Dynamical Systems

Version 0.2

Dennis Pixton

E-mail address: dennis@math.binghamton.edu

Department of Mathematical Sciences
Binghamton University

Contents

CHAPTER 1

Discrete population models

1.1. The basic model

The first example is based on a simple population model. Suppose that a population of rabbits lives on an island (so the population is not affected by immigration from the outside, or emigration to the outside). The population is observed once per year. We'll use x for the number of rabbits, and t for the time, measured in years.

In each year a certain number of rabbits die, and this number is proportional to the number that are present at the start of the year. The proportionality constant c is the "death rate" or "mortality rate". That is, in one year cx rabbits die and are subtracted from the population of x. Similarly, a number of rabbits are born, and this number is also proportional to the number that are present at the start of the year. The proportionality constant b is the "birth rate". That is, in in one year bx rabbits are born and are added to the population of x rabbits. The net effect of this can be summarized as follows:

If the population at some time is x then the population one year later will be $x + bx - cx = (1 + b - c)x$. If we write $r = b - c$, the *net growth rate*, then we can rewrite this as $(1 + r)x$.

Thus we have a function F which specifies how the population changes during one year. In this case, $F(x) = (1 + r)x$. Once we know this function and we know the starting population x we can calculate the population any number of years later by *iterating* the function F. We will often use the alternate parameter $k = 1 + r$. For later reference we present this example as:

EXAMPLE 1.1. Exponential growth: $F(x) = (1 + r)x = kx$, for real numbers x. The parameter r is the net growth rate; it must be > -1. The alternate parameter is $k = 1 + r$, so $k > 0$.

For example, suppose the population is 1000 and the net growth rate is 0.1. Then, after one year, the population will be $F(1000) = (1 + .1) \cdot 1000 = 1100$. To find the population after another year just apply the function F to the population 1100, to get $F(1100) = (1 + .1) \cdot 1100 = 1210$. Then we can find the population after three years by another application: $F(1210) = 1.1 \cdot 1210 = 1331$. Thus if the population is x then, after three years, the population is obtained by iterating F three times,

starting with x, so (using the general formula), we obtain $F(F(F(x)))$. We calculate this by applying the function F three times, starting with x:

$$\begin{aligned} x &\to F(x) = (1+r)x \\ (1+r)x &\to F\Big((1+r)x\Big) = (1+r)\cdot(1+r)x = (1+r)^2x \\ (1+r)^2x &\to F\Big((1+r)^2x\Big) = (1+r)\cdot(1+r)^2x = (1+r)^3x. \end{aligned}$$

It should be clear (and it is easy to prove by induction) that, for any number of years t, the population after t years is $(1+r)^t x$, or $k^t x$. We will use the notation $F^t(x)$ for the t-fold iteration of F, so $F^t(x) = (1+r)^t x$ in this case. In terms of functional composition, $F^t = F \circ F \circ F \circ \cdots \circ F$ (where there are t copies of F), so this really is, in a sense, "F to the t^{th} power". As special cases, F^1 is just F, and F^0 is the identity function; that is, $F^0(x) = x$.

Once we have this formula for $F^t(x)$ we can make some observations about population growth:

(1) Although $(1+r)^t x$ makes mathematical sense for $x < 0$, this doesn't make sense in the model, since populations cannot be negative.
(2) $(1+r)^t x$ is defined mathematically for all real numbers t (with some worries if $1+r$ is 0 or negative), but there is no meaningful sense in which we can iterate a function t times, where t is not a non-negative integer. [We can make sense of F^t when t is a negative integer, by iterating the inverse function F^{-1}, if it exists.]
(3) $F^t(0) = 0$ for all t.
(4) If $r > 0$ then $(1+r) > 1$, so $(1+r)^t \to \infty$ as $t \to \infty$. Hence $F^t(x) \to \infty$ as $t \to \infty$ if $x > 0$.
(5) If $-1 \le r < 0$ then $0 \le (1+r) < 1$, so $(1+r)^t \to 0$ as $t \to \infty$. Hence $F^t(x) \to \infty$ as $t \to \infty$.
(6) If $r = 0$ then $F^t(x) = x$ for all t and all x.

This is a very simplistic model of population growth. Here are a few criticisms:

(1) $F^t(x)$ will generally yield non-integer values.
(2) $F^t(x) \to \infty$ can't happen: eventually there will be no space for the rabbits on the island.
(3) No model can really predict exact numbers of rabbits, since there will always be some random fluctuations.
(4) The model assumes that the net growth rate is constant, independent of both t and x.

1.2. Discrete dynamical systems

Before modifying this model we isolate the features of the model that constitute a *discrete dynamical system*:

In general, we will work with two variables, x and t. We always think of t as time, but the interpretation of x will depend on the particular application. In general, the values of x are called the *states* of the system. In most cases it will be necessary to specify more than just one number to describe the state of the system, so in most cases we will consider x to be a vector, and we'll refer to it as the *state vector* of the system. The *transition rule* for the system is a function F that transforms the state x at one time to the start $F(x)$ one time unit later. Then we can represent the state of the system t time units later by iterating F; that is, we calculate $F^t(x)$.

Essentially, it is the iterated function F^t which constitutes the dynamical system, although we often simply refer to the transition rule F as the dynamical system. It is called a *discrete dynamical system* to emphasize the fact that t is measured in discrete steps, so the values of t are restricted to integers – usually positive integers.

Note that the transition rule must be a function of x alone; it does not depend on t. We will see later how to accommodate transition rules that do depend on t. Also, the transition rule is a well-defined function. Hence, the dynamical system is *deterministic*, in the sense that the values of x and t uniquely determine the value of $F^t(x)$.

It may happen that F has an inverse function F^{-1}, and if this happens we say that the dynamical system is *invertible*. In this case we can make sense of F^t when t is a negative integer, by iterating the function F^{-1}.

The first example above illustrates these features. The state vector x is one dimensional – that is, it is just a number. The transition rule is $F(x) = (1+r)x$, and the state at time t is given by $F^t(x) = (1+r)^t x$ for $t = 0, 1, 2, 3, \ldots$. If $r \neq -1$ this system is invertible, with $F^{-1}(x) = (1+r)^{-1}x$, and the formula $F^t(x) = (1+r)^t x$ holds for all integers.

The first example illustrates some common features of dynamical systems:

First, the domain of F is all real numbers, but only a subset of x values is relevant to the model, since a population cannot be negative. In other examples the domain of F must be restricted – for example, to avoid division by zero or square roots of negative numbers. The domain of F is called the *state space*. The set of x values that are relevant to the model is then a subset of the state space.

Second, the transition rule usually incorporates one or more *parameters*. These are quantities that are constant for a specific instance of the dynamical system, and which, in applications, are either determined empirically or are adjusted externally.

For example, in our original population model the birth and death rates are parameters. The model only depends on their difference, so it is reasonable to talk of just one parameter, the net growth rate, r. It is often necessary to restrict the possible values of the parameters. For example, $1 + r$ cannot be 0 if we want the system to be invertible, and $1 + r$ cannot be less than 0 in our population model, since this leads immediately to negative populations.

Here's an important property of dynamical systems:

PROPOSITION 1.2. *Group action:* $F^t \circ F^s = F^{t+s}$.

1.3. Some variations

We now consider two variations on the original model.

The first variation involves external interaction with the population. We envision a yearly addition to (or subtraction from) the population. For example, if $r < 0$, so the population is dying out, we might add a number of rabbits each year to stabilize the population. If $r > 0$, so the population is incresing without limit, we might remove a number of rabbits each year in order to keep the population from exploding. This gives the following example:

EXAMPLE 1.3. $F(x) = (1+r)x+A = kx+A$, for real numbers x. The parameter A is any real number, and k and $1 + r$ are defined as in Example 1.1.

The effect of this model is that, each year, A rabbits are added to the population. In this model the addition occurs at the end of the year, just before the population is counted. [If the addition occured at the beginning of the year then the transition rule would be $F(x) = (1 + r)(x + A)$. Of course A might be negative, in which case rabbits are removed rather than added.

Iterating a few times gives

$$\begin{aligned}
F^1(x) &= kx + A \\
F^2(x) &= F(F(x)) = F(kx + A) = k(kx + A) + A = k^2x + A + Ak \\
F^3(x) &= F(F^2(x)) = F(k^2x + A + Ak) = k(k^2x + A + Ak)) + A \\
&= k^3x + A + Ak + Ak^2 \\
F^4(x) &= k^4x + A + Ak + Ak^2 + Ak^3.
\end{aligned}$$

The pattern should now be clear; the general form is

$$F^t(x) = k^t x + A\left(1 + k + k^2 + \cdots + k^{t-1}\right).$$

If you look up the *geometric series* you will find a formula for the sum in parentheses. If $k \neq 1$ this allows us to write

$$F^t(x) = k^t x + A \cdot \frac{k^t - 1}{k - 1} = k^t x + A \cdot \frac{1 - k^t}{1 - k}. \tag{1.1}$$

We now look at a second variation on the basic model.

Instead of looking at the effect of adding population from the outside we consider a way of adjusting the growth rate r to take into account the effects of overcrowding. The idea is that if the population x becomes too large then the rabbits on our island will not be able to find enough to eat, so we expect that the death rate will go up. Also, they will be less healthy, so we expect the birth rate to go down. In other words, we expect that the net growth rate will decrease as the population increases.

Our new variation amounts to replacing the constant growth rate r with a function of x, which is approximately equal to r for small values of x but which decreases as x increases. The simplest function that satisfies these conditions is a linear function of the form $r - \alpha x$ where α is a positive constant. There are many other candidates for a variable growth rate, but we will concentrate on this one because of its simplicity. This gives us the next example:

EXAMPLE 1.4. The *logistic model*: $F(x) = (1+r-\alpha x)x = (k-\alpha x)x = kx - \alpha x^2$ for real numbers x. The parameter α is a positive real number, and k and $1 + r$ are defined as in Example 1.1.

The function $kx - \alpha x^2$ is called the *logistic map*. It was introduced the biologist Robert May in 1976, based on the *logistic differential equation* which we will study in the next chapter. The logistic model turns out to have very complex behaviour. It has been the subject of hundreds of research papers since then, and has been a fundamental example in understanding chaotic dynamical systems. The analysis of the logistic equation in the complex domain leads to the the Mandelbrot set and related fractals.

This is not an invertible dynamical system (unless we restrict its domain) since F is a quadratic polynomial, so the equation $F(x) = y$ will usually have two, or zero, solutions. In other words, the function F does not have an inverse.

We can iterate the logistic map:

$$\begin{aligned} F^1(x) &= kx - \alpha x^2 \\ F^2(x) &= F(F(x)) = k(kx - \alpha x^2) - \alpha(kx - \alpha x^2)^2 \\ &= k^2 x - \alpha(k + k^2)x^2 + 2\alpha^2 k x^3 - \alpha^3 x^4. \end{aligned}$$

However, the calculations rapidly become very messy – in fact, $F^t(x)$ is a complicated polynomial of degree 2^t in x, and there is no simple general form for it.

On the other hand, it is easy to calculate as many iterations as we need in a specific case. For example, let $r = .1$ and $\alpha = .0001$. Starting with $x = 2000$ we calculate

$$\begin{aligned} F^1(x) &= 1800 \\ F^2(x) &= 1656 \\ F^3(x) &= 1547 \\ F^4(x) &= 1462 \\ F^5(x) &= 1394 \\ F^6(x) &= 1339 \\ F^7(x) &= 1294 \\ F^8(x) &= 1256 \\ F^9(x) &= 1224 \\ F^{10}(x) &= 1196 \end{aligned} \tag{1.2}$$

Notice that $F^t(2000)$ is decreasing as t increases, although $r > 0$. It is not clear, without further analysis, whether this decrease continues, or what the limiting value is.

1.4. Steady states and limit states

A *steady state* of a dynamical system is a state x_0 so that $F^t(x_0) = x_0$ for all values of t. For example, in our original model, 0 is a steady state, since $F^t(0) = (1 + r)^t \cdot 0 = 0$ for all t.

Since F^t is defined by iterating F, finding a steady state is the same as finding a state for which $F(x_0) = x_0$. In general, such a value x_0 is called a *fixed point* for F. Often such fixed points can be found by just solving an algebraic equation. For example, in our second model, $F(x) = kx + A$, so we can find the fixed points of F by solving $F(x) = x$ for x, as follows:

$$\begin{aligned} kx + A &= x \\ A &= x - kx = (1 - k)x \\ x &= \frac{A}{1 - k} \end{aligned}$$

If A and $1 - k = r$ both have the same sign then $x_0 = \dfrac{A}{1 - k} > 0$ so it represents a population, and this population will not change with time.

Next, suppose that x_1 is a state of a dynamical system. If $F^t(x_1)$ converges to a state x_* as $t \to \infty$ and $F^t(x_1) \neq x_*$ for any t then x_* is called a *limit state* of the system. For example, in our original model, 0 is a limit state if $-1 < r < 0$ since in that case $0 < k = 1 + r < 1$ so $F^t(x) = k^t x \to 0$ for any x. On the other hand, if $k > 1$ then 0 is not a limit state, since $F^t(x) = k^t x \to \pm\infty$ as $t \to \infty$ if $x > 0$. It is true that $F^t(0) \to 0$ as $t \to \infty$ in this case, since 0 is a steady state; but 0 is not a limit state of the system since we required $x_* \neq x_1$ in the definition.

The examples above show that a steady state may – or may not – be a limit state. In the other direction we have:

PROPOSITION 1.5. *If F is continuous then any limit state is a steady state.*

PROOF. Suppose that $F^t(x_1) \to x_*$ as $t \to \infty$. Then, if we replace t with $t + 1$, we also have $F^{t+1}(x_1) \to x_*$. But we can write $F^{t+1}(x_1) = F(F^t(x_1))$. If we take the limit in this equation as $t \to \infty$ and use the fact that F is continuous we get

$$x_* = \lim_{t\to\infty} F^t(x_1) = \lim_{t\to\infty} F(F^t(x_1)) = F\left(\lim_{t\to\infty} F^t(x_1)\right) = F(x_*),$$

so $F(x_*) = x_*$. Since x_* is a fixed point of F it is a steady state. □

So suppose we have found a steady state x_0. How can we tell if it is a limit state? In some important cases we can answer this easily, but we need some definitions.

A steady state x_0 is called an *attractor* or a *sink* if all nearby states stay near to x_0 and also converge to x_0 as $t \to \infty$. In the other direction, a steady state x_0 is called a *repeller* or a *source* if all nearby states that are not equal to x_0 move away from x_0.

More precisely, here are the formal definitions:

(1) A steady state x_0 is a sink if for any positive r there is a positive s so that, for any state x within distance s of x_0, $F^t(x)$ remains within distance r of x_0 for all positive t and converges to x_0 as $t \to +\infty$.
(2) A steady state x_0 is a source if there is a positive constant c so that if x is any state within distance c of x_0 but not equal to x_0 then there is some positive t so that the distance from $F^t(x)$ to x_0 is greater than c.

Execpt in some very unusual cases, a sink is a limit state, but a source is not.

If F is invertible there is a useful connection between sinks and sources:

PROPOSITION 1.6. *Suppose F is continuous and has a continuous inverse. Then a steady state x_0 is a source for F if and only if it is a sink for F^{-1}.*

The proof of Proposition 1.6 requires some ideas from topology, so we will not discuss it.

It is not always easy to decide whether a steady state is a sink or source, but it is easy to check whether a steady state x_0 is *hyperbolic*: This just means that

the derivative of F at x_0 is not equal to ± 1. In other words, a steady state x_0 is hyperbolic if $|F'(x_0)| \neq 1$.

THEOREM 1.7. *Suppose that F has a continuous derivative. Suppose x_0 is a hyperbolic fixed point of F. Then:*

(1) *x_0 is a sink if $|F'(x_0)| < 1$.*
(2) *x_0 is a source if $|F'(x_0)| > 1$.*

If x_0 is not hyperbolic then it may be a sink or a source or neither.

For example, in Example 1.3 we have $F(x) = kx + A$, so $F'(x) = k$. If $k < 1$ then the fixed point $x_0 = \dfrac{A}{1-k}$ is a sink, while if $k > 1$ the fixed point is a source.

PROOF. Suppose that $|F'(x_0)| < 1$. To satisfy the definition of a sink we start with a positive number r.

Pick a number M between $|F'(x_0)|$ and 1, so $|F'(x_0)| < m < 1$. By continuity of F' there is some $\delta > 0$ so that $|F'(c)| \leq M < 1$ if $|x_0 - c| < \delta$. We let s be the minimum of r and δ. If $|x_0 - x| \leq s$ then, using the Mean Value Inequality, Theorem A.5,

$$|x_0 - F(x)| \leq M \cdot |x_0 - x| .$$

Since $M < 1$ this implies that $F(x)$ is closer to x_0 than x is, so it is within s of x_0. This means that we can iterate this argument, so $F^t(x)$ is always within s of x_0. Moreover, under each iteration the distance is reduced by at least a factor of M, so we have $\left|x_0 - F^t(x)\right| \leq M^t \cdot |x_0 - x|$. Taking the limit shows that $\left|x_0 - F^t(x)\right| \to 0$ (since $M < 1$), so $F^t(x) \to x_0$.

The argument for sources is similar. □

Note: The test in Theorem 1.7 is very similar to various tests for convergence of series – for example, the Ratio Test for convergence of infinite series. Note that there is a "borderline" case, with $F'(x_0) \neq \pm 1$, in which the test gives no information.

1.5. Bounce graphs

It is difficult to visualize iteration, or limiting behavior, from the graph of $y = F(x)$. If we start with a value of x_0, then we can calculate $y_0 = F(x_0)$ from the graph. However, we then have to reinterpret this y value as an x value before we can continue. That is, we set $x_1 = y_1$, find x_1 on the x-axis, and then use the graph to calculate $y_1 = F(x_1)$. And so on.

There is a much simpler way to visualize this. We first notice that on the diagonal line D given by $y = x$ there is no difference between x and y values. So instead of starting with x_0 on the x-axis we start with the corresponding point (s_0, x_0) on the line D. We then calculate the next point (x_1, x_1) in a two-step procedure: First move

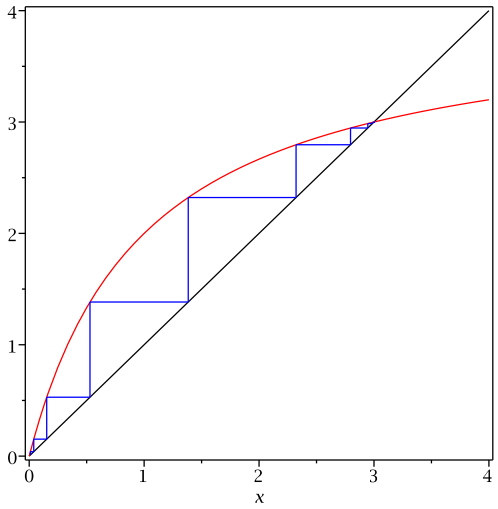


Figure 1.1: $F(x) = \dfrac{4x}{x+1}$

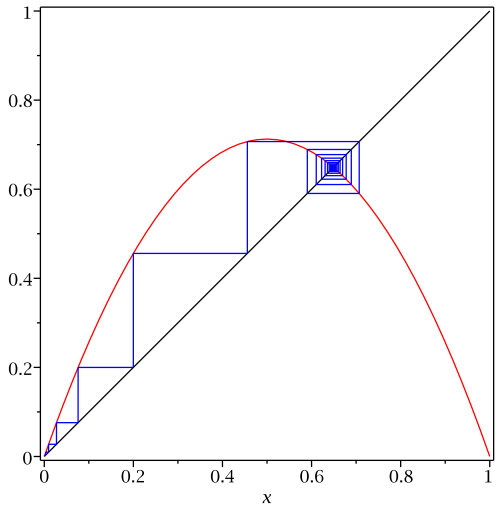


Figure 1.2: $F(x) = 2.85\,x(1-x)$

vertically from (x_0, x_0) until you hit the graph of $y = F(x)$. This determines the point $(x_0, F(x_0))$; sonce $F(x_0) = x_1)$ you are now at the point (x_0, x_1). Now move **horizontally** from (x_0, x_1) until you hit the diagonal line D. You are now at the point (x_1, x_1), so you are ready to start again.

It helps to visualize this by drawing the vertical line segment from (x_0, x_0) to (x_0, x_1), and then the horizontal line segment from (x_0, x_1) to (x_1, x_1). You are, in effect, moving from the diagonal D back to D by "bouncing" off the graph $y = F(x)$, so this is called a "bounce graph". You can continue to find (x_2, x_2), (x_3, x_3), ... by continuing to bounce. **Remember**: The bouncing starts and ends on the diagonal line D.

In Figure 1.1 the dynamical system is specified by $F(x) = \dfrac{4x}{x+1}$. The graph of F crosses the diagonal $y = x$ at two points; these are the fixed points. Some easy algebra shows that the fixed points are 0 and 3. To classify these fixed points we take their derivatives: $F'(x) = \dfrac{4}{(x+1)^2}$. Then $F(0) = 4$ so 0 is a source, and $F'(3) = 1/4$, so 3 is a sink. The bounce graph shows this. First, if x_0 is near 3 and less than 3 then x_t will increase as t increases, and will eventually converge to 3. This is shown in the graph, where $x_0 = .01$, and if you start at a state $x_0 > 3$ and follow a few bounces you will see that x_t now decreases as t increases and converges to 3 from above. On the other hand, if you start with x_0 near 0 and positive then x_t increases as t increases, and eventually x_t will increase until it is greater than 1.

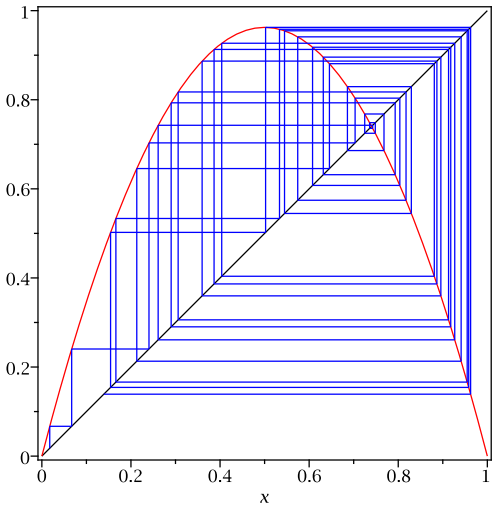


Figure 1.3: $F(x) = 3.85\,x(1-x)$

Similarly, if x_0 is negative and near 0 then x_t will decrease, and it will eventually become smaller than $-.5$.

Figure 1.1 illustrates a very common situation. The curve crosses the diagonal at 0 and at 3, and for x between 0 and 3 the curve is increasing (since $F'(x) > 0$), and it lies above the diagonal. Hence, for any starting point in the interval $(0, 3)$ the bounce graph is increasing and is trapped between 0 and 3. So all it can do is converge to 3.

In Figure 1.2 the dynamical system is specified by $F(x) = 2.85\,x(1-x)$. Again there are two fixed points, one at 0 and the other at $1 - \dfrac{1}{2.85} \approx 0.64912$. The derivative is $F'(x) = 2.85 - 5.7\,x$. At 0 this is 2.85, so 0 is a source. At the other fixed point the derivative is -0.85. This has absolute value less than 1, so 0.64912 is a sink. The bounce graph again shows this. However, there is a new feature here; the negative derivative at the sink forces the bounce graph to *spiral* toward the sink as t increases.

Figure 1.3 shows a "chaotic" bounce graph, obtained by following the dynamical system $F(x) = 3.85\,x(1-x)$. In this case both fixed points, at 0 and at $1 - \dfrac{1}{3.85} \approx 0.7403$, are sources, with spiralling behavior at 0.7403. It is possible for x_t to get close to 0.7403 but, unless it is exactly equal to the fixed point, it will spiral away again. However, it must stay in the interval $[0, 1]$ so, unless it gets "trapped" somewhere else, it may eventually bounce back near 0.7403. It is very difficult to analyze this

behavior precisely. This is an example of the logistic function, and this sort of chaotic picture is one of the main reasons that so much research has used the logistic function as a motivating model.

Exercises

1.1. Use a calculator to compute several values of $F^t(x)$ for $x = 500$, using the same system as for equation (1.2). You should find that the states are now *increasing*, rather than decreasing as in equation (1.2).

1.2. Find the steady states of Example 1.4, using the same parameters as in equation (1.2). Apply Theorem 1.7 to show that one of these is a sink and the other is a source.

1.3. Using Example 1.4: Find the maximum value of $F(x)$ in terms of k and α.

1.4. Example 1.4 has two states x which satisfy $F(x) = 0$. One is 0. Find the other one, in terms of k and α; call this state M. In a population model we need to restrict x to be between 0 and M, since $F(x) < 0$ if $x < 0$ or $x > M$. Demonstrate this graphically, by plotting the equation $y = F(x)$. [This is a parabola. Where does it cross the axis? Is it concave up or down?]

1.5. Example 1.4 generally has two steady states; one is 0.

(a) Determine the other steady state, in terms of k and α.
(b) What conditions on k and α lead to just one steady state?
(c) What conditions on k and α will guarantee that the second steady state is between 0 and M?
(d) What conditions on k and α will guarantee that 0 is a hyperbolic sink?
(e) What conditions on k and α will guarantee that the second fixed point is a hyperbolic sink?

1.6. Suppose $k = 8$ and $\alpha = .001$ in Example 1.4. Choose several starting values x_0 randomly between 0 and M and, for each one, calculate several values of $F^t(x_0)$. You should find that, in essentially all cases, $F^t(x_0)$ becomes negative (stop iterating when this happens). What does this say about the population? Can you explain what is happening?

1.7. Consider the discrete dynamical system defined by the transition rule $F(x) = kx^2$, where k is a positive constant. Find $F^2(x)$, $F^3(x)$, $F^4(x)$. Can you see the general form for $F^t(x)$? What can you say about limit states?

1.8. Let $F(x) = 2e^{-x}$. Plot $y = F(x)$ and $y = x$ on the same axes and show that the dynamical system determined by F has a fixed point x_0 so that $0 < x_0 < 1$. Show that this fixed point is a sink. Start with $x = 0$ and calculate iterations $F^t(x)$ until they are within .05 of the fixed point.

1.9. Newton's Method for solving equations numerically is an example of a dynamical system. The method transforms solving the equation $h(x) = 0$ into iterating

the dynamical system given by $F(x) = x - \dfrac{h(x)}{h'(x)}$, as follows. If x_* is a limit state then x_* is a steady state, so $F(x_*) = x_*$. This means $x_* - \dfrac{h(x_*)}{h'(x_*)} = x_*$, and this simplifies to $h(x_*) = 0$, assuming that $h'(x_*) \neq 0$. Thus limit states of the dynamical system are solutions of $h(x) = 0$, and sinks of the dynamical system correspond to solutions of $h(x) = 0$ that are *stable*, in the sense that any reasonably close approximate solution will increase in accuracy as the iteration continues.

(a) Show that Newton's mathod, applied to solving $x^3 - 2 = 0$, corresponds to iterating $F(x) = \dfrac{2}{3}\left(x + \dfrac{1}{x^2}\right)$

(b) Show that $\sqrt[3]{2}$ is a sink for this dynamical system.

(c) Prepare a rough bounce graph indicating how the iteration proceeds, starting with $x_0 = 0.3$. [First produce a reasonably accurate plot of $y = F(x)$ for $0 \leq x \leq 10$ and $0 \leq y \leq 10$.]

(d) What happens if you start the iteration at $x_0 = -10$?

1.10. Figure 1.4 shows a bounce graph for $F(x) = 3.15\,x(1-x)$. Both fixed points are sources, but the eventual behavior of x_t is not chaotic; in fact it seems that the bounce graph converges to a square. Discuss this example. [You might want to start

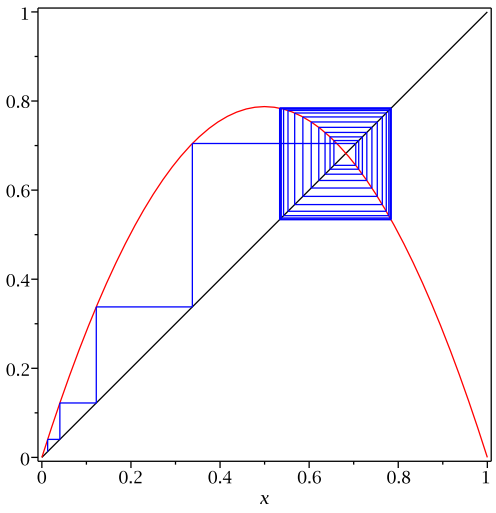


Figure 1.4: $F(x) = 3.15\,x(1-x)$

by concentrating on the corners of the limit square; call them (a, a) and (b, b). What can you say about $F(a)$? What about $F(b)$? How would you determine a and b?.]

1.11. Let $F(x) = x^2 - 2x + 2$. Find the steady states.

1.12. Let $F(x) = \dfrac{x}{2x+1}$, for $x \geq 0$.

(a) Find the steady state. (There is only one; remember $x \geq 0$.)
(b) Calculate $F^t(1)$ for several values of t. You should see a pattern. Can you write a non-recursive formula for $F^t(1)$? What is $\lim_{t\to\infty} F^t(1)$?
(c) Can you generalize part (b) by replacing 1 with any $x > 0$?

1.13. Let $F(x) = \dfrac{x}{x^2+1}$, for any real x.

(a) Find the steady state. (There is only one.)
(b) Show that $F(x) < x$ for all $x \geq 0$.
(c) Why is $\lim_{t\to\infty} F^t(x) = 0$ for any $x \geq 0$?

CHAPTER 2

Continuous population models

2.1. The basic model

We now reconsider Example 1.1 with the assumption that the birth and death properties occur at the same rate throughout the year. In this case we still have a net growth parameter r, which is again interpreted as the net change in the population per unit of population, per unit of time. In other words, starting with a population x, the change in the population after one unit of time (a year in the rabbit example) is rx. But now we reinterpret this to mean that the population changes at a constant rate of r throughout the year. This means that, in a small interval Δt of time, the change in the population is given by

$$\Delta x = rx\Delta t.$$

In fact, this formula is not correct for large values of Δt, but it becomes more precise as Δt becomes smaller. This leads us to rewrite this equation as

$$\frac{\Delta x}{\Delta t} = rx$$

and then take a limit:

EXAMPLE 2.1. $\dfrac{dx}{dt} = rx$, where $x = x(t)$ is a real function of t. The parameter r is the instantaneous net growth rate; it can be any real number.

Just as in the discrete case we want to investigate what happens in the future according to this model. However, there is no transition function F, so there is nothing that we can iterate to determine F^t. Rather, we must determine F^t by solving the *differential equation* $\dfrac{dx}{dt} = rx$. We will often use the notation x' for the derivative of x with respect to t, so this differential equation is more compactly written as $x' = rx$.

There is a notational complication here. In the case of a discrete system we interpreted x as a simple variable, giving our population at a fixed time, and then later values of the population were determined as $F^t(x)$, by iteration of the transition function. However, in the differential equation $x' = rx$, x is a *function* of t, not a simple variable. Our interpretation is that, at an initial time $t = 0$ the function x has

(1) Rewrite $\dfrac{dx}{dt} = f(x,t)$ in differential form: $dx = f(x,t)\,dt$.
(2) Rearrange the equation so that all references to x, *including* dx, are on the left hand side and all references to t, *including* dt, are on the right hand side. If you can't do this then do not continue: this method *fails*.
(3) Find the indefinite integrals of both side. Do not forget to add **one** constant of integration, C.
(4) Solve for x in terms of t and C.
(5) Plug in $t = 0$, remember that $x(0) = x_0$, and solve for C in terms of x_0.
(6) Plug in this value for C in the formula for $x(t)$ and interpret $x(t)$ as $F^t(x_0)$.
(7) You may have performed some dubious steps, like division by zero, so check the answer by verifying that $x(0) = x_0$ and $\dfrac{d}{dt}x(t) = f(x(t),t)$.

Figure 2.1: Solving $\dfrac{dx}{dt} = f(x,t),\ x(0) = x_0$ by separation of variables.

a fixed value, say $x(0) = x_0$, and then $x(t)$, for later time, represents the population t units of time later. In other words,

$$F^t(x_0) = x(t), \quad \text{where } x' = rx \text{ and } x(0) = x_0.$$

Here we resolve the ambiguity in interpreting x by using x_0 to indicate a fixed value of x.

Thus in order to see the future population $F^t(x_0)$ we need to solve the differential equation $x' = rx$, together with the *initial value condition* $x(0) = x_0$. We will follow a simple procedure for doing this, summarized in Figure 2.1. Here are the steps:

(1) Rewrite the equation $\dfrac{dx}{dt} = rx$ in differential form:

$$dx = rx\,dt.$$

(2) Rearrange the equation so that all references to x, *including* dx, are on the left hand side and all references to t, *including* dt, are on the right hand side:

$$\frac{dx}{x} = r\,dt.$$

(In this case it does not really matter whether the parameter r is kept with the t's or with the x's.)

(3) Find the indefinite integrals of both sides.

$$\int \frac{dx}{x} = \int r\,dt$$
$$\ln|x(t)| = rt + C$$

Only one constant of integration is necessary: in effect, the two constants are combined into one and put on the right hand side.

(4) Solve for x in terms of t and C:

$$|x(t)| = e^{rt+C} = e^C e^{rt}$$
$$x(t) = \pm e^C e^{rt} = Be^{rt}.$$

We defined the constant B as $\pm e^C$ to simplify the expression.

(5) Plug in $t = 0$, remember that $x(0) = x_0$, and solve for the constant B in terms of x_0:

$$x_0 = x(0) = Be^{r\cdot 0} = Be^0 = B,$$

so $B = x_0$.

(6) Plug in this value for B in the formula for $x(t)$ and interpret $x(t)$ as $F^t(x_0)$:

$$x(t) = Be^r t = x_0 e^{rt},$$

so $F^t(x_0) = x_0 e^{rt}$.

(7) Check the answer: Plug in $t = 0$ in $x(t) = x_0 e^{rt}$ to get $x(0) = x_0 \cdot e^0 = x_0$, which is correct. Also, $\dfrac{d}{dt}x(t) = \dfrac{d}{dt}x_0 e^{rt} = x_0 \cdot re^{rt} = rx_0 e^{rt}$ and $rx(t) = r \cdot x_0 e^{rt} = rx_0 e^{rt}$, so we have verified that $\dfrac{d}{dt}x(t) = rx(t)$. Therefore our formula for $F^t(x_0)$ is valid for all values of x_0 and t.

Notice that the solution is valid even when $x_0 = 0$. However, the derivation above must exclude $x = 0$ (bacause of the division by 0 in step 2). Also, $x_0 = B = \pm e^C$ and an exponential is never 0, so x_0 can't be 0 in step 5. So we found an extra solution by verifying the answer.

We now have a formula which describes how an initial value of the population changes as a function of t: $F^t(x_0) = x_0 e^{rt}$. As in the discrete case, we interpret both x_0 and t as variables in this expression. However, now t can be any real number, not just an integer.

It is important to realize that the parameter r has a different meaning here than in Example 1.1. For example, in the numeric example following Example 1.1 we used $r = 0.1$ and $x_0 = 1000$ and obtained $1000 \cdot (1 + .1) = 1100$ as the population after one year. However, in our continuous model, the population after one year is $1000 \cdot e^{r\cdot 1} = 1000 \cdot e^{0.1} = 1105.17$.

This difference might be familiar in terms of interest calculations. If we reinterpret 1000 rabbits as 1000 dollars and the island as a savings account, then the net growth rate of .1 corresponds to an interest rate of 10% per year. In the discrete case we consider that this interest is calculated and added to the money in the account once, at the end of one year. A more common practice is to calculate the interest once per quarter; in this case the interest rate per quarter would be $r \cdot \Delta t = 0.1 \cdot \frac{1}{4} = 0.025$, so interest of 2.5% is calculated and added to the account at the end of each quarter. Since the added interest is used when calculating the next interest payment the bank says that this interest is "compounded quarterly"; the effect is that, at the end of four quarters, the amount in the account is $1000 \cdot (1.025)^4 = 1103.81$ instead of 1100. Of course, some accounts might compound the interest monthly, in wich case the amount would be $1000 \cdot \left(1 + \frac{0.1}{12}\right)^{12} = 1000 * (1.008333...)^{12} = 1104.71$ after 12 months. If the interest is compounded daily then the amount after a year is $1000 \cdot \left(1 + \frac{0.1}{365}\right)^{365} = 1105.16$. Many accounts take this subdivision of the year to the limit and say that the interest is "compounded continuously". In this case the amount after one year would be $1000 \cdot e^{0.1} = 1105.17$.

2.2. Continuous dynamical systems

Before modifying this model we isolate the features of the model that constitute a *continuous dynamical system*:

In general, we will work with two variables, x and t. We always think of t as time, but the interpretation of x will depend on the particular application. In general, the values of x are called the *states* of the system. In most cases it will be necessary to specify more than just one number to describe the state of the system, so in most cases we will consider x to be a vector, and we'll refer to it as the *state vector* of the system. The evolution of the system is governed by a differential equation of the form $\frac{dx}{dt} = f(x)$.

As in the discrete case, it is necessary for much of our analysis that the right hand side, $f(x)$, is a function *only* of x. It may depend on parameters, but not on t. We will discuss in a later chapter what happens if t appears on the right hand side.

The general solution of the differential equation expresses x as a function of t together with a constant C of integration. Using the initial value x_0 of x we can write C in terms of x_0, so we can write the formula for $x(t)$ as a function of x_0 and t. This function is written $F^t(x_0)$ and is called the *flow* of the dynamical system. As in the discrete case, F^t transforms the state x_0 at one time to the state $F^t(x_0)$

of the system after t units of time have elapsed. Unlike in the discrete case, time is measured by a real number, not an integer.

In the discrete case the values of $F^t(x_0)$ are calculated by iterating the transition rule F, so there is no problem about defining it as long as the state stays in the domain of the transition rule. In the continuous case we do not calculate $F^t(x_0)$ by a simple iteration procedure, so we shall require some conditions on the differential equation $x' = f(x)$ to ensure that the flow $F^t(x_0)$ can be properly defined.

Just as in the discrete case, we usually need to restrict the set of possible state vectors, either to conform to the process that we are modelling or to avoid mathematical difficulties. Also, the differential equation $x' = f(x)$ usually involves parameters, and the qualitative characteristics of the solution may change as the values of the parameters are varied.

Here are two important principles for dealing with solution curves graphically.

PROPOSITION 2.2. *The graphs of different solutions $x_1(t)$ and $x_2(t)$ do not cross. (This requires the hypotheses of Theorem 2.9.)*

PROPOSITION 2.3. *Group action: $F^t \circ F^s = F^{t+s}$. (This requires the hypotheses of Theorem 2.9.)*

In fact, Porposition 2.3 is the same statement as Proposition 1.2, but with a different meaning, since t is now a continuous variable.

These two propositions require proof, which we postpone until section 2.5.

2.3. Some variations

We now consider the continuous analogs of the two discrete models from section 1.3

The first variation involves external interaction with the population. We envision a steady migration to (or from) the population. This means that in a single year a number A of rabbits move to (if $A > 0$) or from (if $A < 0$) the island. [This is not too reasonable if we think of the rabbits as beinc confined to an island, but it is not too unreasonable if we consider a group of rabbits that is mostly isolated from the larger population.]

We are thinking of this migration as occurring at a steady rate throughout the year, so if A individuals migrate in one year, then $A\Delta t$ will migrate in a time interval of length Δt. Adding this to our derivation of Example 2.1 gives a change of

$$\Delta x = rx\Delta t + A\Delta t$$

for a small change in time. Dividing by Δt and taking the limit as $\Delta t \to 0$ produces the following model:

EXAMPLE 2.4. $\dfrac{dx}{dt} = rx + A$, so $f(x) = rx + A$. The instantaneous growth rate r and the annual migration rate A can be any real numbers.

We follow the procedure in Figure 2.1 to determine the flow $F^t(x_0)$:

(1) Rewrite the equation $\dfrac{dx}{dt} = rx + A$ in differential form:

$$dx = (rx + A)\,dt.$$

(2) Separate the variables:

$$\frac{dx}{rx + A} = dt.$$

(3) Integrate:

$$\int \frac{dx}{rx + A} = \int dt$$
$$\frac{1}{r}\ln|rx + A| = t + C_1.$$

Here C_1 is the constant of integration and the factor of $\frac{1}{r}$ comes from integration using the substitution $u = rx + A$.

(4) Solve for x:

$$\begin{aligned}
\ln|rx + A| = r(t + C_1) = rt + C_2 &\qquad \text{substitute } C_2 = rC_1\\
|rx + A| = e^{rt+C_2} = e^{rt}e^{C_2} = C_3e^{rt} &\qquad \text{exponentiate, substitute } C_3 = e^{C_2}\\
rx + A = \pm C_3e^{rt} = C_4e^{rt} &\qquad \text{substitute } C_4 = \pm C_3\\
x = -\frac{A}{r} + \frac{C_4}{r}e^{rt} = -\frac{A}{r} + Ce^{rt} &\qquad \text{algebra, substitute } C = \frac{C_4}{r}.
\end{aligned}$$

(5) Plug in $t = 0$, $x = x_0$ and solve for C in terms of x_0:

$$x_0 = -\frac{A}{r} + Ce^{r\cdot 0} = -\frac{A}{r} + Ce^0 = -\frac{A}{r} + C$$
$$C = x_0 + \frac{A}{r}.$$

(6) Replace C in the formula for $x(t)$ with its expression in terms of x_0:

$$F^t(x_0) = x(t) = -\frac{A}{r} + Ce^{rt} = -\frac{A}{r} + \left(x_0 + \frac{A}{r}\right)e^{rt}$$

(7) Check your work. Notice that the derivation above assumes that $rx + A \neq 0$, so $x \neq -\dfrac{A}{r}$, so it is a good idea to check that the answer still works if $x_0 = -\dfrac{A}{r}$. It does, so $F^t(x_0)$ is defined for all values of x_0 and t.

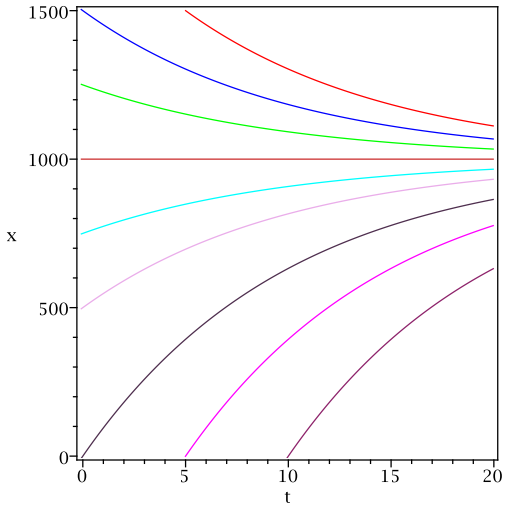


Figure 2.2: $x' = rx + A,\ r = -0.1,\ A = 100$

We can visualize the flow $F^t(x_0)$: We graph a number of solutions $x(t)$ of the differential equation, for different values of x_0, on the same plot, with a horizontal T-axis and a vertical X-axis. The initial value, x_0, for each curve is the x value where the curve crosses the X-axis. we can then see what F^{t_1} does for a fixed value of t_1: Locate an x_0 value on the vertical line $t = 0$, and follow the solution curve $x(t)$ through this point until it meets the vertical line $t = t_1$.

Figures 2.2 and 2.3 provide such a visualization for our example, the first with a negative growth rate and the second with a positive growth rate. For example, it is clear from Figure 2.2 that, in this case, F^{20} compresses the interval $0 \le x \le 1500$ when $t = 0$ to an interval given approximately by $850 \le x \le 1050$ when $t = 20$.

Note that in Figure 2.2 it appears that the migration rate $A = 100$ is sufficient to balance the negative net growth rate of -0.1, and any initial population seems to converge to a stable population of 1000. In Figure 2.3 there is a negative migration rate $A = -100$ and a positive growth rate $r = 0.1$. It is true that the initial population $x_0 = 1000$ remains constant, but any population less than 1000 eventually dies out, while any population greater than 1000 increases without bound.

These graphical features can be easily checked by taking the limit as $t \to \infty$ in the formula for the flow. In both examples the quotient $A/r = -1000$, so

$$F^t(x_0) = -\frac{A}{r} + \left(x_0 + \frac{A}{r}\right) e^{rt} = 1000 + (x_0 - 1000)e^{rt}$$

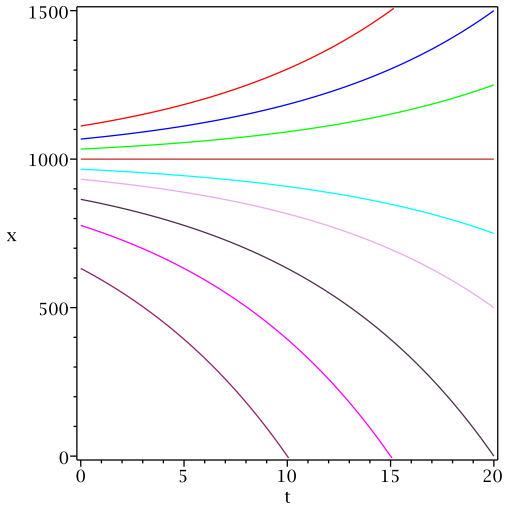


Figure 2.3: $x' = rx + A,\ r = 0.1,\ A = -100$

In Figure 2.2 have $r < 0$ so $e^{rt} \to 0$, and, no matter what the initial condition x_0, we obtain $\lim_{t\to\infty} F^t(x_0) = 1000$. In Figure 2.3 we have $r > 0$ so $e^{rt} \to +\infty$. If $x_0 > 1000$ then $x_0 - 1000$ is positive, so $(x_0 - 1000)e^{rt} \to +\infty$. If $x_0 < 1000$ then $x_0 - 1000$ is negative, so $(x_0 - 1000)e^{rt} \to -\infty$. Since a population cannot be negative we see that, if $0 < x_0 < 1000$, then at some finite time t_* we must have $x(t_*) = 0$. In fact, using logarithms we can solve

$$x(t_*) = 1000 + (x_0 - 1000)e^{0.1\cdot t_*} = 0$$

to get $t_* = 10 \ln\left(\dfrac{1000}{1000 - x_0}\right)$.

Our second variation on the basic population model is a straightforward reinterpretation of Example 1.4 as a continuous system. The idea there was that the growth rate should decrease with increasing population. The simplest version of such a variable growth rate is $r - \alpha x$, and if we consider this growth rate to be applied at a steady rate throughout the year we obtain

$$\Delta x = (r - \alpha x)x\Delta t$$

for the approximate population change over a small time interval Δt. Dividing by Δt and taking the limit as $\Delta t \to 0$ leads to the following:

EXAMPLE 2.5. $\dfrac{dx}{dt} = x(r - \alpha x)$, so $f(x) = x(x - \alpha x)$. The parameters r and α are both positive; α should be small compared with r.

This differential equation is known as the *logistic equation.*

We now follow our standard procedure to determine the flow $F^t(x_0)$ of the logistic equation.

(1) $dx = x(r - \alpha x)dt$.

(2) $\dfrac{dx}{x(r - \alpha x)} = dt$

(3) Using partial fractions (look it up in your calculus book):

$$\begin{aligned}\int \frac{dx}{x(r-\alpha x)} &= \int \left(\frac{\frac{1}{r}}{x} + \frac{\frac{\alpha}{r}}{r - \alpha x} \right) dx \\ &= \frac{1}{r} \ln|x| - \frac{\alpha}{r} \frac{1}{(-\alpha)} \ln|r - \alpha x| \\ &= \frac{1}{r} \left(\ln|x| - \ln|r - \alpha x| \right) \\ &= \int dt = t + C_1\end{aligned}$$

(4) Solve for x, using "difference of logs = log of quotient":

$$\begin{aligned}\frac{1}{r} \left(\ln|x| - \ln|r - \alpha x| \right) &= t + C_1 \\ \ln|x| - \ln|r - \alpha x| &= rt + rC_1 = rt + C_2 \\ \ln \frac{|x|}{|r - \alpha x|} &= rt + C_2 \\ \frac{|x|}{|r - \alpha x|} &= e^{rt + C_2} = C_3 e^{rt} \\ (*) \qquad \frac{x}{r - \alpha x} &= \pm C_3 e^{rt} = C e^{rt} \\ x &= (r - \alpha x) \cdot C e^{rt} = rCe^{rt} - \alpha x C e^{rt} \\ x + \alpha x C e^{rt} &= rCe^{rt} \\ x(1 + \alpha C e^{rt}) &= rCe^{rt} \\ x(t) &= \frac{rCe^{rt}}{1 + \alpha C e^{rt}}\end{aligned}$$

(5) To determine C in terms of x_0 just plug $t = 0$ into equation $(*)$: $C = \dfrac{x_0}{r - \alpha x_0}$.

(6) Plug this into the formula for $x(t)$ and simplify:

$$x(t) = \frac{r\frac{x_0}{r-\alpha x_0}e^{rt}}{1+\alpha\frac{x_0}{r-\alpha x_0}e^{rt}} = \frac{rx_0e^{rt}}{r-\alpha x_0+\alpha x_0e^{rt}}$$

This is $F^t(x_0)$. We can get an alternate form by dividing top and bottom by e^{rt}, so

$$F^t(x_0) = \frac{rx_0e^{rt}}{r-\alpha x_0+\alpha x_0e^{rt}} \quad \text{or} \quad \frac{rx_0}{(r-\alpha x_0)\,e^{-rt}+\alpha x_0}. \tag{2.1}$$

(7) Check your work. It is not easy to plug $x(t)$ into the differential equation and verify that it works; you might want to use a symbolic math package like Maple or Mathematica to do this. On the other hand, during the derivation we divided by both x and $r-\alpha x$, so it is a good idea to check that the constant solutions $x = 0$ and $x = \dfrac{r}{\alpha}$ both work.

Examining the solution shows that there are two steady state solutions, $x = 0$ and $x = \dfrac{r}{\alpha}$. Figure 2.4 is a sample plot of solution curves.

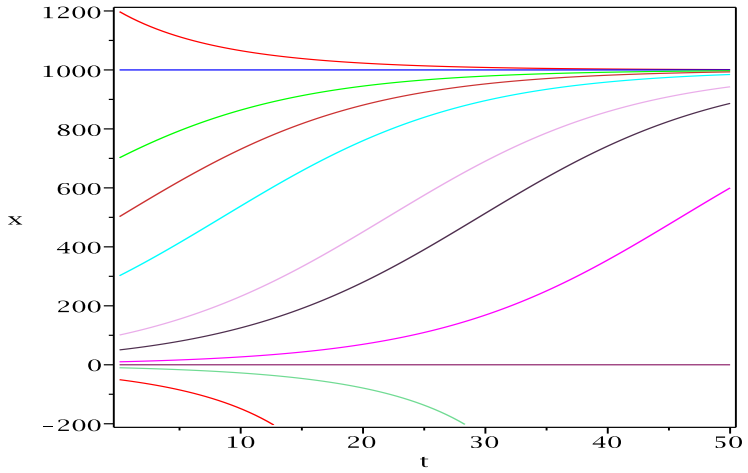


Figure 2.4: $x' = x(r-\alpha x),\ r = 0.1,\ \alpha = 0.0001$

From the plot it seems that all positive initial populations eventually converge to the steady state solution $x = \dfrac{r}{\alpha}$. This is correct, and can be verified by taking the limit of the flow (2.1) as $t \to \infty$: Since $r > 0$ we have $e^{-rt} \to 0$ as $t \to \infty$ so, using

the second form in (2.1), we have

$$\lim_{t\to\infty} F^t(x_0) = \lim_{t\to\infty} \frac{rx_0}{(r-\alpha x_0)\,e^{-rt} + \alpha x_0} = \frac{rx_0}{0+\alpha x_0} = \frac{r}{\alpha}$$

This doesn't work if $x_0 = 0$ because of division by zero. The limit when $x_0 = 0$ is zero, since $x = 0$ is a constant solution.

In the logistic population model the limiting state, $\dfrac{r}{\alpha}$, is called the *carrying capacity* of the system. This is a positive stable population, at which the net growth rate is zero, and any other initial population will decrease (if $x_0 > r/\alpha$) or increase (if $0 < x_0 < r/\alpha$) towards the carrying capacity. This is one explanation for how a population stabilizes over time without external intervention.

There is one strange feature of this limit analysis. This concerns negative initial states x_0. We don't need to worry about these if we are concerned only with this system as a population model, but similar mathematical issues occur in other applications.

Here's the issue: The limit calculations above seem to work just as well if $x_0 < 0$. However, it seems from the plot that the solution $x(t)$ for negative x_0 always decreases, so it stays negative. In fact, it is impossible for such a solution to converge to the carrying capacity, which is possible, for then it would have to cross over the T-axis, and this would violate the basic premise that the system is deterministic, so two different solution curves can't cross.

Here's the resolution: If $x_0 < 0$ then the solution $x(t)$ is *not* defined for all positive values of t. If x_0 is negative then the denominator of the fraction in equation (2.1) becomes zero at some finite value of t, say when $t = t_*$. This means that the solution *only exists for* $0 \le t < t_*$, so it makes no sense to take the limit as $t \to \infty$. In effect the graph of such a function has a vertical asymptote; $x(t) \to -\infty$ on the left of the asymptote and $x(t) \to +\infty$ on the right of the asymptote. The curve to the right of the asymptote converges to the carrying capacity, but this part of the curve is not part of the solution of the differential equation with initial value x_0, since a solution must be defined and satisfy the differential equation on an entire interval starting at $t = 0$.

This value t_* can be determined by setting the denominator to 0 and solving for t_0, using logarithms. The result is $t_* = \dfrac{1}{r}\ln\left(\dfrac{\alpha x_0 - r}{\alpha x_0}\right)$. (This is defined since x_0 is negative, so the numerator and denominator of the fraction are both negative, so the fraction is positive.) Note that the "lifetime" of the solution depends on the value of x_0; values closer to 0 will be defined for a longer interval of t values, but eventually every solution starting at a negative value blows up.

2.4. Steady states and limit states

Just as in the discrete case we define a steady state to be a solution $x(t)$ that is constant as a function of t. We saw several example of steady states in the population examples.

Steady states (also called *equilibrium states*) are usually the first solutions that we should consider when studying a dynamical system. We can find them even without solving for the flow! To do this, just notice that a solution of the differential equation $\frac{dx}{dt} = f(x)$ is a steady state solution if it is constant as a function of time, and that just means that $\frac{dx}{dt} = 0$. Comparing these two conditions, we see that a steady state solution $x(t) = x_0$ of the differential equation is determined by the solution $x = x_0$ of the equation $f(x) = 0$. This equation just refers to x as a variable, ont as a function of t, so solving $f(x) = 0$ is usually just an exercise in algebra.

For example, the steady states of the logistic equation $x' = x(r-\alpha x)$ are obtained by solving the algebraic equation $x(r - \alpha x) = 0$ for x, and we immediately obtain $x = 0$ or $x = \frac{r}{\alpha}$.

We can also adapt the definition of limit state from the discrete case. A state x_1 is a *limit state* of the dynamical system if there is an initial condition x_0, not equatl to x_1, so that the solution $F^t(x_0)$ converges to x_1 as $t \to \infty$. As in the discrete case we have a relation between limit and steady states:

PROPOSITION 2.6. *Any limit state is a steady state. (This requires the hypotheses of Theorem 2.9.)*

However, we will require more details about the solutions of differential equations before we can prove this.

In one dimensional dynamical systems we can actually determine the limit states without solving the differential equation. As an example, consider again the logistic equation $x' = f(x)$ with $f(x) = x(r - \alpha x)$. We already determined the steady state solutions by solving the equation $f(x) = 0$. Now we need to analyze the sign of $f(x)$ for various ranges of x values. We can do this graphically by plotting $y = f(x)$. This curve will cross the X-axis at the points corresponding to the steady states, and in between these states it will be either positive or negative (presuming that f is a continuous function.)

Figure 2.5 shows $y = f(x)$ with the same parameters as in Figure 2.4, where the carrying capacity is 1000. It is clear form the figure that $f(x) > 0$ if $0 < x < 1000$ and $f(x) < 0$ if $x < 0$ or $x > 1000$.

From this we can produce enough of the information in Figure 2.4 to determine limiting behavior. For example, if $0 < x(t) < 1000$ then $x' = f(x) > 0$ so $x(t)$ is

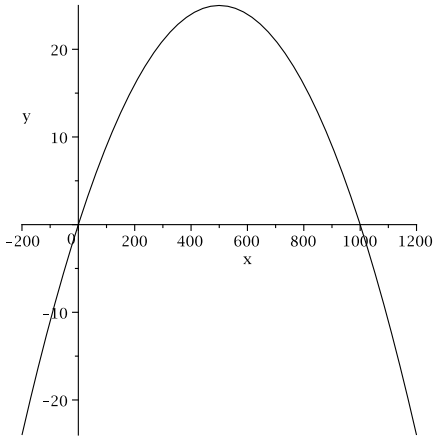


Figure 2.5: $y = f(x) = x(r - \alpha x),\ r = 0.1,\ \alpha = 0.0001$

increasing. so any solution curve that starts with $0 < x_0 < 1000$ will continue to increase as long as it remains less than the steady state solution at $x = 1000$. However, because the dynamical system is deterministic, it is impossible for the solution $x(t)$ to ever cross (or even touch) the steady state solution $x = 1000$. A fundamental fact about the real number system is that any bounded increasing function has a limit, so our solution $x(t)$ must converge to a limit state $x = x_1$, and $x_1 \leq 1000$ because $x(t) < 1000$ for all t. But Proposition 2.6 implies that $x = x_1$ must be a steady state solution. The only two steady state solutions are $x = 0$ and $x = 1000$ and $x_1 \geq x_0 > 0$, so we must have $x_1 = 1000$.

This analysis shows that any solution curve that starts with $0 < x_0 < 1000$ must increase and converge to 1000. A similar analysis shows that any solution curve starting at $x_0 > 1000$ must decrease (because $f(x) < 0$ for $x > 1000$) and must converge to the steady state solution $x = 1000$. We can also conclude that any solution that starts with $x_0 < 0$ must decrease. There is no negative steady state, so such a solution cannot have a finite limit as $t \to \infty$. However, we do not have enough information from this analysis to distinguish between two possibilities: either $x(t)$ is defined for all $t \geq 0$, in which case it will have limit $-\infty$; or it is only defined for a finite range of t values ($0 \leq t < t_*$), in which case it will have a vertical asymptote at $t = t_*$ (this requires some of the theory from the next section).

A more general analysis leads to the same conclusions about the general logistic equation with $f(x) = x(r - \alpha x)$. Solutions with positive initial conditions must converge to the carrying capacity r/α, while solutions with negative initial conditions decrease and do not converge to steady states. Thus the carrying capacity is a limit state, while the other steady state, $x = 0$, is not a limit state.

We can carry over the definitions of sinks and sources from Chapter 1. A steady state solution x_0 is a *sink* if every solution $x(t)$ which starts "near" x_0 must stay near x_0 and converge to x_0 as $t \to \infty$. A steady state solution x_0 is a *source* if every solution which starts "near", but not at, x_0 eventually moves a fixed distance away from x_0. If we look at Figure 2.4 then it is clear that 1000 is a sink and 0 is a source.

We can also translate the criteria for sources and sinks given by Theorem 1.7. First, in the case of continuous dynamical systems we need a different definition for hyperbolicity. If x_0 is a steady state solution of the differential equation $\dfrac{dx}{dt} = f(x)$ then we say x_0 is *hyperbolic* if $f'(x_0) \neq 0$. Here's the analog of Theorem 1.7:

THEOREM 2.7. *Suppose that x_* is a hyperbolic steady state solution of $\dfrac{dx}{dt} = f(x)$. Then:*

(1) *x_* is a sink if $f'(x_*) < 0$.*
(2) *x_* is a source if $f'(x_*) > 0$.*

If x_* is not hyperbolic then it may be a sink or a source or neither. (This requires the hypotheses of Theorem 2.9.)

For example, consider the logistic function, $f(x) = rx - \alpha x^2$, with r and α positive. Then $f'(x) = r - 2\alpha x$. We have already checked that f has fixed points at 0 and r/α. When $x = 0$ we have a source, since $f'(0) = r > 0$, and when $x = r/\alpha$ we have a sink, since $f'(x) = r - 2\alpha \cdot \dfrac{r}{\alpha} = r - 2r = -r < 0$.

We'll give a different argument in section 2.5, but here is an alternate proof, which can be understood as a commentary on Figure 2.4:

PROOF. Suppose x_* is a steady state solution x_* and $f'(x_*) < 0$. Then f is *decreasing* at the point x_*. Hence there are constants $c < x_*$ and $d > x_*$ so that $f(x) > f(x_*) = 0$ if $c < x < x_*$ and $f(x) < f(x_*) = 0$ if $x_* < x < d$. Suppose a solution curve $x(t)$ satisfies $c < x(0) < x_*$. Then $\dfrac{dx}{dt}(0) = f(x(0)) > 0$ so the solution is increasing at $t = 0$. If we follow this solution for $t > 0$ then it must stay between c and x_* unless it hits the solution $x = x_*$ (which can't happen) or it starts decreasing (which can't happen, since the slope is given by $f(x(t))$, which is negative). So $x(t)$ is an increasing bounded function, so it must have a limit. This limit is a steady state, and it cannot be in the interval from c to x_* (since any solution in that interval must be increasing). Hence $x(t) \to x_*$ as $t \to \infty$. The argument is essentially the same if $x(0)$ is between x_* and d.

A similar argument works in case $f(x_*) > 0$. □

2.5. Existence and uniqueness

It is hard to establish general facts about continuous dynamical systems without some very basic results. In fact, the very definition of the flow $F^t(x_0)$ relies on the idea that a differential equation $x' = f(x)$ always has a unique solution $x(t)$ satisfying $x(0) = x_0$, for at least some time interval $[0, T)$ with positive T. That is, it must be possible to follow the solution for a non-empty time interval, for otherwise $x(t)$ is useless for predicting the future; and there can't be two different solutions with the same initial state, for then there is no way to decide which solution to follow.

If the differential equation can be solved explicitly, as we did for examples 2.1, 2.4 and 2.5, then we don't need to worry about existence of a solution. But we do need a general existence criterion to guarantee solutions in case we can't find an explicit solution.

It is harder to establish uniqueness of a solution, since we would need to rule out all other possible solutions. Here is an example where we can find explicit solutions of the differential equation satisfying any initial conditions; but the solutions are not unique:

EXAMPLE 2.8. The differential equation is $\dfrac{dx}{dt} = f(x)$ with $f(x) = 3x^{2/3}$ has solutions satisfying any initial condition, defined for all $t \in \mathbb{R}$. However the solutions are not unique. In fact, for any pair (x_1, t_1) there are infinitely many solutions $x(t)$ satisfying $x(t_1) = x_1$.

To check these properties, we first follow the separation of variables procedure:

(1) $dx = 3t^{2/3}\,dt$.
(2) $\dfrac{dx}{x^{2/3}} = 3dt$.
(3) $\displaystyle\int \frac{dx}{x^{2/3}} = \int x^{-2/3}\,dx = 3x^{1/3} = \int 3dt = 3t + C_1$.
(4) $x^{1/3} = t + C_1/3 = t + C$ so $x(t) = (t + C)^3$.
(5) For $t = 0$, $x(0) = x_0 = (0 + C)^3 = C^3$, so $C = x_0^{1/3}$.
(6) $x(t) = \left(t + x_0^{1/3}\right)^3$.

However, this doesn't allow us to define $F^t(x_0)$ since there are other solutions, so we can't use the rule that $F^t(x_0) = \textit{the}$ solution starting with $x(0) = x_0$, evaluated at time t. In fact, if $x_0 = 0$ then the formula gives $x(t) = t^3$, but there is another solution: $\tilde{x}(t) = 0$ for all t also satisfies the differential equation (just plug it in to the differential equation and check that it works) and satisfies $\tilde{x}(0) = 0$.

So we have two solutions satisfying $x(0) = 0$. To see infinitely many solutions, consider Figure 2.6. The cubics $x = (t+C)^3$ cover all points in the plain (for different values of C) and they are all tangent to the T-axis. We can find more solutions of

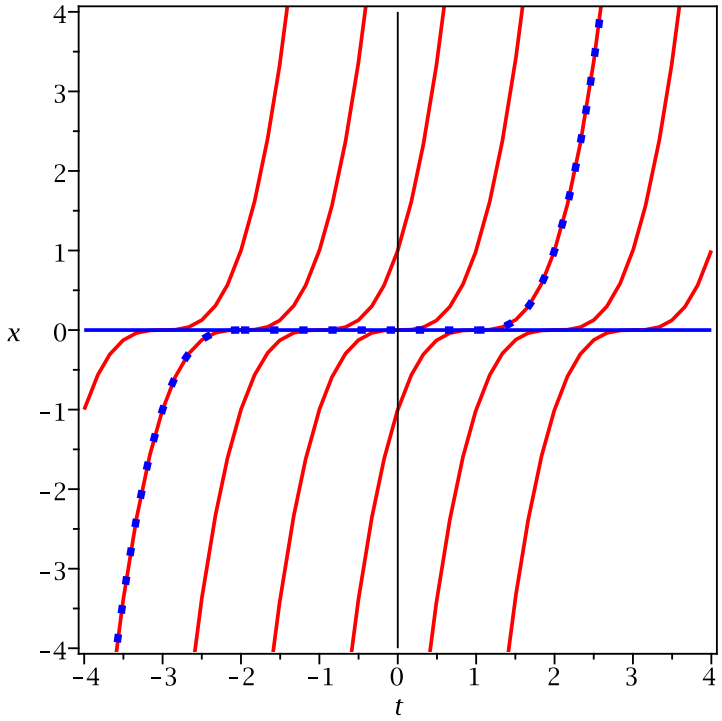


Figure 2.6: Solutions of $x' = 3x^{2/3}$

$x' = x^{2/3}$ with $x(0) = 0$ as follows: Follow a cubic from $t = -\infty$ until it touches the T-axis at a negative t value; then follow the T asis to a point with a positive t value; then leave the T-axis along the cubic which touches the T-axis there and continue on the cubic as $t \to +\infty$. One such solution is indicated on Figure 2.6.

In general, you can find infinitely many solutions through any point (x_1, t_1) by pasting together three partial solution curves: A "half-cubic" through (t_1, x_1), a segment of the T-axis, and another "half-cubic".

You can see from this example that existence of solutions is not enough to guarantee uniqueness. This is reflected in the general theorems regarding existence and uniqueness: The most general existence theorems require much weaker assumptions about the differential equation than the corresponding uniqueness theorems. However, for our purposes, existence is not good for much withot uniqueness, so our basic theorem imposes conditions which imply both existence and uniqueness.

THEOREM 2.9 (Basic Existence and Uniqueness). *Suppose the function f is defined on an open interval I and suppose that its derivative $\dfrac{df}{dx}$ exists and is continuous on I. Suppose that t_1 is fixed. Then*

1. *Existence: For any x_1 in I there is a solution $x(t)$ of $\dfrac{dx}{dt} = f(x),\ x(t_1) = x_1$. This solution is defined for t in some open interval containing t_1. This domain extends indefinitely in either direction, or until $x(t)$ "leaves" the interval (a, b).*
2. *Uniqueness: If $y(t)$ satisfies $\dfrac{dy}{dt} = f(y),\ y(t_1) = x_1$ then $y(t) = x(t)$ as long as $y(t)$ remains in I.*
3. *Continuous dependence on initial conditions: $x(t)$ is a continuous function of the pair (x_1, t).*

There are many variations on this theorem. For example, it is possible to prove the *existence* part of the theorem under the much weaker assumption that f is continuous. However, our formulation covers almost all cases that occur in practice.

Here are some examples:

(1) The logistic equation $x' = x(r - \alpha x)$ with $x(0) = x_0$ has unique solutions for all choices of x_0, since $f(x) = x(r - \alpha x)$ has the continuous derivative $r - 2\alpha x$. The solution is not necessarily defined for all t; see the discussion of Example 2.5.

(2) The equation $x' = x + \sin(x)$ has unique solutions for any initial condition, since $x + \sin(x)$ has the continuous derivative $1 + \cos(x)$. However, we can't find a formula for the solution, since that would require evaluating $\displaystyle\int \frac{dx}{x + \sin(x)}$, and this cannot be expressed in terms of elementary functions.

(3) We know that the equation $x' = 3x^{2/3}$ of Example 2.8 has uniqueness problems. The derivative of $f(x) = 3x^{2/3}$ is $f'(x) = 2x^{-1/3} = \dfrac{2}{x^{1/3}}$, and this is not defined for $x = 0$. Moreover, $f'(x)$ blows up as $x \to 0$, so it is not bounded near 0. However, Theorem 2.9 guarantees unique solutions *if we restrict x to an interval which does not contain* 0. Specifically, there are unique solutions if x is restricted to the interval $(0, \infty)$ (these are the top "half-cubics" in Figure 2.6) or if x is restricted to the interval $(-\infty, 0)$ (the bottom "half-cubics"). In other words, the Theorem only guarantees unique solutions as long as they avoid touching the T axis.

We will not prove Theorem 2.9 in this book; it involves a number of ideas from advanced calculus, plus a large number of iniequalities. However, we want to briefly indicate how it can be used to prove some of the previous results in this chapter.

Existence of $F^t(x_0)$ is now guaranteed: We need to define $F^t(x_0)$ as *the* solution of $\dfrac{dx}{dt} = f(x),\ x(0) = x_0$. We need existence, so there is a solution, and uniqueness, so that the definition of $F^t(x_0)$ is not ambiguous.

Proposition 2.2: If two solution curves touch when $t = t_1$ then uniqueness implies that they are the same for all values of t, as long as they stay in the interval where f' is continuous.

Proposition 2.3: Fix a value for s. Start with the solution $x(t) = F^t(x_0)$, and let $x_1 = F^s(x_0)$. Then $z(t) = F^t(x_1)$ is the solution which starts at x_1 when $t = 0$. We compare this to the function $y(t) = x(t+s)$. This is a solution since (using the chain rule)

$$\frac{dy}{dt}(t) = \frac{d}{dt}x(t+s) = \frac{dx}{dt}(t+s)\frac{d}{dt}\,(t+s) = \frac{dx}{dt}(t+s) = f(x(t+s)) = f(y(t)).$$

Moreover, $y(0) = x(0+s) = x(s) = x_1$. Now $y(t)$ and $z(t)$ are two solutions which satisfy $y(0) = x_1$ and $z(0) = x_1$. By uniqueness, $y(t) = z(t)$ for all t (as long as we stay in the interval I. By the definition of the flow,

$$F^{t+s}(x_0) = x(t+s) = y(t) = z(t) = F^t(x_1) = F^t(F^s(x_0)) = F^t \circ F^s(x_0).$$

Proposition 2.6: This is similar to the proof of Proposition 1.5, but we need Proposition 2.3: If x_* is a limit state in I then $x_* = \lim_{t\to\infty} F^t(x_0)$ for some $x_0 \neq x_*$. Then using continuity of F^s,

$$x_* = \lim_{t\to\infty} F^{s+t}(x_0) = \lim_{t\to\infty} F^s(F^t(x_0)) = F^s\left(\lim_{t\to\infty} F^t(x_0)\right) = F^s(x_*).$$

Since $F^s(x_*) = x_*$ for all values of s, we have a steady state.

We make a detour before looking at Theorem 2.7.

PROPOSITION 2.10. *Let $F^t(z)$ be the flow of $\dfrac{dx}{dt} = f(x)$. If x_0 is a steady state solution then*

$$\frac{\partial}{\partial z}\left(F^t(z)\right)\Big|_{z=x_0} = e^{f'(x_0)t}.$$

(This requires the hypotheses of Theorem 2.9.)

PROOF. In this proof we are conisidering $F^t(z)$ to be a function of two variables, so we write all derivatives as partial derivatives.

Since $F^t(z)$, as a function of t, is a solution of the differential equation we have

$$\frac{\partial}{\partial t}\left(F^t(z)\right) = f\left(F^t(z)\right).$$

Next, differentiate this with respect to z:

$$\frac{\partial}{\partial z}\frac{\partial}{\partial t}\left(F^t(z)\right) = \frac{\partial}{\partial z}\left(f\left(F^t(z)\right)\right).$$

Reverse the order of the partial derivatives on the left and apply the chain rule on the right:

$$(*) \qquad \frac{\partial}{\partial t}\boxed{\frac{\partial}{\partial z}\left(F^t(z)\right)} = f'\left(F^t(z)\right)\cdot\boxed{\frac{\partial}{\partial z}\left(F^t(z)\right)}.$$

Define $X(t)$ to be the boxed quantity above, so $X(t) = \dfrac{\partial}{\partial z}\left(F^t(z)\right)$. We can calculate $X(0)$, since $F^0(z) = z$ for all z, so $X(0) = \dfrac{\partial}{\partial z}\left(F^0(z)\right) = \dfrac{\partial}{\partial z}z = 1$. Now replace the boxed expressions in $(*)$ with $X(t)$ and add this value of $X(0)$ as an initial condition:

$$\frac{d}{dt}X(t) = f'\left(F^t(z)\right)\cdot X(t), \quad X(0) = 1.$$

So we have a differential equation for the unknown function X. We usually can't solve it explicitly, since the right hand side involves the function $F^t(z)$, which is usually too complicated to work with. However, if x_0 is a steady-state solution then we can do something interesting. In this case $F^t(x_0) = x_0$, so $f'\left(F^t(z)\right) = f'(x_0)$ is a constant, say r. Then the differential equation for X becomes the simple equation $X' = rX$, $X(0) = 1$, with the solution $X(t) = e^{rt} = e^{f'(x_0)t}$, which is what we wanted. □

Theorem 2.7: Here is an outline of a different proof. We'll just look at sinks, so suppose that x_* is a steady-state solution and $f'(x_*) < 0$. Let $G = F^1$ and apply Proposition 2.10 with $t = 1$ to get $G'(x_*) = e^{f'(x_*}$. Since $f'(x_*)$ is negative, $0 < G'(x_*) < 1$. By Theorem 1.7, x_* is a sink for G, so $G^n(x) \to x_*$ as $n \to \infty$, as long as x is close to x_*. Now G^n is the same as F^n, so we have $F^t(x) \to x_*$ as $t \to \infty$ if we only want to consider *integer* values of t.

Remark. We still need to cover the general case: $t = n + r$ where n is an integer and r is the fractional part of t, so $0 \le r \le 1$. Then $F^t(x) = F^{n+r}(t) = F^n(F^r(x))$. Since $F^r(x_*) = x_*$ for all r it is reasonable to expect that $F^r(x)$ is close to x_* if $0 \le r \le 1$ and x is even closer to x_*. If this is the case then $F^n(F^r(x))$ will still converge to x_* as $t = n + r \to \infty$. The actual proof that this works requires a number of inequalities plus some ideas from advanced calculus, so we won't give the details.

Exercises

2.1. Find the flow defined by each of the following differential equations:

(a) $\dfrac{dx}{dt} = rx^2$, where r is a non-zero parameter.

(b) $\dfrac{dx}{dt} = -c^2x^2 + A^2$, where A and c are positive parameters.

(c) $\dfrac{dx}{dt} = c^2x^2 + A^2$, where A and c are positive parameters.

(d) $\dfrac{dx}{dt} = \dfrac{r}{x}$, where r is a non-zero parameter.

(e) $\dfrac{dx}{dt} = \dfrac{r}{x} + A$, where A and r are non-zero parameters.

2.2. This problem refers to Exercise 2.1a, so the differential equation is $x' = rx^2$. Assume $r < 0$; you might want to make a substitution like $r = -s$, so that $s > 0$.

(a) Find all steady state solutions.
(b) Verify that the flow $F^t(x_0)$, as found in Exercise 2.1a, decreases as t increases (assuming x_0 is positive).
(c) If x_0 is positive, explain why $x(t)$ never reaches 0.
(d) How long does it take until x has half its initial value? Your answer will involve both x_0 and r.

2.3. Find the flow defined by each of the following defferential equations:

(a) $\dfrac{dx}{dt} = \sec(x)$.

(b) $\dfrac{dx}{dt} = e^{ax}$, where a is a non-zero parameter.

(c) $\dfrac{dx}{dt} = \sqrt{a^2 - x^2}$, where a is a positive parameter and $|x| < a$.

(d) $\dfrac{dx}{dt} = \dfrac{r}{x - a}$ where r and a are non-zero parameters.

2.4. This question concerns the differential equation $\dfrac{dx}{dt} = f(x) = x^2(x^2 - 8x + 12)$. In the following consider all possible initial values $-\infty < x_0 < \infty$. Do not try to solve the differential equation.

(a) Find the steady state solutions.
(b) Determine the intervals on which $f(x)$ is positive or negative.
(c) Sketch a number of solution curves using this information. Indicate the steady state solutions, and indicate the limiting behavior of the other solutions.

(d) Determine $\lim_{t\to\infty} x(t)$ for all possible initial conditions. Your answer will depend on different ranges of the initial conditions.

2.5. Solving a differential equation is based on integration, so it is not surprising that changing variables is a useful technique.

(a) Find a change of variables of the form $y = x + c$, where c is a constant, that transforms the equation of Example 2.4 into the equation of Example 2.1.

(b) Try changes of variable of the form $z = cx$ where c is a non-zero constant on the equation $\dfrac{dx}{dt} = -2x + 4x^2$. You should get an equation of the form $\dfrac{dz}{dt} = Az + Bz^2$.

(1) Can you a choice of c that will make $B = 1$?
(2) Can you find a choice of c that will make $B = 0$?
(3) Can you find a choice of c that will make $A = -4$?

2.6. The logistic equation is often transformed to use the percentage of the carrying capacity y rather than the actual population x as the state variable. (Of course you can't do this until you have first determined that there *is* a carrying capacity.) In other words, $x = Cy$ where $C = \dfrac{r}{\alpha}$ is the carrying capacity. Use this relation between x and y to transform the logistic equation into an equation for y.

2.7. Modify the logistic equation to also incorporate "migration". That is, assume that there is a steady transfer of A individuals per year into or out of the population. (See Example 2.4 for a discussion of migration.) Start with the specific parameters in Figure 2.4 and add migration at the rate of $+20$ or -20 per year (your choice). Find the steady state populations (even if negative) and discuss the limiting behavior. Do not solve the differential equation; however, you will want a calculator to solve for the steady states.

2.8. This question concerns the differential equation $\dfrac{dx}{dt} = f(x) = (x-1)^2(x^2-9)$. In the following consider all possible initial values $-\infty < x_0 < \infty$. Do not try to solve the differential equation.

(a) Find the steady state solutions.
(b) Determine the intervals on which $f(x)$ is positive or negative.
(c) Sketch a number of solution curves using this information. Indicate the steady state solutions, and indicate the limiting behavior of the other solutions.
(d) Determine $\lim_{t\to\infty} x(t)$ for all possible initial conditions. Your answer will depend on different ranges of the initial conditions.

2.9. Radioactive isotopes decay according to the basic law $\dfrac{dx}{dt} = -rx$ where x is the amount of the isotope and r is a constant. Here x is measured in grams and t in years. It is known that the most common isotope of radium has a half life of about 1600 years. This means that if $x(t)$ is the solution of the differential equation with initial condition $x(0) = x_0$ then $x(1600) = \frac{1}{2}x_0$.

(a) Calculate r. [Solve the differential equation, then use logarithms. Give a numeric value for your answer.]
(b) Now suppose that, in an initially empty storage facility, 10 grams of this isotope are deposited each year. Modify the differential equation to model this situation. Explain how you arrived at your modifications.
(c) The amount of radium in the storage facility has a finite limit as $t \to \infty$. Find the limit.
(d) How long does it take until the amount of radium reaches 90% of its limit?

2.10. In each of the following, what restrictions are needed on the domain of x so that any initial value problem with x_0 in this domain will have a unique answer? Explain your answer.

(a) $\dfrac{dx}{dt} = (1 - x^2)^{1/3}$.
(b) $\dfrac{dx}{dt} = (1 - x^2)^{4/3}$.
(c) $\dfrac{dx}{dt} = (1 - x^2)^{-1/3}$.

2.11. This problem refers to Exercise 2.1a, which calculated the flow F^t for $x' = rx^2$. Verify Proposition 2.3 for this flow. That is, calculate $F^{t+s}(x_0)$ and $F^t(F^s(x_0))$ and show that they are equal.

CHAPTER 3

Discrete Linear models

3.1. A stratified population model

A single state variable, as we considered in the first two sections, is usually not enough to describe a system. In this section we will look at some simple examples of multi-dimensional dynamical systems.

Here is a population model that takes into consderation the fact that birth and death rates are connected to age. In this example we consider a plant population with the following characteristics: Plants live at most three years. In their first year they do not produce seeds. THos that survive to the second or third year produce seeds, which sprout in the next spring to become first-year plants. Moreover, the death rate varies from year to year, as does the number of seeds produced per plant. To describe this situation we use three states, x_1, x_2, x_3, to record the number of plants in their first, second or third year. We consider this as a discrete dynamical system, in the sense that we count the plants once per year. Of course, we have to assume that there is some way of identifying a plant as first, second or third year.

Suppose that a fraction, s_1, of the first year plants survive to the second year; as noted above, they do not produce seeds. Suppose that s_2 of the second year plants survive to the third year, and that, on average, each second year plant produces f_2 seeds that will germinate next year to form first year plants. Finally, none of the third year plants survive another year, but, on average, each third year plant produces f_3 seeds.

Suppose we have, as indicated above, a state vector $x = (x_1, x_2, x_3)$ which records the counts of the first, second and third year plants, and suppose the next year's counts are (X_1, X_2, X_3). We can calculate these figures

from the rules above. First year plants next year sprout from the seeds produced by second and third year plants this year, so the number of first year plants will be $X_1 = f_2x_2 + f_3x_3$. Second year plants next year are the first year plants from this year that survive; so there will be $X_2 = s_1x_1$ second year plants. Finally, third year plants next year are the second year plants from this year that survive; so there will be $X_3 = s_2x_2$ third year plants. The transition rule can then be expressed as

$F(x) = X$. In vector terms this is

$$F\left(\begin{bmatrix} x_1 \\ x_2 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} f_2x_2 + f_3x_3 \\ s_1x_1 \\ s_2x_2 \end{bmatrix}.$$

This transition rule is a linear function, so it can be represented as a matrix product:

$$F(x) = Lx, \text{ where } L = \begin{bmatrix} 0 & f_2 & f_3 \\ s_1 & 0 & 0 \\ 0 & s_2 & 0 \end{bmatrix}.$$

In general, the number of age groups may not be three, and we may not be talking about plants; but we can generalize to get the following model for age structured population dynamics:

EXAMPLE 3.1. The state of the system is an m dimensional vector representing populations in m different age groups. For each age group $j < m$ there is a *survival rate* s_j giving the proportion of individuals who survive to the next age group, and for each age group j there is a a *fecundity rate* f_j giving the rate at which new first year individuals are produced. We assume $0 \leq s_j \leq 1$ and $f_j \geq 0$; these are the parameters of the system.

The transition rule is $F(x) = Lx$ where

$$L = \begin{bmatrix} f_1 & f_2 & f_3 & \dots & f_{m-1} & f_m \\ s_1 & 0 & 0 & \dots & 0 & 0 \\ 0 & s_2 & 0 & \dots & 0 & 0 \\ 0 & 0 & s_3 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & s_{m-1} & 0 \end{bmatrix}.$$

This is known as the *Leslie Model*, and L is the *Leslie matrix*.

Before investigating the properties of this system we introduce some terminology and prove a simple lemma that will be useful throughout this section.

A matrix or vector A is *non-negative* if all its entries are non-negative. It is *positive* if all its entries are positive.

LEMMA 3.2. *Suppose A is an $m \times m$ matrix and v is an m-dimensional vector. Then*

(a) *If A is non-negative, $A_{jk} \geq a \geq 0$, and v is non-negative then the j^{th} entry in Ax is $\geq ax_k$.*
(b) *If A and v are non-negative then Av is non-negative.*
(c) *If A is positive and v is non-negative and not zero then Av is positive.*

PROOF. To prove part (a) note that $(Av)_j = \sum_i A_{ji}x_i = A_{jk}x_k + \sum_{i\neq k} A_{ji}v_i$. The last summation is non-negative and the term $A_{jk}v_k$ is $\geq av_k$. Now part (b) follows by selecting $a = 0$. Finally, part (c) follows from part (b) by choosing k so that $v_k \neq 0$, so $v_k > 0$, and choosing a to be the minimum of the entries in A. □

Now, by definition, all Leslie matrices are non-negative, and all state vectors that represent populations are also non-negative. When we iterate the transition rule $F(x) = Lx$ we obtain $F^t(x) = L^t x$. From the lemma we see that the powers L^t are non-negative, as is the future state vectore $L^t x$.

As a concrete example we return to our plant model, and assign some values to the parameters to get the Leslie matrix

$$L = \begin{bmatrix} 0 & 7 & 6 \\ \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix}. \tag{3.1}$$

The first few powers of L are

$$L^1 = \begin{bmatrix} 0 & 7 & 6 \\ 1/4 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix},\ L^2 = \begin{bmatrix} 7/4 & 3 & 0 \\ 0 & 7/4 & 3/2 \\ 1/8 & 0 & 0 \end{bmatrix},\ L^3 = \begin{bmatrix} 3/4 & 49/4 & 21/2 \\ 7/16 & 3/4 & 0 \\ 0 & 7/8 & 3/4 \end{bmatrix},$$
$$L^4 = \begin{bmatrix} 49/16 & 21/2 & 9/2 \\ 3/16 & 49/16 & 21/8 \\ 7/32 & 3/8 & 0 \end{bmatrix},\ L^5 = \begin{bmatrix} 21/8 & 379/16 & 147/8 \\ 49/64 & 21/8 & 9/8 \\ 3/32 & 49/32 & 21/16 \end{bmatrix}.$$

As noted above, each power L^t is non-negative. Moreover, for this example, L^5 is positive. It follows from Lemma 3.2 c applied to each column of L that L^6 is positive, and a simple induction following this pattern shows that L^t is positive for all $t \geq 5$.

Again, by Lemma 3.2 c, $L^t x$ will be positive if $t \geq 5$ and x is any non-negative, non-zero state vector. Hence, even if we start with only first year plants, after at most 5 years we will have plants in all age groups. As an example, suppose the initial state vector is $(100, 0, 0)$. Then

$$x = \begin{bmatrix} 100 \\ 0 \\ 0 \end{bmatrix}, \quad L^5 x = \begin{bmatrix} 262 \\ 77 \\ 9 \end{bmatrix} \text{ (rounded)}.$$

Now notice that the smallest entry on the diagonal in L^5 is $21/16 > 1.3$. If we apply Lemma 3.2 a for a non-negative vector v, with $j = k$, we see that each entry in $L^5 v$ is at lease 1.3 times the corresponding entry in v. Applying this with $v = L^5 x$, we conclude that each entry in $L^5 \cdot L^5 x = L^{10} x$ is at least $1.3 \cdot 9$ (since 9 is the smallest entry in $L^5 x$). If we iterate this argument we see that each entry in $L^{5n} x$ is at least

$(1.3)^{n-1} \cdot 9$. Hence each entry of $A^t x$ is unbounded as $t \to \infty$. (It is true that each entry approaches ∞ as $t \to \infty$. This requires slightly more work, although we will give a different proof later.)

We can see some further details of this behavior if we consider some larger values of t. Further calculation gives

$$y = L^{15}x = \begin{bmatrix} 19625 \\ 3304 \\ 1085 \end{bmatrix}, \quad z = L^{16}x = \begin{bmatrix} 29638 \\ 4906 \\ 1652 \end{bmatrix} \qquad \text{(rounded)}.$$

There are two interesting things to notice here. First,

$$\frac{z_1}{y_1} \approx 1.51, \quad \frac{z_2}{y_2} \approx 1.48, \quad \frac{z_3}{y_3} \approx 1.52$$

Hence z is essentially 1.5 times y. This indicates that the size of each component of the vector $L^t x$ is increasing by a factor of about 1.5 per year. This is much faster than our estimate above, which was 1.3 per 5 years, or $(1.3)^{1/5} \approx 1.05$ per year.

The other thing to notice is that y and z are almost parallel, since z is almost a scalar multiple of y. That is, it appears that y and z are pointing in approximately the same direction. In fact, we can make this explicit by rescaling so that their third components are equal to 1 (there are many other ways to "normalize" the vectors by rescaling). This leads to

$$\frac{1}{y_3} y \approx \begin{bmatrix} 18.09 \\ 3.05 \\ 1.00 \end{bmatrix}, \quad \frac{1}{z_3} z \approx \begin{bmatrix} 17.94 \\ 2.97 \\ 1.00 \end{bmatrix}.$$

So it appears that $L^t x$ is not just growing in magnitude at an exponential rate, but that its direction is converging to the direction of $\begin{bmatrix} 18 \\ 3 \\ 1 \end{bmatrix}$.

We need some more theory to confirm that this is indeed what is happening, and to generalize it.

3.2. Matrix powers, eigenvalues and eigenvectors.

Suppose we have an $m \times m$ matrix M and we need to calculate its powers M^n for $n = 1, 2, \ldots$. This is not generally easy to do, since even for integer matrices there is a lot of variation in the form of the powers. See Exercise 3.4 for some examples.

However, in some cases we can get a very good handle on the powers of a matrix. Suppose the matrix M is *diagonalizable*. This means that there is a basis $\{ v_1, v_2, \ldots, v_m \}$ of $\mathbb{R}^m$ consisting of *eigenvectors* of M; so corresponding to each v_k there is an *eigenvalue* λ_k, satisfying $Mv_k = \lambda_k v_k$. [Even if M is a real matrix the

eigenvalues are, in general, complex numbers; they may not be real. We will discuss later how to deal with complex eigenvalues.]

There are two equivalent ways to see how to use the eigenvalues and eigenvectors of a diagonalizable matrix M to calculate powers of M.

For the first approach, form the matrix P whose columns are the eigenvectors of M; specifically, the k^{th} column of P is v_k. Since the vectors v_k form a basis, the matrix P is invertible. Next, form a diagonal matrix Λ using the eigenvalues as the diagonal entries. It is important to use the eigenvalues in the same order as the eigenvectors, so we should be explicit: the entry of Λ in the k^{th} row and k^{th} column is λ_k, and all other entries of Λ are 0. Then $P^{-1}MP = \Lambda$, or, by solving for M, $M = P\Lambda P^{-1}$. Now we can calculate M^n as

$$\begin{aligned} M^n &= M \cdot M \cdot \ldots M && n \text{ copies of } M \\ &= (P\Lambda P^{-1}) \cdot (P\Lambda P^{-1}) \ldots (P\Lambda P^{-1}) && n \text{ copies of } \Lambda \\ &= P\Lambda(P^{-1}P)\Lambda(P^{-1}P)\ldots(P^{-1}P)\Lambda P^{-1} && \text{associative law} \\ &= P\Lambda \cdot \Lambda \cdot \ldots \Lambda P^{-1} && \text{since } P^{-1}P = I \end{aligned}$$

$$M^n = P\Lambda^n P^{-1}. \tag{3.2}$$

It is easy to calculate Λ^n; it is still a diagonal matrix, with λ_k^n as the k^{th} diagonal entry.

In the second approach we use directly the fact that the eigenvectors form a basis. Hence any vector x in $\mathbb{R}^m$ can be written as a linear combination of the eigenvectors; that is, $x = c_1v_1 + c_2v_2 + \cdots + c_mv_m$. It is not hard to calculate the coefficients c_k: If we form a column vector c with entries given by the coefficients then Pc is the linear combination of the columns v_k of P with the coefficients c_k. Hence we have $Pc = x$ and, given x, we can solve this for c, either by row reduction or by $c = P^{-1}x$. Now we can calculate

$$M^n x = M^n\left(c_1v_1 + c_2v_2 + \cdots + c_mv_m\right) = c_1M^nv_1 + c_2M^nv_2 + \cdots + c_mM^nv_m.$$

But $Mv_k = \lambda_k v_k$, so $M^2v_k = M(Mv_k) = M(\lambda_k v_k) = \lambda_k(Mv_k) = \lambda_k(\lambda_k v_k) = \lambda_k^2 v_k$. Continuing in this way we obtain $M^nv_k = \lambda_k^n v_k$. Plugging this into the equation above produces

$$M^n x = c_1\lambda_1^n v_1 + c_2\lambda_2^n v_2 + \cdots + c_m\lambda_m^n v_m. \tag{3.3}$$

Equations (3.2) and (3.3) greatly simplify analysing iterative matrix multiplication. There are a couple of problems. First, we may need to deal with non-real complex numbers, as noted above. More seriously, there are square matrices that are not diagonalizable. The main criterion for an $m \times m$ matrix to be diagonalizable is that it have m different eigenvalues, for then the corresponding eigenvalues are guaranteed to be linearly independent. If there are repeated eigenvalues then there will

not necessarily be a basis of eigenvectors, and in this case the matrix is not diagonalizable. This occurs rarely, especially with matrices that are determined empirically or experimentally. However, when it does occur the analogs of equations (3.2) and (3.3) are more complicated and harder to analyse. We will take the approach in this book that results will be stated, whenever possible, without assuming that the matrices involved are diagonalizable; but any derivations will assume diagonalizable matrices.

Here is a typical result along these lines. An eigenvalue is *simple* if it has multiplicity 1 as a root of the characteristic equation. (All eigenvalues are simple if the matrix is diagonalizable.) If λ_1 is a simple eigenvalue and $|\lambda_1| > |\lambda|$ for all other eigenvalues λ then λ_1 is called the *dominant eigenvalue*. It is possible that a matrix will fail to have a dominant eigenvalue; for example, the eigenvalues of $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ are 1 and -1. Each is simple, but neither is dominant.

After reading section 3.6 you can show that any dominant eigenvalue of a real matrix must be real. (Exercise 3.13.)

THEOREM 3.3. *Suppose the real matrix M has a dominant eigenvalue λ_1, with v_1 a corresponding eigenvector. Then any vector x can be written uniquely in the form $x = c_1 v_1 + w$ where*

$$\lim_{n\to\infty} \frac{1}{\lambda_1^n} M^n w = 0.$$

Moreover, $\lim_{n\to\infty} \frac{1}{\lambda_1^n} M^n x = c_1 v_1$.

Hence we have the following possibilities, for any vector x:

(a) *If $|\lambda_1| < 1$ then $M^n x \to 0$ as $n \to \infty$.*
(b) *If $|\lambda_1| = 1$ and $c_1 \neq 0$ then $M^n x$ is bounded but does not converge to 0.*
(c) *If $|\lambda_1| > 1$ and $c_1 \neq 0$ then $M^n x$ is not bounded, and its norm converges to ∞.*
(d) *If $|\lambda_1| \geq 1$ and $c_1 = 0$ then nothing can be said about the limiting behavior of $M^n x$ without further information about M and x.*

PROOF. This actually follows easily from equation (3.3), in case M is diagonalizable. We define $c_1 v_1$ as in (3.3), and we define $w = c_2 v_2 + c_3 v_3 + \cdots + c_m v_m$. Then

$$\begin{aligned} \frac{1}{\lambda_1^n} M^n w &= \frac{1}{\lambda_1^n} \left(c_2 \lambda_2^n v_2 + c_3 \lambda_3^n v_3 + \cdots + c_m \lambda_m^n v_m \right) \\ &= \left(\frac{\lambda_2}{\lambda_1} \right)^n c_2 v_2 + \left(\frac{\lambda_3}{\lambda_1} \right)^n c_3 v_3 + \cdots + \left(\frac{\lambda_m}{\lambda_1} \right)^n c_m v_m, \end{aligned}$$

and this clearly converges to 0 since each of the quotients in parentheses is less than 1 in absolute value. The limit of $\frac{1}{\lambda_1^n} M^n x$ is calculated in the same way, except that $c_1 v_1$ is present with a multiplier of $\left(\frac{\lambda_1}{\lambda_1}\right)^n = 1$. The remainder of the theorem follows by multiplying this limit by λ_1^n. □

Note that, in case $|\lambda_1| > 1$ and $c_1 \neq 0$, the fact that $\frac{1}{\lambda_1^n} M^n x$ converges to $c_1 v_1$ can be interpreted as saying that the limiting direction of the vectors $M^n x$ is parallel to the eigenvector v_1.

This theorem explains our numerical results in the previous section about powers L^t of the Leslie matrix (3.1) If we calculate the characteristic polynomial $\det L - \lambda I$ we obtain $-\lambda^3 + \frac{7}{4}\lambda + \frac{3}{4}$. This has the three roots $\lambda_1 = \frac{3}{2}$, $\lambda_2 = -\frac{1}{2}$, $\lambda_3 = -1$. These are the eigenvalues of L, and we can calculate the corresponding eigenvectors as

$$v_1 = \begin{bmatrix} 18 \\ 3 \\ 1 \end{bmatrix}, \; v_2 = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}, \; v_3 = \begin{bmatrix} 8 \\ -2 \\ 1 \end{bmatrix}.$$

Therefore the matrix L is diagonalizable, and $\lambda_1 = \frac{3}{2}$ is the dominant eigenvalue. We can express our initial state x in terms of the eigenvectors, as

$$x = \begin{bmatrix} 100 \\ 0 \\ 0 \end{bmatrix} = \frac{5}{2} v_1 - \frac{25}{2} v_2 + 10 v_3,$$

so $c_1 = 5/2$. Hence we see that $L^t x \to \infty$ as $t \to \infty$, at an approximate exponential rate of $\lambda_1 = 1.5$; and the limiting direction of $L^t x$ is in the same direction as v_1. Both of these correspond closely to what we guessed from the numerical evidence.

Notice that (3.3) shows that $M^n x \to 0$ as $n \to \infty$ if all eigenvalues have absolute value less than 1. This suggests that 0 is a sink, according to the definition in Section 1.4. This is the first part of the following.

PROPOSITION 3.4. *For the dynamical system defined by $F(x) = Mx$:*

(a) 0 *is a sink if and only if all eigenvalues are less than one in absolute value.*
(b) 0 *is a source if and only if all eigenvalues are greater than* 1 *in absolute value.*

The fixed point 0 is called *hyperbolic* if $|\lambda| \neq 1$ for all eigenvalues λ. So sinks and sources are examples of hyperbolic fixed points. Hyperbolic fixed points that are not sources or sinks are called *saddles*.

PROOF. We give the proof for a diagonalizable matrix.

If all the eigenvalues satisfy $|\lambda_j| < 1$ then we have already seen that $M^n x \to 0$ as $n \to +\infty$. We need a little more for a sink, and for that we need to look at (3.2) and (3.3). The constants c_j in (3.3) are determined by the equation $c = P^{-1}x$. If A is the maximum absolute value of any entry in P^{-1} then c_j is a sum of m terms, each of which is some entry of P^{-1} times some entry of x. Therefore $|c_j| \leq mA \max\{\,|x_j|\,\} \leq mA\,\|x\|$. Now let B be the maximum norm of the eigenvectors v_j and let α be the maximum of the absolute values of the eigenvalues λ_j. Then

$$\begin{aligned}\|M^n x\| &= \|c_1\lambda_1^n v_1 + \cdots + c_m\lambda_m^n v_m\| \leq |c_1|\,\alpha^n\,\|v_1\| + \cdots + |c_m|\,\alpha^n\,\|v_m\| \\ &\leq m \cdot (mA\,\|x\|) \cdot \alpha^n \cdot B = m^2AB\alpha^n\,\|x\|\,.\end{aligned}$$

Hence if $r > 0$ is specified and s is defined by $m^2AB \cdot s = r$ and $\|x\| \leq s$, we have

$$\|M^n x\| \leq m^2AB\alpha^n\,\|x\| \leq m^2AB\alpha^n \cdot s \leq \alpha^n \cdot r \leq r$$

since $\alpha < 1$. That is, if x is within s of 0 then $M^n x$ is within r of 0 for all $n \geq 0$, and this verifies the extra condition in the definition of a sink.

Next, suppose that M is not invertible. Then, by exercise 3.15, 0 is a an eigenvalue of M. Choose a corresponding eigenvector v. Then, $M^n(av) = 0$ for all $n > 0$ and all constants a. Thus 0 can't be a source, since we can choose a small enough that av is as close to 0 as desired, but $M^n(av)$ does not move away from 0. So if 0 is a source for M then M is invertible.

Now Proposition 1.6 shows that 0 is a source for M if and only if it is a sink for M^{-1}, and the first part of this proof, together with exercise 3.15, shows that 0 is a sink for M^{-1} if and only if $|\lambda^{-1}| < 1$ for all eigenvalues of M. Taking inverses, this translates to $|\lambda| > 1$. □

3.3. Non negative matrices

In the last section we used eigenvalues and eigenvectors to analyse a specific Leslie model, but we need some guidance as to when such an approach will be appropriate. It turns out that a classic theorem on non-negative matrices is just what we need. In this theorem we use the notion of a *primitive* matrix: this is a matrix A which is non-negative and which has some power A^k which is positive.

THEOREM 3.5 (Perron-Frobeniuus). *Suppose that A is a primitive $m \times m$ matrix. Then A has a dominant eigenvalue, λ_1. Moreover,*

(a) $\lambda_1 > 0$.
(b) *There is a positive eigenvector v_1 corresponding to λ_1.*
(c) *For all non-zero non-negative vectors x, $\lim_{n\to\infty} \frac{1}{\lambda_1^n} A^n x = c_1 v_1$ with $c_1 > 0$.*
(d) *No other eigenvector of A is non-negative.*

Sometimes we will refer to λ_1 as the Perron-Frobenius eigenvalue of A, and to a positive eigenvector as the Perron-Frobenius eigenvector.

We will not prove this result now; it requires several concepts that belong to the realms of analysis and topology.

Use of the Perron-Frobenius Theorem simplifies our previous discussion of plant populations using the Leslie matrix (3.1). If we calculate L^5 we see that L satisfies the hypotheses of the Perron-Frobenius Theorem, so we know that it has a dominant positive real eigenvalue. If we calculate the eigenvalues we see that the dominant eigenvalue is $\lambda_1 = 1.5$, with eigenvector $v_1 = \begin{bmatrix} 18 \\ 3 \\ 1 \end{bmatrix}$. According to the Perron-Frobenius Theorem, for any non-negative non-zero initial state x, all components of $L^t x$ approach ∞ at an approximate exponential rate of $3/2$; and the direction of $L^t x$ converges to the direction of the eigenvector v_1.

In fact, this situation applies to any Leslie model: As long as some power of L is positive then all we need to do to predict the limiting behavior of $L^t x$ is to find the Perron-Frobenius eigenvalue and corresponding eigenvector. If $\lambda_1 > 1$ then the analysis is as above; if $\lambda_1 < 1$ any initial population eventually dies out; and if $\lambda_1 = 1$ any non-zero initial population vector will converge to a stable population vector which points in the direction given by v_1.

For a case in which a general Leslie matrix is guaranteed to have a positive power see Exercise 3.5.

3.4. Networks; more examples

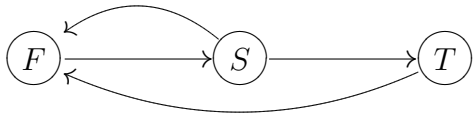
Many dynamical systems models can be described using a type of *network*. There are a number of different formulations of this; we'll use the following.

A *directed graph*, or *digraph*, is a finite collection of objects called *vertices* or *nodes*, together with a collection of ordered pairs of nodes, called *edges*. We say an edge $e = (a, b)$ is the edge *from a to b*; we say the node a is the *tail* or *source* of e and the node b is the *head* or *target* of e. We will usually designate an edge from a to b as $e \to b$ rather than as (a, b). Note that it is possible for an edge to connect a node to itself.

Warning: There are a number of different definitions of digraph available, depending on the context. According to the most prevalent definition, what we are talking about is technically a *pseudograph with no multiple edges*. Use caution when consulting other sources.

We will visualize a digraph as a diagram in which the nodes are represented by numbers or other symbols, and the edges are represented as arrows connecting two

nodes (which may not be different). Here is an example:

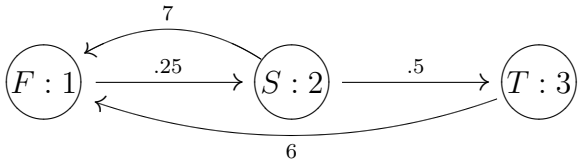


In this digraph there are three nodes, F, S, T, and there are four edges. We will only need digraphs with the property that there can be at most one edge connecting any pair of nodes, so we can identify the edges by their starting and ending nodes. So in this example the nodes are $F \to S$, $S \to T$, $S \to F$ and $T \to F$.

If we number the nodes in a digraph then we can completely specify the digraph by a matrix A in which the ij entry is 1 if there is an edge from i to j, and otherwise is zero. This matrix is called the *adjacency matrix* of the digraph. If we number the nodes in the example above in the order F, S, T then the adjancy matrix is

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

For our uses we need to decorate a digraph with extra information. Specifically, we will attach quantities, called *weights*, to the edges of a digraph; the result is called a *weighted digraph*. The weights can be represented graphically; for example, here is the example above, with weights attached to the edges. We have also added index numbers to the nodes, to make the ordering explicit.



We can modify the adjancy matrix to record the weight information, to construct the *weight matrix* W: we just enter the weight for the edge from i to j in location ij. If there is no edge from i to j we enter 0. If no weights are 0 then we can reconstruct the adjacency matrix from the weight matrix, since in this case $A_{ij} = 1$ if and only if $W_{ij} \neq 0$. Here is the weight matrix for our example:

$$W = \begin{bmatrix} 0 & .25 & 0 \\ 7 & 0 & .5 \\ 6 & 0 & 0 \end{bmatrix}.$$

You have probably already noticed the connection with the stratified population model discussed in section 3.1. For example, we interpret the weighted edge $F \xrightarrow{.25} S$ as representing the fraction of the first year plants (F) that survive be part of the

second year group (S); and the weighted edge $S \xrightarrow{7} F$ represents the number of viable seeds produced by each second year plant.

In general, we can interpret a weighted digraph as specifying the following dynamical system: There is a state vector x, with each entry x_i corresponding to the node in the digraph labelled i. The weight W_{ij} attached to an edge $i \to j$ is interpreted as the fraction of the state variable x_i that is contributed to the next value of the state variable x_j in one time unit. Hence the new value of x_j is the sum of all such contributions, or

$$\text{new } x_j = x_1 \cdot W_{1j} + x_2 \cdot W_{2j} + \cdots + x_m \cdot W_{mj} = \sum_i x_i W_{ij}.$$

In matrix terms, this says that

$$\text{new } \begin{bmatrix} x_1 & x_2 \dots x_m \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \dots x_m \end{bmatrix} \cdot W.$$

However, we are consistently interpreeting state vectors x as *column vectors*, not as *row vectors*. To convert the row vector $\begin{bmatrix} x_1 & x_2 \dots x_m \end{bmatrix}$ into the column vector x we take the transpose, and remember that transposing reverses the order of matrix multiplication, so the new value of x is $\left(\begin{bmatrix} x_1 & x_2 \dots x_m \end{bmatrix} \cdot W\right)^T = W^T \begin{bmatrix} x_1 & x_2 \dots x_m \end{bmatrix}^T = W^T x$. So we have a general procedure for creating a linear dynamical system:

The discrete dynamical system associated to a weighted digraph with vertices $\{ 1, 2, \dots, m \}$ and weight matrix W is

$$F(x) = W^T x \tag{3.4}$$

where x is an m-dimensional state vector.

This is exactly the procedure that leads to the Leslie matrices discussed in sections 3.1, 3.2 and 3.3, and it is easily adapted to similar problems in population dynamics and other studies. See Exercise 3.9 for an ecology problem based on a network.

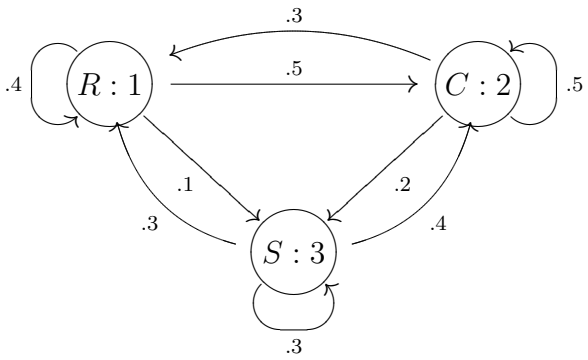
Here is a very different kind of application of digraph methods. We shall consider the weather in Binghamton. We suppose that every day can be classified into exactly one of the following categories:

R: Rainy (or snowy).
C: Cloudy.
S: Sunny.

Historically we know that each day's weather is closely related to the next: For example, if today is rainy then 40% of the time it will be rainy tomorrow; 50% of the time it will be cloudy tomorrow; and so on. There are then nine possible transitions from the weather on one day to the weather on the next day, each with an associated

probability; we can summarize this via a digraph:

(3.5)



(The numbers are just guesses; a study of several years of weather data would lead to more accurate probabilities.)

The corresponding weight matrix is

$$W = \begin{bmatrix} .4 & .5 & .1 \\ .3 & .5 & .2 \\ .3 & .4 & .3 \end{bmatrix}. \tag{3.6}$$

This matrix has two important properties, which constitute the definition of a *stochastic matrix* (more precisely, a *right stocahastic matrix*): W is non-negative, and the entries in each row add up to 1. You can check this in the specific example above, but it is much more generally true: The entries in W must be non-negative since they are probabilities, which cannot be negative; and the entries in any row must add to 1 because these entries give the probabilities for the weather on the next day, and the next day's weather must fit into exactly one of our categories.

The corresponding dynamical system has the form $F(x) = Mx$ where $M = W^T$. We interpret the state vector x as a vector of probabilities, so the j^{th} component of $F^t(x)$ is the probability that the weather will be in category j (R, C or S) after t days. The initial condition x is the actual weather at time $t = 0$; since today is sunny I'll take the initial value of x to be $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. Here are the first few values of $F^t(x)$, rounded:

$$F^1(x) = \begin{bmatrix} 0.300 \\ 0.400 \\ 0.300 \end{bmatrix},\ F^2(x) = \begin{bmatrix} 0.330 \\ 0.470 \\ 0.200 \end{bmatrix},\ F^3(x) = \begin{bmatrix} 0.333 \\ 0.480 \\ 0.187 \end{bmatrix},\ F^4(x) = \begin{bmatrix} 0.333 \\ 0.481 \\ 0.185 \end{bmatrix}$$

Here are two things to notice.

First, the vectors $F^t(x)$ are always *probability vectors*. This means that they are non-negative vectors and the sum of their entries is 1. In this example, the entries

of $F^t(x)$ give the predicted probabilities that the weather will be rainy, cloudy, or sunny at a date t days in the future.

Second, the vectors $F^t(x)$ seem to be converging to a fixed vector, approximately $\begin{bmatrix}0.333\\0.481\\0.185\end{bmatrix}$. This is the limiting probability vector for our system, and we interpret its entries as the probabilities that the weather will be rainy, cloudy, or sunny at any date sufficiently far in the future.

Both of these observations are fairly general properties of stochastic matrices. These, and a number of other basic properties are summarized below.

THEOREM 3.6. *Let u be the m-dimensional non-negative m-dimensional column vector with all entries equal to 1. Then*

(a) *An $m \times m$ matrix W is stochastic if and only if W is non-negative and $Wu = u$.*
(b) *An m-dimensional column vector x is a probability vector if and only if x is non-negative and $u^T v = 1$.*

Now suppose that W is stochastic and let $M = W^T$.

(c) *If k is a non-negative integer then W^k is stochastic.*
(d) *If k is a non-negative integer and x is a probability vector then $M^k x$ is a probability vector.*
(e) *M and W have the same characteristic polynomial, so they have the same eigenvalues.*
(f) *If v is an eigenvector of M corresponding to the eigenvalue λ and $\lambda \neq 1$ then $u^T v = 0$.*
(g) *1 is an eigenvalue of M and all eigenvalues λ of M satisfy $|\lambda| \leq 1$.*

Further, suppose that W is a primitive stochastic matrix. Then

(h) *1 is the dominant eigenvalue of M, and there is a unique positive probability vector p so that $Mp = p$.*
(i) *If x is any probability vector then $M^k x$ converges to p.*
(j) *M^k converges to the matrix in which all columns equal p.*

PROOF. First we need a calculation. If R is an m-dimensional row vector then Ru is a scalar; it is the sum of the entries of R multiplied by 1, since all the entries of u are equal to 1. This is the sum of the entries in R. The same argument applies to a product $u^T C$ where C is a column vector. That is,

$$Ru = \text{ the sum of the entries in } R \qquad \text{if } R \text{ is a row vector} \tag{3.7}$$

$$u^T C = \text{ the sum of the entries in } C \qquad \text{if } C \text{ is a column vector} \tag{3.8}$$

Now here are the arguments for the parts of Theorem 3.6:

(a) This follows from (3.7) applied to each row of W.
(b) This follows from (3.8).
(c) $W^k u = W^{k-1} W u = W^{k-1} u$, using (a). Now repeat the argument $k-1$ more times.
(d) $u^T M^k x = \left(W^k u\right)^T x = u^T x = 1$, using the fact that the transpose of a product is the product of the transposes in the opposite order.
(e) The determinant does not change if you take the transpose, so

$$\det(W - \lambda I) = \det\left((W - \lambda I)^T\right) = \det\left(W^T - \lambda I^T\right) = \det(M - \lambda I).$$

(f) Suppose $Mv = \lambda v$. Then $u^T M v = u^T(\lambda v) = \lambda u^T v$, and $u^T M v = (Wu)^T v = u^T v$. Put these together to get $u^T v = \lambda u^T v$, and then subtract the right hand side from both sides and factor: $(1 - \lambda) u^T v = 0$. Now if $\lambda \neq 1$ the expression in parentheses is not zero, so we can cancel to get $u^T v = 0$.
(g) 1 is an eigenvalue of W with eigenvector u, since $Wu = 1 \cdot u$. Hence 1 is also an eigenvalue of M (but probably with a different eigenvector).

For the second part of (g), suppose that λ is a *real* eigenvalue of M and $|\lambda| > 1$. Pick any positive probability vector w. For a sufficiently small positive number a the vector $z = w + av$ is still positive, and $u^T z = u^T w + a u^T v = 1 + a \cdot 0 = 1$, so z is a probability vector. Now consider $M^k z = M^k w + a\lambda^k v$. This is impossible, since both $M^k z$ and $M^k w$ are probability vectors, so all their entries are between 0 and 1, while the norm of $a\lambda^k v$ is $a\,|\lambda|^k\,\|v\|$, which approaches ∞.

The case of non-real eigenvectors is postponed until section 3.6.

(h) The fact that 1 is a dominant eigenvalue follows from primitivity, part (g), and the Perron-Frobenius Theorem, which also provides a positive eigenvector v corresponding to the eigenvalue 1. Since v is positive the sum of its entries, $u^T v$, is positive. Then define p by normalizing v: $p = \dfrac{v}{u^T v}$.
(i) This follows from the Perron-Frobenius Theorem and part (d).
(j) This is just part (i) applied to the columns of M.

□

A dynamical system of the form $F^t(x) = M^t x$ where x is a probability vector and M is the transpose of a stochastic matrix is called a *Markov chain*.

As an application, let us return to the Binghamton weather example, (3.5), corresponding to the stochastic matrix (3.6). We know that 1 is a simple eigenvalue of $M = W^T$, so we can find the limit probability vector p as follows:

(1) Find a basis for the one-dimensional null space of $M - I$; in other words, find a non-zero solution v of $(M - I)v = 0$. Via row reduction we get $v = \begin{bmatrix} 1.8 \\ 2.6 \\ 1 \end{bmatrix}$.

(2) Divide v by $u^T v = 5.4$, the sum of the entries in v, to get $p = \begin{bmatrix} 0.3333 \\ 0.4815 \\ 0.1852 \end{bmatrix}$ (rounded).

We can interpret this as saying that eventually the probabilities of rainy, cloudy and sunny days will stabilize at 33%, 48% and 19%, as we saw earlier by calculating $M^k x$ until it seemed to stabilize.

You should be aware that there is an alternative to constructing the dynamical system in terms of the transpose of the weight matrix W. In many applications you will find that the weight matrix is not transposed, but that the dynamical system is expressed in terms of the transition rule $F(y) = yW$. In this case, y is the state vector expressed as a row vector; it is the same as the traspose of our x. In this case we iterate F to get $F^t(y) = yW^t$. All the techniques of this chapter can be modified to handle this formulation. For example, the eigenvalue and eigenvector analysis can be adapted, but we need to talk about *left eigenvalues* and *left eigenvectors*.

This alternative approach is usually used for Markov chains, but our original approach is generally used for Leslie matrices. There is little uniformity in other types of models based on matrix powers.

3.5. Google PageRank

In this section we describe the dynamical system that was the major insight in developing Google, and is still the foundation of Google's search technology. This leads to the assignment of a number, the *PageRank*, to each web page in Google's database. This is used to determine part of the presentation order of search results, and is also used internally by Google for various purposes.

The scheme for computing PageRank was discovered by Google's founders, Sergey Brin and Larry Page, while they were graduate students at Stanford. The original publication, written with two Stanford faculty members, Rajeev Motwani and Terry Winograd is [3]. This paper has inspired hundreds of research papers; we will use the basic matrix setup from one of these, [2].

We start with a digraph. The nodes in the network are the pages that Google has found on the web (there are about 10 billion of these in 2009, and the number is growing). If P and Q are two of these pages and there is a link in P that refers to Q then there is an edge in the graph from P to Q. We do not add an edge from a page to itself if the page refers to itself. As far as Google is concerned, this *is* the web, and a surfer browsing the web is moving from page to page along these links.

Now we add weights to the edges. For each page P we count the edges that leave P, and call the total n_P. If $n_P > 0$ then we declare $\frac{1}{n_P}$ to be the weight on each edge leaving P. Let H be the corresponding weight matrix.

Notice that each row of H corresponds to a page P. It is either all zero (if there are no edges leaving P) or the entries are non-negative and sum to 1 (since each non-zero weight in the row is $1/n_P$ and there are n_P such edges. We want to base a Markov chain on H. In effect, we are saying that a surfer looking at page P will next proceed to another page which has a link on the page P, and that s/he will simply pick one of the n_P links at random.

Of course, there are many pages that do not link to any other page; these pages are called *dangling pages*, and the rows corresponding to dangling nodes in the matrix H are entirely 0. But a surfer looking as such a page does not quit; s/he simply types in some other url. So we modify the zero rows of H to take this into account. There are a number of choices here (for example, we could assume that any such surfer will simply jump to a search page or some portal). The most popular choice seems to be to just say that the surfer will jump to some other page on the web, completely at radom. In other words, if P is a dangling node then we replace its row with a row in which every entry equals $\frac{1}{N}$, where N is the total number of pages in the network. You can think of this as a modification of the original digraph: find each dangling node P and add an edge from P to every other node in the graph. (This includes an edge from P to P, because it is simpler to include it than to avoid adding it.)

This modification produces a matrix S, defined by

$$S = H + \frac{1}{N} z u^T.$$

In this formula

(1) u, as in section 3.3, is the N dimensional column vector with all entries equal to 1.
(2) z is the N dimensional column vector defined by the rule $z_P = 1$ if P is a dangling page, and $z_P = 0$ otherwise.

Now S is a stochastic matrix. If we form the dynamical system $F(x) = S^T x$ then we know that 1 is an eigenvalue. However, S^T is not primitive, so there is no reason to expect that 1 is a dominant eigenvalue. In fact, 1 has high multiplicity as an eigenvalue.

There are other problems. For example, there are many situations in which a list of web pages forms a cycle. That is, there are links like $P_1 \to P_2 \to \cdots \to P_k \to P_1$. Moreover, it can happen that none of the pages in this cycle ever contains a link to a page that is not in the cycle. In this case a surfer who enters the cycle can never

leave it. This is unrealistic; a surfer can always type in a url to jump to somewhere completely different.

So here is another modification. We propose that a surfer, randomly, will decide to jump somewhere else. We write $1-\alpha$ for the probability that s/he jumps away, so α is the probability of following one of the links on the current page. To model "jumping somewhere else" we set up a stochastic column vector v, so that v_Q represents the probability of jumping to page Q.

The vector v is called the *personalization vector*, and there are several ways to define it. One way is just to set each entry in v to $\frac{1}{N}$ (N is the total number of pages.) This corresponds to the surfer making a completely random jump. Another possibility is to pick a small number of "major sites" – search engines, portals, wiki, etc – and apportion the values of v among them. The idea is that the surfer will occasionally restart at such a site.

With this last modification we are led to the matrix

$$G = \alpha S + (1-\alpha)uv^T.$$

The effect of the formula is to replace each row of S with a weighted average of that row and the row given by transposing v. A typical value for the weight α is .85.

G is the Google matrix. It turns out that 1 is the dominant eigenvalue for G. The eigenvector corresponding to 1, normalized to be a probability vector, is the PageRank vector π. Thus the probability that a random surfer is looking at page $\#i$ is π_i.

Here is its eigenvalue structure of G.

PROPOSITION 3.7. *G is a stochastic matrix. It is not necessarily primitive, but 1 is a dominant eigenvalue of G^T and all other eigenvalues of G^T satisfy $|\lambda| \leq \alpha$.*

PROOF. First, $v^T u = 1$ by equation (3.7). Hence, remembering that $Su = u$ since S is stochastic,

$$Gu = \alpha Su + (1-\alpha)uv^T u = \alpha u + (1-\alpha)u\left(v^T u\right) = \alpha u + (1-\alpha)u \cdot 1 = u,$$

so G is stochastic.

It is enough to prove the eigenvalue results for G, since G^T has the same characteristic equation as G. Since G is stochastic, u is an eigenvector of G corresponding to the eigenvalue 1.

We now need to consider the other eigenvectors of G. Suppose that x is an eigenvector which is not a scalar multiple of u, corresponding to the eigenvalue λ. We shall show that $\lambda = \alpha\mu$ where μ is an eigenvalue of S. This shows that $|\lambda| \leq \alpha$, since $|\mu| \leq 1$.

The condition that x is an eigenvector is

$$\lambda x = Gx = \alpha Sx + (1-\alpha)uv^T x = \alpha Sx + (1-\alpha)u(v^T x) = \alpha Sx + ku,$$

where k is the scalar factor $(1-\alpha)(v^T x)$. This gives $\lambda x = \alpha S x + ku$, and we can solve this to get

$$(*) \qquad \alpha S x = \lambda x - ku.$$

Now define $y = x + tu$ where t is some scalar. This is non-zero, since x is not a scalar multiple of u. First, calculate $\alpha S y$, using $(*)$:

$$\alpha S y = \alpha S x + t\alpha S u = \lambda x - ku + t\alpha u = \lambda x + (t\alpha - k)u.$$

Now choose t so that the coefficient of u is λt. That is, solve the equation $t\alpha - k = \lambda t$ for t, to get $t = \dfrac{k}{\alpha - \lambda}$. With this choice of t the calculation above becomes $\alpha S y = \lambda x + \lambda t u = \lambda y$. Dividing by α produces $Sy = \mu y$ where $\mu = \dfrac{\lambda}{\alpha}$. Therefore y is an eigenvalue of S with eigenvalue μ, and $\lambda = \alpha\mu$.

This argument does not work if $\lambda = \alpha$, because of the division by $\alpha - \lambda$ in the definition of t. But 1 is an eigenvalue of S, so if $\lambda = \alpha$ then $\lambda = \alpha\mu$ with $\mu = 1$.

This finishes the proof of Proposition 3.7, except for checking that 1 is a simple eigenvector of G.

Checking that an eigenvalue is a simple eigenvalue is discussed in advanced linear algebra courses. We have shown that there are no eigenvectors corresponding to 1 which are not linear multiples of u, since the eigenvalue for such an eigenvector cannot be larger than α in absolute value. However, it is also necessary to rule out the possibility of a *generalized eigenvector* x corresponding to the eigenvalue 1. This is a vector x which satisfies $Gx = x + cu$, where c is a scalar and x is not a scalar multiple of u. But the proof above, starting with $Gx = x + cu$ instead of $Gx = x$, still works to eliminate this possibility; the only difference is that the formula for t now involves c. □

Armed with this eigenvalue structure, we can follow a simple procedure for determining the pagerank vector: Row reduce the matrix $G^T - I$. This will leave one free variable; set it equal to 1 and solve for the other components of the solution. Now normalize to get a probability vector.

However, this is totally unrealistic. Currently, in 2009, N (the number of components) is on the order of 10^{10}. Hence the Google matrix has about 10^{20} entries. To put this in context, if one terabyte drives (10^{12} bytes) are used and only one byte is needed per entry, then the total number of drives needed just to store G is 10^8, or 100 million. Moreover, row reducing G using standard algorithms takes on the order of $N^3 = 10^{30}$ floating point operations. Very fast supercomputers, based on large arrays of fast processors working in parallel, can currently operate at about 1 petaflops, or 10^{15} floating point operations per second. Thus the time needed for

10^{30} floating point operations is about 10^{15} seconds, which is more than ten million years.

Fortunately, there is another method for finding the PageRank vector π. Just start with any probability vector, say x, and calculate $x_k = (G^T)^k x$; the sequence x_k will converge to π, and the convergence rate will be controlled by the second largest eigenvalue, which is at most α. Some analysis shows that it is not necessary to calculate too mamy values of x_k – a few hundred will suffice for reasonable accuracy. Of course we don't want to actually calculate $(G^T)^k$; we just calculate the vectors iteratively by $x_{n+1} = G^T x_n$. Now we can exploit the structure of G. First, there are not many links per page; on average, less than 10. Thus the original matrix H is very *sparse*; it has, on average, no more than 10 links in each row or column. So we can store H very efficiently, by just storing the non-zero entries, and it becomes easy to calculate $H^T x$, since only the non-zero entries in each row of H^T are used. Hence the number of floating point operations needed to caluclate $H^T x$ is on the order of $10N$, or about 10^{11}. Next, consider $S^T x = H^T x + \frac{1}{N} u z^T x$. The factor $\frac{1}{N} z^T x$ is a scalar (the sum of the entries in x corresponding to zero rows in H, divided by N), so the addition involves adding this scalar to each of the entries in $H^T x$. This involves at most N floating point operations. Similarly, $u^T x = 1$, so $G^T x = \alpha S^T x + (1-\alpha) v u^T x = \alpha S^T x + (1-\alpha) v$ so the addition again only involves adding a single vector to $\alpha S^T x$. This accounts for another 10^{10} operations. Altogether, we should expect that it will take no more than 10^{12} floating point operations per calculation of $G^T x$, and so we should be able to calculate 1000 iterations in about 10^{15} operations. This is feasable, even on non-super computers. For example, the fastest PC's can now perform about 10^{10} floating point operations per second, so it would take about 10^5 seconds, which is less than 2 days.

These timing estimates are very rough, but they should indicate why the PageRank procedure can be practical. There have been a number of refinements of the general algorithm over the years, to speed it up and to combat various attempts to artificially manipulate PageRanks. The basic idea is still the same.

3.6. Complex eigenvalues

Even though the problems that we are looking at are stated with only real numbers we often find it necessary to use the complex number system to solve them. One way that complex numbers arise is in finding eigenvalues: you need to solve a polynomial equation, and generally, many of the roots are non-real.

In this section we explain one way to deal with non-real eigenvalues without leaving the real domain. The basic facts about algebra in the complex number system are summarized in Appendix A.2.

Start with a $m \times m$ real matrix A. There is no real difference in finding eigenvalues and eigenvectors in the complex domain. We start by solving the characteristic polynomial of A to find the eigenvalues. The characteristic polynomial has degree m and the Fundamental Theorem of Algebra (Theorem A.6) guarantees that there will be m complex solutions, counted with multiplicity.

The next step is to find an eigenvector v_k corresponding to each eigenvalue λ_k. This is done by row reduction of $A - \lambda_k I$, so if λ_k is non-real then v_k will be non-real.

Here's a simple example: If $A = \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix}$ then the characteristic polynomial is $\lambda^2 - 4\lambda + 5$. Using the quadratic formula we find that the eigenvalues are $\lambda_1 = 2+i$ and $\lambda_2 = 2 - i$. The corresponding eigenvectors are $v_1 = \begin{bmatrix} -1-i \\ 1 \end{bmatrix}$ and $v_2 = \begin{bmatrix} -1+i \\ 1 \end{bmatrix}$. We can use this in calculations of $A^n z$ if we write z as a linear combination of eigenvectors, $z = c_1 v_1 + c_2 v_2$. We can find the coefficients c_1 and c_2 by solving a system of linear equations. That is, if P is the matrix with columns v_1 and v_2 and $c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$ then the equation $c_1 v_1 + c_2 v_2 = z$ becomes $Pc = z$, and we need to solve this for c. In our example, suppose $z = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Then we need to solve

$$Pc = z, \quad \text{which is} \quad \begin{bmatrix} -1-i & -1+i \\ 1 & 1 \end{bmatrix} c = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

We can do this by row reduction or by using the inverse of P, so $c = P^{-1} z$. We expect that c_1 and c_2 will be complex numbers, and that is what we get: $c = \begin{bmatrix} 1 + \frac{3}{2}i \\ 1 - \frac{3}{2}i \end{bmatrix}$. Then, according to (3.3), we can write $A^n z = c_1 \lambda_1^n v_1 + c_2 \lambda_2^n v_2$. This becomes

$$(*) \quad \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \left(1 + \frac{3}{2}i\right)(2+i)^n \begin{bmatrix} -1-i \\ 1 \end{bmatrix} + \left(1 - \frac{3}{2}i\right)(2-i)^n \begin{bmatrix} -1+i \\ 1 \end{bmatrix}.$$

Notice that all the non-real expressions on the right side must simplify to a real answer (since the left hand side is real).

There is a more useful way to handle such calculations, which does noes not require using non-real numbers to get real results. Suppose, in the general case, that A is a real $m \times m$ matrix, and that λ is a non-real eigenvalue, with corresponding eigenvector v. The characteristic polynomial of A has real coefficients, so Proposition A.7 shows that the congugate $\overline{\lambda}$ is also an eigenvalue. In fact, if we conjugate the eigenvector equation $Av = \lambda v$ we obtain $\overline{Av} = A\overline{v} = \overline{\lambda v} = \overline{\lambda}\overline{v}$. This shows directly that $\overline{\lambda}$ is an eigenvalue, and it also demonstrates that $\overline{v}$ is an eigenvector corresponding to $\overline{\lambda}$.

First, we want to replace the non-real eigenvvectors v and $\overline{v}$ with equivalent real vectors. We define x to be the real part of v, and y to be the imaginary part of v. Hence $v = x + iy$ and $\overline{v} = x - iy$. Now solve these for x and y to get $x = \dfrac{1}{2}(v + \overline{v})$ and $y = \dfrac{1}{2i}(v - \overline{v})$, or just use the similar formulas for the real and imaginary parts of a complex number in section A.2. Eigenvectors for different eigenvalues are automatically linearly independent, so v and $\overline{v}$ are linearly independent. Moreover, the formulas relating x and y to v and $\overline{v}$ show that x and y are also linearly independent and have the same span as v and $\overline{v}$. So any real vector that can be expressed as $c_1 v + c_2 \overline{v}$ can also be expressed as $d_1 x + d_2 y$, where now d_1 and d_2 are real.

Next, we need to see what happens to the eigenvalues λ and $\overline{\lambda}$. The vectors x and y are not eigenvectors of A, but Ax and Ay have simple formulas. Write $\lambda = a + ib$ where a and b are real. Then

$$\begin{aligned} Ax &= \tfrac{1}{2}\left(Av + A\overline{v}\right) = \tfrac{1}{2}\left(\lambda v + \overline{\lambda}\overline{v}\right) \\ &= \tfrac{1}{2}\left((a+bi)(x+iy) + (a-ib)(x-iy)\right) \\ &= \tfrac{1}{2}\left(ax + iay + ibx - by + ax - iay - ibx - by\right) \\ &= \tfrac{1}{2}(2ax - 2by) = ax - by. \end{aligned}$$

A similar calculation shows that $Ay = bx + ay$. In summary,

$$Ax = ax - by, \;\; Ay = bx + ay.$$

Thus if we have a vector $z = d_1 x + d_2 y$, written as a linear combination of x and y, then we can calculate Az as a linear combination of x and y: $Ax = d_1(ax - by) + d_2(bx + ay) = (d_1 a + d_2 b)x + (-d_1 b + d_2 a)y$. This is better than working with non-real numbers, but it is not clear that it helps much for $A^n z$. There is one more step: Express the eigenvalue λ in polar coordinates as $\lambda = r(\cos\alpha + i\sin\alpha)$ where $r = |\lambda|$ and $\alpha = \arctan\left(\dfrac{b}{a}\right)$ (see section A.2). In this case $\lambda^n = r^n(\cos(n\alpha) + i\sin(n\alpha))$ using DeMoivre's formula (A.1). If we repeat the calculations above, but replacing a and b with $r^n\cos(n\alpha)$ and $r^n\sin(n\alpha)$ we get

$$A^n x = r^n(\cos(n\alpha)x - \sin(n\alpha)y), \;\; Ay = r^n(\sin(n\alpha)x + \cos(n\alpha)y). \tag{3.9}$$

Applying this with $z = d_1 x + d_2 y$, as above, leads to

$$A^n z = r^n(d_1\cos(n\alpha) + d_2\sin(n\alpha))x + r^n(-d_1\sin(n\alpha) + d_2\cos(n\alpha))y. \tag{3.10}$$

We can now reconsider our example, $A = \begin{bmatrix} 3 & 2 \\ -1 & 1 \end{bmatrix}$. As we saw before, the eigenvalues are $\lambda = 2 + i$ and $\overline{\lambda}$. We write $\lambda = a + ib$ so $a = 2$, $b = 1$. In polar coordinates, $r = |\lambda| = \sqrt{a^2 + b^2} = \sqrt{5} \approx 2.236$, and α is determined as

$\alpha = \arctan\left(\frac{b}{a}\right) = \arctan\left(\frac{1}{2}\right) \approx 0.464$. As before, an eigenvector corresponding to λ is $v = \begin{bmatrix} -1-i \\ 1 \end{bmatrix}$. The vectors x and y are the real and imaginary parts of v, so $x = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and $y = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$. Now, given $z = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, we need to find coefficients d_1 and d_2 so that $z = d_1 x + d_2$. As before, this is a system of linear equations for the coefficients, but the system is real. We can write it as $z = Qd$ where Q is the matrix with columns given by x and y, and $d = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}$. So we need to solve

$$Qd = z, \quad \text{which is} \quad \begin{bmatrix} -1 & -1 \\ 1 & 0 \end{bmatrix} d = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

We can solve this by row reduction or by using the inverse of Q, so $d = Q^{-1}z$. The result is $d = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$. Plugging these values into (3.10) and using $\left(\sqrt{5}\right)^n = \left(5^{1/2}\right)^n = 5^{n/2}$, we get

$$A^n z = 5^{n/2}\left(2\cos(n\alpha) - 3\sin(n\alpha)\right)\begin{bmatrix} -1 \\ 1 \end{bmatrix} + 5^{n/2}\left(-2\sin(n\alpha) - 3\cos(n\alpha)\right)\begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

Compare this with the previous formula, $(*)$. The real result $A^n z$ is calculated without using complex arithmetic. Also, it is easier to see that the magnitude of $A^n z$ is roughly proportional to $5^{n/2}$, and that the direction of $A^n z$ is "rotating" in the plane determined by x and y. The downside is that we need to deal with the trig functions of $n\alpha$, and α is not a "standard" angle like $\pi/4$. In many applications, however, we don't actually need to calculate $\sin(n\alpha)$ and $\cos(n\alpha)$.

Note: In general, it is possible to convert the complex representation $A = P\Lambda P^{-1}$, where Λ is the diagonal matrix with the eigenvalues on the diagonal and P is the matrix with the eigenvectors as columns, to a related form which uses only real numbers. This form uses the matrix Q, where each pair of complex conjugate columns in P is replaced by their real and imaginary parts. This gives a representation $A = Q\Delta Q^{-1}$, but the matrix Δ is not diagonal. This matrix Δ is called the *real canonical form* of A, and it is discussed in advanced linear algebra courses.

We close by finishing the proof of part (g) of Theorem 3.6. The gap in the proof occurred when trying to rule out a non-real eigenvalue λ with $|\lambda| > 1$. The proof in the real case used a vector $z = w + av$ where w is a positive probability vector, a is a small constant, and v is an eigenvalue corresponding to λ. The problem is that we need z to be a positive vector, and this can't work if v is non-real. Instead, we use $z = w + ax$ where x is the real part of v; we also need $u^T x = 0$, and this follows from

$u^T v = 0$. At the end of the argument we need to see that $M^k x$ becomes arbitrarily large. Applying (3.10) with $d_1 = 1$ and $d_2 = 0$ gives

$$M^k x = |\lambda|^k \left(\cos(k\alpha)x - \sin(k\alpha)y\right).$$

The norm $\|\cos(k\alpha)x - \sin(k\alpha)y\|$ is bounded below by the distance from $\cos(k\alpha)x$ to the line through the origin and y, and, using simple geometry, this distance is $|\cos(k\alpha)| \cdot \|x\| \sin\theta$ where θ is the angle between x and y. It is, by a similar argument, bounded below by $|\sin(k\alpha)| \cdot \|y\| \sin\theta$. Since one of $|\cos(k\alpha)|$ and $|\sin(k\alpha)|$ is at least $1/2$, $\|\cos(k\alpha)x - \sin(k\alpha)y\|$ is bounded below by a positive constant m, namely $\frac{1}{2}\sin\theta$ times the minimum of $\|x\|$ and $\|y\|$. Thus the norm of $M^k x$ is bounded below by $m\,|\lambda|^k$, so it approaches ∞ as $k \to \infty$.

Exercises

3.1. In the year 1202 Fibonacci published the mathematics book *Liber abaci*, which was one of the first books to promote the use of Arabic numerals in Europe. Here is one of the exercises from the book:

A certain man put a pair of rabbits in a place surrounded on all sides by a wall. How many pairs of rabbits can be produced from that pair in a year if it is supposed that every month each pair begets a new pair which from the second month on becomes productive?

This is perhaps the first age-structured population system. It is not a Leslie model (for example, the rabbits never die), but it leads to the same kind of dynamical system. Represent the state of the population by a vector $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ where x_1 is the number of juvenile rabbit pairs (less than a month old) and x_2 is the number of mature rabbit pairs (1 month or older). Of course, time is measured in months.

The transition function has the form $F(x) = Ax$ for some 2×2 matrix. What is A? What is the initial state corresponding to Fibonacci's problem? What numbers did Fibonacci invent when he solved the problem? Can you express $A^t x$ in terms of these numbers?

3.2. Replace the Leslie matrix in the discussion of (3.1) with $L = \begin{bmatrix} 0 & 3 & 7/3 \\ 1/4 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix}$, and use the same initial state vector $x = \begin{bmatrix} 100 \\ 0 \\ 0 \end{bmatrix}$. Use Maple to answer the following.

(a) Is some power of L positive?
(b) Calculate $L^{20}x$.
(c) Calculate the eigenvalues of L.
(d) How long will it be until the first component of $L^t x$ is greater than 10000.

3.3. Consider the following generalization of the Leslie matrix in the discussion of (3.1): $L = \begin{bmatrix} 0 & a & b \\ 1/4 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix}$, with $a \geq 0$ and $b \geq 0$. Under what conditions on a and b will 1 be an eigenvalue of L? [Suggestion: Find the characteristic polynomial of L, and then plug $\lambda = 1$ into the characteristic polynomial and set the result equal to 0. This will give you an equation involving a and b. Now find the restrictions necessary so that $a \geq 0$ and $b \geq 0$.]

3.4. Find a general formula for A^n, or at least a non-recursive procedure for determining A^n for any n, for each of the following:,

(a) $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

(b) $A = \begin{bmatrix} 2 & -2 & 8 \\ -1 & 2 & -6 \\ -1 & 1 & -4 \end{bmatrix}$

(c) $A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$

(d) $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$

(e) $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$

(f) $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$

3.5. In the general form of the Leslie matrix in Example 3.1 suppose that all the survival rates s_j and all the fecundity rates f_j for $j > 0$ are positive. Then some power of L is a positive matrix.

3.6. For each of the matrices in Exercise 3.4 find the eigenvalues and decide whether there is a dominant eigenvalue. In cases where there is a dominant eigenvalue, verify that the conclusion of Theorem 3.3 is satisfied. Use $x =$ the unit vector with 1 in the first entry and 0's elsewhere.

3.7. Let $B = \begin{bmatrix} 2 & 1 \\ -1 & 0 \end{bmatrix}$.

(a) What are the eigenvalues and corresponding eigenvectors of B?
(b) Is B diagonalizable?
(c) Calculate a few powers of B. Do you see a pattern? Can you find a non-recursive formula for B^n?

3.8. A is a diagonalizable matrix. Using Maple, I calculated

$$A^{30} = \begin{bmatrix} -11318 & 1617 & 8893 \\ -22582 & 3227 & 17743 \\ -11327 & 1618 & 8900 \end{bmatrix}, \quad A^{31} = \begin{bmatrix} -14127 & 2018 & 11100 \\ -28309 & 4043 & 22242 \\ -14118 & 2017 & 11093 \end{bmatrix} \qquad \text{(rounded)}$$

From this, determine an eigenvalue λ_1 and corresponding eigenvector v_1 for A. You should normalize v_1 so that its third component is 1. You should be able to determine

λ_1 to the nearest hundredth. (Do not use anything except a simple calculator, for doing arithmetic.)

3.9. This is a classical example of a linear dynamical system. We want to predict polution levels in the Great Lakes, assuming all sources of pollution are turned off; the goal is to see how long it will take before the water in the lakes is reasonably free of pollution. (This might be a good time to look at a map of the Great Lakes.)

The basic model is very simplistic: There is an amount of pollutant, measured in tons, dispersed evenly throughout each of the lakes (a different amount for each lake). Dividing this amuount by the volume of the lake, measured in cubic miles, gives the amount of pollution per unit volume, or the concentration, in the lake. If water flows from lake A to lake B then, in one year, some of the pollutant will be removed from lake A and added to lake B. Since the pollutant is assumed to be evenly distributed, the amount of pollutant transferred from lake A to B is just the volume of water that flows from lake A to lake B in one year, times the concentration of pollutant in lake A. If, during the year, some water flows out of a lake, but not to another lake in the system, then the corresponding amount of pollutant is just removed from the lake. If there is inflow into a lake from outside the system then we presume that it does not carry any pollution into the lake.

Here are the current volumes and flow rates, in cubic miles:

Lake	Volume	Inflow	Outflow	Flow
Superior	2900	15		15 → Huron
Michigan	1180	38		38 → Huron
Huron	850	15		68 → Erie
Erie	116	17		85 → Ontario
Ontario	393	14	99	

Suppose the original pollutant totals 5439 tons, for an overall concentration of 1 ton per cubic mile. Select an initial distribution of the 5439 tons among the 5 lakes. You can just assume that all have the same concentration, or you can assume, for example, that the concentration in Superior is only a 0.1 ton per cubic mile, with correspondingly larger concentrations in the other lakes. Just make sure that, in your initial distribution, the total amount of pollutant is 5439, that all initial concentrations are at least 0.1, and that no lake has a concentration greater than 5.

(a) The state vector x is 5 dimensional, and contains the *concentration* of pollutant in each lake. Set up the transition rule for the state vector; it will have the form $F(x) = Ax$ where A is a non-negative matrix. Be careful: For flow from one lake to another you must account for changes in both lakes.

(b) Use Maple to find the first time t at which all the lakes have a concentration less than .01.

(c) Use Maple to determine the largest concentration that will occur in any lake.

I do not expect that you willl need any advanced features of Maple to answer these questions. You just need to calculate values of $F^t(x)$ and inspect the answers. Turn in a copy of your Maple session. If you want, send it to me by email, as a text file, not as a live Maple worksheet.

3.10. There are three towns, North Apple, South Apple, Apple Core (abbreviated N, S, C). There is considerable migration between these towns. In each year

from N: 10% move to S, 10% move to C, and the rest stay in N;
from S: 5% move to N, 15% move to C, and the rest stay in S;
from C: 15% move to N, 20% move to S, and the rest stay in C.

This is a very simple model; we are ignoring births and deaths, and migration from or to the outside.

(a) Draw the network for this Markov chain, and write down the corresponding matrix W. Follow the same scheme as in (3.5) and (3.6).
(b) Find the limiting probability distribution p, using the fact that it is an eigenvector of $M = W^T$ corresponding to the eigenvalue 1. Be sure that the entries in p sum to 1.
(c) Initially there are 1000000 people in each of the cities. What do you expect the population in each city to be after many years?

3.11. Modify Exercise 3.10 to handle births and deaths. That is, adjust the percentage remaining in each town by adding a net growth rate: +5% in N, +3% in S, and −10% in C. Then the matrix W is no longer a stochastic matrix, but it does have a dominant eigenvalue.

(a) Find the dominant eigenvalue and corresponding eigenvector (use Maple).
(b) Based on this eigenvalue and eigenvector, describe what will happen to the population after many years.

3.12. Let $B = \begin{bmatrix} -4 & 4 & 6 \\ 6 & -2 & -6 \\ -8 & 7 & 10 \end{bmatrix}$; its characteristic polynomial is

$$-\lambda^3 + 4\lambda^2 - 14\lambda + 20.$$

The calculations below do not involve any messy numbers, so they can be done by hand. Optionally, use Maple.

(a) Find the eigenvalues. [Hint: 2 is an eigenvalue, so you can factor $2-\lambda$ out of the characteristic polynomial. Now use the quadratic formula. You should get two non-real answers, $\lambda = a + bi$, and $\overline{\lambda}$, where a and b are integers.]
(b) Find the eigenvectors, w, v, $\overline{v}$, corresponding, in order, to the eigenvalues 2, λ, $\overline{\lambda}$.

(c) Let x and y be the real and imaginary parts of v, and write a formula for Bz if the vector z has the form $z = c_1 w + c_2 x + c_3 y$.
(d) Write λ in polar coordinates, and use this to write an expression for $B^k z$.

3.13. Show that a dominant eigenvalue of a real matrix must be real. [Suppose λ is a non-real eigenvalue. Is $\overline{\lambda}$ an eigenvalue? Is $\overline{\lambda} = \lambda$? Is $\left|\overline{\lambda}\right| = |\lambda|$? What goes wrong with the definition of dominant?]

3.14. Let $C = \begin{bmatrix} 0 & -1 \\ 1 & 6/5 \end{bmatrix}$.

(a) Find the eigenvalues of C.
(b) What happens to C^k as $k \to \infty$?
(c) Using Maple, you can check that all entries of C^{62} are within about .02 of the corresponding entries in C. Can you explain this?

3.15. If M is a square matrix then show that:

(a) M is invertible if and only if 0 is not an eigenvalue of M.
(b) If M is invertible then v is an eigenvector of M with eigenvalue λ if and only if v is an eigenvector of M^{-1} with eigenvalue λ^{-1}.

CHAPTER 4

Linear models: continuous version

4.1. The exponential function

There are many definitions of the exponential function e^x for real numbers. It is often defined in Calculus as the inverse of the natural logarithm or as the limit of $\left(1+\frac{x}{n}\right)^n$ as $n \to \infty$; often it is just understood (somewhat circularly) as "e to the x^{th} power". These definitions do not generalize very well.

However, we will find it very useful to work with the following definition:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \dots \tag{4.1}$$

It is shown in Calculus that this series converges absolutely for all real numbers x and that it coincides with the other definitions of e^x.

The basic idea is that this series definition can be used for any objects x for which we can make sense of powers (x^n for positive integers n), vector space operations (scalar multiplication and addition), and limits. We also need a suitable notion of a multiplicative unit, since the series begins with 1, and we would like x^0 to equal this multiplicative unit.

Here is one situation where we can use this idea: Suppose that A is an $m \times m$ matrix. Then we define the *matrix exponential* by the following formula:

$$e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!} = I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \frac{1}{4!}A^4 + \dots \tag{4.2}$$

where I is the $m \times m$ identity matrix.

This definition will make sense as long as the series converges. This is a series of $m \times m$ matrices, so the partial sums $\displaystyle\sum_{n=0}^{N} \frac{A^n}{n!}$ can be calculated; each partial sum is an $m \times m$ matrix. The infinite series is the limit of these partial sums, and we interpret limits of matrices by taking the limits of the entries. The following is the basic justification for the convergence of the series for e^A, but in practice we will have explicit methods for calculating e^A so this is only useful for reassurance:

PROPOSITION 4.1. *For any square matrix A, each entry in e^A is an absolutely convergent series. Hence the series for e^A converges.*

PROOF. Let M be an upper bound for the absolute values of the entries of A. We need an estimate on the size of the entries in A^n.

Suppose B is another $m \times m$ matrix and L is an upper bound for the entries of B. The ik entry in AB is $\sum_j A_{ij}B_{jk}$, and in absolute value this is bounded by $\sum_j |A_{ij}| \cdot |B_{jk}|$. There are m terms in this sum and each term is bounded by $M \cdot L$, so the sum is $\leq mML$. In other words, each entry of AB is bounded, in absolute value, by mML.

Now if we apply this with $B = A$ we see that each entry of $A^2$is bounded, in absolute value, by mM^2. Repeating the argument with $B = A^2$, we see that each entry of A^3 is bounded, in absolute value, by $m \cdot M \cdot mM^2 = m^2M^3$. Proceeding by induction we find that each entry of A^n is bounded by $m^{n-1}M^n$. This is a crude estimate, and we can make it even cruder (and simpler): Each entry of A^n is bounded, in absolute value, by $(mM)^n$.

Now consider any entry in the series for e^A. It is a sum of entries from various powers A^n, divided by $n!$. Replacing each term in this series by its absolute value and using the estimate above, we get a series of the form $\sum_{n=0}^{\infty} \frac{(mM)^n}{n!}$. But this is just the series for e^{mM}, and we know that this converges absolutely. Hence, by the comparison test, the series for each entry of e^A converges absolutely. □

Many of the manipulations with matrix exponentials can be reduced to the corresponding facts about real exponentials, and similar estimates are often necessary. We will not give any further details for such limit arguments.

As a first example we calculate e^A where $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. This was considered in Exercise 3.4; the n^{th} power is $A^n = \begin{bmatrix} 2^{n-1} & 2^{n-1} \\ 2^{n-1} & 2^{n-1} \end{bmatrix}$. Then

$$\begin{aligned} e^A &= I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \frac{1}{4!}A^4 + \dots \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \frac{1}{2!}\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} + \frac{1}{3!}\begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix} + \frac{1}{4!}\begin{bmatrix} 8 & 8 \\ 8 & 8 \end{bmatrix} + \dots \end{aligned}$$

The $1,1$ and $2,2$ entries can be converted into standard exponentials with some algebraic manipulation:

$$\begin{aligned} A_{11} = A_{22} &= 1 + 1 + \frac{1}{2!} \cdot 2 + \frac{1}{3!} \cdot 2^2 + \frac{1}{4!} \cdot 2^3 + \dots \\ &= 1 + \frac{1}{2}\left(2 + \frac{1}{2!} \cdot 2^2 \frac{1}{3!} \cdot 2^3 + \frac{1}{4!} \cdot 2^4 + \dots\right) \\ &= \frac{1}{2} + \frac{1}{2}\left(1 + 2 + \frac{1}{2!} \cdot 2^2 \frac{1}{3!} \cdot 2^3 + \frac{1}{4!} \cdot 2^4 + \dots\right) = \frac{1}{2} + \frac{1}{2} \cdot e^2. \end{aligned}$$

The $1,2$ and $2,1$ entries are the same, except that they do not have the first 1 (which came from the identity matrix in the $1,1$ and $2,2$ entries). Hence we have $A_{21} = A_{12} = \left(\frac{1}{2} + \frac{1}{2}e^2\right) - 1 = -\frac{1}{2} + \frac{1}{2}e^2$. Putting this together, we have

$$e^A = \exp\left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^2 & -\frac{1}{2} + \frac{1}{2}e^2 \\ -\frac{1}{2} + \frac{1}{2}e^2 & \frac{1}{2} + \frac{1}{2}e^2 \end{bmatrix}.$$

Of course, most matrix powers are not as simple as this example, so this is not really a representative result. Even for this case, it is not easy to see what the general form will be.

There is an important case in which it is easy to calculate the matrix exponential. Suppose that Λ is a diagonal matrix, with λ_k in the k,k entry. Then Λ^n is still diagonal, with λ_k^n as the k,k entry. If we put this into the series for e^λ we see that the result is still diagonal, with $\sum_{n=0}^{\infty} \lambda_k^n$ as the k,k entry. In other words,

$$e^\Lambda = \exp\left(\begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_m \end{bmatrix}\right) = \begin{bmatrix} e^{\lambda_1} & 0 & 0 & \dots & 0 \\ 0 & e^{\lambda_2} & 0 & \dots & 0 \\ 0 & 0 & e^{\lambda_3} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & e^{\lambda_m} \end{bmatrix}$$

This example becomes most useful if we consider a diagonalizable matrix A. In this case $A = P\Lambda P^{-1}$ for a diagonal matrix Λ and an invertible matrix P. Remembering that $A^n = P\Lambda^n P^{-1}$, we find

$$e^A = \sum_{n=0}^{\infty} \frac{1}{n!} A^n = \sum_{n=0}^{\infty} \frac{1}{n!} P\Lambda^n P^{-1} = P\left(\sum_{n=0}^{\infty} \frac{1}{n!} \Lambda^n\right) P^{-1} = Pe^\Lambda P^{-1}.$$

Here is another calculation of e^A for $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$: The eigenvalues of A are $\lambda_1 = 0$ and $\lambda_2 = 2$, with corresponding eigenvectors $v_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and $v_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Then the matrix $P = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ has the eigenvectors as columns, and we calculate $P^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$. Hence

$$e^A = Pe^{\Lambda}P^{-1} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \cdot \begin{bmatrix} e^0 & 0 \\ 0 & e^2 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^2 & -\frac{1}{2} + \frac{1}{2}e^2 \\ -\frac{1}{2} + \frac{1}{2}e^2 & \frac{1}{2} + \frac{1}{2}e^2 \end{bmatrix}.$$

In the future we will need to calculate expressions like e^{tA} and $e^{tA}x$ where t is a real number and x is a vector, so we record the following formulas:

PROPOSITION 4.2. *Suppose A is diagonalizable, with eigenvalues $\lambda_1, \ldots, \lambda_m$ and corresponding eigenvectors $v_1, \ldots, v_m$. Let P be the matrix with the eigenvectors as columns and let Λ be the diagonal matrix with the eigenvalues on the diagonal. Then*

(a) $e^{tA} = Pe^{t\Lambda}P^{-1} = P \cdot \begin{bmatrix} e^{\lambda_1 t} & 0 & 0 & \ldots & 0 \\ 0 & e^{\lambda_2 t} & 0 & \ldots & 0 \\ 0 & 0 & e^{\lambda_3 t} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & e^{\lambda_m t} \end{bmatrix} \cdot P^{-1}$.

(b) *If x is expressed as a linear combination of the eigenvectors, $x = c_1v_1 + \cdots + c_mv_m$, then $e^{tA}x = c_1e^{t\lambda_1}v_1 + \cdots + c_me^{t\lambda_m t}v_m$.*

PROOF. Since $tA = P(t\Lambda)P^{-1}$, part (a) is just a restatement of our calculations above.

For part (b), let c be the vector with entries c_j. Then Pc is the linear combination of the columns of P with coefficients given by c; but this linear combination is $x = c_1v_1 + \cdots + c_mv_m$. That is, $x = Pc$. Hence $e^{tA}x = P \cdot e^{t\Lambda}P^{-1}Pc = P \cdot e^{t\Lambda}c$. Now this is the linear combination of the columns of P with coefficients given by $e^{t\Lambda}c$, and this is the column vector with entries $e^{t\lambda_j}c_j$. Therefore $e^{tA}x = c_1e^{t\lambda_1}v_1 + \cdots + c_me^{t\lambda_m t}v_m$. □

There are many properties of the real exponential function, and many of them are still true for the matrix exponential. Here is a list:

PROPOSITION 4.3. *Suppose A and B are $m \times m$ matrices.*

(a) $e^O = I$, *where O is the zero matrix.*
(b) $e^{A+B} = e^A \cdot e^B$ *if A and B commute; that is, $AB = BA$.*
(c) e^A *is invertible, with inverse e^{-A}.*
(d) $e^{(s+t)A} = e^{sA} \cdot e^{tA}$, *if s and t are scalars.*

(e) $\dfrac{d}{dt}\left(e^{tA}\right) = A \cdot e^{tA}$.
(f) *A and e^{tA} commute.*

Warning: $e^{A+B} = e^A \cdot e^B$ is **false** if A and B do not commute. For an example see 4.6.

PROOF. Part (a) is obvious. Part (c) follows from parts (a) and (b), since A and $-A$ commute, so $e^A \cdot e^{-A} = e^{-A} \cdot e^A = e^{A+(-A)} = e^O = I$. Part (d) follows from part (b) since sA and tA commute.

To prove part (b) we multiply the series for e^A and e^B and compare the result to the series for e^{A+B}. Grouping terms of the same degree, and being careful to preserve the order of multiplication,

$$\begin{aligned} e^A \cdot e^B &= \left(I + A + \frac{1}{2}A^2 + \frac{1}{6}A^3 + \dots\right) \cdot \left(I + B + \frac{1}{2}B^2 + \frac{1}{6}B^3 + \dots\right) \\ &= I + (A+B) + \frac{1}{2}\left(A^2 + 2AB + B^2\right) \\ &\qquad + \frac{1}{6}\left(A^3 + 3A^2B + 3AB^2 + B^3\right) + \dots \\ e^{A+B} &= I + (A+B) + \frac{1}{2}(A+B)^2 + \frac{1}{6}(A+B)^3 + \dots \\ &= I + (A+B) + \frac{1}{2}(A^2 + AB + BA + B^2) \\ &\qquad + \frac{1}{6}\left(A^3 + A^2B + ABA + BA^2 + AB^2 + BAB + B^2A + B^3\right) + \dots \end{aligned}$$

It should now be clear why we require that A and B commute: In order to have $e^A e^B = e^{A+B}$ we would need the quadratic terms to agree; but if

$$(A+B)^2 = A^2 + AB + BA + B^2 \quad \text{equals} \quad A^2 + 2AB + B^2$$

then we must have $AB + BA = 2AB$, so $BA = AB$. Also, it should be clear that we have no reason to expect the terms of higher degree to be equal unless A and B commute.

On the other hand, if A and B commute then the parts of the expansions that we have calculated are the same, and this is true for the rest of the terms, using the binomial formula to expand $(A+B)^n$. [A rigorous proof requires a bit more work, since we need to prove that it is possible to multiply series of matrices as we did above.]

To prove part (e) we differentiate the series for e^{tA} term-by-term; this is justified since it is a power series (in each entry) with infinite radius of convergence:

$$\begin{aligned}\frac{d}{dt}\left(e^{tA}\right) &= \frac{d}{dt}\left(\sum_{n=0}^{\infty}\frac{(tA)^n}{n!}\right) = \frac{d}{dt}\left(\sum_{n=0}^{\infty}\frac{t^nA^n}{n!}\right)\\ &= \sum_{n=1}^{\infty}\frac{nt^{n-1}A^n}{n!} = \sum_{n=1}^{\infty}\frac{t^{n-1}A^n}{(n-1)!}\\ &= A\cdot\left(\sum_{n=1}^{\infty}\frac{t^{n-1}A^{n-1}}{(n-1)!}\right) = A\cdot\left(\sum_{k=0}^{\infty}\frac{t^kA^k}{k!}\right) = Ae^{tA}.\end{aligned}$$

To prove part (f) we notice that A commutes with any power of A, since $A\cdot A^n$ and $A^n\cdot A$ are both equal to A^{n+1}. Hence

$$A\cdot\left(\sum_{n=0}^{\infty}\frac{(tA)^n}{n!}\right) = \sum_{n=0}^{\infty}\frac{t^n\,A\cdot A^n}{n!} = \sum_{n=0}^{\infty}\frac{t^n\,A^n\cdot A}{n!} = \left(\sum_{n=0}^{\infty}\frac{(tA)^n}{n!}\right)\cdot A. \qquad \square$$

In the discussion so far we have assumed that our matrices are real matrices. However, if we want to use eigenvalues and eigenvectors to calculate matrix exponentials we have to be prepared to handle e^λ if λ is a non-real eigenvalue. There are no real problems with this – the calculations proceed just as above – except that we need a definition for e^z where z is a complex number. Of course, all we need to do is plug a complex number into the series for the exponential function. The analog of Proposition 4.3 is true for exponentials of complex numbers, and, since complex multiplication is commutative, $e^{z+w} = e^ze^w$ is true for all complex numbers. There are some special rules for exponentials of complex numbers:

PROPOSITION 4.4. *In addition to the analogs of the properties in Proposition 4.3 we have the following. Here z is a complex number and x and y are real.*

(a) $\overline{e^z} = e^{\overline{z}}$.
(b) $e^{x+iy} = e^xe^{iy}$.
(c) *Euler's formula:* $e^{iy} = \cos(y) + i\,\sin(y)$.
(d) $\left|e^{x+iy}\right| = e^x$.

PROOF. Part (a) follows by conjugating the series for e^z. Part (b) is a consequence of $e^{z+w} = e^ze^w$. Part (d) follows from parts (b) and (c), together with $|\cos(y) + i\,\sin(y)| = \sqrt{\cos^2(y) + \sin^2(y)} = 1$.

Part (c) requires the series expansions of the sine and cosine, plus the fact that the sequence i^n follows the periodic pattern $1, i, -1, -i, 1, i, -1, -i, \ldots$:

$$\begin{aligned} e^{iy} &= 1 + (iy) + \frac{1}{2!}(iy)^2 + \frac{1}{3!}(iy)^3 + \frac{1}{4!}(iy)^4 + \frac{1}{5!}(iy)^5 + \frac{1}{6!}(iy)^6 + \ldots \\ &= 1 + iy + -\frac{1}{2!}y^2 - i\frac{1}{3!}y^3 + \frac{1}{4!}y^4 + i\frac{1}{5!}y^5 - \frac{1}{6!}y^6 + \ldots \\ &= \left(1 - \frac{1}{2!}y^2 + \frac{1}{4!}y^4 - \frac{1}{6!}y^6 + \ldots\right) + i\left(y - \frac{1}{3!}y^3 + \frac{1}{5!}y^5 + \ldots\right) \\ &= \cos(y) + i\sin(y). \end{aligned}$$

□

Here's an example of calculating a matrix exponential when the eigenvalues are complex. Let $A = t\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} t & -t \\ t & t \end{bmatrix}$, where t is a real number. This has eigenvalues $\lambda = t + it$ and $\overline{\lambda} = t - it$, with corresponding eigenvectors $v = \begin{bmatrix} i \\ 1 \end{bmatrix}$ and $\overline{v} = \begin{bmatrix} -i \\ 1 \end{bmatrix}$. The change of basis matrix is $P = \begin{bmatrix} i & -i \\ 1 & 1 \end{bmatrix}$, with inverse $P^{-1} = \begin{bmatrix} -\frac{1}{2}i & \frac{1}{2} \\ \frac{1}{2}i & \frac{1}{2} \end{bmatrix}$. Then we have

$$\begin{aligned} \exp\left(\begin{bmatrix} t & -t \\ t & t \end{bmatrix}\right) &= P \cdot \exp\left(\begin{bmatrix} t+it & 0 \\ 0 & t-it \end{bmatrix}\right) \cdot P^{-1} = P \cdot \begin{bmatrix} e^{t+it} & 0 \\ 0 & e^{t-it} \end{bmatrix} \cdot P^{-1} \\ &= \begin{bmatrix} i & -i \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} e^t(\cos(t) + i\sin(t)) & 0 \\ 0 & e^t(\cos(t) - i\sin(t)) \end{bmatrix} \cdot \begin{bmatrix} -\frac{1}{2}i & \frac{1}{2} \\ \frac{1}{2}i & \frac{1}{2} \end{bmatrix} \\ &= \begin{bmatrix} e^t\cos(t) & -e^t\sin(t) \\ e^t\sin(t) & e^t\cos(t) \end{bmatrix} \end{aligned}$$

Exercise 4.6 provides a direct derivation of the general formula for the exponential of the matrix $C = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$:

$$e^C = \exp\left(\begin{bmatrix} a & -b \\ b & a \end{bmatrix}\right) = e^a \begin{bmatrix} \cos(b) & -\sin(b) \\ \sin(b) & \cos(b) \end{bmatrix} = \begin{bmatrix} e^a\cos(b) & -e^a\sin(b) \\ e^a\sin(b) & e^a\cos(b) \end{bmatrix}, \tag{4.3}$$

which agrees with the calculations above.

Here is the analog of (3.9) for matrix exponentials. Suppose that $\lambda = a + ib$ is a non-real eigenvalue with corresponding eigenvector v. Let x and y be the real and imaginary parts of v, so x and y are real vectors and $v = x + iy$. Then

$$e^{tA}x = e^{at}\left(\cos(bt)x - \sin(bt)y\right), \quad e^{tA}y = e^{at}\left(\sin(bt)x + \cos(bt)y\right). \tag{4.4}$$

This is proved as follows. First, (3.9) says that the coefficient of x in $A^n x$ is $r^n \cos(n\alpha)$, which is the real part of λ^n. Thus if we apply (3.9) to every term in $e^{tA}x = \sum_{n=0}^{\infty} \frac{t^n}{n!} A^n x$ and just look at the coefficient of x we get a coefficient of

$$\sum_{n=0}^{\infty} \frac{t^n}{n!} \operatorname{Re}(\lambda^n) = \operatorname{Re}\left(\sum_{n=0}^{\infty} \frac{t^n}{n!} \lambda^n \right) = \operatorname{Re}(e^{t\lambda}) = \operatorname{Re}(e^{at+ibt}) = e^{at}\cos(bt).$$

The other coefficients in (4.4) are calculated the same way.

Finally, if $z = d_1 x + d_2 y$ then the analog of (3.10) is

$$e^{tA}z = e^{at}(d_1 \cos(bt) + d_2 \sin(bt))x + e^{at}(-d_1 \sin(bt) + d_2 \cos(bt))y. \tag{4.5}$$

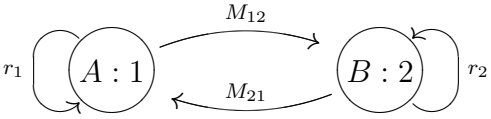
4.2. Some models

We first look at a population model for two (or more) separated populations in which migration occurs, as well as natural population changes doe to births and deaths.

Consider two cities, A and B. For each city there is an intrinsic rate of growth, given by the difference between birth and death rates; these rates may be different. Also, there is a migration rate for individuals moving from A to B, and a migration rate in the other direction; these rates may be different.

If we number the cities as 1 and 2 then we can use the parameters r_1 and r_2 for the intrinsic growth rates, and we can use M_{12} and M_{21} for the migration rates from 1 to 2 and from 2 to 1. We can visualize this information as the following weighted digraph.

(4.6)



We need to interpret this diagram correctly in order to construct the corresponding dynamical system. We use two state variables, x_1 and x_2, for the populations of the two cities. If we consider the changes in x_1 and x_2 during a small time interval Δt then we find, approximately,

$$\begin{aligned} \Delta x_1 &= (r_1 x_1 - M_{12} x_1 + M_{21} x_2)\Delta t = (M_{11} x_1 + M_{21} x_2)\Delta t \\ \Delta x_2 &= (r_2 x_2 - M_{21} x_2 + M_{12} x_1)\Delta t = (M_{12} x_1 + M_{22} x_2)\Delta t, \end{aligned}$$

where $M_{11} = r_1 - M_{12}$ and $M_{22} = r_2 - M_{21}$.

Caution: The multiplier of $x_j \Delta t$ is not just r_j, but r_j minus the rate of migration out of city j.

We divide Δx_1 and Δx_2 by Δt to obtain, in matrix terms,

$$\frac{\Delta x}{\Delta t} = \frac{1}{\Delta t}\begin{bmatrix}\Delta x_1\\ \Delta x_2\end{bmatrix} = \begin{bmatrix}M_{11}x_1 + M_{21}x_2\\ M_{12}x_1 + M_{22}x_2\end{bmatrix} = \begin{bmatrix}M_{11} & M_{21}\\ M_{12} & M_{22}\end{bmatrix}\cdot\begin{bmatrix}x_1\\ x_2\end{bmatrix}.$$

As in the one dimensional case in Chapter 2, this is not correct for large values of Δt, but it becomes more accurate as Δt gets smaller, so we take the limit as $\Delta t \to 0$. So we have our next example:

EXAMPLE 4.5. The state of the system is an m-dimensional vector x representing populations in m cities. The j^{th} city has intrinsic growth rate r_j and the migration rate from the j^{th} city to the k^{th} city is M_{jk}, for $j \neq k$. We assume the migration rates M_{jk} are non-negative. We define M_{jj} to be r_j minus the sum of all migration *out* of the j^{th} city, so

$$M_{jj} = r_j - \sum_{k\neq j} M_{jk}.$$

Finally, we let M be the matrix with entries M_{jk}. Then the dynamical system is defined by the vector differential equation $\dfrac{dx}{dt} = M^T x$.

As we saw in Example 2.1, the one-dimensional differential equation $\dfrac{dx}{dt} = ax$, with initial condition $x(0) = x_0$, has the solution $x(t) = e^{at}x_0$. We can't use the method of solution that we used in Example 2.1, since that involved division and we are using vectors, but we can verify that essentially the same solution works:

PROPOSITION 4.6. *If A is a square matrix then the vector differential equation $\dfrac{dx}{dt} = Ax$, with initial condition $x(0) = x_0$, has the unique solution $x(t) = e^{tA}x_0$.*

PROOF. Define $x(t) = e^{tA}x_0$. Using Proposition 4.3(e), we have

$$\frac{d}{dt}x(t) = \frac{d}{dt}\left(e^{tA}\right)\cdot x_0 = A\cdot e^{tA}x_0 = A\cdot x(t),$$

so $x(t)$ satisfies the differential equation. Also, $x(0) = e^O \cdot x_0 = I \cdot x_0 = x_0$, so $x(t)$ satisfies the initial condition.

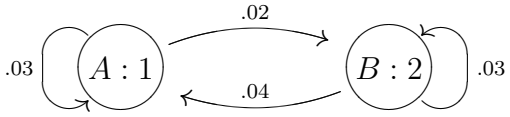
For uniqueness, suppose that $\tilde{x}(t)$ is any other function that satisfies the differential equation and initial conditions, and let $z(t) = e^{-tA}\cdot\tilde{x}(t)$. Since $\tilde{x}(t)$ satisfies the differential equation we have $\dfrac{d}{dt}\tilde{x}(t) = A\tilde{x}(t)$. Using this plus the product rule and Proposition 4.3 (e) we find

$$\frac{d}{dt}z(t) = \frac{d}{dt}\left(e^{-tA}\right)\cdot\tilde{x}(t) + e^{-tA}\cdot\frac{d}{dt}\tilde{x}(t) = -A\cdot e^{-tA}\tilde{x}(t) + e^{-tA}\cdot A\tilde{x}(t).$$

Since A and e^{-tA} commute, by Proposition 4.3 (f), we can simplify this last expression to $-Ae^{-tA}\tilde{x}(t) + Ae^{-tA}\tilde{x}(t) = 0$. Since $\dfrac{dz}{dt} = 0$, the function $z(t) = e^{-tA} \cdot \tilde{x}(t)$ is a constant. Plugging in $t = 0$ and $\tilde{x}(0) = x_0$, we have $z(0) = e^O \cdot x_0 = x_0$, so $z(t) = e^{-tA} \cdot \tilde{x}(t) = x_0$ for all t. Multiplying by e^{tA} gives $\tilde{x}(t) = e^{tA}x_0$. □

Vector differential equations of the form $\dfrac{dx}{dt} = Ax$, where A is a (constant) square matrix, are called *constant coefficient linear differential equations*, and these form the main class of differential equations that can be explicitly solved in terms of elementary functions.

Here is a specific case of Example 4.5. We start with the digraph



Following the recipe of Example 4.5, we have $M_{12} = .02$ and $r_1 = .03$, so $M_{11} = .03 - .02 = .01$. Also, $M_{21} = .04$ and $r_2 = .03$ so $M_{22} = .03 - .04 = -.01$. Therefore $M = \begin{bmatrix} .01 & .02 \\ .04 & -.01 \end{bmatrix}$. If we let $A = M^T$ then the differential equation is

$$\frac{dx}{dt} = Ax, \quad A = \begin{bmatrix} .01 & .04 \\ .02 & -.01 \end{bmatrix}. \tag{4.7}$$

To solve this we need to calculate e^{tA}, so we start by diagonalizing A. We find that the eigenvalues are $\lambda_1 = .03$ and $\lambda_2 = -.03$, with corresponding eigenvectors $v_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. As usual, let P be the matrix with the eigenvectors as columns and let Λ be the diagonal matrix with the eigenvalues on the diagonal. According to Proposition 4.2 (a), $e^{tA} = Pe^{t\Lambda}P^{-1}$. From this we can calculate

$$e^{tA} = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} e^{.03t} & 0 \\ 0 & e^{-.03t} \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix} = \begin{bmatrix} \frac{2}{3}e^{.03t} + \frac{1}{3}e^{-.03t} & \frac{2}{3}e^{.03t} - \frac{2}{3}e^{-.03t} \\ \frac{1}{3}e^{.03t} - \frac{1}{3}e^{-.03t} & \frac{1}{3}e^{.03t} + \frac{2}{3}e^{-.03t} \end{bmatrix}.$$

If we write the initial populations as $x_1(0) = p_1$, $x_2(0) = p_2$ then computing $e^{tA}\begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$ gives the flow $x = F^t(p)$:

$$x_1 = \left(\frac{2}{3}e^{.03t} + \frac{1}{3}e^{-.03t}\right)p_1 + \left(\frac{2}{3}e^{.03t} - \frac{2}{3}e^{-.03t}\right)p_2$$
$$x_2 = \left(\frac{1}{3}e^{.03t} - \frac{1}{3}e^{-.03t}\right)p_1 + \left(\frac{1}{3}e^{.03t} + \frac{2}{3}e^{-.03t}\right)p_2.$$

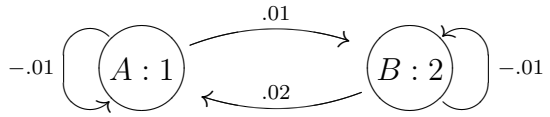
For large t the terms involving $e^{-.03t}$ are negligible, so the solution is approximately

$$x_1 \approx \frac{2}{3}(p_1 + p_2)e^{.03t},\ x_2 \approx \frac{1}{3}(p_1 + p_2)e^{.03t}.$$

So, for large t, both city populations are growing exponentially, with twice as many in city 1 as in city 2.

A simpler way of seeing this is to use Proposition 4.2(b). If the initial vector p is written in terms of the eigenvectors as $p = c_1v_1 + c_2v_2$ then $x = e^{tA}p = c_1e^{.03t} + c_2e^{-.03t}$. From this it is clear that, for large t, $x \approx c_1e^{.03t}v_1$, so $x(t)$ is growing exponentially, at a rate of 3%, in the direction given by $v_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

Here is another "two cities" example:



The corresponding differential equation is

$$\frac{dx}{dt} = Bx, \quad B = \begin{bmatrix} -.02 & .02 \\ .01 & -.03 \end{bmatrix}. \tag{4.8}$$

The matrix B has eigenvalues $\lambda_1 = -.01$ and $\lambda_2 = -.04$, with corresponding eigenvectors $v_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. As above we write the initial condition as $x(0) = p = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$ and then express p as a linear combination of the eigenvectors, so $p = c_1v_1 + c_2v_2$. We can determine c_1 and c_2 from $Pc = p$ by solving for c using row reduction, or by calculating P^{-1}, so that $c = P^{-1}p$. (Since p is arbitrary these two approaches are equivalent, but with a specific choice of p the first is usually simpler.) We obtain $c_1 = \frac{1}{3}p_1 + \frac{1}{3}p_2$, $c_2 = -\frac{1}{3}p_1 + \frac{2}{3}p_2$. Then the solution is $x = e^{tB}p = c_1e^{-.01t}v_1 + c_2e^{-.04}v_2$. From this it is clear that the populations in the two cities are dying out. For large t the two exponentials $e^{-.01t}$ and $e^{-.04t}$ approach zero,

but the second goes to zero faster. Hence, for large t, $x \approx c_1 e^{-.01t} v_1 = c_1 e^{-.01t} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, so the population is very small and the first city has approximately twice as many individuals as the second.

Some simple physics problems lead to constant coefficient linear differential equations. The basis for many physical applications is Newton's Second Law, which says that the position of a simple object of mass m satisfies

$$m\frac{d^2x}{dt^2} = F \tag{4.9}$$

where F is the force acting on the object. The force is generally a function of the position x and the velocity $v = \dfrac{dx}{dt}$, and sometimes depends on t. (We will continue to postpone consideration of systems that depend explicitly on t to a later chapter.) In order to specify a single solution $x(t)$ we must specify initial values for both $x(0)$ and for $\dfrac{d}{dt}x(0)$.

We will always convert higher order differential systems into first order systems by introducing extra state variables to represent derivatives. In case of Newton's law we can use both the position and the velocity v to represent the state of the system. In this case we can replace the second order system (4.9), where x is the position vector, by the first order system

$$\frac{d}{dt}\begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} v \\ \frac{1}{m}F \end{bmatrix}. \tag{4.10}$$

An equivalent formulation is to represent the system in terms of the position and the *momentum* mv.

One of the very first problems encountered in elementary physics is the following example.

Example 4.7. This is a model of an object moving in one dimension, under the influence of a spring. The position x is the displacement of the object from its rest position (where the spring is neither stretched nor compressed) and the force exerted by the spring is proportional to the displacement and acts to move the object back towards its rest position. We assume unit mass, so $m = 1$, and we write, traditionally, k^2 for the spring constant, so the equations are

$$\frac{d}{dt}\begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} v \\ -k^2 x \end{bmatrix} = S \cdot \begin{bmatrix} x \\ v \end{bmatrix}, \quad S = \begin{bmatrix} 0 & 1 \\ -k^2 & 0 \end{bmatrix}.$$

To solve this we first find the eigenvalues and eigenvectors: We find $\lambda_1 = ki$, $\lambda_2 = -ki$ and $v_1 = \begin{bmatrix} -i \\ k \end{bmatrix}$, $v_2 = \begin{bmatrix} i \\ k \end{bmatrix}$. As usual we form the diagonal matrix of eigenvalues

Λ and the change of basis matrix P and calculate $e^t S$. In the calculation we use Euler's formula to evaluate $e^{t\lambda_k}$:

$$\begin{aligned} e^{tS} = P \cdot e^{t\Lambda} \cdot P^{-1} &= \begin{bmatrix} -i & i \\ k & k \end{bmatrix} \cdot \begin{bmatrix} e^{ikt} & 0 \\ 0 & e^{-ikt} \end{bmatrix} \cdot \begin{bmatrix} -i & i \\ k & k \end{bmatrix}^{-1} \\ &= \begin{bmatrix} -i & i \\ k & k \end{bmatrix} \cdot \begin{bmatrix} \cos(kt) + i\sin(kt) & 0 \\ 0 & \cos(kt) - i\sin kt \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{2}i & \frac{1}{2k} \\ -\frac{1}{2}i & \frac{1}{2k} \end{bmatrix} \\ &= \begin{bmatrix} \cos(kt) & \frac{1}{k}\sin(kt) \\ -k\sin(kt) & \cos(kt) \end{bmatrix} \end{aligned}$$

Hence the flow is given by

$$\begin{bmatrix} x(t) \\ v(t) \end{bmatrix} = \begin{bmatrix} \cos(kt) & \frac{1}{k}\sin(kt) \\ -k\sin(kt) & \cos(kt) \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ v_0 \end{bmatrix} = \begin{bmatrix} x_0\cos(kt) + \frac{1}{k}v_0\sin(kt) \\ -kx_0\sin(kt) + v_0\cos(kt) \end{bmatrix}.$$

This kind of solution in general is called a *simple harmonic oscillator*.

Notice that every solution in Example 4.7 is periodic, with period $\dfrac{2\pi}{k}$. Hence any solution will repeat forever. This is an idealized model; in the real world there are always *dissipative* forces at work. These forces include such things as friction, air resistance, metal fatigue, etc., and always work to retard the motion of a physical system, converting its energy into heat. Here is a modified spring model that gives one approach for taking such forces into account.

EXAMPLE 4.8. In some cases a first approximation to frictional forces is a force which opposes the motion and is proportional to the velocity. We modify Example 4.7 by adding term $-rv$ to the force, where r is a positive constant. Hence our differential equation becomes

$$\frac{d}{dt}\begin{bmatrix} x \\ v \end{bmatrix} = \begin{bmatrix} v \\ -k^2x - rv \end{bmatrix} = T \cdot \begin{bmatrix} x \\ v \end{bmatrix}, \quad T = \begin{bmatrix} 0 & 1 \\ -k^2 & -r \end{bmatrix}.$$

It turns out that the behavior of this system is different for different values of the parameters, so we give a numeric example: Let $k = .5$ and $r = .1$. We then calculate the eigenvalues as $\lambda = -0.05 + 0.497i$ and $\overline{\lambda} = -0.05 - 0.497i$; the corresponding eigenvectors are $z = \begin{bmatrix} -0.05 - 0.497i \\ .25 \end{bmatrix}$ and $\overline{z} = \begin{bmatrix} -0.05 + 0.497i \\ .25 \end{bmatrix}$. Rather than finish solving for $x(t)$ (which will be messy) we consider the form of the solution. Eventually $x(t)$ appears as a linear combinations of the complex exponentials $e^{\lambda t}$ and $e^{\overline{\lambda}t}$. Using Euler's formula, we can replace these complex exponentials with $e^{-.05t}\left(\cos(.497t) \pm \sin(.497t)\right)$, and these are in turn linear combinations

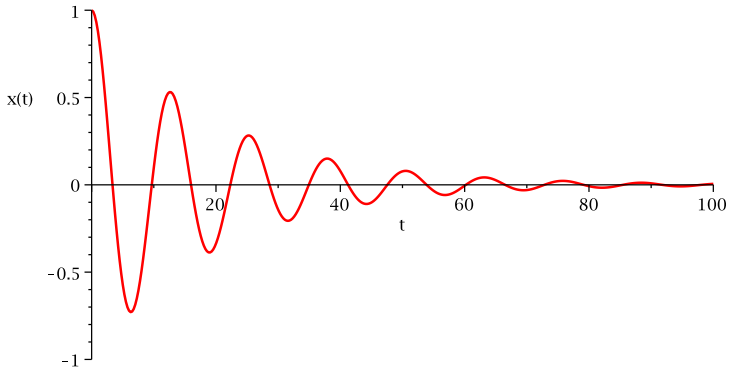


Figure 4.1: $x(t) = c_1 e^{-.05t} \cos(.497t)$

of $e^{-.05t}\cos(.497t)$ and $e^{-.05t}\sin(.497t)$. When we are all finished, there will only be real coefficients, since $x(t)$ is real. Hence

$$x(t) = c_1 e^{-.05t}\cos(.497t) + c_2 e^{-.05t}\cos(.497t) = e^{-.05t}\left(c_1\cos(.497t) + c_2\sin(.497t)\right)$$

for some constants c_1 and c_2. The velocity $v(t)$ will also have this general form; but we can calculate it more quickly from $x(t)$ by differentiating, since $v = \dfrac{dx}{dt}$.

We can now see how $x(t)$ behaves as time increases: It is the product of a periodic function, $c_1\cos(.497t) + c_2\sin(.497t)$, with period $\dfrac{2\pi}{0.497}$, times an exponential $e^{-.05t}$ which converges to 0 as $t \to +\infty$. This system is an example of a *damped harmonic oscillator*: In each period $x(t)$ comes back almost to its starting position, but with a smaller value. For large t the object is essentially stationary at the rest position of the spring. See Figure 4.1 for a sample.

4.3. Phase portraits

A graph like Figure 4.1 can help to visualize a dynamical system, but it shows only a small part of the picture. Usually we can get considerable insight into the overall behavior of a dynamical system by looking at another type of graph.

Suppose we have a vector differential equation $\dfrac{dx}{dt} = f(x)$. This defines a flow F^t as in the one-dimensional case: $F^t(x_0)$ is equal to $x(t)$ where $x(t)$ is the solution of the differential equation which satisfies the initial value condition $x(0) = x_0$. The existence of such a flow requires an m-dimensional version of the Existence and Uniqueness Theorem. We will postpone discussion of this until the next chapter; it

is valid for all systems that we will study. The analogs of Proposition 2.3 and 2.6 are also true.

If x_1 is a fixed state vector then we define the *orbit* or *trajectory* passing through x_1 to be the set of all points that are connected to x_1 by the flow. That is, x_2 is on the orbit through x_1 if $F^{t_1}(x_1) = x_2$ for some time value t_1. We note some general properties of orbits:

PROPOSITION 4.9. *For any state vectors x_1 and x_2:*

(a) *x_1 is on the orbit through x_1.*
(b) *If two orbits intersect then they are the same set.*

PROOF. x_1 is on the orbit through x_1, since $F^0(x_1) = x_1$.

For the second part, suppose that the orbits through x_1 and y_1 have some state in common – say z. Then if y_2 is any point in the orbit through y_1 we can reach y_2 from x_1 as follows: Since z is on the orbit through x_1 then $F^t(x_1) = z$ for some t. Since z is on the orbit through y_1 then $F^s(y_1) = z$ for some s. Since y_2 is on the orbit through y_1 then $F^u(y_1) = y_2$ for some u. Now use Proposition 2.3, twice:

$$F^{u-s+t}(x_1) = F^{u-s}(F^t(x_1)) = F^{u-s}(z) = F^{u-s}(F^s(y_1)) = F^{u-s+s}(y_1) = F^u(y_1) = y_2.$$

This means that y_2 is on the orbit through x_1. Since y_2 was an arbitrary point of the orbit through y_1 this means that the orbit through x_1 contains the orbit through y_1. If we repeat this argument, starting with y_1 , we see that the orbit through y_1 contains the orbit through x_1. So the two orbits are the same set. □

Note: You may have seen this kind of argument before. It uses the relation $x_1 \sim x_2$ defined by the condition that $F^t(x_1) = x_2$ for some t. It follows from Proposition 2.3 that this relation is an *equivalence relation*, and the orbits are the *equivalence classes*.

In practical terms, all that Proposition 4.9 says is that the curves which are described parametrically in state space by the solutions of the differential equation are the orbits. A description of all the orbits is called the *phase portrait* of the dynamical system.

We start with a simple example. Suppose $\dfrac{dx}{dt} = Hx$ where $H = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$. Then the solutions are given by $x = e^{tH}x_0$, and, since H is diagonal, we have $e^{tH} = \begin{bmatrix} e^t & 0 \\ 0 & e^{-t} \end{bmatrix}$. Hence, if $x_0 = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$, the solutions are given by

$$x_1 = c_1 e^t, \quad x_2 = c_2 e^{-t}.$$

The orbit through $(c_1, c_2) = (1, 1)$ is given parametrically as $x_1 = e^t$, $x_2 = e^{-t}$. We can put this into a more recognizable form by eliminating t between the two

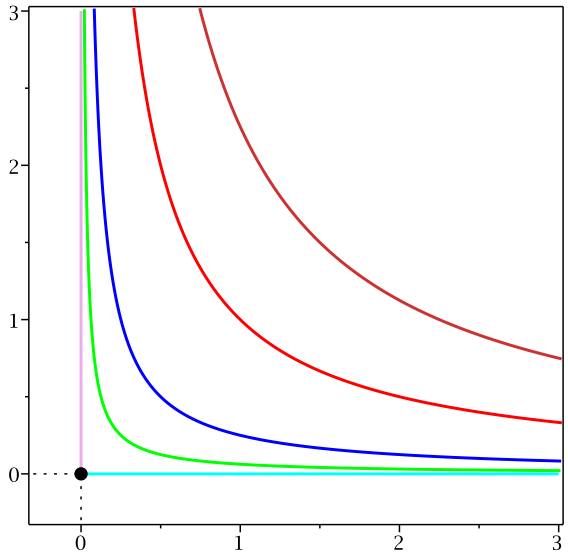


Figure 4.2: $x' = Hx$, $H = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$

equations. If we multiply the equations together then the exponentials will cancel, leaving us with $x_1 x_2 = 1$, which we recognize as a hyperbola, asymptotic to the x_1 and x_2 axes. Notice, though, that a hyperbola has two branches (in this case, in the first and fourth quadrants), but the orbit is just the branch of the hyperbola that passes through $(1, 1)$.

We might ask for the orbit through the point $(2, .5)$, but, since this is on the hyperbola through $(1, 1)$, we do not expect a new orbit. To see this algebraically, note that the orbit with $(c_1, c_2) = (2, .5)$ is given parametrically as $x_1 = 2e^t$, $x_2 = .5e^{-t}$, and if we eliminate t by multiplying the equations, we obtain $x_1 x_2 = 2 \cdot .5 = 1$, so we still have the same hyperbola.

On the other hand, the orbit through $(.5, .5)$ is given parametrically as $x_1 = .5e^t$, $x_2 = .5e^{-t}$, and if we eliminate t between these equations we are left with $x_1 x_2 = .25$, which is a different hyperbola. Again, the orbit is just the branch in the first quadrant.

These orbits, along with the orbits through $(.25, .25)$ and $(1.5, 1.5)$, are shown in Figure 4.2

In fact, three other orbits are shown in Figure 4.2.

First, if the initial point is $(c_1, c_2) = (0, 0)$ then the parametric equations become $x_1 = 0 \cdot e^t = 0$, $x_2 = 0 \cdot e^{-t}$, so $x_1(t) = 0$, $x_2(t) = 0$ for all t. That is, this orbit consists

only of the point $(0,0)$. This is a special kind of orbit, called a *stationary point* (or steady state, as in Chapter 2 When referring to the differential equation $\frac{dx}{dt} = f(x)$ such a point is usually called an *equilibrium point*. As in Chapter 2, any such points can be found by solving the equation $f(x) = 0$. In all our examples in this chapter, $f(x) = Mx$ for some matrix M, so the origin will always be an equilibrium point. There may be other equilibrium points. In the next chapter we will look at systems where $f(x)$ is more complicated than just multiplication by a matrix.

Another orbit shown in figure 4.2 is the positive x_1 axis. This corresponds to the initial condition $(c_1, c_2) = (1, 0)$, which leads to the parametric equations $x_1 = e^t$, $x_2 = 0 \cdot e^{-t} = 0$. This represents the positive x_1 axis since x_2 is 0 for all t, while $x_1 = e^t$ ranges over all positive real numbers as t ranges over $(-\infty, \infty)$. Similarly, the positive x_2 axis is an orbit.

The same kind of argument establishes the negative x_1 and x_2 axes as orbits.

Note: These orbits along the axes *do not contain* the origin, since the origin is another orbit and different orbits are disjoint.

When we look at orbits we lose information about the time dependence of the solutions. We can't do much about this, but we can at least indicate in which direction the variable point $(x_1(t), x_2(t))$ moves along the orbit. This can be done by differentiating the parametric equations for the orbit; the result, written as a vector, points in the direction of motion of $(x_1(t), x_2(t))$. For example, the hyperbola through $(1, 1)$ has parametric equations $x_1 = e^t$, $x_2 = e^{-t}$, and if we differentiate this we obtain $\frac{dx_1}{dt} = e^t$, $\frac{dx_2}{dt} = -e^{-t}$. Then, for example, the point $(1, 1)$ corresponds to $t = 0$ and we have, at this point, the vector $(e^0, -e^{-0}) = (1, -1)$. Hence the point $(x_1(t), x_2(t))$ is moving along the hyperbola in this direction when $t = 0$. We conclude that the point $(x_1(t), x_2(t))$ moves from top to bottom (and left to right) along the hyperbola. This is true when $t = 0$, as we just calculated, and, more generally, for all t.

There is an easier way to see the direction of an orbit: We just calculated the vector with coordinates $\frac{dx_1}{dt}$ and $\frac{dx_2}{dt}$, working with a solution curve as it passes through the point (x_1, x_2). But this is just, in vector terms, $\frac{dx}{dt}$. However, the differential equation is in the form $\frac{dx}{dt} = f(x)$, so we just need to calculate $f(x)$ to find the derivative vector of the solution curve through the point (x_1, x_2). For example, we could calculate the derivative at the point $(1, 1)$ by this method, since $f\left(\begin{bmatrix}1\\1\end{bmatrix}\right) = H \cdot \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1\\-1\end{bmatrix}$.

Using this method, it is easy to see that (x_1, x_2) moves along the second hyperbola in Figure 4.2 in the same direction, from top to bottom (and left to right). The motion along the positive x_1 axis is from left to right, and the motion along the positive x_2 axis is from top to bottom. The origin is an equilibrium point, so it doesn't move at all.

We can also see some limiting behavior. Points on the hyperbolic orbits in the first quadrant go to $+\infty$, asymptotic to the x_1 axis, as $t \to +\infty$, and to $+\infty$, asymptotic to the x_2 axis, as $t \to -\infty$. Points on the positive x_2 axis converge to 0 as $t \to +\infty$, and go to $+\infty$ as $t \to -\infty$. The same kind of thing is true along the positive x_1 axis: points converge to the origin as $t \to -\infty$ and to $+\infty$ as $t \to +\infty$. This is consistent with Proposition 2.6, which says that if a trajectory approaches a limit point in the domain of f then that limit point is an equilibrium point.

Before looking at some more examples we summarize a few facts about phase portraits, in terms of the general equation $\dfrac{dx}{dt} = f(x)$:

(1) Equilibrium points are orbits; they correspond to solutions of the algebraic equation $f(x) = 0$.
(2) The vector $f(x_1)$ points in the direction of motion along the orbit passing through the point x_1.
(3) If a point on an orbit has a limit point x_0 in the domain of f then x_0 is an equilibrium point (assuming the necessary conditions for the Existence and Uniqueness Theorem at x_0).
(4) Orbits are disjoint.

The next example is the "two cities" example 4.5. Several orbits, and the directions of the vector field Ax, are shown in Figure 4.3. This is similar to the picture in Figure 4.2, but the asymptotes of the hyperbolas are at an angle. In fact, the asymptotes are the lines through the eigenvectors v_1 and v_2. The origin is an equilibrium point, since $f(x) = Ax$ where A is a matrix, so $f(0) = 0$. The half lines through the asymptotes are orbits. This is true true for any differential equation of the form $\dfrac{dx}{dt} = Mx$, since, if v is an eigenvector of M corresponding to the eigenvalue λ, then $e^{\lambda t}v$ is a solution of the differential equation. Note that this only gives *positive* multiples of the eigenvector, so the orbit is a half-line. The other half of the line through v is also an orbit, since $-v$ is also an eigenvector, so the set of *positive*multiples of $-v$ is an orbit.

From Figure 4.3 we see that all population vectors in the first quadrant move to be asymptotic to the first eigenvector $v_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ as $t \to +\infty$, as we discovered in our earlier analysis of (4.7).

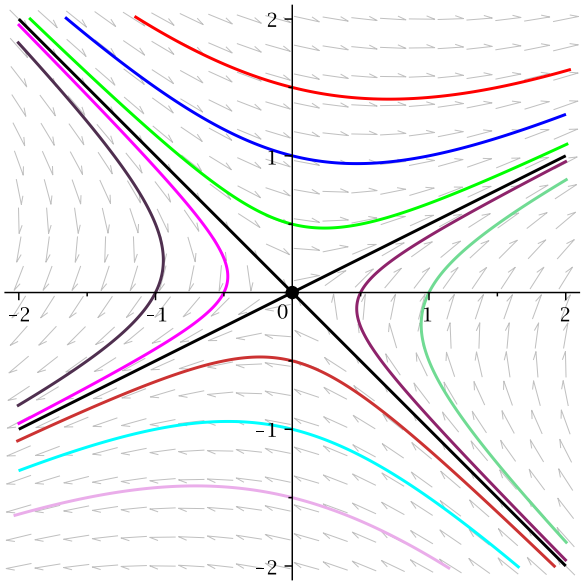


Figure 4.3: $x' = Ax$, $A = \begin{bmatrix} .01 & .04 \\ .02 & -.01 \end{bmatrix}$

Equilibrium points that "look like" Figure 4.2 are called *saddle points*. We will explain exactly what this means in the next chapter, but for now we can characterize this picture by the condition that the eigenvalues are real, with one positive and one negative.

The next example is our second "two cities" example, (4.8). Figure 4.4 shows the equilibrium at the origin, and the straight line orbits along the half lines through the eigenvectors. In this example, all orbits converge to the origin as $t \to +\infty$, as we discovered in our earlier analysis of (4.8). As in that analysis, we see that all population vectors in the first quadrant eventually turn and approach the origin asymptotically to the positive line through the eigenvector $v_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$. In fact, all orbits do this except the origin and the two straight line orbits corresponding to the other eigenvector, $v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

Using the definitions in section 2.4, we see that the origin is a *sink* in Figure 4.4. If we reverse the arrows in Figure 4.4 then we will have an example of a *source*.

The spring model, Example 4.7, has very different solutions from the population models. The eigenvectors are purely imaginary, and the solutions are periodic. The

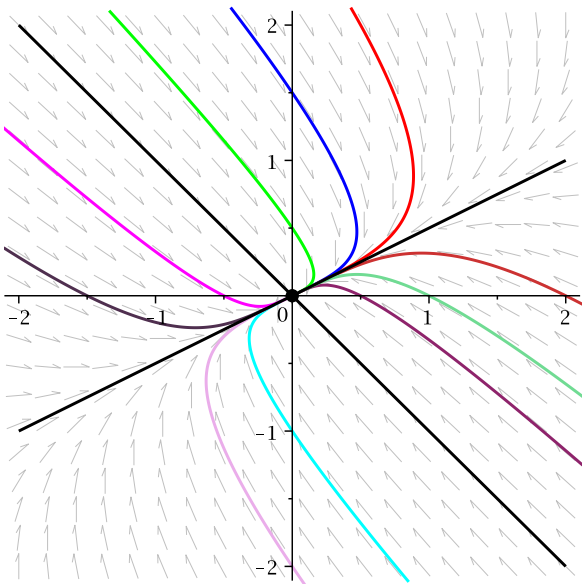


Figure 4.4: $x' = Bx$, $B = \begin{bmatrix} -.02 & .02 \\ .01 & -.03 \end{bmatrix}$

phase portrait is shown in Figure 4.5, with the parameter $k = .5$. The origin is still an equilibrium, but there are no straight-line solutions, since the eigenvectors are not real. If $(x(t), v(t))$ is a solution then both $x(t)$ and $v(t)$ are periodic, with period $2\pi/k = 4\pi$. This means that the solution point $(x(t), v(t))$ will return to exactly its starting position after 4π time units; and then it will simply retrace its path. In other words, the orbit is a closed curve. This is the only way a solution curve can ever intersect itself; if it did anything else than retrace its steps then it would violate the uniqueness theorem. e previously calculated that the solution with initial condition $x(0) = x_0$, $v(0) = 0$ is $x(t) = x_0 \cos(kt) = x_0 \cos\left(\frac{1}{2}t\right)$, $v(t) = -kx_0 \sin(kt) = -\frac{1}{2}x_0 \sin\left(\frac{1}{2}t\right)$. Then

$$\frac{x^2}{x_0^2} + \frac{v^2}{x_0^2/4} = \cos^2\left(\tfrac{1}{2}t\right) + \sin^2\left(\tfrac{1}{2}t\right) = 1.$$

In other words, the orbit through $(x_0, 0)$ is an ellipse, with semi-major axes $|x_0|$ and $\frac{1}{2}|x_0|$.

An equilibrium is sometimes called a *center* if all nearby orbits are closed orbits.

Our last example is the damped oscillator of Example 4.8, with the parameters $k = .5$, $r = .1$. Again the origin is an equilibrium. In our previous analysis we saw

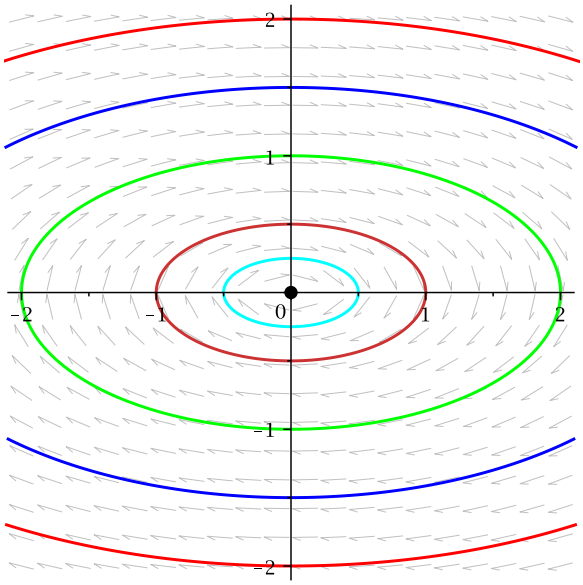


Figure 4.5: $x' = Sx$, $S = \begin{bmatrix} 0 & 1 \\ -k^2 & 0 \end{bmatrix}$, $k = .5$

that all solutions converge to the origin as $t \to +\infty$, so this is another example of a sink. However, the eigenvalues are not real, and this means that the solutions are given by periodic functions times an exponential factor which converges to 0. This forces the solutions to spiral around the origin as they converge to it, as shown in Figure 4.6.

In general, for a vector differential equation of the form $\dfrac{dx}{dt} = Ax$, the origin will always be an equilibrium point, since $A0 = 0$. (There may be other equilibria, corresponding to other solutions of $Ax = 0$.) The analog of Proposition 3.4 in the setting of differential equations is:

PROPOSITION 4.10. *For the dynamical system defined by* $\dfrac{dx}{dt} = Ax$*:*

(a) 0 *is a sink if and only if all eigenvalues have real part less than* 0*.*
(b) 0 *is a source if and only if all eigenvalues have real part greater than* 0*.*

PROOF. The flow of the system is $F^t(x) = e^{tA}x$. Consider this for $t = 1$, so $F^1(x) = e^A x$. According to Proposition 3.4 the origin will be a sink for F^1 if and only if all eigenvalues of $M = e^A$ have absolute value less than 1. The eigenvalues μ of M have the form $\mu = e^\lambda$ where λ is an eigenvalue of A. Since $|\mu| = \left|e^\lambda\right| = e^{\mathrm{Re}(\lambda)}$,

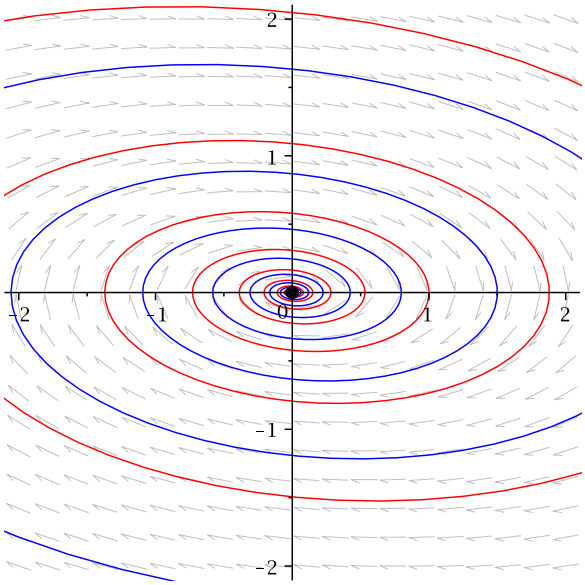


Figure 4.6: $x' = Tx$, $T = \begin{bmatrix} 0 & 1 \\ -k^2 & -r \end{bmatrix}$, $k = .5$, $r = .1$

the condition that all $|\mu| < 1$ translates to the condition that all $\mathrm{Re}(\lambda) < 0$. So 0 will be a sink for F^1 if and only if all eigenvalues of A have real part less than 0. This means that $F^n(x) \to 0$ as $t \to +\infty$ for all initial values x that are near 0 (in fact, this is true for *all* initial vectors). Hence 0 is a sink for the flow of the differential equation.

(Technically, this argument only shows that $e^{At}x \to 0$ as $t \to +\infty$ for *integer* values of t. The case of general real values of t requires a bit more work.)

The argument for sources is entirely similar. □

The fixed point 0 is called *hyperbolic* if $\mathrm{Re}\,\lambda \neq 0$ for all eigenvalues λ. So sinks and sources are examples of hyperbolic fixed points. Hyperbolic fixed points that are not sources or sinks are called *saddles*.

In two dimensions a hyperbolic sink is called a *spiral sink* if it is a sink and, in addition, the eigenvalues of A are not real. A spiral source is defined similarly. In higher dimensions this spiraling behavior is more complicated, and harder to visualize.

Exercises

4.1. Calculate e^A:

(a) $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$. Use the series.

(b) $A = \begin{bmatrix} a & 1 & 0 \\ 0 & a & 1 \\ 0 & 0 & a \end{bmatrix}$. Use the fact that $A = aI + A_1$ where A_1 is the matrix in part (a).

(c) $A = \begin{bmatrix} 1 & 1 \\ 6 & 0 \end{bmatrix}$. Use eigenvalues and eigenvectors.

4.2. Let $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$.

(a) Show that A and B do not commute.

(b) Calculate e^A, e^B and e^{A+B}. You can just use the series for the first two (since almost all the terms are 0). For e^{A+B} you can use eigenvalues and eigenvectors.

(c) Show that $e^{A+B} \neq e^A e^B$.

4.3. (a) $e^{i\pi/2} =$

(b) $e^{1-i\pi/6} =$

(c) $e^{\ln(2)+i\pi} =$

(d) Solve for the real numbers x and y: $e^{x+iy} = 1 + i$. [First write $1 + i$ in polar form.]

4.4. Let $A = \begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix}$. Find the solution to $\dfrac{dx}{dt} = Ax$, $x(0) = x_0 = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ by the following method:

(a) Find the eigenvalues and eigenvectors.

(b) Find the change of basis matrix P and its inverse.

(c) Find the matrix exponential e^{tA} by calculating $Pe^{t\Lambda}P$ where Λ is the diagonal matrix with the eigenvalues on the diagonal.

(d) Now calculate $x(t) = e^{tA}x_0 = e^{tA} \cdot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$

4.5. Let $A = \begin{bmatrix} 1 & 3 \\ 2 & 0 \end{bmatrix}$. Find the solution to $\dfrac{dx}{dt} = Ax$, $x(0) = x_0 = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ by the following method:

(a) Find the eigenvalues, λ_1 and λ_2, and the eigenvectors v_1 and v_2.

(b) x_0 is a linear combination of v_1 and v_2, say $x_0 = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = c_1 v_1 + c_2 v_2$. You know the vectors v_1 and v_2, so this is a system of equations for the unknowns c_1 and c_2 in terms of the "known" constants a_1 and a_2. Solve for c_1 and c_2 by row reduction or another method.

(c) Now $x(t) = c_1 e^{\lambda_1} t v_1 + c_2 e_{\lambda_2 t} v_2$. You know $v_1, v_2, \lambda_1, \lambda_2$ and you have formulas for c_1 and c_2 in terms of a_1 and a_2. Put it all together.

4.6. Derive (4.3) by following this outline:

(a) Let $A = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$ and $B = \begin{bmatrix} 0 & -b \\ b & 0 \end{bmatrix}$. Show that $e^C = e^A \cdot e^B$.

(b) Show that $e^A = e^a I$.

(c) Let $J = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. The powers J^n follow a periodic pattern. What is it?

(d) Since $B = bJ$ we have $B^n = b^n J^n$. Plug this into the series for e^B and use the same idea as in the proof of Euler's formula to show that $e^B = \begin{bmatrix} \cos(b) & -\sin(b) \\ \sin(b) & \cos(b) \end{bmatrix}$.

(e) Finish.

4.7. For each of the following differential equations, determine whether the equilibrium at the origin is a saddle, a source, a sink, a center, or none of these. In case it is a source or sink, determine whether it is a spiral source or sink.

(a) $\dfrac{dx}{dt} = Ax$, $A = \begin{bmatrix} 3 & -2 \\ 1 & 0 \end{bmatrix}$.

(b) $\dfrac{dx}{dt} = Ax$, $A = \begin{bmatrix} 1 & 3 \\ 2 & 0 \end{bmatrix}$.

(c) $\dfrac{dx}{dt} = Ax$, $A = \begin{bmatrix} 2 & -5 \\ 1 & 0 \end{bmatrix}$.

(d) $\dfrac{dx}{dt} = Ax$, $A = \begin{bmatrix} 2 & -1 \\ -4 & 2 \end{bmatrix}$.

(e) $\dfrac{dx}{dt} = Ax$, $A = \begin{bmatrix} 1 & -5 \\ 2 & -1 \end{bmatrix}$.

4.8. This exercise demonstrates some of the borderline cases of phase portraits for linear systems. For each of the following, prepare a sketch of the phase portrait for $\dfrac{dx}{dt} = Ax$ and indicate how it compares to the examples in Figures 4.2 – 4.6.

(a) Repeated eigenvalues: $A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$. This doesn't require Maple: analyze the straight-line solutions.

(b) Zero eigenvalue: $A = \begin{bmatrix} -2 & 1 \\ 2 & -1 \end{bmatrix}$. First find the equilibrium points.

(c) Repeated eigenvalues: $A = \begin{bmatrix} -3 & 4 \\ -1 & 1 \end{bmatrix}$. Use Maple.

(d) Repeated zero eigenvectors: $A = \begin{bmatrix} 2 & 4 \\ -1 & -2 \end{bmatrix}$.

4.9. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, $T = \operatorname{tr} A = a + d$ (the trace of A) and $D = \det A = ad - bc$ (the determinant of A).

(a) Check that the characteristic polynomial of A is $\lambda^2 - T\lambda + D$.
(b) Use the quadratic formula to solve for the eigenvalues in terms of T and D.
(c) If λ_1 and λ_2 are the eigenvalues then the characteristic equation factors as $(\lambda - \lambda_1)(\lambda - \lambda_2)$. Multiply this out and compare with the formula in (a) to conclude that T is the sum of the eigenvalues and D is their product. (This fact holds for $m \times m$ matrices, for any m.)

4.10. Using the results of Exercise 4.9 determine the conditions on T and D that lead to

(a) Periodic orbits (eigenvalues purely imaginary, not 0).
(b) A saddle (eigenvalues are real, one positive, one negative).
(c) A sink (both eigenvalues have real part less than 0).
(d) A spiral sink (a sink with non-real eigenvalues).

4.11. "Two city" population models like Example 4.5 have the form $\dfrac{dx}{dt} = Ax$, where A is 2×2 and the terms off the diagonal (corresponding to migration) are non-negative.

(a) Explain why such a system must have real eigenvalues. You might want to use the results of Exercises 4.9 and/or 4.10.
(b) This is harder: Explain why e^{tA} is a positive matrix for all $t > 0$.

4.12. In Example 4.8 the choice of $k = .5$ and $r = 0$ leads to a center, while $k = .5$ and $r = .1$ leads to a spiral sink. In the following, assume $k = .5$.

(a) Find a formula for the eigenvalues of the system in terms of r.
(b) Show that there is a sink at the origin for all $r > 0$.
(c) There is a value r_c so that there is a spiral sink for $0 < r < r_c$, and a non-spiral sink for $r \geq r_c$. Find r_c.
(d) The solution, in case of a spiral sink, involves trig functions times exponential functions. What is the period of the trig functions, as a function of r?

CHAPTER 5

Non-linear systems

5.1. Orbits and invariants

In this chapter we will look at phase portraits of a number of non-linear systems, mostly in two dimensions. In this section we look at a useful tool for analyzing phase portraits.

Our first example is the following linear system of equations:

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = f(x,y) = \begin{bmatrix} -s & 0 \\ 0 & r \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -sx \\ ry \end{bmatrix} \qquad (r>0,\ s>0) \tag{5.1}$$

In other words,

$$\frac{dx}{dt} = -sx \qquad \frac{dy}{dt} = ry.$$

It is simple to solve these for x and y, since each equation is essentially the same as Example 2.1:

$$x(t) = x_0 e^{-st},\ y(t) = y_0 e^{rt},\ \text{where } x(0) = x_0,\ y(0) = y_0.$$

Therefore we can describe the orbit of the system passing through (x_0, y_0) as the curve with parametric equations $x(t) = x_0 e^{-st},\ y(t) = y_0 e^{rt}$. In a sense, this describes the complete phase portrait.

However, in most cases it is either impossible or uninformative to find such parametric equations for the orbits. Our first job is to find an alternative description of the phase portrait for equation (5.1), *without* using the solution of the system.

The first step in investigating a phase portrait is to locate the equilibrium points. This amounts to solving $f(x,y) = 0$. In our example, this means that $f(x,y) = \begin{bmatrix} -sx \\ ry \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and we solve this to get $x = 0$, $y = 0$. That is, the origin is the only equilibrium for this system. Hence the origin is a one-point trajectory.

The next step is to look for "special" types of orbits. This is not always possible, and there are many different notions of "special" that might apply here. In some cases we can recognize special orbits from the form of the right hand size, $f(x,y)$. This is because, if (x_1, y_1) is any point on a parametric curve, then the tangent direction of the curve is given by $\begin{bmatrix} x' \\ y' \end{bmatrix}$ evaluated at that point. On the other hand,

the parametric curve is an orbit if it satisfies the differential equation at every point (x_1, y_1) on the curve; that is, $\begin{bmatrix} x' \\ y' \end{bmatrix} = f(x_1, y_1)$. Putting these together, we see that the geometric criterion for a curve to be an orbit is that, at each point (x_1, y_1) on the curve, the tangent direction of the curve is given by $f(x_1, y_1)$.

There is an exception to this criterion: If $\begin{bmatrix} x' \\ y' \end{bmatrix} = 0$ then there is no well-defined tangent direction, so we can only apply this reasoning at points where $f(x_1, y_1) \neq 0$. But $f(x_1, y_1) = 0$ means that (x_1, y_1) is an equilibrium point, and hence the single point (x_1, y_1) is an entire orbit.

In summary:

PROPOSITION 5.1. *Suppose C is a curve and the vector $f(x_1, y_1)$ is tangent to C whenever (x_1, y_1) is on C. Then any orbit that starts at a point of C remains on C until it reaches an equilibrium point (which it cannot touch) or it reaches an endpoint of C.*

For example, in our example (5.1), consider points $(x, 0)$ on the x axis. Then $f(x, 0) = \begin{bmatrix} -sx \\ 0 \end{bmatrix}$. Since this is tangent to the x axis we can conclude that the x axis is made up of orbits: The origin is an equilibrium point, so it splits the x axis into the positive and negative x axes, and each of these is an orbit. Similarly, for points $(0, y)$ on the y axis, $f(0, y) = \begin{bmatrix} 0 \\ ry \end{bmatrix}$ is tangent to the y axis, and we conclude that the positive and negative y axes are orbits.

Of course there are infinitely many orbits of the dynamical system and we've only identified five of them. We can write the rest as parametric curves determined by the solutions, as $x = x(t)\, y = y(t)$. However, in many cases it is possible to find the orbits directly, as curves involving only x and y. To do this, first apply the chain rule, so $\dfrac{dy}{dt} = \dfrac{dy}{dx} \cdot \dfrac{dx}{dt}$, From this we calculate $\dfrac{dy}{dx} = \dfrac{dy/dt}{dx/dt} = \dfrac{y'(t)}{x'(t)}$. This does not look too useful, since we need to calculate $x'(t)$ and $y'(t)$, and we are trying to avoid calculating $x(t)$ and $y(t)$. However, $\begin{bmatrix} x' \\ y' \end{bmatrix} = f(x, y)$ and we do not need to calculate $x(t)$ and $y(t)$ in order to use $f(x, y)$.

Here are the steps to apply this idea to example (5.1):

(1) Since $f(x, y) = \begin{bmatrix} -sx \\ ry \end{bmatrix}$, we have $x' = -sx$ and $y' = ry$, and so $\dfrac{dy}{dx} = \dfrac{y'}{x'} = \dfrac{ry}{-sx}$.

(2) Now $\dfrac{dy}{dx} = \dfrac{ry}{-sx}$ is a differential equation for y as a function of x, so we can try to solve it by separation of variables. First write it in differential form:

$$dy = -\frac{ry}{sx}\,dx.$$

(3) Separate variables:

$$\frac{dy}{ry} = -\frac{dx}{sx},$$

and rewrite this with 0 on the right:

$$\frac{dx}{sx} + \frac{dy}{ry} = 0.$$

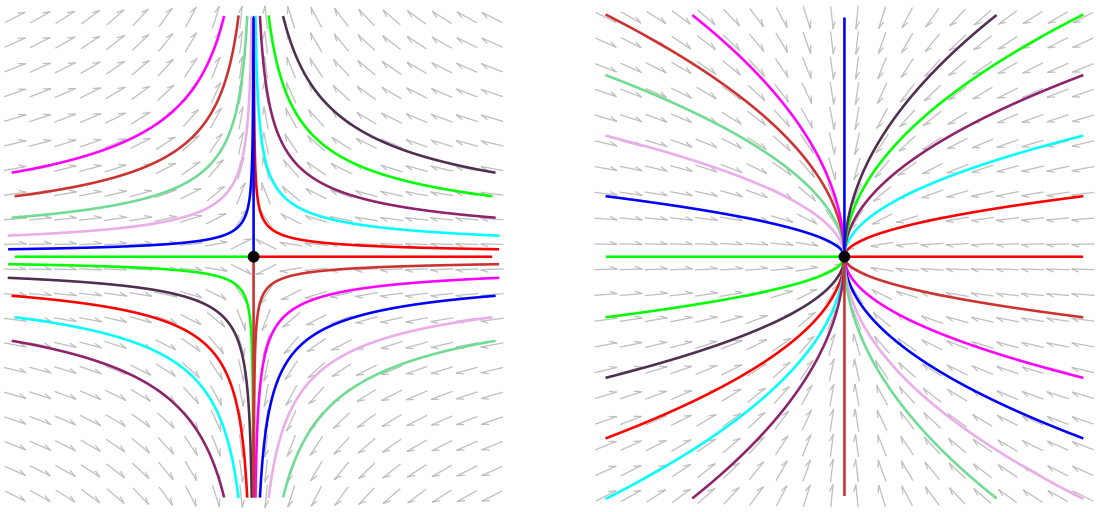
(4) Integrate both sides. The terms on the left can both be integrated, since *each* involves *only one variable*. Integrating on the right gives 0, but there is a constant of integration since these are indefinite integrals:

$$\begin{aligned}
\int \frac{dx}{sx} + \int \frac{dy}{ry} &= C_1 \\
\frac{1}{s}\ln|x| + \frac{1}{r}\ln|y| &= C_1 \\
r\ln|x| + s\ln|y| &= rsC_1 = C_2 && \text{multiply by } rs \\
e^{r\ln|x|}e^{s\ln|y|} &= e^{c_2} = C && \text{exponentiate and use } e^{a+b} = a^a e^b \\
|x|^r\,|y|^s &= C && \text{since } e^{ab} = (e^b)^a \text{ and } e^{\ln u} = u.
\end{aligned}$$

We now have a general solution for all the orbits: $|x|^r\,|y|^s = C$ where C is a constant. The value of C is determined by any point (x_0, y_0) on the curve: just plug it in to get $C = |x_0|^r\,|y_0|^s$.

For example, suppose that $r = 1$ and $s = 2$. We will first look in the first quadrant so we can forget about the absolute values. Then the equation for the orbits is $xy^2 = C$. If we want the orbit through $(2, 2)$ then we plug in $x_0 = 2$, $y_0 = 2$ to get $C = x_0 y_0^2 = 2 \cdot 2^2 = 8$, so the orbit passing through $(2, 2)$ is given by $xy^2 = 8$. This is not a hyperbola (since a hyperbola would have the form $xy = C$) but it has the same general shape: it is asymptotic to the y axis as $t \to +\infty$ and it is asymptotic to the x axis as $t \to -\infty$. Notice that the orbit is the branch of $xy^2 = 8$ which which contains $(2, 2)$; there are points in the fourth quadrant that satisfy $xy^2 = 8$ but these are on a different branch of this "hyperbola-like" curve.

If we want the orbit through $(-2, 3)$ then we need to do the same thing, but remembering the absolute values, and we get $|x|\,|y|^2 = 18$. As long as the orbit remains in the second quadrant we will have $|x| = -x$ and $|y| = y$, so the equation becomes $(-x)y^2 = 18$, or $xy^2 = -18$. In fact, the orbit *must* remain in the second



(a) Equation (5.1), $s = 2, r = 1$

(b) Equation (5.2), $s = 2, r = 1$

Figure 5.1: A saddle and a sink, with orthogonal eigenvectors.

quadrant: It can never touch the x or y axis since the x and y axes are made up of orbits, and orbits can never touch each other.

The phase portrait for this system is shown in Figure 5.1(a).

There is another way to think of the general equation $|x|^r |y|^s = C$ for the orbits. The procedure above naturally produced this equation in the form $\varphi(x, y) = C$, where φ is the function of two variables defined by $\varphi(x, y) = |x|^r |y|^s$. This function has the following property: If (x_1, y_1) and (x_2, y_2) are two points on the *same* orbit then $\varphi(x_1, y_1) = \varphi(x_2, y_2)$. This is because both (x_1, y_1) and (x_2, y_2) satisfy the *same* equation $\varphi(x, y) = C$, so $\varphi(x_1, y_1) = C$ and $\varphi(x_2, y_2) = C$. In this sense, φ is called an *invariant* of the dynamical system, since it does not change if (x, y) moves along an orbit. Equivalent names that are used for such a function, especially in physics applications, are *conserved quantity* or *constant of the motion*.

Here is a translation of the invariance property of φ into differential equations. If φ is an invariant and $x = x(t)$, $y = y(t)$ is any solution curve then $\varphi(x(t), y(t))$ does not change as t changes, since all the points $(x(t), y(t))$ are on the same orbit. An alternative way to say this is $\dfrac{d}{dt}\varphi(x(t), y(t)) = 0$, and, by the chain rule, we can rewrite this as $\dfrac{\partial\varphi}{\partial x}\dfrac{dx}{dt} + \dfrac{\partial\varphi}{\partial y}\dfrac{dy}{dt} = 0$. Now $\dfrac{dx}{dt}$ and $\dfrac{dy}{dt}$ are given by the right hand side $f(x, y)$ of the differential equation, so we can summarize this as:

PROPOSITION 5.2. *Consider the differential equation* $\dfrac{d}{dt}\begin{bmatrix}x\\y\end{bmatrix} = f(x,y)$, *and write the vector* $f(x,y)$ *as* $\begin{bmatrix}P(x,y)\\Q(x,y)\end{bmatrix}$. *Then a differentiable function* φ *is an invariant if and only if*

$$P\frac{\partial\varphi}{\partial x} + Q\frac{\partial\varphi}{\partial y} = 0.$$

For example, in equation (5.1) we have $P = -sx$ and $Q = ry$ and $\varphi = x^r y^s$ (ignoring the absolute values). Then

$$P\frac{\partial\varphi}{\partial x} + Q\frac{\partial\varphi}{\partial y} = (-sx)\cdot(rx^{r-1}y^s) + (ry)\cdot(sx^r y^{s-1}) = -srx^r y^s + rsx^r y^s = 0.$$

Here is another example:

$$\frac{d}{dt}\begin{bmatrix}x\\y\end{bmatrix} = f(x,y) = \begin{bmatrix}-s & 0\\0 & -r\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}-sx\\ry\end{bmatrix} \qquad (r>0,\ s>0) \tag{5.2}$$

As before, the origin is the only equilibrium and the positive and negative x and y axes are orbits. The only difference between this and equation (5.1) is in the sign of y'. If you change r to $-r$ in the calculations for (5.1) you will eventually get to the invariant $\varphi(x,y) = |x|^{-r}|y|^s$. For example, if $r=1$ and $s=2$ we get (ignoring the absolute values) $\varphi(x,y) = xy^{-2}$, so the orbits are described by equations of the form $xy^{-2} = C$. For example, the orbit through $(2,2)$ has $C = x_0 y_0^{-2} = 2\cdot 2^{-2} = \frac{1}{2}$, so its equation is $xy^{-2} = \frac{1}{2}$. If we rewrite this as $x = \frac{1}{2}y^2$ we recognize it as a parabola passing through the origin and opening to the right, with axis given by the positive x axis. Of course the orbit through $(2,2)$ acnnot go through the origin (since the origin is an orbit) so the orbit through $(2,2)$ is just the part of the parabola which lies in the first quadrant, not including the origin. The other half of the parabola, in the fourth quadrant, is another orbit.

If we write the general orbit as $x = Cy^2$ then we can see the y axis orbits by setting $C = 0$ to get $x = 0$. However, we cannot choose C to produce the x axis solutions. To do this we find *another* invariant. In fact, once we have one invariant we can find many others. For example, if $\varphi(x,y)$ is an invariant then so is $e^{\varphi(x,y)}$, since if $\varphi(x,y)$ doesn't change along an orbit then $e^{\varphi(x,y)}$ can't change either. That is,

LEMMA 5.3. *If* $\varphi(x,y)$ *is an invariant and* g *is any function then* $\psi(x,y) = g(\varphi(x,y))$ *defines another invariant.*

If we apply this with $g(u) = u^{-1}$ and $\varphi(x) = xy^{-2}$ we get $\psi(x,y) = x^{-1}y^2$. (Technically we have to worry a bit about the domains, but now that we have a

candidate for an invariant we can check it via Proposition 5.2.) Now the general equation for an orbit can be written as $x^{-1}y^2 = C$, or $y^2 = Cx$. This now gives the x axis solution if we choose $C = 0$.

The phase portrait for this system is shown in Figure 5.1(b).

There is another property of constants of the motion that this example illustrates. Notice that $\varphi(x, y)$ is not defined at the origin – and neither is $\psi(x, y)$. Notice, also that the origin is a sink, since the eigenvalues $-r$ and $-s$ are negative. This is not coincidence:

PROPOSITION 5.4. *Suppose that φ is an invariant which is defined and continuous for all states near an equilibrium (x_1, y_1), and let $A = \varphi(x_1, y_1)$. If $(x(t), y(t))$ is a solution which has (x_1, y_1) as a limit (for $t \to +\infty$ or $t \to \infty$) then $\varphi(x(t), y(t)) = A$ for all t.*

PROOF. $\lim_{t\to+\infty} \varphi(x(t), y(t)) = \varphi(x(t), y(t))$, since φ doesn't change as t changes. On the other hand, by continuity, $\lim_{t\to\infty} \varphi(x(t), y(t)) = \varphi(x_1, y_1) = A$. The argument for the limit as $t \to -\infty$ is similar. □

All states near a sink or source (x_1, y_1) converge to (x_1, y_1) as $t \to +\infty$ or as $t \to -\infty$. So:

COROLLARY 5.5. *Suppose that φ is an invariant which is defined and continuous for all states near a sink or source (x_1, y_1). Then φ is constant near (x_1, y_1).*

A constant function (e.g., $\varphi(x, y) = 17$ for all values of x and y) is obviously a constant of the motion, but it is not a useful one. We want $\varphi(x, y) = C$ to determine an orbit, but this is not the case for a constant function (e.g., *every* point satisfies $\varphi(x, y) = C$ if φ is a constant function with value 17), so a constant function is useless as a constant of the motion. According to Corollary 5.5, a *useful* constant of the motion is either undefined or not continuous at any sink or source.

5.2. A predator-prey model

Our first example of a non-linear two-dimensional system is a further modification of the "rabbit island" models from chapter 2.

We will use y to represent the number of rabbits on the island. Then a very simple first approximation to modelling their population is the equation $\dfrac{dy}{dt} = ry$ where r is a positive parameter, representing the net growth rate of the rabbit population. This was covered in Example 2.1; we find that $y(t) = e^{rt}y_0$, so any positive initial population will grow exponentially fast.

On similar grounds, suppose we represent the number of foxes on the island as x. If there is nothing edible on the island (except, perhaps, carrots) then we would

expect the foxes to die off. As a first approximation to this we can use a differential equation like $\dfrac{dx}{dt} = -sx$ where s is a positive parameter. Here $-s$ represents the net growth rate of the fox population, and it is negative since we expect that the death rate will be larger than the birth rate. Solving the differential equation leads to $x(t) = e^{-st}x_0$, so any initial population of foxes will die off exponentially fast.

Now suppose both rabbits and foxes are present on the island. In this case we need to model their interactions: foxes will now find something edible on the island, so we would expect the net growth rate for foxes to rise. Of course, at the same time we would expect the net growth rate for rabbits to fall. More specifically, we replace the net growth rate r for rabbits by $r - \beta x$ where β is a positive constant. That is, the net growth rate for rabbits should decrease if foxes are present, and this decrease should be proportional to the number of foxes. On the other hand, we replace the net growth rate $-s$ for foxes by $-s + \alpha x$, where α is a positive constant, so the net growth rate for foxes will increase, and the amount of the increase is proportional to the number of rabbits.

Putting everything together we have the following model. This model was first proposed in the 1920's (independently) by Volterra and Lotka, and this and similar models are known as *Lotka-Volterra* systems.

EXAMPLE 5.6. Then the *predator-prey* system is

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = f(x,y) = \begin{bmatrix} (-s + \alpha y)x \\ (r - \beta x)y \end{bmatrix}.$$

Here x and y represent populations of the predator (foxes) and the prey (rabbits); r,s,α,β are positive parameters.

It is not possible to solve this system of equations using elementary functions. We will instead try to find out as much as possible about the phase portrait of the system.

The first thing to do is to locate the equilibrium points. To do this we set $f(x,y) = 0$ and solve for (x,y). Using the definition of $f(x,y)$ in Example 5.6, we get

$$\left\{\begin{array}{c} (-s + \alpha y)x = 0 \\ \text{and} \\ (r - \beta x)y = 0 \end{array}\right\} \text{ which is equivalent to } \left\{\begin{array}{lll} x = 0 & \text{or} & y = \dfrac{s}{\alpha} \\ & \text{and} & \\ y = 0 & \text{or} & x = \dfrac{r}{\beta} \end{array}\right\}$$

Since the combinations $x = 0$ with $x = \dfrac{r}{\beta}$ and $y = 0$ with $y = \dfrac{s}{\alpha}$ are impossible, there are only two equilibrium points:

$$(0,0) \text{ and } \left(\frac{r}{\beta}, \frac{s}{\alpha}\right).$$

We also notice that the positive and negative x axes are orbits, since, for points $(x,0)$ on the x axis, $f(x,0) = \begin{bmatrix} -sx \\ 0 \end{bmatrix}$ is tangent to the x axis, so we can apply Proposition 5.1. Similarly, the positive and negative y axes are orbits.

We have now found six trajectories of the system (the two equilibria and the four semi-axes). It is always important to locate the equilibrium points, as we shall see, but the semi-axes are orbits only because of special features of this system.

The pattern of straight line solutions through the origin is just like what we found for a saddle equilibrium, as in Figure 5.1(a). But our predator-prey system is non-linear and we can't expect the same phase portrait as for a linear system. In fact, we know that the phase portrait is different, since our system has two equilibria, but a linear hyperbolic system has either one equilibrium or, if 0 is an eigenvalue, infinitely many.

However, we can look at the equilibrium point at the origin as follows. First, rewrite $f(x,y)$:

$$f(x,y) = \begin{bmatrix} (-s+\alpha y)x \\ (r-\beta x)y \end{bmatrix} = \begin{bmatrix} -sx \\ ry \end{bmatrix} + \begin{bmatrix} \alpha xy \\ -\beta xy \end{bmatrix} = \begin{bmatrix} -s & 0 \\ 0 & r \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \alpha xy \\ -\beta xy \end{bmatrix}. \tag{5.3}$$

That is, we have written $f(x,y)$ as the sum of two terms. The first has the form $A\begin{bmatrix} x \\ y \end{bmatrix}$ where A is the constant matrix $\begin{bmatrix} -s & 0 \\ 0 & r \end{bmatrix}$; it is called the *linearization* of f at the origin. The second term involves only "higher powers" of x and y. If x and y are "small enough" then the entries in this second term, αxy and $-\beta xy$, will be much smaller than the entries $-sx$ and ry in the linearization. In this case it is natural to compare the solutions of the non-linear differential equation $\dfrac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = f(x,y)$ to the solutions of the *linearized* equation $\dfrac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = A\begin{bmatrix} x \\ y \end{bmatrix}$. We will see, in the next section, a general theorem which says, in case A is hyperbolic, that the solutions "near" an equilibrium of the non-linear system "look like" the solutions of the linearized system near the origin.

In our case the matrix $A = \begin{bmatrix} -s & 0 \\ 0 & r \end{bmatrix}$ has eigenvalues $-s$ and r, so, since neither of these has zero real part, the matrix is hyperbolic. Near the origin, the trajectories

of the predator-prey system are approximately the same as the trajectories of the linearized system $\frac{d}{dt}\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}-s & 0\\0 & r\end{bmatrix}\begin{bmatrix}x\\y\end{bmatrix}$ near the origin, and this linear system has a saddle equilibrium at the origin. This type of system was analyzed in section 5.1: Its trajectories consist of the origin, the x and y semi-axes, and "hyperbola-like" curves asymptotic to the axes, similar to the trajectories in Figure 5.1(a).

One way to see how the rest of the orbits fit together is to use an invariant for the system. So our next step is to try to find a constant of the motion for the predator-prey system. We'll follow the technique used in section 5.1:

$$\begin{aligned}
\frac{dy}{dx} &= \frac{dy/dt}{dx/dt} && \\
&= \frac{(r-\beta x)y}{(-s+\alpha y)x} && \text{since } \frac{d}{dt}\begin{bmatrix}x\\y\end{bmatrix} = \begin{bmatrix}(-s+\alpha y)x\\(-s+\alpha y)x\end{bmatrix} \\
dy &= \frac{(r-\beta x)y}{(-s+\alpha y)x}\,dx && \text{write in differential form} \\
\frac{-s+\alpha y}{y}\,dy &= \frac{r-\beta x}{x}\,dx && \text{separate variables} \\
\left(-\frac{r}{x}+\beta\right)dx + \left(-\frac{s}{y}+\alpha\right)dy &= 0 && \text{0 on the right} \\
\left(\frac{r}{x}-\beta\right)dx + \left(\frac{s}{y}-\alpha\right)dy &= 0 && \text{multiply by } -1 \\
r\ln|x| - \beta x + s\ln|y| - \alpha y &= C_1 && \text{integrate} \\
e^{r\ln|x|-\beta x+s\ln|y|-\alpha y} &= e^{C_1} = C && \text{exponentiate} \\
e^{r\ln|x|}e^{-\beta x}e^{s\ln|y|}e^{-\alpha y} &= C && \text{properties of exponents} \\
|x|^r e^{-\beta x}|y|^s e^{-\alpha y} &= C && \text{using } e^{a\ln b} = \left(e^{\ln b}\right)^a = b^a
\end{aligned}$$

So we have found a constant of the motion: $\varphi(x,y) = |x|^r e^{-\beta x}|y|^s e^{-\alpha y}$.

The first thing we notice about φ is that it is zero exactly on the x and y axes, and we have already seen that the origin and the 4 semi-axes are orbits. We concentrate on the first quadrant (which is all that is relevant in a population model), so we can ignore the absolute values. For $x > 0$ the function $x^r e^{-\beta x}$ is non-negative; it approaches 0 as $x \to \infty$; and it has only one critical point, at $x = \dfrac{r}{\beta}$, where it has its maximum. The function $y^s e^{-\alpha y}$ has a similar description, with a maximum when $y = \dfrac{s}{\alpha}$. The conserved quantity $\varphi(x,y)$ is the product of these two functions,

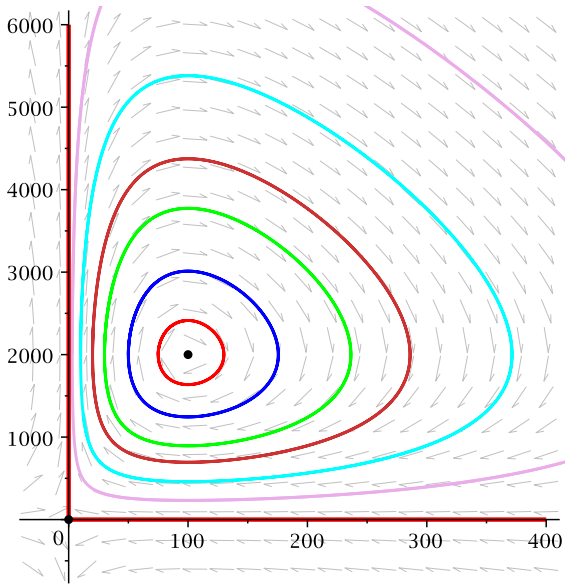


Figure 5.2: Example 5.6 with $r = .1$, $s = .2$, $\alpha = .0001$, $\beta = .001$. Time is measured in months.

so it has a strict maximum in the first quadrant at $(x_0, y_0) = \left(\frac{r}{\beta}, \frac{s}{\alpha}\right)$. Let's write $M = \varphi(x_0, y_0)$ for this maximum value. Then $\varphi(x, y) = M$ has only one solution in the first quadrant, since there is only one point where φ has its maximum, and all other values of φ are smaller than the maximum value. But $\varphi(x, y)$ is equal to M at all points of the orbit through (x_0, y_0), so this single point must be a one-point orbit.Of course, we already knew this, since $\left(\frac{r}{\beta}, \frac{s}{\alpha}\right)$ is the second equilibrium point.

The rest of the first quadrant is covered by the level curves $\varphi(x, y) = c$ with $0 < c < M$. It should be clear that these are closed curves surrounding the equilibrium at $\left(\frac{r}{\beta}, \frac{s}{\alpha}\right)$: Think of the graph of $z = f(x, y)$; over the first quadrant this surface is a "hill", and our level curves $\varphi(x, y) = c$ can be thought of as "slicing" this surface by a horizontal plane. These curves are the orbits of the dynamical system, since any orbit that starts on a curve of the form $\varphi(x, y)$ must stay there unless it reaches an equilibrium point, and there are no equilibrium points which satisfy $0 < \varphi(x, y) < M$.

The various features of the phase portrait in the first quadrant are illustrated in Figure 5.2. Since the orbits corresponding to $\varphi(x, y) = c$ for $0 < c < M$ are

closed orbits they correspond to periodic solutions. Unlike the periodic orbits in the linear case, Example 4.7, the periods depend on the orbit. It is easy to see that this must be so, since as c gets closer to 0 the orbit must pass very close to the origin. However, the vectors $f(x, y)$, for (x, y) near the origin, are very small, and so the point $(x(t), y(t))$ moves very slowly while it remains near the origin. This means that the period of the closed orbits must become very large as the orbits get close to the origin.

There is no expression for the period in terms of c, but in the next section we shall see that the period is approximately $\dfrac{2\pi}{\sqrt{rs}}$ for orbits close to the equilibrium point $\left(\dfrac{r}{\beta}, \dfrac{s}{\alpha}\right)$.

5.3. Linearization

The first step in analyzing a phase portrait is usually to find and classify the equilibrium points. Suppose the differential equation is written in the vector form $\dfrac{dx}{dt} = f(x)$. It is relatively straightforward to find the critical points: we just write $f(x) = \begin{bmatrix} f_1(x, y) \\ f_2(x, y) \end{bmatrix}$ and solve the two equations $f_1(x, y) = 0,\ f_2(x, y) = 0$ simultaneously.

Graphically, the equations $f_1(x, y) = 0$ and $f_2(x, y) = 0$ define curves in the xy plane, and the equilibrium points are the intersections of these curves. These curves are called *nullclines* and are often useful in understanding the geometry of the phase portrait. The first, with equation $f_1(x, y) = 0$, is the set of points where the the vector field $f(x, y)$ has zero x coordinate; these are the points where the solution curves have a *vertical* tangent. Hence this is called the *vertical nullcline*. Similarly, $f_2(x, y) = 0$ defines the set of points at which the solution curves have a horizontal tangent, so this set is called the *horizontal nullcline*. A more general type of curve is an *isocline*; this is a curve where the solution curves all have a given slope.

As an example of finding and classifying equilibria we start with a modification to the predator-prey system of Example 5.6. In that system the rabbit population, in the absence of predators, will grow exponentially. This is unrealistic, and we already saw, in Example 2.5, a more reasonable one-species model, in which the population growth is modified by a "crowding term". We add this idea to our predator-prey system, so that the net growth rate for the prey population is reduced not only by a linear term proportional to the number of foxes but also by a linear term proportional to the number of rabbits:

EXAMPLE 5.7. The modified predator-prey system is

$$\frac{d}{dt}\begin{bmatrix}x\\y\end{bmatrix} = f(x,y) = \begin{bmatrix}(-s+\alpha y)x\\(r-\beta x-\delta y)y\end{bmatrix}.$$

Here x and y represent populations of the predator (foxes) and the prey (rabbits); r,s,α,β,δ are positive parameters.

We shall assume that

$$\frac{r}{\delta} > \frac{s}{\alpha}. \tag{5.4}$$

Exercise 5.10 describes the situation when this condition is not satisfied.

We start our investigation of this system by drawing the nullclines. The vertical nullcline is defined by setting the first component of $f(x,y)$ to zero, so $(-s+\alpha y)x = 0$. Since the right hand side is 0 the only way that this can happen is if one of the factors on the left is 0, and this leads to $x = 0$ or $y = s/\alpha$. Thus the horizontal nullcline is the union of the two straight lines with these equations. Similarly the horizontal nullcline is defined by $(r-\beta x-\delta y)y = 0$, so it is the union of the two straight lines with equations $y = 0$ and $\beta x + \delta y = r$.

We can find the equilibria by solving the equations for the nullclines simultaneously, as we did for Example 5.6. There are four possibilities for solutions, since each nullcline is the union of two straight lines. The pair of lines $x = 0$ and $y = 0$ determines the equilibrium at the origin, $O = (0,0)$. If $x = 0$ and $\beta x + \delta y = r$ we find that $y = r/\delta$, and this gives the second equilibrium, $C = \left(0, \frac{r}{\delta}\right)$. If we solve $y = s/\alpha$ and $\beta x + \delta y = r$ we find $x = r/\beta - s\delta/(\alpha\beta)$, so a third equilibrium is at $S = \left(\frac{r}{\beta} - \frac{s\delta}{\alpha\beta}, \frac{s}{\alpha}\right)$. The x coordinate of S can be factored as $\frac{\delta}{\beta}\cdot\left(\frac{r}{\delta} - \frac{s}{\alpha}\right)$, and this is positive by the condition (5.4). Hence S is in the first quadrant. The fourth combination is $y = 0$ and $y = s/\alpha$, and these lines do not intersect.

Figure 5.3 shows the nullclines and the equilibria, for a typical choice of parameters. With these parameters $C = (0, 5000)$ and $S = (60, 2000)$. The point $(0, 2000)$ is **not** an equilibrium point since it is not an intersection point of the two nullclines: it is on the vertical nullcline, but it is not on the horizontal nullcline. Similarly, the point $(100, 0)$ is **not** an equilibrium point.

We have shown horizontal and vertical arrows along the nullclines. The directions of the arrows are determined from the vector field f. For example, on the x axis, which is part of the horizontal nullcline, we have $y = 0$ so $\frac{dx}{dt} = (-s+\beta y)x = -sx$. This is negative for $x > 0$ and positive for $x < 0$, so the field points to the left for $x > 0$ and to the right for $x < 0$. In general, along a nullcline the direction of the arrows can only change at the equilibrium points. For example, along the

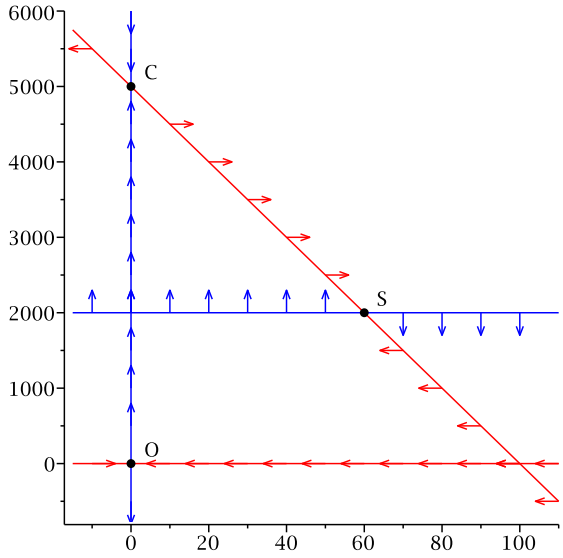


Figure 5.3: Nullclines for Example 5.7, with $r = .1$, $s = .2$, $\alpha = .0001$, $\beta = .001$, $\delta = .0002$.

horizontal nullcline the direction of the vector field is determined by the sign of its x component, and, by continuity, this sign can only change at points where the x component becomes 0. Hence, to determine the direction of the arrows along a nullcline it is only necessary to divide it into pieces by cutting it at the equilibrium points, and then to check one point on each piece.

Warning: The nullclines are not, in general, solution curves. In Figure 5.3 the x and y axes are unions of trajectories, as is the equilibrium point S, but at other points on the nullclines the solutions **cross** the nullclines.

Our next step in analyzing the phase portrait is to consider a linear approximation to the vector field near each equilibrium. We did this in the last section using equation (5.3), which exhibits $f(x, y)$ as a linear function plus terms of "higher order". We need a more systematic way of doing this. This is provided by the Linear Approximation Theorem from multi-variable calculus; see Theorem A.11.

The idea is as follows: For the vector differential equation $\dfrac{dz}{dt} = f(z)$ we first calculate the matrix derivative Df. Then, at each equilibrium point (x_0, y_0) we calculate the matrix $A = f(x_0, y_0)$. The *linearized system* near (x_0, y_0) is the system

$\frac{d}{dt}\begin{bmatrix}x\\y\end{bmatrix} = A\begin{bmatrix}x\\y\end{bmatrix}$. In chapter 4 we saw how to describe the phase portrait of the linearized system in terms of the eigenvalues and eigenvectors of A. We can then ask whether the orbits near the equilibrium point (x_0, y_0) of the original system are approximated by the orbits of the linearized system near the origin. In most cases there is a very close relation, as summarized in the following theorem:

THEOREM 5.8 (The 2 dimensional Linearization Theorem). *Suppose $f(x_0, y_0) = 0$ and $A = Df(x_0, y_0)$, and define the two systems:*

$$\textit{Original: } \frac{d}{dt}\begin{bmatrix}x\\y\end{bmatrix} = f(x, y) \quad \textit{Linearized: } \frac{d}{dt}\begin{bmatrix}x\\y\end{bmatrix} = A\begin{bmatrix}x\\y\end{bmatrix}.$$

If A is hyperbolic then the orbits of the original system near (x_0, y_0) "look like" the orbits of the linearized system near $(0, 0)$. In particular:

(a) *If all eigenvalues of A have real part less than 0 then (x_0, y_0) is a sink for the original system.*
(b) *If all eigenvalues of A have real part greater than 0 then (x_0, y_0) is a source for the original system.*
(c) *If the eigenvalues λ_1, λ_2 of A are real, with $\lambda_1 < 0 < \lambda_2$, then (x_0, y_0) is a saddle equlibrium.*
(d) *In the saddle case, let v_1 and v_2 be eigenvectors corresponding to λ_1 and λ_2. Then the stable curve through (x_0, y_0) has tangent vector at (x_0, y_0) parallel to v_1, and the unstable curve through (x_0, y_0) has tangent vector at (x_0, y_0) parallel to v_2.*

Moreover, if the eigenvalues of A are non real, so $\lambda = a \pm bi$ with $b \neq 0$, then the orbits near (x_0, y_0) spiral around (x_0, y_0), and the spiralling period of the orbit starting at (x_1, y_1) approaches $\frac{2\pi}{|b|}$ as (x_1, y_1) approaches (x_0, y_0). This spiralling behaviour occurs even if A is not hyperbolic.

Following this theorem, we say that an equilibrium point (x_0, y_0) is *hyperbolic* if the matrix $Df(x_0, y_0)$ is hyperbolic.

We need to explain the meaning of several terms in this theorem.

The sense in which the orbits of the original system "look like" the orbits of the linearized sytem near the origin is, technically, that there is a transformation $(u, v) = F(x, y)$ defined near (x_0, y_0) so that every piece of an orbit of the original system near (x_0, y_0) is transformed to a piece of an orbit of the linearized system in the UV plane. The precise definition is very technical, and we don't need it here.

We say an equilibrium point (x_0, y_0) is a *saddle* if there are two curves passing through the point which are unions of orbits, so that

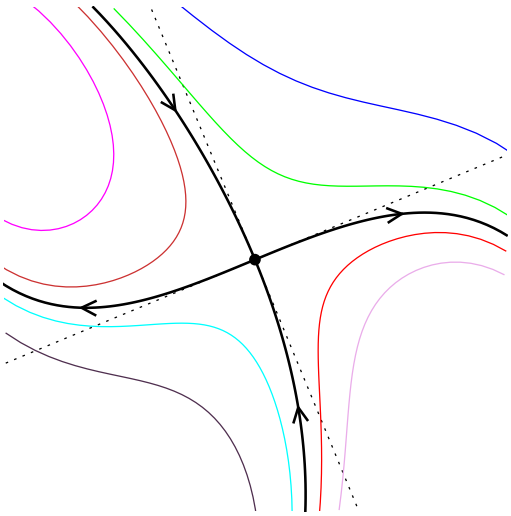


Figure 5.4: A non-linear saddle.

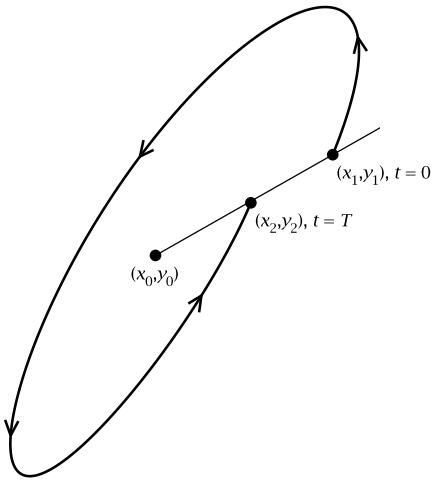


Figure 5.5: The spiralling period at (x_1, y_1) is T.

(a) The points on one curve (the *stable* curve) limit to (x_0, y_0) as $t \to +\infty$.
(b) The points on the other (the *unstable* curve) limit to (x_0, y_0) as $t \to -infty$.
(c) No other points have limit (x_0, y_0) as t approaches either $\pm\infty$. In fact, there is some distance c so that all points near (x_0, y_0) which are not on the stable or unstable curve move at least c units away from (x_0, y_0), in both positive and negative time.

See Figure 5.4, where the stable and unstable curves are shown, as well as the tangent lines in the direction of the eigenvectors.

Finally, we say that the orbit through (x_1, y_1) *spirals around* (x_0, y_0) if the following geometric condition is satisfied: Draw the ray starting at (x_0, y_0) and passing through (x_1, y_1). Now follow the orbit in positive time, starting at (x_1, y_1). We require that this orbit leaves the ray in one direction (clockwise or counterclockwise), and, at some future time T, it crosses the ray again in the same direction. The smallest such positive T is called the *spiralling period* of the orbit starting at (x_1, y_1). This will be the ordinary period of the orbit if it is a closed orbit; but the orbit may not be closed, in which case it spirals towards or away from (x_0, y_0). In general, the spiralling period is different for different orbits, and it is different at different points on the same orbit if the orbit is not closed.

Figure 5.5 illustrates an equilibrium point (x_0, y_0) and the ray from (x_0, y_0) through the point (x_1, y_1). The orbit starting at (x_1, y_1) at time $t = 0$ spirals around

the equilibrium until it again hits the ray at the point (x_2, y_2), at time $t = T$. Then T is the spiralling period for the orbit starting at (x_1, y_1).

We now apply the linearization procedure to the system of Example 5.7. We first calculate Df. The first column of this matrix is $\dfrac{\partial f}{\partial x}$ and the second is $\dfrac{\partial f}{\partial y}$. Differentiating $\begin{bmatrix} (-s+\alpha y)x \\ (r-\beta x-\delta y)y \end{bmatrix} = \begin{bmatrix} -sx+\alpha xy \\ ry-\beta xy-\delta y^2 \end{bmatrix}$ with respect to x and y gives

$$Df(x,y) = \begin{bmatrix} -s+\alpha y & \alpha x \\ -\beta y & r-\beta x-2\delta y \end{bmatrix}.$$

Next, we consider each equilibrium point:

The first equilibrium is at the origin. Plugging $(0,0)$ into the formula for Df gives $Df(0,0) = \begin{bmatrix} -s & 0 \\ 0 & r \end{bmatrix}$. This has eigenvalues $-s$ and r, so this equilibrium point is a hyperbolic saddle. The corresponding eigenvectors are $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, so we expect that the stable and unstable curves at the origin will be tangent to the x and y axes. In fact, we already know that the x and y axes are unions of orbits.

The second equilibrium is $C = (0, r/\delta)$, so

$$Df(C) = Df\left(0, \frac{r}{\delta}\right) = \begin{bmatrix} -s+\alpha\cdot r/\delta & 0 \\ \beta\cdot r/\delta & r-2r \end{bmatrix} = \begin{bmatrix} -s+\alpha\cdot r/\delta & 0 \\ -\beta\cdot r/\delta & -r \end{bmatrix}.$$

The eigenvalues are $\lambda_1 = -s+\alpha\cdot\dfrac{r}{\delta} = \alpha\left(\dfrac{r}{\delta}-\dfrac{s}{\alpha}\right)$ and $\lambda_2 = -r$. Condition (5.4) implies that λ_1 is positive, so C is another hyperbolic saddle. The second eigenvector is $v_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, and this is consistent with the fact that the y axis is a union of solutions. The eigenvector corresponding to λ_1 is messier: $v_1 = \begin{bmatrix} 1 \\ -r\beta/\left(r\delta+(r\alpha-s\delta)\right) \end{bmatrix}$. It follows from (5.4) that the slope of this vector is less than 0 and greater than $-\beta/\delta$, which is the slope of the line $\beta x+\delta y = r$. Thus we expect the unstable curve at C in the first quadrant to start at C at a negative slope, but greater than the slope of the horizontal nullcline passing through C.

The third equilibrium is $S = \left(\dfrac{r}{\beta}-\dfrac{\delta s}{\alpha\beta}, \dfrac{s}{\alpha}\right)$, so

$$\begin{aligned} Df(S) &= \begin{bmatrix} -s+\alpha\cdot s/\alpha & \alpha\left(r/\beta-\delta s/(\alpha\beta)\right) \\ -\beta\cdot s/\alpha & r-\beta\left(r/\beta-\delta s/(\alpha\beta)\right)-2\delta\cdot s/\alpha \end{bmatrix} \\ &= \begin{bmatrix} 0 & (\alpha r-\delta s)/\beta \\ -\beta s/\alpha & -\delta s/\alpha \end{bmatrix} \end{aligned}$$

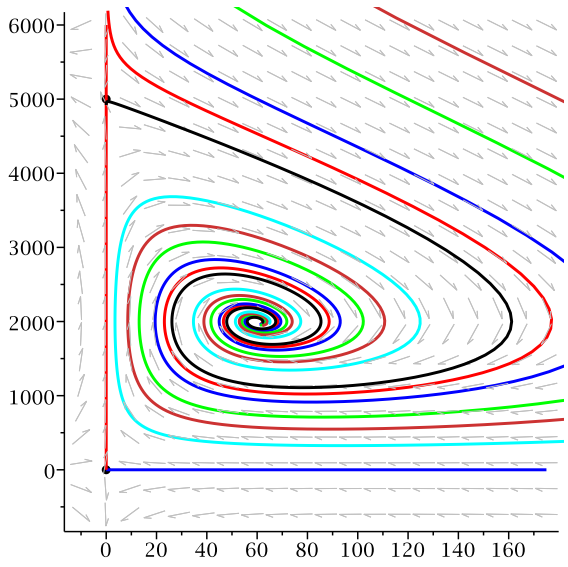


Figure 5.6: Example 5.7, with $r = .1$, $s = .2$, $\alpha = .0001$, $\beta = .001$, $\delta = .00002$. Time is measured in months.

The $1, 2$ entry in this matrix may be written as $\dfrac{\alpha\delta}{\beta}\left(\dfrac{r}{\delta} - \dfrac{s}{\alpha}\right)$, so, by condition (5.4), it is positive. Hence this matrix has negative trace and positive determinant, and, using exercise 4.9, we conclude that both eigenvalues have negative real part. Therefore S is a hyperbolic sink. Further calculation shows that the sink is a spiral sink as long as $\dfrac{r}{\delta} - \dfrac{s}{\alpha} > \dfrac{s\delta}{4\alpha^2}$.

Figure 5.6 shows the phase portrait in the first quadrant, with the same parameters as in Figure 5.3; the scale has been changed to accommodate more orbits. In the figure the sink at S is obvious. The stable and unstable curves corresponding to the saddle at the origin lie on the x and y axes, and any orbit that passes close to the origin has a "hyperbola-like" shape. For example, points starting in the first quadrant near $(100, 0)$ will move to the left, tracking the orbit on the stable curve on the x axis, and will slowly rise as they get near the origin. Eventually they will come close enough to the y axis that they will begin to track the unstable curve on the y axis. However, the curve will not continue to approach the y axis, since it will eventually come under the influence of the sink or the saddle at $C = (0, 5000)$. Points close to C will similarly track one of the stable curves at C, which converge to C along the y axis from above or below. When they get close enough to C they will

start moving away from C, tracking the unstable curve at C. This unstable curve is indicated on the Figure: It starts at C in the direction given by the eigenvector of the linearized system which corresponds to the positive eigenvalue at C, and eventually it spirals toward the sink S.

The figure illustrates another feature of this example. Any orbit which starts in the open first quadrant converges to the sink S as $t \to +\infty$. It requires some work to prove this, because it is necessary to explain why orbits that start with a large y coordinate will move far to the right, but will eventually move back towards the origin with a y value less than the y value at the sink. We'll postpone this argument, but we'll formalize this picture with a definition. If p is a sink of a dynamical system then we define the *basin of attraction* of p to be the set of all points (x, y) which approach p as $t \to +\infty$. By the Hartman-Grobman Theorem we know that every point sufficiently near p is in the basin, but the basin is usually much more extensive. In this case the basin is the entire open first quadrant.

It is also interesting to ask where points came from. In this example, almost all points in the open first quadrant "go to ∞" as $t \to -\infty$. In fact, in most cases these orbits blow up at a finite negative value of t. The only exceptions are the sink S, which stays fixed for all t; and the points on the unstable curve of the saddle C, which converge to C as $t \to -\infty$.

5.4. Conservation of energy and the pendulum problem

A classic physics problem is the simple pendulum. This system is confined to a vertical plane. A mass m is attached to a massless, rigid rod of length L; the other end of the rod is attached to a fixed point and is free to pivot about this point. There is no friction, and the only force acting on the mass is gravity. See Figure 5.7.

We use the angle x between the pendulum shaft and the vertical direction to describe the motion of the pendulum. Since the motion of the pendulum is constrained to lie on the circle of radius L about the pivot point, we can use the arc length xL between the pendulum and its rest point as the position variable. The effective force acting on the pendulum is the tangential component of the gravitational force $\begin{bmatrix} 0 \\ -mg \end{bmatrix}$, which has magnitude $mg \sin x$. Applying Newton's Second Law, we are led to the following model:

Example 5.9. The ideal pendulum: The motion is governed by the equation $\dfrac{d^2x}{dt^2} = -a^2 \sin x$ where $a^2 = \dfrac{g}{L}$. The initial value problem for this equation is $x(0) = x_0$, $\dfrac{dx}{dt}(0) = v_0$ where x_0 is the initial angular displacement and v_0 is the initial angular velocity.

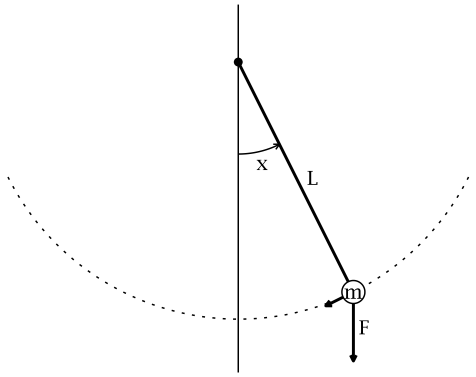


Figure 5.7: The pendulum.

As we did in the spring example 4.7, we replace this second order equation with a system of first order differential equations, by introducing the angular velocity $y = \dfrac{dx}{dt}$ as a second variable. Then $\dfrac{dy}{dt} = \dfrac{d^2x}{dt^2} = -a^2 \sin x$, so the system version of this model is

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} = -a^2 \sin x. \tag{5.5}$$

It is not possible to find the solution curves of this system in terms of elementary functions of t.

We start analyzing this system by examining its equilibrium points. Since $\dfrac{dx}{dt} = y$ the vertical nullcline is defined by $y = 0$, so this is just the x axis. Since $\dfrac{dy}{dt} = -a^2 \sin x$ the horizontal nullcline is defined by $\sin x = 0$. This corresponds to infinitely many vertical lines, defined by $x = n\pi$ where n is an integer. Hence the equilibria are at the points $(n\pi, 0)$. When $n = 0$ this is just the origin, so $x = 0$, $y = 0$. This makes physical sense, since it corresponds to the pendulum hanging straight down (the angle x is 0) and at rest (the angular velocity y is 0). The points $(2k\pi, 0)$ correspond to the same physical situation, since an angle of $2k\pi$ also describes the pendulum hanging straight down. The equilibria of the form $((2k + 1)\pi, 0)$ correspond to the physical situation where the pendulum is pointing straight up, and the angular velocity is 0. This makes some physical sense: remember that the shaft of

the pendulum is rigid, so it is possible, although very unlikely, for it to balance in this position and remain there for all time.

We can calculate the linearization of the system at these equilibria. Writing $f(x,y) = \begin{bmatrix} y \\ -a^2 \sin x \end{bmatrix}$ we obtain $Df(x,y) = \begin{bmatrix} 0 & 1 \\ -a^2 \cos x & 0 \end{bmatrix}$, so

$$Df(2k\pi, 0) = \begin{bmatrix} 0 & 1 \\ -a^2 & 0 \end{bmatrix}, \quad Df((2k+1)\pi, 0) = \begin{bmatrix} 0 & 1 \\ a^2 & 0 \end{bmatrix}.$$

At the even multiples of π the eigenvalues are $\pm ai$, with zero real part, so these equilibria are not hyperbolic and the linearization process is not applicable. According to Theorem 5.8 the solution curves near these equilibria must cycle around the origin, approaching a rotation with period $\dfrac{2\pi}{a}$, but the linearization cannot tell us whether the orbits are closed.

At the odd multiples of π the eigenvalues are a and $-a$, so these equilibria are saddles. The corresponding eigenvectors are $\begin{bmatrix} 1 \\ a \end{bmatrix}$ and $\begin{bmatrix} 1 \\ -a \end{bmatrix}$, so the unstable curves at these equilibria will have slope a, while the stable curves will have slope $-a$.

We need more information before we can describe the behavior near the other set of equilibria. In fact, we can get a full picture of the phase portrait by finding a constant of the motion.

Most idealized models of simple physical systems have a particular constant of the motion, which is known in physics as the *total energy* of the system. In our case this can be defined as $\varphi(x,y) = \frac{1}{2}y^2 - a^2 \cos x$. In physical terms this is the sum of the *kinetic energy* $K = \frac{1}{2}y^2$ and the *potential energy* $V = -a^2 \cos x$, and these are related to the differential equation by $\dfrac{dx}{dt} = \dfrac{\partial K}{\partial y}$ and $\dfrac{dy}{dt} = -\dfrac{\partial V}{\partial x}$. If we ignore the physical meaning of these quantities we can simply verify

$$\frac{dx}{dt} = \frac{\partial \varphi}{\partial y}, \quad \frac{dy}{dt} = -\frac{\partial \varphi}{\partial x}. \tag{5.6}$$

But these equations immediately imply that φ is a constant of the motion, since

$$\frac{d}{dt}\varphi(x(t), y(t)) = \frac{\partial \varphi}{\partial x}\frac{dx}{dt} + \frac{\partial \varphi}{\partial y}\frac{dy}{dt} = \frac{\partial \varphi}{\partial x}\frac{\partial \varphi}{\partial y} + \frac{\partial \varphi}{\partial y}\left(-\frac{\partial \varphi}{\partial x}\right) = 0.$$

So each orbit of the system lies in a single level curve, of the form $\varphi(x,y) = c$. To analyze these level curves, first note that $-\cos x \geq -11$ and $y^2 \geq 0$ so $\varphi(x,y) = \frac{1}{2}y^2 - a^2 \cos x \geq 0 - a^2(1) = -a^2$. So $-a^2$ is a lower bound for φ; moreover, it is the minimum value of φ since $\varphi(2k\pi, 0) = -a^2 \cos(2k\pi) = -a^2$. Hence the level set defined by $\varphi = -a^2$ consists only of the points $(2k\pi, 0)$, so each of these points is a single orbit. We already knew this, since each of these is an equilibrium point.

Since $-a^2$ is the minimum value of φ the level sets $\varphi = c$ for $c < -a^2$ are empty.

Suppose now that $c > -a^2$. Then $a^2\cos(x) + c > 0$ so the equation $\varphi(x,y) = \frac{1}{2}y^2 - a^2\cos x = c$ can be solved for y in terms of x, as

$$(*) \qquad y = \pm\sqrt{2a^2\cos(x) + 2c}.$$

If $c > a^2$ then $2a^2\cos(x) + 2c > 2a^2(-1) + 2c > 0$ for all values of x, so the level curve $\varphi = c$ consists of two components, each defined for all x, one above the x axis and one below. Thus each of these components is an orbit.

If $-a^2 < c < a^2$ then the functions in $(*)$ are not defined for all x. To see this, let $x_c = \cos^{-1}\left(-\frac{c}{a^2}\right)$. This is defined since $|c/a^2| < 1$, and is in the interval $(0,\pi)$. Since the cosine function is even and is decreasing on $(0,\pi)$ we have $2a^2\cos(x) + 2c > 0$ for $|x| < x_c$ and $2a^2\cos(x) + 2c < 0$ for $x_c < |x| \le \pi$. Then we can describe the level curve as follows. First, there is a closed loop which starts at $(-x_c, 0)$, follows the graph of the positive function in $(*)$ until it hits the x axis again at $(x_c, 0)$, and then returns to the point $(-x_c, 0)$ along the graph of the negative function in $(*)$. By periodicity, this same picture is repeated infinitely often, so there are closed loops surrounding $(2k\pi, 0)$ for all k. Each of these closed loops is a closed orbit of the system.

There is one more case: $c = a^2$. In this case we use the trig identity $1 + \cos x = 2\cos^2(x/2)$ to simplify $(*)$:

$$y = \pm\sqrt{2a^2\left(\cos(x) + 1\right)} = \pm\sqrt{4a^2\cos^2\left(\frac{x}{2}\right)} = \pm 2a\left|\cos\left(\frac{x}{2}\right)\right|.$$

The saddle points $((2k+1)\pi, 0)$ are on this level set, since $\cos\left((2k+1)\pi/2\right) = 0$. If we cut the level set by taking out these points we are left with arcs of $\cos(x/2)$ and $-\cos(x/2)$ for x between consecutive odd multiples of π. Hence each such arc is an orbit of the system, and the orbit converges to saddle points as $t \to \pm\infty$. Hence each such arcs represents, simultaneously, a stable curves of one saddle and an unstable curve of another saddle. Curves which are simultaneously stable and unstable curves of saddles (or of the same saddle) are called *saddle connections*.

All these features of the phase portrait are illustrated in Figure 5.8. The closed orbits correspond to the pendulum swinging back and forth periodically. The non-closed orbits, like the one through $(0,5)$, correspond to the pendulum swinging entirely around the pivot, and continuing periodically. The orbit through $(0,4)$ converges to $(\pi, 0)$ in forward time, so it is one of the unstable curves for the saddle at $(\pi, 0)$. In backward time is converges to $(-\pi, 0)$ so it is one of the stable curves of the saddle at $(-\pi, 0)$. The forward orbit starting at $(0,4)$ corresponds to the motion of the pendulum if it starts at its rest position, $x = 0$, with an angular velocity $y = 4$. This is exactly enough velocity so that the pendulum will approach $(\pi, 0)$ as $t \to +\infty$. Physically, the pendulum swings counterclockwise, slowing down as it

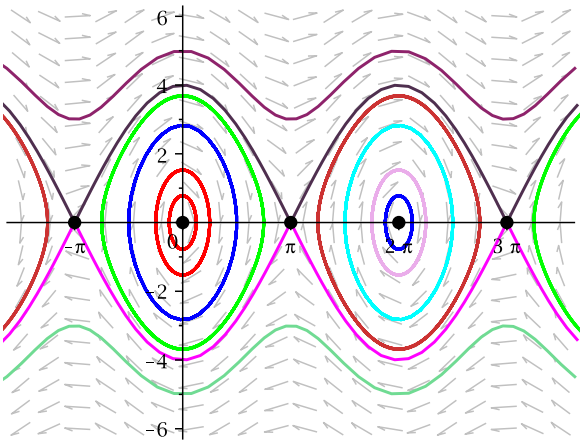


Figure 5.8: $\dfrac{dx}{dt} = y$, $\dfrac{dy}{dt} = -a^2 \sin(x)$, with $a = 2$.

approaches the balanced position where the shaft points straight up, with zero angular velocity. As $t \to -\infty$ the orbit approaches $(-\pi, 0)$. This is the same physical position, but the pendulum is approaching it clockwise.

Notice how the stable and unstable curves separate the phase portrait into different regions: Closed inside inside, non-closed outside. This is often true, and, for this reason, a stable or unstable curve is often called a *separatrix*.

Like the spring model (Example 4.7), this model ignores any dissipative forces, like friction. As in Example 4.8 we can add a generic dissipative force which opposes the action, so is in the direction opposite to y. This leads to:

EXAMPLE 5.10. The pendulum model with friction: The modified equation of motion is $\dfrac{d^2x}{dt^2} = -a^2 x - ry$ where $a^2 = \dfrac{g}{L}$ and $r > 0$.

As before, we treat this as a system of equations:

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} = -a^2 \sin(x) - ry.$$

For equilibrium points we need $y = 0$ and $-a^2 \sin(x) - ry = 0$. Plugging $y = 0$ into the second equation leads to $\sin(x) = 0$, so $x = n\pi$, just as in the original model. We now have $f(x, y) = \begin{bmatrix} y \\ -a^2 \sin(x) - ry \end{bmatrix}$, so $Df(x, y) = \begin{bmatrix} 0 & 1 \\ -a^2 \cos(x) & -r \end{bmatrix}$. At the

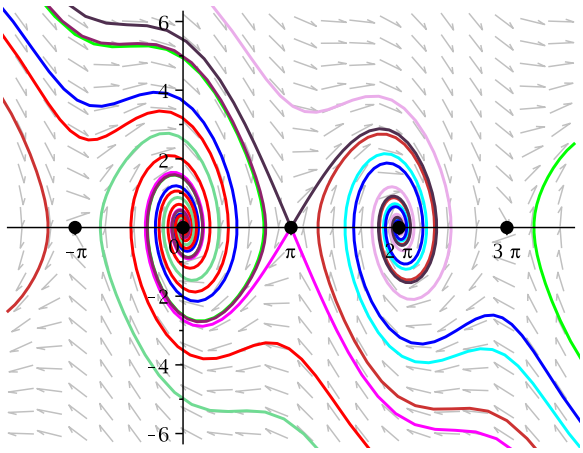


Figure 5.9: $\dfrac{dx}{dt} = y$, $\dfrac{dy}{dt} = -a^2 \sin(x) - ry$, with $a = 2$ and $r = 0.7$.

equilibria this becomes

$$Df(2k\pi, 0) = \begin{bmatrix} 0 & 1 \\ -a^2 & -r \end{bmatrix}, \quad Df((2k+1)\pi, 0) = \begin{bmatrix} 0 & 1 \\ a^2 & -r \end{bmatrix}.$$

At the odd multiples of π the derivative matrix has determinant $-a^2 < 0$. Since the determinant is the product of the eigenvalues we must have one positive and one negative eigenvector. Thus the points $((2k+1)\pi, 0)$ are still saddles. At the even multiples of π the derivative matrix has determinant $a^2 > 0$ and trace $-r < 0$. It follows from Exercise 4.10 that the points $(2k\pi, 0)$ are sinks. Further calculation gives the eigenvalues as $\frac{1}{2}\left(-r \pm \sqrt{r^2 - 4a^2}\right)$. The eigenvalues are non-real if $r < 2a$, so in this case the trajectories spiral toward the equilibrium point. For $r \geq 2a$ the eigenvalues are real, so the trajectories do not spiral.

Figure 5.9 shows a number of orbits for a choice of a and r for which the sinks show spirals. Consider one orbit, for example the one passing near $(0, 4)$. Here is a physical interpretation of this orbit: The pendulum starts at its rest position, with angular velocity $y \approx 4$, so it starts moving counterclockwise. The pendulum does not have enough velocity to carry it to, or beyond, the vertical position corresponding to the saddle at $(\pi, 0)$. Instead it reaches approximately a horizontal position with $y = 0$, then starts swinging clockwise. The pendulum will continue to swing back

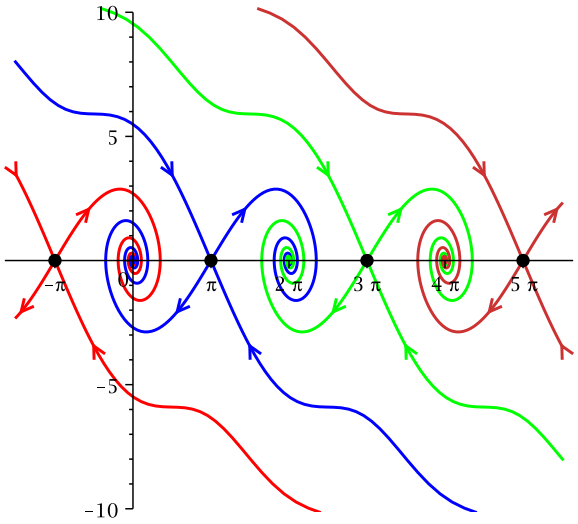


Figure 5.10: The separatrices in Figure 5.9.

and forth, but, unlike in the idealized model, the motion does not repeat. It will, instead, oscillate around the rest position and steadily converge to the rest position.

A physical explanation for this decay of the motion is that the pendulum loses energy. We can see this mathematically, using the energy function from the idealized model, $\varphi(x, y) = \frac{1}{2}y^2 - a^2 \cos x$. We calculate the time derivative of φ along orbits of the modified system to get

$$\begin{aligned}\frac{d\varphi}{dt} &= \frac{\partial\varphi}{\partial x}\frac{dx}{dt} + \frac{\partial\varphi}{\partial y}\frac{dy}{dt} \\ &= a^2 \sin(x) \cdot y + y \cdot (-a^2 \sin(x) - ry) \\ &= -ry^2\end{aligned}$$

This is less than 0 if $y \neq 0$, so φ decreases as t increases, unless the orbit is one of the equilibrium points, where φ is constant. We therefore expect any orbit to converge to one of the minima of φ, and these are the sinks $(2k\pi, 0)$. The only exceptions are the equilibria and the stable curves that converge to the saddles.

In Figure 5.10 we show just the separatrices for a number of saddles, on a somewhat larger scale. Notice that the stable curves form the boundaries of the basins of attraction for the sinks. For example, the points on the Y axis between approximately $(0, -5)$ and $(0, 5)$ converge to the sink at $(0, 0)$ as $t \to +\infty$. However, the points between approximately $(0, 5)$ and $(0, 10)$ converge to the sink at $(0, 2\pi)$, and those just above this interval converge to the sink at $(4\pi, 0)$.

An orbit starting at $(0, 7)$ corresponds to the following physical motion: The pendulum starts at its rest position with an angular velocity $y = 7$. This is fast enough that the pendulum will swing counterclockwise through an entire circle, returning to its rest position, which is now represented by $x = 2\pi$, but with a smaller angular velocity. At this velocity it will not be energetic enough to complete another full circle, so it will oscillate back and forth around its rest position, eventually converging to the rest position. Similarly, an orbit starting at about $(0, 12)$ will correspond to the pendulum being energetic enough to complete two full circles before converging to its rest position, which is now represented by $x = 4\pi$.

5.5. Limit cycles

Here is a basic tool that is only available for analysing differential equations in two dimensions:

THEOREM 5.11 (Poincaŕe-Bendixson). *Suppose* $\dfrac{dx}{dt} = f(x)$ *is a differential equation defined for* x *in some region of the plane* $\mathbb{R}^2$. *Suppose that* S *is a non-empty, bounded, closed, forward invariant subset of the domain of* f. *Then* S *contains an equilibrium point or a closed orbit. In fact, any orbit that starts in* S *is a closed orbit, or converges to a limit cycle, or comes arbitrarily close to an equilibrium point as* $t \to \infty$.

The proof of this theorem requires some topology, and we will not attempt to describe the proof. However, we need to explain the terms used in the theorem, and then show how it is used.

A periodic orbit C is called a *limit cycle* if all nearby orbits on at least one side of C converge to C as $t \to \infty$. This can only happen if nearby orbits spiral toward C.

A set S in $\mathbb{R}^k$ is *bounded* if there is a constant R so that $\|x\| \leq R$ for all x in S. Equivalently, the set S lies inside or on the circle of radius R around the origin. This condition is usually easy to check. For example, the square with vertices at $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$ is bounded since it lies inside or on the circle centered at the origin of radius $\sqrt{2}$. (Draw a picture.)

A set S in $\mathbb{R}^k$ is *closed* if it contains all its limit points. This means that if x_n is a sequence of points in S which converges to a point x_* in $\mathbb{R}^k$ then x_* must lie in S.

Most sets S that we will be interested in will be described by a finite number of conditions which are equalities or "weak" inequalities (using $\leq$ or $\geq$ instead of the stronger versions, $<$ or $>$). If these conditions involve only continuous functions then the set is closed. This is just because if a sequence of points x_n satisfies such conditions and $x_n \to x_*$ then functional values $f(x_n)$ converge to $f(x_*)$ (since the

functions involved are continuous) and equalities and weak inequalities are preserved by limits. For example, the set described by $x^2 + y^2 \leq 4$ and $xy \geq 1$ is closed.

There is an important topological property of sets called *compactness*, and a subset of R^k is compact if and only if it is both closed and bounded. We will occasionally use the term "compact" instead of "closed and bounded".

The last condition in the Poincaré-Bendixson Theorem is that S must be "forward invariant". This is only meaningful in terms of a particular differential equation, which determines the flow F^t. Then we say that a set S is *forward invariant* if every point of S stays in S in future time: that is, if x_0 is any point in S then $F^t(x_0)$ lies in S for all $t \geq 0$. We can also, of course, define *backward invariant* sets. The unmodified term *invariant* sometimes means "forward invariant" and sometimes "forward and backward invariant", depending on the context.

Usually the hard part in applying the Poincaré-Bendixson Theroem is checking that the set S is invariant. But there is a common situation that is not too hard. Consider a simple curve, B, with no endpoints. Then B separates nearby points into two sets (the two "sides" of the curve). If we call these two sides U and V then we say an orbit *crosses B from U to V* if it is on B for a particular t value, $t = t_0$; it is in U for t near t_0 but less than t_0; and it is in V for t near t_0 but greater than t_0.

PROPOSITION 5.12. *Suppose that the boundary of a set S consists of finitely many curves. Suppose that, except for finitely many points on the boundary, any trajectory that meets the boundary crosses the boundary from the outside to the inside. Then S is forward invariant.*

In this situation we say that the vector field (or the flow) *points into* S, even though it might actually be tangent to the boundary at a finite number of points.

PROOF. No orbits can leave S, except possibly at the finitely many exceptional points. But it follows from continuity of the flow that if one orbit leaves S then all nearby orbits must leave S. This would mean that infinitely many orbits would leave S, contradicting the hypothesis. □

In simple cases we can see graphically that a vector field points into a region, but we need a method that we can apply more generally. The following gives a way of seeing how orbits cross a curve.

LEMMA 5.13. *Suppose the curve B has no endpoints and satisfies $\varphi = c$ for some differentiable function φ. Let A_+ be the set where $\varphi > c$ and let A_- be the set where $\varphi < c$. If an orbit meets B at the point (x_1, y_1) then it*

(a) *crosses B from A_- to A_+ if $\dfrac{d\varphi}{dt} > 0$ at (x_1, y_1), or*

(b) *crosses B from A_+ to A_- if $\dfrac{d\varphi}{dt} < 0$ at (x_1, y_1).*

PROOF. Suppose the solution curve $(x(t), y(t))$ meets B when $t = t_0$. If $\dfrac{d\varphi}{dt}(t_0) > 0$ then $\varphi(x(t), y(t))$ is increasing at t_0. Since $\varphi(x(t_0), y(t_0)) = c$ it must be true that $\varphi(x(t), y(t)) < c$ for t near t_0 and less than t_0, and $\varphi(x(t), y(t)) > c$ for t near t_0 and greater than t_0. But $\varphi(x(t), y(t)) < c$ means that $(x(t), y(t))$ is in A_-, and $\varphi(x(t), y(t)) > c$ means that $(x(t), y(t))$ is in A_+.

The argument when $\dfrac{d\varphi}{dt}(t_0) < 0$ is similar. □

Here's an example. Suppose $\dfrac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = f(x, y) = \begin{bmatrix} -2x + \cos(y) \\ -2y + \sin(x) \end{bmatrix}$, and let S be the closed disk of radius 2 centered at the origin. Then S is closed and bounded. It is also forward invariant. To see this, let $\varphi(x, y) = x^2 + y^2$. Then the boundary of S is the circle of radius 2 centered at the origin, with equation $\varphi(x, y) = x^2 + y^2 = 4$. At points of the boundary,

$$\begin{aligned}
\frac{d\varphi}{dt} &= \frac{d}{dt}\left(x^2 + y^2\right) \\
\frac{d\varphi}{dt} &= 2x\frac{dx}{dt} + 2y\frac{dy}{dt} \\
&= 2x\left(-2x + \cos(y)\right) + 2y\left(-2y + \sin(x)\right) \\
&= -4(x^2 + y^2) + 2x\cos(y) + 2y\sin(x) \\
&= -16 + 2x\cos(y) + 2y\sin(x) && \text{since } x^2 + y^2 = 4 \\
&\leq -8 && \text{since } 2x\cos(y) \leq 2\,|x| \leq 4 \\
& && \text{and } 2y\sin(x) \leq 2\,|y| \leq 4
\end{aligned}$$

Hence, using Lemma 5.13, the vector field on the boundary points to the region where $\varphi(x, y) = x^2 + y^2 < 4$, and this is *inside* S. By Proposition 5.12, S is forward invariant. So the Poincaré-Bendixson Theorem guarantees that there is either an equilibrium point or a closed orbit (possibly both, and possibly more than one) in S. In fact, according to Theorem 5.16 there is an equilibrium point in S. This equilibrium point can be found by solving $f(x, y) = 0$, but this cannot be done without using numerical methods. Maple gives $(x, y) \approx (.486, 0.234)$.

Our main example was discovered by Balthasar van der Pol in 1920. He found it while experimenting with vacuum tubes; we do not give the derivation, which requires knowledge of electrical circuits. It exhibits an isolated periodic orbit, so it is very different qualitatively from the examples that we have seen so far. This feature has made the system useful in other applications, for example in an influential model of neuron excitation due to FitzHugh and Nagumo in the early 1960's.

EXAMPLE 5.14. The van der Pol equation is $\dfrac{d^2x}{dt^2} + \mu(x^2-1)\dfrac{dx}{dt} + x = 0$, where μ is a positive parameter.

We can recast this as a system by introducing the variable $v = \dfrac{dx}{dt}$ as in the pendulum model; this leads to

$$\frac{dx}{dt} = v, \quad \frac{dv}{dt} = -\mu(x^2-1)v - x.$$

We shall study a different conversion, which is easier to analyze. We first introduce the auxiliary function $F(x) = \mu\left(\frac{1}{3}x^3 - x\right)$, which is an antiderivative of the factor $\mu(x^2-1)$ in van der Pol's equation. In terms of this function we define a new variable $y = \dfrac{dx}{dt} + F(x)$, so $\dfrac{dx}{dt} = y - F(x)$. Also,

$$\frac{dy}{dt} = \frac{d}{dt}\left(\frac{dx}{dt} + F(x)\right) = \frac{d^2x}{dt^2} + F'(x)\frac{dx}{dt} = \frac{d^2x}{dt^2} + \mu(x^2-1)\frac{dx}{dt},$$

which matches the first two terms in van der Pol's equation. Hence van der Pol's equation becomes $\dfrac{dy}{dt} + x = 0$, or $\dfrac{dy}{dt} = -x$. We summarize our convertied system as

$$\frac{dx}{dt} = y - F(x), \ \frac{dy}{dt} = -x, \ F(x) = \mu\left(\tfrac{1}{3}x^3 - x\right). \tag{5.7}$$

We start analyzing this system by identifying the nullclines. The horizontal nullcline is given by $-x = 0$, so it is the y axis. The vertical nullcline is given by $y - F(x) = 0$, so it is the cubic curve with equation $y = F(x) = \mu\left(\frac{1}{3}x^3 - x\right)$.

The cubic intersects the y axis only at the origin, so this is the only equilibrium point. The derivative matrix for the vector field $\begin{bmatrix} y - F(x) \\ -x \end{bmatrix}$ is $\begin{bmatrix} -F'(x) & 1 \\ -1 & 0 \end{bmatrix}$. Plugging in the origin and remembering that $F'(x) = \mu(x^2-1)$, we get $\begin{bmatrix} \mu & 1 \\ -1 & 0 \end{bmatrix}$. The eigenvalues of this matrix are $\frac{1}{2}\left(\mu \pm \sqrt{\mu^2 - 4}\right)$. These are non-real with positive real part $\frac{1}{2}\mu$ if $0 < \mu < 2$. If $\mu \geq 2$ then they are real and positive, since $\sqrt{\mu^2-4} < \mu$. Hence the origin is a source, and it spirals if $\mu < 2$.

The general features of the phase portrait are similar for any positive value of μ. For convenience we fix $\mu = 1$ for the rest of the analysis, so the equations are

$$\frac{dx}{dt} = y - F(x), \ \frac{dy}{dt} = -x, \ F(x) = \tfrac{1}{3}x^3 - x. \tag{5.8}$$

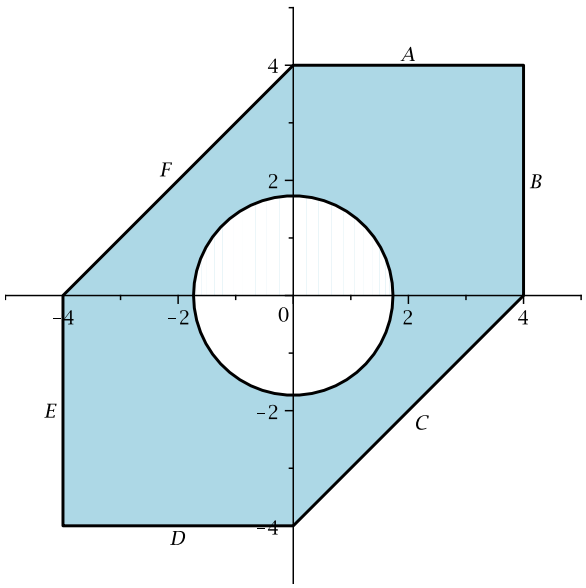


Figure 5.11: The invariant set S for the van der Pol system (5.7), with $\mu = 1$.

We shall apply the Poincaré-Bendixson Theorem to the set S illustrated in Figure 5.11. This set is bounded on the outside by a hexagon and on the inside by the circle of radius $\sqrt{3}$, centered at the origin. Precisely, the set S is defined by the inequality $x^2 + y^2 \geq 3$, together with linear inequalities corresponding to the six sides of the hexagon. Hence S is closed, and clearly it is bounded. We must show that it is forward invariant. We shall apply Lemma 5.13 to each of the 7 boundary curves.

First we examine the circle $x^2 + y^2 = 3$. Define $\varphi(x, y) = x^2 + y^2$ and calculate

$$\begin{aligned}\frac{d\varphi}{dt} &= \frac{d}{dt}\left(x^2 + y^2\right) = 2x\frac{dx}{dt} + 2y\frac{dy}{dt} \\ &= 2x\left(y - F(x)\right) + 2y\left(-x\right) = 2xy - 2xF(x) - 2xy \\ &= -2xF(x) = -2x\left(\tfrac{1}{3}x^3 - x\right) = -\tfrac{2}{3}x^2(x^2 - 3).\end{aligned}$$

Now on the circle $x^2 + y^2 = 3$ each coordinate, in absolute value, is no greater than the radius, $\sqrt{3}$. Hence $x^2 - 3 \leq 0$, so $\dfrac{d\varphi}{dt} \geq 0$ on the circle. In fact, there are only four points on the circle where $\dfrac{d\varphi}{dt} = -\frac{2}{3}x^2(x^2 - 3) = 0$, namely $(\pm\sqrt{3}, 0)$, where $x^2 - 3 = 0$, and $(0, \pm\sqrt{3})$, where $x^2 = 0$. Hence $\dfrac{d\varphi}{dt} > 0$ except at these four points,

so the vector field on the circle points into the region where $\varphi = x^2 + y^2 > 3$, which is *inside* S.

Next we consider the sides of the hexagon. We refer to them as A through F, as indicated in Figure 5.11. We will ignore the vertices, since Lemma 5.13 applies to curves without endpoints.

Side A: Let $\varphi_A(x, y) = y$, so this side satisfies $\varphi_A = 4$. We have $\dfrac{dy}{dt} = -x$ and this is negative on side A since $x > 0$ (remember, we are ignoring the vertex $(0, 4)$.) Hence the vector field on side A points into the region where $\varphi < 4$, or $y < 4$, and this is *inside* S.

Side B: Let $\varphi_B(x, y) = x$, so this side satisfies $\varphi_B = 4$. We have $\dfrac{d\varphi}{dt} = \dfrac{dx}{dt} = y - F(x)$. Since $x = 4$ and $y \leq 4$, this becomes $y - F(4) = y - \left(\frac{1}{3} \cdot 64 - 4\right) \leq 4 - 52/3 = -40/3 < 0$. Hence the vector field along side B points into the region where $\varphi < 4$, or $x < 4$, and this is *inside* S.

Side C: This side has the equation $y = x - 4$. Let $\varphi_C(x, y) = y - x$, so this side satisfies $\varphi_C = -4$. We calculate

$$\frac{d\varphi_C}{dt} = \frac{d}{dt}(y - x) = \frac{dy}{dt} - \frac{dx}{dt} = -x - (y - F(x)) = -x - y + (\tfrac{1}{3}x^3 - x) = \tfrac{1}{3}x^3 - 2x - y.$$

On side C we have $y = x - 4$, so we substitute this for y and we get $\dfrac{d\varphi_C}{dt} = \frac{1}{3}x^3 - 3x + 4$. Let $g(x) = \frac{1}{3}x^3 - 3x + 4$. In order to see that the vector field points into S on side C we need to see that $g(x)$ is positive for all corresponding x; that is, for $0 < x < 4$. So we find the minimum value of g: Calculate $g'(x) = x^2 - 3$, set it equal to 0, and solve for x. This gives $x = \sqrt{3}$ (we ignore $-\sqrt{3}$ since $x > 0$). Since $g''(x) = 2x > 0$ for $0 < x < 4$, we see that g is concave upwards, so $x = \sqrt{3}$ gives the minimum value of g on this interval. Hence

$$g(x) \geq g(\sqrt{3}) = \tfrac{1}{3} \cdot (\sqrt{3})^3 - 3\sqrt{3} + 4 = \tfrac{1}{3} \cdot 3\sqrt{3} - 3\sqrt{3} + 4 = 4 - 2\sqrt{3} \approx .536 > 0.$$

Thus $\dfrac{d\varphi_C}{dt} = g(x) > 0$ at points of side C, so the vector field along side C points into the region where $\varphi_C > 4$. This is the region where $y - x > 4$, or $y > x - 4$, which is *inside* S.

The other three sides of S are handled similarly.

Therefore, we have shown that the vector field points into S at all points of the boundary except $(\pm\sqrt{3}, 0)$, $(0, \pm\sqrt{3})$, and the six vertices of the hexagon. By Proposition 5.12 the set S is invariant.

We can now apply the Poincaré-Bendixson Theorem to this example: There must be an equilibrium point or closed orbit in S. The only equilibrium point of the system

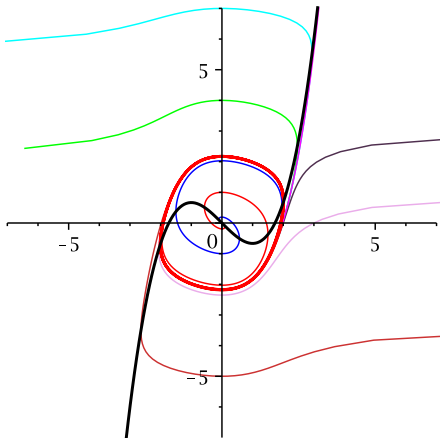


Figure 5.12: Orbits of the van der Pol system (5.7), with $\mu = 1$, together with the cubic curve $y = F(x)$.

is at the origin, and this is not in S. Hence there is a closed orbit of the system inside S. More detailed analysis shows that there is only one closed orbit, and that all other orbits (except the equilibrium point) spiral towards this closed orbit as $t \to +\infty$. This is illustrated in Figure 5.12.

From the graph it seems that orbits that cross the cubic $y = F(x)$ at points where $|x| > \sqrt{3}$ must follow close to the cubic until it crosses the x axis. This behavior is verified in Exercise 5.12. Similar analysis shows that any orbit that starts far from the origin must eventually cross the cubic, as illustrated.

Here are two addenda to the Poincaré-Bendixson Theorem:

THEOREM 5.15. *Suppose* $\dfrac{dx}{dt} = f(x)$ *is a differential equation defined for* x *in some region of the plane* $\mathbb{R}^2$. *Suppose that* S *is a non-empty, bounded, closed subset of the domain of* f, *and that* S *does not contain an equilibrium point or closed orbit. Then any orbit that enters* S *at some time* t_0 *must leave* S *at some time* $t_1 > t_0$.

THEOREM 5.16. *Suppose that* $\dfrac{dx}{dt} = f(x)$ *is a differential equation defined for* x *in some region of the plane* $\mathbb{R}^2$. *Suppose that* S *is a non-empty, bounded, closed, forward invariant subset of the domain of* f, *and that the boundary of* S *is a simple closed curve. Then* S *contains an equilibrium point.*

A *simple closed curve* is a closed curve without self-intersections. Theorem 5.16 applies in the special case that S consists of a closed orbit plus its interior.

5.6. Stability

There are several distinct notions of *stability* in dynamical systems. We will give two of them:

The first definitions refer to a fixed dynamical system. If C is a non-empty compact invariant set we say that C is *stable* if all orbits that start near C stay near C for all $t \geq 0$, and we say that C is *asymptotically stable* if, in addition, all orbits that start near C converges to C as $t \to +\infty$.

To make this precise, let F^t be the flow of the system. Define $d_C(x)$ to be the distance of a point x to the set C; this is the minimum distance from x_1 to a point in C. Then C is stable if for any positive number r there is a positive number δ so that if $d(x_1) < \delta$ implies $d(F^t(x_1)) < r$ for all $t \geq 0$. If, in addition, there is some positive number δ_0 so that $\lim_{t\to+\infty} d_C(F^t(x_1)) = 0$ whenever $d_C(x_1) < \delta_0$ then C is asymptotically stable.

If C consists of a single equilibrium point x_0 then the definition of asymptotically stable is the same as the definition of a sink. For an example of an equilibrium point that is stable, but not asymptotically stable, see the origin in the pendulum example, Figure 5.8.

If C consists of a single closed orbit then the definition of asymptotically stable is the same as the definition of two-sided limit cycle. On the other hand, each closed orbit surrounding the origin in Figure 5.8 is stable, but not asymptotically stable.

In higher dimensions we will see other examples of stable invariant sets.

The second definition is concerned with the idea that features of the phase portrait will persist when the dynamical system itself is perturbed. There is a formal definition for an entire dynamical system to be *structurally stable*, but we will instead concentrate on features of the phase portrait. We say that a dynamical system is *structurally stable* near a compact invariant set C if every nearby system has a compact invariant set C' near C, and the orbits near and on C correspond to orbits near and on C'. The precise definition is complicated, and the main results about structural stability require a fair amount of analysis, so we will skip this and just record a simple version of the main theorem.

THEOREM 5.17. *The differential equation* $\dfrac{dx}{dt} = f(x)$ *is structurally stable near any hyperbolic equilibrium point.*

The definitions above make sense for discrete dynamical systems as well. To formulate the analog of Theorem 5.17 we need a definition of hyperbolicity: A fixed point x_0 of a discrete dynamical system F is *hyperbolic* if the linearization $DF(x_0)$ is a hyperbolic matrix as defined in Section 3.2. Then the main theorem becomes:

THEOREM 5.18. *Suppose F is an invertible discrete dynamical system. Then F is structurally stable near any hyperbolic fixed point.*

5.7. Bifurcations

Suppose a dynamical system depends on a parameter α. If the dynamical system is structurally stable for some value of α, say $\alpha = \alpha_0$, then small changes in α will cause only small changes in the phase portrait, and will not change the qualitative features of the system. For example, suppose that x_0 is a hyperbolic equilibrium point for the system when $\alpha = \alpha_0$. Then there will still be a unique hyperbolic equilibrium point near x_0 when α is near α_0, and this point will be of the same type (source, sink or saddle) as x_0. Similarly, hyperbolic limit cycles (see the definition in the next chapter) will persist.

On the other hand, the phase portrait may be different for other values of α. For example, when $\alpha = \alpha_1$ the system might also be structurally stable, but there may be a different number of equilibrium points than when $\alpha = \alpha_0$ or their type (source, sink or saddle) might be different. In such a case there must be some value of the parameter, say $\alpha = \alpha'$, so that the phase portraits do not change qualitatively for $\alpha_0 \leq \alpha < \alpha'$, but so that the phase portrait for α' is different. Such a parameter value is called a *bifurcation* value.

The system cannot be structurally stable for $\alpha = \alpha'$, since there are values of α arbitrarily close to α' for which the phase portrait is different from the phase portrait for $\alpha = \alpha'$. In simple cases the phase portraits for $\alpha' < \alpha \leq \alpha_1$ will not change qualitatively, so that the phase portrait for $\alpha = \alpha'$ represents the only transitional form between $\alpha = \alpha_0$ and $\alpha = \alpha_1$. However, the bifurcation picture can be much more complicated, with many – possibly infinitely many – bifurcation values between α_0 and α_1.

Our next example is taken from a recent bio-engineering paper: T. Gardner, C. Cantor, J. Collins, *Construction of a genetic toggle switch in Escherichia coli* [1].

This system models the behavior of two genes, which we call G_U and G_V, inserted via a plasmid in the bacterium *Escherichia coli*. These genes code for proteins, which we denote P_U and P_V, respectively. The protein P_U is a *repressor* for the gene G_V; that is, the presence of P_U in a cell inhibits the activity of gene G_V. Similarly, P_V is a repressor for the gene G_U. This behavior can be manipulated externally, by controlling one of two chemicals called *inducers*. The inducer I_U enhances the expression of G_U by interfering with the repressor P_V. Similarly, I_V is the inducer for gene G_V.

The paper describes experimental verification for behavior based on this basic model. The relative concentrations of the two repressors can be visually monitored

since the expression of one of the genes also causes a green fluorescent protein to be produced. Specifically, the authors used the following.

EXAMPLE 5.19. There are two non-negative state variables, u and v, which are the concentrations of the repressor proteins P_U and P_V. The simplified differential equations for the concentrations are

$$\frac{du}{dt} = -u + \frac{\alpha}{1+v^\beta}, \quad \frac{dv}{dt} = -v + \frac{\gamma}{1+u^\delta}$$

where α, β, γ and δ are positive parameters. The time scale has been adjusted so that the coefficients of u and v in these equations are -1.

Here is a short explanation of the form of the equations. Consider the equation for $\frac{du}{dt}$. The first term, $-u$, reflects the fact that the protein P_U is rapidly broken down within the cell, so that it will tend to disappear unless it is recreated. The second term, $\frac{\alpha}{1+v^\beta}$, represents the production of the protein P_U. This expression has a maximum value of α when v is 0, and it decreases to 0 as v increases. This general picture is reasonable, since the presence of the protein P_V represses the production of the protein P_U. The more precise form of this expression, including the meaning of the exponent β, is based on properties of specific bio-chemical reactions.

The effect of the inducer I_U is incorporated in the parameter α. As the amount of I_U in the cell is increased the parameter α will increase, since a higher concentration of I_U makes it more likely that the gene G_U will be able to produce the protein P_U.

The equation for $\frac{dv}{dt}$ has the same general interpretation.

The experiment showed that it is possible to make the gene G_U "dominant" by adding the I_U inducer, so that the concentration u of the protein P_U became very high and the concentration v of P_V was almost 0. This situation persisted after the inducer was removed from the bacterial culture. It was then possible to switch this situation, making the G_V gene dominant, by adding the inducer I_V to the system; again this configuration persisted after the inducer was removed. This explains the title of the article: The authors were able to turn their bacterial culture into a toggle switch, which could be switched between the "G_U dominant" and "G_V dominant" states, and which would stay in one state until switched to the other.

The parameters used in the experimental model are quite specific, corresponding to known characteristics of the particular genes that were used; for example, $\beta = 1$, $\gamma = 2.5$. However, the qualitative picture does not depend on the actual values of the parameters, within a fairly broad range. In order to focus on the qualitative picture we will simplify the situation, taking $\beta = \delta = 2$ and restricting α and γ to values between 0 and 10.

We need to analyze the phase portrait for various values of the parameters α and γ. To find the equilibrium points we plot the nullclines and find the points of intersection. The nullclines have the equations

$$u = \frac{\alpha}{1+v^2}, \quad v = \frac{\gamma}{1+u^2} \tag{5.9}$$

If we eliminate v between these equations we are left with a fifth degree polynomial equation for u, and each value of u is related to a single value of v by the second equation in (5.9). Hence there are at most 5 equilibrium points. In fact, there are at most 3 equilibrium points since two solutions of the fifth degree equation are non-real. There are no closed orbits: see Exercise 5.15.

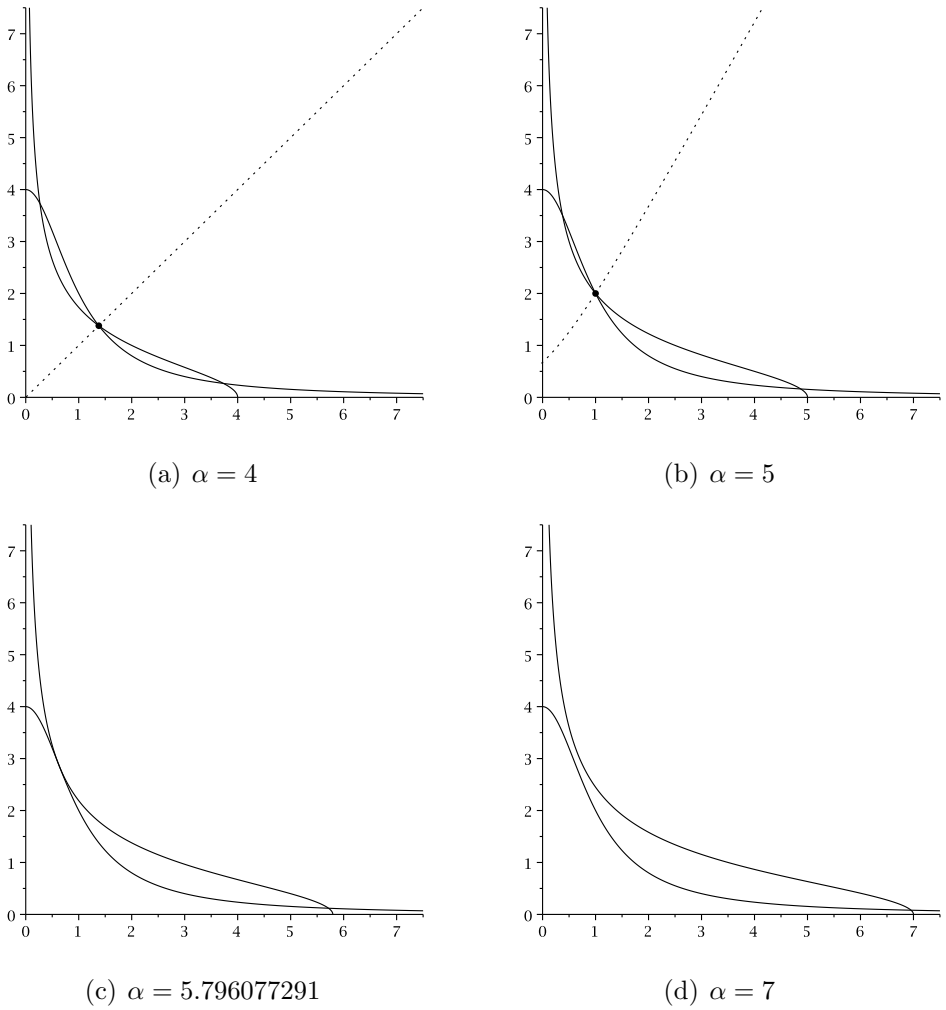
We start with Figure 5.13(a), showing the nullclines when $\alpha = 4$ and $\gamma = 4$. In this case there are three equilibrium points. If we calculate the eigenvalues of the linearizations we find that two equilibria are near $(4, 0)$ and $(0, 4)$, and these are sinks. The third equilibrium is on the line $u = v$ and it is a saddle. If we consider only points along the line $u = v$ then the vector field $f(u, v)$ becomes

$$f(u, u) = \left(-u + \frac{4}{1+u^2},\ -u + \frac{4}{1+u^2} \right).$$

Hence the vector field, at all points on the line $u = v$, points in a direction parallel to the line $u = v$. That is, the line $u = v$ is an invariant set, so it consists of orbits: the saddle point and the corresponding stable curves. Hence the line $u = v$ is a separatrix, dividing the plane into the basins of attraction for the two sinks. This separatrix is also shown in the figure.

We interpret $\alpha = 4$, $\gamma = 4$ as the "natural" state of the bacterial colony. Very quickly, in each individual cell, the state vector (u, v) will converge to one of the two sinks. There will, of course, be some cells that are on, or very close to, the line $u = v$, but this will be a very small fraction of the total population. Hence we expect that about 50% of the cells will show "G_U dominance", so they will have a (u, v) value close to $(4, 0)$, while the other 50% will show "G_V dominance", with a (u, v) value close to $(0, 4)$.

We now consider what happens as the inducer I_U is gradually added to the culture. This will have the net effect of increasing α. Since the system for $\alpha = 4$ was structurally stable (the equilibria are hyperbolic) we expect that the phase portrait will be qualitatively the same for α near 4. The nullclines for $\alpha = 5$ are shown in Figure 5.13(b), and the general features are the same as for $\alpha = 4$: There are two sinks and a saddle, and the stable curves of the saddle separate the plane into the two basins of attraction. It is not as easy to describe the separatrix, but, from the figure, it has shifted upwards, so that somewhat more than 50% of the cells will find themselves below the separatrix and hence attracted to the sink near $(5, 0)$.



(a) $\alpha = 4$ (b) $\alpha = 5$

(c) $\alpha = 5.796077291$ (d) $\alpha = 7$

Figure 5.13: Nullclines for Example 5.19. The curves are given by (5.9) with $\gamma = 4$ and α as shown. The dotted curve in plots (a) and (b) are the separatrices corresponding to the saddle points.

However, if α is increased enough then we see a different phase portrait: See Figure 5.13(d) for $\alpha = 7$. The nullclines now have only one point of intersection, which is the sink near $(7, 0)$. Thus we expect all cells to eventually converge to a state with a (u, v) value near $(7, 0)$. That is, all cells will be "G_U dominant".

Since the systems for $\alpha = 4$ and $\alpha = 7$ are structurally stable, with different phase portraits, there must be a bifurcation value for α between 4 and 7. Graphically, this

value of α is the value where the two nullclines have a transition from three points of intersection to one point of intersection. Increasing α has the effect of pulling the nullcline $v = \dfrac{\alpha}{1+v^2}$ to the right, and the last α value for which there is an intersection near the point $(0, 4)$ occurs when the two nullclines have a point of tangency. This observation gives a method for determining this value of α numerically: We need a point of intersection, so the two nullcline equations must be satisfied. Also, the nullclines must be tangent at the point of intersection, so we have a third equation, expressing the fact that the two nullclines have the same slopes. This amounts to the equation

$$\frac{d}{du}\left(\frac{4}{1+u^2}\right) = \left[\frac{d}{dv}\left(\frac{\alpha}{1+v^2}\right)\right]^{-1}.$$

Solving this equation together with the two nullcline equations (using Maple) gives the bifurcation value $\alpha \approx 5.796077291$, and this is shown in Figure 5.13(c). The equilibrium point is at about $(0.6238,\ 2.8794)$ and it is not hyperbolic; one of the eigennvalues is 0.

Graphically, the bifurcation occurs as the two equilibrium points on the left come closer together, then "coalesce" when the nullclines become tangent, and then "cancel out" as the nullclines pull apart. Calculations show that this movement of the equilibria is accompanied by a weakening of one of the eigenvalues at each point: at the sink the eigenvalues change from $-.5$, -1.5 to 0, -2, and at the saddle they change from 0.3106, -2.3106 to 0, -2.

From these diagrams we can interpret the "toggle switch" behavior in the experiments. As indicated above, as the I_U inducer is increased the bacterial population will move from a configuration evenly divided between G_U dominant and G_V dominant states to one in which essentially all bacteria show G_U dominance. If the extra I_U inducer is removed then the phase portrait will return from Figure 5.13(d) to Figure 5.13(a). However, since essentially all the bacteria are in a state very close to the sink on the right, they will stay near that sink as α decreases. Hence, even when the level of the I_U inducer has fallen to its initial values, essentially all the bacteria will have (u, v) values near $(4, 0)$, so they will show G_U dominance. This initially "sets" the toggle switch. It should then be clear that manipulating the other inducer, I_V, will push essentially all the bacteria to the sink near the v axis, and after the extra inducer is removed the bacteria will mostly stay near this sink, so the bacteria will show G_V dominance. This is how we switch the toggle between its two basic configurations.

Exercises

5.1. Using the same method as in section 5.1, find a constant of the motion for each of the following:

(a) $\dfrac{dx}{dt} = x^2(y-1)$, $\dfrac{dy}{dt} = xy$.

(b) $\dfrac{dx}{dt} = (x-1)(y-1)$, $\dfrac{dy}{dt} = \dfrac{1}{x+1}$.

(c) $\dfrac{dx}{dt} = \sin x \cos y$, $\dfrac{dy}{dt} = \tan x$.

5.2. Find a constant of the motion for the spring system, Example 4.7.

5.3. Section A.4 in the appendix reviews a method for solving an "exact differential equation". Use this method to find a constant of the motion for the following:

(a) $\dfrac{dx}{dt} = y - 2x$, $\dfrac{dy}{dt} = x^2 + 2y$.

(b) $\dfrac{dx}{dt} = y - x^2$, $\dfrac{dy}{dt} = x(1 - x + 2y)$.

5.4. Use Maple to determine the periods of several of the closed orbits in Figure 5.2.

5.5. This problem refers to the system of Figure 5.2. The island managers want to ensure that there are always an adequate number of rabbits, since they make a living by selling bunnies at Easter (this is reflected, of course, in the growth rate r). You will need to use Maple to answer this completely, but you can get a first approximation by looking at the graph.

(a) What are the maximum and minimum values for the rabbit population, if initially there are 50 foxes and 2000 rabbits?

(b) The island is in a nature preserve, but the managers have special permission to arrange a one-time only reduction of 25 in the number of foxes. Why doesn't it make sense to schedule this reduction when the foxes are at their minimum value?

(c) Approximately what is the optimum time to remove 25 foxes, in order to move the system to a state with the greatest minimum number of rabbits?

(d) After this reduction, what will be the minimum number of rabbits?

5.6. Here is yet another Rabbit Island scenario. Instead of introducing predators (foxes) to the island we introduce competitors: woodchucks. Woodchucks and rabbits eat the same kind of thing (carrots?) so we expect that the net growth rate for rabbits will be reduced by an increased number of rabbits (because of crowding) and also by an increased number of woodchucks (competition). Of course, the same thing holds from the woodchuck point of view.

This leads to the following general model: We represent the rabbit population by y and the woodchuck population by x, and we propose the following differential equation:

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} (s - \alpha x - \beta y)x \\ (r - \gamma x - \delta y)y \end{bmatrix}.$$

The parameters $r, s, \alpha, \beta, \gamma, \delta$ are all positive.

In this exercise, use the following values:

$$r = .1,\ s = .05,\ \alpha = 0.0006,\ \beta = 0.0004,\ \gamma = 0.002,\ \delta = 0.0004$$

You can use Maple, if you want, to prepare a phase portrait, but you don't really need it. You might want to use Maple to calculate eigenvalues, but it's not necessary.

(a) There are four equilibrium points, all in the closed first quadrant. Sketch the x and y nullclines and find the equilibrium points.
(b) Describe the orbits on the x and y axes. Explain why an orbit can't leave the first quadrant.
(c) The equilibrium points are all hyperbolic. Classify them. You should find 1 saddle, 1 source, 2 sinks. Is there any spiralling behavior?
(d) What can you say about the basins of attraction for the sinks? You can ignore all but the first quadrant.
(e) What can you say about the stable and unstable curves starting at the saddle?
(f) According to this model, can rabbits and woodchucks co-exist in the long run?

5.7. For each of the following, sketch the nullclines; find the equilibrium points; find the linearizations at the equilibria; and classify the equilibria.

(a) $\dfrac{dx}{dt} = 4y - y^3,\ \dfrac{dy}{dt} = x + y.$
(b) $\dfrac{dx}{dt} = x^2 - y^2,\ \dfrac{dy}{dt} = y + x^2 - 2.$
(c) $\dfrac{dx}{dt} = xy + 6,\ \dfrac{dy}{dt} = x^2 - 7 - y.$

5.8. Start with the second order differential equation $\dfrac{d^2x}{dt^2} + 2x^3 = 0.$

(a) Convert this to a system of differential equations, using x and $y = \dfrac{dx}{dt}$ as the variables.
(b) This system has only one equilibrium, at the origin. Calculate the linearization at the origin and verify that the system is not hyperbolic.
(c) Show that $\varphi(x, y) = x^4 + y^2$ is a constant of the motion.

(d) What can you say about the orbits?

5.9. Start with the second order differential equation $x\dfrac{d^2x}{dt^2} - \dfrac{dx}{dt} + x^2 - x^3 = 0$.

(a) Convert this to a vector differential equations in the form $\dfrac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = f(x, y)$, using x and $y = \dfrac{dx}{dt}$ as the variables. What happens when $x = 0$?

(b) Here's another way to convert the second order DE to a system: Divide the equation by x^2, notice that $\dfrac{1}{x}\dfrac{d^2x}{dt^2} - \dfrac{1}{x^2}\dfrac{dx}{dt} = \dfrac{d}{dt}\left(\dfrac{1}{x}\dfrac{dx}{dt}\right)$, and use the variables x and $z = \dfrac{1}{x}\dfrac{dx}{dt}$. Find the differential equations in terms of x and z.

(c) Part (b) produces a differential equation for x and z with one equilibrium. Find the equilibrium and classify it.

(d) What happens in the x-z system when $x = 0$?

5.10. Describe what happens in Example 5.7 when condition (5.4) is replaced by the condition

$$r/\delta < s/\alpha.$$

In particular,

(a) Find and classify the equilibria.

(b) Sketch the phase portrait in the first quadrant.

(c) What does this system predict about the long term populations of rabbits and foxes?

If you prefer, do this for a specific numeric example: Use the parameters in Figure 5.6, but change δ to $\delta = .0001$.

5.11. Consider the following equations:

$$\frac{dx}{dt} = 4x - 3y - x(x^2 + y^2), \quad \frac{dy}{dt} = 3x + 4y - y(x^2 + y^2).$$

(a) There are unique constants a and b so that $a > 0$ and $x = a\cos(bt)$, $y = a\sin(bt)$ is a solution of this system. Plug these formulas for x and y into the differential equations and determine a and b. Thus one orbit of this system is a circle of radius a, centered at the origin.

(b) Let $\varphi(x, y) = x^2+y^2 = r^2$. Calculate $\dfrac{d\varphi}{dt}$ and write it in the form $r^2(A-Br^2)$ for some constants A and B.

(c) Use part (b) to explain why your solution from part (a) is a two-sided limit cycle.

(d) Sketch the phase portrait.

5.12. This problem refers to the van der Pol system with $\mu = 1$, and it explains why the orbits which cross the cubic $y = F(x) = \frac{1}{3}x^3 - x$ stay very close to the cubic until they hit the x axis.

Suppose y_1 is any positive number. Define the region R to have the following boundary curves:

top:	$y = y_1$
left:	$y = F(x)$
right:	$y = F(x) - 2/3$
bottom:	$y = 0$

(a) Make a careful sketch of the region R; you can choose $y_1 = 4$ (or any larger value). The left and the right curves will be very close to each other.

(b) Check that the left and right curves intersect the x axis at $x = \sqrt{3}$ and $x = 2$.

(c) Using the facts that $\dfrac{dy}{dt} = -x < 0$ for $x > 0$ and $\dfrac{dx}{dt} = y - F(x) = 0$ on the cubic, check that solution curves enter R on its top and left boundaries, and exit on the bottom boundary.

(d) Let $\psi(x, y) = y - F(x)$. Check that $\dfrac{d\psi}{dt} = -x - F'(x)\psi(x, y)$ along any solution curve.

(e) Check that $\dfrac{d\psi}{dt} \geq 0$ on the right boundary of R. Explain why this implies that solutions enter R on the right boundary. [Notice that the left boundary corresponds to $\psi = 0$ and the right boundary corresponds to $\psi = -2/3$.]

(f) Using $\dfrac{dy}{dt} < 0$, explain why R cannot contain any equilibrium points or closed orbits.

(g) Finally, explain why any solution curve which hits the cubic at any point (x_1, y_1) in the first quadrant must intersect the x axis between $\sqrt{3}$ and 2.

5.13. Start with the differential equations $\dfrac{dx}{dt} = x^2,\ \dfrac{dy}{dt} = -y$.

(a) This system has exactly one equilibrium point, at the origin. Show that this is not a hyperbolic equilibrium by calculating the eigenvalues at the origin.

(b) Sketch the flow on the x axis.

(c) Now perturb the equations slightly, to $\dfrac{dx}{dt} = x^2 + \varepsilon,\ \dfrac{dy}{dt} = -y$, where ε is a small constant. Show that if ε is positive then the system has no equilibrium points.

(d) Show that, if ε is negative, say $\varepsilon = -\delta^2$, the system has two equilibrium points, both hyperbolic.
(e) Sketch the flow on the x axis for the perturbed system with negative ε.

5.14. Start with the differential equations $\dfrac{dx}{dt} = x^3,\ \dfrac{dy}{dt} = -y$.

(a) This system has exactly one equilibrium point, at the origin. Show that this is not a hyperbolic equilibrium by calculating the eigenvalues at the origin.
(b) Sketch the flow on the x axis.
(c) Consider the perturbed system $\dfrac{dx}{dt} = x^3 + \varepsilon x,\ \dfrac{dy}{dt} = -y$, where ε is a small constant. How many equilibrium points does the perturbed system have? Your answer will depend on the sign of ε.
(d) Any small perturbation of this system must have at least one equilibrium point. To show this, draw the square S with vertices $(\pm 1, \pm 1)$, and show graphically that the flow on the boundary of S always points into the square. This will still be true if the system is slightly perturbed. Now cite a theorem in section 5.5.

5.15. The dynamical system of Example 5.19, with any choice of parameters, has no closed orbits. Show this by contradiction: If there is a closed orbit, apply the Divergence Theorem to the vector field $f(u, v)$ and the region enclosed by the closed orbit. [The Divergence Theorem in two dimensions is also known as Green's Theorem. There are two versions of Green's Theorem – you need the one that relates the divergence of f to the *normal* component of f on the boundary curve.]

5.16. Here is another bifurcation in Example 5.19, with $\beta = \delta = 2$. Suppose the inducers are simultaneously adjusted, so that $\alpha = \gamma$ remains true as α varies. The phase portrait for $\alpha = \gamma = 4$ was analyzed in the text; see Figure 5.13(a).

(a) Find and analyze the equilibrium points when $\alpha = \gamma = 1$.
(b) There is a bifurcation when $\alpha = \gamma = \alpha_1$ for some α_1 between 1 and 4. Find α_1.
(c) Find and analyze the equilibrium points at the bifurcation.

CHAPTER 6

Chaos

6.1. The Hénon Map

For many years there was a general feeling among mathematicians that "most" dynamical systems are "simple". For example, for differential equations the expectation was that most systems had the following features:

(1) There is a discrete set of equilibrium points and closed orbits. (A collection of sets is "discrete" if there only finitely many of the sets meet any bounded region.)
(2) The equilibrium points and closed orbits are hyperbolic. (Hyperbolicity for closed orbits will be defined later.)
(3) "Almost all" orbits either converge to ∞ or converge to a sink or attracting closed orbit. (An attracting closed orbit is the higher-dimensional analog of a two-sided limit cycle.)
(4) The system is structurally stable.

Of course, not all systems satisfy these conditions. For example, the spring example, 4.7, does not have a discrete set of closed orbits; the closed orbits and half the equilibria are not hyperbolic; there are no sinks or attracting closed orbits; and the system is not structurally stable. However, a small change in the equations changes the spring to the "sticky" spring example, 4.8, and this satisfies all the conditions for a simple dynamical system. This is the sense in which it was thought that "most" systems are simple: If a system is not simple then it sshould become simple after an arbitrarily small perturbation.

Also notice that, in example 4.8, all orbits except the saddle points and their stable curves converge to one of the sinks as $t \to +\infty$. We can describe this as follows: Each sink or attracting closed orbit has a basin of attraction, consisting of all points that converge to it as $t \to +\infty$. Then the condition that "almost all" orbits either converge to ∞ or converge to a sink or attracting closed orbit means:

(1) Each basin of attraction is *open*. This means that if an orbit converges to a sink or attracting closed orbit then every nearby orbit will converge to the same place.
(2) The set of points that converge to infinity or lie in some basin of attraction is *dense*. This means that if an orbit does not converge to ∞ or a sink or attracting

closed orbit then there are orbits starting arbitrarily close to it which do onverge to ∞ or lie in a basin.

The van der Pol equation, 5.14, is an example in which all points, except the source at the origin, lie in the basin of attraction of the limit cycle. If we run van der Pol's system backwards then the closed orbit becomes a repeller and the origin becomes a sink. The basin of attraction of the sink is the interior of the closed orbit, and all points outside the closed orbit converge to ∞ as $t \to \infty$.

In fact, this picture holds for two-dimensional differential equations, but in general something much more complex can happen. The conditions for simple dynamics can be violated in a *persistent* way, so that the violations cannot be fixed by small perturbations of the system. A specific kind of violation is called "chaos".

Chaos was discovered – and largely forgotten – several times. It was first described by Henri Poincaré in about 1890 in his theoretical analysis of the "three body problem" of classical mechanics. It was rediscovered experimentally by Balthasar van der Pol in the 1920's in the "periodically forced" van der Pol equation, which occured 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.

We'll describe the horseshoe, and analyze some of its chaotic features, in the next section. Here we will look at a simple example, due to Michel Hénon in 1976, that exhibits a horseshoe. This example is now known as the Hénon map.

This is a discrete dynamical system, defined on the plane. Its general form is

$$F(x, y) = (A + By + Cx^2,\ Dx) \tag{6.1}$$

where A, B, C and D are non-zero parameters. We will concentrate on parameters that determine a horseshoe. Hénon was looking for more complex dynamics, and he used parameters that determine a *strange attractor*, which we will describe later.

We will work with the specific map H defined by

$$H(x, y) = (\alpha - \beta y - x^2,\, x), \text{ with } \alpha = 7.5,\, \beta = .8. \tag{6.2}$$

We will frequently use α and β instead of their numeric values when it will simplify the calculations.

H is an invertible dynamical system. To see this we need to determine the inverse function, H^{-1}. So write $H(x, y) = (X,\, Y)$ and then solve for x and y in terms of X and Y:

$$\begin{aligned}
H(x, y) &= (X,\, Y) && \\
(\alpha - \beta y - x^2,\, x) &= (X,\, Y) && \text{definition of } H(x, y) \\
\alpha - \beta y - x^2 &= X \text{ and } x = Y && \text{equating components} \\
\alpha - \beta y - Y^2 &= X \text{ and } x = Y && \text{substitute } x = Y \\
y &= \alpha\beta^{-1} - \beta^{-1}X - \beta^{-1}Y^2 \text{ and } x = Y && \text{solve for } y \\
(x,\, y) &= (Y,\, \alpha\beta^{-1} - \beta^{-1}X - \beta^{-1}Y^2) && \text{as ordered pairs} \\
H^{-1}(X, Y) &= (Y,\, \alpha\beta^{-1} - \beta^{-1}X - \beta^{-1}Y^2) && \text{definition of inverse.}
\end{aligned}$$

If we now replace the variables X and Y with x and y, we have the following formula for H^{-1}:

$$H^{-1}(x, y) = (y,\, \alpha\beta^{-1} - \beta^{-1}x - \beta^{-1}y^2) = (y,\, 9.375 - 1.25x - 1.25y^2). \tag{6.3}$$

Notice that this is another Hénon function, except that the roles of x and y have been interchanged.

We can start to investigate this system by looking at the orbits of points. If $p_0 = (x_0, y_0)$ is an initial point then the *orbit* of p_0 is the set of all iterates of p_0 under H and H^{-1}. That is, the orbit of p_0 consists of the bi-infinite sequence

$$\ldots,\ p_{-2} = H^{-2}(p_0),\ p_{-1} = H^{-1}(p_0),\ p_0,\ p_1 = H(p_0),\ p_2 = h^2(p_0),\ \ldots$$

For example, if $p_0 = (0, 0)$ then

$$\begin{aligned}
p_{-2} &= H^{-2}(0, 0) = H^{-1}(H^{-1}(0, 0)) = H^{-1}(0, 9.375) \approx (9.375, -100.488) \\
p_{-1} &= H^{-1}(0, 0) = (0,\, \alpha\beta^{-1}) = (0, 9.375) \\
p_0 &= (0,\, 0) \\
p_1 &= H(0, 0) = (\alpha, 0) = (7.5, 0) \\
p_2 &= H^2(0, 0) = H(H(0, 0)) = H(7.5, 0) = (-48.75,\, 7.5).
\end{aligned}$$

It appears that $H^k(p_0)$ is going rapidly to ∞ as $k \to +\infty$, and also as $k \to -\infty$.

However, not alll orbits go to infinity. For example, H has fixed points, as follows:

$$\begin{aligned} H(x,y) &= (x,\, y) \\ (\alpha - \beta y - x^2,\, x) &= (x,\, y) \\ \alpha - \beta y - x^2 &= x \text{ and } y = x \\ \alpha - \beta x - x^2 &= x && \text{substitute } x = y \\ x^2 + (\beta+1)x - \alpha &= 0 && \text{rearrange} \\ x = y &= \tfrac{1}{2}\left(-(\beta+1) \pm \sqrt{(\beta+1)^2 + 4\alpha}\right) && \text{quadratic formula.} \end{aligned}$$

With $\alpha = 7.5$ and $\beta = .8$ this gives the fixed points

$$(1.982707061,\ 1.982707061), \quad (-3.782707061,\ -3.782707061) \quad \text{(approximately)}.$$

If p_0 is one of these fixed points then $H^n(p_0) = p_0$ for all n, so the orbit starting at p_0 consists of only one point.

More generally, suppose F is any invertible dynamical system. A *periodic point* of F is a point x_0 so that, for some $m > 0$, $F^m(x_0) = x_0$. We say m is a *period* of the periodic point x_0 if $m > 0$ and $F^m(x_0) = x_0$. The *least period* of a periodic point x_0 is the smallest period of x_0. For example, if x_0 is a fixed point then $F^m(x_0) = x_0$ for all $m > 0$, so x_0 has period m for all positive m, but it has least period 1.

Here are some simple facts about periodic points:

PROPOSITION 6.1. *For any dynamical system F:*

(a) *If x_0 has period m then it has period km for all $k > 0$.*
(b) *If x_0 has least period m and n is any other period of x_0 then $n = km$ for some integer k.*
(c) *If x_0 has period m then $F^j(x_0)$ also has period m, for any non-negative j. If F is invertible then j can be negative.*
(d) *If F is invertible then x_0 is invertible if and only if the orbit of x_0 is finite; and in this case the least period of x_0 is the number of points in its orbit. (If F is not invertible then the same is true, but just considering the forward orbit of x_0).*

Returning to H, we can look for periodic points. For points of period 2 we need to solve the equation $H^2(x,y) = (x,\, y)$, or

$$(\alpha - \alpha^2 - \beta x + 2\alpha\beta y + 2\alpha x^2 - \beta^2 y^2 - 2\beta y x^2 - x^4,\ \alpha - \beta y - x^2) = (x,\, y).$$

This leads to a fourth degree equation for x. Fortunately, this is solvable: We already know two solutions, since the fixed points are also points of period 2. So we can factor out the linear terms corresponding to the x coordinates of the two fixed

points, leaving a quadratic equation. It turns out that the roots of this are real, so we get two points of least period 2:

$$(3.15166605,\ -1.35166605), \quad (-1.35166605,\ 3.15166605) \quad \text{(approximately)}.$$

Remember that these two points represent one orbit: H transformas each of these points into the other.

It becomes more difficult to try to find periodic points of higher order, since the degree of the equations becomes very large. For example, to calculate the points of period 5 we would need to solve an equation of degree 32.

We would like to determine all the periodic points. More generally, we would like to determine which orbits remain bounded. To do this it is necessary to consider the geometry of the Hénon map.

We can calculate $H(p)$ where p is a point. Now we want to see how to calculate $H(S)$ where S is a set of points. By definition, $H(S)$ is the set of all points of the form $H(p)$, where p is in S. For this reason $H(S)$ is sometimes called the *image* of S under H. Let's first consider the image of a curve, C. Here is a procedure for determining the curve $H(C)$:

(1) Write C parametrically, as $(x, y) = (x(t), y(t))$.
(2) Calculate $H(x(t), y(t))$; then the parametric form of the image $H(C)$ is $(x, y) = H(x(t), y(t))$.
(3) If desired, eliminate t in this parametric form to get an equation of $H(C)$ in terms of x and y.

For the first example, let C be a vertical straight line with equation $x = c$ for some constant c. Then this can be written parametrically by setting $y = t$, so the parametric form is $x = c,\ y = t$. Now calculate $H(c, t) = (\alpha - \beta t - c^2,\ c)$. So the parametric equations for $H(C)$ are $x = \alpha - \beta t - c^2,\ y = c$. As t varies x will take on all possible values, but y will stay at c. So the image curve is the horizontal line with equation $y = c$. So H transforms every vertical line into a horizontal line. Moreover, every vertical segment will be transformed to a horizontal segment, so to determine the image of a vertical segment it is only necessary to determine where the endpoints go.

Next, suppose C is a horizontal straight line with equation $y = c$. This can be written parametrically as $x = t,\ y = c$, and $H(t, c) = (\alpha - \beta c - t^2,\ t)$. Hence the parametric equations for $H(C)$ are $x = \alpha - \beta c - t^2,\ y = t$. We can eliminate t form these equations: Since $y = t$ we just replace t with y in the first equation to get $x = \alpha - \beta c - y^2$. Since $\alpha - \beta c$ is jsut a constant, we recognize this as the equation of a parabola, with axis equal to the x axis, opening to the right. So H transforms the horizontal line $y = c$ into the parabola

$$(*) \qquad x = \alpha - \beta c - y^2.$$

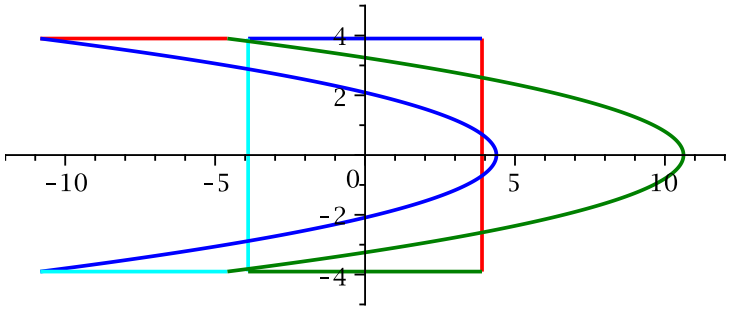


Figure 6.1: Q and $H(Q)$.

Now if we have a geometric figure S, bounded by curves, we can calculate $H(S)$ by calculating the image of its boundary curves using the procedures above. We will be very interested in the square Q centered at the origin, with sides at $(\pm s, \pm s)$, where s has the specific value 3.9. Now the right hand side of Q is a vertical segment, from $(s, -s)$ to (s, s), so, by the discussion above, H transforms this to a horizontal segment with endpoints at

$$H(s, -s) = (\alpha - \beta \cdot (-s) - s^2,\ s) = (-4.59,\ 3.9)$$
$$H(s, s) = (\alpha - \beta \cdot s - s^2,\ s) = (-10.83,\ 3.9).$$

Similar considerations show that the image of the left side of S is the horizontal segment between $(-4.59,\ -3.9)$ and $(-10.83,\ -3.9)$.

Next we consider the top, which is a segment of the horizontal line $y = s$. Replacing c with s in $(*)$, the image of this horizontal line is the parabola

$$x = \alpha - \beta \cdot s - y^2 = 7.5 - .8 \cdot 3.9 - y^2 = 4.38 - y^2.$$

Notice that the vertex of this parabola is at $(4.83, 0)$, and this is to the *right* of Q. The image of the bottom of Q also lies on a parabola, with equation $x = \alpha - \beta \cdot (-s) - y^2 = 10.62 - y^2$. The images of the top and bottom of Q are determined as the portions of these parabolas that are bounded by the images of the corners of Q.

Hence the boundary of $H(Q)$ consists of two horizontal line segments, connected by two portions of parabolas, and $H(Q)$ itself is the region inside this region. This is illustrated in figure 6.1. Note that some of the points in $H(Q)$ lie outside Q. It is shown in Exercise 6.2 that these points will never re-enter Q under H^n, for any $n > 0$. A little more work shows that these points go to ∞ as $n \to +\infty$.

The choice of $s = 3.9$ was dictated by the requirement that the images of the left and right sides of Q are strictly to the left of Q, and the images of the top has

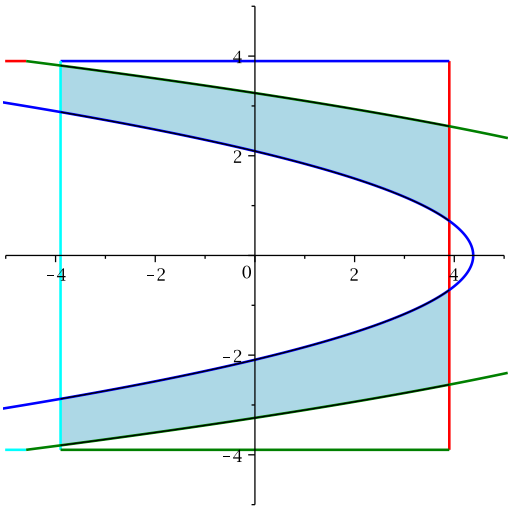


Figure 6.2: $Q \cap H(Q)$.

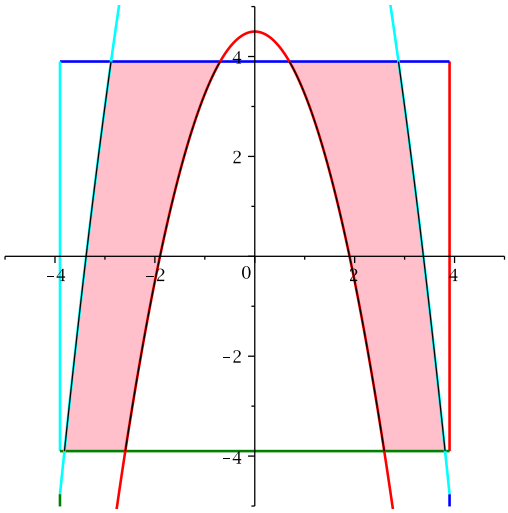


Figure 6.3: $Q \cap H^{-1}(Q)$.

its vertex strictly to the right of Q. This has the consequence that the intersection $Q \cap H(Q)$ consists of two disjoint connected regions, each extending from the left side of Q to the right edge. This is shown in Figure 6.2.

Recall that the inverse function H^{-1} has the same general form as H, if we interchange the roles of x and y. If we repeat the analysis above we find that the image of Q under H^{-1} looks very much like the image of Q under H, but now the parabolic curves have their axes on the y axis, and they open downward. If we calculate $Q \cap H^{-1}(Q)$ we are led to Figure 6.3. There are again two connected components of this intersection, but now they extend from the top to the bottom of Q. The points of $H^{-1}(Q)$ that leave Q under H^{-1} never return under applications of H^n for $n < 0$, and, in fact, they go to ∞ as $n \to -\infty$.

It follows that all points that have bounded orbits must lie in Q. In fact, their entire orbits are in Q. Since H and H^{-1} transform orbits into the same orbit, any orbit that lies entirely in Q must lie entirely in both regions indicated in Figures 6.2 and 6.3. That is, any orbit that lies entirely in Q must lie in the intersection of these regions, $H^{-1}(Q) \cap Q \cap H(Q)$. This intersection consists of 4 components, as shown in Figure 6.4.

If we plot the fixed points of H we will find that one is in the lower left component of $H^{-1}(Q) \cap Q \cap H(Q)$, and the other is in the upper right component. The periodic orbit of least period 2 has one point in each of the other two components.

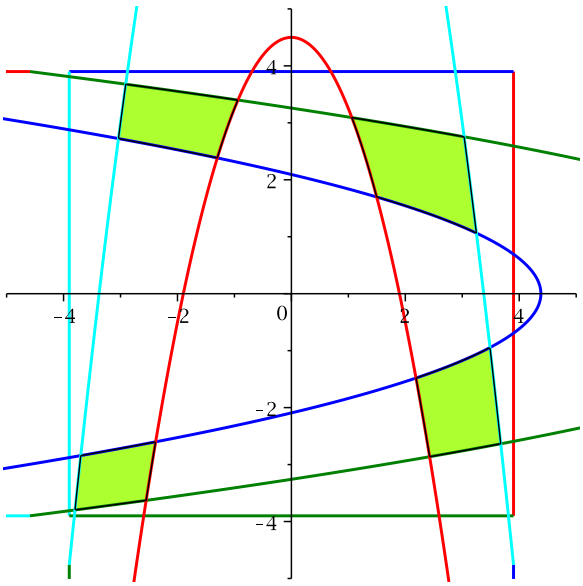


Figure 6.4: $H^{-1}(Q) \cap Q \cap H(Q)$.

6.2. The horseshoe

Even though it was introduced before the Hénon map, we can consider the horseshoe as a "linearization" of the Hénon map. The idea is to describe a system that captures the geometry of the Hénon map, but which is simple enough to analyze directly. Once the horseshoe is understood a great variety of more "natural" dynamical systems – like the Hénon map – can be studied as distortions of the horseshoe.

The horseshoe map can be described geometrically as follows. We start with a rectangle Q, with sides parallel to the x and y axes. Then we transform Q in four steps:

(1) Shrink Q vertically by a constant factor, μ, to give a new rectangle, R_1.
(2) Stretch R_2 horizontally by a constant factor, λ, to produce the rectangle R_2.
(3) Bend R_2 in a thin vertical subrectangle at the middle of R_2 to produce a U-shaped figure, U.
(4) Now slide U back over the initial Q, and call the result H. If μ is small enough, λ is big enough, and the bending is done carefully the resulting figure describes the horseshoe map.

This process is illustrated in Figure 6.5. The critical features are:

(a) The bent region is entirely outside Q.

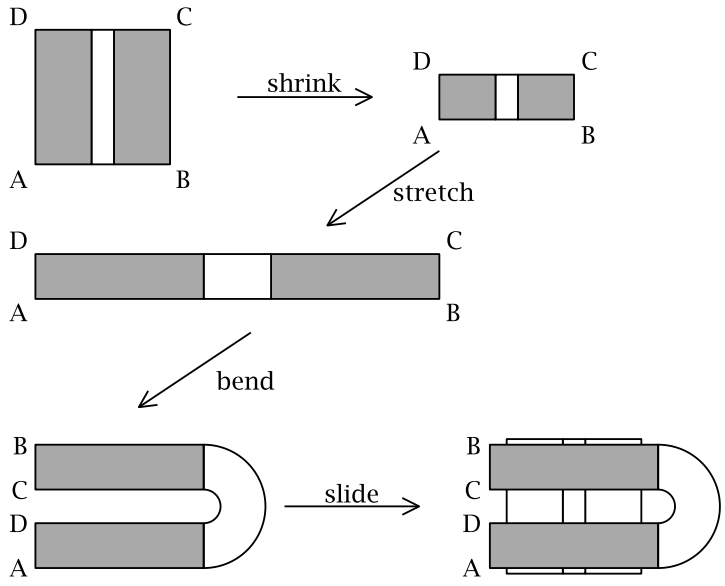


Figure 6.5: Making a horseshoe.

(b) The straight sides of H are rectangles parallel to the sides of Q, stretching entirely accross Q from left to right, and lying between the top and bottom sides of Q.

It is necessary for this construction that $\mu < \frac{1}{2}$, since otherwise there the straight sides of H will be too thick to fit between the top and bottom of Q. It is also necessary that $\lambda > 2$, for otherwise the two straight sides of H will not stretch entirely across Q. Moreover, it is necessary that $\lambda * (w - b) > 2w$, where w is the width of Q and b is the width of the rectangle where bending takes place.

On the parts of Q that map to the straight sides of H we can give a precise description of the transformation F. The actions of vertical shrinking by a factor of μ and horizontal stretching by a factor of λ are linear transformations, and correspond to multiplication by the matrices $\begin{bmatrix} 1 & 0 \\ 0 & \mu \end{bmatrix}$ and $\begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix}$, so the first two steps in the description of the horseshoe map is the linear transformation defined by the product

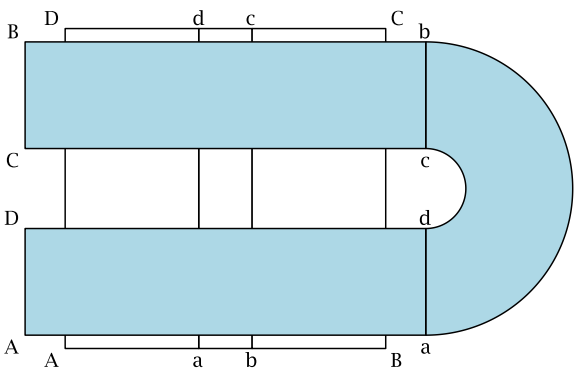


Figure 6.6: The horseshoe map.

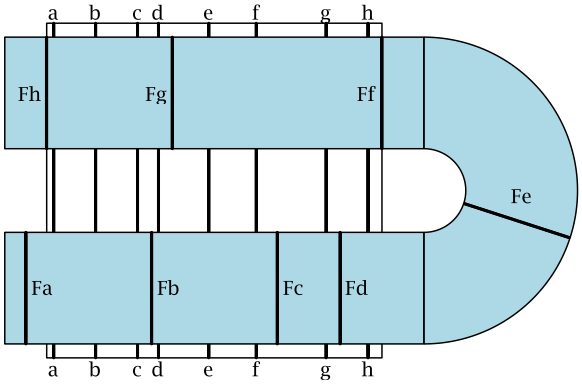


Figure 6.7: Images of vertical lines.

of these matrices, or $\begin{bmatrix}\lambda & 0\\ 0 & \mu\end{bmatrix}$. Bending as shown in Figure 6.5 will not affect the transformation to the left of the bending region, but on the right of the bending region the image rectangle has been rotated by 180°. A rotation of 180° is the linear transformation corresponding to the matrix $\begin{bmatrix}-1 & 0\\ 0 & -1\end{bmatrix}$, so to the right of the bending region the effect of steps 1–3 is given by the product $\begin{bmatrix}-1 & 0\\ 0 & -1\end{bmatrix}\cdot\begin{bmatrix}\lambda & 0\\ 0 & \mu\end{bmatrix} = \begin{bmatrix}-\lambda & 0\\ 0 & -\mu\end{bmatrix}$. Finally, the last step is a translation, which corresponds to adding constant vectors to our linear transformations. A function which is defined by a linear transformation followed by a translation is called an *affine* transformation.

The remaining points of Q are in the center rectangle where bending occurs. All we need to know about these points is that F maps them into the "bent" part of H, and this part is disjoint from Q.

Summarizing: Q is the union of three subrectangles, L, B, R, arranged from left to right, and there are constant vectors v_L and v_R so that

$$F(x,y) = \begin{bmatrix}\lambda & 0\\ 0 & \mu\end{bmatrix}\cdot\begin{bmatrix}x\\ y\end{bmatrix} + v_L \qquad \text{if } (x,y)\in L, \tag{6.4}$$

$$F(x,y) = \begin{bmatrix}-\lambda & 0\\ 0 & -\mu\end{bmatrix}\cdot\begin{bmatrix}x\\ y\end{bmatrix} + v_R \qquad \text{if } (x,y)\in R, \tag{6.5}$$

$$F(x,y) \notin Q \qquad \text{if } (x,y)\in B. \tag{6.6}$$

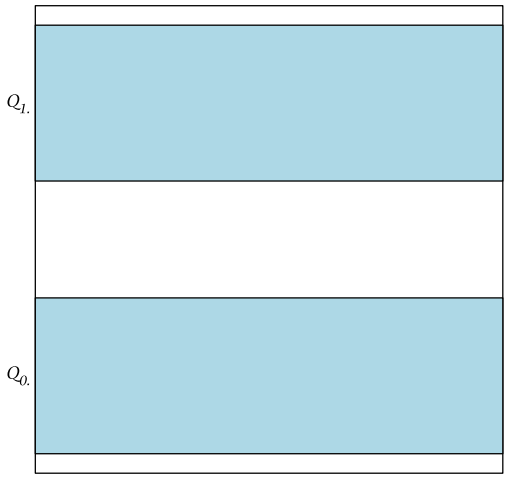


Figure 6.8: $F(Q) \cap Q$.

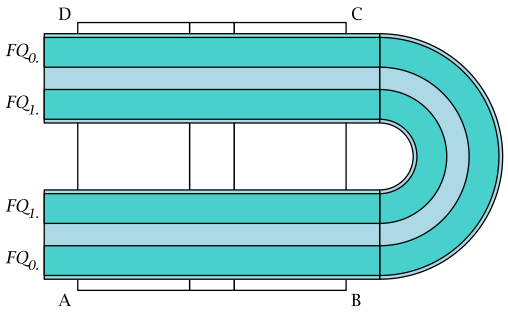


Figure 6.9: The image of $F(Q) \cap Q$.

This is illustrated in Figure 6.6. The bending takes place in the rectangle $B = abcd$. The other subrectangles are $L = AadD$ and $R = bBCc$. Notice that the image of the rectangle $bBCc$ is rotated 180° and that the image of the rectangle $abcd$ is entirely outside Q.

In all the examples in this section Q is actually a square, and $\lambda = 2,\ \mu = \frac{1}{3}$.

Equations (6.4) and (6.5) describe the action of F as multiplication by a diagonal matrix, followed by a translation. Multiplication by one of these diagonal matrices amounts to vertical and horizontal scale changes, by μ and λ, possibly with a change in orientation. Hence this operation converts a vertical segment of length ℓ to a new vertical segment of length $\mu \cdot \ell$, and the translation part of the transformation converts this vertical segment to another vertical segment of the same length. Figure 6.7 shows the images of the vertical segments $a, b, \ldots, h$ under F. The segments a and e are mapped outside of Q, and the segments f, g, h are mapped to the upper leg of the horseshoe. Note that f, g, h are arranged from left to right, but their images appear from right to left; also the orientation on each of these images has been reversed. These are the expected consequences of the change in orientation in equation (6.5).

We are interested in points which remain in Q under all iterations of F, and we start with $F(Q) \cap Q$. From Figure 6.6 it is clear that this intersection consists of two rectangles which stretch from the left boundary of Q to the right boundary; these are shown in Figure 6.8. The lower of these two rectangles is labeled $Q_{0\,\cdot}$ and the

upper is labeled $Q_{1.}$. The subscripts are *not* numbers, but labels, and the decimal point in each subscript is part of the label. There will be many more labels that are based on this initial labeling of the two components of $F(Q) \cap Q$.

We now consider applying F to these two sets. In each case the result is a thinner version of $F(Q)$, as shown in Figure 6.9. Each of these images intersects Q in a horizontal rectangle, so there are four of these. In set terms we are looking at $Q \cap F(F(Q) \cap Q) = F^2(Q) \cap F(Q) \cap Q$, and this is a subset of $F(Q) \cap Q = Q_{0.} \cup Q_{1.}$. It should be clear that two of these four rectangles lie in $Q_{0.}$ and two in $Q_{1.}$. We label these new rectangles by the rule

$$Q_{st.} = F(Q_{s.}) \cap Q_{t.}, \quad \text{where } s, t \in \{ 0, 1 \}.$$

This is illustrated in Figure 6.10. We then continue this process. For example, $F(Q_{00.})$ consists of two even thinner rectangles, one of which lies in $Q_{0.}$ and the other in $Q_{1.}$, and we label these as $Q_{000.}$ and $Q_{001.}$. There are eight such rectangles, as shown in Figure 6.11.

In general we define horizontal rectangles $Q_{\rho.}$ where ρ is a sequence of 0's and 1's. There are 2^k such rectangles at the k^{th} generation, when the sequence ρ has length k, and we get the $(k+1)^{\text{st}}$ generation of rectangles by the rule

$$Q_{\rho 0.} = F(Q_{\rho.}) \cap Q_{0.}, \quad Q_{\rho 1.} = F(Q_{\rho.}) \cap Q_{1.}$$

The union of all these rectangles is $F^k(Q) \cap F^{k-1}(Q) \cap \cdots \cap F^2(Q) \cap F(Q) \cap Q$. Notice that each of these rectangles has width w and height $\mu^k h$ where Q has width w and height h.

Next we consider images of Q under the inverse function F^{-1}. If we are careful constructing F then we can arrange that it is invertible, and we will find that F^{-1} transforms Q to a horseshoe figure, but now oriented vertically. Hence the intersection $Q \cap F^{-1}(Q)$ will consist of two vertical rectangles, which we shall call $Q_{.0}$ and $Q_{.1}$; these are illustrated in Figure 6.12. We can see this from considerations of vertical segments under F: In Figure 6.7 notice that the vertical lines labeled f and h map into the right and left edges of Q. Let S be the rectangle bordered by f and h and the top and bottom of Q. Then $F(S) \subset Q$, and applying F^{-1} to this inclusion shows that $S \subset F^{-1}(Q)$. In fact, S is the rectangle labeled $Q_{.1}$ in Figure 6.12.

Here's another way to see what $Q \cap F^{-1}(Q)$ looks like. If we apply F^{-1} to $F(Q) \cap Q = Q_{0.} \cup Q_{1.}$ we get

$$Q \cap F^{-1}(Q) = F^{-1}\left(F(Q) \cap Q\right) = F^{-1}(Q_{0.} \cup Q_{1.}) = F^{-1}(Q_{0.}) \cup F^{-1}(Q_{1.})$$

Now F^{-1} on the set $Q_{0.}$ acts as multiplication by $\begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}^{-1} = \begin{bmatrix} \lambda^{-1} & 0 \\ 0 & \mu^{-1} \end{bmatrix}$, followed by a translation. Now $Q_{0.}$ is a rectangle of width w and height μh (where Q has

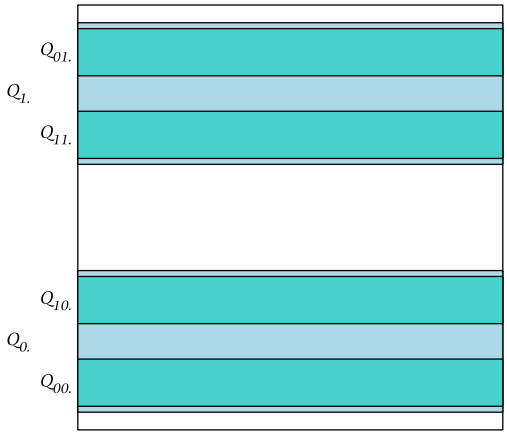


Figure 6.10: $F^2(Q) \cap F(Q) \cap Q$.

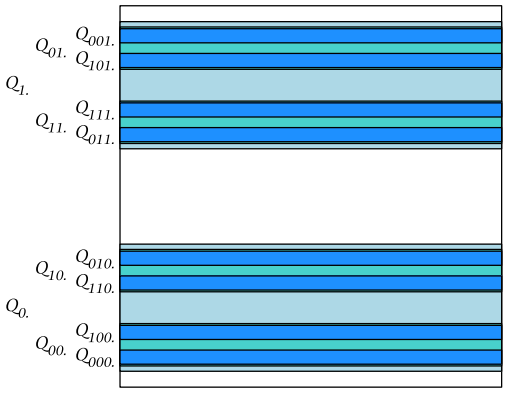


Figure 6.11: $F^3(Q) \cap F^2(Q) \cap F(Q) \cap Q$.

dimensions $w \times h$), so multiplication by the matrix $\begin{bmatrix} \lambda^{-1} & 0 \\ 0 & \mu^{-1} \end{bmatrix}$ transforms this into a rectangle of width λ^{-1} and height $\mu^{-1}\mu h = h$. After the translation part of F^{-1} we have a rectangle of width $\lambda^{-1}w$ stretching from the top to the bottom of Q, and this is what we labeled $Q_{\,.0}$. So $F^{-1}(Q_{0\,.}) = Q_{\,.0}$, or $Q_{0\,.} = F(Q_{\,.0})$. The rule here is that application of F^{-1} to $Q_{0\,.}$ *moves the decimal point one position to the left* to produce $Q_{\,.0}$; or, in the other direction, application of F to $Q_{\,.0}$ *moves the decimal point one position to the right* to produce $Q_{0\,.}$. This gives a simple way to define the rectangles $Q_{\,.\rho}$ where ρ is a sequence of 0's and 1's of length k: we start with $Q_{\rho.}$ and move the decimal point k positions by applying F^{-1} k times. That is,

$$Q_{\,.\rho} = F^{-k}(Q_{\rho\,.}) \quad \text{where } \rho \text{ has length } k.$$

Then, if the length of the sequence ρ is fixed at k, there are 2^k such sequences, each of width $\lambda^k w$ and height h, and the union of all these rectangles is

$$\begin{aligned} F^{-k}(F^k(Q) \cap F^{k-1}(Q) \cap \cdots \cap F(Q) \cap Q) \\ &= F^{-k}(F^k(Q)) \cap F^{-k}(F^{k-1}(Q)) \cap \cdots \cap F^{-k}(F(Q)) \cap F^{-k}(Q) \\ &= Q \cap F^{-1}(Q) \cap F^{-2}(Q) \cap \cdots \cap F^{-k+1}(Q) \cap F^{-k}(Q). \end{aligned}$$

Figure 6.13 illustrates this fir $k = 4$.

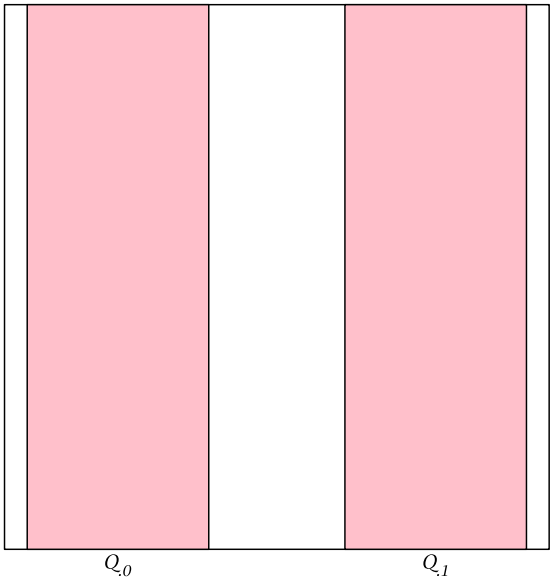


Figure 6.12: $Q \cap F^{-1}(Q)$.

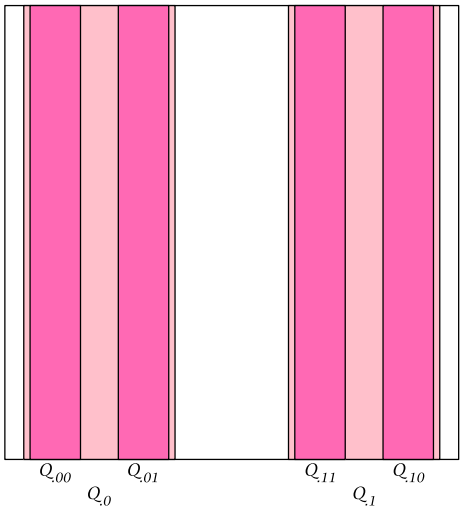


Figure 6.13: $Q \cap F^{-1}(Q) \cap F^{-2}(Q)$.

There is one more extension to this notation. If we have a horizontal rectangle $Q_{\rho\cdot}$ and a vertical rectangle $Q_{\cdot\tau}$ then their intersection is labeled as follows:

$$Q_{\rho\cdot\tau} = Q_{\rho\cdot} \cap Q_{\cdot\tau}$$

If ρ has length j and τ has length k then $Q_{\rho\cdot}$ has height $\mu^j h$ and $Q_{\cdot\tau}$ has width $\lambda^{-k} w$, so $Q_{\rho\cdot\tau}$ has height $\mu^j h$ and width $\lambda^k w$. Since $\lambda > 1$ and $0 < \mu < 1$ the dimensions of $Q_{\rho\cdot\tau}$ approach 0 as j and k both tend to ∞. There are 2^{j+k} such rectangles, and their union is the intersection

$$F^j(Q) \cap F^{j-1}(Q) \cap \cdots \cap F(Q) \cap Q \cap F^{-1}(Q) \cap \cdots \cap F^{-k+1}(Q) \cap F^{-k}(Q). \tag{6.7}$$

Note that the "shift the decimal point" rule still applies. That is,

$$F^n(Q_\sigma) = Q_{\mathrm{sh}^n(\sigma)}$$

where σ is a sequence of 0's and 1's with a decimal point, and *sh* is the *shift* operation on such sequences: Applying sh^n to σ moves the decimal point n places to the right if $n \geq 0$, and it moves the decimal point $|n|$ to the left if $n < 0$.

See Figure 6.14 for $j = k = 1$ and 6.15 for $j = k = 2$.

We need to explain the significance of the intersections in (6.7). Suppose we have a point p_0, and we label the orbit of p_0 as $p_n = F^n(p_0)$ for integers n. Then $p_0 \in F^k(Q)$ is equivalent to saying $F^{-k}(p_0) \in F^{-k}(F^k(Q)) = Q$, or $p_{-k} \in Q$. Hence

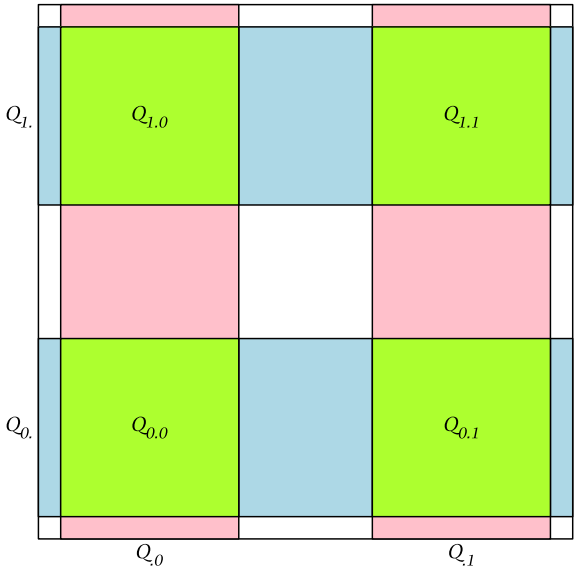


Figure 6.14: $F(Q) \cap Q \cap F^{-1}(Q)$.

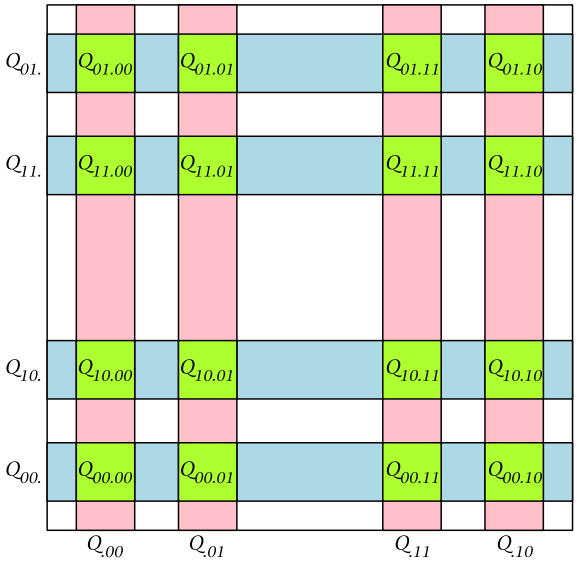


Figure 6.15: $F^2(Q) \cap F(Q) \cap Q \cap F^{-1}(Q) \cap F^{-2}(Q)$.

p_0 is in the set (6.7) if and only if p_n is in Q for each $n = -j, -j+1, \ldots, k-1, k$. That is, p_0 is in the set (6.7) if and only if the part of its orbit from p_{-j} to p_k is in Q.

There is one last part of this labeling procedure. Suppose we start with a *bi-infinite* sequence of 0's and 1's with a decimal point; this looks like $\sigma = \ldots \sigma_{-2}\sigma_{-1} \cdot \sigma_1\sigma_2 \ldots$. Corresponding to this we want to define a *point* P_σ as the limit of the rectangles that we have defined so far.

The notation for defining P_σ in general is clumsy, but the idea is simple, so we'll illustrate it with an example. If $\sigma = \ldots 111 \,.\, 000 \ldots$ then we have a sequence of rectangles obtained by truncating σ after n places both on the left and on the right of the decimal point: $Q_{1\,.\,0}$, $Q_{11\,.\,00}$, $Q_{111\,.\,000}$, and so on. These rectangles are *nested*, meaning that each new rectangle is contained in the last: $Q_{1\,.\,0} \supset Q_{11\,.\,00} \supset Q_{111\,.\,000} \supset \ldots$. Moreover, each rectangle is non-empty and both the width and the height limit to 0. In this case there is exactly one point which is in all the rectangles in the sequence; specifically, it can be defined as the limit of the centers of the rectangles. This limiting point is called P_σ.

Suppose $p_0 = P_\sigma$; we'll again use $\sigma = \ldots 111 \,.\, 000 \ldots$. Since p_0 is in $Q_{11\ldots1\,.\,0\ldots00}$, where there are n symbols on the left and the right of the decimal point, we see that the orbit of p_0 from p_{-n} through p_n is in Q. Since n was arbitrary we see that the entire orbit of p_0 lies in Q. Converseley, if the entire orbit of a point p_0 lies in Q then

we can define a sequence σ by the following rule:

$$\begin{aligned} \sigma_n &= s \qquad \text{if } n > 0 \text{ and } p_n = F^n(p_0) \in Q_{\cdot s} \\ \sigma_{-n} &= s \qquad \text{if } n > 0 \text{ and } p_{-n} = F^{-n}(p_0) \in Q_{s\cdot}. \end{aligned}$$

It follows from the definitions that $P_\sigma = p_0$.

These two constructions show how to obtain a point from a sequence, and a sequence from a point; and it is not hard to see that these are inverse operations. We think of the sequence σ as an *encoding* of the point p_0.

To summarize this correspondence we make some definitions.

First, Ω is defined as the set of points p_0 so that the entire orbit of p_0 lies in Q. In other words, a point of Q is not in Ω if it leaves Q at some time n, either positive or negative.

Next, Σ is defined to be the set of all bi-infinite sequences of 0's and 1's, with a decimal point. We can define a distance between points in Σ as follows: If $\sigma = \ldots\sigma_{-2}\sigma_{-1}\cdot\sigma_1\sigma_2\ldots$ and $\tau = \ldots\tau_{-2}\tau_{-1}\cdot\tau_1\tau_2\ldots$ are two *different* sequences in Σ then we can find the position k closest to the decimal point at which $\sigma_k \neq \tau_k$. (It is possible that both $\sigma_k \neq \tau_k$ and $\sigma_{-k} \neq \tau_{-k}$, but this is not necessary.) We define the distance between σ and τ to be 2^{-n}, where $n = |k|$. For example, the distance between $\ldots 10011\cdot 01110\ldots$ and $\ldots 11011\cdot 01110\ldots$ is 2^3, since the sequences differ in the third position to the left of the decimal place, but not in any positions closer to the decimal point. The following simple fact gives another way of thinking about this distance.

LEMMA 6.2. *The distance between σ and τ is less than 2^{-d} if and only if $\sigma_k = \tau_k$ for all k saatisfying $-d \leq k \leq d$.*

Finally, we can summarize the discussion so far:

THEOREM 6.3. *The procedure above describes a bijection (one-one correspondence) between Ω and Σ. This correspondence is continuous in both directions, in the sense that the points P_σ and P_τ are close if and only if the sequences σ and τ are close. The function F on points of Ω corresponds to the shift transformation on sequences in Σ in the sense that $F^n(P_\sigma) = P_{\mathrm{sh}^n(\sigma)}$.*

Now we can use this correspondence to analyze the behavior of F on the set Ω. The method is generally known as "symbolic dynamics".

THEOREM 6.4.
(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 points 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, no matter how small, 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 position. If this position is m positions to the right of the decimal point then we set $n = m - 1$, but if it is m places to the left of the decimal place then we set $n = -m + 1$. With this choice of n the sequences $\mathrm{sh}^n(\sigma)$ and $\mathrm{sh}^n(\tau)$ differ in a position that is just 1 place from the decimal point, so the distance between $\mathrm{sh}^n(\sigma)$ and $\mathrm{sh}^n(\tau)$ is $\frac{1}{2}$. For example, if $\sigma = \ldots 1100101\,.\,00010\ldots$ and $\tau = \ldots 0101101\,.\,00010\ldots$ then the sequences differ in the fourth place to the left of the decimal point, so $m = 4$ and $n = -3$. Then sh^{-3} moves the decimal point 3 places to the left, so $\mathrm{sh}^{-3}(\sigma) = \ldots 1100\,.\,10100010\ldots$ and $\mathrm{sh}^{-3}(\tau) = \ldots 0101\,.\,10100010\ldots$, and these differ in the first point to the left of the decimal point.

If we set $d_0 = \frac{1}{2}$ then we have shown that the transformation sh has sensitive dependence on initial conditions on the set Σ.

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, so we can find an integer n as above. Tracing through the correspondences, we see that $F^n(p)$ or $F^n(q)$ are in *different* rectangles in the list $Q_{0\,.\,0}$, $Q_{0\,.\,1}$, $Q_{1\,.\,0}$, $Q_{1\,.\,1}$. Thus the definition of sensitive

dependence on initial conditions is satisfied if we define d_0 to be the minimum distance between any two of these sets.

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 $\mathrm{sh}^k(\sigma) = \sigma$. This means that the sequence σ does not change if the decimal point is moved k units to the right, so the symbols in σ starting at position $k+1$ must be the same as the symbols starting at position 1. Hence the block of symbols in positions 1 through k to the right of the decimal point must be repeated in positions $k+1$ through $2k$. Replacing this argument with sh^{nk} shows that this block of symbols must be repeated in positions $nk+1$ through $(n+1)k$ if $n > 0$, and in positions $|n|\,k+1$ through $(|n|+1)k$ to the left if $n <= 0$. That is, $sigma = \ldots\beta\beta\,.\,\beta\beta\ldots$ where β is a fixed block of smbols of length k. Conversely, this recipe converts any block β of length k into a periodic sequence of period k. Notice that there are 2^k such blocks, so there are 2^k periodic sequences of period k.

For example, if $k = 4$ and $\beta = 1000$ 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 $\beta = 100\ldots0$, 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 $\pi_\beta = \ldots\beta\beta\,.\,\beta\beta\ldots$ for the periodic sequence defined by a finite block β of 0's and 1's in this manner.

Here are the first few cases:

Period 1: We have $\pi_0 = \ldots000\,.\,000\ldots$ and $\pi_1 = \ldots111\,.\,111\ldots$. 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 $\ldots0000\,.\,0000\ldots$ obtained by repeating 00, and this is the same as the sequence π_0. Thus π_{00} and π_{11}, although they have period 2, have *least* period 1. There are two sequences of least period 2, namely

$$\pi_{01} = \ldots10101\,.\,01010\ldots, \quad \pi_{10} = \ldots01010\,.\,10101\ldots$$

Notice that $\mathrm{sh}(\pi_{01}) = \pi_{10}$ and $\mathrm{sh}(\pi_{10}) = \pi_{01}$, so these two points make up one orbit.

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. The remaining 12 blocks of size 4 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 6.6.

The homoclinic points of any periodic orbit are dense: If p is a periodic point then a point q is called a *homoclinic point* corresponding to p if q is not on the orbit of p, but $F^n(p)$ 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 π_β 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,

$$\tau = \dots \beta\beta\tau_{-M} \dots \tau_{-1} \cdot \tau_1\tau_1 \dots \tau_N\beta\beta \dots$$

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 decimal point, with one of the sequences in the orbit of π_β. Thus the distance between $\mathrm{sh}^K(\tau)$ and points in the orbit of π_β will tend to 0 as $K \to +\infty$ and also as $K \to -\infty$, so τ is a homoclinic point corresponding to the orbit of π_k.

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} of ρ by requiring the symbols in τ to coincide with the symbols in ρ within N places of the decimal point, both to the left and to the right, 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 the sequence σ to the right of the decimal point; 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 that the symbol n places to the left of the decimal point is the same as the symbol n places to the right.

To see that the orbit of this sequence is dense, consider a sequence τ in Σ, and choose $N \geq 0$. Let β be the block of terms in τ extending N units on each side of the decimal point. This block occurs in σ somewhere, since *every* finite block occurs in σ. So there is some M so that $\sigma_{M-N+1} \dots \sigma_{M+N} = \beta$. So, if we write $\rho = \mathrm{sh}^{-M}(\sigma)$, then the block of terms in ρ extending N places on either side of the decimal point is equal to β, which is the corresponding block of τ. Hence the distance between $\rho = \mathrm{sh}^{-M}(\sigma)$ and τ is less than 2^{-N}. 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 6.4 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 of dimension. 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 $Q_{0.0}$ onto all of Ω. Complicated sets – in particular, fractals – that are defined by iteration often show self-similarity.

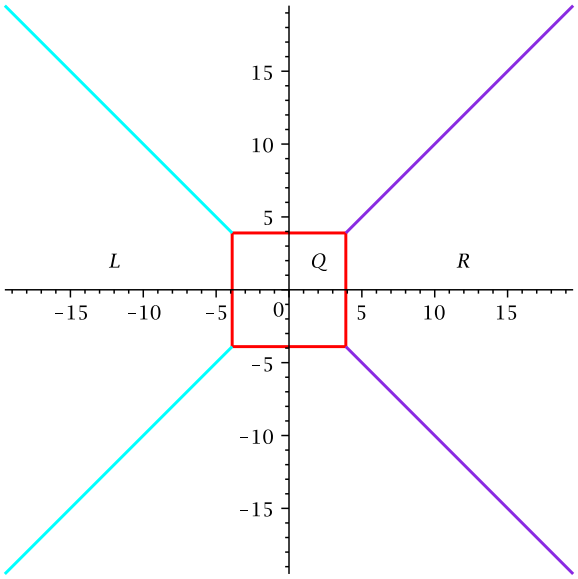


Figure 6.16: Q and the diagonal lines $y = x$ and $y = -x$.

Exercises

6.1. Let H be the the Hénon map discussed in section 6.1. Find the images of the diagonal lines $y = x$ and $y = -x$ under H. You must find non-parametric forms for these image curves. You should find that they are parabolas.

6.2. Let H be the the Hénon map discussed in section 6.1.

(a) Consider the region L in Figure 6.16 extending to the *left* of the square Q, both above and below the x axis. This region is bounded by the left side of the square and two diagonal lines.

Sketch the image of L under H. Note that the image of the left side of Q was already determined (see Figure 6.1), and the equations for the diagonal lines were determined in Exercise 6.1.

(b) You should be able to see from your sketch that $H(L)$ is contained in L. Suppose that p is any point in L. Explain why $H^n(p)$ can never be in Q, for any $n > 0$.

(c) Let R be the similar region to the right of Q. By symmetry it follows that $H(R)$ is the reflection of $H(L)$ through the x axis. Sketch $H(R)$. You should see that $H(R)$ is contained in L. Suppose that p is any point in R. Explain why $H^n(p)$ can never be in Q, for any $n > 0$.

(d) Suppose that p is in Q but $H(p)$ is not in Q. Explain why $H^n(p)$ can never be in Q, for any $n > 0$. (Look at Figure 6.1. Where is $H(p)$?)

6.3. Let H be the the Hénon map discussed in section 6.1. Show that the fixed points of H are saddle points. (A fixed point of a discrete dynamical system is a saddle point if one of the eigenvalues of the linearization at the fixed point has absolute value less than 1, and the other has absolute value greater than 1.)

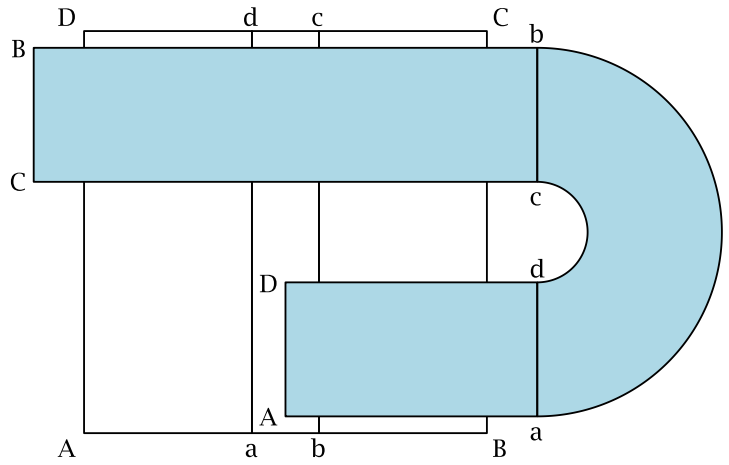
6.4. For the horseshoe map:

(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.

6.5. (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?

6.6. Generalize Exercise 6.5 to show that, if σ is any sequence in Σ and N is arbitrary then there is a periodic sequence within distance 2^{-N} of σ.

6.7. 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 $Q_{.\beta}$ where β is a suitable block of 0's and 1's of length ≤ 4. This will not be the same as in the full horseshoe. For example, $Q_{.00}$ is empty since there

are no points that start in $Q_{.0}$ and are in $Q_{.0}$ after one iteration of F. [You do not need to draw the empty rectangles.]

(b) Draw the horizontal rectangles $Q_{\beta .}$ where β has length ≤ 3.

6.8. Using the broken horseshoe of Exercise 6.7 we define the set Ω_1 of points that lie in Q under all iterations, and we encode x in Ω_1 by the bi-infinite sequence σ as for the full horseshoe. Not all sequences can be obtained in this manner; for example, the sequence $\ldots 000.0000 \ldots$ corresponds to a point that remains in $Q_{.0}$ under all iterations, but no point in $Q_{.0}$ remains in $Q_{.0}$ for even one iteration.

Define

$$\Sigma_1 = \{\, \sigma \colon \text{the block 00 does not occur in } \sigma \,\}.$$

In other words, the condition for σ to be in Σ_1 is that there cannot be two consecutive 0's anywhere in σ.

The following are two 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 (with decimal point) so that 00 does not occur in σ. Show that Q_σ is non-empty.

6.9. Using the definition of Σ_1 in Exercise 6.8, 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?

Bibliography

[1] T. S. Gardner, C. R. Cantor, and J. J. Collins, *Construction of a genetic toggle switch in Escherichia coli*, Nature **403** (Jan 2000), no. 6767, 339–42.

[2] Ilse C. F. Ipsen and Rebecca S. Wills, *Mathematical properties and analysis of Google's PageRank*, Bol. Soc. Esp. Mat. Apl. SeMA (2006), no. 34, 191–196. MR 2296216

[3] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report 1999-66, Stanford InfoLab, 1999. Previous number = SIDL-WP-1999-0120.

APPENDIX A

Review

A.1. Calculus

We need a few facts from calculus.

If we look at the graph of a continuous function defined on a *closed* and *bounded* interval then expect to see a "connected" curve, with a highest and a lowest point. That is the content of the Intermediate Value and Extreme Value Theorems:

THEOREM A.1 (Intermediate Value Theorem). *If f is continuous on $[a, b]$ and v is between $f(a)$ and $f(b)$ then there is some c in $[a, b]$ so that $f(c) = v$.*

THEOREM A.2 (Extreme Value Theorem). *If f is continuous on $[a, b]$ then there are numbers c and d in $[a, b]$ so that $f(c) \leq f(x) \leq f(d)$ is true for all x in $[a, b]$.*

The Fundamental Theorem of Calculus has two parts, and together they say that integration and differentiation are inverse operations:

THEOREM A.3 (Fundamental Theorem of Calculus).
Part 1: If G is differentiable on the interval $[a, b]$ and $g = G'$ then

$$\int_a^b g(t)\,dt = G(b) - G(a).$$

Part 2: If g is continuous on the interval $[a, b]$ then g is integrable on $[a, b]$, the function G defined by $G(x) = \int_a^x g(t)\,dt$ is differentiable, and

$$G'(x) = \frac{d}{dx}\int_a^x g(t)\,dt = g(x).$$

Part 1 is the basis for evaluating definite integrals by finding antiderivatives, and part 2 is used extensively to define differentiable functions by integration.

The derivative is defined as a limit of difference quotients $\dfrac{\Delta y}{\Delta x}$. The Mean Value Theorem says that any difference quotient *is* a derivative – but it doesn't say *where* you should take the derivative:

THEOREM A.4 (MVT: Mean Value Theorem). *Suppose that f is dirrerentiable on the open interval (a, b) and continuous at the endpoints a and b. Then there is a number c in (a, b) so that*

$$\frac{f(b) - f(a)}{b - a} = f'(c).$$

If f is differentiable on the *closed* interval $[a, b]$ then the continuity assumption is automatically satisfied, since differentiability implies continuity. The MVT as stated is slightly more general.

The MVT is used frequently in constructing inequalities, as follows. First, if s and t are in $[a, b]$ and not equal then the MVT applies equally well to the interval between s and t, so $\dfrac{f(s) - f(t)}{s - t} = f'(w)$ where w is some number between s and t. If we multiply this out and take absolute values we get $|f(s) - f(t)| = |f'(w)| \cdot |s - t|$; this holds even if $s = t$ since then both sides are 0. Now we usually have no way of finding the value $f'(w)$ since we have no way of finding w. However, it is often the case that we can find an upper bound for $|f'(w)|$, and then we can convert an equality involving a number w that we don't know to an inequality involving an upper bound that we can estimate. This reasoning proves the following, which is often more useful than the MVT:

THEOREM A.5 (Mean Value Inequality). *Under the hypotheses of the MVT, suppose that $|f'(w)| \leq M$ for all numbers w in (a, b). Then, for all numbers s and t in $[a, b]$,*

$$|f(s) - f(t)| \leq M \cdot |s - t|.$$

Here's a typical example: The derivative of $\ln t$ is $1/t$. If $t \geq 1$ then $\dfrac{1}{t} \leq 1$. Hence the Mean Value Inequality, applied to the interval $[1, x]$ and using the upper bound $M = 1$ for the derivative, says that $|\ln(x) - \ln(1)| \leq 1 \cdot |x - 1|$. Since $\ln(1) = 0$ and $\ln(x)$ and $x - 1$ are non-negative we can simplify this to get the inequality $\ln(x) \leq x - 1$, valid for all $x \geq 1$.

A.2. Complex numbers

This section is a summary of practical algebra in the complex number system.

Technically, a complex number is an ordered pair of real numbers; in other words, the set $\mathbb{C}$ of complex numbers is the same as R^2, the usual XY plane. Instead of writing a complex number z as (x, y) we will generally write it as $x + iy$. The *real part* of z is x and the *imaginary part* of z is y.

Caution: The imaginary part of a complex number is a **real** number.

Complex numbers of the form $x + 0i = x$ are just real numbers, while complex numbers of the form $0 + iy = iy$, with $y \neq 0$, are called *pure imaginary* numbers.

In this form the rules of algebra, for addition, subtraction, and multiplication can be summarized as:

(1) Use the commutative, associative and distributive laws as for real numbers.
(2) Remember that 0 and 1 are units for addition and multiplication, respectively.
(3) Use $i^2 = -1$ to reduce powers of i.

For example, $i^3 = i^2 \cdot i = -1 \cdot i = -i$. Also,

$$\begin{aligned}(2+i)(3-2i) &= 2 \cdot 3 + 2 \cdot (-2i) + i \cdot 3 + i \cdot (-2i) \\ &= 6 - 4i + 3i - 2i^2 = 6 - i - 2 \cdot (-1) = 8 - i.\end{aligned}$$

There is an important operation for complex numbers called *conjugation*, which is defined by "changing the sign of the imaginary part". Specifically, if $z = x + iy$ then $\overline{z} = x - iy$. The main properties of congugation are:

(1) $\overline{(z \pm w)} = \overline{z} \pm \overline{w}$.
(2) $\overline{z \cdot w} = \overline{z} \cdot \overline{w}$.
(3) $\overline{\overline{z}} = z$.
(4) z is real if and only if $z = \overline{z}$.
(5) The real and imaginary parts of z are $\dfrac{1}{2}(z + \overline{z})$ and $\dfrac{1}{2i}(z - \overline{z})$.
(6) $z \cdot \overline{z} = x^2 + y^2$. This is real and non-negative, and is 0 if and only if $z = 0$.

Using the conjugate we can calculate multiplicative inverses, so we can divide. The idea is to "rationalize the denominator" by multiplying top and bottom by the conjugate of the denominator. For example,

$$\frac{2+i}{3+2i} = \frac{(2+i)(3-2i)}{(3+2i)(3-2i)} = \frac{8-i}{3^2-(2i)^2} = \frac{8-i}{9-(-4)} = \frac{8}{13} - \frac{i}{13}.$$

We can also use conjugation to define the *absolute value* or *modulus* of a complex number as $|z| = \sqrt{z \cdot \overline{z}} = \sqrt{x^2 + y^2}$. For example, $|3 + 2i| = \sqrt{9+4} = \sqrt{13}$.

Caution: The formula for $|z|$ is $\sqrt{x^2 + y^2}$, **not** $\sqrt{x^2 + (iy)^2}$. Remember that y is real.

The absolute value has many of the same properties as the absolute value of real numbers:

(1) $|zw| = |z| \cdot |w|$.
(2) $|z| \geq 0$, and $|z| = 0$ if and only if $z = 0$.
(3) $|z + w| \leq |z| + |w|$.

Here are a few things to be careful about:

(1) $|z|^2 = z^2$ is true for real numbers, but is **not** generally true for complex numbers.
(2) $|z| = \pm z$ is true for real numbers, but is **not** generally true for complex numbers.

There is no useful way to define an order relation for complex numbers; so $z \leq w$ has no meaning. All we can do is compare magnitudes, using absolute values.

Instead of using Cartesian coordinates, we can represent complex numbers using polar coordinates. The usual formulas relating Cartesian and polar coordinates are

$$\begin{aligned} x &= r\cos\theta & r &= \sqrt{x^2+y^2} \\ y &= r\sin\theta & \theta &= \arctan\left(\frac{y}{x}\right) \end{aligned}$$

Applying them to $z = x + iy$, we see that $z = r\cos\theta + ir\sin\theta = r(\cos\theta + i\sin\theta)$, where $r = |z|$. The polar angle θ is called the *argument* of z. It can be any angle if $z = 0$, and otherwise it is defined up to the addition of an integer multiple of 2π. If we multiply two complex numbers in polar coordinates and remember the addition formulas for the sine and cosine, we are led to

$$r_1(\cos\theta_1 + i\sin\theta_1)\cdot r_2(\cos\theta_2 + i\sin\theta_2) = r_1r_2(\cos(\theta_1+\theta_2) + i\sin(\theta_1+\theta_2)).$$

That is, when you multiply two complex numbers you multiply their absolute values, but you add their arguments. In geometric terms this means that multiplication of points in the plane (i.e., complex numbers) by a fixed complex number z is the transformation defined by a scale change of $r = |z|$, followed by a rotation around the origin through the angle θ.

We can apply this to calculating z^n. We write $z = r\cos\theta + i\sin\theta$ in polar coordinates. Then $|z^n| = |z|^n = r^n$, and the argument of z^n is obtained by adding n copies of the argument of z, to give $n\theta$. So we have

$$z^n = (r(\cos\theta + i\sin\theta)) = r^n(\cos(n\theta) + i\sin(n\theta) \tag{A.1}$$

This is known as *De Moivre's formula*.

Complex numbers were introduced to solve polynomial equations – indeed, the quadratic equation often gives non-real solutions for quadratic equations. There are similar, but much more complicated, formulas for solving third and fourth degree polynomial equations, and these also involve complex numbers. Mathematicians suspected for many years that complex numbers were enough to solve all polynomial equations. This was first proved by Gauss in 1799, although there were some gaps in his proof; several different completely rigorous proofs have since appeared. It is important enough to be called the Fundamental Theorem of Algebra:

THEOREM A.6 (Fundamental Theorem of Algebra). *If $P(z)$ is a polynomial of degree $m > 0$ with complex coefficients then $P(z)$ has exactly m ccomplex roots $\alpha_1, \alpha_2, \ldots, \alpha_m$ (listed with multiplicity). In other words, $P(z)$ factors as $c(z-\alpha_1)(z-\alpha_2)\ldots(x-\alpha_m)$ where c is the coefficient of z^m in $P(z)$.*

Even if the original polynomial has real coefficients there will probably be non-real roots. The following gives a way of organizing them.

PROPOSITION A.7. *If $P(z)$ is a polynomial of degree $m > 0$ with real roots then the roots of $P(z)$ can be listed as $r_1, r_2, \ldots, r_p, c_1, \overline{c}_1, c_2, \overline{c}_2, \ldots c_q, \overline{c}_q$ (listed with multiplicity) where $p + 2q = m$, the roots r_i are real, and the roots c_i and $\overline{c}_i$ are non-real. In other words, complex roots occur in conjugate pairs.*

PROOF. The key observation is that $\overline{P(z)} = P(\overline{z})$, which is true since conjugation preserves sums and products and leaves the coefficients of $P(z)$ unchanged, since these coefficients are real. From this we see that if $P(z) = 0$ then $P(\overline{z}) = 0$. It is then clear that each non-real root z is matched up with its conjugate $\overline{z}$, and so we can list the roots as in the proposition. □

A.3. Partial derivatives

For a function f of two or more variables the *partial derivatives* are defined by differentiating with respect to one variable, while treating all other variables as constants. For example, if $f(x, y) = x^2 + y^3 - \sin(3x + 5y)$ then the partial derivatives are

$$\frac{\partial f}{\partial x} = 2x - 3\cos(3x + 5y), \quad \frac{\partial f}{\partial y} = 3y^2 - 5\cos(3x + 5y).$$

Higher order derivatives are defined by iterating this process. For example, for this example we have

$$\begin{aligned}
\frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x}\left(2x - 3\cos(3x + 5y)\right) = 2 + 9\sin(3x + 5y) \\
\frac{\partial^2 f}{\partial x \partial y} &= \frac{\partial}{\partial x}\left(3y^2 - 5\cos(3x + 5y)\right) = 15\sin(3x + 5y) \\
\frac{\partial^2 f}{\partial y^2} &= \frac{\partial}{\partial y}\left(3y^2 - 5\cos(3x + 5y)\right) = 6y + 25\sin(3x + 5y) \\
\frac{\partial^2 f}{\partial y \partial x} &= \frac{\partial}{\partial y}\left(2x - 3\cos(3x + 5y)\right) = 15\sin(3x + 5y)
\end{aligned}$$

In general, working with partial derivatives requires slightly stronger assumptions than in the corresponding single variable situations, and it is simplest to formulate these extra assumptions in terms of continuity of the partial derivatives. We say a function is a C^1 function if it has first order partial derivatives with respect to all its variables, and these partials are *continuous*. We say a function is a C^2 function if it has continuous first and second order partials, and so on.

Also, working with partial derivatives, especially if the function f is a vector function, leads to very cumbersome notation. We can often simplify this by using

the *matrix derivative*, also known as the *Jacobian matrix*: Suppose f is a function of n variables and its values are m-dimensional vectors, which we regard as **column vectors**. If we calculate the partial derivative of f with respect to one of its variables we will get a m dimensional vector. We can then form the $m \times n$ matrix that has these partial derivatives as its columns. This matrix is written as Df, or as $Df(a)$ if the derivatives are all taken at the point a.

For example, if $f(x,y) = \begin{bmatrix} x^2 \\ xy + y^2 \end{bmatrix}$ then

$$\frac{\partial f}{\partial x} = \begin{bmatrix} 2x \\ y \end{bmatrix},\ \frac{\partial f}{\partial y} = \begin{bmatrix} 0 \\ x + 2y \end{bmatrix},\ Df = \begin{bmatrix} 2x & 0 \\ y & x + 2y \end{bmatrix},\ Df(-1,3) = \begin{bmatrix} -2 & 0 \\ 3 & 5 \end{bmatrix}.$$

Here are some of the main facts about partial derivatives:

THEOREM A.8. *If f is C^1 then f is continuous; if f is C^2 then f is also C^1.*

THEOREM A.9 (Equality of mixed partials). *If f is C^2 then the "mixed partials" with respect to any two varialbes are equal. That is,* $\dfrac{\partial^2 f}{\partial x_i \partial x_j} = \dfrac{\partial^2 f}{\partial x_j \partial x_i}$.

THEOREM A.10 (Chain rule). *If f and g are C^1 functions then the composition $f \circ g$ is C^1. If a is in the domain of g and $g(a)$ is in the domain of f then $D(f \circ g)(a) = Df(g(a))Dg(a)$.*

THEOREM A.11 (Linear approximation). *If f is C^1 and a is in the domain of f then*

$$f(x) = f(a) + Df(a)(x - a) + R(x)$$

where the remainder term $R(x)$ satisfies

$$\|R(x)\| = \varepsilon(x)\,\|x - a\|\,, \quad \lim_{x \to a} \varepsilon(x) = 0.$$

The condition on the remainder means that $R(x)$ is very small in comparison to $x - a$, if x is near a.

A.4. Exact differential equations

Here is a technique for finding a constant of the motion for a system of equations of the form $\dfrac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = f(x,y)$. We are going to treat this as an "exact differential equation". A similar method is covered in Calc III, or in Physics courses, in terms of "finding a potential function for a conservative force field".

Start by writing $f(x,y) = \begin{bmatrix} P(x,y) \\ Q(x,y) \end{bmatrix}$, so $\dfrac{dy}{dx} = \dfrac{dy/dt}{dx/dt} = \dfrac{Q(x,y)}{P(x,y)}$, and separate variables as outlined in section 5.1. The result is $Q(x,y)\,dx - P(x,y)\,dy = 0$. A

constant of the motion φ will satisfy $d\varphi = \dfrac{\partial\varphi}{\partial x}\,dx + \dfrac{\partial\varphi}{\partial y}\,dy = 0$ along these trajectories. This suggests that we try to find a function φ satisfying $d\varphi = Q\,dx - P\,dy$. If this is possible then $\dfrac{\partial\varphi}{\partial x} = Q$ and $\dfrac{\partial\varphi}{\partial y} = -P$, so

$$\frac{\partial Q}{\partial y} = \frac{\partial}{\partial y}\left(\frac{\partial\varphi}{\partial x}\right) = \frac{\partial^2\varphi}{\partial y\partial x} = \frac{\partial^2\varphi}{\partial x\partial y} = \frac{\partial}{\partial x}\left(\frac{\partial\varphi}{\partial y}\right) = -\frac{\partial P}{\partial x}.$$

Hence this method can't work unless the *integrability condition* $\dfrac{\partial Q}{\partial y} = -\dfrac{\partial P}{\partial x}$ is satisfied. Note that the integrability condition can be written in the form $\dfrac{\partial P}{\partial x} + \dfrac{\partial Q}{\partial y} = 0$, which, in the language of vector calculus, says that the *divergence* of the vector field f is zero.

At this stage you should check the integrability condition, and don't bother to continue unless it is satisfied. However, you can try to algebraically modify the equation $Q(x,y)\,dx - P(x,y)\,dy = 0$ to a different form $Q_1(x,y)\,dx - P_1(x,y)\,dy = 0$, and then you can try again. There are infinitely many ways to rewrite the equation, since you can multiply by any function of (x,y).

If the integrability condition is satisfied then you can continue. Start with one of the equations, say $\dfrac{\partial\varphi}{\partial x} = Q(x,y)$, and integrate **with respect to** x. The result will be $\varphi(x,y)$, involving an arbitrary constant **with respect to** x. That is, the arbitrary constant will in fact be a function of y, say $g(y)$. Now plug this expression for φ, including the unknown $g(y)$, into $\dfrac{\partial\varphi}{\partial y} = -P(x,y)$. All x terms should cancel out from this equation, leaving a differential equation for $g(y)$ in terms of y. (If not, you made a mistake; try again.) Solve this differential equation for $g(y)$; you don't need to keep the constant of integration. Since you know $g(y)$ you have now determined $\varphi(x,y)$. Congratulations: This is a constant of the motion.

Here's an example: Suppose $\dfrac{dx}{dt} = 2x - y^2$, $\dfrac{dy}{dt} = x - 2y$, so $P(x,y) = 2x - y^2$ and $Q(x,y) = x - 2y$. To find the differential equation for the constant of the motion, calculate

$$\frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{x-2y}{2x-y^2}.$$

Multiply this out and gather all terms on one side of the equation:

$$(x-2y)\,dx + (y^2-2x)\,dy = 0. \tag{A.2}$$

The equations for φ are $\dfrac{\partial\varphi}{\partial x} = Q(x,y) = x - 2y$, $\dfrac{\partial\varphi}{\partial y} = -P(x,y) = y^2 - 2x$. The integrability condition $\dfrac{\partial Q}{\partial y} = -\dfrac{\partial P}{\partial x}$ is satisfied; both sides are equal to -2. Hence it is OK to continue, so integrate $\dfrac{\partial\varphi}{\partial x} = x - 2y$ with respect to x to find

$$\varphi = \int (x - 2y)\,dx = \tfrac{1}{2}x^2 - 2xy + g(y).$$

Here $g(y)$, an unknown function of y, is a constant as long as we consider x to be the variable. Now plug this expression for φ into the other equation, $\dfrac{\partial\varphi}{\partial y} = y^2 - 2x$, to get

$$\frac{\partial\varphi}{\partial y} = \frac{\partial}{\partial y}\left(\tfrac{1}{2}x^2 - 2xy + g(y)\right) = -2x + g'(y) = y^2 - 2x.$$

This simplifies to $g'(y) = y^2$, which integrates to $g(y) = \frac{1}{3}y^3$ (setting the constant of integration to 0). So

$$\varphi(x,y) = \tfrac{1}{2}x^2 - 2xy + g(y) = \tfrac{1}{2}x^2 - 2xy + \tfrac{1}{3}y^3.$$

Notice that we were fortunate in the way we wrote the differential equation (A.2). We might have written it, for example, as

$$dy - \frac{x - 2y}{2x - y^2}\,dx = 0, \quad \text{or} \quad \frac{dx}{2x - y^2} - \frac{dy}{x - 2y} = 0,$$

and neither of these satisfies the integrability condition.

APPENDIX B

Maple notes

This chapter contains some notes to help you get started using Maple.

B.1. Basic operations

Here are a few general tips.

Use Worksheets, not Documents. You can the default format to Worksheet on the "Options" dialog, under the "Interface" tab.

There are several styles for sending input to Maple, selected by the drop-down at the top of the worksheet. The default is "2D math", so the input to raise 2 to the 100^{th} power appears as 2^{100}. Since it is not always clear how to input things like fractions, exponents, differential equations, etc., I have represented input expressions as "1D math". This is a text-only input format, so it is clear which keystrokes are necessary to indicate things like exponents, fractions, etc. For example, the input to raise 2 to the 100^{th} power, using "1D math", appears as `2^100`.

If you want, you can change the default input method to "1D math" on the "Options" dialog, under the "Display" tab, where it is called "Maple notation". You can always switch back and forth for an individual command using the drop-down menus on the main display, where it is called "Maple input".

Maple comes with an extensive help system, and these notes are not meant as a replacement. You can browse all the help by choosing "Maple Help" on the "Help" menu. You can find help on a particular topic by searching for it in "Maple Help". You can also enter the command `help(int)` or `?int`, for example, to get help on the `int` procedure. You can get the same effect if you position your cursor on a term in the worksheet and then hit the "F2" key or look under the "Help" menu.

In "1d math" mode, each command must end with a colon or a semicolon; the semicolon is not needed in "2d math" mode. You execute the command by hitting the `Enter` or `Return` key while your cursor is anywhere on the command. If the line ends with a semicolon then the result of the command will be displayed; if it ends with a colon then the result will not be displayed, unless there is an error. You can put several commands on one line, each terminated by a semicolon or colon.

The entire worksheet is live. You can go back and modify earlier commands and re-execute them, and you can insert new lines wherever you want using the `[>` button on the main display. This can be confusing, since the results that you see depend on the order in which you do things. If you save a worksheet and then reload it you need to re-execute all the commands. The `!!!` button on the main display will do this.

Maple deals in mathematical objects: numbers, expressions, equations, functions, graphs, lists, sets, matrices, etc. This approach sometimes causes confusion. For example, `x = 3` is an *equation*; it does *not* set x equal to the value 3.

Maple uses the symbol `:=` to give a *name* to something. For example, after `x := 3` is executed, `x` is no longer a variable, but just a name for the value 3. You can tell Maple to forget that you assigned a value to `x` by the command `unassign('x')`. The quotes are necessary – otherwise, Maple will interpret this as `unassign(3)`, which doesn't make sense.

B.2. Linear algebra

Here is an annotated Maple session that demonstrates some of Maple's facilities for linear algebra calculations.

Maple has built-in support for vectors and matrices. Lists bracketed by `<...>` become column vectors:

```
>   <a,b,c>;
```

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Inside `<...>` a vertical bar `|` starts a new column, so you get a matrix:

```
>   <a,b|C,D>;
```

$$\begin{bmatrix} a & C \\ b & D \end{bmatrix}$$

So you can get a row vector by:

```
>   <a|b|c>;
```

$$\begin{bmatrix} a & b & c \end{bmatrix}$$

If you want to enter a matrix in row order, rather than column order, then you can put the rows in a column:

```
>   <<a|b>,<C|D>>;
```

$$\begin{bmatrix} a & b \\ C & D \end{bmatrix}$$

The `<...>` construction actually translates to a construct using the `Vector` or `Matrix` procedure. For example, `<<a|b>,<C|D>>` becomes `Matrix([[a,b],[C,D]])`. Maple will sometimes rewrite a matrix this way.

The construction `<a,b;C,D>` is shorthand for `<<a|b>,<C|D>>`. For example, the matrix in (3.1) can be entered like this:

```
>   L:=<0,7,6;1/4,0,0;0,1/2,0>;
```

$$L := \begin{bmatrix} 0 & 7 & 6 \\ 1/4 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix}$$

You can also enter matrices and vectors using the "Matrix" widget on the side panel of the main display. This works best in "2D math" mode.

Maple knows how to add or subtract matrices, and how to multiply or divide by scalars. You can also multiply matrices times other matrices or vectors. Maple requires that you use a star (`*`) for scalar multiplication, and a period (`.`) for matrix multiplication. You can also calculate matrix powers using `^`; this works for negative powers as well, if the matrix is invertible.

This calculates the 20^{th} power of `L` and saves the result as `L20`:

```
>   L20:=L^20;
```

$$L20 := \begin{bmatrix} \frac{1569891841}{1048576} & \frac{1176370305}{131072} & \frac{783897345}{131072} \\ \frac{261299115}{1048576} & \frac{1569891841}{1048576} & \frac{523646805}{524288} \\ \frac{174548935}{2097152} & \frac{261299115}{524288} & \frac{21753081}{65536} \end{bmatrix}$$

Maple does exact arithmetic if it can. It is easier to read something like this if you convert it to decimal form. The `evalf` procedure will convert anything to decimal form:

```
> evalf(L20);
```

$$\begin{bmatrix} 1497.165528 & 8974.993172 & 5980.662117 \\ 249.1942549 & 1497.165528 & 998.7770176 \\ 83.23141813 & 498.3885098 & 331.9256744 \end{bmatrix}$$

You can refer to the ij entry in a matrix using `[i,j]`. For example, this divides `L20` by its $3,1$:

```
> L20/L20[3,1];
```

$$\begin{bmatrix} \frac{3139783682}{174548935} & \frac{342216816}{3173617} & \frac{228042864}{3173617} \\ \frac{9501786}{3173617} & \frac{3139783682}{174548935} & 12 \\ 1 & \frac{19003572}{3173617} & \frac{696098592}{174548935} \end{bmatrix}$$

Use `evalf(L20/L20[3,1])` to see this in decimal form. Alternatively, after you calculate `L20/L20[3,1]` you can use `evalf(%)`, since `%` always refers to the last thing that Maple *calculated* (not necessarily the calculation in the preceeding line).

Maple has basic programming facilities. You will not need to do any programming for this course, but it is sometimes useful to do simple iteration. Here is a statement that will print the first through twentieth powers of `L`:

```
> for k from 1 to 20 do k,evalf(L^k);
```

Maple has a number of *packages* of procedures for more specialized mathematics. In order to use a package you need to either tell Maple to activate the procedures in the package or you have to refer to them individually using a somewhat cumbersome syntax. The following enables the `LinearAlgebra` package:

```
> with(LinearAlgebra);
```

The output is the list of procedures in the package; it should cover more than you saw in your Linear Algebra course. You can get an overview of the package in the help system:

```
> help(LinearAlgebra);
```

For example, here is the Characteristic polynomial of L, using λ as the variable:

```
> P:=CharacteristicPolynomial(L,lambda);
```

$$P := \lambda^3 - \frac{7}{4}\lambda - \frac{3}{4}$$

Here `P` is an *expression*, not a function. Some Maple procedures will require that you use an expression, and some will require that you use a function, or some other type of object. You need to check the documentation.

Here's one way to get a function corresponding to P:

```
> FP:=lambda->lambda^3-(7/4)*lambda-3/4;
```

$$FP := \lambda \mapsto \lambda^3 - \frac{7}{4}\lambda - \frac{3}{4}$$

The `solve` routine solves equations. The following solves the *equation* $P = 0$ for the variable λ:

```
> solve(P=0,lambda);
```

$$-\frac{1}{2}, \frac{3}{2}, -1$$

Maple tries to solve equations exactly if they are written exactly. However, if any of the numbers in a problem is written with a decimal point (which signals an approximation) then Maple uses numeric methods. You can also force a numeric solution by using `fsolve`.

Once you have the eigenvalues you can solve the eigenvector equation to get the corresponding eigenvector. For example, for the eigenvalue $frac32$ it is necessary to solve the equation $\left(L - \frac{3}{2}I\right) v = 0$. To do this, first define the 3×3 identity matrix:

```
> I3:=IdentityMatrix(3);
```

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(In Maple, `I` means the complex number i, so you can't use it as a variable name.)

Now find a basis for the null space of $L - \frac{3}{2}I$:

```
> NullSpace(L-3/2*I3);
```

$$\left\{ \begin{bmatrix} 18 \\ 3 \\ 1 \end{bmatrix} \right\}$$

Even better, you can get the eigenvalues and the corresponding eigenvectors in one step by

```
> Eigenvectors(L);
```

$$\begin{bmatrix} \frac{3}{2} \\ -\frac{1}{2} \\ -1 \end{bmatrix}, \begin{bmatrix} 18 & 2 & 8 \\ 3 & -1 & -2 \\ 1 & 1 & 1 \end{bmatrix}$$

Here is one more variant. Let's define a new matrix with a variable entry:

```
> unassign('a'); M:=<0,7,6;a,0,0;0,1/2,0>;
```

$$M := \begin{bmatrix} 0 & 7 & 6 \\ a & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix}$$

(The `unassign` command erases any value that may have been assigned to `a`.)

Now we can assign a value to `a`, and calculate with that value. For example, if $a = 1/4$ then M is equal to L, so

```
> a:=1/4; EL := Eigenvectors(M);
```

$$a := \frac{1}{4}$$

$$EL := \begin{bmatrix} -1 \\ \frac{3}{2} \\ -\frac{1}{2} \end{bmatrix}, \begin{bmatrix} 8 & 18 & 2 \\ -2 & 3 & -1 \\ 1 & 1 & 1 \end{bmatrix}$$

We can assign a different value to `a` and repeat the command, since `M` has not been redefined, so it still has `a` as a variable. So we'll change a from $1/4$ to $1/5$ and calculate the eigenvalues again:

```
> a:=1/5; EL := Eigenvectors(M);
```

This produces many lines of output, since Maple solves the characteristic equation to get the eigenvalues exactly. The characteristic equation is a cubic equation, and it doesn't factor using rational numbers when $a = 1/5$. Maple's answer is exact, but

very messy. If we were dealing with 5×5 or larger matrices then there is usually no reasonable formula for the eigenvalues in terms of elementary functions.

However, if we replace 1/5 with .2 then Maple calculates using numeric approximations rather than exact arithmetic:

```
> a:=.2;EL := Eigenvectors(M);
```

$$a := 0.2$$

$$EL := \begin{bmatrix} 1.357231 + 0.0\,I \\ -0.814419 + 0.0\,I \\ -0.542812 + 0.0\,I \end{bmatrix}$$

$$\begin{bmatrix} 0.987893 + 0.0\,I & 2.492243 + 0.0\,I & 6.523453 + 0.0\,I \\ 0.145575 + 0.0\,I & -0.612029 + 0.0\,I & -2.403578 + 0.0\,I \\ 0.053629 + 0.0\,I & 0.375746 + 0.0\,I & 2.214006 + 0.0\,I \end{bmatrix}$$

Maple actually shows more digits than this. If you do not want to see all the digits, go to the "Precision" tab on the "Options" dialog, check the "Round screen display" box, and set the display precision to your liking. Don't change the "Round calculation" setting.

Notice that the entries in `EL` are complex numbers, although the imaginary parts are zero so they are actually real. This is to be expected, since polynomial equations usually have some complex roots. If you redo this with $a = .1$ then two of the eigenvalues, and their corresponding eigenvectors, will be non-real.

The output of the `Eigenvectors` command is a *list* of two elements: a column vector containing the eigenvalues and a matrix whose columns are the eigenvectors. Since the result is stored in the variable `EL` you can get these two components as `EL[1]` and `EL[2]`.

Since M has 3 distinct eigenvalues we can write $M = P\Lambda P^{-1}$ where P is the matrix with the eigenvectors as columns and λ is the diagonal matrix with thhe eigenvalues on the diagonal, in the same order as the corresponding eigenvectors. (See equation (3.2).) Notice that P is just the second component of `EL`, and we can get Λ from the first component of `EL`. So

```
> Lambda:=DiagonalMatrix(EL[1]);
```

$$\Lambda := \begin{bmatrix} 1.357231 + 0.0\,I & 0 & 0 \\ 0 & -0.814419 + 0.0\,I & 0 \\ 0 & 0 & -0.542812 + 0.0\,I \end{bmatrix}$$

```
> P:= EL[2];
```

$$P := \begin{bmatrix} 0.987893 + 0.0\,I & 2.492243 + 0.0\,I & 6.523453 + 0.0\,I \\ 0.145575 + 0.0\,I & -0.612029 + 0.0\,I & -2.403578 + 0.0\,I \\ 0.053629 + 0.0\,I & 0.375746 + 0.0\,I & 2.214006 + 0.0\,I \end{bmatrix}$$

```
> P.Lambda.P^(-1);
```

$$\begin{bmatrix} -6.810346 \cdot 10^{-10} + 0.0\,I & 7.0 + 0.0\,I & 6.0 + 0.0\,I \\ 0.200000 + 0.0\,I & 1.470675 \cdot 10^{-10} + 0.0\,I & 2.246900 \cdot 10^{9} + 0.0\,I \\ -1.020024 \cdot 10^{-11} + 0.0\,I & 0.500000 + 0.0\,I & 3.396713 \cdot 10^{-11} + 0.0\,I \end{bmatrix}$$

The result is M, up to some very small error terms.

Actually, it is possible to get much greater accuracy. There is a Maple variable, Digits, which controls the number of digits that Maple uses in numeric calculations. The default is 10, but you can set it to 20, or 100, or ...

B.3. Differential equations

This annotated Maple session uses the `DEtools` package to plot and calculate various aspects of two systems of equations.

Load the `DEtools` package, and look at the help page:

```
> with(DEtools):
```

```
> help(DEtools);
```

The first system is a constant coefficient system:

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -2 & 2 \\ 1 & -3 \end{bmatrix}.$$

(This is the system shown in Figure 4.4, with a different time scale.)

You can define the system as a *set* (using `{...}`) or as a *list* (using `[...]`) of differential equations. It will be convenient to have a set later:

```
> LSys := \{diff(x(t),t) = -2*x(t)+2*y(t), diff(y(t),t) = x(t)-3*y(t)\};
```

$$LSys := \left\{ \frac{d}{dt}x\left(t\right) = 2\,y\left(t\right) - 2\,x\left(t\right), \frac{d}{dt}y\left(t\right) = x\left(t\right) - 3\,y\left(t\right) \right\}$$

There is nothing special about the name `LSys`. I used it because this is a linear system of equations.

Maple can solve this exactly:

```
> dsolve(LSys);
```

$$\left\{x\left(t\right) = \_C1\,\mathrm{e}^{-t} + \_C2\,\mathrm{e}^{-4\,t}, y\left(t\right) = 1/2\,\_C1\,\mathrm{e}^{-t} - \_C2\,\mathrm{e}^{-4\,t}\right\}$$

The symbols `_C1` and `_C2` are arbitrary constants. You can get a solution in terms of the initial conditions by specifying initial conditions *inside* the set `LSys`. If we want the solutions in terms of x_0 and y_0 then we need to add `x(0)=x_0, y(0)=y_0` too the set `LSys`. To avoid rewriting the entire set we use the `union` operator:

```
>   dsolve(LSys union \{x(0)=x_0,y(0)=y_0\});
```

$$\left\{x\left(t\right) = \left(2/3\,y\_0 + 2/3\,x\_0\right)\mathrm{e}^{-t} + \left(-2/3\,y\_0 + 1/3\,x\_0\right)\mathrm{e}^{-4\,t},\right.$$
$$\left.y\left(t\right) = 1/2\,\left(2/3\,y\_0 + 2/3\,x\_0\right)\mathrm{e}^{-t} - \left(-2/3\,y\_0 + 1/3\,x\_0\right)\mathrm{e}^{-4\,t}\right\}$$

Now we are going to draw some orbits in the phase portrait. We will need to specify initial conditions for the different solutions. This must be set up as a list or a set; we'll use a list, because we might want to color the curves differently and in a list the order is fixed. Each initial condition is specified as a list, containing initial values for the variables.

```
>   LIcs := [[x(0)=1,y(0)=0],[x(0)=3,y(0)=0],[x(0)=0,y(0)=1],
  [x(0)=0,y(0)=3],[x(0)=3,y(0)=3]];
```

$$LIcs := [[x\left(0\right) = 1, y\left(0\right) = 0], [x\left(0\right) = 3, y\left(0\right) = 0], [x\left(0\right) = 0, y\left(0\right) = 1],$$
$$[x\left(0\right) = 0, y\left(0\right) = 3], [x\left(0\right) = 3, y\left(0\right) = 3]]$$

These are just sample initial conditions; you can add more to get a more detailed phase portrait.

Now we ask for a plot. The `DEplot` procedure has several forms (look at the help); the form we are going to use requires that first six arguments, in order, are:

(1) The dynamical system.
(2) The variables to plot; this is `[x(t),y(t)]`.
(3) The range for t, in the form `t=firstT .. lastT`.
(4) The range for x, as above.
(5) The range for y, as above.
(6) The initial conditions.

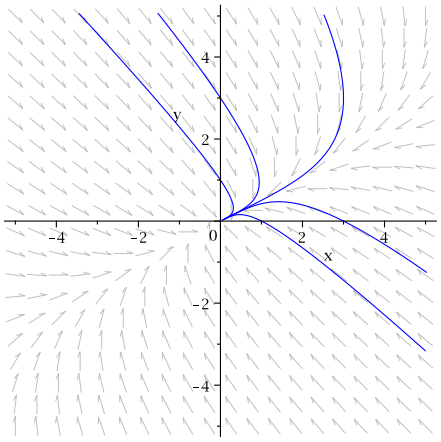
There are also many options that can be specified after the first six arguments. We'll use:

(1) `color`, the color of the arrows,
(2) `linecolor`, the color of the trajectories.
(3) `thickness`, the thickness of the trajectories.
(4) `numpoints`, the number of points to plot per curve. Use a higher number if the curves seem jagged. The default is 49.

Another useful option is `obsrange`, which indicates whether trajectories that leave the plotting area should be redrawn if they re-enter. This is not necessary for the current example. See help for `DEplot`, and for `plot`, for other options.

Here is the plot statement. The x and y ranges are arbitrary (but they should include the initial contidiosn). The t range was determined by experiment.

```
> DEplot(LSys,[x(t),y(t)],t=-1..3,x=-5..5,y=-5..5,LIcs,color=gray,
  linecolor=blue,thickness=1,numpoints=500);
```



Next we look at the non-linear system

$$\frac{d}{dt}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} y - 2x \\ -x + .1y^3 \end{bmatrix}.$$

First define the system:

```
> NSys := {diff(x(t),t)=y(t)-2*x(t), diff(y(t),t)=-x(t)+.1*y(t)^3};
```

$$NSys := \left\{ \frac{d}{dt}x\left(t\right) = y\left(t\right) - 2\,x\left(t\right), \frac{d}{dt}y\left(t\right) = -x\left(t\right) + 0.1\,\left(y\left(t\right)\right)^{3} \right\}$$

You can ask Maple for an analytic solution using `dsolve`, but you won't get anything useful. However, it is a good idea to locate the equilibria:

```
> solve({y-2*x=0,-x+.1*y^3=0});
```

$$\{x = 0.0, y = 0.0\}, \{x = 1.118033989, y = 2.236067978\}, \{x = -1.118033989, y = -2.236067978\}$$

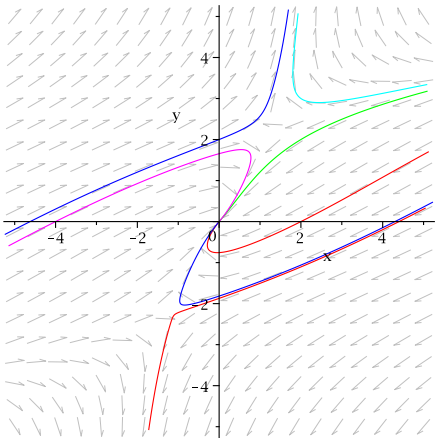
We will draw a phase portrait, so we need some initial conditions. We'll specify them using the alternate style: `[x(0)=2,y(0)=3]` is written as `[0,2,3]`. These initial values were determined by experimentation: I specified one or two, looked at the graph, decided to put in a couple more or make adjustments, etc.

```
> NIcs := [[0,0,2],[0,2,0],[0,2,2],[0,2,3],[0,-4,0],[0,4.3,0],
  [0,4.4,0]];
```

$$NIcs := [[0, 0, 2], [0, 2, 0], [0, 2, 2], [0, 2, 3], [0, -4, 0], [0, 4.3, 0], [0, 4.4, 0]]$$

In the plot command I asked for an assortment of colors for the orbits. (Search the help for `colornames` for a list of the colors that Maple understands.)

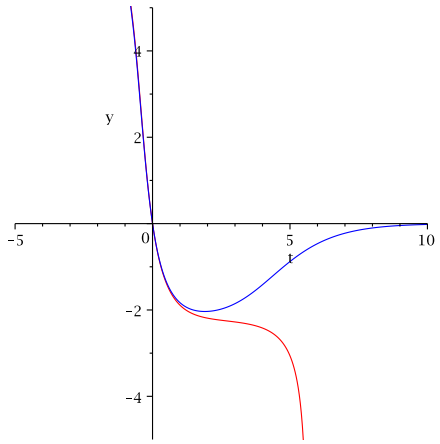
```
> DEplot(NSys,[x(t),y(t)],t=-5..10,x=-5..5,y=-5..5,NIcs,color=gray,
  linecolor=[blue,red,green,cyan,magenta,blue,red],thickness=1,
  numpoints=500);
```



From the graph it seems that there is something interesting going on for orbits through approximately $(4.3, 0)$. The orbit going through $(4.3, 0)$ (colored blue) seems to converge to the origin, while the orbit through $(4.4, 0)$ (colored red) seems to go to ∞ in the third quadrant.

To look at this more carefully I ran a different DEplot command, with a few changes. I changed the initial conditions, since I only wanted to look at two solutions. Instead of defining a variable for the initial conditions I put the list directly in the `DEplot` command. I also changed the `linecolor` option, since it must have as many colors as there are initial conditions. Finally, I used the `scene` option to specify a graph of y versus t.

```
> DEplot(NSys,[x(t),y(t)],t=-5..10,x=-5..5,y=-5..5,
  [[x(0)=4.4,y(0)=0],[x(0)=4.3,y(0)=0]],color=gray,
  linecolor=[red,blue],thickness=1,numpoints=400,scene=[t,y]);
```



From this it seems that, as $t \to \infty$, $y(t) \to 0$ on one of the solutions but $y \to -\infty$ on the other. It is hard to quantify this from the graph, but Maple will give us a numeric version of the solution. This uses the `dsolve` procedure as above, but with a `numeric` option. Since we are looking at only one solution curve it is necessary to specify a single initial condition, and this version of `dsolve` requires that the differential equations and the initial conditions be specified in a single set. We defined `NSys` as a set of differential equations, so we just add the initial conditions to this set using the `union` operator as above:

```
> orb1 := dsolve(NSys union {x(0)=4.4,y(0)=0}, numeric, [x(t),y(t)]);
```

$$orb1 := \text{proc}\,(x\_rkf45)\,\ldots\,\text{endproc}$$

The result is a *procedure*, which we can use like a function to find points on the solution curve. For example (after some experimentation):

```
> orb1(5);
```

$$[t = 5., x(t) = -1.34498191430922986, y(t) = -3.06222234749248834]$$

```
> orb1(5.6);
```

$$[t = 5.6, x(t) = -1.97384367918946269, y(t) = -7.27741244337618465]$$

```
> orb1(5.69);
```

$$[t = 5.69, x(t) = -2.60991725995272894, y(t) = -27.6725723616629864]$$

```
> orb1(5.696);
```

$$[t = 5.696, x(t) = -2.83579696162862049, y(t) = -96.8877322592397405]$$

```
> orb1(5.7);
```

Error, (in orb1) cannot evaluate the solution further right of 5.6965326, probably a singularity

It looks like the solution through $(4.4, 0)$ blows up, at $t \approx 5.6965326$.

We can set up the solution through $(4.3, 0)$ similarly:

```
> orb2 := dsolve(NSys union {x(0)=4.3,y(0)=0}, numeric, [x(t),y(t)]);
```

Then, after some experimentation with `orb2`, we see that $x(10) \approx -0.0161$ and $x(11) \approx -0.0069$.

Suppose we want to determine t so that $x(t) = -.01$. We can't do this analytically, since we don't have a formula for $x(t)$. However, from the calculations of $x(10)$ and $x(11)$ it appears that t is between 10 and 11. To find a better approximation to t we can try searching further, but it is simpler to prepare a table of values. Rather than writing a loop to do this, we use another version of `dsolve` that computes a table of values. As part of the input we need to specify a sequence of t values to use, in the format `output = array([list of t values])`. In the following we use Maple's `seq` function to fill in the t values. The `seq` function generates a sequence from a formula, and we'll generate the t values at 0.1 intervals between 10 and 11. I told Maple to round the answers and display 10 digits: To do this, click on the "Tools > Options" menu item, and select the "Precision" tab.

```
> dsolve(NSys union {x(0)=4.3, y(0)=0}, numeric, [x(t),y(t)],
  output = array([seq(10+i/10, i = 0 .. 10)]));
```

$$\left[\begin{array}{c} \left[\begin{array}{ccc} t & x\,(t) & y\,(t) \end{array}\right] \\ \left[\begin{array}{ccc} 10.0 & -0.0161789040 & -0.0187261898 \\ 10.1000000000 & -0.0148697645 & -0.0171746947 \\ 10.2000000000 & -0.0136632824 & -0.0157489123 \\ 10.3000000000 & -0.0125517681 & -0.0144389579 \\ 10.4000000000 & -0.0115280660 & -0.0132356876 \\ 10.5000000000 & -0.0105855389 & -0.0121306775 \\ 10.6000000000 & -0.0097180050 & -0.0111161205 \\ 10.7000000000 & -0.0089197184 & -0.0101847950 \\ 10.8000000000 & -0.0081853621 & -0.0093300578 \\ 10.9000000000 & -0.0075100013 & -0.0085457726 \\ 11.0 & -0.0068890505 & -0.0078262601 \end{array}\right] \end{array}\right]$$

Apparently $x(t) = -.01$ is satisfied for t somewhere between 10.5 and 10.6. If you regenerate the table with steps of .01 between 10.5 and 10.6 then you'll find that t is between 10.56 and 10.57.