Hyperbolic dynamical systems, chaos, and Smale's horseshoe: a guided tour*

Yuxi Liu

Abstract

This document gives an illustrated overview of several areas in dynamical systems, at the level of beginning graduate students in mathematics. We start with Poincaré's discovery of chaos in the three-body problem, and define the Poincaré section method for studying dynamical systems. Then we discuss the long-term behavior of iterating a diffeomorphism in $\mathbb{R}^n$ around a fixed point, and obtain the concept of hyperbolicity. As an example, we prove that Arnold's cat map is hyperbolic.

Around hyperbolic fixed points, we discover chaotic homoclinic tangles, from which we extract a source of the chaos: Smale's horseshoe. Then we prove that the behavior of the Smale horseshoe is the same as the binary shift, reducing the problem to symbolic dynamics. We conclude with applications to physics.

1 The three-body problem

1.1 A brief history

The first thing Newton did, after proposing his law of gravity, is to calculate the orbits of two mass points, M_1, M_2 moving under the gravity of each other. It is a basic exercise in mechanics to show that, relative to one of the bodies M_1, the orbit of M_2 is a conic section (line, ellipse, parabola, or hyperbola).

The second thing he did was to calculate the orbit of three mass points, in order to study the sun-earth-moon system. Here immediately he encountered difficulty. And indeed, the three body problem is extremely difficult, and the n-body problem is even more so.

We will not give the full history of how Poincaré studied the n-body problem. A good overview is in [Chenciner, 2015]. In short, in 1887, the king of Sweden proposed a prize for a solution to the n-body problem, and Poincaré failed to solve it, but discovered one reason why the problem is so difficult. He found that chaos is everywhere in the n-body problem, even in the most basic version – the planar, circular, restricted three-body problem.

We shall not say more about the general n-body problem, other than showing Figure 1, which demonstrates the beautiful chaos.

1.2 The planar, circular, restricted three-body problem

Refer to Figure 2. Consider the earth, the moon, and a spacecraft moving around them. The earth and moon circle each other in roughly circular orbits, and the spacecraft moves under the effect of their gravity, but does not affect the moon or the earth because it is so small.

*Project paper for the course Ergodic Theory taught by Tanja Schindler at Australian National University in 2019 Semester 1.

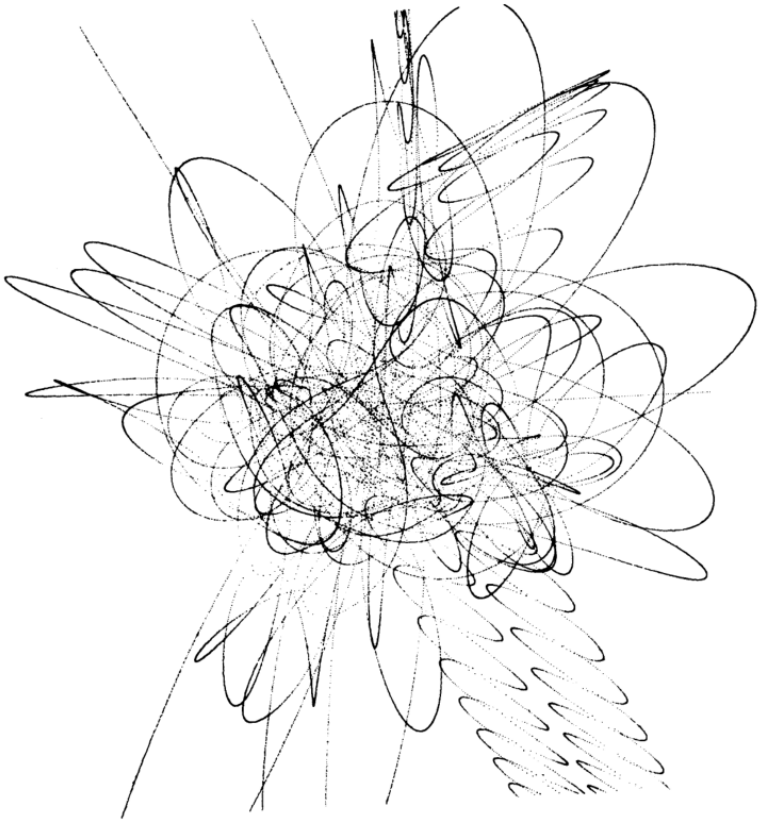

Figure 1: A typical three-body problem trajectory. Figure from [Hut, 1993].

Viewed at the barycenter, the earth and the moon both move in circular orbits with the same angular velocity Ω. So, if we use a rotating barycenter frame, the earth and the moon are both static. Then it only remains to study the motion of the satellite.

Let the plane of rotation be the xy-plane, and the axis of rotation be the z-axis. Then, when viewed in the rotating frame, the satellite experiences three forces: attraction to the earth, to the moon, and a centrifugal force parallel to the xy-plane. Then we see that if the satellite has initial position (x, y, z) and velocity (v_x, v_y, v_z), and $z = 0, v_z = 0$, then it will always remain so, since there is no force pulling it along the z-axis.

Thus we have obtained the planar (in the xy-plane), circular (the earth-moon system moves circularly), restricted (the satellite is too small to affect the earth-moon system), three-body problem. This is as simple as it could possible be, and it still contains chaos.

1.3 Reducing to two dimensions

Given any initial conditions $x(0), y(0)$, $(v_x(0), v_y(0))$ of the satellite, there exists a unique solution $(x(t), y(t))$ that satisfies its Newtonian equations of motion. This is how one could solve the orbit of the satellite, but there are other ways to do so. One could instead consider the **phase space** of the satellite

$$(x, y; v_x, v_y) \in \mathbb{R}^2 \times \mathbb{R}^2. \tag{1}$$

Then, given a path in the phase space, parametrized by some arbitrary number s:

$$(x(s), y(s); v_x(s), v_y(s)), \tag{2}$$

we can recover $(x(t), y(t))$ by solving the ODE

$$\frac{dx}{ds}\frac{ds}{dt} = v_x, \quad \frac{dy}{ds}\frac{ds}{dt} = v_y, \tag{3}$$

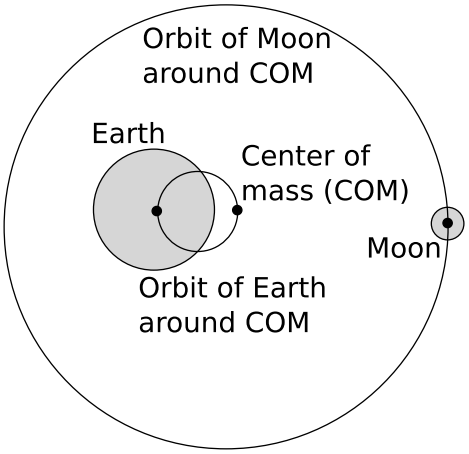

Figure 2: The earth-moon system. We drew it in inkscape.

which gives[1] $\frac{ds}{dt}$ as a function of s and can be solved to give $t = f(s)$, which can be inverted to give $s = f^{-1}(t)$, which can then be plugged into $(x(s), y(s))$ to give $(x(t), y(t))$.

Thus, we only need to solve for orbits in the phase space that satisfies the equations of motion. The phase space has 4 dimensions, and the curves have 1 dimension. We must cut away 3 dimensions.

At each point $(x, y; v_x, v_y)$ of the phase space, there is a corresponding energy of the satellite. The explicit formula is not important, but it looks like

$$\text{Energy} = \text{Kinetic energy} + \text{Gravitational energy} + \text{Centrifugal force field energy}$$

By conservation of energy, any orbit in the phase space must remain on an equal-energy surface $E = E_0$, which has $4 - 1 = 3$ dimensions. Call such a surface T. The surface can be coordinatized by (x, y, v_x), where v_y can be solved from $(x, y, v_x), E = E_0$.

We could impose a new coordinate system of T such that one of the coordinates is circular. For example, we could replace y by θ, so that $y = x \tan \theta$. Then we have the coordinates (x, v_x, θ) for the surface T. Under such coordinates, it look like a solid donut $\mathbb{R}^2 \times S^1$, where S^1 is the circle. In general, these coordinates can't cover the whole T smoothly. For example, $(x, y) = (0, 0)$ cannot be covered smoothly. But such singularities are rare (of measure 0), and we can ignore them in this document.

Then, we are concerned with curves in a torus. Let's consider one such curve $(x(s), v_x(s), \theta(s))$. If we are lucky, $\theta(s)$ is monotonically increasing, and so the curve would circle around the torus. But we might get unlucky and the curve would fall back without completing a circle. This can be fixed by another coordinate transform into action-angle coordinates, but we won't give the details here. Suffice to say that in the action-angle coordinates, the angle θ increases with constant speed.

Then, we can study the curves as they move around the torus by studying their intersections with the plane $\theta = 0$. This is called the method of **Poincaré section**. See Figure 3 for an illustration.

Let the plane of Poincaré section be P, then any point $p \in P$, there is a unique curve through p, so we can trace the curve around the torus by one cycle and land at $p' \in P$. This gives a bijection $F : P \to P$, which by Liouville theorem, is area-preserving. F is smooth, since the equations of motion are smooth. So is its inverse, by reversing the direction of time. Thus F is a diffeomorphism.

[1]Except when $\frac{d}{ds}(x, y) = 0$, that is, when the parametrization is degenerate. We assume that this does not happen.

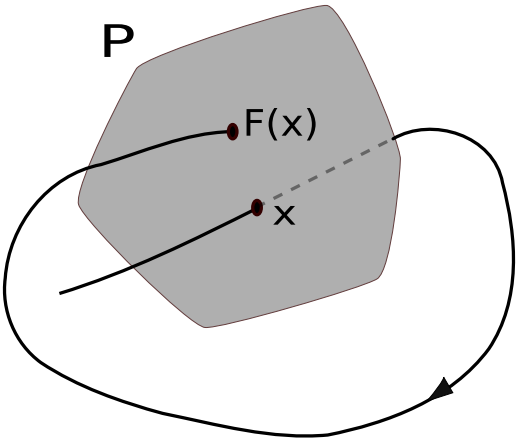

Figure 3: The method of Poincaré section. The three-dimensional space represents the three-dimensional slice of phase space under consideration, and the curves are solutions to the equations of motion in phase space. The curves are roughly circular, since we use the action-angle coordinates. We take a two-dimensional plane P across the phase space, and study the intersections of the curves with the plane, which defines a map $F : P \to P$. Figure taken from `https://en.wikipedia.org/wiki/File:Poincare_map.svg`, and edited in Inkscape.

Thus, we have reduced a three-body problem into a problem of measure-preserving diffeomorphisms on the plane. Poincaré did this, and discovered a chaotic complexity:

> When we try to represent the figure formed by these two curves and their intersections in a finite number, each of which corresponds to a doubly asymptotic solution, these intersections form a type of trellis, tissue, or grid with infinitely serrated mesh. Neither of these two curves must ever cut across itself again, but it must bend back upon itself in a very complex manner in order to cut across all of the meshes in the grid an infinite number of times. The complexity of this figure will be striking, and I shall not even try to draw it. [Poincaré, 1899, volume III, chapter XXXIII, section 397]

2 Hyperbolicity of diffeomorphisms

2.1 Iteration around a fixed point

Consider a general diffeomorphism F on $\mathbb{R}^n$. At each point x, F locally behaves as a linear map, the linear differential map $M = DF_x : T_x\mathbb{R}^n \to T_{F(x)}\mathbb{R}^n$, or more simply, $DF_x : \mathbb{R}^n \to \mathbb{R}^n$. If $F(x) = x$, that is, x is a fixed point of F, then we can consider the effect of iterating F upon a neighborhood of x. The long-term behavior of iterating F depends critically on the eigenvalues of $M = dF_x$.

M has n eigenvalues $\lambda_1, \cdots, \lambda_n \in \mathbb{C}$ with eigenvectors $v_1, \cdots, v_n \in \mathbb{C}^n$. For convenience, we assume that there is no multiplicity.[2]

Since M is real, its complex eigenvalues are the roots of the real equation

$$\det(\lambda I - M) = 0,$$

[2]Our conclusions are valid for general M, though the proof would not be as geometric, but rather rely on the Jordan normal form of the matrix.

so if we have nonreal λ, its conjugate $\overline{\lambda}$ is also an eigenvalue. Correspondingly, we have eigenvector $v \in \mathbb{C}^n$, such that $Mv = \lambda v$. Since $\overline{Mv} = M\overline{v} = \overline{\lambda}\overline{v}$, we have that $\overline{v}$ is the eigenvector of $\overline{\lambda}$.

Let $v_r = Re(v), v_i = Im(v)$, so $v = v_r + iv_i$. If v_r, v_i are $\mathbb{R}$-linearly dependent, then it's easy to show that λ must be real, so v_r, v_i are linearly independent. Decompose λ into polar components $\lambda = re^{i\theta}$, then we have

$$Mv_r = r(\cos(\theta)v_r - \sin(\theta)v_i), \quad Mv_i = r(\cos(\theta)v_i + \sin(\theta)v_r).$$

So the effect of M on the 2-dimensional real subspace $\langle v_r, v_i\rangle_{\mathbb{R}} \subset \mathbb{R}^n$ is a "rotation" and a scaling by r. Since v_r and v_i might not be of the same length, so the rotation is really an "elliptical rotation". All in all, the effect of M on $\langle v_r, v_i\rangle_{\mathbb{R}}$ is an "elliptical spiraling".

Thus, M splits $\mathbb{R}^n$ into a direct sum of linear subspaces, as

$$\mathbb{R}^n = \left(\bigoplus_{j=1}^{k} \langle v_j\rangle_{\mathbb{R}}\right) \bigoplus \left(\bigoplus_{j=k+1}^{k+l} \langle Re(v_j), Im(v_j)\rangle_{\mathbb{R}}\right), \tag{4}$$

where we have arranged so that $\lambda_1, \cdots \lambda_k$ are real, and

$$\lambda_{k+1} = \overline{\lambda_{k+l+1}}, \cdots \lambda_{k+l} = \overline{\lambda_{k+l+l}} = \overline{\lambda_n},$$

are nonreal, conjugate pairs. The action of M on $\mathbb{R}^n$ is a scaling by λ_j on $\langle v_j\rangle_{\mathbb{R}}$, and an elliptical spiraling on $\langle Re(v_j), Im(v_j)\rangle_{\mathbb{R}}$.

Now, if $|\lambda| < 1$, and λ is real, then the corresponding $\langle v\rangle_{\mathbb{R}}$ is shrunken by M; if λ is nonreal, then the corresponding $\langle Re(v), Im(v)\rangle_{\mathbb{R}}$ is rotated-and-shrunken by M. Similarly for $|\lambda| = 1, |\lambda| > 1$.

Thus we obtain a splitting of $\mathbb{R}^n = E_{<1} \oplus E_{=1} \oplus E_{>1}$, corresponding to the three cases of $|\lambda|$. The effect of M is to rotate-and-shrink $E_{<1}$, rotate-and-expand $E_{>1}$, and merely rotate $E_{=1}$. All rotations, as noted, are elliptical.

Since as we iterate M, all points of $E_{<1}$ fall toward 0, so we call it the **stable subspace** of M, and $E_{>1}$ the **unstable subspace** of M. Points in $E_{=1}$ circle around 0, neither approaching nor escaping it, so it's called the **central subspace**. The usual notations for them are E^s, E^u, E^c.

There is another way to define the three subspaces without explicitly calculating the eigenvectors.

Consider some eigenvalue $|\lambda| < 1$, and its eigenvector v. If λ is real, then

$$\forall w \in \langle v\rangle_{\mathbb{R}}, \quad \|M^n w\| \leq |\lambda|^n \|w\|.$$

If λ is nonreal, then $v = v_r + iv_i$,

$$\forall w \in \langle v_r, v_i\rangle_{\mathbb{R}}, \quad \|M^n w\| \leq c|\lambda|^n \|w\|,$$

where

$$c = \frac{\text{long axis}}{\text{short axis}} \quad \text{of the ellipse } \{v_r \sin(\theta) + v_i \cos(\theta) \,|\, \theta \in [0, 2\pi]\}.$$

The proof is visual: each iteration of M shrinks the ellipse by $|\lambda|$.

Thus, for big enough c and some $\lambda \in (0, 1)$, we have

$$\forall n = 1, 2, \cdots, \forall w \in E^s, \|M^n w\| \leq c\lambda^n \|w\|.$$

Similarly, for big enough c and some $\lambda \in (0, 1)$, we have

$$\forall n = 1, 2, \cdots, \forall w \in E^u, \|M^{-n} w\| \leq c\lambda^n \|w\|.$$

And for the subspace between the two extremes, the growth behavior under iterations of M is neither exponentially increasing nor decreasing. That is, for any $c, d > 0$, and $0 < \lambda, \xi < 1$, we do not have

$$\forall n \in \mathbb{Z}, \forall w \in E^c, \|M^n w\| \leq c\lambda^n \|w\|,$$

or

$$\forall n \in \mathbb{Z}, \forall w \in E^c, \|M^{-n} w\| \leq d\xi^n \|w\|.$$

We say that x is a hyperbolic fixed point of F iff $E^c = 0$, that is, $\mathbb{R}^n = E^s \oplus E^u$, and all points in $\mathbb{R}^n$ either fall toward 0 or ∞ under iteration of dF_x.

The name "hyperbolic" can be understood in many ways. One way is to note that the prototypical hyperbolic fixed point is $(0, 0)$ for the map $F(x, y) = (2x, \frac{1}{2}y)$ on $\mathbb{R}^2$. Another way, which would make more sense after reading the next section, is to note that the geodesic flow on a hyperbolic plane is hyperbolic on the whole plane.

2.2 Hyperbolicity in general

In general, consider a diffeomorphism F on smooth manifold M. We wish to generalize the concept of hyperbolicity when F fixes not a point x, but a set Λ. For example, suppose we have $M = \mathbb{R}^2 \times S^1$, and $F(x, y, e^{i\theta}) = (2x, \frac{1}{2}y, e^{i(\theta+\phi)})$, then we want to say that M is hyperbolic on the central circle $\{(0, 0)\} \times S^1$.

Let $\Lambda \in M$ be F-invariant, that is, $F(\Lambda) = \Lambda$. We say that Λ is a **hyperbolic set** of F if we can split the sub-vector-bundle $T_\Lambda M$ to a direct sum into a **stable bundle** E^s and **unstable bundle** E^u, such that both are F-invariant:

$$\forall x \in \Lambda, dF_x(E^s_x) = E^s_{F(x)}, dF_x(E^u_x) = E^u_{F(x)}. \tag{5}$$

And there exists $c > 0$, $\lambda \in (0, 1)$, such that

$$\begin{aligned} \forall x \in \Lambda,\ \forall n = 1, 2, \cdots, \forall w \in E^s_x, &\|d(F^n)_x w\| \leq c\lambda^n \|w\| \\ \forall w \in E^u_x, &\|d(F^{-n})_x w\| \leq c\lambda^n \|w\|. \end{aligned} \tag{6}$$

In the definition of hyperbolicity, λ quantifies the hyperbolicity. The closer λ is to 0, the more hyperbolic F is on Λ. The fact that $\lambda < 1$ gives a uniform bound to the hyperbolicity of F on Λ. If there is no such uniform bound, then we would stray right into the study of **nonuniform hyperbolicity**. Another generalization is to not require the clean split into a stable and unstable bundle, but allow a central bundle as well. This is the start of **partial hyperbolicity**. And they can even be combined into the study of **nonuniform partial hyperbolicity**. The very strong reader who wishes to pursue these heavy complications should consult [Barreira and Pesin, 2006].

2.3 Example: Anosov diffeomorphisms

If M is a hyperbolic set of F, then we say that F is an **Anosov diffeomorphism** on M.

As an example, consider the area-preserving **Arnold cat map** on the 2-dimensional torus. Let

$$M = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \tag{7}$$

then define:

$$F : T^2 \to T^2, \quad F(x, y) = M \begin{bmatrix} x \\ y \end{bmatrix} \mod 1. \tag{8}$$

The map is illustrated in Figure 4.

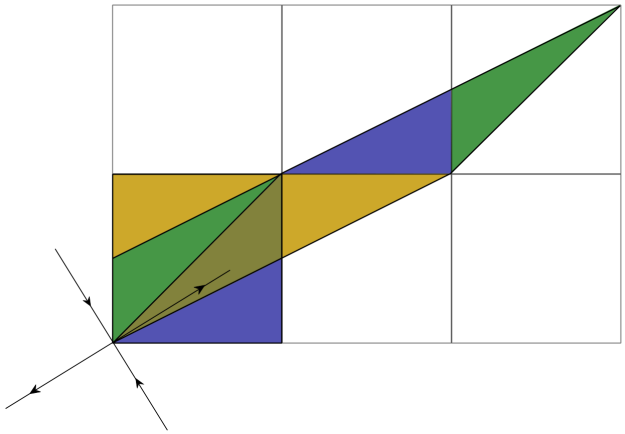

Figure 4: Arnold cat map. The effect of one iteration of the map on the square is shown. The stable and unstable manifolds of the fixed point $(0, 0)$ are also shown. From Wikipedia `https://en.wikipedia.org/wiki/File:Arnoldcatmap.svg`

The eigenpairs of M are

$$v_1 = \begin{bmatrix} \frac{1+\sqrt{5}}{2} \\ 1 \end{bmatrix}, \lambda_1 = \frac{1}{2}(3+\sqrt{5}); \quad v_2 = \begin{bmatrix} \frac{1-\sqrt{5}}{2} \\ 1 \end{bmatrix}, \lambda_2 = \frac{1}{2}(3-\sqrt{5}).$$

Thus, the unstable and stable bundles of F are

$$E^u = T^2 \times \langle v_1 \rangle_{\mathbb{R}}, \quad E^s = T^2 \times \langle v_2 \rangle_{\mathbb{R}}.$$

Theorem 2.1. *The Arnold cat map is strongly mixing.*

Proof We show that the map has Lebesgue spectrum, then by Theorem 2.12 of [Walters, 2000], it is strongly mixing.

Define $f_{(a,b)} : T^2 \to \mathbb{C}$ by

$$f_{(a,b)}(x, y) = \exp(2\pi i(ax + by)).$$

As in Fourier analysis, $f_{(a,b)}$ with $(a, b) \in \mathbb{Z}^2$ is an orthonormal basis of $L^2(T^2)$. It's also clear that $f_{(a,b)} \circ F^n = f_{(a,b)M^n}$.

Define an equivalence relation on $\mathbb{Z}^2$ by $(a, b) \sim (c, d)$ iff $\exists n \in \mathbb{Z}$, such that $(a, b)M^n = (c, d)$. Any equivalence class under this relation is either finite, in which case it corresponds to a periodic orbit, or infinite, in which case it is of the form

$$\{(a, b)M^n | n \in \mathbb{Z}\} \cong \mathbb{Z}.$$

Consider any periodic point (a, b), such that $\exists n > 0, (a, b)M^n = (a, b)$. But since M has eigenvalues with absolute value $\neq 1$, $(a, b) = (0, 0)$.

Thus, the equivalence relation splits $\mathbb{Z}^2$ into a countable number of classes, one being $\{(0, 0)\}$, and all the others being of the form

$$\{(a, b)M^n | n \in \mathbb{Z}\} \cong \mathbb{Z}.$$

Then, we take one representative from each of these infinite classes, to get $(a_1, b_1), (a_2, b_2), \cdots$.

Then define $f_0 = f_{(0,0)} = 1, f_j = f_{(a_j,b_j)}$ where $j = 1, 2, \cdots$, then we obtain an orthonormal basis of $L^2(T^2)$ in the form of

$$\begin{array}{ccccc} & & f_0 & & \\ \cdots & f_1 \circ F^{-1} & f_1 & f_1 \circ F^1 & \cdots \\ \cdots & f_2 \circ F^{-1} & f_2 & f_2 \circ F^1 & \cdots \end{array} \tag{9}$$

which shows that F has a countable Lebesgue spectrum. □

Note that the proof works in general for any n-by-n matrix M with integer entries, whose inverse also has integer entries, and with absolute values of eigenvalues $\neq 1$. In this way, we obtain a whole family of measure-preserving Anosov diffeomorphisms on tori.

2.4 Stable and unstable manifolds

Consider a hyperbolic fixed point x of a diffeomorphism F on a smooth manifold M. Consider all points that, under iteration of F, converges to x. Call them the **stable manifold** of x:

$$W^s(x) = \{y \in M | \lim_{n\to\infty} F^n(y) = x\}, \tag{10}$$

and similarly define the **unstable manifold** of x:

$$W^u(x) = \{y \in M | \lim_{n\to\infty} F^{-n}(y) = x\}. \tag{11}$$

Locally at x, the behavior of F is approximated by the behavior of dF_x, the local linear approximation[3]. So split T_xM into the stable and unstable subspaces $T_xM = E^s \oplus E^u$, and we should expect that, locally at x, W^u should look just like E^u, but slightly curved. Similarly for W^s being a curved version of E^s. See Figure 5. This observation can be rigorously stated as the **Hadamard–Perron theorem**, but we will not do so here. The rigorously stated and proven version can be found in the classic book on dynamical systems, [Katok and Hasselblatt, 1995, section 6.2].

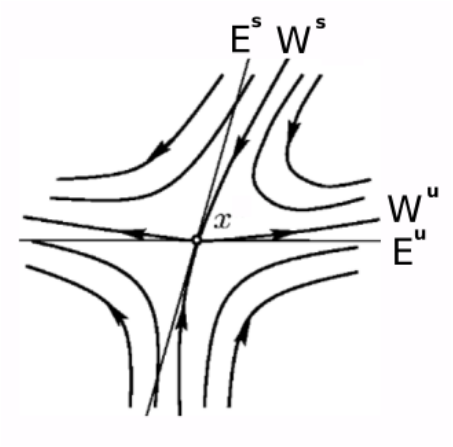

Figure 5: The Hadamard–Perron theorem. Here W^s, W^u denotes the stable and unstable manifolds. They are slightly curved versions of E^s, E^u, the stable and unstable subspaces. We took a picture from `https://math.stackexchange.com/questions/1241375` and edited it in inkscape.

Theorem 2.2. *W^u and W^s cannot have self-intersections.*

Proof (sketch) If there is a self-intersection of W^s at some point y, then we iterate F on a small neighborhood U of y, until $W^s \cap U$ becomes close enough to x. Since F is a diffeomorphism, the self-intersection of $W^s \cap U$ is preserved after the iteration. Since by Hadamard–Perron theorem, locally W^s looks just like E^s around x, and there is no self-intersection of E^s. Contradiction.

Same argument works for W^u, with F^{-1} instead of F. □

[3]This is the discrete case of the **Hartman–Grobman theorem**.

Also, consider two different fixed points x, x', with their own stable and unstable manifolds. $W^s(x)$ and $W^s(x')$ cannot have intersections either, since otherwise the intersection point would have to converge to boh x and x'. Similarly, $W^u(x)$ and $W^u(x')$ are disjoint.

There can however be intersections between stable and unstable manifolds, and this is a source of great complexity.

Suppose $W^u(x)$ and $W^s(x)$ intersect transversely at y, then we say that y is a **homoclinic intersection** of x. Similarly, if $x \neq x'$, and $y \in W^u(x) \cap W^s(x')$, then y is a **heteroclinic intersection** of x, x'.

Both homoclinic and heteroclinic intersections are prevalent in complicated dynamical systems, such as in the Poincaré section in the three-body problem. They can generate extremely complicated behaviors, called the **tangle**. A schematic drawing is Figure 6. There are many good explanations of the structure of the tangle, so we will not explain here. The reader can simply search online, or read [Sussman and Wisdom, 2015, Chapter 4.3].

We will only consider the case where the intersection is **transverse**. It is possible that the stable and unstable manifold are tangent at the intersection, which makes the dynamics even more complex. Tangent intersections are studied in the context of structural stability, where a dynamical system is varied gradually, causing the stable and unstable manifolds to change shape, and some transverse intersection could become tangential, which marks a sudden change in the behavior of the dynamical system.

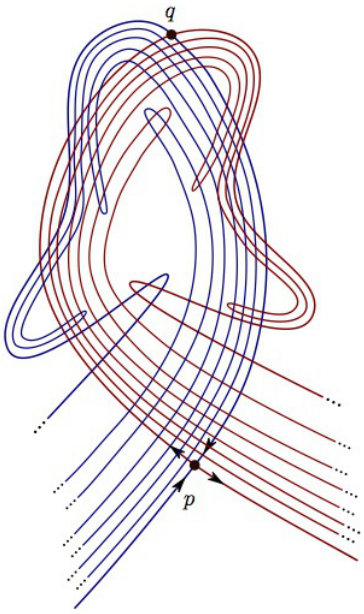

Figure 6: A schematic drawing of homoclinic tangle. p is the hyperbolic fixed point and q is a homoclinic intersection of p. Picture from `https://www.mat.univie.ac.at/~bruin/VO_DS2018.html`.

Poincaré called the stable and unstable manifolds the "asymptotic manifolds", because the points on these manifolds approaches the fixed points asymptotically. Poincaré saw the complexity of the tangle, even without computer simulations, testifying his amazing mind.

Homoclinic tangles can be absolutely beautiful. We cannot resist presenting some more in Figures 7, 8.

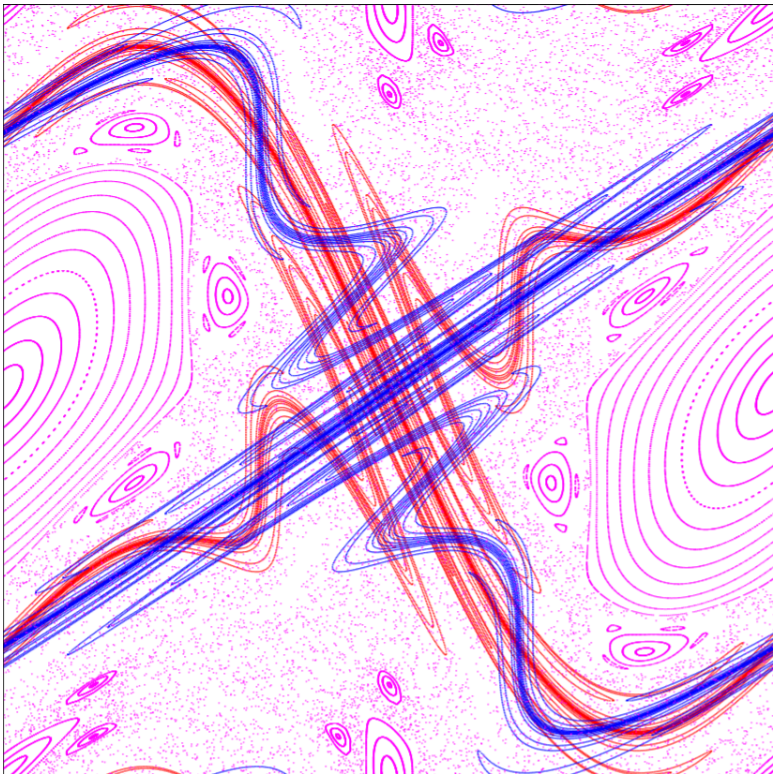

Figure 7: A homoclinic tangle computed by Carles Simò. Figure from [Chenciner, 2015, Figure 18].

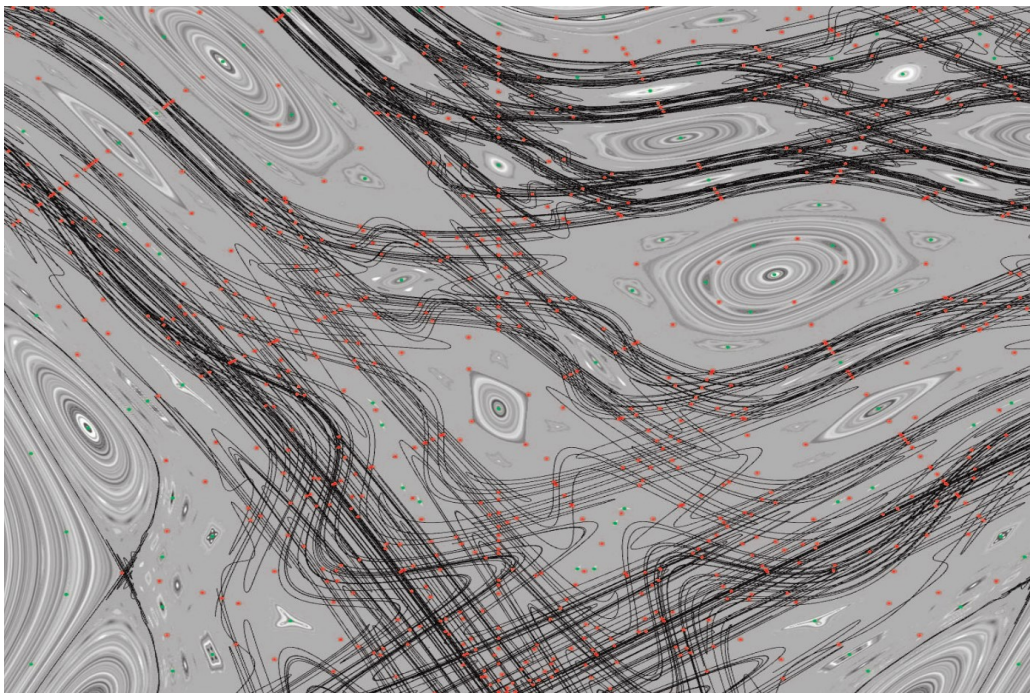

Figure 8: A homoclinic tangle in the so-called Chirikov standard map, which we will not define here, but it is a standard example in dynamical systems. Details can be found in the citation. The green dots are elliptic stable fixed-points, and the red dots are hyperbolic unstable fixed-points. Figure from [Tricoche et al., 2011, Figure 7(b)].

3 Smale horseshoe

3.1 A brief history

In 1960, Stephen Smale was studying dynamical systems on the beaches of Rio de Janeiro, and discovered the **horseshoe map**. This map turned out to exist in many dynamical systems, to imply the existence of chaos. This was first published in [Smale, 1967] and it caused a great flourishing of hyperbolic dynamics in the 60s and 70s. For a lively and detailed story of the discovery, told straight from the horse's (that is, Stephen Smale's) mouth, read [Smale, 1998].

3.2 Finding the horseshoe

The Smale horseshoe is an abstraction that highlights certain features of the homoclinic tangle. Even this abstraction, as we will see, already contains a rich chaos. From this, one can appreciate that the full behavior of the homoclinic tangle is vastly more complicated.

There are in fact several ways to find a horseshoe in a homoclinic tangle. One way is to consider as in Figure 9. Take a small ball around a hyperbolic fixed point x, and iterate F on the ball forwards to obtain U, stretched along the stable manifold $W^s(x)$, and iterate F backwards to obtain V, stretched along the unstable manifold $W^u(x)$. Then we obtain something that is topologically the same as stretching a square, then folding it back to itself.

Let this stretch-and-fold map be F, and the square be S. We really don't need F to be measure-preserving, or be well-defined outside of S, since we are only concerned with the topological effect of iterating F on a certain subset of S. So, define F on S by horizontally compressing the square to width λ, then stretching to length γ, then fold back onto itself, as in Figure 10.

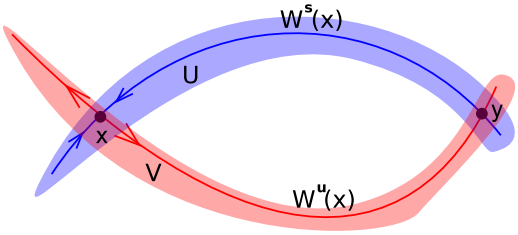

Figure 9: Finding a horseshoe in a homoclinic tangle. x is a hyperbolic fixed point, and y is a homoclinic intersection. A horseshoe emerges when we iterate a ball around x forwards and backwards. We drew this in Inkscape.

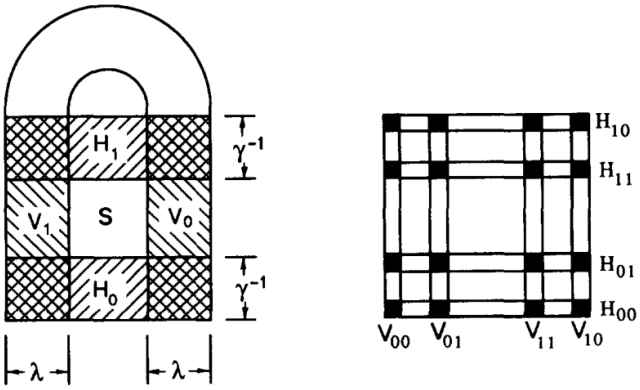

Figure 10: A horseshoe map on a square. Figure taken from [Holmes, 1990, Figure 8a].

3.3 The horseshoe dynamical system

It is inconvenient when points in S moves outside of S, because we have not defined F outside of S, so we restrict our attention to points that never leave S, that is, we consider the invariant set of the horseshoe

$$\Lambda = \{x \in \mathbb{R}^2 | \forall i \in \mathbb{Z}, F^i(x) \in S\} = \bigcap_{i\in\mathbb{Z}} F^i(S). \tag{12}$$

As shown in the picture, $S \cap F(S)$ is made of two vertical stripes V_0, V_1, whose preimages are two horizontal stripes H_0, H_1, that is, $S \cap F^{-1}(S) = H_0 \cup H_1$. After two iterations forwards and backwards, we find that $\cap_{i=-2}^{2} F^i(S)$ is made of 16 little squares, and after a countable number of iterations, we obtain the 2-dimensional Cantor set, shown in Figure 11 .

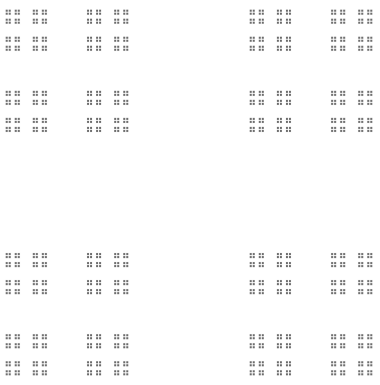

Figure 11: The 2-dimensional Cantor set. Figure from `https://en.wikipedia.org/wiki/File:Cantor_dust.png`.

Thus

Theorem 3.1. *The invariant set Λ of the horseshoe map is homeomorphic to the two-dimensional Cantor set.*

We consider the dynamical system (Λ, F). There is not much of a horseshoe left anymore, just a cloud of Cantor dust folding upon itself repeatedly.

Since F acts on S by stretching the vertical direction and squeezing the horizontal direction, the vertical directions are the "unstable directions", and the horizontal directions are the "stable directions", that is,

Theorem 3.2. *$\Lambda \subset S$ is a hyperbolic set of F, with stable and unstable bundles*

$$E^s = \Lambda \times (\mathbb{R} \times \{0\}), \quad E^u = \Lambda \times (\{0\} \times \mathbb{R}). \tag{13}$$

Now we define a way to encode the orbit of any $x \in \Lambda$ under F as an infinite binary sequence $(\cdots b_{-1} b_0 b_1 \cdots) \in \{0,1\}^{\mathbb{Z}}$. Let b_i be the code of the vertical stripe that $F^i(x)$ belongs to. So for example, knowing that $b_1 = 0$ means $x \in H_0$, and in general, knowing $b_1 \cdots b_n$ allows us to pin down x in a horizontal stripe of height γ^{-n}, and knowing $b_{-n+1} \cdots b_0$ allows us to pin down x in a vertical stripe of width λ^n. Thus, this binary encoding is a bijection between Λ and $\{0,1\}^{\mathbb{Z}}$.

If we impose the product topology of $\{0,1\}^{\mathbb{Z}}$, then the encoding map is a homeomorphism. This can be proved by noting that the topology on Λ has a subbase defined by

$$\{\Lambda \cap H | n \in \mathbb{N}, H \text{ is one of the horizontal stripes making up } S \cap F^{-n}(S)\}$$

$$\cup \{\Lambda \cap V | n \in \mathbb{N}, V \text{ is one of the vertical stripes making up } S \cap F^{n}(S)\},$$

which is mapped by the encoding map to a subbase of $\{0,1\}^{\mathbb{Z}}$

$$\left\{\prod_{i\in\mathbb{Z}} S_i \middle| n\in\mathbb{N}, b_1\cdots b_n\in\{0,1\}, S_i=\{b_i\} \text{ if } i=1,\cdots n, \text{ else } S_i=\{0,1\}\right\}\bigcup$$

$$\left\{\prod_{i\in\mathbb{Z}} S_i \middle| n\in\mathbb{N}, b_0\cdots b_{-n+1}\in\{0,1\}, S_i=\{b_i\} \text{ if } i=-n+1,\cdots 0, \text{ else } S_i=\{0,1\}\right\}.$$

Let the encoding map be $f:\Lambda\to\{0,1\}^{\mathbb{Z}}$, and the left-shift map on $\{0,1\}^{\mathbb{Z}}$ be σ, then we have $f\circ F=\sigma\circ f$, thus we get

Theorem 3.3. *The dynamical system (Λ, F) is topologically conjugate to the binary shift system $(\{0,1\}^{\mathbb{Z}},\sigma)$.*

Thus, we have finally reduced the Smale horseshoe dynamics to a problem of **symbolic dynamics**. We can then easily prove the following:

Theorem 3.4. *The invariant set A of the horseshoe contains: (1) two fixed points, and orbits of every finite period $k\in\mathbb{N}$, and a countable infinity of periodic orbits; (2) an uncountable infinity of nonperiodic orbits, among which are countably many homoclinic points and countably many heteroclinic points, and (3) uncountably many dense orbits.*

Proof We prove these for the binary shift system instead, since they are conjugate.

(1) The only two fixed points are $(0)'$ and $(1)'$ where $()'$ means periodic repeat. Similarly we can obtain periodic points of all orders as $(00\cdots01)'$. In general, every periodic orbit can be specified by a finite binary sequence.

(2) All other points are not periodic, and since $\{0,1\}^{\mathbb{Z}}$ is uncountable, there are an uncountble infinity of nonperiodic points.

A point is a homoclinic point of the fixed point $(1)'$ iff it has both ends being repeated 1, and there are only countably many of these. Similarly for the homoclinic points of $(0)'$, and the two kinds of heteroclinic points.

(3) Enumerate all the countably many finite binary sequences as

$$\cdots b_{-2}, b_{-1}, b_0, b_1, b_2, \cdots$$

then for any $(\cdots c_{-2}c_{-1}c_0c_1c_2\cdots)\in\{0,1\}^{\mathbb{Z}}$,

$$(\cdots c_{-2}b_{-2}c_{-1}b_{-1}c_0b_0c_1b_1c_2\cdots)$$

has an orbit that approaches every point of $\{0,1\}^{\mathbb{Z}}$ arbitrarily closely, thus it has a dense orbit. □

Our construction of the Smale horseshoe dynamical system (Λ, F) is rather unrealistic, since usually when we encounter horseshoes in actual homoclinic tangles, as in Figure 9, the horseshoe is curvy and deformed. However, it can be proved that even if we deform the horseshoe map, the dynamical system is still conjugate to the binary shift system. That is, the horseshoe dynamical system is **structurally stable**.

Proof (sketch) In proving the conjugacy, we only used some qualitative geometric properties of the map, which are preserved under any small perturbation. Thus, the theorem is true for any deformation of the horseshoe map that preserves these qualitative geometric properties. □

So we obtain the general **Smale–Birkhoff homoclinic theorem**:

Theorem 3.5. *Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be a diffeomorphism with a hyperbolic saddle fixed point x, with a homoclinic intersection y. Then, $\exists N \in \mathbb{N}$, F^N has a hyperbolic set Λ, such that (Λ, F^N) is topologically conjugate to the binary shift $(\{0,1\}^{\mathbb{Z}}, \sigma)$.*

A hyperbolic saddle fixed point usually imply a homoclinic intersection, which implies a Smale horseshoe, which has a chaotic dynamic that's conjugate to the binary shift system.

In fact, we can consider horseshoes of far more general shapes, and they generate even richer structures of chaos. A starting point would be [Kennedy et al., 2001]. Cool pictures of such general horseshoes can be found in studies of "pruned horseshoes", such as [de Carvalho and Hall, 2002].

4 Chaos in physics

After such a long trek into the heart and hoof of mathematical chaos, we can get back to physics.

By numerical simulation of the solar system, chaos has been discovered in the motion of the planets, especially Mercury and Mars. In fact, according to [Laskar and Gastineau, 2009], there is a 1% chance that Mercury's eccentricity could exceed 0.7 within 5 billion years, which could cause all kinds of dire outcomes, such as collision with Earth.

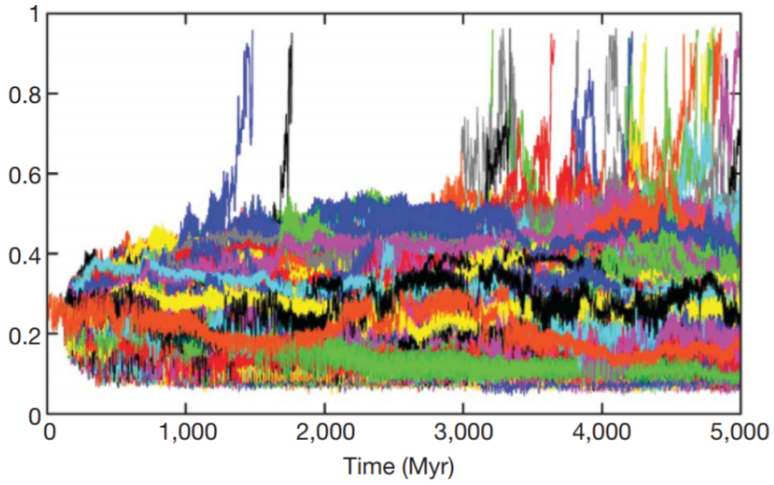

Figure 12: Mercury's eccentricity over 5 billion years can be seen to be chaotic. Each line represents one of the 2501 runs of the simulation, each run differing only slightly in the initial conditions. Figure from [Laskar and Gastineau, 2009].

The tokamak is a device for confining hot plasma with magnetic field, and is one of the leading candidates for fusion energy. The main problem with tokamak is the difficulty in keeping the plasma from hitting the wall, as hot plasma is apt to drift out of confinement. Homoclinic tangles have been numerically predicted to exist [Roeder et al., 2003] and experimentally observed [Evans et al., 2005]. In this way, understanding of homoclinic tangles allows one to better understand the dynamics of confined plasma and the heating of the tokamak wall, and get closer to fusion energy.

Chaos theory became famous in the 1980s, and one main example used to illustrate the chaos in the world is the weather system. Unsurprisingy, horseshoes can be found in the study of hurricanes [Du Toit and Marsden, 2010], which is reminiscent of the classic question of chaos [Lorenz, 1972]:

Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?

Finally, while we have been exclusively dealing with classical mechanics, horseshoes and chaos can also be found in quantum mechanics. Quite notable is [Cvitanović, 1991], which uses the aforementioned "pruned" horseshoes to study classical and quantum chaos.

References

[Barreira and Pesin, 2006] Barreira, L. and Pesin, Y. (2006). Smooth ergodic theory and nonuniformly hyperbolic dynamics. *Handbook of dynamical systems*, 1:57–263.

[Chenciner, 2015] Chenciner, A. (2015). Poincaré and the three-body problem. In *Henri Poincaré, 1912–2012*, pages 51–149. Springer.

[Cvitanović, 1991] Cvitanović, P. (1991). Periodic orbits as the skeleton of classical and quantum chaos. *Physica D: Nonlinear Phenomena*, 51(1-3):138–151.

[de Carvalho and Hall, 2002] de Carvalho, A. and Hall, T. (2002). How to prune a horseshoe. *Nonlinearity*, 15(3):R19.

[Du Toit and Marsden, 2010] Du Toit, P. C. and Marsden, J. E. (2010). Horseshoes in hurricanes. *Journal of Fixed Point Theory and Applications*, 7(2):351–384.

[Evans et al., 2005] Evans, T., Roeder, R., Carter, J., Rapoport, B., Fenstermacher, M., and Lasnier, C. (2005). Experimental signatures of homoclinic tangles in poloidally diverted tokamaks. In *Journal of Physics: Conference Series*, volume 7, page 174. IOP Publishing.

[Holmes, 1990] Holmes, P. (1990). Poincaré, celestial mechanics, dynamical-systems theory and "chaos". *Physics Reports*, 193(3):137–163.

[Hut, 1993] Hut, P. (1993). Binary-single-star scattering. III-numerical experiments for equal-mass hard binaries. *The Astrophysical Journal*, 403:256–270.

[Katok and Hasselblatt, 1995] Katok, A. and Hasselblatt, B. (1995). *Introduction to the modern theory of dynamical systems*, volume 54. Cambridge university press.

[Kennedy et al., 2001] Kennedy, J., Koçak, S., and Yorke, J. A. (2001). A chaos lemma. *The American Mathematical Monthly*, 108(5):411–423.

[Laskar and Gastineau, 2009] Laskar, J. and Gastineau, M. (2009). Existence of collisional trajectories of Mercury, Mars and Venus with the Earth. *Nature*, 459:817–819.

[Lorenz, 1972] Lorenz, E. (1972). Predictability: does the flap of a butterfly's wing in Brazil set off a tornado in Texas? Address at the American Association for the Advancement of Science, 139th Meeting.

[Poincaré, 1899] Poincaré, H. (1899). Les méthodes nouvelles de la mécanique céleste.

[Roeder et al., 2003] Roeder, R., Rapoport, B., and Evans, T. (2003). Explicit calculations of homoclinic tangles in tokamaks. *Physics of Plasmas*, 10(9):3796–3799.

[Smale, 1967] Smale, S. (1967). Differentiable dynamical systems. *Bulletin of the American mathematical Society*, 73(6):747–817.

[Smale, 1998] Smale, S. (1998). Finding a horseshoe on the beaches of Rio. *The Mathematical Intelligencer*, 20(1):39–44.

[Sussman and Wisdom, 2015] Sussman, G. J. and Wisdom, J. (2015). *Structure and interpretation of classical mechanics*. Mit Press.

[Tricoche et al., 2011] Tricoche, X., Garth, C., and Sanderson, A. (2011). Visualization of topological structures in area-preserving maps. *IEEE transactions on visualization and computer graphics*, 17(12):1765–1774.

[Walters, 2000] Walters, P. (2000). *An introduction to ergodic theory*, volume 79. Springer Science & Business Media.