Chapter 7

Chaotic Attractors and Persistent Chaos

Lecture 28

a. Trapping regions. The gap between what has been proved regarding the family of logistic maps and what is believed to be true based on numerical results remains substantial; it becomes even wider when we consider the local map (5.7) for the FitzHugh–Nagumo model.

Despite the lack of rigorous results, the empirical evidence overwhelmingly suggests that what was proved in very restricted circumstances (one-dimensional quadratic maps in a limited parameter range) holds much more generally, as is suggested, for example, by the bifurcation diagram in Figure 5.6. We see a period-doubling cascade leading to the onset of chaos at A_∞, beyond which there are windows of stability surrounded by maps with chaotic behaviour (although as mentioned at the end of the previous chapter, even within these windows of stability the map exhibits transient chaos).

Throughout all this, the map $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ has three fixed points, $\mathbf{0}$, $\mathbf{p}_1$, and $\mathbf{p}_2$, two of which began life as stable fixed points, and then lost their stability at the bifurcation values A_1 and A_1'. During the period-doubling cascade, all trajectories wind up approaching a

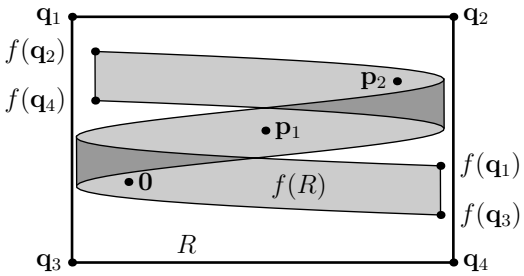


Figure 7.1. A trapping region for the FitzHugh–Nagumo map.

stable periodic orbit of length 2^n; after the onset of chaos, however, there are no stable periodic orbits to approach, except in the windows of stability. So where do the orbits go?

Even if trajectories of f are not asymptotically stable or periodic, it may happen that they are confined within some bounded region. This is certainly the case if there is some bounded region $R \subset \mathbb{R}^2$ for which $f(R) \subset R$, since then the forward trajectory of any point in R remains in R. The following definition codifies this idea, adding a small topological requirement.

Definition 7.1. An open set R is a *trapping region* for the map f if $\overline{R}$ is compact and $\overline{f(R)} \subset R$.

We will explore the dynamical consequences of the existence of a trapping region after first proving that such a region exists for the FitzHugh–Nagumo map with certain values of A.

Proposition 7.2 (Orendovici–Pesin)**.** *Fix parameters $0 < \gamma < 1$ and $\beta > 0$. There exists an open rectangle $R = (a, b) \times (c, d) \subset \mathbb{R}^2$ and $A' > 0$ such that for $\alpha > 0$ sufficiently small, θ sufficiently near $1/2$, and $0 < A < A'$, the rectangle R is a trapping region for the local map* (5.7) *for the FitzHugh–Nagumo model. Furthermore, R contains the three fixed points $\mathbf{0}$, $\mathbf{p}_1$, and $\mathbf{p}_2$.*

Proof. *(See* [**OP00**] *and* [**PY04**]*.)* We will guarantee containment of the fixed points by taking $a < 0 < 1 < b$ and appropriate values of c and d.

Let $\mathbf{q}_1 = (a, d)$, $\mathbf{q}_2 = (b, d)$, $\mathbf{q}_3 = (a, c)$, and $\mathbf{q}_4 = (b, c)$ be the four corners of R, as in Figure 7.1. In order for R to be a trapping region, $f(\mathbf{q}_2)$ must lie below the line $v = d$ and $f(\mathbf{q}_3)$ must lie above the line $v = c$. That is, we require that

$$\beta b + \gamma d < d, \qquad \beta a + \gamma c > c,$$

or equivalently,

$$d > \frac{\beta b}{1-\gamma}, \qquad c < \frac{\beta a}{1-\gamma}.$$

Thus $\mathbf{q}_2$ must lie above the line $v = \beta u/(1-\gamma)$ and $\mathbf{q}_3$ must lie below it; observe that this is the line which passes through the three fixed points 0, $\mathbf{p}_1$, and $\mathbf{p}_2$, and hence for $a < 0 < 1 < b$ all three fixed points lie in R.

In order to guarantee that R is a trapping region, it remains to show that the images of both the top and bottom edges of R are themselves in R. If we write $k(t) = t - At(t-\theta)(t-1)$, then these images are as follows:

$$\{(k(u) - \alpha d, \beta u + \gamma d) \mid a \le u \le b\},$$
$$\{(k(u) - \alpha c, \beta u + \gamma c) \mid a \le u \le b\}.$$

That the v-coordinate will lie between c and d follows from the linearity of the v-component of the map and the fact that the corners $\mathbf{q}_i$ are mapped into R. Thus we require only the following inequalities for every $u \in [a, b]$:

$$a < k(u) - \alpha d < k(u) - \alpha c < b. \tag{7.1}$$

The idea now is to prove these in the unperturbed case where $\alpha = 0$ and $\theta = 1/2$, and then to use a continuity argument to extend the result to small values of α and values of θ near $1/2$. In the unperturbed case, (7.1) amounts to choosing a and b such that

$$a < k(t_1) < k(t_2) < b, \tag{7.2}$$

$$a < k(b), \tag{7.3}$$

$$k(a) < b, \tag{7.4}$$

where t_1 and t_2 are the unique local minimum and local maximum, respectively, of the cubic polynomial k. We see immediately from the form of $k(t)$ that $k(t_1)$ decreases as A increases and $k(t_2)$ increases

as A increases; a little computation shows that for $A = 8$ we have $k(t_1) \approx -.207$ and $k(t_2) \approx 1.207$. Thus taking $a = -.21$ and $b = 1.21$, we see that (7.2) is satisfied for $0 < A < 8$.

Now the inequality $k(a) < b$ may be written as

$$a - Aa(a-\theta)(a-1) < b,$$

or equivalently,

$$A < \frac{a-b}{a(a-\theta)(a-1)} \approx 7.98,$$

where the computation is for the particular case $\theta = 1/2$. Similarly, $k(b) < a$ is equivalent to

$$A < \frac{b-a}{b(b-\theta)(b-1)} \approx 7.98.$$

Thus taking $A' = 7.5$ to give ourselves a bit of room to play with, we see that (7.2)–(7.4) all hold for every $0 < A < A'$, with $\theta = 1/2$ and $\beta = 0$, and hence R is a trapping region for these particular parameter values.

Finally, note that all the inequalities which guarantee that R is a trapping region involve only continuous functions of the parameters; in particular, they all hold for values of θ sufficiently close to $1/2$ and values of $\beta > 0$ sufficiently small. □

b. Attractors. What are the dynamical implications of the existence of a trapping region? First, we observe that a trajectory which enters a trapping region will never leave it—hence the name. Furthermore, the property $\overline{f(R)} \subset R$ is inherited by the images of R, thanks to the following exercise.[1]

Exercise 7.1. Given a continuous map f and an arbitrary domain R such that $\overline{R}$ is compact, show that $\overline{f(R)} = f(\overline{R})$.

Using the result of the exercise, we see that

$$\overline{f^2(R)} = \overline{f(f(R))} = f(\overline{f(R)}) \subset f(R).$$

Continuing in this way, we obtain a nested sequence of compact sets

$$\overline{R} \supset \overline{f(R)} \supset \overline{f^2(R)} \supset \cdots \supset \overline{f^n(R)} \supset \overline{f^{n+1}(R)} \supset \cdots .$$

[1]Observe that $f(R)$ may not be open (see Figure 7.1) and hence may not be a trapping region in its own right.

We may "take the limit" of this sequence by taking the intersection of all these sets, and we obtain

$$\Lambda = \bigcap_{n\geq 0} f^n(R). \tag{7.5}$$

The intersection Λ is called an *attractor* for the map f.

Theorem 7.3. *Let R be a trapping region for a continuous map f, and define an attractor Λ by (7.5). Then Λ has the following properties:*

(1) Λ is compact and non-empty.

(2) Λ is f-invariant: $f(\Lambda) = \Lambda$.

(3) Λ is the largest f-invariant subset of R; that is, $Z \subset \Lambda$ for every f-invariant set $Z \subset R$.

(4) Λ attracts every orbit of f which enters R: $\omega(x) \subset \Lambda$ for every $x \in R$.

(5) Λ contains all the fixed points and periodic points of f in R.

Proof.

(1) Λ is the intersection of nested compact sets, and hence is compact and non-empty.

(2) Since $f(\overline{R}) \subset R$, we have

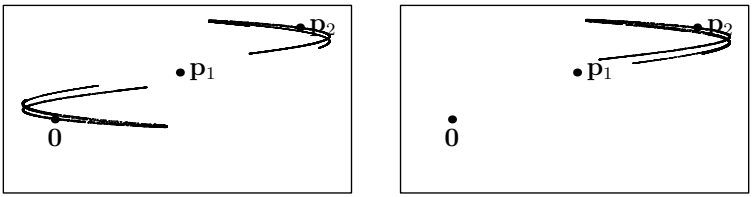
$$f(\Lambda) = f\left(\bigcap_{n\geq 0} f^n(\overline{R})\right) = \bigcap_{n\geq 0} f(f^n(\overline{R})) = \bigcap_{n\geq 1} f^n(\overline{R}) = \Lambda.$$

(3) If $Z \subset R$ is f-invariant ($f(Z) = Z$), then for every $n \in \mathbb{N}$ we have $Z = f^n(Z) \subset f^n(R)$, and hence $Z \subset \Lambda$.

(4) The set $\omega(x)$ is f-invariant, so by Property (3), $\omega(x) \subset \Lambda$.

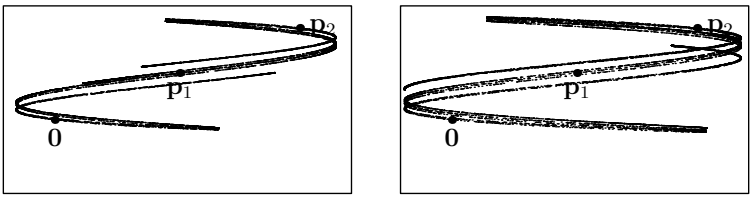
(5) If $p \in R$ is a fixed point, then $\{p\}$ is f-invariant, so Property (3) applies. Similarly, if $f^k(p) = p$, then $\{p, f(p), \ldots, f^{k-1}(p)\}$ is f-invariant. □

Exercise 7.2. Let R be an open domain such that $\overline{f(R)} \subset \overline{R}$. Show that Theorem 7.3 remains true in this case.



$A = 6.1$ $A = 6.4$

Figure 7.2. The attractor for the FitzHugh–Nagumo map with $\alpha = .01, \beta = .02, \theta = .51, \gamma = .2$, and varying A.



$A = 6.7$ $A = 7.5$

Figure 7.3. Changes in the attractor as A increases.

We return now to the specific example of the FitzHugh–Nagumo system, for which Proposition 7.2 guarantees the existence of a trapping region R and, hence, an attractor Λ. Figure 7.2 shows a numerically computed approximation to the attractor for two different parameters of A, drawn by plotting long orbit segments to approximate $\omega(x)$. Notice that for certain values of A, corresponding to the gap in the bottom half of Figure 5.6, the attractor lies entirely in the top right quadrant of the trapping region.

As A continues to increase, the attractor "grows", as shown in Figure 7.3, to occupy more and more[2] of the trapping region R; here one also sees the "grainy" structure which is inevitably associated with the orbit-plotting method of producing such images. The human eye, on viewing Figure 7.3, immediately wants to connect the dots and view Λ as a union of curves, rather than simply a collection of points.

[2]One may legitimately ask, though, in just what sense it becomes larger, beyond the obvious statement that its diameter increases. Throughout all of this, the Lebesgue measure of Λ is zero; thus a useful quantification of the attractor's size ought to involve some dimensional quantity.

Indeed, this is exactly what one ought to do. For the values of A where the attractor shown in Figures 7.2 and 7.3 appears, all three fixed points are hyperbolic; that is, they have one expanding and one contracting direction. We saw in Lecture 22 that for each such fixed point $\mathbf{p}$ there exists a local unstable curve W^u_ε through $\mathbf{p}$ which is tangent to the unstable eigenvector. This curve is expanding in the sense that the image $f(W^u_\varepsilon)$ is a curve which contains W^u_ε.

Taking the union of the curves $f^n(W^u_\varepsilon)$, we obtain, as we did before, the global unstable curve W^u, which is f-invariant and contained in R; by Theorem 7.3, this implies that $W \subset \Lambda$.

In fact, it is conjectured (and widely believed) that for an appropriate range of the parameters, W^u is dense in Λ for the FitzHugh–Nagumo system; however, this remains an open problem.

Lecture 29

a. The Smale–Williams solenoid. From the discussion in the previous lecture, it is apparent that the FitzHugh–Nagumo model is very rich in intricate and interesting behaviour, but is also quite difficult to analyse. We thus turn our attention to simpler model examples, which exhibit a similar richness of behaviour but are rather more tractable.

Our first such example is a map from the solid torus to itself. Abstractly, the solid torus is

$$P = D^2 \times S^1,$$

the direct product of a disc and a circle. Writing the disc as

$$D^2 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\},$$

and the circle as $S^1 = \mathbb{R}/2\pi\mathbb{Z}$, we may use coordinates (x, y, θ) on P; x and y give the coordinates on the disc, and θ is the angular coordinate on the circle.

We may visualise P via its embedding in $\mathbb{R}^3$ as the standard torus of revolution together with the region it encloses:

$$\rho(P) = \left\{ (x, y, z) \in \mathbb{R}^3 \;\middle|\; \left(\sqrt{x^2 + y^2} - 2\right)^2 + z^2 \leq 1 \right\};$$

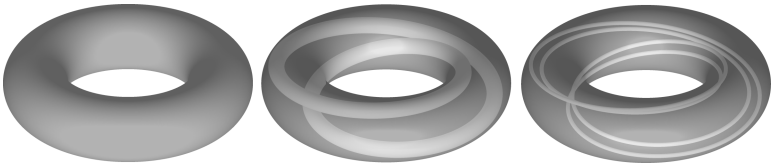


Figure 7.4. A map from the solid torus to itself.

here $\rho\colon D^2 \times S^1 \to \mathbb{R}^3$ is the map given by

$$\rho(x, y, \theta) = ((2 + x)\cos\theta, (2 + x)\sin\theta, y).$$

Fixing parameters $r \in (0, 1)$ and $\alpha, \beta \in (0, \min\{r, 1 - r\})$, we define a map $f\colon P \to P$ by

$$f(x, y, \theta) = (\alpha x + r\cos\theta, \beta y + r\sin\theta, 2\theta). \tag{7.6}$$

The images of P under the first two iterates f and f^2 are shown in Figure 7.4. The action of f on P may be described as follows:

(1) Take the torus and slice it along a disc so that it becomes a tube.
(2) Squeeze this tube so that its cross-sections are no longer circles of radius 1, but ellipses with axes of length α and β.
(3) Stretch the tube along its axis until it is twice its original length.
(4) Wrap the resulting longer, skinnier tube twice around the z-axis within the original solid torus.
(5) Glue the ends of the tube together.

We see that $f(P) \subset \operatorname{int} P$, and so we may repeat the procedure in the previous section, obtaining an attractor by taking the intersection of all images of P:

$$\Lambda = \bigcap_{n \geq 0} f^n(P). \tag{7.7}$$

The attractor Λ is known as the *Smale–Williams solenoid*; in order to investigate the structure of Λ, we look at a vertical cross-section of the solid torus $P = D^2 \times S^1$ by fixing the angular coordinate θ and considering the disc $D^2 \times \{\theta\}$.

From Figure 7.4, it is clear that the image $f(P)$, which is a long skinny tube wrapped twice around the z-axis, intersects this disc in

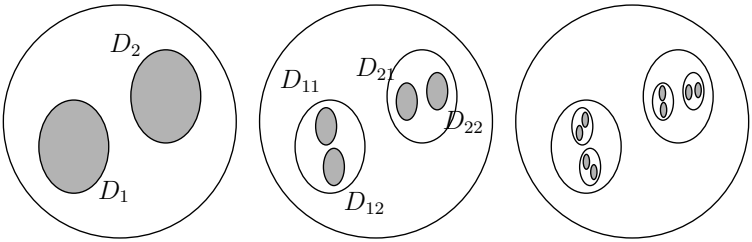


Figure 7.5. A cross-section of the Smale–Williams solenoid.

two ellipses D_1 and D_2, whose axes have length α and β (see Figure 7.5). The second image $f^2(P)$ is an even longer and skinnier tube which is wrapped *four* times around the z-axis, and intersects the disc in four ellipses D_{11}, D_{12}, D_{21}, and D_{22}, whose axes have lengths α^2 and β^2.

Continuing in this manner, we see that $f^n(P) \cap (D^2 \times \{\theta\})$ is the union of 2^n ellipses $D_{w_1 \ldots w_n}$ whose axes have lengths α^n and β^n. By now the reader should not be too shocked to discover that this is yet another example of a Cantor-like construction;[3] the basic sets at each step are the ellipses just mentioned, and the cross-section $C = \Lambda \cap (D^2 \times \{\theta\})$ is a Cantor set obtained as the intersection of the basic sets at all levels.

Exercise 7.3. Consider the cross-section $\theta = 0$ of the solid torus P, and describe the location of the centres of the ellipses $D_{w_1 \ldots w_n}$.

Each basic set is the intersection of a tube with the disc $D^2 \times \{\theta\}$; as n increases, the diameters of the tubes decrease exponentially, and so upon passing to the limit set C, we see that each point in C is contained in precisely one curve which meets $D^2 \times \{\theta\}$ transversely (indeed, orthogonally). Thus in a neighbourhood of each cross-section (a slice out of the torus), the attractor is the direct product $C \times (-\varepsilon, \varepsilon)$. However, this product structure is only local; if we follow one of these curves all the way around the torus, we will in general return to a different point of C than the one we left (see Figure 7.6).

[3] Indeed, one could obtain the exact construction shown in Figure 1.20 by modifying f so that $f(P)$ wraps around the z-axis *three* times, and allowing $\alpha = \beta$ to depend on θ.

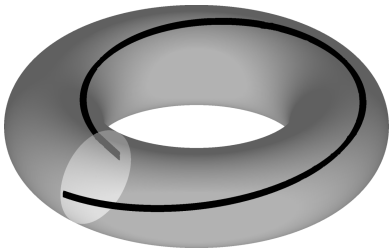


Figure 7.6. Following an unstable curve once around the torus.

The local product structure of the attractor Λ has more than just a geometric significance; it also helps us describe the dynamics of the map f. Through each point $\mathbf{p} = (x, y, \theta) \in \Lambda$, we have a disc $W^s = D^2 \times \{\theta\}$ and a curve $W^u_\varepsilon = \{(x, y)\} \times (-\varepsilon, \varepsilon)$, as shown in Figure 7.6. The former is contracting while the latter is repelling, as follows: given $\mathbf{q} \in W^s$, we have

$$d(f(\mathbf{p}), f(\mathbf{q})) \leq \max\{\alpha, \beta\} d(\mathbf{p}, \mathbf{q}),$$

while for $\mathbf{q}' \in W^u_\varepsilon$, the orbits are driven further apart:

$$d(f(\mathbf{p}), f(\mathbf{q}')) = 2d(\mathbf{p}, \mathbf{q}').$$

Thus every point looks like a saddle; it has two stable directions (forming the disc) and one unstable direction (the curve). Notice, however, that since $\mathbf{p}$ may not be fixed, the reference to which or from which the orbit $\{f^n(\mathbf{q})\}$ is attracted or repelled is not the point $\mathbf{p}$ itself, but the trajectory of $\mathbf{p}$.

b. Uniform hyperbolicity. The behaviour exhibited by the Smale–Williams solenoid Λ, wherein hyperbolic behaviour exists at *every* point, not just fixed points, is an important enough and widespread enough phenomenon to warrant the following general definition.

Definition 7.4. Let $U \subset \mathbb{R}^d$ be open, and let $f\colon U \to f(U)$ be $\mathcal{C}^1$ with $\mathcal{C}^1$ inverse (see Appendix). A compact f-invariant set $\Lambda \subset U$ is called *hyperbolic* if for every $\mathbf{x} \in \Lambda$, there exists a direct sum decomposition $\mathbb{R}^d = E^s(\mathbf{x}) \oplus E^u(\mathbf{x})$ such that the subspaces $E^s(\mathbf{x})$ and $E^u(\mathbf{x})$ have the following properties:

(1) *Uniform contraction/expansion:* There exist $\lambda \in (0,1)$ and $C > 0$, independent of $\mathbf{x}$, such that for every $n \geq 0$, $\mathbf{v}^s \in E^s(\mathbf{x})$, and $\mathbf{v}^u \in E^u(\mathbf{x})$, we have

$$\|Df^n(\mathbf{x})\mathbf{v}^s\| \leq C\lambda^n\|\mathbf{v}^s\|,$$
$$\|Df^{-n}(\mathbf{x})\mathbf{v}^u\| \leq C\lambda^n\|\mathbf{v}^u\|.$$

(2) *Invariance of stable and unstable subspaces:* For every $\mathbf{x} \in \Lambda$, we have $Df(\mathbf{x})E^s(\mathbf{x}) = E^s(f(\mathbf{x}))$ and $Df(\mathbf{x})E^u(\mathbf{x}) = E^u(f(\mathbf{x}))$.

(3) There exists an open neighborhood U of Λ such that

$$\Lambda = \bigcap_{n\in\mathbb{Z}} f^n(U).$$

Roughly speaking, this definition says that the hyperbolicity we previously observed at hyperbolic fixed points can be found at every point of Λ. (Indeed, if $\Lambda = \{\mathbf{p}\}$ is a single fixed point, then the definition reduces to the definition of a hyperbolic fixed point.)

The first condition states that the map is contracting in the direction of E^s and expanding in the direction of E^u (since contraction along backward orbits corresponds to expansion along forward orbits). The second condition accounts for the fact that although most points $\mathbf{x} \in \Lambda$ are not fixed, the directions in which expansion and contraction occur should still be consistent along an orbit.

The third condition means that Λ is *locally maximal*—that is, if $Z \subset U$ is an invariant set, then $Z \subset \Lambda$. Traditionally, the definition of hyperbolic sets includes only the first two conditions, but many principal results (for example, Theorem 7.7 below) require the third condition as well, and so we will only consider locally maximal hyperbolic sets.

The hyperbolicity just described represents a fundamentally new type of behaviour compared with the Morse–Smale systems we found for the logistic map in the period-doubling cascade, where we observed hyperbolic behaviour only at a finite number of fixed points. Here, by contrast, the hyperbolicity is ubiquitous.

Definition 7.5. A hyperbolic set Λ is a *hyperbolic attractor* if there exists an open set $U \supset \Lambda$ such that $\Lambda = \bigcap_{n\geq 0} f^n(U)$. The open set $\bigcup_{n\geq 0} f^{-n}(U)$ is the *basin of attraction* for Λ.

Figure 7.7. The north-south map.

The Smale–Williams solenoid is an important example of a hyperbolic attractor; the basin of attraction in this case is the entire solid torus.

Exercise 7.4. The "north-south" map f of the unit sphere is defined so that trajectories move in the directions shown in Figure 7.7, from the north pole (which is a repelling fixed point) to the south pole (which is an attracting fixed point). Find the largest hyperbolic set and the attractor for f. Describe the basin of attraction.

What does the pervasive hyperbolicity just described mean for the dynamics of f? If $\mathbf{p}$ and $\mathbf{q}$ are two points in Λ which do not lie on the same stable disc W^s or unstable curve W^u_ε, then repeated iteration by f will decrease the distance between $f^n(\mathbf{p})$ and $f^n(\mathbf{q})$ in the stable direction (corresponding to the coordinates x and y) but will increase it in the unstable direction (corresponding to θ). In particular, the trajectory of $\mathbf{q}$ is repelled from the trajectory of $\mathbf{p}$.

So almost every pair of trajectories moves apart under the action of f; however, Λ is bounded, so they cannot move too far apart. Indeed, the definition of a hyperbolic set is given in terms of local properties (expanding and contracting subspaces for the linear map $Df(\mathbf{x})$), and so it ceases to give any information about the relationship of the two trajectories once they are no longer close. After they separate, a similar line of reasoning shows that the trajectory of $\mathbf{q}$ is constantly being repelled from whatever trajectories it finds itself near at any given time, and eventually is repelled back towards the

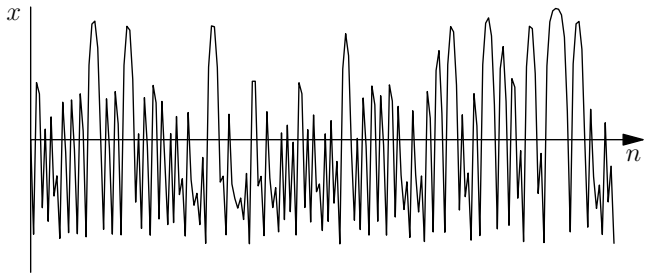


Figure 7.8. The x-coordinates of a trajectory as a chaotic signal.

trajectory of $\mathbf{p}$; at this point it is once again repelled from the trajectory of $\mathbf{p}$, and the whole cycle repeats itself.

This behaviour, this unending dispersal and return, is characteristic of hyperbolic dynamics. If we plot the x-coordinate of the trajectory of a point $\mathbf{p} \in \Lambda$ as a function of n, we see a chaotic signal, without periodicity or pattern, as shown in Figure 7.8. Various quantitative properties of this signal and of the dispersal and return of orbits of f are related to the dimensional quantities we have studied. For example, the rate at which nearby trajectories are repelled is given by the Lyapunov exponent of the map, which relates the entropy and the dimension. Furthermore, the statistical properties of recurrence (the times at which the trajectory beginning at $\mathbf{p}$ returns to a neighbourhood of $\mathbf{p}$), and of the correlations between measurements of a chaotic signal at different times, turn out to be related to various dimensional quantities of the hyperbolic set, and also to the multifractal analysis mentioned at the end of Chapter 4.

c. Symbolic dynamics. Another fundamental manifestation of the chaotic nature of the map $f\colon \Lambda \to \Lambda$ is its connection to symbolic dynamics. In Lecture 15 we saw that one-dimensional expanding Markov maps can be modeled by one-sided subshifts of finite type; this correspondence put at our disposal all the machinery of symbolic dynamics, allowing us to investigate the original map in terms of entropy, invariant measures, and so on.

These expanding maps had the characteristics of uniformly hyperbolic behaviour but were non-invertible; hence the one-sided shift

spaces were good models for their dynamics. It turns out that if we use a *two-sided* shift space, on which the shift map is invertible, then a similar correspondence is possible for invertible uniformly hyperbolic maps.

Definition 7.6. Given $k \in \mathbb{N}$, the *two-sided symbolic space on k symbols* is the space

$$\begin{aligned}\Sigma_k &= \{1, \dots, k\}^{\mathbb{Z}} \\ &= \{w = (w_j)_{j=-\infty}^{\infty} \mid w_j \in \{1, \dots, k\} \text{ for every } j \in \mathbb{Z}\},\end{aligned}$$

where it is convenient to write members of Σ_k in the form

$$w = (\dots, w_{-2}, w_{-1} | w_0 | w_1, w_2, \dots)$$

in order to highlight which entry of the sequence is the "centre" (this was not necessary in the one-sided case because the sequence had a beginning). The *full two-sided shift on k symbols* is Σ_k together with the shift map

$$\begin{aligned}\sigma \colon \Sigma_k &\to \Sigma_k, \\ (\dots, w_{-2}, w_{-1} | w_0 | w_1, w_2, \dots) &\mapsto (\dots, w_{-1}, w_0 | w_1 | w_2, w_3, \dots).\end{aligned}$$

As in the one-sided case, we may fix $a > 1$ and define a metric on Σ_k by

$$d_a(w, w') = \sum_{j \in \mathbb{Z}} \frac{|w_j - w'_j|}{a^{|j|}}.$$

This metric once again induces a topology in which the open sets are unions of cylinders;[4] the latter take the form

$$C_{w_{-n} \dots w_{-1} | w_0 | w_1 \dots w_n} = \{w' \in \Sigma_k \mid w'_j = w_j \text{ for all } -n \le j \le n\}.$$

More generally, we may consider cylinders of the form

$$C_{w_a \dots w_b} = \{w' \in \Sigma_k \mid w'_j = w_j \text{ for all } a \le j \le b\},$$

where $a \le b \in \mathbb{Z}$ are arbitrary.

[4] As in the one-sided case, cylinders may not actually be open balls for small values of a. In this case, we need to take $a > 3$ to guarantee that $B(w, a^{-n}) = C_{w_{-n} \dots w_{-1} | w_0 | w_1 \dots w_n}$.

For the one-sided shift, every sequence has k preimages, because the shift "forgets" the first element of the sequence. The fundamental novelty of the two-sided shift is that nothing is forgotten; all the elements of the sequence remain, with a shifted reference point, and so the map is invertible. This makes the two-sided shift well suited for modeling invertible maps, such as the Smale–Williams solenoid, while the one-sided shift is well suited for modeling non-invertible maps, such as one-dimensional Markov maps.

The idea, then, is to partition Λ into disjoint sets $\Lambda_1, \dots, \Lambda_k$, and to code trajectories of f by recording which partition element the iterate $f^n(x)$ lands in. (This is exactly what was done for one-dimensional Markov maps, using the forward trajectory with $n \geq 0$.) Thus to a trajectory $\{f^n(x)\}$ we associate the sequence

$$w = (\dots, w_{-2}, w_{-1} | w_0 | w_1, w_2, \dots) \in \Sigma_k,$$

where w_j is such that $f^j(x) \in \Lambda_{w_j}$ for each $j \in \mathbb{Z}$. Conversely, we may begin with a sequence $w \in \Sigma_k$ and look for a point whose trajectory is given by w; the set of all such points is

$$\bigcap_{j \in \mathbb{Z}} f^{-j}(\Lambda_{w_j}). \tag{7.8}$$

What sequences in Σ_k do we obtain as codings of trajectories in Λ? The answer depends on which partition we choose; for example, if $i, j \in \{1, \dots, k\}$ are such that $f(\Lambda_i) \cap \Lambda_j = \emptyset$, then no sequence w which contains the symbol i followed by the symbol j can correspond to a trajectory in Λ; these sequences are not admissible.

In general, it is not possible to find a partition such that all sequences are admissible, just as we found for one-dimensional Markov maps, where not every map could be modeled by the full shift. However, there is the following remarkable result: for a uniformly hyperbolic map, it *is* possible to find sets $\{\Lambda_1, \dots, \Lambda_k\}$ (with disjoint interiors) such that for the $k \times k$ transition matrix A given as in (3.15) by

$$a_{ij} = \begin{cases} 0 & f(\Lambda_i) \cap \Lambda_j = \emptyset, \\ 1 & f(\Lambda_i) \cap \Lambda_j \neq \emptyset, \end{cases} \tag{7.9}$$

the admissible sequences are precisely those which lie in

$$\Sigma_A = \{w \in \Sigma_k \mid a_{w_j w_{j+1}} = 1 \text{ for every } j \in \mathbb{Z}\}.$$

This is made precise by the following theorem, which relates the map $f\colon \Lambda \to \Lambda$ to the two-sided subshift $\sigma\colon \Sigma_A \to \Sigma_A$, and which is the fundamental vehicle for most of what is known about dynamics on uniformly hyperbolic sets.

Theorem 7.7. *Let Λ be a hyperbolic set for f. Then there exists a cover of Λ by sets $\Lambda_1, \ldots, \Lambda_k \subset \Lambda$ with disjoint interiors such that for the $k \times k$ transition matrix A given by* (7.9) *and the coding map $h\colon \Sigma_A \to \Lambda$ given by* (7.8)*, the following hold:*

(1) h is continuous, onto, and one-to-one on a residual set.[5]

(2) The following diagram commutes:

$$\begin{array}{ccc} \Sigma_A & \xrightarrow{\ \sigma\ } & \Sigma_A \\ \big\downarrow{\scriptstyle h} & & \big\downarrow{\scriptstyle h} \\ \Lambda & \xrightarrow{\ f\ } & \Lambda \end{array} \tag{7.10}$$

Proof. See [**KH95**] or [**BS02**]. □

This correspondence between hyperbolic sets and subshifts of finite type provides a bridge via which key symbolic results can be transported to the hyperbolic regime. In this manner, symbolic dynamics can be used to establish many properties of hyperbolic maps which are characteristic of chaos (see Lecture 31(c)).

Lecture 30

a. Dimension of direct products. Having discussed some of the qualitative properties of hyperbolic attractors, such as the Smale–Williams solenoid Λ, and the implications of these properties for the dynamics of f, we turn our attention to quantitative questions. In particular, since Λ has a fractal structure, we ask the natural question: what is the Hausdorff dimension of Λ?

[5] Recall that a set is *residual* if it is a countable intersection of open dense sets, and hence comprises "almost everything" in a topological sense. In this case, the residual set in question is the union $\bigcup_{n\in\mathbb{Z}} f^n(B)$, where B is the union of the boundaries of the sets Λ_i.

Locally, Λ is a direct product of a Cantor set and an interval; since we know the Hausdorff dimension for both of these sets, we would like to have a general expression for $\dim_H(A \times B)$ in terms of $\dim_H A$ and $\dim_H B$.

Intuitively, we expect dimension to be additive with respect to direct products; after all, the direct product of $\mathbb{R}^d$ and $\mathbb{R}^p$ is $\mathbb{R}^{d+p}$, and so it seems natural to conjecture that in general,

$$\dim_H(A \times B) = \dim_H A + \dim_H B. \tag{7.11}$$

Remark. When A and B are subsets of the Euclidean spaces $\mathbb{R}^d$ and $\mathbb{R}^p$, their direct product $A \times B$ lies in $\mathbb{R}^{d+p}$, and hence inherits the Euclidean metric. From an abstract point of view, when A and B are arbitrary metric spaces, there are a number of natural metrics that one might use. Chief among these are the following: here $x, x' \in A$ and $y, y' \in B$, and d_A and d_B are the metrics on A and B, respectively.

$$\begin{aligned} d_1((x,y),(x',y')) &= d_A(x,x') + d_B(y,y'), \\ d_2((x,y),(x',y')) &= \sqrt{d_A(x,x')^2 + d_B(y,y')^2}, \\ d_\infty((x,y),(x',y')) &= \max\{d_A(x,x'), d_B(y,y')\}. \end{aligned}$$

The standard metric that we obtain in the Euclidean case is d_2; however, if we endow Euclidean space with the alternate metric (2.11), we obtain d_1, and similarly (2.12) leads to d_∞. As in Exercise 2.5, the three metrics d_1, d_2, and d_∞ are strongly equivalent, and hence any of the three may be used for computations of Hausdorff dimension.

Exercise 7.5. Using the product measures $m_H(\cdot,\alpha) \times m_H(\cdot,\beta)$, show that for any two sets A and B,

$$\dim_H(A \times B) \geq \dim_H A + \dim_H B. \tag{7.12}$$

Exercise 7.5 establishes one half of (7.11). However, the reverse inequality is not true in general; a counterexample to this effect was first produced by Besicovitch.[6]

Example 7.8. We produce two sets $A, B \subset [0,1]$ which both have Hausdorff dimension equal to 0, but are large enough that

$$A + B = \{x + y \mid x \in A, y \in B\} \supset [0,1].$$

[6]The example we give here is slightly less general than the one given in [**BM45**], but it follows the same idea.

Since the map $f\colon \mathbb{R}^2 \to \mathbb{R}$ given by $f(x,y) = x + y$ is Lipschitz, this will imply that

$$\dim_H(A \times B) \geq \dim_H f(A \times B) = \dim_H(A + B) \geq \dim_H[0,1] = 1.$$

To construct A and B, we first fix an increasing sequence of positive integers n_k, which is to satisfy a certain growth condition to be defined below. We write $x = 0.x_1x_2x_3\cdots$ for the binary expansion of $x \in [0,1]$; if x is a dyadic rational, so that its binary expansion terminates in an infinite string of zeros or ones, we choose the expansion which ends in zeros. Now define sets $Z_k \subset [0,1]$ as follows:

$$Z_k = \{x \in [0,1] \mid x_i = 0 \text{ for all } n_k < i \leq n_{k+1}\}.$$

Finally, define A and B by

$$A = Z_1 \cap Z_3 \cap Z_5 \cap \cdots,$$
$$B = Z_2 \cap Z_4 \cap Z_6 \cap \cdots.$$

That is, if we think of the sequence $\{n_k\}$ as partitioning $\mathbb{N}$ into a sequence of intervals $(n_k, n_{k+1}]$, then A is the set of numbers whose binary digits x_i are zero whenever i lies in an odd interval; similarly, numbers in B have binary expansions which vanish on the even intervals.

It is immediately apparent that any $w \in [0,1]$ may be written as $w = x + y$ where $x \in A$ and $y \in B$; simply take x and y to be the numbers whose binary expansions are given by

$$x_i = \begin{cases} 0 & i \in (n_{2k-1}, n_{2k}], \\ 1 & i \in (n_{2k}, n_{2k+1}], \end{cases} \qquad y_i = \begin{cases} 1 & i \in (n_{2k-1}, n_{2k}], \\ 0 & i \in (n_{2k}, n_{2k+1}]. \end{cases}$$

Thus we need only choose n_k such that $\dim_H A = \dim_H B = 0$. Observe that for odd values of k, the set $Z_1 \cap Z_3 \cap \cdots \cap Z_k$ is a union of 2^{m_k} intervals of length 2^{-n_k}, where

$$m_k = (n_2 - n_1) + (n_4 - n_3) + \cdots + (n_{k-1} - n_{k-2}).$$

Since $A \subset Z_1 \cap Z_3 \cap \cdots \cap Z_k$ for every odd k, we thus have a family of covers of A by intervals of length 2^{-n_k}, whence

$$\dim_H A \leq \underline{\dim}_B A \leq \varliminf_{k\to\infty} \frac{\log 2^{m_k}}{-\log 2^{-n_k}} = \varliminf_{k\to\infty} \frac{m_k}{n_k}.$$

Since m_k only depends on $n_1, \dots, n_{k-1}$ and not on n_k itself, we may choose n_k to be increasing rapidly enough that this last quantity tends to 0, and hence $\dim_H A = 0$. Similar considerations give us $\dim_H B = 0$, and hence we have strict inequality in (7.12).

Besicovitch also proved that (7.11) *does* hold under an additional assumption on the sets involved.

Theorem 7.9. *If A is such that $\dim_H A = \underline{\dim}_B A = \overline{\dim}_B A$, then equality holds in* (7.11) *for any B.*

Proof. This is a consequence of Exercise 7.5 and the following general inequality, which holds for arbitrary A and B:

$$\dim_H(A \times B) \le \overline{\dim}_B A + \dim_H B. \tag{7.13}$$

To prove (7.13), we fix $s > \overline{\dim}_B A$ and $t > \dim_H B$; thus

$$\overline{\lim_{\varepsilon \to 0}} \frac{\log N(A,\varepsilon)}{-\log \varepsilon} < s,$$
$$\lim_{\varepsilon \to 0} m_H(B,t,\varepsilon) = 0.$$

In particular, there exists $\varepsilon_0 > 0$ such that for all $0 < \varepsilon < \varepsilon_0$,

$$N(A,\varepsilon) < \varepsilon^{-s}, \tag{7.14}$$
$$m_H(B,t,\varepsilon) < 1. \tag{7.15}$$

Now by (7.15) there exists an ε-cover $\{U_i \mid i \in \mathbb{N}\}$ of B such that $\sum_i (\operatorname{diam} U_i)^t < 1$. For each i, (7.14) guarantees the existence of a cover $\{U_{i,j} \mid 1 \le j \le N(A, \operatorname{diam} U_i)\}$ of A such that $\operatorname{diam} U_{i,j} = \operatorname{diam} U_i$ for all j. Using the metric d_∞ on the direct product $A \times B$, we see that

$$\operatorname{diam}(U_{i,j} \times U_i) = \max\{\operatorname{diam} U_{i,j}, \operatorname{diam} U_i\} = \operatorname{diam} U_i < \varepsilon,$$

and hence

$$\begin{aligned} m_H(A \times B, s+t, \varepsilon) &\le \sum_{i,j} \operatorname{diam}(U_{i,j} \times U_i)^{s+t} = \sum_{i,j} (\operatorname{diam} U_i)^{s+t} \\ &= \sum_i N(A, \operatorname{diam} U_i)(\operatorname{diam} U_i)^{s+t} \\ &< \sum_i (\operatorname{diam} U_i)^{-s} (\operatorname{diam} U_i)^{s+t} \\ &= \sum_i (\operatorname{diam} U_i)^t < 1. \end{aligned}$$

Taking the limit as $\varepsilon \to 0$, we see that $m_H(A \times B, s+t) < \infty$, and hence $\dim_H(A \times B) \le s+t$. Since $s > \overline{\dim}_B A$ and $t > \dim_H B$ were arbitrary, this establishes (7.13). □

Exercise 7.6. Let C_1 and C_2 be the limit sets of two Moran constructions on the line. Show that $\dim_H(C_1 \times C_2) = \dim_H C_1 + \dim_H C_2$.

b. Quantifying the attractor. Returning to the Smale–Williams solenoid Λ, we write

$$\Lambda(\theta, \varepsilon) = \Lambda \cap (D^2 \times (\theta - \varepsilon, \theta + \varepsilon))$$

for the ε-wedge of the attractor around angle θ. Since Λ can be written as a finite union of such wedges, we can compute $\dim_H \Lambda$ by computing $\dim_H \Lambda(\theta, \varepsilon)$.

Writing $C = \Lambda \cap (D^2 \times \{\theta\})$ for a cross-section of the attractor, we recall that $\Lambda(\theta, \varepsilon)$ is homeomorphic to $C \times (-\varepsilon, \varepsilon)$. In fact, the homeomorphism can be chosen to be bi-Lipschitz, and so

$$\dim_H \Lambda(\theta, \varepsilon) = \dim_H(C \times (-\varepsilon, \varepsilon)).$$

Theorem 7.9 only requires coincidence of the Hausdorff and box dimensions for *one* of the two sets A and B. Since these quantities coincide for the interval $(-\varepsilon, \varepsilon)$, we have

$$\dim_H \Lambda(\theta, \varepsilon) = (\dim_H C) + 1,$$

and since $\dim_H C$ does not depend on θ or ε,

$$\dim_H \Lambda = (\dim_H C) + 1. \tag{7.16}$$

Of course, we still need to compute $\dim_H C$. In the simplest case where $\alpha = \beta$, the construction of C is exactly of the sort dealt with by Moran's theorem, and we have

$$\dim_H C = \frac{\log 2}{-\log \alpha}.$$

Thus we get

$$\dim_H \Lambda = 1 + \frac{\log 2}{-\log \alpha} = \log 2 \left(\frac{1}{\log 2} + \frac{1}{-\log \alpha} \right); \tag{7.17}$$

this is reminiscent of (4.17), which related Hausdorff dimension and topological entropy for a one-dimensional Markov map with constant slope. Indeed, one may show that the topological entropy of the Smale–Williams solenoid is $h_{\text{top}}(\Lambda, f) = \log 2$, and that the measure of maximal entropy is the product of the $(1/2, 1/2)$-Bernoulli measure on W^s and Lebesgue measure on W^u_ε.

In (4.17), the scaling factor relating the Hausdorff dimension and the topological entropy was the reciprocal of the Lyapunov exponent. In (7.17), this factor is the somewhat odd-looking expression $(1/\log 2) - (1/\log \alpha)$. What are we to make of this?

c. Lyapunov exponents in multiple dimensions. When f is a one-dimensional map, the definition of the Lyapunov exponent is relatively simple: at a given point x, the Lyapunov exponent $\lambda_f(x)$ is the asymptotic rate of expansion along the orbit of x (provided the limit exists). For maps in more than one dimension, the situation is somewhat more complicated, as f may have different rates of expansion in different directions. Indeed, from the definition of a hyperbolic set Λ we see that along the orbit of any point $\mathbf{x} \in \Lambda$, there are some directions in which f is expanding (corresponding to a positive Lyapunov exponent) and some in which f is contracting (corresponding to a negative Lyapunov exponent).

Thus in general, the Lyapunov exponent depends not only on the point $\mathbf{x}$ but also on the direction $\mathbf{v}$ in which expansion is measured.

Definition 7.10. Let $U \subset \mathbb{R}^d$ be the domain of an invertible differentiable map $f\colon U \to U$. Given a point $\mathbf{x} \in U$ and a vector $\mathbf{v} \in \mathbb{R}^d$, the *forward Lyapunov exponent* of f at the point $\mathbf{x}$ in the direction

of $\mathbf{v}$ is

$$\lambda_f^+(\mathbf{x},\mathbf{v}) = \lim_{k\to\infty} \frac{1}{k} \log \|Df^k(\mathbf{x})\mathbf{v}\|,$$

and the *backward Lyapunov exponent* is

$$\lambda_f^-(\mathbf{x},\mathbf{v}) = \lim_{k\to\infty} \frac{1}{k} \log \|Df^{-k}(\mathbf{x})\mathbf{v}\|,$$

provided the limits exist. If $\lambda_f^+(\mathbf{x},\mathbf{v}) = -\lambda_f^-(\mathbf{x},\mathbf{v})$, then we call this value the *Lyapunov exponent* of f at the point $\mathbf{x}$ in the direction of $\mathbf{v}$, and denote it by $\lambda_f(\mathbf{x},\mathbf{v})$.

Example 7.11. Let $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ be the linear map defined by the matrix $\left(\begin{smallmatrix} \alpha & 0 \\ 0 & \beta \end{smallmatrix}\right)$, where $0 < \alpha < 1 < \beta$. Then the trajectories of f are as shown in Figure 5.3, and there are three possibilities for $\lambda(\mathbf{0},\mathbf{v})$:

(1) $\mathbf{v}$ lies along the x-axis. In this case, $\lambda_f^+(\mathbf{0},\mathbf{v}) = -\lambda_f^-(\mathbf{0},\mathbf{v}) = \log\alpha < 0$, so the Lyapunov exponent exists and is negative.

(2) $\mathbf{v}$ lies along the y-axis. In this case, $\lambda_f^+(\mathbf{0},\mathbf{v}) = -\lambda_f^-(\mathbf{0},\mathbf{v}) = \log\beta > 0$, so the Lyapunov exponent exists and is positive.

(3) $\mathbf{v}$ does not lie along either axis. In this case, $\lambda_f^+(\mathbf{0},\mathbf{v}) = \log\beta$, as the forward trajectory is repelled vertically, and $\lambda_f^-(\mathbf{0},\mathbf{v}) = -\log\alpha$, as the backward trajectory is repelled horizontally. Thus the Lyapunov exponent does not exist.

As Example 7.11 shows, the directions in which the Lyapunov exponents of a linear map exist are the eigenspaces of the map; in such a direction, the Lyapunov exponent is the logarithm of the corresponding eigenvalue. By finding a basis for $\mathbb{R}^d$ consisting of generalised eigenvectors, we can decompose $\mathbb{R}^d$ as the direct sum of subspaces along which the Lyapunov exponents exist. Such subspaces are called *Lyapunov subspaces*, and are the generalisation of eigenspaces to nonlinear maps.

Exercise 7.7. Compute the Lyapunov exponent of the linear map from $\mathbb{R}^2$ to itself given by the rotation matrix

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Exercise 7.8. Compute the Lyapunov exponents of the north-south map of the unit sphere (see Exercise 7.4), where the rate of expansion

around the north pole is $\alpha > 1$, and the rate of contraction around the south pole is $\beta < 1$.

Returning to the Smale–Williams solenoid $f\colon \Lambda \to \Lambda$, we see that f has two Lyapunov subspaces at any given point. The first subspace is tangent to the local unstable curve W^u_ε; any vector in this direction is expanded by a factor of 2 under the action of f, and so the corresponding Lyapunov exponent is $\log 2 > 0$. The second subspace is orthogonal to the first, and contains the stable disc W^s; any vector in this subspace is contracted by a factor of α under the action of f, and so the corresponding Lyapunov exponent is $\log \alpha < 0$.

Now we may interpret the ratio in (7.17) in terms of Lyapunov exponents; it is the sum of the reciprocals of the absolute values of the Lyapunov exponents in the expanding and contracting directions.

d. The non-conformal case. So far, we have studied the quantitative properties of the map f given in (7.6) only in the case $\alpha = \beta$. In this case, the map is *conformal*: every stable direction has the same rate of contraction, and every unstable direction has the same rate of expansion.

For $\alpha \neq \beta$, we are in the non-conformal case, which is much more difficult. Because the basic sets in Figure 7.5 are no longer similar to the basic sets at previous steps, we cannot use Moran's theorem. This case was studied by the German mathematician Hans Bothe, who considered a more general class of maps f, in which the functions $\cos\theta$ and $\sin\theta$ in (7.6) are replaced by arbitrary periodic functions $z_1(\theta)$ and $z_2(\theta)$, which changes the geometry of how the image $f(P)$ wraps around the z-axis. Bothe obtained a general formula for the Hausdorff dimension of the attractor for "typical" functions z_1 and z_2, but it was not until 1997 that the Hungarian mathematician Károly Simon proved that sin and cos belong to this "typical" class [**Sim97**]. He established that for the Smale–Williams solenoid Λ with $\beta < \alpha < 1/8$, we have[7]

$$\dim_H \Lambda = 1 + \frac{\log 2}{-\log \alpha} = \log 2 \left(\frac{1}{\log 2} + \frac{1}{-\log \alpha} \right).$$

[7]This was later extended to include all $\alpha < 1/2$ by Jörg Schmeling.

Somewhat surpisingly, the smaller value β, which corresponds to a direction of faster contraction, does not affect the Hausdorff dimension of the attractor! Thus only one of the two negative Lyapunov exponents plays a role in this particular situation.

e. The attractor for the FitzHugh–Nagumo map. Returning to the local map (5.7) for the FitzHugh–Nagumo model, we recall that for a particular range of values of A, the map f has a trapping region R, as shown in Figure 7.1. This ensures the existence of an attractor $\Lambda \subset R$ as in (7.5), and it is natural to ask what features Λ shares with the Smale–Williams solenoid, since both are attractors.

We saw that the Smale–Williams solenoid is hyperbolic, and hence Theorem 7.7 applies, allowing us to use all the tools of symbolic dynamics to study the solenoid and obtain many properties which are characteristic of hyperbolicity and chaotic behaviour, such as density of periodic points.

The attractor Λ for the FitzHugh–Nagumo system is more difficult to study, because it is not a hyperbolic set; there are no uniformly contracting and expanding subspaces which satisfy the definition of hyperbolicity. However, this does not preclude the possibility that there may exist two Lyapunov subspaces at "typical" points $\mathbf{x} \in \Lambda$, which are given by vectors $\mathbf{v}^s$ and $\mathbf{v}^u$ such that the map f is asymptotically contracting in the direction given by $\mathbf{v}^s$ and asymptotically expanding in the direction given by $\mathbf{v}^u$; that is, $\lambda(\mathbf{x}, \mathbf{v}^s) < 0 < \lambda(\mathbf{x}, \mathbf{v}^u)$.

Computer simulations strongly suggest that this is in fact the case, but no rigorous proofs are available.