped onto A' after some time $t > 0$, then $m(A') < m(A)$.

Although the dissipative system in the previous paragraph settles into a stable equilibrium, we learn in a first course in differential equations that this need not always be the case: a harmonic oscillator that is both damped *and* periodically forced approaches a stable periodic orbit, not a fixed point. Here again, though, there is no amplification of initial measurement errors, and the system remains

[51] See §§7–13 of "Les probabilités dénombrables et leurs applications arithmétiques", *Rendiconti del Circolo Matematico di Palermo* **27** (1909), pages 247–271.

predictable, so the non-invariance of Lebesgue measure does not matter. We are led to the following two questions.

(1) Are there dissipative systems that exhibit the "frequent unpredictability" we observed for the standard map?
(2) If such systems exist, is it possible to study them using the tools of ergodic theory, despite the fact that Lebesgue measure is not invariant?

The answer to both questions is a resounding "yes". We will address the first question in this section; the second will need to wait for a later chapter and the tools of thermodynamic formalism.

As with §1.9 regarding Poincaré's work, the material presented in this section contains a fair amount of history, describing how Edward Lorenz observed the phenomenon of "sensitive dependence" in a simplified weather model, which he initially presented at a meteorology conference in Tokyo in November 1960, and how he went on to give a description in a 1963 paper of what would eventually become known as a "strange attractor" exhibiting "chaotic" behavior. References for the material in this section include:

[Lor62]: The written version of Lorenz's 1960 presentation, published as "The Statistical Prediction of Solutions of Dynamic Equations", *Proc. Internat. Symp. on Numerical Weather Prediction*, Tokyo, 1962.

[Lor63]: Lorenz's most-cited paper: "Deterministic Nonperiodic Flow", *J. Atmospheric Sci.* **20** (1963), no. 2, 130–141.

[Lor93]: A book expanding on a series of three lectures (the Jessie and John Danz Lectures) Lorenz gave at the University of Washington in 1990: "The Essence of Chaos", Univ. of Washington Press, 1993. Pages 130–146 give Lorenz's own account of the story below.

To describe the context of Lorenz's work, let us start by identifying two possible approaches to making predictions.[52]

- *Statistical:* use the patterns in past "time series" observations to make predictions about the future, without reference to an underlying model that governs how different variables evolve in time.[53]
- *Dynamic:* make predictions about the future by developing a mathematical model that describes the time evolution of the system, measuring the current state of the system as accurately as possible, and then finding the solution of the model that corresponds to the observed initial condition.

In the 1950s, the most well-studied type of statistical forecasting used linear methods, and some meteorologists argued that these could perform just as well as any

[52] A more detailed description of these two approaches was given by Lorenz in "The Predictability of Hydrodynamic Flow", *Transactions of the New York Academy of Sciences*, Ser. II, Vol. 25, No. 4, Feb. 1963, pp. 409–433, which also appears as the final chapter of "Simplified dynamic equations and their use in the study of atmospheric predictability", Air Force Cambridge Research Laboratories, Office of Aerospace Research, United States Air Force, 1963.

[53] In Lorenz's words [Lor93, p. 130], statistical forecasting is "based on observations of what has happened in the past, rather than on physical principles".

other approach. In 1956, Edward Lorenz, who had recently taken over the Statistical Forecasting Project in the Department of Meteorology at MIT, and whose training was in dynamic forecasting, began "investigating the feasibility of forecasting sea-level pressures by linear regression methods", with the goal of "obtaining a numerical solution of a set of deterministic equations, and then investigating the predictability of this solution by linear regression methods" [Lor62]. The idea was that "if these formulas could really match up to any other forecasting scheme, they would have to perform perfectly, since one could easily 'predict' the 'data' perfectly simply by running the computer program a second time" [Lor93, p. 132].

This led to the task of determining which set of equations to study numerically. In 1955, Norman Phillips had performed "a numerical simulation of the atmosphere's general circulation for a period of 1 month", in which the "instantaneous state of the modeled atmosphere is determined by roughly 500 numbers".[54] Given the numerical experiments he had in mind, Lorenz wanted a model that had fewer variables, but would still display "the irregular fluctuations which characterize the atmosphere, and which make statistical forecasting so difficult" [Lor62, p. 632]. In particular, it should avoid not only the stable periodic behavior of the forced damped harmonic oscillator, but also stable *quasi-periodic* behavior, which can also be well-predicted by linear methods.

To aid his investigations, Lorenz obtained a "small" computer – "about the size of a large desk" – in whose programming he was assisted by Margaret Hamilton[55] from 1959–1961, and then by Ellen Fetter.[56] After some experimentation, he began studying a simplified model of global atmospheric circulation with 14 variables, before eventually eliminating two more variables to obtain a 12-dimensional system.[57]

[54]Phillips's work was published as "The general circulation of the atmosphere: A numerical experiment", *Quart. J. Roy. Meteor. Soc.* **82** (1956), 123–164. The quotes here are from pages 41 and 47 in John M. Lewis's paper "Clarifying the Dynamics of the General Circulation: Phillips's 1956 Experiment", *Bulletin of the American Meteorological Society* **79**, No. 1, 1998, pp. 39–60, which describes a good deal more of the history and details.

[55]Who would go on to head teams that wrote software for the Apollo program and for Skylab.

[56]Many accounts have focused solely on Lorenz, but the project relied heavily on the work of Hamilton and Fetter, as Lorenz himself acknowledged in both [Lor62] (for Hamilton) and [Lor63] (for Fetter). For a more detailed version of their role, see "The Hidden Heroines of Chaos", Joshua Sokol, *Quanta Magazine*, May 20, 2019 – `https://www.quantamagazine.org/the-hidden-heroines-of-chaos-20190520/` – which suggests that by today's standards, both women would likely have been included as co-authors.

[57]Lorenz describes the details of his general model in "Energy and Numerical Weather Prediction", *Tellus* (1960), 12:4, 364–373, and various other simplifications in "Maximum Simplification of the Dynamic Equations", *Tellus* (1960), 12:3, 243–254. The 14-dimensional model is described in detail in "The Mechanics of Vacillation", *J. Atmos. Sci.* **20** (1963), 448–464. The 12-dimensional system is described in [Lor62], but full details are not given there, and in fact I have not been able to find the exact relationship between this and the 14-dimensional system. In [Lor62, p. 631], Lorenz points out a similarity between his model and one studied by Kirk Bryan ("A Numerical Investigation of Certain Features of the General Circulation", *Tellus* **11**:2 (1959), 163–174), whose paper suggests that this work also involved Lorenz.

In the abstract of [Lor62], Lorenz described the physical interpretation of this system in the following way:

> A two-layer baroclinic model, in which the flow pattern in each layer is expressed by six terms in a double Fourier series, has been integrated on a small electronic computer. The model contains frictional damping, and thermal forcing which is constant with time but variable with latitude and longitude.

REMARK 1.127. It is perhaps worth mentioning here a similarity to the work of Poincaré, who also used trigonometric series, and began his investigations by studying the behavior of truncations of those series. In Lorenz's work, the effect of the truncation is to only allow circulation at very large scales.

The main conclusion of Lorenz's 1960 presentation was that after experimenting with various values of the parameters in the model, he had succeeded in producing behavior that could not be adequately predicted by linear regression methods beyond very short time scales (one or two "days of weather" in the model).

In fact, Lorenz had observed a much more general phenomenon, and was aware of this fact. Following his talk, Lorenz was asked if he had run simulations with slightly different initial conditions to see how much the resulting forecasts varied. In replying, he said, "We found that this error grew and continued to grow at an exponential rate", and concluded, "at least for this particular set of equations there is a limit to how far you can forecast" [Lor62, p. 635]. He would later describe the discovery this way [Lor93, p. 134]:

> At one point I decided to repeat some of the computations in order to examine what was happening in greater detail. I stopped the computer, typed in a line of numbers that it had printed out a while earlier, and set it running again. I went down the hall for a cup of coffee and returned after about an hour, during which time the computer had simulated about two months of weather. The numbers being printed were nothing like the old ones.

Upon further inspection, Lorenz worked out what had happened: since the computer stored six significant digits but only printed three,[58] he had not restarted it in exactly the same configuration as the previous run, but had introduced a small change in the initial conditions, which was then amplified. As his comments after his 1960 talk made clear, he realized the consequences of this observation for prediction: although dynamic forecasting techniques produced good results for substantially longer than the linear methods he was comparing them to, they still encountered fundamental limitations.

Lorenz decided that to get at the heart of the matter, it would be best to present this new realization using the simplest possible example; after all, a 12-dimensional

[58] As described on p. 423 of Lorenz's 1963 *Trans. New York Acad. Sci.* paper referenced in Footnote 52 above, or p. 55 of "On the prevalence of aperiodicity in simple systems", in Global Analysis, *Lecture Notes in Mathematics* **755**, Springer, 1979.

system is still larger than we can properly visualize. However, he struggled to simplify his system further without losing the sensitive dependence that he sought to illustrate.[59] Then sometime in 1961, Barry Saltzman showed Lorenz a system of 7 equations in whose solutions he had observed some nonperiodic behavior. Lorenz noticed that 4 of the 7 variables quickly became small, and focused on the other 3, leading him to the following system:

Lec 26
Mon, Mar 24

$$\begin{aligned} \dot{x}_1 &= -\sigma x_1 + \sigma x_2, & \sigma &= 10, \\ \dot{x}_2 &= r x_1 - x_2 - x_1 x_3, & r &= 28, \\ \dot{x}_3 &= x_1 x_2 - b x_3, & b &= 8/3. \end{aligned} \tag{1.127}$$

Saltzman's equations modeled "convective fluid motion driven by heating from below, such as might occur locally over warm terrain, instead of the global atmospheric circulation, which is driven mainly by horizontal differences in heating" [Lor93, p. 137]. To put it another way, consider a layer of fluid between two horizontal plates maintained at constant temperatures, with the bottom plate hotter than the top one. If the difference in temperature is small, the fluid will remain motionless, and heat will flow by conduction from the bottom plate to the top one. For a larger temperature difference, one observes convection cells: rotating vortices carrying warm fluid up and cool fluid down.

REMARK 1.128. This situation had been studied in 1916 by Lord Rayleigh,[60] and Saltzman's equations followed the same approach described in Remark 1.127: use Fourier series and remove higher-order terms to obtain a low-dimensional system. Consequently, (1.127) ignores small-scale behavior, and thus does not fully represent the physical situation described. A physical model that is more accurately described by (1.127) is the "chaotic waterwheel" built by Willem Malkus.[61]

The following physical interpretations of the variables are taken essentially verbatim from [Lor63, p. 135]:

- x_1 is proportional to the intensity of the convective motion;
- x_2 is proportional to the temperature difference between the ascending and descending currents (when x_1 and x_2 have the same sign, warm fluid is rising and cold fluid is descending);
- x_3 is proportional to the distortion of the vertical temperature profile from linearity.

The parameter b in (1.127) is related to the shape of the convection cells. The parameter r is proportional to the *Rayleigh number*, which is proportional to the

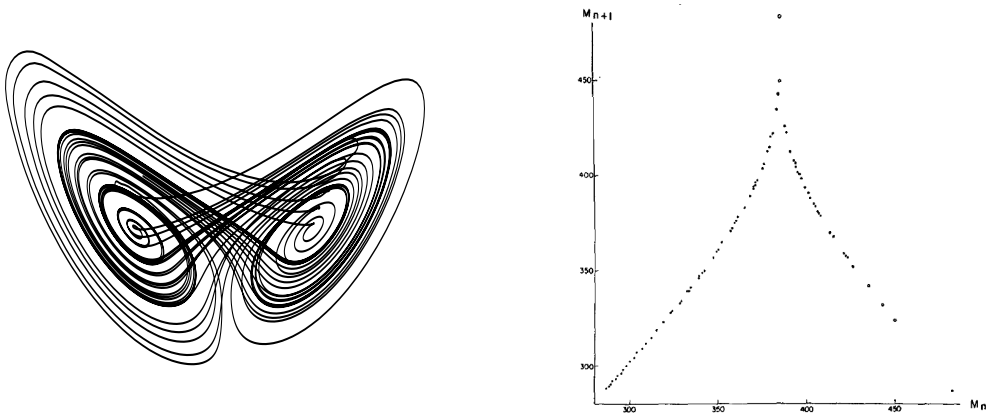
[59] Eventually, in 1983, Lorenz found such a simplification to 3 equations, which is presented on page 102 of "Irregularity: a fundamental property of the atmosphere", *Tellus* 1984, 36A, pp. 98–110. But the famous Lorenz system in (1.127) is the one suggested by Saltzman.

[60] "On convective currents in a horizontal layer of fluid when the higher temperature is on the under side", *Phil. Mag.* **32** (1916), 529–546.

[61] See §9.1 of Steven Strogatz's book "Nonlinear Dynamics and Chaos", or p. 56–57 of [Lorenz 1979]. In [Lor93, p. 143], Lorenz also mentions Louis Howard and Ruby Krishnamurti as having built versions of this waterwheel.

temperature difference between the bottom and top plates. The parameter σ is the *Prandtl number*, which gives the ratio of viscosity to thermal conductivity. As we will see, the parameter values in (1.127) lead to the chaotic behavior that Lorenz observed, but other paramater values can give other qualitative behavior. Indeed, Lorenz remarked [Lor93, p. 137] on his luck that Salzman had chosen a Prandtl number ($\sigma = 10$) closer to that of water than of air (for which $\sigma = 1$ would have been a natural choice), since in the latter case he would not have observed chaotic behavior.

FIGURE 1.54. The Lorenz attractor and an interval map.



The main results in [Lor63] are as follows.

- Numerical approximations to the solutions of (1.127), one of which is shown in Figure 1.54, display the sensitive dependence on initial conditions that Lorenz had observed earlier in the 12-variable atmospheric model, and also display nonperiodic behavior, meaning that neither periodicity nor quasiperiodicity was typically observed.
- These two facts are related: any system with recurrence and nonperiodicity must display sensitive dependence.[62] Indeed, if no recurrent orbit is quasiperiodic to within $\epsilon > 0$, then for every $t \in \mathbb{R}$ there exists $s \in \mathbb{R}$ such that $d(f_s(f_t x), f_s x) > \epsilon$. If x has a recurrent orbit, then $f_t x$ returns arbitrarily close to x, so the inequality guarantees sensitive dependence.
- All solutions of (1.127) appear to converge quite quickly to a certain "infinite complex of surfaces, each extremely close to one or the other of two merging surfaces"; these two merging surfaces can be seen in the first part of Figure 1.54.

[62] This fact also appeared in §V.11 of V.V. Nemytskii and V.V. Stepanov, "Qualitative Theory of Differential Equations", *Princeton University Press*, 1960, although they do not focus on the sensitive dependence property, or even formulate it: they deduce that "Lyapuov stability" (failure of sensitive dependence) implies quasiperiodicity. Similar ideas were explored earlier by Philip Franklin, "Almost periodic recurrent motions", *Math. Z.* **30**, 325–331, 1929, and by A.A. Markov, "Stabilität im Liapounoffschen Sinne und Fastperiodizität", *Math. Z.* **36**, 708–738, 1933.

- A substantial part of the dynamics of the flow can be described in terms of a one-dimensional map from an interval to itself, whose graph has a tent-like shape, as shown in the second part of Figure 1.54. This picture is taken directly from [Lor63], and shows the points with coordinates (M_n, M_{n+1}), where M_n is the nth local maximum of the x_3-coordinate as we move along the orbit.

We will soon describe how Lorenz's experimental descriptions were strengthened to rigorous theorems using *geometric Lorenz models*, but first we make some elementary observations.

▶ EXERCISE 1.45. Prove that there exists $R > 0$ such that if $\|x\| \geq R$ and $\{f_t\}_t$ is the flow induced by (1.127), then $\frac{d}{dt}\|f_t(x) - (0, 0, r + \sigma)\|\big|_{t=0} < 0$. *(Hint: it will be easier to differentiate the square of the norm.)*

Exercise 1.45 shows that if $B \subset \mathbb{R}^3$ is a sufficiently large ball centered at $(0, 0, r+\sigma)$, then B is forward-invariant: $f_t(B) \subset B$ for every $t \geq 0$. In particular, every trajectory that enters B eventually converges to the compact invariant set

$$X := \bigcap_{t \geq 0} f_t(B),$$

which we call the *Lorenz attractor*.

Since the vector field F defining the right-hand side of (1.127) has $\operatorname{div}(F) = -\sigma - 1 - b < 0$, the system is dissipative: given any bounded region $U \subset \mathbb{R}^3$, we have $\operatorname{Leb}(f_t(U)) = e^{-(\sigma+1+b)t} \operatorname{Leb}(U) \to 0$ as $t \to \infty$. Taking $U = B$, we see that the Lorenz attactor has Lebesgue measure 0.

REMARK 1.129. Given the apparently chaotic nature of the Lorenz system, it is natural to ask whether we can study it using ergodic theory. For this we need an invariant measure. With the standard map, it was natural to use Lebesgue measure, but here, things are not so simple, since Lebesgue measure is not invariant. It is then a substantial question to determine which measure we should use instead: this will lead us eventually to the theory of *Sinai–Ruelle–Bowen measures*.

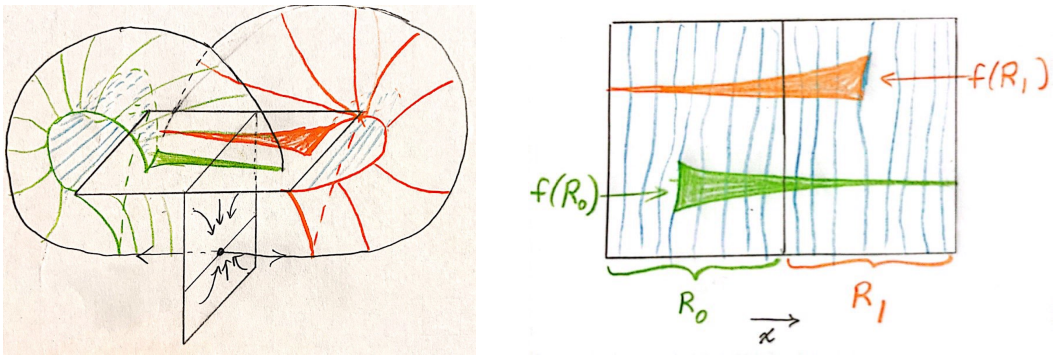
▶ EXERCISE 1.46. Prove that if μ is an invariant Borel probability measure for (1.127), then $\mu(X) = 1$.

▶ EXERCISE 1.47. Writing $F(x) \in \mathbb{R}^3$ for the vector field in (1.127), solve $F(x) = 0$ to find all equilibrium points of the system. Compute DF and find its eigenvalues at these equilibrium points to determine their stability properties.

If you complete Exercise 1.47, you should find that the origin is the only equilibrium point, and that $DF(\mathbf{0})$ has one positive eigenvalue and two negative ones, so that the linearized system has a one-dimensional expanding subspace and a two-dimensional contracting subspace. One can prove a higher-dimensional analogue of the Hadamard–Perron Theorem 1.15 and show that the original, nonlinear system, has the following:

- an invariant curve tangent to the expanding subspace, along which orbits are repelled from $\mathbf{0}$;

FIGURE 1.55. The geometric Lorenz model and its Poincaré return map.



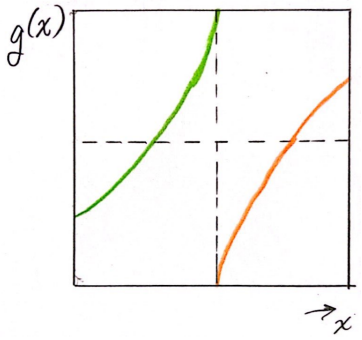
(a) (b)

- an invariant surface tangent to the contracting subspace, along which orbits are attracted towards **0**.

These *unstable and stable manifolds* are shown in Figure 1.55(a), along with an indication of how the numerics suggest the flow behaves away from the origin. The key features are that there is a surface R (in this case it is the horizontal rectangle shown, which lies in the plane $x_3 = r-1$) with the property that the stable manifold of **0** divides R into two pieces, R_0 and R_1, each of which flows along one of the two "circular tunnels" shown in the figure, until it returns to R as a long, narrow region that is roughly triangular, as shown in Figure 1.55(b).

The *geometric Lorenz models* were introduced by V. Afraimovich, V. Bykov, and L. Shil'nikov,[63] and also by J. Guckenheimer and R. Williams.[64] They describe a class of three-dimensional that share the qualitative behavior in the previous paragraph, together with some further details, such as the existence of a foliation of R by curves that are contracted under the Poincaré return map, so that the future of an orbit of the flow is essentially determined by an orbit of the interval map shown in Figure 1.56.

FIGURE 1.56. An interval map.



It is not too difficult to show that the geometric Lorenz models satisfy the various properties observed numerically by Lorenz. However, one

[63] "On the appearance and structure of the Lorenz attractor", *Dokl. Acad. Sci. USSR* **234**, 336–339, 1977.

[64] "Structural stability of Lorenz attractors", *Publ. Math. IHES* **50**, 59–72, 1979, and R.F. Williams, "The structure of the Lorenz attractor", *Publ. Math. IHES* **50**, 73–99, 1979.

must still prove that the Lorenz system itself is indeed a geometric Lorenz model. This was eventually done via a combination of two results: Warwick Tucker used *rigorous* numerical computations combined with a normal form analysis near the origin to prove that the behavior of the Lorenz system itself is "robust",[65] at which point a result of C. Morales, M.J. Pacifico, and E. Pujals[66] shows that it has the necessary properties to be a geometric Lorenz model. A more detailed account of all these ideas is given by Marcelo Viana's review article: "What's new on Lorenz strange attractors", *The Mathematical Intelligencer* **22**(3), 6–19, 2000.[67]

1.13. Markov measures

Lec 27
Wed, Mar 26

It is instructive to compare the Poincaré return map of a geometic Lorenz model in Figure 1.55)(b) with the horseshoe map in Figure 1.40. We see that certain features of the horseshoe appear here as well: there are two regions $R_0, R_1 \subset R$ that the return map f carries diffeomorphically to images $f(R_i)$, preserving horizontal and vertical cones, expanding horizontal distance, and contracting vertical distances. However, there are two important differences:

- the regions R_0 and R_1 *cover* R (except for the single curve separating them), and every orbit that is not on the stable manifold of $\mathbf{0}$ returns to R, unlike the horseshoe where a large set of orbits did not return;
- the images $f(R_i)$ do not stretch all the way across R in the horizontal direction, as they did for the horseshoe.

For now, we focus on this second difference. In §1.8.2, we used the "completely crossing" property of the horseshoe's branches to deduce that there was a topological conjugacy between the dynamics of the horseshoe and the dynamics of the full shift $A^{\mathbb{Z}}$. This no longer holds for the Poincaré return map of the Lorenz attractor: while we can still code orbits using a bi-infinite sequence in $\{0,1\}^{\mathbb{Z}}$, which can visually be interpreted as tracking the sequence of lobes that the orbit winds around, there will be some sequences that do not code any orbit of the Lorenz system. Indeed, since the number of successive windings around a single lobe is bounded, there exists $k \in \mathbb{N}$ such that no coding associated to the attractor includes either of the words 0^k or 1^k. There are other forbidden words as well.

We conclude that the Lorenz flow is not coded by the full shift $\{0,1\}^{\mathbb{Z}}$, but by some *subshift* $X \subset \{0,1\}^{\mathbb{Z}}$. We will return later to the question of describing this subshift precisely, and restrict our attention at the moment to the question of what this means for ergodic theory. For the horseshoe, we obtained a one-parameter

[65] "The Lorenz attractor exists", *C.R. Acad. Sci. Paris* **328**, Série I, Mathématique, 1197–1202, 1999.

[66] "On C^1 robust singular transitive sets for three-dimensional flows", *C.R. Acad. Sci. Paris* **326**, Série I, 81–86, 1998.

[67] A shorter, less technical, overview of this result is given by Ian Stewart, "The Lorenz attractor exists", *Nature* **406**, 948–949, 2000. For an earlier overview of the Lorenz system across a wide range of parameter values, see C. Sparrow, "The Lorenz equations: bifurcations, chaos and strange attractors", volume 41 of *Applied Mathematical Sciences*, Springer Verlag, 1982.

family of invariant measures via the Bernoulli measures on $\{0,1\}^{\mathbb{Z}}$. However, this approach does not produce invariant measures on the Lorenz attractor:

▶ EXERCISE 1.48. Prove that if A is a finite set, $X \subsetneq A^{\mathbb{Z}}$ is closed and σ-invariant, and μ is a Bernoulli measure for some probability vector p whose entries are all positive, then $\mu(X) = 0$.

We will not develop a full theory of invariant measures for the Lorenz attractor in this section, but will take the first step by introducing the family of *Markov measures* on a shift space $A^{\mathbb{Z}}$, which bear similarities to Bernoulli measures but have greater flexibility.[68]

DEFINITION 1.130. Given $d \in \mathbb{N}$, a $d \times d$ matrix P is *stochastic* if its rows are probability vectors.

A $d \times d$ stochastic matrix P defines the following *Markov process*: we consider a system that can be in one of d states, which are labeled with elements from $A := \{1, \dots, d\}$. If the system is currently in state $a \in A$, then for each $b \in A$, P_{ab} gives the probability that it will be in state b at the next time step. To put it another way, the state of the system at each time step is chosen according to the probability distribution given by the ath row of the matrix P, where a is the state at the previous time step.

Vanishing entries of P correspond to forbidden transitions: if $P_{ab} = 0$, then the Markov process has 0 probability of being in state b immediately after being in state a.

REMARK 1.131. One could also consider Markov processes in which the next probability distribution depends not only on the most recent state a, but on the most recent m states – this would give a Markov chain of *order* m, or with *memory* m. The important thing is that m is finite. We will restrict our attention to the case $m = 1$. Formally, the case $m > 1$ can be reduced to this one by replacing A with A^m as the state space.

We want to produce an invariant probability measure μ on $A^{\mathbb{N}}$ that is consistent with the Markov process defined by P in the following sense: given any $a, b \in A$ and any word $w \in A^*$ that ends with the symbol a, we have

$$\mu[wb] = \mu[w]P_{ab}. \tag{1.128}$$

Let us introduce the following notation: given $k \in \mathbb{N}$ and $w \in A^k$, write

$$\mathbf{P}(w) = P_{w_1 w_2} P_{w_2 w_3} \cdots P_{w_{k-1} w_k} \tag{1.129}$$

for the probability that we witness all of the transitions given in the word w, conditioned on starting with the symbol w_1. Then iterating (1.128), we see that any measure μ consistent with the Markov process must satisfy

$$\mu[w] = \pi_{w_1} \mathbf{P}(w) = \pi_{w_1} P_{w_1 w_2} \cdots P_{w_{k-1} w_k} \tag{1.130}$$

[68] Later, we will need to go beyond the space of Markov measures and consider *Gibbs measures*. But for now, Markov measures will illustrate some of the main ideas.

for the probability vector π defined by $\pi_a := \mu[a]$, which gives the initial probability distribution on A. Conversely, every probability vector induces a measure μ:

PROPOSITION 1.132. *Let P be a $d \times d$ stochastic matrix, and $\pi \in \mathbb{R}^d$ a probability (row) vector. As in Exercise 1.29, let $\mathcal{C}$ denote the set of all cylinders in $A^{\mathbb{N}}$, and define a function $\ell \colon \mathcal{C} \to [0,1]$ by $\ell[w] = \pi_{w_1}\mathbf{P}(w)$. Then ℓ is countably additive, and thus by the Carathéodory Extension Theorem 1.73, there exists a unique Borel probability measure μ on $A^{\mathbb{N}}$ such that $\mu[w] = \ell[w]$ for every $w \in A^*$.*

PROOF. Mimic Exercise 1.30. □

DEFINITION 1.133. The measure μ in Proposition 1.132 is the *Markov measure* associated to P and π.

We see that Markov measures incorporate a limited amount of time-dependence: the future is allowed to depend on the present, but not on the past. Thus they represent a step towards the kinds of measures we expect to need when we study deterministic dynamical systems that exhibit stochastic behavior.

PROPOSITION 1.134. *The Markov measure μ from Proposition 1.132 is shift-invariant if and only if $\pi P = \pi$.*

PROOF. Given $w \in A^*$, we have $\sigma^{-1}[w] = \bigsqcup_{a \in A}[aw]$, so (1.130) gives

$$\mu(\sigma^{-1}[w]) = \sum_{a \in A} \mu[aw] = \sum_{a \in A} \pi_a P_{aw_1} \mathbf{P}(w) = (\pi P)_{w_1} \mathbf{P}(w). \tag{1.131}$$

If $\pi P = \pi$, then this last expression is equal to $\mu[w]$, so $\sigma_*\mu$ agrees with μ on cylinders; by Proposition 1.93, this implies that μ is shift-invariant. Conversely, if μ is shift-invariant, then for each $b \in A$ we can take $w = b$ in (1.131) and obtain

$$\pi_b = \mu[w] = \mu(\sigma^{-1}[w]) = (\pi P)_b,$$

so $\pi = \pi P$. □

Now suppose we are given a stochastic matrix P, and want to study the r5statistical behavior of sequences in $A^{\mathbb{N}}$ with respect to the transition probabilities given by P. This requires us to find an eigenvector $\pi = P\pi$ so that (P, π) defines a shift-invariant Markov measure, which we can study using ergodic theory. With this in mind, the following questions present themselves.

- Does every stochastic matrix P admit a probability vector π with $\pi P = \pi$?
- If such a probability vector exists, is it unique?
- Are these Markov measures ergodic?

In $\mathbb{R}^d$, denote the positive orthant by

$$\Omega^d := \{x \in \mathbb{R}^d : x_j \geq 0 \text{ for all } 1 \leq j \leq d\}, \tag{1.132}$$

and the space of probability vectors by

$$\Delta^{d-1} := \{x \in \Omega^d : x_1 + \cdots + x_d = 1\}. \tag{1.133}$$

This is a $(d-1)$-dimensional simplex. When d is fixed, we will sometimes suppress it from the notation, simply writing Ω and Δ.

The question of finding eigenvectors in Ω^d is one that we already addressed in §1.4 in the case $d = 2$. In that setting, we observed that existence follows quickly from a topological fixed point theorem, while stronger results require us to find a metric in which the linear map acts as a contraction. The story here is similar. We start by observing that P maps Δ^{d-1} to itself by right multiplication.

LEMMA 1.135. *If P is a $d \times d$ stochastic matrix and $\pi \in \Delta^{d-1}$, then $\pi P \in \Delta^{d-1}$.*

PROOF. For each $j \in \{1, \dots, d\}$ we have $(\pi P)_j = \sum_{i=1}^d \pi_i P_{ij} \geq 0$, so $\pi P \in \Omega^d$. Moreover,

$$\sum_{j=1}^d (\pi P)_j = \sum_{j=1}^d \sum_{i=1}^d \pi_i P_{ij} = \sum_{i=1}^d \pi_i \sum_{j=1}^d P_{ij} = \sum_{i=1}^d \pi_i = 1,$$

using the fact that rows of P are probability vectors. Thus $\pi P \in \Delta^{d-1}$. □

Since $\Delta = \Delta^{d-1}$ is homeomorphic to a disc, Brouwer's fixed point theorem implies that the continuous map $P\colon \Delta \to \Delta$ has a fixed point π. We can also avoid appealing to Brouwer by taking an arbitrary $q \in \Delta$, considering the sequence $q^n := \frac{1}{n} \sum_{k=0}^{n-1} qP^k \in \Delta$, and choosing any limit point $\pi = \lim_{j\to\infty} q^{n_j}$, which exists by compactness of Δ. Then

$$\pi P - \pi = \lim_{j\to\infty} \frac{1}{n_j} \sum_{k=0}^{n_j-1} (qP^{k+1} - qP^k) = \lim_{j\to\infty} \frac{1}{n_j}(qP^{n_j} - q) = 0,$$

so π is a fixed point of P.

We will soon see how to determine when there is a metric on Δ in which P acts as a contraction, so that π is unique and $qP^n \to \pi$ exponentially fast. First we examine the question of when the invariant Markov measure μ associated to (P, π) is ergodic.

A natural approach is to mimic the argument from Example 1.120 that Bernoulli measures are ergodic as a consequence of Proposition 1.119 by verifying (1.120): that is, given cylinders $[v], [w] \subset A^{\mathbb{N}}$ and $c > 0$, we want to understand when $\mu(\sigma^{-n}[v] \cap [w]) \geq c\mu[v] \cdot \mu[w]$.

To this end, first observe that $\mathbf{P}$ from (1.129) has the following multiplicativity property: given any $u, v \in A^*$ and $a \in A$, we have

$$\mathbf{P}(uav) = \mathbf{P}(ua)\mathbf{P}(av). \tag{1.134}$$

Moreover, by properties of matrix multiplication, given $a, b \in A$ and $k \geq 0$, we have

$$\sum_{u\in A^k} \mathbf{P}(aub) = \sum_{u\in A^k} P_{au_1} P_{u_1u_2} \cdots P_{u_{k-1}u_k} P_{u_kb} = (P^{k+1})_{ab}. \tag{1.135}$$

Thus we see that $(P^k)_{ab}$ represents the probability that in k steps we go from a to b, conditioned on starting with a, when we are not concerned with what happens in between.

Now given any $v, w \in A^*$ and $n \geq |w|$, we can use the fact that $\sigma^{-n}[v] \cap [w] = \bigsqcup_{u \in A^{n-|w|}} [wuv]$ to obtain the following: writing $a = w_{|w|}$ for the last symbol of w, $b = v_1$ for the first symbol of v, and $k = n - |w|$ for the number of symbols that appear between w and v when we have a sequence in $\sigma^{-n}[v] \cap [w]$, we get

$$\begin{aligned}\mu(\sigma^{-n}[v] \cap [w]) &= \sum_{u \in A^k} \mu[wuv] = \sum_{u \in A^k} \pi_{w_1} \mathbf{P}(wuv) \\ &= \sum_{u \in A^k} \pi_{w_1} \mathbf{P}(w)\mathbf{P}(aub)\mathbf{P}(v) = \pi_{w_1} \mathbf{P}(w)(P^{k+1})_{ab}\mathbf{P}(v).\end{aligned}$$

We record this in the following form: given a word w ending in $a \in A$, and a word v beginning with $b \in A$, for every $k \geq 0$ we have

$$\mu(\sigma^{-(k+|w|)}[v] \cap [w]) = \mu[w] \cdot (P^{k+1})_{ab} \cdot \mathbf{P}(v). \tag{1.136}$$

Lec 28
Fri, Mar 28

Considering the case when $w = a$ and $v = b$, we see that if μ is ergodic, then given any $a, b \in A$ with $\pi_a > 0$ and $\pi_b > 0$, we must have $(P^n)_{a,b} > 0$ for some $n \in \mathbb{N}$. Motivated by this, we make the following definition.

DEFINITION 1.136. A stochastic matrix P is *primitive* if there exists $n \in \mathbb{N}$ such that every entry of P^n is positive: $(P^n)_{a,b} > 0$ for every $a, b \in A$.

REMARK 1.137. Later, we will examine a sense in which every Markov process can be expressed in terms of primitive processes, and we will see that the following ergodicity result can be extended to a broader class of stochastic matrices.

PROPOSITION 1.138. *If P is a primitive stochastic matrix and $\pi = \pi P$ is a probability eigenvector, then the Markov measure $\mu = \mu_{P,\pi}$ is ergodic.*

PROOF. By Proposition 1.119, it suffices to prove that there exists $c > 0$ such that for every $v, w \in A^*$, there exists $N = N_{v,w}$ such that for every $n \geq N_{v,w}$, we have

$$\mu(\sigma^{-n}[v] \cap [w]) \geq c\mu[v] \cdot \mu[w]. \tag{1.137}$$

We will use primitivity of P together with (1.136). With $n \in \mathbb{N}$ as in the definition of primitivity, we see that $c := \min\{(P^n)_{ab} : a, b \in A\} > 0$. Thus for every $\ell \in \mathbb{N}$ and every $a, b \in A$, we have

$$(P^{n+\ell})_{ab} = \sum_{i \in A} (P^\ell)_{ai}(P^n)_{ib} \geq \sum_{i \in A} (P^\ell)_{ai} c = c,$$

where the last equality uses the fact that P^ℓ is a stochastic matrix. Now for every $k \geq n$, (1.136) gives

$$\mu(\sigma^{-(k+|w|)}[v] \cap [w]) \geq \mu[w] \cdot c \cdot \mathbf{P}(v) \geq c\mu[w]\mu[v],$$

where the last inequality uses the fact that $\mu[v] = \pi_{v_1}\mathbf{P}(v) \leq \mathbf{P}(v)$. This proves (1.137), and Proposition 1.119 shows that μ is ergodic. □

Our experiences in §1.4 suggest that we might in fact expect the eigenvector π to be unique, and to witness exponential convergence to π under the iterates P^k. This will turn out to be true provided P is primitive, and once we prove it, the arguments in the proof of Proposition 1.138 will show that (recalling that $b = v_1$)

$$\mu(\sigma^{-(k+|w|)}[v] \cap [w]) = \mu[w] \cdot (P^{k+1})_{ab} \cdot \mathbf{P}(v) \to \mu[w]\pi_b \mathbf{P}(v) = \mu[w] \cdot \mu[v],$$

where the convergence is exponentially fast: this *exponential mixing* property gives a strong sense in which Markov chains exhibit stochastic behavior and asymptotic independence of the future from the past.

To prove uniqueness and exponential convergence via the Banach fixed point theorem, we must find a metric on Δ in which P acts as a contraction. This can be done using ideas from projective and hyperbolic geometry. In order to unify this with the continuation of our discussion from §1.4.1, we consider a more general setting where L is any $d \times d$ matrix with nonnegative entries and the property that for some N, we have $(L^N)_{ab} > 0$ for all $a, b \in A$. Then L maps the positive orthant $\Omega = \Omega^d$ into itself, and we once again consider the projectivization $\mathbb{P}\Omega$, as we did in §1.4.

Now we look for a metric on $\mathbb{P}\Omega$ in which L becomes a contraction. For the time being, we restrict our attention to the case $d = 2$, and recognizing the value of circuitous exploration, we start by describing a metric that does not work.

Given $v = \left(\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}\right)$, the line $[v] \in \mathbb{P}\Omega$ has slope $s([v]) := \frac{v_2}{v_1} \in [0, \infty]$, and we could equip $\mathbb{P}\Omega$ with a metric[69] coming from this bijection $s \colon \mathbb{P}\Omega \to [0, \infty]$ by writing $d([v], [w]) = |s(v) - s(w)|$.

To see whether $L = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ is a contraction in this metric, start by writing L in "slope coordinates" as

$$f := s \circ L \circ s^{-1} \colon y \mapsto s \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 \\ y \end{pmatrix} = s \begin{pmatrix} a + by \\ c + dy \end{pmatrix} = \frac{c + dy}{a + by}. \tag{1.138}$$

Since f is differentiable, given $s_1, s_2 \in [0, \infty]$ we have

$$|f(s_2) - f(s_1)| = \left| \int_{s_1}^{s_2} f'(y)\, dy \right| \leq (\sup_y |f'(y)|)|s_2 - s_1|. \tag{1.139}$$

From (1.138) we get

$$f'(y) = \frac{(a + by)d - (c + dy)b}{(a + by)^2} = \frac{ad - bc}{(a + by)^2}, \tag{1.140}$$

so that in the specific case $L = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ we have $|f'(y)| = \frac{1}{(2+y)^2} \leq \frac{1}{4}$ for all $y \in [0, \infty]$, which together with (1.139) implies that $d(L[v], L[w]) \leq \frac{1}{4} d([v], [w])$ for all $[v], [w] \in \mathbb{P}\Omega$. This allows us to apply the Banach fixed point theorem and deduce uniqueness and stability of the eigenspace in the first quadrant. However, this does not work as broadly as we might hope.

[69] Note that we allow our metric to take the value ∞.

▶ EXERCISE 1.49. Prove that the matrix $L = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 3 \end{smallmatrix}\right)$ is not a contraction in the "slope metric" described above, but that replacing "slope" $\frac{v_2}{v_1}$ with "reciprocal slope" $\frac{v_1}{v_2}$ would yield a metric in which L is a contraction. Then prove that $L = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$ is not a contraction in either of these metrics.

Observe that the parametrization $s\colon \mathbb{P}\Omega \to [0,\infty]$ represents $[v] \in \mathbb{P}\Omega$ by the y-coordinate at which the line $[v]$ intersects the vertical line $x = 1$, and that the reciprocal slope parametrization suggested in Exercise 1.49 follows a similar procedure with the horizontal line $y = 1$. Some examination of the computations in Exercise 1.49 should reveal that the problems (failure of contraction) arise where these lines intersect the coordinate axes orthogonally at the boundary of Ω. This suggests that a successful metric ought to "expand" the Euclidean metric near the boundary of Ω.

If you are familiar with hyperbolic geometry, recall that the hyperbolic metric on the upper half-plane does just this: it takes the Euclidean metric and divides it by the distance from the boundary, obtaining $ds^2 = (dx^2 + dy^2)/y^2$. In particular, given two points on the same vertical line in the hyperbolic plane whose vertical coordinates are y_1 and y_2, the distance between them is $|\log(y_1/y_2)|$. Motivated by this, we consider for each $v \in \Omega$ the quantity

$$t([v]) := \log s([v]) = \log v_2 - \log v_1, \tag{1.141}$$

and then define the *Hilbert metric*[70] Θ on $\mathbb{P}\Omega$ by

$$\Theta(v,w) = |t(v) - t(w)| = \left|\log \frac{v_2 w_1}{v_1 w_2}\right|. \tag{1.142}$$

Now let $L = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ be an invertible matrix with $a, b, c, d \geq 0$. (If L is not invertible, then it collapses $\mathbb{P}\Omega$ to a single point, and our work is done.) If we write $g = t \circ L \circ t^{-1}\colon [-\infty,\infty] \to [-\infty,\infty]$ for the map $L\colon \mathbb{P}\Omega^2 \to \mathbb{P}\Omega^2$ written in these "logarithmic slope" coordinates, then we see from (1.138) that

$$g(x) = \log \frac{c + de^x}{a + be^x} = \log(c + de^x) - \log(a + be^x),$$

and differentiating gives

$$g'(x) = \frac{de^x}{c + de^x} - \frac{be^x}{a + be^x} = \frac{(ad - bc)e^x}{(c + de^x)(a + be^x)} = \frac{ad - bc}{ace^{-x} + (ad + bc) + bde^x}.$$

[70] As we will later see, this is a specific case of a metric that can be placed on any bounded convex set in Euclidean space by taking the line connecting two points x, y, writing a, b for the points where this line intersects the boundary of the set, and then defining the distance to be the logarithm of the cross-ratio of the four points a, x, y, b. David Hilbert used this in "Neue Begründung der Bolyai–Lobatschefskychen Geometrie", *Math. Ann.* **57**, 137–150, 1903. The use of the cross-ratio to define a metric in projective geometry goes back to Arthur Cayley, 1859, "Sixth memoir on quantics", and was used for the hyperbolic plane by Felix Klein, 1871, "Ueber die sogenannte Nich-Euclidische Geometrie".

In light of (1.139), we want an upper bound for

$$\gamma := \sup_{x\in\mathbb{R}} |g'(x)|. \tag{1.143}$$

This requires a lower bound for the denominator

$$h(x) := ace^{-x} + (ad+bc) + bde^{x}.$$

First we deal with the degenerate case where one or more of a,b,c,d vanishes. If $ad=0$, then $h(x)\geq bc>0$ (since L is invertible), so $|g'(x)|\leq \frac{bc}{bc}=1$. Similarly, if $bc=0$, then $h(x)\geq ad>0$ and $|g'(x)|\leq \frac{ad}{ad}=1$. Thus in these cases we cannot conclude that L is a contraction, although at least it does not expand any distances.

Now suppose $a,b,c,d>0$. Then h is strictly convex and goes to ∞ as $x\to\pm\infty$, so it achieves its minimum at a unique critical point obtained by solving

$$0=h'(x)=bde^{x}-ace^{-x} \quad\Leftrightarrow\quad e^{x}=\sqrt{\frac{ac}{bd}} \quad\Rightarrow\quad ace^{-x}=bde^{x}=\sqrt{abcd}.$$

We conclude that

$$\inf_{x\in\mathbb{R}} h(x) = ad+2\sqrt{abcd}+bc = (\sqrt{ad}+\sqrt{bc})^2,$$

and writing $|ad-bc| = |\sqrt{ad}-\sqrt{bc}|(\sqrt{ad}+\sqrt{bc})$, we see that the supremum in (1.143) is given by

$$\gamma = \frac{|ad-bc|}{(\sqrt{ad}+\sqrt{bc})^2} = \frac{|\sqrt{ad}-\sqrt{bc}|}{\sqrt{ad}+\sqrt{bc}}. \tag{1.144}$$

This is enough to show that $\gamma<1$ and thus L is a contraction with respect to Θ. It is useful, though, to record this in a slightly different form. Observe that γ is determined by the single quantity $\kappa := \frac{ad}{bc}$, as can be seen by dividing its numerator and denominator by $\sqrt{bc}$ to obtain

$$\gamma = \frac{|\sqrt{\kappa}-1|}{\sqrt{\kappa}+1}. \tag{1.145}$$

Since $\binom{a}{b} = L\binom{1}{0}$ and $\binom{c}{d} = L\binom{0}{1}$, we have

$$|\log\kappa| = \left|\log\frac{ad}{bc}\right| = \Theta\left(L\tbinom{1}{0}, L\tbinom{0}{1}\right) = \sup_{v,w\in\Omega}\Theta(Lv,Lw) = \operatorname{diam}_\Theta(L\Omega).$$

Suppose $\kappa\geq 1$; the case $\kappa<1$ is similar. Then writing $D := \operatorname{diam}_\Theta(L\Omega)$, we have $\kappa = e^D$, and (1.145) gives

$$\gamma = \frac{e^{D/2}-1}{e^{D/2}+1} = \frac{e^{D/4}-e^{-D/4}}{e^{D/4}+e^{D/4}} = \tanh(D/4).$$

We record the conclusions of the above discussion in the following form.

Lec 29
Mon, Mar 31

THEOREM 1.139. *Equip the first quadrant $\Omega\subset\mathbb{R}^2$ with the Hilbert metric Θ defined in* (1.142). *Let $L=\left(\begin{smallmatrix}a&b\\c&d\end{smallmatrix}\right)$ be invertible and let $D=|\log\frac{ad}{bc}| = \operatorname{diam}_\Theta(L\Omega)$. Then $\Theta(Lv,Lw)\leq\tanh(\frac{D}{4})\Theta(v,w)$ for all $v,w\in\Omega$. In particular, if $D<\infty$, then L is a contraction.*

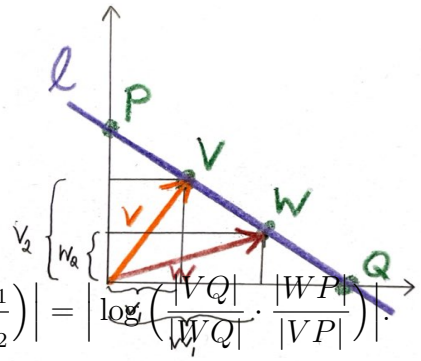
Stripped of the notation, the key observation here is the following, which was first made by Garrett Birkhoff:[71]

In the Hilbert metric, boundedness implies contraction.

We can extend Theorem 1.139 to higher dimensions by equipping the positive orthant $\Omega = \Omega^d \subset \mathbb{R}^d$ with a suitable *Hilbert projective metric*.[72]

To motivate the definition of this metric, let us give another geometric interpretation of the metric Θ on $\mathbb{P}\Omega^2$ from (1.142). Given $v, w \in \Omega^2 \setminus 0$ that are not scalar multiples of each other, let V, W be the points in the plane that lie at the tips of the vectors v, w, and let $\ell \subset \mathbb{R}^2$ be the line passing through V and W. Let P, Q be the points where ℓ intersects the coordinate axes.[73] Then as Figure 1.57 shows, we have

FIGURE 1.57. Θ via cross-ratio.



$$\Theta(v,w) = \left|\log \frac{v_2 w_1}{v_1 w_2}\right| = \left|\log\left(\frac{v_2}{w_2}\cdot\frac{w_1}{w_2}\right)\right| = \left|\log\left(\frac{|VQ|}{|WQ|}\cdot\frac{|WP|}{|VP|}\right)\right|. \tag{1.146}$$

The quantity in brackets is well-known in projective geometry: it is the *cross-ratio* of the four collinear points P, V, W, Q.

Now we can define a projective metric Θ on the positive orthant $\Omega = \Omega^d \subset \mathbb{R}^d$: given $v, w \in \Omega \setminus 0$, let V, W be the points where the lines $\mathbb{R}v$ and $\mathbb{R}w$ intersect the simplex $\Delta = \Delta^{d-1} \subset \Omega$, and let P, Q be the points where the line ℓ through V and W intersects the boundary $\partial\Omega$. Then $\Theta(v, w)$ is defined by the last expression in (1.146) as the absolute value of the logarithm of the cross-ratio of the points P, V, W, Q. (We allow $\Theta = \infty$.)

►► EXERCISE 1.50. Prove that Θ defines a complete metric on $\mathbb{P}\Omega$.

To compute Θ in practice, it is often useful to use the following characterization. Given $v, w \in \mathbb{R}^2$, write $v \preceq w$ if $v_i \le w_i$ for all i; equivalently, if $w - v \in \Omega^2$. This

[71] This is not the same Birkhoff we encountered twice earlier, in §1.9 and 1.11: George David Birkhoff (1884–1944), whose work we described there, was the father of Garrett Birkhoff (1911–1996), who is responsible for the ideas here, which can be found in: Garrett Birkhoff, "Extensions of Jentzsch's Theorem", *Trans. Amer. Math. Soc.* **85**:1, 219–227, 1957.

[72] Everything described here works for a broader class of closed convex cones Ω, which can lie in a finite- or infinite-dimensional vector space. We will describe this later when we discuss the Ruelle–Perron–Frobenius Theorem, but for now we stick to the simplest setting for concreteness.

[73] Figure 1.57 shows P and Q as lying on the positive coordinate axes, which form $\partial\Omega$. Depending on the relative lengths of v, w, one of these points could lie on a negative axis, or at infinity. In the latter case, the corresponding ratio in (1.146) should be treated as 1.

gives a partial order on $\mathbb{R}^2$. Now referring to Figure 1.57, observe that the ratio $\alpha := \frac{|WQ|}{|VQ|} = \frac{w_2}{v_2}$ has the following property:

- for every $t \le \alpha$, we have $w - tv \in \Omega$ and $tv \preceq w$;
- for every $t > \alpha$, we have $w - tv \notin \Omega$ and $tv \not\preceq w$.

Thus we have

$$\alpha = \sup\{t \ge 0 : w - tv \in \Omega\} = \sup\{t > 0 : tv \preceq w\}. \tag{1.147}$$

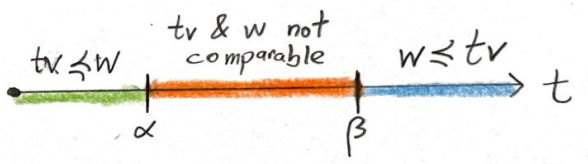
Similarly, the ratio $\beta := \frac{|WP|}{|VP|} = \frac{w_1}{v_2}$ can be given as

$$\beta = \inf\{t > 0 : tv - w \in \Omega\} = \inf\{t > 0 : w \preceq tv\}. \tag{1.148}$$

One way to interpret this is as follows: for sufficiently small t, we have $tv \preceq w$, and for sufficiently large t, we have $w \preceq tv$ (unless v lies on $\partial\Omega$), but there is some interval of t for which w and tv are not comparable in the partial order $\preceq$. The quantities α and β are the endpoints of this interval, and referring to (1.146), we see that

$$\Theta(v, w) = \Big| \log \frac{\beta(v, w)}{\alpha(v, w)} \Big|. \tag{1.149}$$

FIGURE 1.58. Defining Θ via a partial order.



▶▶ EXERCISE 1.51. Let L be a $d \times d$ matrix with positive entries.

(1) Given $v, w \in \Omega^d$, compute $\Theta(v, w)$ and $\Theta(Lv, Lw)$ in terms of the entries of v, w, and L.
(2) Prove that $D := \operatorname{diam}_\Theta(L\Omega^d) < \infty$, and give an expression for D in terms of the entries of L.

Now Theorem 1.139 has the following consequence.

THEOREM 1.140 (Birkhoff contraction theorem for positive matrices). *Let $L\colon \mathbb{R}^d \to \mathbb{R}^d$ be a linear map taking Ω^d into itself, and let $D = \operatorname{diam}_\Theta(L\Omega^d)$. Then $\Theta(Lv, Lw) \le \tanh(\frac{D}{4})\Theta(v, w)$ for all $v, w \in \Omega \setminus 0$. In particular, if $D < \infty$, then L is a contraction in the Hilbert metric.*

PROOF. When $d = 2$, this is just a reformulation of Theorem 1.139. We show that the case $d > 2$ can be reduced to this one.

Given $v, w \in \Omega^d \setminus 0$ that are not scalar multiples of each other, let $X_1 = \operatorname{span}(v, w) \subset \mathbb{R}^d$ be the two-dimensional subspace they span, and similarly, let

$X_2 = \operatorname{span}(Lv, Lw)$. For $i = 1, 2$, let $T_i \colon X_i \to \mathbb{R}^2$ be a linear isomorphism that carries $X_i \cap \Omega^d$ onto Ω^2. We see from (1.147) and (1.148) that for any $x, y \in X_i$, we have

$$\alpha(T_i x, T_i y) = \alpha(x, y) \quad \text{and} \quad \beta(T_i x, T_i y) = \beta(x, y),$$

from which we deduce that

$$\Theta_{\Omega^2}(T_i x, T_i y) = \Theta_{\Omega^d}(x, y).$$

Then $\tilde{L} = T_2 \circ L \circ T_1^{-1} \colon \mathbb{R}^2 \to \mathbb{R}^2$ is a linear map for which $\operatorname{diam}_{\Omega^2}(\tilde{L}\Omega^2) = \operatorname{diam}_{\Omega^d}(L(X_1 \cap \Omega^d)) \le D$, so the case $d = 2$ implies the result. □

Now we are in a position to answer our earlier questions about uniqueness of eigenvectors for primitive stochastic matrices, via the following more general result.

THEOREM 1.141 (Perron–Frobenius Theorem (primitive case)). *Let L be a $d \times d$ matrix with nonnegative entries, and suppose there exists $N \in \mathbb{N}$ such that all entries of L^N are positive. Then the following are true.*

(1) There exist $v, w \in \Omega^d$ and $\lambda > 0$ such that $Lv = \lambda v$ and $w^T L = \lambda w^T$.
(2) All entries of v, w are positive, and up to scalar multiples, v, w are the only right and left eigenvectors in Ω^d.
(3) The eigenvalue λ is simple (one-dimensional eigenspace), and is the largest eigenvalue: every other eigenvalue ν has $|\nu| < \lambda$.
(4) Given any $u \in \mathbb{R}^d$, the sequence $L^n u$ converges to the subspace $\mathbb{R}v$ exponentially fast, in the following sense: defining a projection $P \colon \mathbb{R}^d \to \mathbb{R}v$ by $Pu = \frac{\langle w,u\rangle}{\langle w,v\rangle} v$, there exist $C > 0$ and $\chi \in (0, 1)$ such that for every $u \in \mathbb{R}^d$ and every $n \in \mathbb{N}$, we have

$$\|\lambda^{-n} L^n u - Pu\| \le C\chi^n \|u - Pu\|. \tag{1.150}$$

PROOF. By Exercise 1.51, we can apply Theorem 1.140 to the positive matrix L^N: writing $D = \operatorname{diam}_\Theta(L^N \Omega^d) < \infty$ and $\gamma = \tanh(\frac{D}{4}) \in (0, 1)$, we have

$$\Theta(L^N v, L^N w) \le \gamma \Theta(v, w) \quad \text{for all } v, w \in \Omega \setminus 0. \tag{1.151}$$

We will deduce the conclusions of the theorem from (1.151). First observe that given any $v, w \in \Omega^d$ and any $n \ge 0$, we can write $n = kN + j$ for some $k \ge 0$ and $0 \le j < N$, so

$$\Theta(L^n v, L^n w) = \Theta((L^N)^k (L^j v), (L^N)^k (L^j w)) \le \gamma^k \Theta(L^j v, L^j w) \le \gamma^k \Theta(v, w).$$

Writing $\chi = \gamma^{1/N} \in (0, 1)$, observe that $k \ge \frac{n}{N} - 1$, so $\gamma^k \le \gamma^{\frac{n}{N} - 1} = \gamma^{-1} \chi^n$, and we have

$$\Theta(L^n v, L^n w) \le \gamma^{-1} \chi^n \Theta(v, w). \tag{1.152}$$

Since $(\mathbb{P}\Omega, \Theta)$ is a complete metric space (Exercise 1.50), we conclude that L has a unique fixed point $[v] \in \mathbb{P}\Omega$. This proves the parts of Statement 1 and 2 relating to the right eigenvector v. The statements about the left eigenvector w follow by working with the transpose of L.

To prove Statements 3 and 4, we first point out that $\langle w, v\rangle > 0$, so

$$\mathbb{R}^d = \mathbb{R}v \oplus w^\perp. \tag{1.153}$$

Moreover, given $u \in w^\perp$, we have

$$\langle w, Lu\rangle = w^T Lu = (w^T L)u = (\lambda w^T)u = 0,$$

so the decomposition in (1.153) is L-invariant. Let $L_0 = \lambda^{-1}L$ and observe that $L_0 v = v$, so the affine subspace $Y := v + (w^\perp)$ is L_0-invariant.

Now both Statements 3 and 4 will follow from the following claim: there exists $C > 0$ such that with χ as in (1.152), we have

$$\|L_0^n z\| \le C\chi^n \|z\| \text{ for all } z \in w^\perp \text{ and } n \ge 0. \tag{1.154}$$

Indeed, once we prove this, then given any $u \in \mathbb{R}^d$ we can write $u = av + z$ for some $a \in \mathbb{R}$ and $z \in w^\perp$, and observe that

$$Pu = \frac{\langle w, av + z\rangle}{\langle w, v\rangle} v = \frac{a\langle w, v\rangle + \langle w, z\rangle}{\langle w, v\rangle} v = av,$$

so $z = u - av = u - Pu$, and since $\lambda^{-n}L^n(Pu) = Pu$, we have

$$\|\lambda^{-n}L^n u - Pu\| = \|\lambda^{-n}L^n z\| \le C\chi^n\|z\| = C\chi^n\|u - Pu\|,$$

where the inequality uses (1.154). This proves Statements 3 and 4, so it remains to prove (1.154).

We do this using (1.152) and the following observation, which is a consequence of the fact that $\Theta|_Y$ is obtained by scaling the Euclidean metric by a factor that is bounded near v.

LEMMA 1.142. *There exists $\delta > 0$, depending only on v, w, such that:*

- *if $z \in w^\perp$ satisfies $\|z\| \le \delta$, then $\Theta(v, v + z) \le 2\|z\|$; and*
- *if $z \in w^\perp$ satisfies $\Theta(v, v + z) \le 2\delta$, then $\|z\| \le 2\Theta(v, v + z)$.*

PROOF. Exercise. □

Now given any $z \in w^\perp$, let $c = \delta/\|z\|$, so $\|cz\| = \delta$. Then Lemma 1.142 gives $\Theta(v, v + cz) \le 2\|cz\| = 2\delta$. Observe that $L_0^n v = v$ and

$$L_0^n(v + cz) = L_0^n v + L_0^n(cz) = v + c(L_0^n z) \text{ for every } n \ge 0.$$

Using this together with (1.152) gives

$$\Theta(v, v + c(L_0^n z)) \le 2\|cz\|\gamma^{-1}\chi^n.$$

Moreover, $\Theta(v, v + c(L_0^n z)) \le \Theta(v, v + cz) \le 2\delta$, so Lemma 1.142 gives

$$\|c(L_0^n z)\| \le 4\|cz\|\gamma^{-1}\chi^n.$$

Dividing both sides by c gives (1.154) with $C = 4\gamma^{-1}$. This completes the proof of Theorem 1.141. □

Lec 30
Wed, Apr 2

REMARK 1.143. A common approach to proving the Perron–Frobenius Theorem is to deduce Statement 4 as a consequence of Statement 3 using facts about eigenvalues and eigenvectors of matrices. The approach here, which goes the other way around, offers several benefits.

- It can easily be adapted to deal with products of sequences of positive matrices, not just powers of a single matrix.
- As you saw in Exercise 1.51, the approach via the Hilbert projective metric and Birkhoff's contraction theorem provides a concrete way to determine the contraction ratio $\tanh(\frac{D}{4})$ in terms of the entries of the matrix.
- It adapts well to the infinite-dimensional setting we will need later on.

REMARK 1.144. By Gelfand's formula for the spectral radius, we see that the Perron–Frobenius eigenvalue is given by $\lambda = \lim_{n\to\infty} \|L^n\|^{1/n}$; this can also be deduced from (1.150). Taking logarithms gives

$$\log \lambda = \lim_{n\to\infty} \frac{1}{n} \log \|L^n\|. \tag{1.155}$$

This procedure for extracting an exponential growth rate is one that we will encounter repeatedly.

Now we can return to the setting of Markov measures associated to a stochastic matrix P. Write $\mathbf{1}$ for the vector that has a 1 in every position: the property of being a stochastic matrix can be rewritten as $P\mathbf{1} = \mathbf{1}$, so $\mathbf{1}$ is the right Perron–Frobenius eigenvector of P, and the corresponding eigenvalue is $\lambda = 1$. The left Perron–Frobenius eigenvector is the stationary distribution $\pi = \pi P$ that we sought, and (1.150) provides a uniform contraction property. Observing that for every $q \in \Delta$ we have $\langle q, 1\rangle = 1 = \langle \pi, 1\rangle$, we conclude that the projection in Statement 4 of Theorem 1.141 maps every element of Δ to π. Thus we obtain the following.

COROLLARY 1.145. *Let P be a $d \times d$ primitive stochastic matrix. Then there exists a unique $\pi \in \Delta$ such that $\pi P = \pi$. All entries of π are positive, and there exist $C > 0$ and $\chi \in (0, 1)$ such that given any $q \in \Delta$ and $n \geq 0$, we have*

$$\|qP^n - \pi\| \leq C\chi^n \|q - \pi\|. \tag{1.156}$$

In particular, when q is the ath standard basis vector, qP^n is the ath row of P^n, so the rows of P^n converge to π exponentially fast:

$$|(P^n)_{ab} - \pi_b| \leq C\chi^n \text{ for all } n \geq 0. \tag{1.157}$$

Combining (1.136) and (1.157), we see that when μ is the invariant Markov measure associated to a primitive stochastic matrix P, given any words $v, w \in A^*$, we have (writing a for the last symbol of w and b for the first symbol of v):

$$\begin{aligned} |\mu(\sigma^{-n}[v] \cap [w]) - \mu[v] \cdot \mu[w]| &= \left|\mu[w]\left((P^{n-|w|+1})_{ab} - \pi_b\right)\mathbf{P}(v)\right| \\ &\leq \mu[w]\mathbf{P}(v)C\chi^{n-|w|+1}. \end{aligned}$$

Thus $\mu(\sigma^{-n}[v] \cap [w]) \to \mu[v] \cdot \mu[w]$ exponentially fast, which strengthens the estimates we used in the proof of ergodicity, and can be understood as providing a sense in which the correlation between the present and the future decays exponentially fast. We will return to this idea later on.

1.14. Chapter summary

We conclude this chapter with a brief overview of the main ideas and results that have appeared. These will return later, in more sophisticated situations.

(1) The mechanism driving apparent randomness in deterministic systems is expansion in phase space.
(2) Expansion in phase space leads to contraction in auxiliary dynamics, such as projective space or the space of curves tangent to unstable cones.
(3) Cone invariance can also be used to obtain contraction in the Hilbert metric under a finite diameter condition.
(4) These auxiliary contractions can be used to construct stable and unstable manifolds near hyperbolic fixed points (Hadamard–Perron Theorem 1.15) and to prove exponential convergence to the top eigendirection of a positive matrix (Perron–Frobenius Theorem 1.141).
(5) By "packing" the space of all orbit segments of a given length with enough "good" orbit segments, we can obtain strong results on the asymptotic behavior of a typical orbit (Birkhoff Ergodic Theorem 1.105).
(6) Dynamical systems with hyperbolic behavior can be coded by shift spaces.

CHAPTER 2

Uniform hyperbolicity and SRB measures

The previous chapter described some examples of dynamical systems exhibiting hyperbolic phenomena, and some of the fundamental results describing their behavior. Many natural questions remain.

(1) If a system has a hyperbolic linearization at a fixed point, then near that fixed point, the full system behaves like its linearization (Hadamard–Perron Theorem). Is any similar statement true for fixed points at which the linearization is not hyperbolic, such as $(\frac{1}{2}, 0)$ for the standard map?
(2) The Poincaré–Birkhoff–Smale Theorem shows that a transverse homoclinic point leads to a horseshoe on which the dynamics behave hyperbolically; however, this horseshoe set has zero Lebesgue measure. Are there "hyperbolic sets" with positive Lebesgue measure?
(3) Our initial discussions suggested that expansion in phase space is the mechanism driving the appearance of stochastic phenomena in deterministic systems. The Birkhoff Ergodic Theorem gave a precise sense in which deterministic systems can be described probabilistically, via the Strong Law of Large Numbers. However, some of the systems satisfying this theorem, such as irrational circle rotations, do not display any phase space expansion. From the point of view of ergodic theory, what is the difference (if any) between expanding and non-expanding behavior?
(4) The Birkhoff Ergodic Theorem provides results about typical trajectories with respect to an invariant measure. When Lebesgue measure is invariant, this gives information about Lebesgue-a.e. point. What happens for systems where Lebesgue measure is not invariant?

The first question leads to KAM theory, which we will not discuss at present. The remaining three questions will receive partial answers in this chapter. Of course there are many other questions one could ask as well.

The standard map and the Lorenz flow both exhibit expanding behavior, but there are complications:

- they also display contracting behavior in some directions;
- the expansion we witness does not occur at every point along an orbit, but rather occurs in fits and starts, with orbits appearing to alternate between regular and chaotic behavior. Moreover, in the standard map it appears that some orbits do not experience any chaotic behavior.

Because of these difficulties, we will set these examples aside for now, and will focus on simpler classes of systems in which the phenomenon of phase space expansion can be studied more easily; this will lead us to discuss *uniformly expanding maps*. Once we have studied these, we will reintroduce the contracting direction and study *uniformly hyperbolic sets*. The phenomenon of *non-uniform hyperbolicity* suggested by the second complication above is one that we will not have time to study in this course.

2.1. The one-sided full shift as a template

Given a finite alphabet S with $\#S = d \geq 2$, the one-sided full shift $\sigma\colon S^{\mathbb{N}} \to S^{\mathbb{N}}$ displays several phenomena that are worth keeping in mind as a template for the behavior we will study in expanding and hyperbolic systems.

- Writing $\mathcal{M}_\sigma$ for the set of all σ-invariant Borel probability measures on $S^{\mathbb{N}}$, the space $\mathcal{M}_\sigma$ is extremely large: it contains periodic measures, Bernoulli measures, and Markov measures (recall that these latter two classes provide uncountable families of invariant measures), as well as many others that we have not yet encountered.
- The Benoulli measure μ associated to the probability vector $(\frac{1}{d}, \dots, \frac{1}{d})$ "maximizes our ignorance" in the sense that for each n, it gives equal weight to every $w \in S^n$: given any $v, w \in S^n$, we have $\mu[v] = \mu[w] = d^{-n}$.
- If we wish to consider trajectories that are distributed according to a rule with prescribed transition probabilities between different states of the system, as encoded by a stochastic matrix P, then there is a unique stationary probability vector π for which the corresponding Markov measure is invariant.

A substantial part of the study of ergodic theory for hyperbolic dynamical systems consists of determining to what extent one can establish analogues of these statements.

We have already seen that in the case of an irrational rotation, the space of invariant measures can be very small, consisting of a single element. If we consider continuous maps on compact metric spaces, this is as small as it can be:

Definition 2.1. Let X be a compact metric space and $f\colon X \to X$ a continuous map. We will denote the space of Borel probability measures on X by $\mathcal{M} = \mathcal{M}(X)$, and the space of f-invariant measures by $\mathcal{M}_f = \mathcal{M}_f(X) \subset \mathcal{M}(X)$.

Definition 2.2. Given $\mu_n, \mu \in \mathcal{M}$, we say that $\mu_n \to \mu$ *in the weak* topology* if $\int \varphi \, d\mu_n \to \int \varphi \, d\mu$ for every $\varphi \in C(X)$.

We will see later that the weak* topology is induced by a metric (the Wasserstein metric). For now we merely observe that $\mathcal{M}$ is compact in the weak* topology as a consequence of the Banach–Alaoglu theorem from functional analysis.

Theorem 2.3 (Krylov–Bogolyubov). *Let X be a compact metric space and $f\colon X \to X$ a continuous map. Then $\mathcal{M}_f(X) \neq \emptyset$.*

PROOF. This is the same argument as in the paragraph following the proof of Lemma 1.135, except that now we work in infinite dimensions. Let $m \in \mathcal{M}(X)$ be arbitrary, and let $\mu_n := \frac{1}{n}\sum_{k=0}^{n-1} f_*^k m$. By weak* compactness of $\mathcal{M}(X)$ there exists a limit point $\mu = \lim_{j\to\infty} \mu_{n_j}$, and given any $\varphi \in C(X)$, we have

$$\begin{aligned}\int \varphi\, d(f_*\mu) - \int \varphi\, d\mu &= \lim_{j\to\infty} \frac{1}{n_j} \sum_{k=0}^{n_j-1} \Big(\int \varphi\, d(f_*^{k+1}\mu) - \int \varphi\, d(f_*^k \mu)\Big) \\ &= \lim_{j\to\infty} \frac{1}{n_j}\Big(\int \varphi\, d(f_*^{n_j}\mu) - \int \varphi\, d\mu\Big) = 0.\end{aligned}$$

It follows that $f_*\mu = \mu$. □

2.2. Expanding maps

DEFINITION 2.4. Let (X, d) be a compact metric space and $f\colon X \to X$ a continuous map. We say that f is *uniformly expanding* (or just *expanding*) if there exist $N \in \mathbb{N}$, $\epsilon > 0$, and $\gamma \in (0, 1)$ such that given any $x, y \in X$ satisfying $d(x, y) < \epsilon$, we have

$$d(x, y) \le \gamma d(f^N x, f^N y). \tag{2.1}$$

In the following examples, it suffices to take $N = 1$.

EXAMPLE 2.5. The one-sided shift map $\sigma\colon S^{\mathbb{N}} \to S^{\mathbb{N}}$ is expanding. The two-sided shift $\sigma\colon S^{\mathbb{Z}} \to S^{\mathbb{Z}}$ is not.

EXAMPLE 2.6. Let $X = S^1 = \mathbb{R}/\mathbb{Z}$ be the unit circle, and let $f\colon X \to X$ be the *doubling map* given by $f(x) = 2x \pmod 1$. Then f is uniformly expanding. More generally, if $f\colon X \to X$ is any C^1 map satisfying $\inf_x |f'(x)| > 1$, then f is uniformly expanding.

EXAMPLE 2.7. Given any $\theta \in S^1$, the circle rotation $R_\theta\colon S^1 \to S^1$ given by $R_\theta(x) = x + \theta \pmod 1$ is an isometry, and hence is not expanding.

As Examples 2.5–2.7 show, we should not expect expanding maps to be injective, in contrast to our examples from Chapter 1. However, a point cannot have too many preimages:

▶ EXERCISE 2.1. Prove that if X is a compact metric space and $f\colon X \to X$ is expanding, then there exists $D \in \mathbb{N}$ such that $\#f^{-1}(x) \le D$ for every $x \in X$. *(Hint: First prove that if $\epsilon > 0$ is as in Definition 2.4, then $f|_{B(y,\epsilon)}$ is injective for every $y \in X$. Then use compactness.)*

If $f(y) = x$, then we will often refer to $f|^{-1}_{B(y,\epsilon)}$ as a *branch* of f^{-1}. Condition (2.1) can be interpreted as saying that each branch of f^{-N} is a contraction. This may seem like an odd way of defining "expanding", but of course it is equivalent to the expansion condition $d(f^N x, f^N y) \ge \gamma^{-1} d(x, y)$, and in practice the contraction inequality is used in more proofs.

REMARK 2.8. In the case $X = S^{\mathbb{N}}$, the branches can be globally indexed: for each $a \in S$, the map $g_a \colon S^{\mathbb{N}} \to [a]$ defined by $g_a(x) = ax$ is a branch of σ^{-1}, and is a bijection. This is not possible for more general expanding maps. (Can you see why it does not work for the doubling map?)

In order to iterate the contraction inequality (2.1), one needs more than just proximity of x and y; one also needs their forward orbits to remain close for some period of time.

DEFINITION 2.9. Given $x \in X$, $n \in \mathbb{N}$, and $\epsilon > 0$, the *Bowen ball* (or *dynamical ball*) of radius ϵ and order n centered at x is

$$B_n(x, \epsilon) := \{y \in X : d(f^k y, f^k x) < \epsilon \text{ for all } 0 \le k < n\}. \tag{2.2}$$

▶ EXERCISE 2.2. Prove that f is uniformly expanding if and only if there exist $C, \lambda, \epsilon > 0$ such that given any $x \in X$, $n \in \mathbb{N}$, and $y \in B_n(x, \epsilon)$, we have

$$d(x, y) \le Ce^{-\lambda n} d(f^n x, f^n y). \tag{2.3}$$

(Hint: mimic the argument in §1.13 for deriving (1.152) *from* (1.151)*.)*

There is a very important subtlety to observe regarding the inverse branches $f|^{-1}_{B(y,\epsilon)}$. One might naturally form a mental picture in which near the iterate $y \mapsto f(y)$, we view f as being defined on a neighborhood of y, and f^{-1} as being defined on a neighborhood of $f(y)$. However, nothing in Definition 2.4 requires that $f(B(y, \epsilon))$ contains a neighborhood of $f(y)$.

To investigate this issue more closely, let us give it a name.

DEFINITION 2.10. An expanding map $f \colon X \to X$ is *locally onto* if $\epsilon > 0$ in Definition 2.4 can be chosen such that $f(B(y, \epsilon)) \supset B(f(y), \epsilon)$ for every $y \in X$.

REMARK 2.11. The locally onto condition will be essential for various proofs later on. Because of this, some authors include it in the definition of "expanding".[1] I have personally witnessed enough confusion as a result of this practice that I prefer to state it explicitly, as a separate condition. In some places in the literature, the term "Ruelle expanding" is used to refer to maps that are both expanding and locally onto.

If we attempt to picture an expanding map that is not locally onto, we may at first encounter difficulty; this is because in Euclidean spaces, no such example exists.

PROPOSITION 2.12. *Given $d \in \mathbb{N}$ and $\gamma \in (0, 1)$, let $U \subset \mathbb{R}^d$ be an open set and $f \colon U \to \mathbb{R}^d$ a C^1 map such that $\|x - y\| \le \gamma \|fx - fy\|$ for all $x, y \in U$. Then $f(U)$ is open.*

PROOF. The expansion condition guarantees that $\|Df(x)(v)\| \ge \gamma^{-1}\|v\|$ for every $x \in U$ and $v \in \mathbb{R}^d$, so each $Df(x)$ is invertible. Thus the conclusion follows from the Inverse Function Theorem. □

Lec 31
Fri, Apr 4

Proposition 2.12 implies that every C^1 expanding map on a compact Riemannian manifold[2] is locally onto. However, things are different when we consider symbolic dynamics:

DEFINITION 2.13. Given a finite set S, a *one-sided shift space* on S is a closed σ-invariant set $X \subset S^{\mathbb{N}}$. We also call X a *(one-sided) subshift*.

DEFINITION 2.14. Given a set $\mathcal{F} \subset S^*$ of words over an alphabet S, the *subshift determined by forbidding words in* $\mathcal{F}$ is the set $X \subset S^{\mathbb{N}}$ comprising all infinite sequences $x \in S^{\mathbb{N}}$ with the property that x does not contain any element of $\mathcal{F}$.

EXAMPLE 2.15. When $S = \{0, 1\}$ and $\mathcal{F} = \{11\}$, the subshift determined by forbidding words in $\mathcal{F}$ is the set of all infinite sequences $x \in \{0, 1\}^{\mathbb{N}}$ such that the symbol 1 never appears twice in a row.

DEFINITION 2.16. A subshift $X \subset S^{\mathbb{N}}$ is a *subshift of finite type* (SFT) if there exists a *finite* set of forbidden words $\mathcal{F} \subset S^*$ that determines X.

Observe that the shift map $\sigma\colon X \to X$ is uniformly expanding for every subshift $X \subset S^{\mathbb{N}}$. However, it is *not* locally onto in general:

▶▶ EXERCISE 2.3. Let $X \subset S^{\mathbb{N}}$ be a subshift. Prove that $\sigma\colon X \to X$ is locally onto if and only if X is an SFT.

▶▶ EXERCISE 2.4. Find an example of an expanding map that is not locally onto by finding an infinite set $\mathcal{F} \subset S^*$ such that the subshift $X \subset S^{\mathbb{N}}$ it determines is (1) nonempty, and (2) not determined by any finite set $\mathcal{F} \subset S^*$.

2.3. Absolute continuity and physical measures

Let X be a compact metric space and $f\colon X \to X$ a continuous map. Given $x \in X$ and $n \in \mathbb{N}$, consider the *empirical measure*

$$\delta_{x,n} := \frac{1}{n} \sum_{k=0}^{n-1} \delta_{f^k(x)} = \frac{1}{n} \sum_{k=0}^{n-1} f_*^k \delta_x. \tag{2.4}$$

As in the proof of the Krylov–Bogolyubov Theorem, the sequence $(\delta_{x,n})_{n\in\mathbb{N}}$ has at least one weak*-accumulation point μ, and every such μ is an invariant measure. If there is *exactly* one such μ – that is, if the sequence of empirical measures along the orbit of x converges in the weak* topology – then we say that x is a *generic point* for μ. To put it another way:

[1] See, for example, §11.2 of "Foundations of Ergodic Theory" by Viana and Oliveira.

[2] If you have not studied Riemannian manifolds before, it is enough for purposes of the present course to mentally substitute the words "a d-dimensional torus $\mathbb{R}^d/\mathbb{T}^d$" every time you see the phrase "a compact Riemannian manifold".

DEFINITION 2.17. Given $\mu \in \mathcal{M}_f(X)$, a point x is *generic* for μ if

$$\frac{1}{n} S_n \varphi(x) \to \int \varphi \, d\mu \text{ for all } \varphi \in C(X). \tag{2.5}$$

Let G_μ denote the set of generic points for μ.

The Birkhoff Ergodic Theorem implies that if μ is ergodic, then $\mu(G_\mu) = 1$, so μ-a.e. point is generic.

►► EXERCISE 2.5. On $\{0,1\}^{\mathbb{N}}$, let δ_0 and δ_1 be the delta-measures associated to the sequences of all 0s and all 1s, respectively, and let $\mu = \frac{1}{2}(\delta_0 + \delta_1)$. Prove that G_μ is nonempty even though μ is not ergodic. *(Hint: consider* $x = 0^{a_1}1^{a_1}0^{a_2}1^{a_2}\cdots$ *for a suitable sequence* $a_1 \ll a_2 \ll \cdots$*.)*

Of particular importance is the case when M is a compact Riemannian manifold and $f\colon M \to M$ is a C^1 map, in which case we can study Lebesgue measure: that is, we write m for the probability measure obtained by normalizing the Riemannian volume form[3] on M, and ask whether m is (1) invariant and (2) ergodic. If it is both invariant and ergodic, then $m(G_m) = 1$, so (2.5) describes the asymptotic behavior of Lebesgue-a.e. trajectory. This motivates the following definition.

DEFINITION 2.18. Given a compact Riemannian manifold M and a C^1 map $f\colon M \to M$, an ergodic invariant measure $\mu \in \mathcal{M}_f$ is *physical* if $m(G_\mu) > 0$, where m is normalized Lebesgue measure on M.

If Lebesgue measure is invariant and ergodic, then it is the unique physical measure. If invariance or ergodicity fails, then we are led to ask whether a physical measure exists, and if so, whether it is unique.

PROPOSITION 2.19. *If $\mu \in \mathcal{M}_f$ is ergodic and satisfies $\mu \ll m$, then μ is physical.*

PROOF. The Birkhoff Ergodic Theorem implies that $\mu(G_\mu) > 0$, so $m(G_\mu) > 0$ by the definition of absolute continuity. □

Thus we are let to search for an *absolutely continuous invariant probability measure (ACIP)*, and in particular, for an ergodic ACIP.

The standard map and the Lorenz flow (or at least its time-1 map) fit into this framework of C^1 maps on compact manifolds; note that although both are defined on Euclidean space, which is not compact, the Lorenz attractor itself is compact, and the periodicity of the standard map lets us consider it as a map on the (compact) torus. Both of these are common mechanisms for obtaining a compact system from a map on Euclidean space.

DEFINITION 2.20. Let $U \subset \mathbb{R}^q$ be open and bounded. If $f\colon U \to U$ is continuous and has the property that $\overline{f(U)} \subset U$, then we say that U is a *trapping region* for f, and refer to the compact invariant set $X = \bigcap_{n\geq 0} \overline{f^n(U)}$ as an *attractor*.

[3] If these words are mysterious to you, replace them with the following example: when $M = \mathbb{R}^q/\mathbb{Z}^q$ is the q-dimensional torus, m is just the measure inherited from Lebesgue measure on $\mathbb{R}^q$.

REMARK 2.21. In the setting of Definition 2.20, it is common for the attractor X to have Lebesgue measure 0. When this happens, there can be no ACIP, since every invariant measure is supported on the attractor; we encountered this already with the Lorenz flow. For such systems we will eventually need to work with the broader class of *Sinai–Ruelle–Bowen (SRB) measures*. For expanding maps, however, we will see that ACIPs suffice.

DEFINITION 2.22. Let $g\colon \mathbb{R}^q \to \mathbb{R}^q$ be a continuous map with the property that $g(x+\mathbb{Z}^q) \subset g(x)+\mathbb{Z}^q$ for every $x \in \mathbb{R}^q$. Then the map $f\colon \mathbb{T}^q \to \mathbb{T}^q$ given by $f(x+\mathbb{Z}^q) = f(x)+\mathbb{Z}^q$ is well-defined and continuous. We refer to f as the *toral map induced by* g.

We will study expanding toral maps induced by expanding maps on $\mathbb{R}^q$. Similarly to our study of horseshoes, it will be helpful to start by considering the linear case, and then later to turn our attention to nonlinear examples.

EXAMPLE 2.23. Consider the q-dimensional torus $M = \mathbb{T}^q = \mathbb{R}^q/\mathbb{Z}^q$, and let A be any $q \times q$ integer matrix whose eigenvalues all lie outside the unit circle (every eigenvalue λ has $|\lambda| > 1$). Then the linear map $A\colon \mathbb{R}^q \to \mathbb{R}^q$ is uniformly expanding in the Euclidean norm on $\mathbb{R}^q$, and maps $\mathbb{Z}^q$ into itself, so it passes to a map f on $M = \mathbb{T}^q$ given by $f(x+\mathbb{Z}^q) = Ax+\mathbb{Z}^q$. This map is expanding in the metric

$$d(a+\mathbb{Z}^q, b+\mathbb{Z}^q) = \min\{\|y-z\| : y \in a+\mathbb{Z}^q, z \in b+\mathbb{Z}^q\}. \tag{2.6}$$

►► EXERCISE 2.6. Let $f\colon \mathbb{T}^q \to \mathbb{T}^q$ be the toral endomorphism (linear toral map) induced by a $q \times q$ integer matrix A satisfying $\det A \neq 0$. (Note that we do not place any expanding assumption on A.) Let $D = |\det A|$, and prove that every point on the torus has exactly D preimages under f.

As a consequence of Exercise 2.6, there exists $\epsilon > 0$ such that given any $x \in \mathbb{T}^q$, we can write $f^{-1}(x) = \bigsqcup_{i=1}^{D} U_i^x$, where $f|_{U_i^x}\colon U_i^x \to B(x,\epsilon)$ is a homeomorphism for each i. We will write $b_i^x := f|_{U_i^x}^{-1}\colon B(x,\epsilon) \to U_i^x$ for the corresponding inverse branches.[4]

Now we restrict our attention to the expanding case. Since the determinant is the product of the eigenvalues, we must have $D > 1$, so f is not invertible and every point has multiple preimages.

PROPOSITION 2.24. *Given any expanding toral endomorphism as in Example 2.23, Lebesgue measure m is invariant and ergodic.*

PROOF. For invariance, we observe that since b_i^x is affine and its linear part is A^{-1}, which has determinant $\pm D^{-1}$, for every measurable $E \subset B(x,\epsilon)$ we have

$$m(b_i^x(E)) = D^{-1} m(E). \tag{2.7}$$

Summing over the inverse branches gives

$$m(f^{-1}(E)) = m(E) \text{ for all measurable } E \subset \mathbb{T}^q \text{ with } \operatorname{diam}(E) < \epsilon. \tag{2.8}$$

[4]Note that the labeling of these branches may vary discontinuously in x.

Lec 32
Mon, Apr 7

Since every measurable set can be written as a disjoint union of measurable sets with diameter $< \epsilon$, this implies that m is f-invariant.

For ergodicity, fix a set E such that $f^{-1}(E) = E$ and $m(E) < 1$. We will prove that $m(E) = 0$ via the following strategy.

(1) Partition M into ℓ sets $X_1, \dots, X_\ell$ on which inverse branches b_a can be defined for each $a \in S := \{1, \dots, D\}$. For each $n \in \mathbb{N}$, this gives a partition $\{b_w(X_j) : w \in S^n, 1 \le j \le \ell\}$, where $b_w := b_{w_n} \circ \cdots \circ b_{w_1}$.
(2) Argue that since $m(E) < 1$ and $\operatorname{diam} b_w(X_j) \to 0$ as $|w| \to \infty$, the relative measure of E in these sets must become arbitrarily small for some choice of w, j.
(3) Use the invariance of E and the linearity of A to deduce that there exists j such that $m(E \cap X_j) = 0$.
(4) Taking $N \in \mathbb{N}$ such that $f^N(X_j) = M$, deduce that $m(E) = 0$.

Step 1: *Partitions.*

Say that a set $Z \subset M$ is *δ-separated* if $d(y, z) \ge \delta$ for every $y, z \in Z$ with $y \ne z$. Say that Z is *maximal* if there does not exist a δ-separated set that properly contains Z. In this case we must have $\bigcup_{z \in Z} B(z, \delta) = M$ (so Z is *δ-spanning*), since if there were a point $y \in M \setminus \bigcup_{z \in Z} B(z, \delta)$, then $Z \cup \{y\}$ would be δ-separated, violating maximality.

With the previous paragraph in mind, let $Z = \{z_1, \dots, z_\ell\}$ be a maximal $\epsilon/2$-separated set. (Observe that Z is finite because M is compact.) We will define a partition $\{X_1, \dots, X_\ell\}$ that is *adapted* to Z in the sense that

$$B(z_j, \epsilon/4) \subset X_j \subset B(z_j, \epsilon/2) \text{ for every } j \in \{1, \dots, \ell\}. \tag{2.9}$$

Writing $C_j := B(z_j, \epsilon/4)$ (the "core") and $B_j := B(z_j, \epsilon/2)$, we do this via the following iterative procedure:

$$Y_0 := \bigcup_{j=1}^{\ell} C_j, \qquad X_j := C_j \cup (B_j \setminus Y_{j-1}), \qquad Y_j := Y_{j-1} \cup X_j. \tag{2.10}$$

Thus Y_0 consists of all the "cores", which can only be assigned to the predetermined X_j, and Y_j consists of these cores together with everything that has been assigned to some X_i for $i \le j$. Then X_j consists of the appropriate core C_j together with all "available" points in B_j.

Now given $a \in S := \{1, \dots, D\}$, we define $b_a \colon M \to M$ by the condition that $b_a|_{X_j} = b_a^{x_j}$; that is, b_a represents a globally defined inverse branch, which is measurable but not necessarily continuous, although its restriction to each X_j is continuous.

Given any $n \in \mathbb{N}$, we consider the following partition of M:

$$\mathcal{P}_n := \{b_w(X_j) : w \in S^n, 1 \le j \le \ell\}, \quad \text{where } b_w := b_{w_n} \circ \cdots \circ b_{w_1}.$$

Since f is uniformly expanding and $\operatorname{diam}(X_j) < \epsilon$ for every j, we have $\max_{C\in\mathcal{P}_n} \operatorname{diam} C \to 0$ as $n \to \infty$.

STEP 2: *The set E has small relative measure.*

The partitions $\mathcal{P}_n$ can be used to approximate the invariant set E:

LEMMA 2.25. *There exist $\mathcal{R}_n \subset \mathcal{P}_n$ such that writing $R_n := \bigcup_{C\in\mathcal{R}_n} C$, we have $m(R_n \triangle E) \to 0$ as $n \to \infty$.*

PROOF. Fix $\epsilon > 0$. There exist a compact set $K \subset E$ and an open set $U \supset E$ such that $m(U \setminus K) < \epsilon$. Let $\gamma = \inf\{d(x, y) : x \in K, y \in U^c\}$, and choose n sufficiently large that $\operatorname{diam} C < \gamma$ for every $C \in \mathcal{P}_n$. Then let $\mathcal{R}_n := \{C \in \mathcal{P}_n : C \cap K \neq \emptyset\}$. Each $C \in \mathcal{R}_n$ has $C \cap U^c = \emptyset$, so we have $K \subset R_n \subset U$, which implies that $R_n \triangle E \subset U \setminus K$, completing the proof. □

We claim that the following quantity tends to 0 as $n \to \infty$:

$$\xi_n := \min\left\{\frac{m(E \cap C)}{m(C)} : C \in \mathcal{P}_n\right\}.$$

Indeed, let $\mathcal{R}_n \subset \mathcal{P}_n$ and $R_n \subset M$ be given by Lemma 2.25. Then

$$m(E \cap R_n^c) \geq \sum_{C\in\mathcal{P}_n\setminus\mathcal{R}_n} m(E \cap C) \geq \xi_n m(R_n^c).$$

Since $m(E\cap R_n^c) \leq m(E\triangle R_n^c) \to 0$ and $m(R_n^c) \to m(E^c) > 0$, we must have $\xi_n \to 0$. It follows that there exists $C_n \in \mathcal{P}_n$ such that $\frac{m(E\cap C_n)}{m(C_n)} \to 0$.

STEP 3: *From small scale to large scale.*

For each n, we have $f^n(C_n) = X_{j(n)}$ for some $j(n) \in \{1, \dots, \ell\}$. Since E is invariant, we have $E = f^{-1}(E)$, and applying f to both sides gives $f(E) = f(f^{-1}(E)) = E \cap f(M) = E$ since f is onto. Iterating gives $f^n(E) = E$. Similarly, we have $f^n(E^c) = E^c$, and we obtain

$$X_{j(n)} \cap E = X_{j(n)} \setminus E^c = f^n(C_n) \setminus f^n(E^c) \subset f^n(C_n \setminus E^c) = f^n(C_n \cap E).$$

(For the inclusion: given $x \in f^n(C_n \setminus E^c)$, we have $x = f^n(y)$ for some $y \in C_n$, and since $x \notin f^n(E^c)$ we must have $y \notin E^c$, so $x \in f^n(C_n \setminus E^c)$.

Since $f^n|_{C_n}$ scales Lebesgue measure by D^n, we conclude that

$$\frac{m(X_{j(n)} \cap E)}{m(X_{j(n)})} \leq \frac{m(f^n(C_n \cap E))}{m(f^n(C_n))} = \frac{m(C_n \cap E)}{m(C_n)}. \tag{2.11}$$

This last quantity tends to 0 as $n \to \infty$, and since $m(X_{j(n)}) \leq 1$, we get

$$\lim_{n\to\infty} m(X_{j(n)} \cap E) = 0. \tag{2.12}$$

Since $j(n)$ can take only finitely many values, there exists $i \in \{1, \dots, \ell\}$ such that $j(n) = i$ for arbitrarily large n, and thus

$$m(X_i \cap E) = 0. \tag{2.13}$$

STEP 4: *From local to global.*

To complete the proof, it remains to use (2.13) to show that $m(E) = 0$. Start by observing that since all eigenvalues of the $q \times q$ matrix A exceed 1 in absolute value, there exists $N \in \mathbb{N}$ such that $A^N B(0, \epsilon/4)$ contains the set $[-\frac{1}{2}, \frac{1}{2}]^q \subset \mathbb{R}^q$. Using the fact that $X_i \supset B(z_i, \epsilon/4)$, this implies that

$$f^N X_i = M. \tag{2.14}$$

This implies that $E = f^N(X_i) \cap f^N(E) \subset f^N(X_i \cap E)$, and since f maps Lebesgue null sets to Lebesgue null sets, we conclude that $m(E) = 0$, which proves ergodicity and completes the proof of Proposition 2.24. □

Lec 33
Wed, Apr 9

The argument in the proof of Proposition 2.24 can be adapted to a broader setting, including a large class of *nonlinear* expanding maps. To this end, we identify the main ingredients that were used.

- Uniform expansion guaranteed that the partitions $\mathcal{P}_n$ have elements whose diameter tends to 0, and thus can be used to approximate the set E in the sense of Lemma 2.25, leading to the conclusion that there exist $C_n \in \mathcal{P}_n$ satisfying $\frac{m(E \cap C_n)}{m(C_n)} \to 0$.
- The map f is onto (used in Step 3), and in fact has the stronger *topological exactness* property in (2.14): given any open set $U \subset M$, there exists $N \in \mathbb{N}$ such that $f^N(B(x, \delta)) = M$.
- Every inverse branch of f scales Lebesgue measure m by a constant, so that as in (2.11), we have

$$\frac{m(f^n(C_n \cap E))}{m(f^n(C_n))} = \frac{m(C_n \cap E)}{m(C_n)}, \tag{2.15}$$

  and moreover, if $Y \subset M$ has $m(Y) = 0$, then $m(fY) = 0$ as well (this is used in the last part of Step 4).

The first of these three ingredients generalizes immediately: the argument given in the proof of Proposition 2.24 works just as well if Lebesgue measure m is replaced by any other Borel probability measure μ, and for any set E with $m(E) < 1$, whether or not μ and E are invariant.

The remaining ingredients can be generalized as well, but require more work. The second ingredient, topological exactness, is true for every expanding map on a connected manifold, but we will need a new proof. The third ingredient, scaling by a constant, is not true in general, and will require us to impose a condition on the regularity of Df in order to obtain a *bounded distortion* property.

We will carry all of this out in the following sections. Before doing so, we conclude this section by describing the strategy for producing an ACIP when f is a nonlinear expanding map on a compact connected Riemannian manifold M with normalized Lebesgue measure m.

REMARK 2.26. If you prefer not to think about general Riemannian manifolds, you can think about everything that follows in the setting of expanding maps on the torus, by taking $M = \mathbb{T}^q$ as in Example 2.23, with metric is given by (2.6). In this case every tangent space T_xM is naturally identified with $\mathbb{R}^q$, and m comes from Lebesgue measure on $\mathbb{R}^q$.

Let $\mathcal{M}_{\text{ac}} \subset \mathcal{M}$ denote the set of all Borel probability measures μ such that $\mu \ll m$. Then our goal is to find an invariant measure in $\mathcal{M}_{\text{ac}}$. A key first step is the following.

▶ EXERCISE 2.7. Let M be a compact Riemannian manifold with Lebesgue measure m, and let $f\colon M \to M$ be a C^1 map such that $\det DF(x) \neq 0$ for every $x \in M$. Then $f_*m \ll m$.

We see from Exercise 2.7 that $f_*\mu \ll m$ whenever $\mu \ll m$, so f_* maps $\mathcal{M}_{\text{ac}}$ to itself, and we seek a fixed point of this mapping.

By the Radon–Nikodym theorem, there is a bijective correspondence $\pi\colon \mathcal{M}_{\text{ac}} \to \Delta := \{h \in L^1(m) : h \geq 0, \int h\, dm = 1\}$. Thus there is a map $L\colon \Delta \to \Delta$ such that the following diagram commutes.

(2.16)
$$\begin{array}{ccc} \mathcal{M}_{\text{ac}} & \xrightarrow{f_*} & \mathcal{M}_{\text{ac}} \\ \downarrow{\scriptstyle \pi} & & \downarrow{\scriptstyle \pi} \\ \Delta & \xrightarrow{L} & \Delta \end{array}$$

The map L is the *Ruelle–Perron–Frobenius transfer operator*. To find an ACIP, we will write a formula for L and exhibit an invariant subset of Δ on which L acts as a contraction in the Hilbert metric.

All of this is completely analogous to the process we went through when we discussed Markov measures in §1.13: we define a class of measures within which we want to find an invariant measure, find a description of that class in terms of a "simpler" space on the pushforward under the dynamics admits a concrete formula, and then impose an appropriate Hilbert metric on that class under which we obtain a contraction. The dictionary is as follows.

	Markov measures for stochastic matrix P	ACIPs
class of measures	"compatible" with P	absolutely continuous
simpler space	simplex	positive unit ball in L^1
restricted action	right multiplication by P	RPF transfer operator
fixed point	probability eigenvector	invariant density function

2.4. Hölder regularity and bounded distortion

In the next section, we will prove a more general version of the ergodicity result from Proposition 2.24, but before doing so, we need to formulate a condition that will replace the scale-changing step (2.15), in which we can no longer expect to have equality when f is nonlinear.

To this end, let $f\colon M \to M$ be an expanding map on a compact Riemannian manifold, and let m be normalized Lebesgue measure on M. Let $\epsilon > 0$ be the locally expanding scale, so $f|_{B(x,\epsilon)}$ is injective and expanding for every $x \in M$.

Definition 2.27. The expanding map f has *bounded distortion* with respect to Lebesgue measure if there exists $Q > 0$ such that given any $x \in M$, $n \in \mathbb{N}$, and any $A, B \subset B_n(x, \epsilon)$, we have

$$\frac{m(f^n(A))}{m(f^n(B))} = Q^{\pm 1}\frac{m(A)}{m(B)}, \tag{2.17}$$

where the notation $X = Q^{\pm 1}Y$ means $Q^{-1}Y \le X \le QY$.

In the ergodicity argument, bounded distortion will allow us to recover (2.15) up to a factor of $Q^{\pm 1}$, which is sufficient for its role in the proof.

To verify bounded distortion, we need more restrictive conditions on f; the bounded distortion property does not hold for every C^1 expanding map.

Definition 2.28. Given two metric spaces (X, d) and (Y, ρ), a function $g\colon X \to Y$ is *Hölder continuous* if there exist $\alpha \in (0, 1]$ and $C > 0$ such that for every $x, y \in X$, we have

$$\rho(g(x), g(y)) \le C d(x, y)^\alpha. \tag{2.18}$$

If g is Hölder continuous with $\alpha = 1$, then we say that g is *Lipschitz*.

Definition 2.29. Given a compact Riemannian manifold M, we say that the map $f\colon M \to M$ is $C^{1+\alpha}$ if it is C^1 and if its derivative Df is α-Hölder continuous for some $\alpha > 0$.

Proposition 2.30. *Let M be a compact Riemannian manifold and $f\colon M \to M$ a $C^{1+\alpha}$ uniformly expanding map. Then f has bounded distortion with respect to Lebesgue measure.*

Proof. Define a function $J\colon M \to \mathbb{R}$ by

$$J(x) := |\det Df(x)|, \tag{2.19}$$

and observe that J is Hölder continuous since f is $C^{1+\alpha}$. Since J is bounded away from 0 (a consequence of the uniformly expanding property), it follows that $\log J$ is Hölder continuous as well: there exists $C > 0$ such that

$$|\log J(x) - \log J(y)| \le C d(x, y)^\alpha \text{ for all } x, y \in M. \tag{2.20}$$

Given any $x \in M$ and $n \in \mathbb{N}$, we have

$$|\det Df^n(x)| = \prod_{k=0}^{n-1} J(x) = e^{\mathsf{S}_n \log J(x)},$$

and given any $A \subset B_n(x,\epsilon)$, the usual change-of-variables formula gives

$$m(f^n(A)) = \int_A |\det Df^n(y)|\, dm(y) = \int_A e^{\mathsf{S}_n \log J(y)}\, dm(y). \tag{2.21}$$

For every $y \in A$, we have $y \in B_n(x,\epsilon)$, and the uniform expansion property (in its "backwards contraction") form guarantees that $d(f^{n-k}(y), f^{n-k}(x))$ is exponentially small in terms of k. More precisely, taking N such that each inverse branch of f^N contracts distances by a factor of $\gamma \in (0,1)$, we see that given any integers $\ell \geq 0$ and $r > 0$ such that $\ell N + r \leq n$, we have

$$d(f^{n-(\ell N+r)}(x), f^{n-(\ell N+r)}(y)) \leq \gamma^\ell d(f^{n-r}(x), f^{n-r}(y)) < \epsilon\gamma^\ell,$$

and thus

$$|\log J(f^{n-(\ell N+r)}(x)) - \log J(f^{n-(\ell N+r)}(y))| \leq C\epsilon^\alpha \gamma^{\alpha\ell}.$$

Since every $k \in \{1,\dots,n\}$ can be written as $k = \ell N + r$ for some $\ell \geq 0$ and $r \in \{1,\dots,N\}$, we conclude that

$$|\mathsf{S}_n \log J(x) - \mathsf{S}_n \log J(y)| \leq N \sum_{\ell=0}^{\infty} C\epsilon^\alpha \gamma^{\alpha\ell} = \frac{NC\epsilon^\alpha}{1-\gamma^\alpha} =: K < \infty.$$

Combining this with (2.21) gives

$$m(f^n(A)) = \int_A e^{\pm K} e^{\mathsf{S}_n \log J(x)}\, dm(y) = e^{\pm K} e^{\mathsf{S}_n \log J(x)} m(A).$$

Thus given any $A, B \subset B_n(x,\epsilon)$, we have

$$\frac{m(f^n(A))}{m(f^n(B))} = e^{\pm 2K} \frac{e^{\mathsf{S}_n \log J(x)} m(A)}{e^{\mathsf{S}_n \log J(x)} m(B)} = e^{\pm 2K} \frac{m(A)}{m(B)},$$

which proves Proposition 2.30 with $Q = e^{2K}$. □

The Hölder continuity property plays a central role in many arguments in hyperbolic dynamics, typically following a similar structure to the above proof: hyperbolicity guarantees that two nearby trajectories are in fact exponentiall close, which together with Hölder continuity of some function φ allows us to bound a difference of ergodic sums $\mathsf{S}_n\varphi$ by a geometric series.

Although the following result plays no role in the proofs to come, it is perhaps worth mentioning here that in the definition of expanding maps, one can take $N = 1$ in (2.1), and $C = 1$ in (2.3), by passing to an *adapted metric* that does not change the Hölder class, in the following sense.

DEFINITION 2.31. Two metrics d and ρ on X are *Hölder equivalent* if both of the identity maps $(X,d) \to (X,\rho)$ and $(X,\rho) \to (X,d)$ are Hölder continuous. Equivalently, there are $\alpha \in (0,1]$ and $C > 0$ such that

$$\rho(x,y) \le Cd(x,y)^\alpha \quad \text{and} \quad d(x,y) \le C\rho(x,y)^\alpha \text{ for all } x,y \in X. \tag{2.22}$$

PROPOSITION 2.32. *Let (M,d) be a compact Riemannian manifold and $f\colon X \to X$ a C^1 expanding map. There exists a metric ρ on X that is Hölder equivalent to d and in which f is uniformly expanding with $N = 1$; that is, there exist $\epsilon > 0$ and $\chi \in (0,1)$ such that for every $x, y \in X$ with $d(x,y) < \epsilon$, we have*

$$\rho(x,y) \le \chi\rho(fx,fy). \tag{2.23}$$

PROOF. Roughly speaking, the idea is to make the following definition when $x \approx y$: consider the *separation time* $s(x,y) = \min\{n \in \mathbb{N} : d(f^n x, f^n y) \ge \epsilon\}$, and then let $\rho(x,y) = 2^{-s(x,y)}$. Then when $\rho(x,y) \le \frac{1}{2}$, we have

$$s(x,y) = s(fx,fy) + 1 \quad \Rightarrow \quad \rho(x,y) = \frac{1}{2}\rho(fx,fy). \tag{2.24}$$

This does not quite do the job, because s is not continuous in x and y, so ρ is not equivalent to d. We can fix it by replacing the sequence $(d(f^n x, f^n y))_n$ in the definition of $s(x,y)$ with its linear interpolation: for each $t \ge 0$, write $t = n + r$ where $n \in \mathbb{N}_0$ and $r \in [0,1)$, and let

$$\begin{aligned} D_t(x,y) &:= (1-r)d(f^n x, f^n y) + rd(f^{n+1}x, f^{n+1}y), \\ s(x,y) &:= \inf\{t \ge 0 : D_t(x,y) \ge \epsilon\}. \end{aligned}$$

Now if we once again define $\rho(x,y) := 2^{-s(x,y)}$, we see that when $\rho(x,y) \le \frac{1}{2}$, we have $D_t(x,y) < \epsilon$ for all $t \in [0,1)$, so the reasoning in (2.24) still holds. It remains to show that ρ is Hölder equivalent to d.

First observe that $\rho(x,y) = 2^{-s(x,y)} = e^{-s(x,y)\log 2}$. Given $x, y \in X$ and $n = \lfloor s(x,y) \rfloor - 1$, we have $y \in B_n(x,\epsilon)$, so by Exercise 2.2, we see that

$$\begin{aligned} d(x,y) &\le Ce^{-\lambda n} d(f^n x, f^n y) < Ce^{-\lambda n}\epsilon \\ &\le Ce^{-\lambda(s(x,y)-2)}\epsilon = C\epsilon e^{2\lambda}\rho(x,y)^{\lambda/\log 2}. \end{aligned}$$

This proves that the identity map $(X,\rho) \to (X,d)$ is Hölder continuous. For the other direction, let $L = \max\{\|Df(x)\| : x \in M\}$, so that $d(fx,fy) \le Ld(x,y)$ for every $x, y \in M$: then given any $x, y \in M$, we have

$$d(f^k x, f^k y) \le L^k d(x,y) \text{ for all } k \ge 0.$$

To get an upper bound on ρ in terms of d, we must bound $s(x,y)$ below, so we are interested in those k for which $L^k d(x,y) < \epsilon$. The following inequalities are all

equivalent.

$$\begin{aligned} L^k d(x,y) &< \epsilon \\ L^{-k} &> d(x,y)/\epsilon \\ 2^{-k\frac{\log L}{\log 2}} &> d(x,y)/\epsilon \\ 2^{-k} &> \epsilon^{-\beta} d(x,y)^\beta \qquad \text{where } \beta = \log 2/\log L \end{aligned}$$

Taking k to be maximal such that $L^k d(x,y) < \epsilon$, we see that the last inequality fails for $k+1$, so

$$2^{-(k+1)} \le \epsilon^{-\beta} d(x,y)^\beta.$$

At the same time, $s(x,y) \ge k$, so we have

$$\rho(x,y) = 2^{-s(x,y)} \le 2^{-k} \le 2\epsilon^{-\beta} d(x,y)^\beta,$$

which completes the proof. □

2.5. Ergodicity for nonlinear expanding maps

Lec 34
Fri, Apr 11

This section is devoted to the proof of the following result, which gives a version of ergodicity even *without* requiring invariance of Lebesgue measure,[5] and implies that if an ACIP exists, it must be ergodic and unique. In the next section, we will prove existence.

THEOREM 2.33. *Let M be a compact connected Riemannian manifold and $f\colon M \to M$ a $C^{1+\alpha}$ expanding map. Then given any f-invariant set $E \subset M$, we either have $m(E) = 0$ or $m(E^c) = 0$, where m denotes Lebesgue measure on M.*

COROLLARY 2.34. *Let M and f be as in Theorem 2.33. If $\mu \in \mathcal{M}_f$ is absolutely continuous with respect to Lebesgue measure, then μ is ergodic. In particular, there can be at most one ACIP.*

PROOF. If $E = f^{-1}(E)$, then Theorem 2.33 implies that $m(E) = 0$ or $m(E^c) = 0$, and since $\mu \ll m$, we conclude that $\mu(E) = 0$ or $\mu(E^c) = 0$, implying ergodicity. If μ, ν are two ACIPs, then $\frac{1}{2}(\mu+\nu)$ is an ACIP as well, so it must be ergodic, which is only possible if $\mu = \nu$. □

In the proof of Theorem 2.33, we will use the bounded distortion property from Proposition 2.30, as well as the following fact, which complements Exercise 2.7.

▶ EXERCISE 2.8. Let M be a compact Riemannian manifold with Lebesgue measure M, and let $f\colon M \to M$ be a C^1 map. Given any measurable set $E \subset M$ with $m(E) = 0$, we have $m(f(E)) = 0$.

We will also need the following concept.

DEFINITION 2.35. The map $f\colon M \to M$ is *topologically exact* if for every nonempty open set $U \subset M$, there exists $k \ge 1$ such that $f^k(U) = M$.

[5] One could just as well give the definition of "ergodic" for all Borel probability measures, not just invariant measures.

LEMMA 2.36. *If M is a compact connected Riemannian manifold and $f\colon M \to M$ is a C^1 expanding map, then f is topologically exact.*

PROOF. Thanks to Proposition 2.32, we can assume that the metric d witnesses expansion in a single step: there exists $\chi \in (0,1)$ and $\epsilon > 0$ such that $d(x,y) \le \chi d(fx, fy)$ whenever $y \in B(x,\epsilon)$. Now we argue as follows:

$$\text{expanding} \;\Rightarrow\; \text{locally onto} \;\Rightarrow\; \text{covering map} \;\Rightarrow\; \text{topologically exact.}$$

For the first step, for every $x \in M$, by Proposition 2.12 we see that $f(B(x,\epsilon))$ is open and thus contains a ball $B(fx, \epsilon_x)$. These open balls cover the compact set $f(M) \subset M$, so there is a finite subcover. Replacing ϵ by the minimum of the associated ϵ_x, we obtain the following properties:

- if $x \in M$ and $y \in B(x,\epsilon)$, then $d(x,y) \le \chi d(fx, fy)$;
- $f(B(x,\epsilon)) \supset B(fx,\epsilon)$.

Given $y \in M$, let $f^{-1}(y) = \{x_1, \dots, x_D\}$. For each i, the map $b_i := f|^{-1}_{B(x_i,\epsilon)}$ carries $B(y,\epsilon)$ homeomorphically to a neighborhood of x_i. In particular, $f^{-1}(B(y,\epsilon)) = \bigsqcup_{i=1}^{D} b_i(B(y,\epsilon))$, so f is a covering map.

Consequently, every curve can be lifted under f: given any continuous $c\colon [0,1] \to M$ and any $x \in f^{-1}(c(0))$, there is a unique continuous $\tilde{c}\colon [0,1] \to M$ such that $f \circ \tilde{c} = c$ and $\tilde{c}(0) = x$. This is a standard fact from topology, which can be readily proved by observing that if $I \subset [0,1]$ is a sufficiently small interval that contains some t for which $\tilde{c}(t)$ has already been determined, then we must have $\tilde{c}|_I = b \circ c|_I$ for the unique inverse branch b that satisfies $b(c(t)) = \tilde{c}(t)$. Since each inverse branch contracts distance by χ, we see that if c is rectifiable, then $\tilde{c}$ is as well, with $\ell(\tilde{c}) \le \chi \ell(c)$.

Now we can prove topological exactness. Given any nonempty open set $U \subset M$, there exist $x \in M$ and $\delta > 0$ such that $B(x,\delta) \subset U$. Take $k \in \mathbb{N}$ sufficiently large that $(1 + \operatorname{diam} M)\chi^k < \delta$. Given any $y \in M$, we will prove that $f^{-k}(y) \cap B(x,\delta) \neq \emptyset$; this will imply that $f^k(U) = M$.

Since M is connected, there exists a C^1 curve $c\colon [0,1] \to M$ with length at most $1 + \operatorname{diam} M$ such that $c(0) = f^k(x)$ and $c(1) = y$. Lifting c by the covering map f^k, we obtain a C^1 curve $\tilde{c}\colon [0,1] \to M$ with length at most $(1 + \operatorname{diam} M)\chi^k < \delta$ such that $\tilde{c}(0) = x$ and $\tilde{c}(1) \in f^{-k}(y)$. It follows that $\tilde{c}(1) \in B(x,\delta) \cap f^{-k}(y)$, completing the proof. □

Now we have all the tools to prove Theorem 2.33, following the proof of Proposition 2.24. Thus we let $E \subset M$ be any f-invariant set with $m(E^c) > 0$, and aim to prove that $m(E) = 0$.

Start by constructing partitions $\mathcal{P}_n$ exactly as in Step 1 there: obtain X_j as in (2.9)–(2.10), and let $\mathcal{P}_n$ be the partition of M into sets of the form $b_w(X_j)$, where b_w is the inverse branch of f^n associated to the word $w \in S^n$.

Next, observe that the argument in Step 2 of that proof works just as well here, since it requires only that

$$\max\{\operatorname{diam} C : C \in \mathcal{P}_n\} \to 0 \text{ as } n \to \infty,$$

which remains true in the nonlinear expanding setting. Thus there exist partition elements $C_n \in \mathcal{P}_n$ such that

$$\frac{m(E \cap C_n)}{m(C_n)} \to 0 \text{ as } n \to \infty. \tag{2.25}$$

The first part of Step 3 also goes through, giving for each $n \in \mathbb{N}$ some $j(n) \in \{1, \dots, \ell\}$ such that

$$\frac{m(X_{j(n)} \cap E)}{m(X_{j(n)})} \le \frac{m(f^n(C_n \cap E))}{m(f^n(C_n))}.$$

Now in place of the equality in (2.11), we use the bounded distortion property from Proposition 2.30, which relied on the $C^{1+\alpha}$ property of f, to conclude that

$$\frac{m(X_{j(n)} \cap E)}{m(X_{j(n)})} \le Q \frac{m(C_n \cap E)}{m(C_n)},$$

where Q is independent of n. Combined with (2.25), this implies that there exists $i \in \{1, \dots, \ell\}$ such that $m(X_i \cap E) = 0$.

It follows from Exercise 2.8 that $m(f^k(X_i \cap E)) = 0$ for every $k \in \mathbb{N}$. By invariance of E and E^c (and surjectivity of f), we have

$$f^k(X_i) \cap E = f^k(X_i) \setminus E^c f^k(X_i) \setminus f^k(E^c) \subset f^k(X_i \setminus E^c) = f^k(X_i \cap E),$$

so $m(f^k(X_i) \cap E) = 0$ for every k. By Lemma 2.36, f is topologically exact, so there exists k such that $f^k(X_i) = M$. We conclude that $m(E) = 0$, which completes the proof of Theorem 2.33.

2.6. The Ruelle–Perron–Frobenius theorem

Now we prove existence of an ACIP for a $C^{1+\alpha}$ expanding map $f\colon M \to M$. By Exercise 2.7, f_* maps $\mathcal{M}_{\text{ac}}$ to itself, and since the Radon–Nikodym Theorem allows us to identify $\mathcal{M}_{\text{ac}}$ with the set of densities

$$\Delta := \Big\{ h \in L^1(m) : h \ge 0, \int_M h \, dm = 1 \Big\}, \tag{2.26}$$

we can define the *Ruelle–Perron–Frobenius transfer operator* $L\colon \Delta \to \Delta$ by the condition that the diagram in (2.16) commutes, so that

$$\text{if } h = \frac{d\mu}{dm}, \text{ then } Lh = \frac{d(f_*\mu)}{dm}. \tag{2.27}$$

▶ Exercise 2.9. Prove that the RPF operator is characterized by the fact that for every $\varphi \in L^1(m)$, we have

$$\int \varphi \cdot (Lh) \, dm = \int (\varphi \circ f) \cdot h \, dm. \tag{2.28}$$

(In fact, it suffices to have (2.28) for every continuous φ.)

REMARK 2.37. Recalling that the pullback map $\varphi \mapsto \varphi \circ f$ is called the *Koopman operator*, we see from (2.28) that the RPF operator and the Koopman operator are dual to each other. The Koopman operator treats functions as observables (measurements), while the RPF operator treats them as densities (probability distributions).

Lec 35
Mon, Apr 14

Thus we can produce an ACIP by finding a fixed point of the RPF operator L. As with the corresponding question for Markov measures – namely, existence of a probability eigenvector – this can be done either by a "pushforward and average" procedure (see the paragraph following the proof of Lemma 1.135, or the proof of the Krylov–Bogolyubov Theorem 2.3), or by finding an appropriate subset of Δ on which L acts as a contraction in some metric.

In both approaches, the first step is to find an explicit formula for L that can be used in place of the implicit definitions in (2.27) and (2.28). To this end, let $A \subset M$ have diameter less than ϵ, so we can continuously define inverse branches $b_i \colon A \to M$ for $1 \le i \le D$ for which $b_i^{-1} = f|_{b_i(A)}$, and recall that $J(y) = |\det Df(y)|$, so we have

$$\begin{aligned}(f_*\mu)(A) = \mu(f^{-1}A) &= \mu\Big(\bigsqcup_{i=1}^{D} b_i(A)\Big) = \sum_{i=1}^{D} \mu(b_i A) \\ &= \sum_{i=1}^{D} \int_{b_i A} h(y)\, dm(y) = \sum_{i=1}^{D} \int_A h(b_i x) |\det Db_i(x)|\, dm(x) \\ &= \int_A \sum_{i=1}^{D} \frac{h(b_i x)}{|\det Df(b_i x)|}\, dm(x) = \int_A \sum_{y \in f^{-1}x} \frac{h(y)}{J(y)}\, dm(x).\end{aligned}$$

From this last expression we deduce that

$$(Lh)(x) = \sum_{y \in f^{-1}(x)} \frac{h(y)}{J(y)} = \sum_{y \in f^{-1}(x)} h(y) e^{-\log J(y)}. \tag{2.29}$$

Iterating, we obtain

$$(L^n h)(x) = \sum_{y \in f^{-n}(x)} h(y) e^{-\mathsf{S}_n \log J(y)}. \tag{2.30}$$

Now we return to the pushforward and average procedure. Consider the sequence of measures

$$\mu_n := \frac{1}{n} \sum_{k=0}^{n-1} f_*^k m, \tag{2.31}$$

for which we have $\mu_n \ll m$ with

$$\frac{d\mu_n}{dm} = \frac{1}{n} \sum_{k=0}^{n-1} L^k \mathbf{1}. \tag{2.32}$$

THEOREM 2.38. *Let M be a compact connected Riemannian manifold with normalized Lebesgue measure m, and let $f\colon M \to M$ be a $C^{1+\alpha}$ expanding map. Then the measures μ_n defined in (2.31) converge in the weak* topology to an ergodic invariant measure μ, which is the unique ACIP.*

PROOF. It suffices to prove that if $n_j \nearrow \infty$ is such that $\mu_{n_j} \to \mu$ in the weak* topology, then $\mu \ll m$. Indeed, weak* compactness of $\mathcal{M}$ guarantees that some such subsequence converges, and once we know that the limit must be an ACIP, Corollary 2.34 guarantees that it is ergodic and the unique ACIP. This uniqueness in turn guarantees that any subsequence μ_{n_j} has a subsubsequence converging to μ, from which we deduce that the whole sequence converges to μ.

Thus we prove that $\mu = \lim_j \mu_{n_j} \ll m$. This requires us to control the densities $L^k\mathbf{1}$. By the bounded distortion result in Proposition 2.30, there exists $K > 0$ such that for every $y \in M$, $n \in \mathbb{N}$, and $y' \in B_n(y, \epsilon)$, we have

$$|\mathsf{S}_n \log J(y) - \mathsf{S}_n \log J(y')| \le K. \tag{2.33}$$

Now given any $x \in M$ and any $x' \in B(x, \epsilon)$, there is a bijective correspondence $\pi\colon f^{-n}(x) \to f^{-n}(x')$ with the property that $\pi(y) \in B_n(y, \epsilon)$ for every $y \in f^{-n}(x)$; to see this, it suffices to enumerate the inverse branches of f^n as $\{b_j\}_{j=1}^{D^n}$ and put $\pi(b_j x) = b_j(x')$. Using this correspondence, we have

$$L^n\mathbf{1}(x') = \sum_{y \in f^{-n}x} e^{-\mathsf{S}_n \log J(\pi y)} \le \sum_{y \in f^{-n}x} e^K e^{-\mathsf{S}_n \log J(y)} = e^K L^n\mathbf{1}(x).$$

Since M is compact and connected, there exists $\ell \in \mathbb{N}$ such that given any $x, y \in M$, there is a sequence of points $x_0, x_1, \dots, x_\ell$ with $x_0 = x$ and $x_\ell = y$ satisfiying $d(x_{i-1}, x_i) < \epsilon$ for every i. This implies that

$$L^n\mathbf{1}(y) = L^n\mathbf{1}(x) \prod_{i=1}^{\ell} \frac{L^n\mathbf{1}(x_i)}{L^n\mathbf{1}(x_{i-1})} \le e^{\ell K} L^n\mathbf{1}(x).$$

Writing $Q = e^{\ell K}$ and observing that $\inf L^n\mathbf{1} \le 1$ because $\int L^n\mathbf{1}\, dm = 1$, we conclude that

$$L^n\mathbf{1}(x) \le Q \text{ for all } x \in M. \tag{2.34}$$

Since $L^n\mathbf{1} = \frac{d(f^n_* m)}{dm}$, we conclude that for every measurable $E \subset M$, and every nonnegative measurable function φ on M, we have

$$(f^n_* m)(E) \le Q m(E) \quad \text{and} \quad \int \varphi\, d(f^n_* m) \le Q \int \varphi\, dm. \tag{2.35}$$

This implies that $\int \varphi\, d\mu_n \le Q \int \varphi\, dm$ for every continuous φ, and passing to the weak* limit $\mu_{n_j} \to \mu \in \mathcal{M}_f$, we have

$$\int \varphi\, d\mu = \lim_{j\to\infty} \int \varphi\, d\mu_{n_j} \le Q \int \varphi\, dm. \tag{2.36}$$

Given any closed $A \subset M$, there are continuous functions $\varphi_k \searrow \mathbf{1}_A$, and thus

$$\mu(A) = \int \mathbf{1}_A \, d\mu = \lim_{k\to\infty} \int \varphi_k \, d\mu \le Q \lim_{k\to\infty} \int \varphi_k \, dm = Qm(A). \tag{2.37}$$

Now given any measurable $E \subset M$ with $m(E) = 0$, (2.37) gives $\mu(A) = 0$ for every closed $A \subset E$, and thus

$$\mu(E) = \sup\{\mu(A) : A \subset E \text{ is closed}\} = 0.$$

We conclude that $\mu \ll m$, as claimed. This proves Theorem 2.38. □

In fact, Theorem 2.38 can be strengthened in several ways:

- the invariant density $h = \frac{d\mu}{dm}$ is Hölder continuous, not just measurable;
- the sequence $f_*^n m$ converges to μ, even without averaging as in (2.31);
- the convergence happens exponentially fast.

Let us continue to use the metric provided by Proposition 2.32, and assume that d witnesses expansion in a single step: there exists $\chi \in (0,1)$ and $\epsilon > 0$ such that $d(x,y) \le \chi d(fx, fy)$ whenever $y \in B(x,\epsilon)$.

Lec 36
Wed, Apr 16

Assume that both $\log J$ and $\log h$ are locally Hölder continuous with exponent α, so that there exist constants C_J and C_h such that given any $x \in M$ and $y \in B(x,\epsilon)$, we have

$$|\log J(x) - \log J(y)| \le C_J d(x,y)^\alpha \quad \text{and} \quad |\log h(x) - \log h(y)| \le C_h d(x,y)^\alpha.$$

Our goal is to get an upper bound on the Hölder constant of $\log Lh$. Given $x \in M$ and $x' \in B(x,\epsilon)$, let $\pi\colon f^{-1}(x) \to f^{-1}(x')$ be the pairing described after (2.33). Then we have

$$h(\pi y) = e^{\pm C_h d(y,\pi y)^\alpha} h(y) = e^{\pm C_h \chi^\alpha d(x,x')^\alpha} h(y),$$
$$J(\pi y) = e^{\pm C_J d(y,\pi y)^\alpha} J(y) = e^{\pm C_J \chi^\alpha d(x,x')^\alpha} h(y).$$

From these we deduce that

$$\frac{Lh(x)}{Lh(x')} = \frac{\sum_{y\in f^{-1}x} h(y)/J(y)}{\sum_{y\in f^{-1}x} h(\pi y)/J(\pi y)} = e^{\pm (C_h + C_J)\chi^\alpha d(x,x')^\alpha}.$$

In particular, $\log Lh$ is Hölder continuous with constant $(C_h + C_J)\chi^\alpha$, and we have proved the following result.

LEMMA 2.39. *Let M be a compact connected Riemannian manifold, and $f\colon M \to M$ a $C^{1+\alpha}$ expanding map. Let $J = |\det Df|$, and let C_J be the α-Hölder constant of $\log J$.*

Given $K > 0$, let Ω_K be the set of functions $\varphi\colon M \to (0,\infty)$ such that $\log\varphi$ is (K,α)-Hölder continuous. Then

$$L\Omega_K \subset \Omega_{(K+C_J)\chi^\alpha}. \tag{2.38}$$

If K is sufficiently large (an elementary calculation shows that $K > C_J/(\chi^{-\alpha}-1)$ will do the job), then Lemma 2.39 shows that L maps Ω_K into itself, so that $L^k\mathbf{1} \in \Omega_K$ for every $k \in \mathbb{N}$. It follows that the densities $\frac{d\mu_n}{dm}$ are equicontinuous.

▶ EXERCISE 2.10. Use the preceding paragraph to prove that the unique ACIP μ has $\frac{d\mu}{dm} \in \Omega_K$.

In fact, the space Ω_K defined in Lemma 2.39 provides us with a setting where Birkhoff's Contraction Theorem can be applied. First we define the metric, recalling that $\Omega = \Omega_K$ carries a partial order defined by writing $g \preceq h$ if and only if $h - g \in \Omega$. Following (1.147)–(1.149), given $g, h \in \Omega = \Omega_K$, we write

$$\alpha = \sup\{t \geq 0 : h - tg \in \Omega\} = \sup\{t > 0 : tg \preceq h\},$$
$$\beta = \inf\{t > 0 : tg - h \in \Omega\} = \inf\{t > 0 : h \preceq tg\}.$$

Recall the interpretation of this in Figure 1.58. Then the Hilbert metric on $\Omega = \Omega_K$ is

$$\Theta(g, h) = \Big| \log \frac{\beta(g,h)}{\alpha(g,h)} \Big|. \tag{2.39}$$

One can prove that Θ gives a complete metric on the projectivization of Ω_K, although we omit the details here.[6]

Moreover, following exactly the same argument as in Theorem 1.140, one can use Theorem 1.139 to deduce the following:

THEOREM 2.40. *If $L\colon C(M) \to C(M)$ is a linear map that maps Ω_K into itself and has the property that $A := \operatorname{diam}_{\Theta_K}(L\Omega_K) < \infty$, then L contracts Θ_K-distances by a factor of $\tanh(A/4)$.*

Thanks to (2.38) and Theorem 2.40, in order to deduce that L is a contraction on Θ_K for sufficiently large K, our only remaining task is to prove the following.

LEMMA 2.41. *Given $K > 0$, let Θ_K be the Hilbert metric on Ω_K. Then given any $R \in (0, K)$, we have $\operatorname{diam}_{\Theta_K} \Omega_R < \infty$.*

PROOF. Let $\mathcal{P} = \{(x, y) : x, y \in M,\ y \neq x,\ d(x, y) < \epsilon\}$. Given a pair $(x, y) \in \mathcal{P}$, consider the linear functional $\ell_{x,y}\colon C(M) \to \mathbb{R}$ given by

$$\ell_{x,y}(g) = e^{Kd(x,y)^\alpha} g(x) - g(y). \tag{2.40}$$

Then the cone Ω_K is characterized as

$$\Omega_K = \{g \in C(M) \setminus \{\mathbf{0}\} : g \geq 0 \text{ and } \ell_{x,y}(g) \geq 0 \text{ for all } (x, y) \in \mathcal{P}\}. \tag{2.41}$$

Note that membership in Ω_K also requires $g > 0$ everywhere, but this is guaranteed by the conditions in (2.41): by compactness and connectedness of M, there exists $q \in \mathbb{N}$ such that given any $x, y \in M$, there exist points $\{x_i\}_{i=0}^q$ such that $x_0 = x$, $x_q = y$, and $d(x_i, x_{i+1}) < \epsilon$ for every i, and thus every g satisfying the conditions in (2.41) also has the property that

$$g(y) \leq Qg(x) \text{ for every } x, y \in M, \text{ where } Q = qK\epsilon^\alpha. \tag{2.42}$$

Since g is not the zero function, this implies that $g > 0$ everywhere.

[6] See Proposition 12.3.3 of the book by Viana and Oliveira, for example.

Now given $g, h \in \Omega_R$, we want to bound $\Theta_{\Omega_K}(g,h)$. To this end, we first compute $\alpha(g,h)$ and $\beta(g,h)$ by using the functionals $\ell_{x,y}$ to characterize the set of $t \geq 0$ such that $h - tg \in \Omega_K$ and $tg - h \in \Omega_K$. Given $(x,y) \in \mathcal{P}$, we have

$$\ell_{x,y}(h - tg) = \ell_{x,y}(h) - t\ell_{x,y}(g).$$

Using (2.41), we see that

$$\begin{aligned} h - tg \in \Omega_K \quad &\Leftrightarrow \quad \ell_{x,y}(h) \geq t\ell_{x,y}(g) \text{ for all } (x,y) \in \mathcal{P}, \text{ and} \\ tg - h \in \Omega_K \quad &\Leftrightarrow \quad \ell_{x,y}(h) \leq t\ell_{x,y}(g) \text{ for all } (x,y) \in \mathcal{P}. \end{aligned}$$

It follows that $\alpha(g,h)$ and $\beta(g,h)$ are the infimum and supremum, respectively, of the set

$$\left\{ \frac{\ell_{x,y}(h)}{\ell_{x,y}(g)} : (x,y) \in \mathcal{P} \right\}. \tag{2.43}$$

Recalling (2.39), we conclude that

$$\Theta_K(g,h) = \log \sup \left\{ \frac{\ell_{x,y}(h)}{\ell_{x,y}(g)} \frac{\ell_{x',y'}(g)}{\ell_{x',y'}(h)} : (x,y), (x',y') \in \mathcal{P} \right\}. \tag{2.44}$$

Lec 37
Fri, Apr 18

To bound $\ell_{x,y}(h)$ for $(x,y) \in \mathcal{P}$, let $\rho := d(x,y)^\alpha \in [0, \epsilon^\alpha)$, and observe that since $h \in \Omega_R$, we have $h(y) \geq e^{-R\rho} h(x)$, from which we obtain

$$\ell_{x,y}(h) = e^{K\rho} h(x) - h(y) \leq h(x)(e^{K\rho} - e^{-R\rho}). \tag{2.45}$$

Similarly, $g(y) \leq e^{R\rho} g(x)$, so

$$\ell_{x,y}(g) \geq g(x)(e^{K\rho} - e^{R\rho}). \tag{2.46}$$

Combining (2.45) and (2.46) gives

$$\frac{\ell_{x,y}(h)}{\ell_{x,y}(g)} \leq \frac{h(x)(e^{K\rho} - e^{-R\rho})}{g(x)(e^{K\rho} - e^{R\rho})} \leq s \frac{h(x)}{g(x)}, \tag{2.47}$$

where

$$s := \sup_{\rho \in (0, \epsilon^\alpha)} \frac{e^{K\rho} - e^{R\rho}}{e^{K\rho} - e^{-R\rho}}. \tag{2.48}$$

Since $K > R$, the function inside the supremum is continuous on $(0, \infty)$. It approaches $\frac{K+R}{K-R}$ as $\rho \to 0$ (as a first-order Taylor expansion of the numerator and denominator reveals), and we conclude that $s < \infty$. A similar inequality to (2.47) holds with the roles of h, g reversed, and thus (2.44) gives

$$\Theta_K(g,h) \leq \log \left(s^2 \sup_{x,x' \in M} \frac{h(x)}{g(x)} \frac{g(x')}{h(x')} \right). \tag{2.49}$$

By (2.42), we have $\frac{g(x')}{g(x)} \leq Q$, and similarly $\frac{h(x)}{h(x')} \leq Q$, so

$$\Theta_K(g,h) \leq s^2 Q^2. \tag{2.50}$$

This holds for every $g, h \in \Omega_R$, and completes the proof of Lemma 2.41. □

▶▶ EXERCISE 2.11. Prove that the function inside the supremum in (2.48) is decreasing in ρ, and thus $s = \frac{K+R}{K-R}$.

Combining (2.38), Theorem 2.40, and Lemma 2.41, we conclude that $\operatorname{diam}_{\Theta_K}(L\Omega_K) < \infty$, and thus there exists $\gamma \in (0,1)$ such that for every $g, h \in \Omega_K$, we have

$$\Theta_K(Lg, Lh) \le \gamma \Theta_K(g, h). \tag{2.51}$$

REMARK 2.42. Using the explicit expressions in (2.38), Theorem 2.40, and Exercise 2.11, we in fact have $\gamma = \tanh((\frac{Q(K+R)}{2(K-R)})^2)$, where $R = (K + C_J)\chi^\alpha$ and $Q = qK\epsilon^\alpha$, with $q \approx \operatorname{diam} M/\epsilon$ (see the paragraph before (2.42) for the precise definition).

Using (2.51), one can prove the following (we omit the detailed argument), which should be compared to the Perron–Frobenius Theorem 1.141 for primitive nonnegative matrices.

THEOREM 2.43 (Ruelle–Perron–Frobenius Theorem for Lebesgue measure). *Let M be a compact connected Riemannian manifold with normalized Lebesgue measure m, and let $f\colon M \to M$ be a $C^{1+\alpha}$ expanding map. Let $L\colon C(M) \to C(M)$ be the RPF operator. Then the following are true.*

(1) There exists a unique $h \in C(M)$ such that $Lh = h$, and the measure μ defined by $\mu(E) = \int_E h\, dm$ is the unique ACIP;

(2) There exist constants $C > 0$ and $\gamma \in (0,1)$ such that given any locally $(1,\alpha)$-Hölder continuous[7] function $g\colon M \to \mathbb{R}$, writing $m(g) = \int g\, dm$, we have

$$\|L^n g - m(g)h\| \le C\gamma^n \quad \text{for all } n \ge 0. \tag{2.52}$$

In particular, applying (2.52) to $g = 1$ gives a precise sense in which $f_*^n m \to \mu$ exponentially fast.

2.7. Comparing the ACIP to Markov measures

Let us adopt the following notational convention: given quantities X, Y that may depend on n, x, ϵ, we will write $X \asymp Y$ if there exists $Q = Q(\epsilon) > 0$ such that $Q^{-1}Y \le X \le QY$ for every x, n. Now we make the following observations regarding the unique ACIP μ for an expanding map f on a compact connected manifold M.

▶ EXERCISE 2.12. Given $x \in M$ and $n \in \mathbb{N}$, prove that f^n is injective on the corresponding Bowen ball

$$B_n(x,\epsilon) = \{y \in M : d(f^k y, f^k x) < \epsilon \text{ for all } 0 \le k < n\}$$

(recall Definition 2.9), and that

$$f^n(B_n(x,\epsilon)) = B(f^n(x),\epsilon) \quad \text{and} \quad B_n(x,\epsilon) = b_{x,n}(B(f^n(x),\epsilon)),$$

where $b_{x,n}$ is the inverse branch of f^n defined in a neighborhood of x.

[7] That is, $|g(x) - g(y)| \le d(x,y)^\alpha$ for all $x, y \in M$ with $d(x,y) < \epsilon$.

Using Exercise 2.12 and the fact that $Df^n(z) = e^{\mathsf{S}_n \log J(z)}$, we obtain

$$\begin{aligned}\mu(B_n(x,\epsilon)) &= \int_{B_n(x,\epsilon)} h(y)\, dm(y) \\ &= \int_{f^n(B_n(x,\epsilon))} h(b_{x,n}(z))|\det Db_{x,n}(z)|\, dm(z) \\ &= \int_{B(f^n(x),\epsilon)} h(b_{x,n}(z)) e^{-\mathsf{S}_n \log J(b_{x,n}(z))}\, dm(z) \\ &\asymp e^{-\mathsf{S}_n \log J(x)} m(B(f^n(x),\epsilon)),\end{aligned}$$

where the last line uses the fact that $0 < \inf h < \sup h < \infty$ and the bounded distortion property of $\mathsf{S}_n \log J$. Moreover, since $z \mapsto m(B(z,\epsilon))$ is a continuous function on the compact set M, it is bounded away from 0 and ∞, and we conclude that

$$\mu(B_n(x,\epsilon)) \asymp e^{-\mathsf{S}_n \log J(x)}. \tag{2.53}$$

We can make a similar observation regarding Markov measures. Given a $q \times q$ primitive stochastic matrix P and writing $S = \{1,\dots,q\}$, let us define a function $\varphi\colon S^{\mathbb{N}_0} \to \mathbb{R} \cup \{-\infty\}$ by

$$\varphi(x) = \log P_{x_0 x_1}. \tag{2.54}$$

Then we see that

$$P_{x_0x_1} P_{x_1x_2} \cdots P_{x_{n-1}x_n} = e^{\mathsf{S}_n \varphi(x)}.$$

Moreover, if we choose $\epsilon > 0$ such that $d(x,y) < \epsilon$ if and only if $x_0x_1 = y_0y_1$, then we see that

$$B_n(x,\epsilon) = [x_0x_1\cdots x_n].$$

Writing π for the unique left probability eigenvector of P, and $\mu_{P,\pi}$ for the corresponding invariant Markov measure, we have

$$\mu_{P,\pi}(B_n(x,\epsilon)) = \pi_{x_0} P_{x_0x_1} P_{x_1x_2} \cdots P_{x_{n-1}x_n} = \pi_{x_0} e^{\mathsf{S}_n \varphi(x)} \asymp e^{\mathsf{S}_n \varphi(x)}. \tag{2.55}$$

Observe that both (2.53) and (2.55) can be written in the form

$$\mu(B_n(x,\epsilon)) \asymp e^{\mathsf{S}_n \varphi(x)} \tag{2.56}$$

for an appropriate *potential function* φ, which in the case of the absolutely continuous invariant measure is the *geometric potential* $\varphi(x) = -\log J(x)$. Eventually we will see how (2.56) is a special case of a more general *Gibbs property*. For now, we explore further connections between the ACIP and Markov measures.

Lec 38
Mon, Apr 21

One such connection is between the stochastic matrix P and the RPF operator L. In both cases, the question of finding an invariant probability measure within the equivalence class of interest reduced to the problem of finding a fixed point of the appropriate linear operator. In fact, both can be viewed as specific cases of a more general Ruelle–Perron–Frobenius operator.

Given an expanding map f, let $\varphi(x) = -\log J(x) = -\log|\det Df(x)|$. Then the RPF operator can be written as

$$(Lg)(x) = \sum_{y\in f^{-1}(x)} g(y)e^{\varphi(y)}. \tag{2.57}$$

Now given a $q \times q$ stochastic matrix P, let $\varphi(x) = \log P_{x_0x_1}$ as in (2.54). Embed the set Δ^{q-1} of probability vectors into $C(S^{\mathbb{N}_0})$ by associating to each $v \in \Delta^{q-1}$ the function τv that depends only on the first coordinate:

$$(\tau v)(x) = v_{x_0}.$$

Then observe that τv and $\tau(vP)$ are related as follows:

$$\tau(vP)(x) = (vP)_{x_0} = \sum_{a\in S} v_a P_{ax_0} = \sum_{a\in S} (\tau v)(ax)e^{\varphi(ax)}.$$

Defining L as in (2.57), we see that $\tau(vP) = L(\tau v)$, so the following diagram commutes.

$$\begin{array}{ccc} \Delta^{q-1} & \xrightarrow{\cdot P} & \Delta^{q-1} \\ \downarrow{\scriptstyle \tau} & & \downarrow{\scriptstyle \tau} \\ C(S^{\mathbb{N}_0}) & \xrightarrow{L} & C(S^{\mathbb{N}_0}) \end{array} \tag{2.58}$$

In other words, right-multiplication by the stochastic matrix P represents the restriction of the appropriate RPF operator to the subspace of functions that depend only on the first coordinate.

To develop this point of view further and describe a result that generalizes both the Perron–Frobenius Theorem 1.141 and the Ruelle–Perron–Frobenius Theorem 2.43 for Lebesgue measure, we need to restrict to a subset of the full shift. Given a $q \times q$ stochastic matrix P, define a $q \times q$ matrix A with entries in $\{0, 1\}$ by

$$A_{ij} = \begin{cases} 1 & \text{if } P_{ij} > 0, \\ 0 & \text{if } P_{ij} = 0, \end{cases} \tag{2.59}$$

and then consider the set

$$X_A := \{x \in S^{\mathbb{N}_0} : A_{x_nx_{n+1}} = 1 \text{ for all } n \geq 0\}. \tag{2.60}$$

This is an example of a *subshift of finite type (SFT)*.

▶ Exercise 2.13. Prove that X_A is compact, σ-invariant (meaning that $\sigma(A) \subset A$), and that if P is primitive, then $X_A = \operatorname{supp}\mu$, where $\mu = \mu_{P,\pi}$ is the invariant Markov measure determined by P and its Perron–Frobenius probability eigenvector $\pi = \pi P$.

Remark 2.44. It is often useful to interpret X_A in terms of a directed graph with vertex set $S = \{1, \dots, q\}$, where there is an edge from $i \in S$ to $j \in S$ if and only if $A_{ij} = 1$. Then X_A is naturally identified with the space of all infinite walks on this graph.

▶ EXERCISE 2.14. Prove that if A is a $q \times q$ matrix with entries in $\{0,1\}$, and X_A is the associated SFT, then $\sigma\colon X_A \to X_A$ is expanding and locally onto. If in addition A is primitive, prove that σ is topologically exact.

From now on we let X be a compact metric space and $f\colon X \to X$ a continuous map that is expanding, locally onto, and topologically exact. Observe that this class of systems includes both the examples we have focused on so far: expanding maps on connected compact Riemannian manifolds, and SFTs.

Given a continuous potential function $\varphi\colon X \to \mathbb{R}$, the *Ruelle–Perron–Frobenius (RPF) operator* associated to φ is the map $L_\varphi\colon C(X) \to C(X)$ defined as in (2.57) by

$$(L_\varphi g)(x) = \sum_{y \in f^{-1}(x)} g(y) e^{\varphi(y)}. \tag{2.61}$$

We would like to interpret L_φ as acting on a space of densities, but in order to do this, we need an appropriate reference measure, such as Lebesgue measure for expanding maps on manifolds. This reference measure should bear a similar relationship to the potential function φ as Lebesgue measure bears to the geometric potential $-\log J$. In the case of Lebesgue measure, the key fact was the change of variables formula: when $E \subset M$ has small diameter (so that $f|_E$ is injective), we have

$$m(f(E)) = \int_E |\det Df(x)|\, dm(x) = \int_E e^{\log J(x)}\, dm(x) = \int_E e^{-\varphi(x)}\, dm(x).$$

Equivalently, when restricting to small sets on which f is injective, we have $m \circ f \ll m$, with

$$\frac{d(m \circ f)}{dm} = e^{-\varphi} \quad \Rightarrow \quad \varphi = -\log \frac{d(m \circ f)}{dm}.$$

Thus we want a measure with respect to which φ behaves as a negative log Jacobian, just as the geometric potential does for Lebesgue. If m is such a measure for an expanding map f on a compact metric space X, then arguments similar to those at the beginning of §2.6 would show that given any $\psi \in C(X)$ and $g \in L^1(X,m)$, we have

$$\int \psi \cdot (L_\varphi g)\, dm = \int (\psi \circ f) \cdot g\, dm. \tag{2.62}$$

Thus m would be the measure with respect to which the RPF operator L_φ defined in (2.61) is dual to the Koopman operator. It is helpful to interpret m as a fixed point by observing that when $\psi \equiv 1$, (2.62) becomes

$$\int L_\varphi g\, dm = \int g\, dm \text{ for all } g \in L^1(X,m).$$

With this in mind, define a linear map $L_\varphi^*\colon C(X)^* \to C(X)^*$ on the space of bounded linear functionals on $C(X)$ by

$$(L_\varphi^* m)(g) = m(Lg).$$

If you prefer the "inner product" notation $\langle g, m\rangle = m(g)$, this becomes

$$\left\langle g, L_\varphi^* m\right\rangle = \langle Lg, m\rangle .$$

Recalling that the Riesz Representation Theorem identifies $C(X)^*$ with the space of finite signed measures on X, and abusing notation mildly by using m to represent both a measure and its corresponding linear functional, we see that the dual operator L_φ^* is characterized by

$$\int g\, d(L_\varphi^* m) = \int (L_\varphi g)\, dm. \tag{2.63}$$

By following ideas similar to those in the proof of the earlier Ruelle–Perron–Frobenius Theorem 2.43, one can obtain the following result, whose proof we omit.

THEOREM 2.45 (Ruelle–Perron–Frobenius Theorem). *Let X be a compact metric space and $f\colon X \to X$ a continuous map that is expanding, locally onto, and topologically exact. Then given any α-Hölder continuous potential $\varphi\colon X \to \mathbb{R}$, the associated RPF operator $L = L_\varphi$ and its dual L^* satisfy the following.*

(1) There exists a unique $\lambda > 0$ and Borel probability measure $m \in \mathcal{M}(X)$ such that $L^* m = \lambda m$.
(2) There exists a unique $\lambda > 0$ and $h \in C(X)$ such that $Lh = \lambda h$ and $\int h\, dm = 1$, where m is the eigenmeasure from Statement 1; moreover, the eigenvalue λ is the same as in Statement 1.
(3) The measure $\mu \in \mathcal{M}(X)$ defined by $\mu(E) = \int_E h\, dm$ is the unique f-invariant measure that is absolutely continuous with respect to m.
(4) The eigenmeasure m gives positive weight to every open set, and the eigenfunction h is Hölder continuous and positive everywhere.
(5) There exist constants $C > 0$ and $\gamma \in (0, 1)$ such that given any locally $(1, \alpha)$-Hölder continuous function $g\colon M \to \mathbb{R}$, we have

$$\|L^n g - m(g)h\| \le C\gamma^n \quad \text{for all } n \ge 0. \tag{2.64}$$

Lec 39
Wed, Apr 23

Now one can follow arguments similar to those at the beginning of this section and conclude that if μ is the RPF measure provided by Theorem 2.45, then we have the following more general version of (2.56):

$$\mu(B_n(x, \epsilon)) \asymp \lambda^{-n} e^{\mathsf{S}_n \varphi(x)}. \tag{2.65}$$

It is customary to write $P = \log \lambda$ and make the following definition.

DEFINITION 2.46. A measure μ is a *Gibbs measure* for $f\colon X \to X$ and $\varphi\colon X \to \mathbb{R}$ at scale $\epsilon > 0$ if there exists a constant $Q > 0$ such that given any $x \in X$ and $n \in \mathbb{N}$, we have

$$Q^{-1} \le \frac{\mu(B_n(x, \epsilon))}{e^{-nP + \mathsf{S}_n \varphi(x)}} \le Q. \tag{2.66}$$

We conclude that every RPF measure for a Hölder continuous potential is also a Gibbs measure for that potential. In fact, the converse is also true, but we will not prove this.

2.8. Markov measures as RPF measures

Let P be a $q \times q$ primitive stochastic matrix. Let A be the associated 0-1 transition matrix as in (2.59), and $X = X_A$ the corresponding SFT as in (2.60). In the previous section we observed that writing $\varphi(x_0x_1x_2\cdots) = \log P_{x_0x_1}$, the corresponding RPF operator $L = L_\varphi$ preserves the subspace of $C(S^{\mathbb{N}_0})$ comprising functions that depend only on the first symbol, and that its action on this space was given in terms of right-multiplication by P.

What about the dual operator L^*? Recalling (2.63), given any measure m on X and any legal word $w \in S^k$, we have

$$(L^*m)[w] = \int \mathbf{1}_{[w]}\, d(L^*m) = \int L\mathbf{1}_{[w]}\, dm. \tag{2.67}$$

Writing $i \to j$ when $A_{ij} = 1$, we can write

$$L\mathbf{1}_{[w]}(x) = \sum_{a \to x_0} \mathbf{1}_{[w]}(ax) e^{\varphi(ax)} = \sum_{a \in S} \mathbf{1}_{[w]}(ax) P_{ax_0}. \tag{2.68}$$

Observe that $\mathbf{1}_{[w]}(ax) = 0$ for all $a \neq w_1$, so the sum can have at most one nonzero term. If $k > 1$, we write $\sigma w := w_2w_3\cdots w_k$ and use (2.67) to obtain

$$L\mathbf{1}_{[w]}(x) = \mathbf{1}_{[\sigma w]}(x) P_{w_1w_2} \quad \Rightarrow \quad (L^*m)[w] = P_{w_1w_2} m[\sigma w]. \tag{2.69}$$

If $k = 1$ so that $w = w_k$ is a single symbol, we have $L\mathbf{1}_{[w]}(x) = P_{w_kx_0}$, and thus

$$(L^*m)[w] = \int P_{w_kx_0}\, dm(x) = \sum_{j \in S} P_{w_kj} m[j]. \tag{2.70}$$

From (2.69) we see that the weight of L^*m on k-cylinders is determined by the weight of m on $(k-1)$-cylinders whenever $k \geq 2$, and from (2.69) we see that if we associate to each measure m the vector of weights it gives 1-cylinders, then L^* acts on these vectors via left-multiplication by P.

In particular, if m is an eigenmeasure, so that $L^*m = m$ (recall that $\lambda = 1$ since P is a stochastic matrix), then for every $w \in S^k$ with $k \geq 2$, (2.69) gives

$$m[w] = P_{w_1w_2} m[\sigma w] = \cdots = P_{w_1w_2} P_{w_2w_3} \cdots P_{w_{k-1}w_k} m[w_k]. \tag{2.71}$$

Moreover, since the vector whose entries are all equal to 1 is a right eigenvector for P (since it is a stochastic matrix), we see that

$$m[w] := P_{w_1w_2} P_{w_2w_3} \cdots P_{w_{k-1}w_k} \tag{2.72}$$

is an eigenmeasure of L^*. Recalling (1.129), we see that our discussion in §1.13 of "measures consistent with the Markov process" can be reframed as "measures absolutely continuous with respect to the eigenmeasure m".

Having considered potential functions given in terms of Jacobians and of transition probabilities, let us consider the simplest possible potential function: $\varphi = 0$.

Continuing to work on the SFT $X = X_A$ associated to a primitive 0-1 matrix A, we see that (2.69) and (2.70) are replaced by the following for a legal word $w \in S^k$:

$$(L^*m)[w] = \begin{cases} A_{w_1w_2} m[\sigma w] & \text{if } k > 1, \\ \sum_{j \in S} A_{w_k j} m[j] & \text{if } k = 1. \end{cases} \tag{2.73}$$

By the Perron–Frobenius Theorem, there exist left and right eigenvectors ℓ and r, normalized so that $\sum_i \ell_i r_i = 1$, and an eigenvalue λ, such that

$$\ell A = \lambda \ell, \qquad Ar = \lambda r. \tag{2.74}$$

Following the same reasoning as above, to get an eigenmeasure m satisfying $L^*m = \lambda m$, we can take $m[i] = r_i$ for each $i \in S$, and define $m[w]$ for $w \in S^k$ with $k > 1$ by the condition that

$$m[w] = \lambda^{-1}(L^*m)[w] = \lambda^{-1} A_{w_1w_2} m[\sigma w].$$

If w is a legal word, then $A_{w_1w_2} = 1$, so we have

$$m[w] = \lambda^{-1} m[\sigma w], \tag{2.75}$$

and iterating this $(k-1)$ times gives

$$m[w] = \lambda^{-(k-1)} r_{w_k}. \tag{2.76}$$

For the eigenfunction, we can once again observe (just as we did for Markov measures) that the RPF operator L preserves the space of functions that depend only on the first symbol, and that it acts on this space as right multiplication by A, so the eigenfunction is

$$h(x) = \ell_{x_0}. \tag{2.77}$$

In particular, given $w \in S^k$ and $x \in [w]$, we have $h(x) = \ell_{w_1}$, so the RPF measure for $\varphi = 0$ on $X = X_A$ is given by

$$\mu[w] = \ell_{w_1} r_{w_k} \lambda^{-(k-1)}. \tag{2.78}$$

This is the *Parry measure* associated to the SFT X_A.

The Parry measure is a Gibbs measure: indeed, writing

$$C = \max\{\ell_i r_j : i, j \in S\} \quad \text{and} \quad Q = \lambda \max(C, C^{-1}),$$

we see from (2.78) that for every $k \in \mathbb{N}$ and $w \in S^k$, we have

$$\mu[w] = Q^{\pm} \lambda^{-k} = Q^{\pm 1} e^{-kP}, \quad \text{where } P = \log \lambda. \tag{2.79}$$

Moreover, the Parry measure is in fact a Markov measure, although this is not obvious from (2.78). Consider the matrix P and row vector π defined by

$$P_{ij} = \frac{A_{ij} r_j}{\lambda r_i}, \quad \pi_i = \ell_i r_i. \tag{2.80}$$

Then π is a probability vector by our normalization of ℓ and r. We claim that P is a stochastic matrix and that π is its left probability eigenvector. Indeed, a simple computation shows that for every $i \in S$, we have

$$\sum_{j \in S} P_{ij} = \frac{\sum_{j \in S} A_{ij} r_j}{\lambda r_i} = \frac{(Ar)_i}{\lambda r_i} = 1,$$

since $Ar = \lambda r$, and similarly, for every $j \in S$,

$$(\pi P)_j = \sum_{i \in S} \pi_i P_{ij} = \sum_{i \in S} \frac{\ell_i r_i A_{ij} r_j}{\lambda r_i} = \frac{\sum_{i \in S} \ell_i A_{ij} r_j}{\lambda} = \frac{(\ell A)_j r_j}{\lambda} = \ell_j r_j = \pi_j.$$

Now given any $k \in \mathbb{N}$ and $w \in S^k$, the Markov measure associated to P and π satisfies

$$\mu_{P,\pi}[w] = \pi_{w_1} \prod_{j=1}^{k-1} P_{w_j w_{j+1}} = \ell_{w_1} r_{w_1} \prod_{j=1}^{k-1} \frac{A_{w_j w_{j+1}} r_{w_{j+1}}}{\lambda r_{w_j}}.$$

If the word w is illegal for the SFT X_A, then one of the factors $A_{w_j w_{j+1}}$ will vanish, so $\mu_{P,\pi}[w] = 0$. Otherwise these factors are all equal to 1, and we have

$$\mu_{P,\pi}[w] = \ell_{w_1} r_{w_1} \prod_{j=1}^{k-1} \frac{r_{w_{j+1}}}{\lambda r_{w_j}} = \ell_{w_1} r_{w_k} \lambda^{-(k-1)} = \mu[w],$$

so $\mu_{P,\pi}$ is equal to the Parry measure μ.

2.9. Measure-theoretic entropy and the variational principle

The fact that the Parry measure is the Gibbs measure for $\varphi = 0$, as in (2.79), means that for every $k \in \mathbb{N}$, all legal words $w \in S^k$ have roughly the same weight, where "roughly" means "up to a multiplicative factor that is uniformly bounded in k". In fact, the Parry measure is the *only* invariant measure with this property, so it is the unique measure that strikes this particular balance between invariance (allowing us to use ergodic theory) and "maximum ignorance", in the sense that with no prior information about which legal words are most likely, we would like to assign an equal probability to all words of a given length (all outcomes of the first k experiments).

This can be formalized by the concept of an *information function*. Suppose we have a probability space and wish to assign to each event (measurable subset) a quantity I, which will represent the information we gain if we observe this event to occur. (Or if you prefer, how surprised we would be to observe the event.) It is reasonable to ask that I satisfy the following conditions.

- The information I depends only on the probability p that the event occurs: events with the same probability carry the same amount of information.
- The more likely the event, the less information it carries: $I\colon [0,1] \to [0,\infty]$ is strictly decreasing.
- Events with probability 1 carry no information: $I(1) = 0$.

- If we observe two independent events occuring, then we gain the same amount of information as if we observed them both separately and added the resulting information gains: $I(pq) = I(p) + I(q)$.

▶ EXERCISE 2.15. Show that every function satisfying all of these conditions has the form $I(p) = -\log_b p$ for some base $b > 1$.

Lec 40
Fri, Apr 25

With Exercise 2.15 in mind, we will consider the information function $I\colon (0,1] \to [0,\infty)$ given by[8]

$$I(p) = -\log p. \tag{2.81}$$

Now let $X = S^{\mathbb{N}}$ and fix a shift-invariant Borel probability measure $\mu \in \mathcal{M}_\sigma(X)$. Given $k \in \mathbb{N}$, we will write $H_k(\mu)$ for the "expected information gained from observing the first k symbols":

$$H_k(\mu) = \int_X I(\mu[x_1 \cdots x_n])\, d\mu(x). \tag{2.82}$$

Partitioning X into the cylinders $[w]$ for $w \in S^k$, we see that

$$H_k(\mu) = \sum_{w \in S^k} \int_{[w]} I(\mu[w])\, d\mu(x) = \sum_{w \in S^k} -\mu[w] \log \mu[w]. \tag{2.83}$$

The quantity $H_k(\mu)$ is the *(static) entropy* of the measure μ associated to the partition into k-cylinders. One can prove the following.

- $H_1(\mu) \le H_2(\mu) \le H_3(\mu) \le \cdots$.
- For every $n, k \in \mathbb{N}$, we have $H_{n+k}(\mu) \le H_n(\mu) + H_k(\mu)$, with equality if and only if for every $v \in S^n$ and $w \in S^k$, we have $\mu[vw] = \mu[v]\mu[w]$.

This last property is called *subadditivity* of the sequence $(H_k(\mu))_k$, and has the following important consequence.

LEMMA 2.47 (Fekete's Lemma). *Let $(a_n)_n$ be a sequence of real numbers such that $a_{n+k} \le a_n + a_k$ for all n, k. Then $\lim_{n\to\infty} \frac{1}{n} a_n$ exists and is equal to $\inf_n \frac{1}{n} a_n$.*

PROOF. Exercise. Start by observing that given any $n, k \in \mathbb{N}$, we have $H_{nk}(\mu) \le nH_k(\mu)$, so $\frac{1}{nk} H_{nk}(\mu) \le \frac{1}{k} H_k(\mu)$. This does not quite imply that $k \mapsto \frac{1}{k} H_k(\mu)$ is nonincreasing (and indeed monotonicity may fail), but gets you close enough that the rest is not so hard. □

From Fekete's Lemma, we see that the limit

$$h(\mu) := \lim_{n\to\infty} \frac{1}{n} H_n(\mu) \tag{2.84}$$

exists. This is the *(dynamical) measure-theoretic*[9] *entropy* of (X, σ, μ). It represents the (linear) rate at which we gain information (on average) as we make successive observations of the symbols of $x \in X$.

[8] We use the natural logarithmic base; it is also common to define information in base 2.

[9] The term *metric entropy* is sometimes used as well, despite that fact that we are not dealing with a metric space.

Now suppose that $X \subset S^{\mathbb{N}}$ is a compact shift-invariant subset (meaning that $\sigma X \subset X$); we say that X is a *shift space*. This includes the case where X is an SFT, but also includes many other examples besides.

For each $k \in \mathbb{N}$, consider the set

$$\mathcal{L}_k := \{w \in S^k : [w] \cap X \neq \emptyset\},$$

which comprises all legal words of length k. The *language* of X is

$$\mathcal{L} := \bigcup_{k \in \mathbb{N}} \mathcal{L}_k. \tag{2.85}$$

Let μ be a shift-invariant measure supported on X, so that $\mu(X) = 1$. Given any $k \in \mathbb{N}$, the quantities $p_w := \mu[w]$ for $w \in \mathcal{L}_k$ form a probability vector with $\#\mathcal{L}_k \leq (\#S)^k$ entries, whose entropy is

$$H_k(\mu) = \sum_{w \in \mathcal{L}_k} -p_w \log p_w = \sum_{w \in \mathcal{L}_k} -\mu[w] \log \mu[w].$$

▶ EXERCISE 2.16. Given any probability vector $p \in \Delta^{N-1}$, we have

$$H(p) := \sum_{j=1}^{N} -p_j \log p_j \leq \log N, \tag{2.86}$$

with equality if and only if $p_j = \frac{1}{N}$ for all j.

From Exercise 2.16, we see that

$$H_k(\mu) \leq \log \#\mathcal{L}_k, \tag{2.87}$$

with equality if and only if every word in $\mathcal{L}_k$ has equal weight under μ.

▶ EXERCISE 2.17. Prove that $\#\mathcal{L}_{n+k} \leq (\#\mathcal{L}_n)(\#\mathcal{L}_k)$ for every $n, k \in \mathbb{N}$.

Combining Exercise 2.17 with Fekete's Lemma 2.47 proves existence of the following limit:

$$h(X) := \lim_{k \to \infty} \frac{1}{k} \log \#\mathcal{L}_k. \tag{2.88}$$

This is the *topological entropy* of the shift space X. By (2.87), we see that

$$h(\mu) \leq h(X) \text{ for all } \mu \in \mathcal{M}_\sigma(X). \tag{2.89}$$

Exercise 2.16 suggests that there should be a connection between measures achieving (or coming close to) equality in (2.89), and measures for which all words of length k have roughly equal weights.

With this in mind, let us now specialize to the case where A is a primitive 0-1 matrix, $X = X_A$ is the corresponding SFT, and μ is its Parry measure. We will first compute the topological entropy $h(X)$, then the measure-theoretic entropy $h(\mu)$.

For every $w \in S^k$, we have

$$A_{w_1 w_2} A_{w_2 w_3} \cdots A_{w_{k-1} w_k} = \begin{cases} 1 & w \in \mathcal{L}_n, \\ 0 & \text{otherwise}, \end{cases}$$

and thus

$$\#\mathcal{L}_k = \sum_{w \in S^n} A_{w_1 w_2} A_{w_2 w_3} \cdots A_{w_{k-1} w_k} = \sum_{i,j} (A^{k-1})_{ij}.$$

▶ EXERCISE 2.18. Use the Perron–Frobenius Theorem 1.141 to prove that there exists a constant $C > 0$ such that $\sum_{i,j}(A^{k-1})_{ij} = C^{\pm 1}\lambda^k$ for every $k \in \mathbb{N}$, and deduce that $h(X) = \log \lambda$.

From (2.89), the Parry measure μ has $h(\mu) \le h(X) = \log \lambda$. We show that in fact, we have equality.

By (2.79), we have $\mu[w] \le Q\lambda^{-k}$ for all $w \in \mathcal{L}_k$. Using monotonicity of the information function, this implies that

$$I(\mu[w]) \ge I(Q\lambda^{-k}) = k \log \lambda - \log Q.$$

Summing over $w \in \mathcal{L}_k$ gives

$$H_k(\mu) = \sum_{w \in \mathcal{L}_k} \mu[w] I(\mu[w]) \ge \sum_{w \in \mathcal{L}_k} \mu[w](k \log \lambda - \log Q) = k \log \lambda - \log Q.$$

Dividing both sides by k and sending $k \to \infty$ gives

$$h(\mu) \ge \log \lambda, \tag{2.90}$$

and since the other inequality was proved in (2.89), we conclude that

$$h(\mu) = h(X) = \log \lambda. \tag{2.91}$$

We say that μ is a *measure of maximal entropy (MME)*. In fact, one can prove that it is the *only* MME for $X = X_A$. Thus the Parry measure can be characterized in three ways:

- it is the RPF measure associated to $\varphi = 0$;
- it is the unique Gibbs measure associated to $\varphi = 0$;
- it is the unique measure of maximal entropy.

It turns out that the unique ACIP for expanding maps admits a similar range of characterizations. Let X be a compact connected Riemannian manifold, and $f\colon X \to X$ an expanding $C^{1+\alpha}$ map. Given an invariant Borel probability measure $\mu \in \mathcal{M}_f(X)$, we would like to once again define the measure-theoretic entropy $h_f(\mu)$ as the rate of increase of expected information.

For this we require a partition. Let $\xi = \{C_1, \dots, C_N\}$ be a finite partition of X, and to each $x \in X$ associate a sequence $\pi(x) \in S^{\mathbb{N}_0}$ by the condition that $f^n(x) \in C_{\pi(x)_n}$. That is, $\pi(x)_n \in S$ simply records which partition element the iterate $f^n(x)$ lands in. Then $\pi\colon X \to S^{\mathbb{N}_0}$ *codes* the dynamics of f: we have $\sigma \circ \pi = \pi \circ f$, and we can define the *dynamical entropy of μ relative to the partition* ξ by

$$h_f(\mu, \xi) := h(\pi_* \mu). \tag{2.92}$$

Some partitions may not "see" all the entropy of the system.[10] Thus we define the *measure-theoretic entropy*[11] of (μ, f) to be

$$h_f(\mu) := \sup\{h_f(\mu, \xi) : \xi \text{ is a finite partition of } X\}. \tag{2.93}$$

With some work, one can prove the following *Margulis–Ruelle inequality*:

$$h_f(\mu) \le \int \log J \, d\mu \quad \text{for all } \mu \in \mathcal{M}_f(X), \tag{2.94}$$

where $J(x) = |\det Df(x)|$ is the Jacobian determinant. Moreover, it turns out that equality holds in (2.94) if and only if $\mu \ll m$, where m is Lebesgue measure. Thus each of the equivalent criteria in the following list characterizes the ACIP μ.

- It is the RPF measure associated to $\varphi(x) = -\log J(x)$.
- It is the unique Gibbs measure associated to $\varphi(x) = -\log J(x)$.
- It is the unique measure that maximizes the quantity $h_f(\mu) + \int \varphi \, d\mu$, where $\varphi(x) = -\log J(x)$. (We say that it is the *equilibrium state*, or *equilibrium measure*, for this potential.)
- It is the unique absolutely continuous invariant measure.
- It is the unique physical measure in the sense of Definition 2.18.

2.10. Sinai–Ruelle–Bowen measures

Lec 41
Mon, Apr 28

Now we turn our attention to the question of physical measures for invertible dynamical systems. To this end, let M be a compact Riemannian manifold and $f\colon M \to M$ a $C^{1+\alpha}$ diffeomorphism, meaning that f is invertible and both f and f^{-1} are $C^{1+\alpha}$.

This class of systems includes the standard map and the time-1 map of the Lorenz flow.[12] For the standard map, we already saw that Lebesgue measure is invariant, and thus we already have a physical measure. For the Lorenz system, the question of finding a physical measure turns out to involve complications that are best avoided here,[13] and so for now we confine our attention to a simpler class of examples.

Consider a matrix $A \in SL(2, \mathbb{Z})$ with eigenvalues $\lambda^{\pm 1}$, where $0 < \lambda < 1 < \lambda^{-1}$. The example $A = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ is good to keep in mind. The linear map $A\colon \mathbb{R}^2 \to \mathbb{R}^2$ given by left multiplication by A preserves the integer lattice $\mathbb{Z}^2$, and thus this passes to a diffeomorphism $g = g_A\colon \mathbb{T}^2 \to \mathbb{T}^2$. This preserves Lebesgue measure m, but if we perturb the system slightly, we obtain a diffeomorphism f that need not preserve Lebesgue measure, and we consider the question of whether f has a physical measure.

[10] Think about what happens if $\xi = \{X\}$.

[11] Sometimes called *Kolmogorov–Sinai entropy*.

[12] For the standard map, M is the torus. For the Lorenz flow, the natural phase space is $\mathbb{R}^3$, which is not compact, but by adding a point at infinity we can consider this system as living on the three-dimensional sphere.

[13] This is due to the *non-uniform* nature of the hyperbolic behavior in the system.

Recall that given $\mu \in \mathcal{M}_f$, we write

$$G_\mu = \left\{ x \in M : \mathsf{A}_n \varphi(x) \to \int \varphi \, d\mu \text{ for all } \varphi \in C(M) \right\}$$

for the set of generic points for μ, and we say that μ is *physical* if $m(G_\mu) > 0$. So far, all of our examples of physical measures have been absolutely continuous with respect to m, and we were able to obtain absolutely continuous invariant measures by studying the evolution of density functions under the dynamics: expansion in phase space led to a "smoothing out" of the density function, and this controlled regularity allowed us to prove a contraction result that yielded convergence to the ACIP.

However, life is different when there is a contracting direction: pushing forward an absolutely continuous measure under a contracting map has the effect of making its density function *less* regular! Consider Gaussian measure μ on $\mathbb{R}$ under the map $x \mapsto \frac{1}{2}x$ to see this effect. In this example, we see that the measures $f_*^n \mu$ converge in the weak* topology to an atomic measure at the stable fixed point. This measure is physical, but is not absolutely continuous.

Thus it appears that physicality should be associated with absolute continuity in the presence of expansion, but not in the presence of contraction. To make this idea more precise, let $g = g_A$ be the linear map of the torus from above, and let $f \approx g$ be close to it in the C^1 topology, so that at every point $x \in \mathbb{T}^2$, we have $f(x) \approx g(x)$ and $Df(x) \approx Dg(x)$.

Just as in our proof of the Hadamard–Perron Theorem 1.15, the eigenspaces for A lie inside cones K^u and K^s, which are invariant under A and A^{-1}, respectively. Provided Df is close enough to $Dg = A$, we see that these cones are invariant under Df and Df^{-1} as well. Moreover, this is true not just near a fixed point, but at *every* point on $\mathbb{T}^2$.

Following the same arguments as in the Hadamard–Perron Theorem, we see that at every $x \in \mathbb{T}^2$, the following sets are one-dimensional subspaces of $\mathbb{R}^2$:

$$E_x^u := \bigcap_{n \geq 0} D(f^n)_{f^{-n}x} K^u, \qquad E_x^s := \bigcap_{n \geq 0} D(f^{-n})_{f^n x} K^s.$$

Moreover, these distributions satisfy

$$E_x^u \subset K^u, \qquad E_x^s \subset K^s, \qquad \mathbb{R}^2 = E_x^u \oplus E_x^s,$$

and we also have the following invariance property:

$$Df(x) E_x^u = E_{f(x)}^u \quad \text{and} \quad Df(x) E_x^s = E_{f(x)}^s.$$

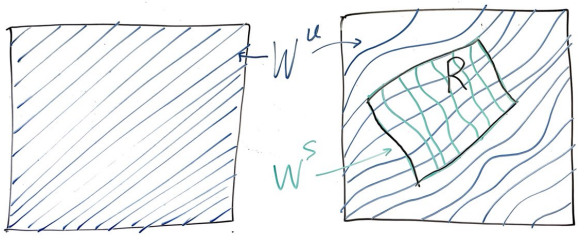
From the cone contraction and expansion estimates, we also see that given $\gamma \in (\lambda, 1)$, the cones can be chosen narrow enough such that for every $v^u \in E^u$ and $v^s \in E^s$, we have

$$\|Df(x) v^u\| \geq \gamma^{-1} \|v^u\| \quad \text{and} \quad \|Df(x) v^s\| \leq \gamma \|v^s\|.$$

A map f with these properties (invariant splitting into distributions with uniform contraction and expansion estimates) is called an *Anosov diffeomorphism*, and is often referred to as *uniformly hyperbolic*.

REMARK 2.48. The cone argument just sketched implies that the set of Anosov diffeomorphisms is open in the C^1 topology: if g is Anosov, then so is any C^1-small perturbation of g.

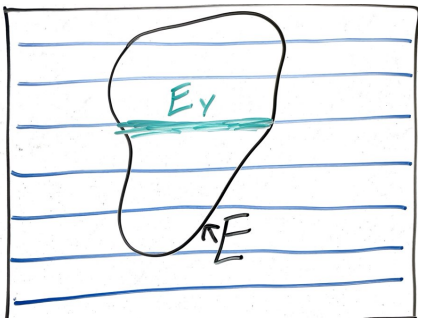
FIGURE 2.1. Leaves of W^u for $g = g_A$ and $f \approx g$, and a "rectangle" R.



Continuing to follow the arguments from the Hadamard–Perron Theorem, one can prove that there exist f-invariant foliations W^u and W^s that are tangent to $E^{u,s}_x$ at every $x \in M$. For the linear toral map $g = g_A$, leaves of these foliations are (images of) lines, while for the perturbations $f \approx g$, the leaves are curves whose tangent vectors lie in the corresponding cones; see Figure 2.1.

Figure 2.1 also illustrates a "rectangle" R, which is a set that can be written as a union of local unstable leaves W^u_{loc} and also as a union of local stable leaves W^s_{loc} in such a way that each W^u_{loc} meets each W^s_{loc} in a single point. We encountered such dynamical rectangles earlier, in our study of the horseshoe map. Here they appear *everywhere* in the system.

FIGURE 2.2. Disintegrating a measure.



Given a Borel probability measure μ, we can write $\mu|_R$ as a combination of a family of *conditional measures* μ^u_x on the leaves of W^u. See Figure 2.2, which shows

how we have

$$m_2(E) = \int_{[0,1]} m_1(E_y)\, dm_1(y) = \int_{[0,1]^2} m_1(E_y)\, dm_2(x,y).$$

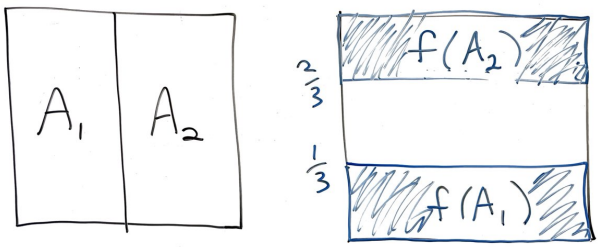
DEFINITION 2.49. An invariant measure μ is a *Sinai–Ruelle–Bowen (SRB) measure* if $\mu_x^u \ll m_x^u$ for μ-a.e. x, where m_x^u denotes leaf volume.

▶ EXERCISE 2.19. Every SRB measure is a physical measure. (First prove that if ν is invariant and $x \in G_\nu$, then $W_x^s \subset G_\nu$.)

One can prove that each of the following limits exists, agrees, and is the unique SRB measure.

- $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f_*^k m$
- $\lim_{n\to\infty} f_*^n m$
- $\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f_*^k m_x^u$
- $\lim_{n\to\infty} f_*^n m_x^u$

FIGURE 2.3. Developing singularities in the stable direction.



To construct an SRB/physical measure as any of the four limits just listed, one may hope to follow the ideas in the construction of an ACIP for an expanding map, but there is a new problem now. For an expanding map, when an absolutely continuous measure is pushed forward under the dynamics, its density transforms according to the RPF operator, and expansion in phase space leads to a contraction in these auxiliary dynamics: the density becomes "more regular" in an appropriate sense. However, for an Anosov diffeomorphism, this is no longer true: while the density becomes more regular in the unstable direction, it becomes *less* regular in the stable direction, along which the dynamics may be converging to something irregular like a Cantor set, as shown in Figure 2.3.

The classical method for dealing with this irregularity involves *Markov partitions*: by coding the system using a carefully chosen partition of the phase space M, it is possible to find an SFT $X \subset A^{\mathbb{Z}}$ and a continuous onto map $h\colon X \to M$ such that $h \circ \sigma = f \circ h$, and such that h is injective on a "sufficiently large" set. Then one can pass from the two-sided SFT X to the corresponding one-sided SFT $X^+ \subset A^{\mathbb{N}}$ by "forgetting" the negative indices. (One must also find an "equivalent version"

of the geometric potential function $J^u(x) = |\det Df|_{E^u_x}|$ that only depends on the positive indices; this can be done provided $f \in C^{1+\alpha}$.) Then on the one-sided SFT, we can apply the RPF Theorem 2.45 to obtain a Gibbs measure μ for J^u, which is the unique SRB measure.

REMARK 2.50. More recently, an alternative approach via *anisotropic Banach spaces* has been developed, where one works with the RPF operator directly over the manifold M, without the use of Markov partitions. This requires replacing the space of Hölder continuous densities with a more sophisticated Banach space of objects that behave like functions in the unstable direction, and like distributions in the stable direction, so that the RPF operator still acts as a contraction in an appropriate sense.

THEOREM 2.51. *Let M be a compact smooth Riemannian manifold and $f\colon M \to M$ a topologically transitive Anosov diffeomorphism with geometric potential function $J^u(x) = |\det Df|_{E^u_x}|$. Given an f-invariant Borel probabiliy measure μ on M, the following properties are equivalent.*

- *$\mu^u_x \ll m^u_x$ for μ-a.e. x.*
- *The measure μ is physical.*
- *The measure μ is the RPF measure for $-\log J^u$.*
- *The measure μ is the Gibbs measure for $-\log J^u$.*
- *The measure μ is the* equilibrium measure *for $-\log J^u$: for any $\nu \in \mathcal{M}_f$, we have $h(\nu) \le \int \log J^u \, d\nu$, with equality if and only if ν is SRB.*

There is exactly one measure that satisfies one (and hence all) of these properties; this is the unique SRB measure for f.