CHAPTER 6

Higher dimensions and chaos

6.1. The horseshoe

Chaos was discovered – and largely forgotten – several times. It was first described by Poincaré in about 1890 in his theoretical analysis of the "three body problem" of classical mechanics. It was rediscovered experimentally by van der Pol in the 1920's in the "periodically forced" van der Pol equation, in his studies of vacuum tubes. It was analyzed by John Littlewood and Mary Cartwight in the 1940's, who were led to the periodically forced van der Pol equation while doing war-time radar research. They gave a complicated explanation of part of the very strange phase portrait that underlies van der Pol's experiments, and Nathan Levinson gave a simpler explanation of these phenomena in the 1950's.

But by about 1960 most of this work was unknown to almost all mathematicians, to the extent that Steve Smale published conjectures that, in effect, said that chaos did not happen. Fortunately, Levinson told him about some of this history, and Smale disproved his own conjectures by creating the horseshoe map. Smale and his coworkers developed a very clear analysis, based on simple geometry, of this example which illustrated many of the features that came to be codified under the heading of "chaotic dynamics". This time the mathematical community embraced the notion of chaos, and since then there have been thousands of papers demonstrating chaos in a wide variety of dynamical systems. In almost all cases chaos has been established by detecting a copy of Smale's horseshoe embedded in the phase portrait of the dynamical system.

The horseshoe system is easiest to explain as a discrete dynamical system in two variables. We'll discuss later how the horseshoe shows up in a three dimensional system of differential equations.

We will not give formulas for all of the horseshoe system, since only some of the details are important for the analysis. We can describe the map as follows; refer to Figure 6.1: Start with a rectangle in the plane, divided into three subrectangles labeled V_0, B and V_1, with with two semi-circular endcaps, labeled A and C. Call this figure R. Imagine transforming this figure by the following sequence of operations:

(1) Shrink the rectangle vertically.

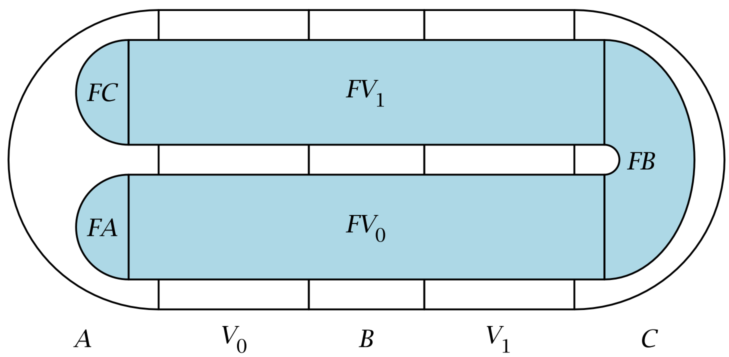


FIGURE 6.1. The horseshoe, showing the original region R and its image $F(R)$.

(2) Shrink the endcaps horizontally, until they regain their semi-circular form.
(3) Stretch the rectangles V_0 and V_1.
(4) Bend the figure within the rectangle B so it forms a U-shaped figure, the "horseshoe". The elongated rectangles V_0 and V_1, after this bending operation, are horizontal.

With some care the horseshoe will now fit back inside the original figure R; the images of the caps A and C will be inside the original A, and the image of B will be inside the cap C. The net effect is a function F which is defined on R. It can be extended as a function defined on the entire plane, but we will not need this extension. It is an injection (a one-one function), so the inverse function F^{-1} is defined, at least on the image $F(R)$ of R. It is easy to construct F to be continuous, and, with some care, it can be constructed to be C^1 or, in fact, C^k for any k. We assume that the vertical shrinking and horizontal stretching on V_0 and V_1 are given by affine maps, so

$$F(x) = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix} x + v_0 \qquad \text{if } x \in V_0, \tag{6.1}$$

$$F(x) = \begin{bmatrix} -\lambda & 0 \\ 0 & -\mu \end{bmatrix} x + v_1 \qquad \text{if } x \in V_1, \tag{6.2}$$

where v_0 and v_1 are constant vectors and $0 < \mu < 1 < \lambda$. In the illustration $\mu < \frac{1}{2}$ and $\lambda > 2$. The matrix for F on V_1 has negative entries since the map on V_1 involves rotation through 180° after the shrinking and stretching corresponding to μ and λ.

It is also easy to arrange that F will have a sink in A, so that all points in A are in the basin of attraction of this sink. Since F maps C into A, all points of C are also in the basin of attraction of the sink. Since F maps B into C, all points in B are in the basin of attraction of the sink. This basin of attraction is not interesting in the discussion of the chaotic behavior of F, and in general we will not be interested in any points which leave V_0 or V_1 under any iterate F^n. This leaves us with the following set:

$$\Omega = \{\, x\colon\, F^n(x) \in V_0 \cup V_1 \text{ for all integers } n \,\}\,. \tag{6.3}$$

This is a complicated set, and the action of F on points of Ω is also complicated. However, it can be represented in terms of an abstract encoding, and this encoding makes it easy to describe the chaotic features of the horseshoe.

In this section we will describe the encoding and see how to use it, and in the next section we'll see how to construct it.

There are three ingredients in this encoding: A set of codes, a transformation on these codes, and a way of measuring distance between codes.

First, the set Σ is the set of all *bi-infinite* sequences of 0's and 1's. In general, such a sequence can be written as $\sigma = \left\langle\, \sigma_k\colon\, -\infty < k < \infty \,\right\rangle$. That is, there is a *term* σ_k corresponding to each integer, including negative integers. For example, we could have a sequence defined by $\sigma_k = 0$ for $k \le 0$ and $\sigma_k = 1$ for $k \ge 0$. We would like to write this sequence simply as $\ldots 0001111\ldots$, but this representation cannot keep track of the subscript values. In particular, there is no way to tell whether the first 1 in $\ldots 0001111\ldots$ is the 0^{th} term of the sequence (as we intended), or the 5^{th} term, or the -17^{th} term, etc. To make it possible to use this kind of notation we introduce a decimal point "." which has no significance except as a separator between the terms of with negative subscripts and the terms with non-negative subscripts: the decimal point will always be written just before the term of index 0. With this convention our example can now be written unambiguously as $\ldots 000\,.\,1111\ldots$.

The second ingredient is a transformation of codes. This is a function sh (pronounced "shift") defined on sequences in Σ by the rule

$$\operatorname{sh}(\sigma) = \sigma' \text{ where } \sigma'_k = \sigma_{k+1} \text{ for all } k.$$

This is the formal definition, but it has a very simple interpretation. For example, the term in the 0^{th} position in σ' is $\sigma'_0 = \sigma_{0+1} = \sigma_1$. That is, the 1^{testst} term in σ becomes the 0^{th} term in σ'. Continuing, we see that the 2^{nd} term in σ becomes the 1^{testst} term in σ', the 3^{rd} term in σ becomes the 2^{nd} term in σ', and so on. This leads to the following picture:

$$\sigma = \ldots\sigma_{-3}\sigma_{-2}\sigma_{-1}\,.\,\sigma_0\sigma_1\sigma_2\sigma_3\ldots \mapsto \sigma' = \ldots\sigma_{-3}\sigma_{-2}\sigma_{-1}\sigma_0\,.\,\sigma_1\sigma_2\sigma_3\ldots$$

We can describe this in two ways. Either we can think of the decimal point as fixed, and sh shifts the entire sequence σ one place to the left, or we can think of the terms of the sequence as fixed, and sh simply oves the decimal point one place to the right. We will prefer the second interpretation; it easily generalizes as:

LEMMA 6.1. *If $k \geq 0$ then*
(a) $\mathrm{sh}^k(\sigma)$ is obtained from σ by moving the decimal point k places to the right.
(b) $\mathrm{sh}^{-k}(\sigma)$ is obtained from σ by moving the decimal point k places to the left.

The third ingredient is a notion of distance between sequences in Σ. The idea is that two sequences σ and τ should be close if they agree on a large range of subscripts centered at the origin. Formally, we define the distance between σ and τ to be 0 if $\sigma = \tau$. Otherwise, we must have $\sigma_n \neq \tau_n$ for some integer n. We let N be the smallest value of $|n|$ for which $\sigma_n \neq \tau_n$, and we define the distance between σ and *tau* to be

$$d(\sigma, \tau) = 2^{-N}.$$

In other words, N is the largest integer such that $\sigma_n = \tau_n$ for all n with $|n| < N$. For example, $\ldots 110\,.\,1001\ldots$ and $\ldots 101\,.\,0111\ldots$ differ in the 0^{th} position, so $N = 0$ and their distance is $2^0 = 1$. For another example, let $\sigma = \ldots 1010101\,.\,001001001\ldots$ and $\tau = \ldots 1010101\,.\,001000001\ldots$ The closest place to the center where these sequences differ occurs at $\sigma_5 = 1$, $\tau_5 = 0$, so $N = 5$ and the distance is $d(\sigma, \tau) = 1/32$.

Now we have the ingredients for the main theorem.

THEOREM 6.2. *There is a function h from Ω to Σ so that*
(a) h is a bijection.
(b) For all integers k and all points p in Ω, $h(F^k(p)) = \mathrm{sh}^k(h(p))$.
(c) Both h and h^{-1} are continuous.
(d) $\sigma = h(p)$ if and only if, all integers k, $F^k(p)$ lies in V_{σ_k}.

Here are some explanations of the different properties of h:

(a) A bijection is the same as a one-to-one correspondence. Thus each p in Ω corresponds to a unique sequence $\sigma = h(p)$ in Σ, and each sequence σ in Σ corresponds to a unique point $p = h^{-1}(\sigma)$ in Ω.
(b) $h(F^k(p)) = \mathrm{sh}^k(h(p))$ means that the sequence corresponding to $F^k(p)$ is the sequence $\mathrm{sh}^k(\sigma)$, so the action of F^k on Ω is "the same as" the action of sh^k on Σ. In this context h is called a *conjugacy* from F^k to sh^k.
(c) The continuity condition means (in a precise ε-δ formulation) that two points in Ω are close if and only if their corresponding sequences are close in Σ.
(d) If p is in Ω then, by the definition of Ω, each iterate $F^k(p)$ lies in $V_0 \cup V_1$, so we can define a sequence σ by defining $\sigma_k = 0$ if $F^k(p)$ is in V_0, and $\sigma_k = 1$ if $F^k(p)$ is in V_1. Property (d) says that this rule is the definition of $h(p)$.

This theorem will be proved in the next section. The following is a list of consequences:

THEOREM 6.3. *(a) F has sensitive dependence to initial conditions on Ω.*
(b) Ω contains infinitely many periodic points.
(c) The periodic points are dense in Ω.
(d) The homoclinic points of any periodic point are dense in Ω.
(e) Ω contains a dense orbit.
(f) F is topologically transitive on Ω.
(g) F has positive entropy on Ω.
(h) Ω is a fractal.
(i) Ω is self-similar.

There are many studies of chaos and most of them use some or all of these properties of Ω as their definition of a chaotic system. We will explain some of these properties below.

Sensitive dependence on initial conditions: If we are in a system in which a point p is attracted to a sink in forward time and to a source in backwards time then nearby points stay close to p for all iterations. Technically, for any positive ε there is a positive δ so if q is within distance δ of p then $F^n(q)$ will be within distance ε of $F^n(p)$ for all integers n. "Sensitive dependence on initial conditions" says that this does not happen for any point in Ω. Technically, the condition is this: There is a positive constant d_0 so that, if p and q are any two different points in Ω, then for some n the points $F^n(p)$ and $F^n(q)$ are at least distance d_0 apart. This is one of the most striking properties of a chaotic system: Although the system is completely deterministic, any difference in initial conditions will eventually be magnified to be greater than the fixed quantity d_0. This means that, in a real system exhibiting sensitive dependence, precise long-term predictions are essentially impossible, since initial conditions can only be specified with finite precision.

To prove that F shows sensitive dependence on initial conditions in Ω we translate the question to Σ, where it is easy to answer: If σ and τ are different sequences then they differ in at least one term, so we can determine m so that $\sigma_m \neq \tau_m$. We set $n = -m$, and notice that $\mathrm{sh}^n(\sigma)$ and $\mathrm{sh}^n(\tau)$ differ in the 0^{th} term, so their distance is 1. Hence the function sh exhibits sensitive dependence on initial conditions with the constant $d_0 = 1$.

Let's translate this back to F on Ω. If p and q are distinct points of Ω then they correspond to different sequences σ and τ as above. Then there is an integer n so that $\mathrm{sh}^n(\sigma)$ and $\mathrm{sh}^n(tau)$ differ in the 0^{th} term. These shifted sequences correspond to $F^n(p)$ and $F^n(q)$, and the 0^{th} term in the shifted sequences specify the initial rectangles V_0 or V_1 which contain $F^n(p)$ and $F^n(q)$. Thus one of these points is in

V_0 and one is in V_1, and the definition of sensitive dependence is satisfied for F on Ω with d_0 being the distance between V_0 and V_1. Note that we can interpret sensitive dependence as follows: Unless we know p to "infinite precision" we cannot know which rectangle, V_0 or V_1, will contain each iterate $F^n(p)$.

F has infinitely many periodic points: A point p is periodic with period k if $k > 0$ and $F^k(p) = p$. Then the corresponding sequence σ is periodic under the shift, so $\mathrm{sh}^k(\sigma) = \sigma$. Remembering that the n^{th} term in $\mathrm{sh}^k(\sigma)$ is σ_{n+k}, we see that a sequence σ is periodic, with period k, if and only if

$$\sigma_{n+k} = \sigma_n$$

for all n. That is, the terms in the sequence σ are periodic with period k. If we know the terms $\sigma_0\sigma_1 \ldots \sigma_{k-1}$ then all the terms of σ are determined by this periodicity: To determine σ_N then subtract a multiple of k from N (or add one) so that the result, r, is between 0 and $k-1$. Then $\sigma_N = \sigma_r$. In other words, σ is obtained by repeating the block $\sigma_0\sigma_1 \ldots \sigma_{k-1}$ forever, in positive and negative time. For example, if $k = 4$ and $\sigma_0\sigma_1\sigma_2\sigma_3 = 1000$ then

$$\sigma = \ldots\ 1000\,1000\,.\,1000\,1000\,1000\ \ldots$$

is a periodic sequence of period 4. Clearly, we can specify infinitely many periodic sequences using this idea: For a sequence of period k we can choose the block $100\ldots0$, where there are $k-1$ zeros.

We can be very precise about the periodic points in Ω by looking at the periodic sequences in Σ. We will use the notation π_β for the periodic sequence defined by a finite block β of 0's and 1's in this manner. That is, if β has length k then π_β has β in positions 0 through $k-1$, and then this block is repeated forever. Here are the first few cases:

Period 1: We have $\pi_0 = \ldots 000\,.\,000\ldots$ and $\pi_1 = \ldots 111\,.\,111\ldots$. These are the two fixed sequences.

Period 2: There are 4 blocks of length 4, so there are 4 sequences of period 4. However, π_{00} is the sequence $\ldots 0000\,.\,0000\ldots$ obtained by repeating 00, and this is the same as the sequence π_0. This is reasonable, since if σ is a fixed sequence, meaning $\mathrm{sh}(\sigma) = \sigma$, then $\mathrm{sh}^k(\sigma) = \mathrm{sh}^{k-1}(\mathrm{sh}(\sigma)) = \mathrm{sh}^{k-1}(\sigma)$, and a simple induction shows that $\mathrm{sh}^k(\sigma) = \sigma$ for all k. Thus π_{00} and π_{11}, although they have period 2, have *least* period 1. There are two sequences of least period 2, namely

$$\pi_{01} = \ldots 10101\,.\,01010\ldots, \quad \pi_{10} = \ldots 01010\,.\,10101\ldots$$

Notice that $\mathrm{sh}(\pi_{01}) = \pi_{10}$ and $\mathrm{sh}(\pi_{10}) = \pi_{01}$. This illustrates another general fact: If σ is a periodic sequence of period k then $\mathrm{sh}^n(\sigma)$ is also periodic of period k. Moreover, the *orbit* of σ consists of the sequences σ, $\mathrm{sh}(\sigma)$, $\mathrm{sh}^2(\sigma), \ldots, \mathrm{sh}^{k-1}(\sigma)$. This orbit has k elements if k is the least period of σ; otherwise this list of iterates has repetitions.

So we can summarize the period 2 situation: There are 4 sequences of period 2; there are 2 sequences of least period 2; there are two orbits of size 1 and one orbit of size 2.

Period 3: There are 8 blocks of size 3. Thus we have the fixed sequences $\pi_{000} = \pi_0$ and $\pi_{111} = \pi_1$, corresponding to two orbits of size 1. The sequences π_{001}, π_{010}, π_{100} all have least period 3 and form a periodic orbit of size 3, and the sequences π_{011}, π_{110}, π_{101} form another such orbit.

Period 4: There are $16 = 2^4$ blocks of size 4. As above, π_{0000} and π_{1111} are the fixed sequences. We also see that π_{0101} and π_{1010} are equal to the sequences π_{01} and π_{10} of period 2, and this illustrates another general fact: If σ is a periodic sequence of period k then σ is also a periodic sequence of period m where m is any positive multiple of k. There are 12 blocks of size 4 that we have not yet considered, and they define 12 periodic sequences of least period 4, grouped as 3 orbits of size 4.

Periodic points are dense: This means that, if p is any point in Ω and ε is any positive number then there is a periodic point within ε of p. We prove this by translating to Σ and then using Exercise 6.3.

The homoclinic points of any periodic orbit are dense: If p is a periodic point then a point q is called a *homoclinic point* of the orbit of p if q is not on this orbit, and $F^n(p)$ converges to the orbit of p as $n \to +\infty$ and also as $n \to -\infty$. It is easy to find homoclinic points after we translate to Σ. The periodic point p corresponds to a sequence σ consisting of an infinitely repeated block, β. We take a sequence τ which repeats β periodically in indices that are far away from 0; it is irrelevant what the terms in τ look like near 0. Schematically,

$$\tau = \ldots \beta\beta\tau_{-M} \ldots \tau_{-1} . \tau_1\tau_1 \ldots \tau_N\beta\beta \ldots$$

If $|K|$ is much larger than N or M then $\mathrm{sh}^K(\tau)$ will have its decimal point in one of the regions consisting of repetitions of the block β, and it will agree, for a large range of indices centered at the origin, with a sequence of all β's in a point in the orbit of σ. Thus the distance between $\mathrm{sh}^K(\tau)$ and points in the orbit of σ will tend to 0 as $|K| \to \infty$, so τ is homoclinic to the orbit of σ.

Saying that the homoclinic points of the orbit of p are dense means that, arbitrarily close to any point r in Ω, we can find a point which is homoclinic to the orbit of p. To prove this we translate the problem to Σ. The point r corresponds to a sequence ρ. Using the terminology above, we construct a homoclinic point τ within distance 2^{-N-1} of ρ by defining $\tau_k = \rho_k$ for $-N \le k \le N$, and then filling in the rest of τ with copies of β.

There is a dense orbit: This means that there is a single point p in Ω so that, for any other point q in Ω and any positive ε there is some n so that $F^n(p)$ is within ε of q. We translate this to Σ, where we need to define a sequence σ with the same property. We define σ by stringing together all finite blocks of 0's and 1's. Here's how: In locations 1 and 2 we put 0 and 1, the two blocks of size 1. In the next 8 locations we put 00, 01, 10, and 11, the 4 blocks of size 2. This is followed by the 8 blocks of length 3, the 64 blocks of length 4, and so on. This process defines σ_k for $k \geq 1$; the first 50 terms are

$$0\,1\,00\,01\,10\,11\,000\,001\,010\,011\,100\,101\,110\,111\,0000\,0001\,0010\,0011$$

We define the terms with negative subscripts by reflection, so $\sigma_{-k} = \sigma_k$. How we define σ_0 is not important; for definiteness we set it to 0.

To see that the orbit of this sequence is dense, consider a sequence τ in Σ, and choose $N \geq 0$. Let βbe the block of terms in τ of length $2N + 1$ centered at τ_0, so $\beta = \tau_{-N} \dots \tau_N$. This block occurs in σ somewhere, since *every* finite block occurs in σ. So there is some M so that $\sigma_{M-N} \dots \sigma_{M+N} = \beta$. So, if we write $\rho = \mathrm{sh}^{-M(\sigma)}$, then the block of terms in ρ of length $2N + 1$ centered at ρ_0 is equal to β, which is the corresponding block of τ. Hence the distance between $\rho = \mathrm{sh}^{-M}(\sigma)$ and τ is no more than 2^{N+1}. This shows that we can find points on the orbit of σ as close as desired to any given point τ in Σ, so the orbit of σ is dense.

Actually, this argument shows slightly more: Both the positive orbit and the negative orbit or σ are dense, since we can choose M to be either positive or negative.

The remaining parts of Theorem 6.3 require more background, so we just give short descriptions.

F is topologically transitive: This is closely related to having a dense orbit; it says that any open set has a dense orbit.

F has positive entropy: There are several notions of entropy in dynamical systems. A system with positive entropy shows essentially random behavior on long time scales.

Ω is a fractal: For most sets it is possible to define a number called the *Hausdorff dimension*. For simple sets this is the usual notion. For example, smooth curves have Hausdorff dimension 1 and smooth surfaces have Hausdorff dimension 2. Sets with non-integral Hausdorff dimension are much more complicated; they are called *fractals*. The Hausdorff dimension of Ω is bigger than 0 but less than 1.

Ω is self similar: A set is *self-similar* if small parts of the set are geometrically similar to each other. In effect, the fine structure of the set is the same at any scale. Ω has this property; for example, there is a linear map which transforms the part of Ω which is inside $V_0 \cap V_1$ onto all of Ω. Complicated sets – in particular, fractals – that are defined by iteration often show self-similarity.

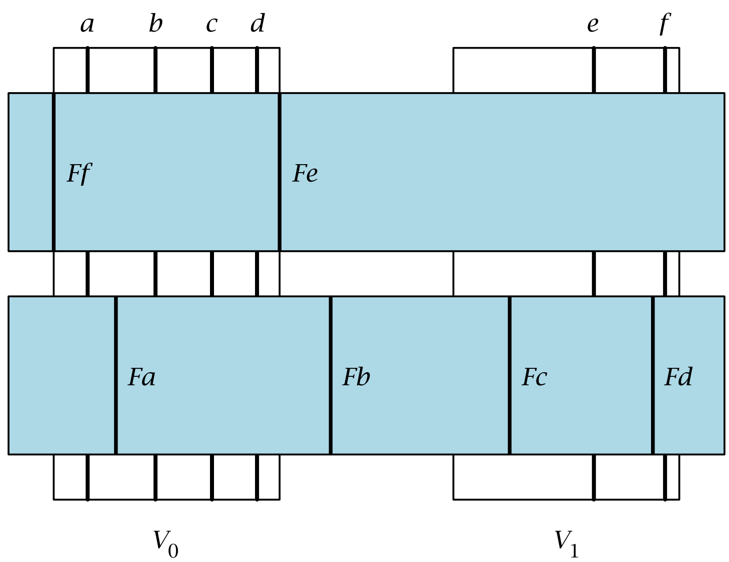


FIGURE 6.2. Six vertical segments $a, b, \ldots, f$ and their images.

6.2. Encoding the horseshoe

In this section we will give some of the details in the proof of Theorem 6.2. We will continue to use the notation of Figure 6.1 and of equations (6.1) and (6.2) to describe the horseshoe map, and (6.3) to define Ω. The correspondence h between Ω and Σ is defined the condition in part (d) of Theorem 6.2. We need to see that h is a bijection, that h and h^{-1} are continuous, and that the conjugacy condition in part (b) of Theorem 6.2 is satisfied.

We start by describing the images of vertical segments in $V_0 \cup V_1$ under F. Figure 6.2 shows several vertical lines and their images. Consider the vertical segments a, b, c, d in V_0. The image of each such segment is a vertical segment in the rectangle $F(V_0)$. This is just a consequence of the form (6.1) for F on V_0, since both translations and diagonal matrices transform vertical lines to vertical lines. Moreover, as we move vertical segments in V_0 from left to right their images move continuously from left to right in $F(V_0)$. So the left side of V_0 is mapped into A, the segment a is mapped into V_0, b is mapped into B, c and d are mapped into V_1, and the right side of V_0 is mapped into C. Similarly, the vertical segments in V_1 are mapped to vertical segments in $F(V_1)$. As the segments in V_1 move from left to right their images move from right to left, because of the negative entry $-\lambda$ in the form (6.2).

Notice that the segments e and f map into the right and left sides of V_0 respectively, so the subrectangle of V_1 bounded by e and f maps into V_0. Similarly, by

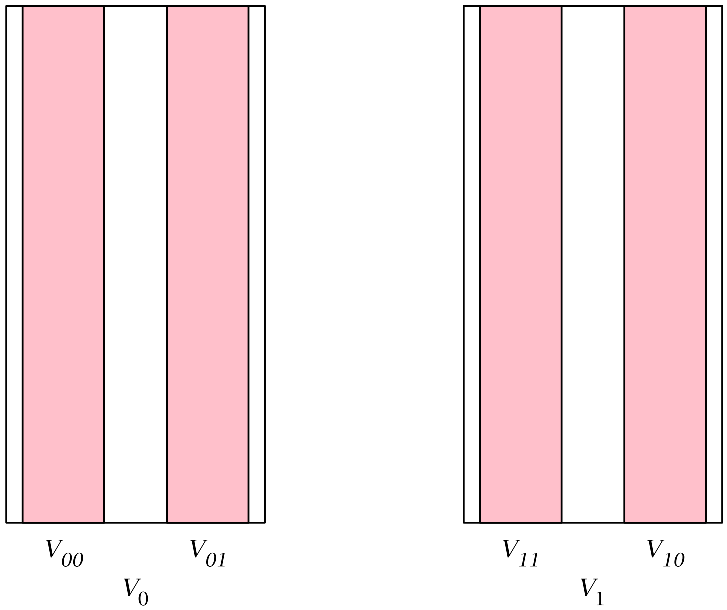


FIGURE 6.3. The points which stay in $V_0 \cup V_1$.

positioning vertical segments in V_1 so that they map into the boundaries of V_1 we will find a subrectangle of V_1 wich maps into V_1. We can do the same in V_0. We end with two subrectangles in each of V_0 and V_1 which map into either V_0 or V_1, as shown in Figure 6.3. Moreover, the points in $V_0 \cup V_1$ that are not in one of these subrectangles map into $A \cup B \cup C$. In other words, the points in the four subrectangles are exactly the points which stay in $V_0 \cup V_1$.

We can repeat this process. Consider the subrectangle labeled V_{00}. If p is a point in V_{00} then $F(p)$ is in V_0. We next ask whether $F^2(p) = F(F(p))$ lies in $V_0 \cup V_1$. Arguing as above, we see that each vertical segment in V_{00} maps under F^2 to a short vertical segment, and these image segments range from A through $V_0 \cup B \cup V_1$ and into C. Hence the points of V_{00} which map into V_0 under F^2 form a vertical rectangle in V_{00}, as do the points in V_{00} which map into V_1 under F^2. The same picture holds in the other three subrectangles V_{01}, V_{10}, V_{11}, and we find that the points which stay in $V_0 \cup V_1$ under both F and F^2 form eight thin vertical rectangles, as shown in Figure 6.4.

Here is the explanation of the numbering system for these rectangles: Look at V_{10}. This is a subrectangle of V_1, and consists of the points in V_1 which are mapped into V_0. That is, a point p is in V_{10} if $F^0(p) = p \in V_1$ and $F(p) \in V_0$. The rectangle V_{100} is a subrectangle of V_{10} and consists of the points in V_{10} which map into V_0

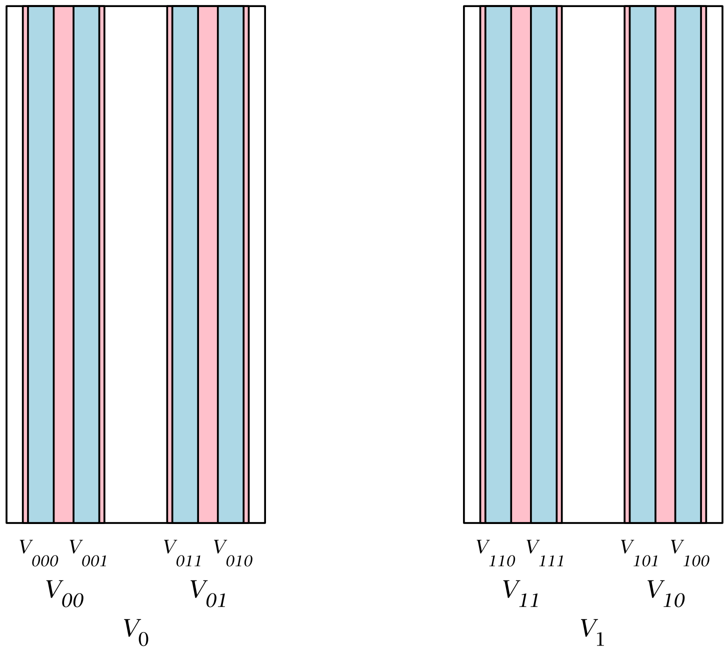


FIGURE 6.4. The points which stay in $V_0 \cup V_1$ for two iterations.

under F^2. That is, p is in V_{101} if $F^0(p) = p \in V_1$, $F^1(p) \in V_0$, and $F^2(p) \in V_0$. So the subscript 100 of V_{100} tells where the image of V_{100} is under the iterates F^0, F^1, and F^2.

To generalize this picture we use *finite* sequences of 0's and 1's. The *length* of a sequence σ is the number of terms σ, and it is written $|\sigma|$. For example, if $\sigma = 010011$ then $|\sigma| = 6$. (There is a sequence of length zero, but we will not need to talk about it in this section.)

There is a way of combining two of these sequences, called *concatenation*, which is defined by just pasting the two sequences together. For example, if $\sigma = 010011$ and $\tau = 00100$ then the concatenation of σ and τ, written $\sigma\tau$, is the sequence $010011\,00100$. Concatenation is not a commutative operation in general; in our example the concatenation of τ and σ in the opposite order is $\tau\sigma = 00100\,010011$.

We will use the following related terminology: If a sequence ρ can be factored as $\rho = \sigma\tau$ then we say that σ is a *prefix* of ρ and τ is a *suffix* of ρ. For example, if $\rho = 001011$ then 00 is a prefix of ρ, and 011 is a suffix.

We can refer to the individual terms in a sequence if we know where the numbering starts or ends. For describing vertical rectangles we will use sequences that are numbered starting at 0, so the terms of $\sigma = 010011$ will be labeled as $\sigma = \sigma_0\sigma_1\sigma_2 \ldots \sigma_5$.

Using this notation we define the generalized vertical rectangles by the condition

$$p \in V_\sigma \iff F^k(p) \in V_{\sigma_k} \text{ for } 0 \le k \le n, \text{ where } |\sigma| = n+1. \tag{6.4}$$

Note the similarity to the condition in Theorem 6.2(d).

We summarize the properties of these sets:

LEMMA 6.4. *Suppose the rectangles V_0 and V_1 have width w and height h. Then*

(a) V_σ is a vertical rectangle, of height h and width $\lambda^{n1}w$, where $|\sigma| = n+1$.
(b) V_σ contains V_τ if σ is a prefix of τ.
(c) V_σ and V_τ are disjoint unless σ is a prefix of τ or τ is a prefix of σ.
(d) The image of $V_{\sigma_0\sigma_1\cdots\sigma_k}$ under F lies in $V_{\sigma_1\sigma_2\cdots\sigma_k}$.
(e) A point in $V_0 \cup V_1$ remains in $V_0 \cup V_1$ for n iterations if and only if it lies in some V_σ where $|\sigma| = n+1$.

The proof just codifies the geometry discussed above. The formula for the width of V_σ comes from the observation that each time F is applied to a rectangle in $V_0 \cup V_1$ the width of the rectangle is multiplied by λ; and, after n iterations, the rectangle stretches across either V_0 or V_1.

Now we need to consider the points that start in $V_0 \cup V_1$ and remain there for *all* positive iterates of F. To describe these we use infinite sequences $\sigma = \sigma_0\sigma_1\sigma_2\ldots$. The prefix relation still makes sense when applied to a finite sequence and an infinite sequence.

If σ is such a sequence we can define V_σ by an obvious modification of (6.4): we require $F^k(p) \in V_{\sigma_k}$ for *all* $k \ge 0$. Geometrically, these sets are rectangles of zero width – that is:

LEMMA 6.5. *If σ is an infinite sequence of 0's and 1's then V_σ is a vertical segment, and the analogs of the statements in Lemma 6.4 are still true.*

PROOF. The only real issue is whether V_σ is non-empty. In fact, it is defined by the following limiting process. First, if τ is any finite prefix of σ, then V_σ is a subset of V_τ. We write L_n and R_n for the left and right sides of V_τ, where τ is the prefix of σ of length n. These are vertical line segments and, as n increases, L_n moves to the right and R_n moves to the left. We let l_n be the x coordinate of L_n and we let r_n be the x coordinate of R_n. Then l_n is a non-decreasing bounded sequence of real numbers, so it has a limit l_*. Similarly, r_n has a limit r_*. Since the distance between l_n and r_n is the width of V_τ and this width converges to 0 as $n \to +\infty$, we have $l_* = r_*$. Then V_σ is the limit of the vertical segments L_n and R_n; that is, it is the vertical segment with x coordinate l_* (or r_*). □

This construction gives a lot of information about points that stay in $V_0 \cup V_1$ for all positive iterations. We can specify an arbitrary future for a point, in the form

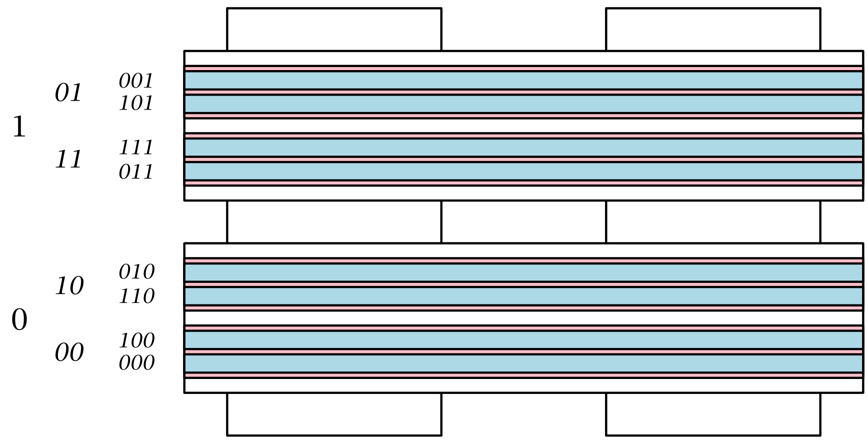


FIGURE 6.5. H_σ for length of $\sigma \le 3$. The rectangles are identified by their subscripts.

of an infinite sequence σ, and we are guaranteed that there is a vertical segment consisting of points that have that future. For example, there is a vertical segment consisting of points that stay in V_0 for all positive iterations: we just take $\sigma_k = 0$ for all k. Similarly, there is a vertical segment of points in V_0 which has a future which alternates between V_1 and V_0: we take $\sigma = 010101\ldots$.

To complete this picture we also need to analyze points that stay in $V_0 \cup V_1$ for all negative iterations. We define sets H_σ using the analog of (6.4) for negative time:

$$p \in H_\sigma \iff F^k(p) \in V_{\sigma_k} \text{ for } -n \le k \le -1 \text{ where } n = |\sigma| \,. \tag{6.5}$$

Notice that we are using *negative* indexes for the subscripts of the sets H_σ. For example, when we write H_σ with $\sigma = 010011$ then we understand that the entries in σ are indexed as $\sigma_{-6}\sigma_{-5}\ldots\sigma_{-1}$. Thus a point p in H_{010011} satisfies $F^{-6}(p) \in V_{\sigma_{-6}} = V_0$, $F^{-5}(p) \in V_{\sigma_{-5}} = V_1, \ldots, F^{-1}(p) \in V_{\sigma_{-1}} = V_1$.

Consider H_0. According to the definition, this is the set of points p for which $F^{-1}(p) \in V_0$. But the condition $F^{-1}(p) \in V_0$ is equivalent to the condition $p \in F(V_0)$. That is, H_0 is just the image of V_0 under F. This easily generalizes:

LEMMA 6.6. *Suppose σ is a sequence of length k. Then H_σ is the image of V_σ under F^k.*

For example, not only do we have $H_0 = F(V_0)$ but $H_{01} = F^2(V_{01})$, $H_{011} = F^3(V_{011})$, and so on. Suppose that σ has length n. Then V_σ has width $\lambda^{-n+1}w$ and each iteration of F multiplies horizontal lengths by λ, so $F^n(V_\sigma) = H_\sigma$ has length $\lambda^n \cdot \lambda^{-n+1}w = \lambda w$. So all the horizontal rectangles H_σ have the same width as H_0 and H_1. Since H_σ lies in H_0 or H_1 we have the layout shown in Figure 6.5.

The basic properties for these rectangles follow immediately from the corresponding properties for the vertical rectangles, Lemma 6.4. Thus we have

Lemma 6.7. *Suppose the rectangles V_0 and V_1 have width w and height h. Then*
(a) H_σ is a horizontal rectangle, of height $\mu^n h$ and width λw, where $n = |\sigma|$.
(b) H_σ contains H_τ if σ is a suffix of τ.
(c) H_σ and H_τ are disjoint unless σ is a suffix of τ or τ is a suffix of σ.
(d) The image of $H_{\sigma_{-k}\sigma_{-k+1}\cdots\sigma_{-1}}$ under F^{-1} lies in $H_{\sigma_{-k}\sigma_{-k+1}\cdots\sigma_{-2}}$.
(e) A point is in $V_0 \cup V_1$ for n iterations of F^{-1} if and only if it lies in some H_σ where $n = |\sigma|$.

Next, as we did for the sets V_σ, we extend the definition of H_σ to infinite sequences σ. Now we have $\sigma = \ldots\sigma_{-3}\sigma_{-2}\sigma_{-1}$, and H_σ is defined by the condition that $F^{-k}(p) \in V_{\sigma_k}$ for all $k \leq -1$. The analog of Lemma 6.5 is true, with a similar proof:

Lemma 6.8. *If σ is an infinite sequence of 0's and 1's then H_σ is a horizontal segment, and the analogs of the statements in Lemma 6.7 are still true.*

We are now ready to put all this together. A point p is in Ω if and only if $F^k(p)$ is in $V_0 \cup V_1$ for all integers k. This is equivalent to saying that $F^k(p)$ is in $V_0 \cup V_1$ for all $k \geq 0$, and also that $F^k(p)$ is in $V_0 \cup V_1$ for all $k < 0$. By the infinite analogs of Lemma 6.4(e) and Lemma 6.7(e), this is in turn equivalent to the statement that p lies in V_τ and also in H_ρ, for some infinite sequences τ and ρ. That is, p is in the intersection $H_\rho \cap V_\tau$. Moreover, any choice of infinite sequences ρ and τ determines a unique point p in Ω, since H_ρ and V_τ are horizontal and vertical line segments, so they intersect in exactly one point.

We can concatenate the two infinite sequences $\rho = \ldots\rho_{-3}\rho_{-2}\rho_{-1}$ and $\tau = \tau_0\tau_1\tau_2\ldots$ to form a single *bi-infinite* sequence $\sigma = \rho\,.\,\tau$. (We are using the decimal point as in Section 6.1 to mark the 0^{th} term.)

Now we can show that the function h defined in Theorem 6.2(d) is a bijection. In fact, we have just described the inverse function: To each σ in Σ there is a corresponding point $p = g(\sigma)$ in Ω, defined by

(6.6) $$g(\sigma) = p \iff \sigma = \rho\,.\,\tau \text{ and } p \text{ is the intersection of } H_\rho \text{ and } V_\tau.$$

It is a simple matter to show that g is indeed a two-sided inverse of h, so h is a bijection and $h^{-1} = g$.

We think of the sequence $\sigma = h(p)$ as an *encoding* of the point p, and we think of the point $p = h^{-1}(\sigma)$ as the *decoding* of the sequence σ.

Now here is the encoding of the action of F, establishing the conjugacy condition in Theorem 6.2(b):

LEMMA 6.9. *If $p \in \Omega$ corresponds to $\sigma = h(p) \in \Sigma$ then, for any integer k, $F^k(p)$ corresponds to* $\mathrm{sh}^k(\sigma)$.

PROOF. Let $F^k(p) = q$, and let $\tau = \mathrm{sh}^k(\sigma)$. We need to show that the encoding of q is *tau*. From the definition of the encoding we have $F^n(p) \in V_{\sigma_n}$ for all n. We can replace the arbitrary integer n with $n+k$ to get $F^{n+k}(p) \in V_{\sigma_{n+k}}$ for all n. But $F^{n+k}(p) = F^n(F^k(p)) = F^n(q)$, and $\sigma_{n+k} = \tau_n$. So we have $F^n(q) \in V_{\tau_n}$ for all n. But then, by definition, the encoding of q is τ. □

Finally, we demonstrate continuity for h and h^{-1}.

Suppose that p is in Ω and ε is a positive number. Choose N so that $2^{-N} < \varepsilon$. Let d_0 be the distance between V_0 and V_1. By continuity at p of the finitely many functions F^n for $|n| < N$, there is a positive number δso that, if q is within δ of p then the distance from $F^n(q)$ and $F^n(p)$ is less than d_0, for all n with $|n| < N$. It follows that, for each such n, $F^n(p)$ and $F^n(q)$ are in the same rectangle, V_0 or V_1. Hence the sequences $h(p)$ and $h(q)$ agree in all terms indexed by n with $|n| < N$, so the distance from $h(p)$ to $h(q)$ in Σ is no more than $2^{-N} < \varepsilon$. This establishes continuity of h.

Now suppose that σ is in Σ and ε is a positive number. Choose N large enough that

$$\sqrt{(\mu^N h)^2 + (\lambda^{-N} w)^2} < \varepsilon \tag{6.7}$$

and let $\delta = 2^{-N-1}$. Let $\rho = \sigma_{-N}\sigma_{-N+1}\dots\sigma_{-1}$ and $\tau = \sigma_0\sigma_1\dots\sigma_N$. If σ' is a sequence within δ of σ then σ' and σ agree in all terms indexed by n with $|n| \leq N$. Hence both $h^{-1}(\sigma)$ and $h^{-1}(\sigma')$ lie in $H_\rho \cap V_\tau$. This is a rectangle, of height $\mu^N h$ and width $\lambda^{-N} w$. The diameter of this rectangle, according to (6.7), is less than ε, so $h^{-1}(\sigma')$ is within distance ε of $h^{-1}(\sigma)$. This establishes continuity of h^{-1} at σ.

6.3. Time dependent systems

So far we have concentrated on systems of differential equations of the form $\dfrac{dx}{dt} = f(x)$ where, as indicated, the vector field $f(x)$ depends *only* on the position vector x. Such a system is called an *autonomous* or *time indepentent* system of differential equations. A more general type of system has the form $\dfrac{dx}{dt} = f(x,t)$ where now the vector field $f(x,t)$ depends explicitly on the time variable t. This more general type of system is called a *non-autonomous* or *time dependent* system. All of our analysis so far has been concerned with autonomous systems. Some of the basic facts, like a suitable existence and uniqueness theorem, are still valid. On the other hand, our basic ideas of the flow, solution curves, and orbits all need to be modified.

Here is a very simple example, which can be easily solved by integration:

$$\frac{dx}{dt} = t, \ \frac{dy}{dt} = x \qquad x(t_0) = x_0, \ y(t_0) = y_0. \tag{6.8}$$

To solve this, first separate variables in the first equation, to get $dx = t\,dt$, and then integrate: $x(t) = \int t\,dt = \frac{1}{2}t^2 + C$. The constant is determined from the initial conditions, so, plug in $t = t_0$ and $x(t_0) = x_0$ and solve for C: $x_0 = \frac{1}{2}t^2 + C$ so $C = x_0 - \frac{1}{2}t_0^2$. Now plug this value of C into the formula for $x(t)$:

$$x(t) = \tfrac{1}{2}t^2 + C = \tfrac{1}{2}t^2 + x_0 - \tfrac{1}{2}t_0^2 = x_0 + \tfrac{1}{2}\left(t^2 - t_0^2\right).$$

Now do the same thing with the second differential equation to find $y(t)$. First write the equation in differential form and integrate: $dy = x\,dx$ so $y(t) = \int x(t)\,dt$. In more complicated systems of equations we would be stuck right here, since we need to know $x(t)$ as a function of t in order to perform the integration. But this example is simple enough that we have already solved for $x(t)$, so just plug in our formula:

$$y(t) = \int x(t)\,dx = \int \left(x_0 + \tfrac{1}{2}t^2 - \tfrac{1}{2}t_0^2\right)\,dt = x_0 t + \tfrac{1}{6}t^3 - \tfrac{1}{2}t_0^2 t + C.$$

Now plug in $t = t_0$ and $y(t_0) = y_0$ and solve for C, and then rewrite the equation for $y(t)$ with this value of C. After some algebra we find:

$$y(t) = y_0 + x_0(t - t_0) + \tfrac{1}{6}t^3 - \tfrac{1}{2}t_0^2 t + \tfrac{1}{3}t_0^3.$$

Now that we have determined $x(t)$ and $y(t)$ we can define $G(x_0, y_0, t)$ as the solution with initial conditions $x(0) = x_0$, $y(0) = y_0$, just as we did for autonomous systems. In vector terms, if we write $z = \begin{bmatrix} x \\ y \end{bmatrix}$ and $z_0 = [(]\ x_0$
y_0), we have

$$G(z_0, t) = \begin{bmatrix} x_0 + \frac{1}{2}t^2 \\ y_0 + x_0 t + \frac{1}{6}t^3 \end{bmatrix}.$$

We do not write this as $G^t(x_0, y_0)$ since G **is not a flow**. The defining condition for a flow is that $G^s \circ G^t = G^{s+t}$. In our notation this would be $G(G(z_0, t), s) = G(z_0, s+t)$,

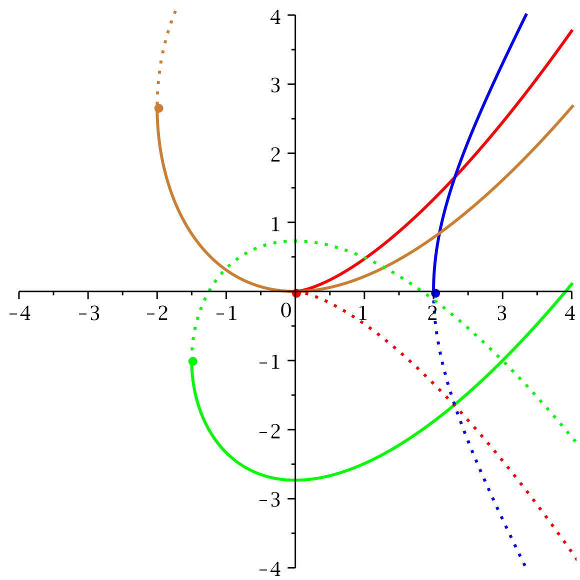


FIGURE 6.6. Four solution curves of the system (6.8). The dotted parts of the curve correspond to negative time.

and this is **not true**. Gor example, suppose $z_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $t = 1$ and $s = 1$. Then

$$G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, 1\right) = \begin{bmatrix} 0 + \frac{1}{2} \cdot 1^2 \\ 0 + 0 \cdot 1 + \frac{1}{6} \cdot 1^3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{6} \end{bmatrix}$$

$$G\left(G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, 1\right), 1\right) = G\left(\begin{bmatrix} \frac{1}{2} \\ \frac{1}{6} \end{bmatrix}, 1\right) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2} \cdot 1^2 \\ \frac{1}{6} + \frac{1}{2} \cdot 1 + \frac{1}{6} \cdot 1^3 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{5}{6} \end{bmatrix}$$

$$G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, 1 + 1\right) = G\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, 2\right) = \begin{bmatrix} 0 + \frac{1}{2} \cdot 2^2 \\ 0 + 0 \cdot 2 + \frac{1}{6} \cdot 2^3 \end{bmatrix} = \begin{bmatrix} 2 \\ \frac{4}{3} \end{bmatrix},$$

so $G(G(z_0, t), s) \neq G(z_0, s + t)$.

Not only do we **not** get a flow from the solution curves, but the solution curves do not look like the solution curves of an autonomous system in a very fundamental way: Different solution curves can intersect, and a single non-periodic solution curve can have self-intersections. See Figure 6.6.

It should be clear that the notion of orbit is not supported by these solution curves. In particular, there are no equilibrium points of this system. If we set $f(x, y, t) = 0$ and solve we find $t = 0$, $x = 0$, and y is arbitrary. But this can't correspond to an equilibrium point, since every solution curve of (6.8) has $x(t) = \frac{1}{2}t^2$ plus a constant, so there are no solutions that are constant.

There is a simple way to reinterpret a time-dependent system as an autonomous system. We change t to a new variable, s, which corresponds to absolute time, as measured by a fixed clock. Now we are free to reuse the variable t, and we interpret it as relative time, giving the difference between s and the initial time s_0, so $s = s_0 + t$. This leads to an autonomous system of equations plus initial conditions, where the independent variable is t and the independent variables are the original variables, plus the new variable s.

A solution to this reinterpreted system of equations will have the form $x(t)$, $s(t)$ where $s(t) = t_0 + t$. Then the corresponding solution to the original time-dependent system is given by $x(s - s_0)$. Formally,

Theorem 6.10. *$X(s)$ is a solution of the time-dependent initial value problem*

$$\frac{dX}{ds} = f(X, s), \quad X(s_0) = x_0$$

if and only if $X(s) = x(s - s_0)$ where $x(t)$, $s(t)$ is a solution of the autonomous initial value problem

$$\frac{dx}{dt} = f(x, s),\ \frac{ds}{dt} = 1, \quad x(0) = x_0,\ s(0) = s_0.$$

The proof of this is very simple; it is essentially just a change of variables.

We apply this transformation to our example (6.8). Forst we change the t variable on the right hand side to the new independent variable s, and we replace the initial value t_0 with s_0. Then we add the differential equation $\dfrac{ds}{dt} = 1$, and we change the initial conditions to indicate that the starting time is $t = 0$. This gives us the autonomous version

$$\frac{dx}{dt} = s,\ \frac{dy}{dt} = x,\ \frac{ds}{dt} = 1 \qquad x(0) = x_0,\ y(0) = y_0,\ s(0) = s_0. \tag{6.9}$$

The solution of this system is obtained by first solving for s, to get $s = s_0 + t$, and then proceeding as we did for the non-autonomous version. This produces

$$\begin{aligned} x(t) &= x_0 + s_0 t + \tfrac{1}{2}t^2 \\ y(t) &= y_0 + x_0 t + \tfrac{1}{2}s_0 t^2 + \tfrac{1}{6}t^3 \\ s(t) &= s_0 + t \end{aligned}$$

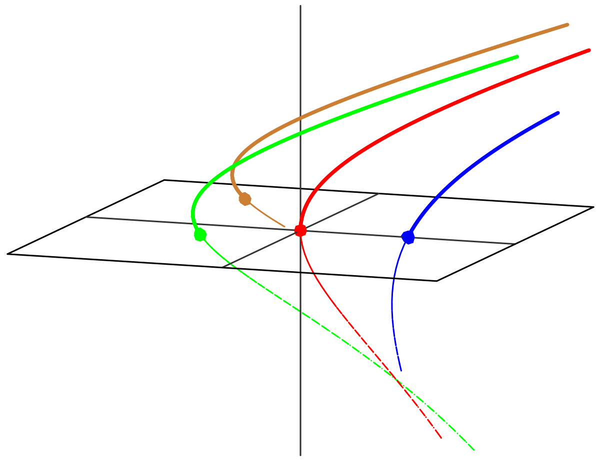


FIGURE 6.7. Solution curves of the autonomous version (6.9) of (6.8), using the same initial conditions as in Figure 6.6. The plane shown is the xy plane, and the vertical line is the s axis.

Now, since we are working with an autonomous system, we cna use these solution curves to define a flow:

$$F^t\left(\begin{bmatrix}x_0\\y_0\\s_0\end{bmatrix}\right)=\begin{bmatrix}x_0+s_0t+\frac{1}{2}t^2\\y_0+x_0t+\frac{1}{2}s_0t^2+\frac{1}{6}t^3\\s_0+t\end{bmatrix}.$$

Since we have a flow we know that the solution curves do not intersect, that they do not intersect each other unless they are periodic, and we can define orbits so we can thin about the phase portrait. However, the phase portrait is in three dimensions (x, y, s) rather than two. Figure 6.7 shows the same solution curves as in Figure 6.6, but interpreted in three dimensions.

6.4. Strange attractors

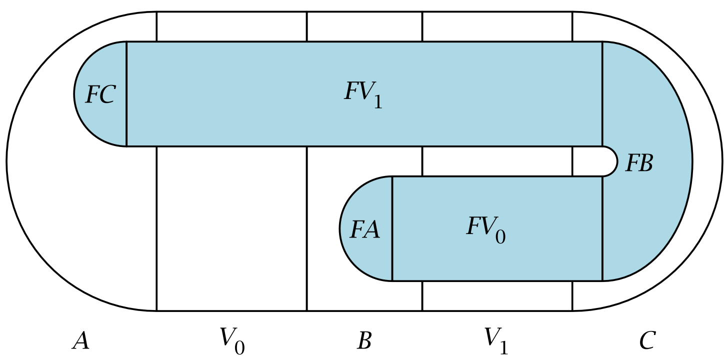
Exercises

6.1. For the horseshoe map:
(a) How many periodic points are there of period 6?
(b) How many periodic points are there of least period 6?
(c) How many periodic orbits are there of period 6? Give subtotals for the different orbit sizes.

6.2. (a) Find a periodic sequence of distance at most 2^{-6} from $\ldots 000.111\ldots$. What is its period?
(b) Suppose N is a positive integer and find a periodic sequence of distance at most 2^{-N} from $\ldots 000.111\ldots$. What is its period?

6.3. Generalize Exercise 6.2 to show that, if σ is any sequence in Σ and N is arbitrary then there is a periodic point within distance 2^{-N} of σ.

6.4. There are many variations on the basic horseshoe picture. Here's a "broken horseshoe":



(a) Draw the vertical rectangles that arise in this example. Specifically, show V_σ and H_σ where $\sigma$is a suitable string of 0's and 1's of length ≤ 4. This will not be the same as in the full horseshoe. For example, V_{00} is empty since there are no points that start in V_0 and are in V_0 after one iteration of F. [You do not need to draw the empty rectangles.]
(b) Draw the horizontal rectangles H_σ where σ has length ≤ 3.

6.5. Using the broken horseshoe of Exercise 6.4 we define the set Ω_1 of points that lie in $V_0 \cup V_1$ under all iterations, and we encode x in Ω_1 by the bi-infinite sequence σ so that $F^n(x)$ lies in V_{σ_n} for all n. Not all sequences can be obtained in this manner; for example, the sequence $\ldots 000.0000\ldots$ corresponds to a point that ramains in V_0 under all iterations, but no point in V_0 remains in V_0 for even one iteration.

Define

$$\Sigma_1 = \{\, \sigma \colon \sigma \text{ does not contain consecutive 0's} \,\}\,.$$

In other words, the condition for σ to be in Σ_1 is that whenever $\sigma_k = 0$ then $\sigma_{k+1} \neq 0$.

The following are three of the steps needed to show that sequences in Σ_1 are exactly the encodings of points in Ω_1. (The rest of the proof is the same as for the full horseshoe.)

(a) Suppose x is in Ω_1; show that the encoding of x is in Σ_1.

(b) Suppose σ is a *finite* sequence so that 00 does not occur in σ. Show that V_σ is non-empty.

(c) For a finite sequence sequence as in part (b), show that H_σ is non-empty.

6.6. Using the definition of Σ_1 in Exercise 6.5, find the number of periodic strings of period p, for $p \leq 5$ (at least). For $p = 1, 2, 3$ you should get 1, 3, 4. Do you see a pattern?