

THE SMALE HORSESHOE AS A FRACTAL STRUCTURE IN DYNAMICAL SYSTEMS

MARK BRANSON

ABSTRACT. This paper provides an in-depth look at the two forms (attracting and repelling) of the Smale Horseshoe map. The fractal structure of the map is examined and the relationship to the homoclinic tangency is explained. The author also provides several examples where the horseshoe map appears in dynamical systems and its relationship to chaoticity in those systems.

1. INTRODUCTION

The map now known as the Smale Horseshoe map was first studied by Smale in [SML]. The simplest form of the map, from which it gains the name "horseshoe" is a map from a square region (we will use the unit square $[0, 1] \times [0, 1]$, denoted as S for convenience) in \mathbb{R}^2 to \mathbb{R}^2 given by applying the three following operations:

Step 1.: Contract the square by a factor of λ in the vertical direction, where $0 < \lambda < \frac{1}{2}$, such that $S \rightarrow [0, 1] \times [0, \lambda]$.

Step 2.: Expand the rectangle obtained by a factor of μ in the horizontal direction, where $2 + \epsilon < \mu$ (the need for this ϵ factor is explained in step 3), such that $[0, 1] \times [0, \lambda] \rightarrow [0, \mu] \times [0, \lambda]$.

Step 3.: This produces a rectangle of dimensions $\mu \times \lambda$. Take this rectangle and bend it so that it crosses the original square S in two sections. Note that this produces a bend in the rectangle, as shown in Figure 1. The ϵ in step 2 is added to account for the extra length needed to create this bend, as well as any extra on the other side of the square. We will refer to this rectangle as $f[S]$.

Step 4.: This process is then repeated, only using $f[S]$ rather than the unit square. Although the operations are not as simple to describe mathematically, the same basic ideas are used - $f[S]$ is contracted in one direction, stretched in the other, and spread over itself with a single bend. The n th iteration of this process will be called $f^n[S]$. $f^2[S]$ is shown in Figure 2.

Note. *Although this paper will work primarily with the horseshoe which crosses the original square in a linear fashion, this is not a requirement of the map. Horseshoe maps which appear in applications are rarely so regular, but the behavior can be very similar. For an example of a nonlinear horseshoe map, consider the map which takes two parallel edges to two parabolas, and then crosses the square. While this map does not cross the horseshoe in a linear fashion, its behaviour will still be that of the standard horseshoe, distorted by the nonlinearity.*

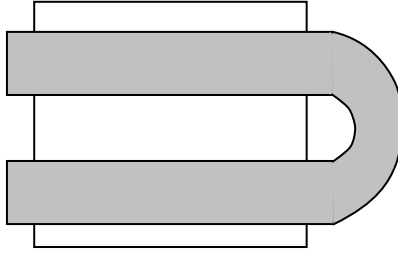


FIGURE 1. The horseshoe map after a single iteration (in light grey), $f[S]$, superimposed on S .

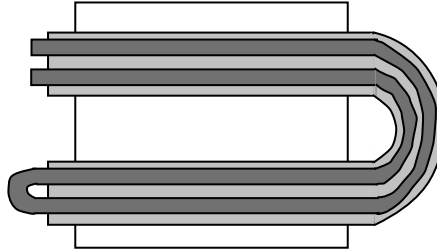


FIGURE 2. The horseshoe map after two iterations (in dark grey), $f^2[S]$ superimposed on the figure of $f[S]$.

Clearly, at the n th iteration, the rectangle will have a width of λ^n , and a length of μ^n . Thus, as one dimension goes to 0, the other dimension goes to ∞ . If we restrict our study to the limit set, the intersection of all iterations, denoted by Δ_+ :

$$(1) \quad \Delta_+ = \lim_{n \rightarrow \infty} \Delta_n = \bigcap_{i \geq 0} f^i[S]$$

the image that emerges is that of an interval crossed with a Cantor-like set. This is clear from the process presented above. In the first iteration, Δ_1 , two basic intervals of dimensions $1 \times \lambda$ are created. In the second iteration, Δ_2 , there are 4 basic intervals of dimensions $1 \times \lambda^2$. In fact, it is trivial to show that Δ_n consists of 2^n basic intervals of dimensions $1 \times \lambda^n$. This is obviously an interval cross a Cantor-like set.

This map, however, is only one part of the construction which makes up Smale's horseshoe map. We also wish to examine the preimages of our mapping, $f^{-k}[S]$. If we examine the first iteration of the horseshoe more closely, we can easily see that the preimage of the area mapped back into the square, which we will denote Δ_{-1} , is simply that of two vertical strips, as Figure 3 shows.

The reader notes that this inverse set Δ_{-1} closely resembles Δ_1 rotated by $\frac{-\pi}{2}$ (of course, this is only in the linear case - in the nonlinear case, the rotation will

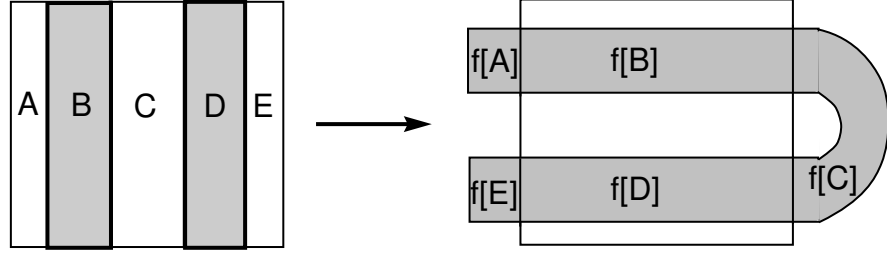


FIGURE 3. Clearly $\Delta_{-1} = B \cup D$ is mapped to Δ_1 by $f[S]$.

not be so regular). A close examination of the preimage of any Δ_k shows the same behaviour. Therefore, we see that the limit set of $f^{-1}[S]$, denoted by Δ_- ,

$$(2) \quad \Delta_- = \lim_{n \rightarrow \infty} \Delta_n = \bigcap_{i \geq 0} f^{-i}[S]$$

is exactly a Cantor-like set cross an interval. Note that the length of the initial basic interval will be $\frac{1}{\mu}$, since an expansion by μ takes it to a rectangle of length 1.

Smale's horseshoe map actually consisted of the iterated application of both $f[S]$ and $f^{-1}[S]$. Since Δ_+ consists of a interval cross a Cantor-like set, and Δ_- consists of a Cantor-like set cross an interval, the points in Δ , where

$$(3) \quad \Delta = \Delta_+ \cap S \cap \Delta_-$$

consists of a Cantor-like set cross a Cantor-like set. This map and its associated trajectories formed the basis for Smale's work.

2. FRACTAL STRUCTURE

Clearly the limit set of the map has some fractal structure. By looking at this structure, we can more clearly understand the system itself.

First of all, we find the Hausdorff dimension of the limit set Δ . Since Δ consists of a Cantor-like set cross a Cantor-like set, this is an easy task.

Theorem 2.1. *If we consider the map $f[S]$, where f is the horseshoe map that contracts by λ and expands by μ , where $0 < \lambda < \frac{1}{2}$ and $\mu > 2 + \epsilon$, the Hausdorff dimension of Δ is $\log 2 \left(\frac{1}{\log \mu} - \frac{1}{\log \lambda} \right)$.*

Proof. First, we wish to find the Hausdorff dimension of each Cantor set. We will begin with the vertical set, which we will denote E . This set has basic intervals of length λ . Since this is a Cantor-like set, the Hausdorff dimension s satisfies the equation $\lambda^s + \lambda^s = 1$. Therefore $2\lambda^s = 1$, so $\lambda^s = \frac{1}{2}$. If we take the common log of each side, we get $s \log \lambda = \log \frac{1}{2}$. Dividing each side by $\log \lambda$ obtains $s = \frac{\log \frac{1}{2}}{\log \lambda}$. This implies that $\dim_H[E] = -\frac{\log 2}{\log \lambda}$. A similar proof applied to the horizontal set, which we will denote F , with $\frac{1}{\mu}$ replacing λ , obtains a result of $\dim_H[F] = \frac{\log 2}{\log \mu}$.

To obtain a Hausdorff dimension for the limit set Δ , we apply Corollary 7.4 from [FAL]. Since both E and F are Cantor-like sets, $\dim_H[F] = \overline{\dim}_B[F]$, and thus $\dim_H[E \times F] = \dim_H[E] + \dim_H[F]$. Therefore, we obtain a result of $\dim_H[\Delta] = \dim_H[E \times F] = \log 2 \left(\frac{1}{\log \mu} - \frac{1}{\log \mu} \right)$. ■

Since the Cantor set itself is an uncountable, nowhere dense set in \mathbb{R} , our limit set Δ forms an uncountable, nowhere dense set in \mathbb{R}^2 . This can be seen by looking at the symbolic space of Δ . Since $f[S]$ consists of 2^n basic intervals at step n , with each basic interval at step n containing two basic intervals at step $n + 1$, the symbolic space of Δ_+ is clearly Σ_2^+ . Similarly, if we look at Δ_- , we will see that basic intervals increase in the same way as $n \rightarrow -\infty$. Thus, we can represent Δ_- as Σ_2^+ . Clearly, if we wish to represent Δ in the symbolic space, we can represent it as Σ_2 , where nonnegative terms in the n th place represent basic intervals in the $(n + 1)$ th step (the adjustment by one is to account for the lack of a zero term), while negative terms in the $-n$ th place represent basic intervals in the $-n$ th step. Since Σ_2^+ is uncountably infinite (the binary map takes it to the interval $[0, 1]$), and Σ_2 can be considered as a pair of such binary numbers, it can be mapped to the region $[0, 1] \times [0, 1]$. However, this set has Lebesgue measure 0, since it is a product of two sets of Lebesgue measure 0, and thus Δ is nowhere dense.

3. DYNAMICAL STRUCTURE

Although the structure of the limit set Δ is interesting alone, in the same way that the Cantor set is, Δ is more interesting when studied as the invariant set of $f[S]$. Since $f[S]$ is equivalent to the shift map in the symbolic space, periodic points are those which repeat the same word infinitely in any direction. For example, there are two fixed points, represented in Σ_2 as $\{\dots 11111 \dots\}$ and $\{\dots 22222 \dots\}$. There are exactly two points of order exactly two (of course the fixed points have order two), namely $\{\dots 1212.1212 \dots\}$ and $\{\dots 2121.2121 \dots\}$ ¹. In general, the number of periodic points of order dividing n in the invariant set Δ will be the number of words of length n . Since there are two choices for each character of n , the number of such words is precisely 2^n . Thus, Δ has an uncountably infinite number of periodic points. This is easy to see, since each point in Δ corresponds to a subset of the natural numbers (think of each word as a subset containing the natural numbers corresponding to places where the word has a 1), and Cantor's theorem tells us that the set of subsets of \mathbb{N} , $\mathcal{P}(\mathbb{N})$, is uncountable. Also, those periodic points have arbitrarily high period - namely, for any $n \in \mathbb{N}$, there exists a periodic point in Δ with period n .

Note. *Although there are an uncountable number of periodic points, there are only a countable number of periodic orbits - these are the words of finite length up to shifting.*

Although Δ has an uncountably infinite number of periodic points, this is not to say that there are no nonperiodic points in Δ . In fact, the number of nonperiodic points is also uncountably infinite. This can be seen by a proof similar to Cantor's

¹The '.', placed to divide the negative and nonnegative halves of the symbolic representation, is significant in this case, as the two points are identical up to shift by a single digit

proof of the uncountability of \mathbb{R} using infinite-length random words (which correspond to nonperiodic points) in Σ_2 rather than infinite-length random decimals in \sum_{10} . The details of the proof follow in exactly the same way, so the proof itself is left to the reader.

Note here that the periodic points are dense in Σ_2 , since any infinite length sequence can be represented as the limit of a sequence of finite length sequences. Thus, any point in Σ_2 is the limit of a sequence of periodic points. Similarly, the nonperiodic points are also dense in Σ_2 . Since both sets are dense in the symbol-space, they must be dense in Δ .

The next step in analyzing the system is to look at its stability. Obviously, any point not in the invariant set Δ will eventually be mapped out of the square, so the invariant set itself is a repeller. However, the behaviour for points close to Δ is still not simple. Although the point will eventually be repelled from the square, this will only be after N iterations, where $N \geq n$, such that the point is in a basic interval of the n th iteration. Of course, since a point not in Δ can be arbitrarily close to Δ , so the number N can be arbitrarily large.

Because the point will eventually begin to behave in a normal manner, after following an orbit of the map for N steps, the horseshoe map is said to produce *intermittent chaos*. Once the point leaves a chaotic orbit, though, it may not continue to behave normally forever. Points can be recaptured by chaotic orbits. Thus, the graph of a particular point's position vs. the number of iterations of $f[S]$ can show periods of chaotic behaviour interspersed with periods of normal behaviour.

4. RELATIONSHIP TO THE HOMOCLINIC TANGENCY

The most significant contribution of the Smale Horseshoe to dynamics, which led to Smale's investigation of the map, is its relationship to the homoclinic tangency. The homoclinic tangency, discovered by Poincaré while investigating the famous three-body problem of celestial mechanics, was one of the earliest instances of mathematical chaos [TAR].

Of course, as is often the case with Poincaré's discoveries, the homoclinic tangency is a nontrivial matter. To understand the structure of the tangency, we will first need some definitions.

Definition 1. *An invariant manifold of a map is a set of points X such that, $\forall x \in X, f[x] \in X$.*

Definition 2. *Every fixed point, with the exception of centers, will be an element of some invariant manifold. Given a fixed point Y_0 in the invariant manifold, the manifold Y is called **stable** if, $\forall y \in Y, f^n[y] \rightarrow Y_0$ as $n \rightarrow \infty$. Similarly, a manifold is called **unstable** if, $\forall y \in Y, f^{-n}[y] \rightarrow Y_0$ as $n \rightarrow \infty$.*

Definition 3. *A fixed point is called **hyperbolic** if it is the intersection of one or more stable manifolds and one or more unstable manifolds.*

Definition 4. *A **homoclinic point** is a point which lies on a stable manifold and an unstable manifold of the same fixed point. Namely, if X is a stable manifold and Y is an unstable manifold, with hyperbolic fixed point $z \in X, Y$, and $\exists w \in X, Y$ such that $w \neq z$, then w is a homoclinic point.*

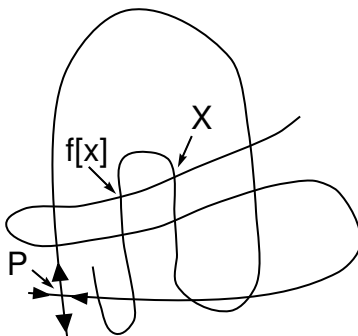


FIGURE 4. Part of a homoclinic tangency. P is the hyperbolic fixed point, and the stable and unstable manifolds are indicated by arrows. X is a homoclinic point, and thus $f[X]$, the forward iterate of X , is a homoclinic point.

The existence of the homoclinic tangency, as well as the origin of its name, comes from a simple theorem of Poincaré.

Theorem 4.1 (Poincaré). *If there exists a single homoclinic point on a stable and an unstable invariant manifold corresponding to a particular hyperbolic fixed point, then there exist an infinite number of homoclinic points on the same invariant manifolds.*

Proof. By induction. Our base case is the initial homoclinic point. Thus, assume that there exist n homoclinic points for these invariant manifolds. Let X be the stable manifold and Y be the unstable manifold, and let v be the homoclinic point farthest from the fixed point along the unstable manifold. Since X and Y are invariant manifolds, $f[v] \in X, Y$. Since $f[v] \in X, Y$, $f[v]$ is either a homoclinic point or the fixed point. Since f takes a point on the unstable manifold Y away from the fixed point, $f[v]$ cannot be the fixed point. Thus, it is a homoclinic point. For the same reason, it is not one of the n homoclinic points that we already have. Thus, there exist $n + 1$ homoclinic points. Therefore, the number of homoclinic points on the corresponding invariant manifolds is infinite. ■

The homoclinic tangency is, quite simply, the tangled intersection of such invariant manifolds with homoclinic points. In order to intersect in such a way, the stable and unstable manifolds must loop back upon one another, as shown in Figure 4.

Of course, the actual situation is much more complicated than this, as the curves must loop an infinite number of times to form an infinite collection of homoclinic points. However, the image serves to illustrate the basic situation, and to motivate the connection with the Smale Horseshoe. One quickly notes a resemblance between the illustration of the homoclinic tangency in Figure 4 and the horseshoe map. In fact, this resemblance is more than superficial. To illustrate this connection, first detailed in [SML], we look at a square containing the hyperbolic fixed point on the graph of the homoclinic tangency. This square is shown in Figure 5.

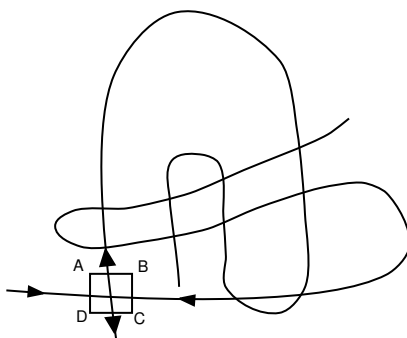


FIGURE 5. A square $ABCD$ containing the hyperbolic fixed point

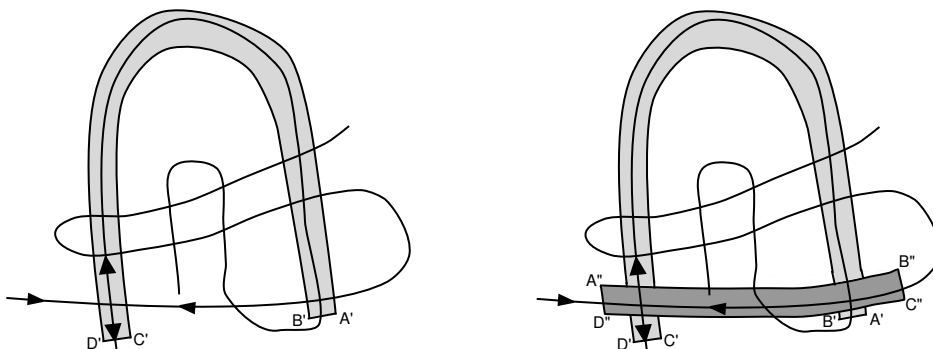


FIGURE 6. An image, $A'B'C'D'$, of the square, and its intersection with the preimage, $A''B''C''D''$ of the square.

Next, we want to look at what happens to this square as we iterate the map in each direction. Clearly, as we iterate it in the positive direction, the square will stretch out along the unstable manifold, always containing a section of the unstable manifold (since those points on the intersection of the unstable manifold and the square $ABCD$ will always be mapped to the unstable manifold). The image of the square, $f^n[ABCD]$, $n > 0$, is shown in Figure 6.

Similarly, if we iterate in the negative direction, the square will stretch out along the stable manifold, always containing some section of the stable manifold for the same reason. The preimage of the square, $f^m[ABCD]$, $m < 0$, is also shown in Figure 6.

This map contracts and expands (albeit not as evenly) in exactly the same way as the horseshoe map. In fact, the homoclinic tangency is the same topological structure as the horseshoe map [WEI]. This can be seen by looking at the horseshoe map and producing the homoclinic tangency. Recall that the horseshoe map

had two invariant sets, one in the positive direction and one in the negative direction. While we considered only the intersection with the square S , we now wish to consider these sets in their entirety, including points outside of S . In this case, the two sets will be convoluted lines intersecting the square in an infinite number of line segments. As we know, they cross in an uncountably infinite number of points. The set in the positive direction is clearly our unstable manifold - points within it are mapped to it for all iterations of $f[S]$. Similarly, the set in the negative direction is the stable manifold. Thus, the points at which they cross, Δ , are either hyperbolic fixed points or homoclinic points. Since Δ contains only two fixed points (the points that correspond to $\{\dots 11.11\dots\}$ and $\{\dots 22.22\dots\}$ in Σ_2), all of the other points in Δ must be homoclinic points of one of those two fixed points. In fact, they must be homoclinic points of both, since these fixed points share the same stable and unstable manifolds.

Consider an alternative map, in which these two fixed points do not share the same invariant manifolds. In this case, then the points of crossing will be *heteroclinic points*.

Definition 5. *A heteroclinic point is a point which lies on a stable manifold and an unstable manifold of different hyperbolic fixed points.*

However, it can be easily seen that a similar structure will form, as any one heteroclinic point will also guarantee an infinite number of such points. This structure is called the *heteroclinic tangency*. If we look at a square centered on one of the fixed points in the same way, we can construct a horseshoe-like map which is equivalent to the heteroclinic tangency. However, this is not the Smale horseshoe, and will have different behaviour.

5. ALTERNATIVE HORSESHOE MAPS

In this section, we will briefly discuss some other maps which have similar structure to the horseshoe map, and have been called horseshoe maps by various people in various publications.

First of all, we will discuss the simplest (and most similar analogue) to the normal horseshoe map, the generalized n -horseshoe. The generalized n -horseshoe has very similar properties to the normal horseshoe. Its invariant set also forms a homoclinic tangency.

Definition 6. *The generalized n -horseshoe is a horseshoe map formed with the steps seen in Section 1, with the following two exceptions:*

- 1: In Step 2, μ should be greater than $n + \epsilon$.
- 2: In Step 3, rather than bending once, the rectangle should be bent $n-1$ times, and ϵ should be large enough to accommodate these bends and still allow the horseshoe to pass through the rectangle n times.

One notes that the definition above makes no distinction about how the bends are made, thus allowing either of the 3-horseshoes shown in Figure 7. However, these maps are not significantly different - one can easily create a one-to-one correspondence between them in the symbolic space \sum_3 by exchanging the digits 2 and 3 in each infinite sequence.

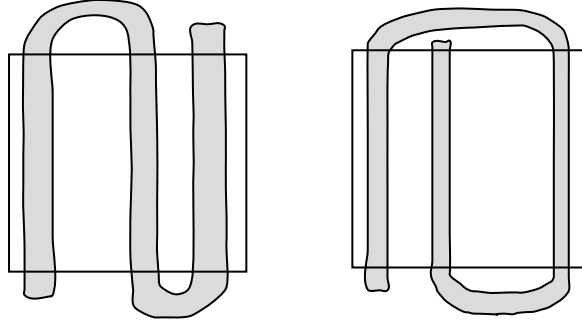


FIGURE 7. Two different versions of the generalized 3-horseshoe.

As stated above, the dynamics of the n -horseshoe are very similar to the standard horseshoe. As far as fractal structure is concerned, the Hausdorff dimension can be determined in precisely the same way.

Theorem 5.1. *If we consider the map $f[S]$, where f is a generalized n -horseshoe map that contracts by λ and expands by μ , where $0 < \lambda < \frac{1}{2}$ and $\mu > n + \epsilon$, the Hausdorff dimension of Δ is $\log n \left(\frac{1}{\log \mu} - \frac{1}{\log \mu} \right)$.*

Another significant variation on the horseshoe map which is sometimes called the Smale horseshoe [TAB], this one with very different behaviour, is the horseshoe attractor.

Definition 7. *The horseshoe attractor is a horseshoe map formed with the steps seen in Section 1, with the following two exceptions:*

- 1: In Step 2, μ should be less than $2 - \epsilon$.
- 2: In Step 3, when the bent horseshoe is mapped to the unit square, the entire horseshoe is mapped to the interior of the square with no points of S mapped to $\mathbb{R}^2 \setminus S$.

In the case of the horseshoe attractor, only positive powers of $f[S]$ are considered, since the preimages of every $f^n[S]$ are exactly S . The reader will quickly realize that rather than producing a hyperbolic invariant set, this map will produce an invariant set which consists of a one-dimensional stable manifold. Thus, this manifold will be an attractor for all the points in S . However, the attractor still contains an uncountably infinite number of nonperiodic orbits, and still produces chaotic behaviour. In fact, unlike the intermittent chaos of the horseshoe map itself, the horseshoe attractor produces persistent chaos. Any random point picked in S is guaranteed to be attracted by a nonperiodic point and fall into a persistently chaotic behaviour, since the nonperiodic orbits are dense. The horseshoe attractor is what is called a *strange attractor*.

Definition 8. *An attracting set is called a **strange attractor** if the set has measure zero and fractal structure.*

6. THE SMALE HORSESHOE IN CHAOTIC SYSTEMS

As was previously mentioned, the Smale horseshoe appears prominently in many natural systems that exhibit chaotic behaviour. One such example, the Fitzhugh-Nagumo model of the nerve, was discussed in class and thus will not be covered here. Another example is the one which initially sparked interest in the homoclinic tangency - namely, the three-body problem.

The problem is surprisingly simple when one considers how difficult it has proven to solve. It concerns the actions of three bodies upon each other in celestial mechanics. While the gravitational interaction of two bodies has been well understood for some time, the perturbation produced by other bodies (like the pull of the sun in the simple Earth-Moon system) has proven difficult to determine, even when restrictions were placed upon their motion². When Poincaré discovered the homoclinic tangency, he proved that the system (for certain initial values, at least) was chaotic and unable to be determined analytically.

Of course, Poincaré's discovery led to the advent of dynamics and chaos theory in physics, and is thus, in a way, the progenitor of the other examples we will consider. One such example is the damped forced pendulum. For sufficiently large ϵ , where ϵ is the coupling parameter, the behaviour of the pendulum is modeled by a strange attractor. However, the Smale horseshoe manifests for a particular square in the map, given proper initial parameters.

Another example where the homoclinic tangle, and thus the horseshoe map, appears are the famous Lorenz equations.

$$\begin{aligned}\dot{X} &= \sigma(Y - X) \\ \dot{Y} &= -XZ + rX - Y \\ \dot{Z} &= XY - bZ\end{aligned}$$

If one takes a fixed $b = \frac{8}{3}$ and $\sigma = 10$ and varies the third parameter r , then for values of r between (approximately) 13.926 and 24.74, a homoclinic point forms in the Lorenz map ([TAB]). Of course, with a single homoclinic point comes infinitely many such homoclinic points, and the homoclinic tangency forms. For $r > 24.74$, the map begins to resemble a three dimensional analogue of the horseshoe attractor mentioned in Section 5.

REFERENCES

- [ASY] Alligood, K.T., Sauer, T.D., Yorke, J.A., *Chaos: An Introduction to Dynamical Systems*, Springer-Verlag, New York: 1996 pp. 216-221, 409-413
- [FAL] Falconer, K. *Fractal Geometry: Mathematical Foundations and Applications*, John Wiley & Sons, Chichester: 1995
- [Nit] Nitecki, Z., *Differentiable Dynamics: An Introduction to the Orbit Structure of Diffeomorphisms*, The M.I.T. Press, Cambridge, MA: 1971
- [SML] Smale, S., "Differentiable Dynamical Systems," *Bull. of the Amer. Math. Soc.*, vol. 73 (1967), pp. 747-817
- [TAB] Tabor, M., *Chaos and Integrability in Nonlinear Dynamics*, John Wiley & Sons, New York: 1989
- [TAR] Tuffillaro, N., Abbott, T., Reilly, J., *An Experimental Approach to Nonlinear Dynamics and Chaos*, Addison-Wesley, Redwood City, CA

²Poincaré actually worked with the restricted three-body problem, where the three bodies are in a single plane and one is small enough that it does not affect the motion of the others

[WEI] Weisstein, E., "Homoclinic Tangle", "Smale Horseshoe", "Homoclinic Point", Eric Weisstein's World of Mathematics, <http://mathworld.wolfram.com>, 2002