# Math 118, Spring 2,000

Shlomo Sternberg

October 10, 2000

# Contents

# Chapter 1

# Iterations and fixed points

## 1.1 Square roots

Perhaps the oldest algorithm in recorded history is the Babylonian algorithm (circa 2000BCE) for computing square roots: If we want to find the square root of a positive number $a$ we start with some approximation, $x_0 > 0$ and then recursively define

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{a}{x_n}\right). \tag{1.1}$$

This is a very effective algorithm which converges extremely rapidly. Here is an illustration. Suppose we want to find the square root of 2 and start with the really stupid approximation $x_0 = 99$. We apply (1.1) recursively thirteen times to obtain the values

99.00000000000000
49.51010101010101
24.77524840365297
12.42798706655775
6.29445708659966
3.30609848017316
1.95552056875300
1.48913306969968
1.41609819333465
1.41421481646475
1.41421356237365
1.41421356237309
1.41421356237309

to fourteen decimal places. For the first seven steps we are approximately dividing by two in passing from one step to the next, also (approximately) cutting the error - the deviation from the true value - in half. After line eight the accuracy improves dramatically: the ninth value, $1.416\ldots$ is correct to two decimal

places. The tenth value is correct to five decimal places, and the eleventh value is correct to eleven decimal places.

To see why this algorithm works so well (for general $a$), first observe that the algorithm is well defined, in that we are steadily taking the average of positive quantities, and hence, by induction, $x_n > 0$ for all $n$. Introduce the *relative error* in the $n-$th approximation:

$$e_n = \frac{x_n - \sqrt{a}}{\sqrt{a}}$$

so

$$x_n = (1 + e_n)\sqrt{a}.$$

As $x_n > 0$, it follows that

$$e_n > -1.$$

Then

$$x_{n+1} = \sqrt{a}\frac{1}{2}(1 + e_n + \frac{1}{1 + e_n}) = \sqrt{a}(1 + \frac{1}{2}\frac{e_n^2}{1 + e_n}).$$

This gives us a recursion formula for the relative error:

$$e_{n+1} = \frac{e_n^2}{2 + 2e_n}. \tag{1.2}$$

This implies that $e_1 > 0$ so after the first step we are always overshooting the mark. Now $2e_n < 2 + 2e_n$ so (1.2) implies that

$$e_{n+1} < \frac{1}{2}e_n$$

so the error is cut in half (at least) at each stage and hence, in particular,

$$x_1 > x_2 > \cdots,$$

the iterates are steadily decreasing. Eventually we will reach the stage that

$$e_n < 1.$$

From this point on, we use the inequality $2 + 2e_n > 2$ in (1.2) and we get the estimate

$$e_{n+1} < \frac{1}{2}e_n^2. \tag{1.3}$$

So if we renumber our approximation so that $0 \leq e_0 < 1$ then (ignoring the $1/2$ factor in (1.3)) we have

$$0 \leq e_n < e_0^{2^n}, \tag{1.4}$$

an exponential rate of convergence.

If we had started with an $x_0 < 0$ then all the iterates would be $< 0$ and we would get exponential convergence to $-\sqrt{a}$. Of course, had we been so foolish as to pick $x_0 = 0$ we could not get the iteration started.

## 1.2 Newton's method

This is a generalization of the above algorithm to find the zeros of a function $P = P(x)$ and which reduces to (1.1) when $P(x) = x^2 - a$. It is

$$x_{n+1} = x_n - \frac{P(x_n)}{P'(x_n)}. \tag{1.5}$$

If we take $P(x) = x^2 - a$ then $P'(x) = 2x$ the expression on the right in (1.5) is

$$\frac{1}{2}\left(x_n + \frac{a}{x_n}\right)$$

so (1.5) reduces to (1.1).

Also notice that if $x$ is a "fixed point" of this iteration scheme, i.e. if

$$x = x - \frac{P(x)}{P'(x)}$$

then $P(x) = 0$ and we have a solution to our problem. To the extent that $x_{n+1}$ is "close to" $x_n$ we will be close to a solution (the degree of closeness depending on the size of $P(x_n)$).

In the general case we can not expect that "most" points will converge to a zero of $P$ as was the case in the square root algorithm. After all, $P$ might not have *any* zeros. Nevertheless, we will show in this section that if we are "close enough" to a zero - that $P(x_0)$ is "sufficiently small" in a sense to be made precise - then (1.5) converges exponentially fast to a zero.

### 1.2.1 The guts of the method.

Before embarking on the formal proof, let us describe what is going on on the assumption that we know the existence of a zero - say by graphically plotting the function. So let $z$ be a zero for the function $f$ of a real variable, and let $x$ be a point in the interval $(z - \mu, z + \mu)$ of radius $\mu$ about $z$. Then

$$-f(x) = f(z) - f(x) = \int_x^z f'(s)ds$$

so

$$-f(x) - (z - x)f'(x) = \int_x^z (f'(s) - f'(x))ds.$$

Assuming $f'(x) \neq 0$ we may divide both sides by $f'(x)$ to obtain

$$\left(x - \frac{f(x)}{f'(x)}\right) - z = \frac{1}{f'(x)} \int_x^z (f'(s) - f'(x))ds. \tag{1.6}$$

Assume that for all $y \in (z - \mu, z + \mu)$ we have

$$|f'(y)| \geq \rho > 0 \tag{1.7}$$

$$|f'(y_1) - f'(y_2)| \leq \delta |y_1 - y_2| \tag{1.8}$$

$$\mu \leq \rho/\delta. \tag{1.9}$$

Then setting $x = x_{old}$ in (1.6) and letting

$$x_{new} := x - \frac{f(x)}{f'(x)}$$

in (1.6) we obtain

$$|x_{new} - z| \leq \frac{\delta}{\rho} \int_{x_{old}}^{z} |s - x_{old}| ds = \frac{\delta}{2\rho} |x_{old} - z|^2.$$

Since $|x_{old} - z| < \mu$ it follows that

$$|x_{new} - z| \leq \frac{1}{2}\mu$$

by 1.9). Thus the iteration

$$x \mapsto x - \frac{f(x)}{f'(x)} \tag{1.10}$$

is well defined. At each stage it more than halves the distance to the zero and has the quadratic convergence property

$$|x_{new} - z| \leq \frac{\delta}{2\rho} |x_{old} - z|^2.$$

As we have seen from the application of Newton's method to a cubic, unless the stringent hypotheses are satisfied, there is no guarantee that the process will converge to the nearest root, or converge at all. Furthermore, encoding a computation for $f'(x)$ may be difficult. In practice, one replaces $f'$ by an approximation, and only allows Newton's method to proceed if in fact it does not take us out of the interval. We will return to these points, but first rephrase the above argument in terms of a vector variable.

### 1.2.2 A vector version.

Now let $f$ a function of a vector variable, with a zero at $z$ and $x$ a point in the ball of radius $\mu$ centered at $z$. Let $v_x := z - x$ and consider the function

$$t :\mapsto f(x + tv_x)$$

which takes the value $f(z)$ when $t = 1$ and the value $f(x)$ when $t = 0$. Differentiating with respect to $t$ using the chain rule gives $f'(x + tv_x)v_x$ (where $f'$ denotes the derivative =(the Jacobian matrix) of $f$. Hence

$$-f(x) = f(z) - f(x) = \int_0^1 f'(x + tv_x)v_x dt.$$

This gives

$$-f(x) - f'(x)v_x = -f(x) - f'(x)(z - x) = \int_0^1 [f'(x + tv_x) - f'(x)]v_x dt.$$

Applying $[f'(x)]^{-1}$ (which we assume to exist) gives the analogue of (1.6):

$$\big(x - [f'(x)]^{-1} f(x)\big) - z = [f'(x_)]^{-1} \int_0^1 [f'(x + tv_x) - f'(x)]v_x dt.$$

Assume now

$$\|[f'(y)]^{-1}\| \leq \rho^{-1} \tag{1.11}$$
$$\|f'(y_1) - f(y_2\| \leq \delta\|y_1 - y_2\| \tag{1.12}$$

for all $y, y_1, y_2$ in the ball of radius $\mu$ about $z$, and assume also that (1.9) holds. Setting $x_{old} = x$ and

$$x_{new} := x_{old} - [f'(x_{old})]^{-1} f(x_{old})$$

gives

$$\|x_{new} - z\| \leq \frac{\delta}{\rho} \int_0^1 t\|v_x\|\|v_x\|dt = \frac{\delta}{2\rho}\|x_{old} - z\|^2.$$

From here on the argument is the same as in the one dimensional case.

### 1.2.3 Implementation.

We return to the one dimensional case.

In numerical practice we have to deal with two problems: it may not be easy to encode the derivative, and we may not be able to tell in advance whether the conditions for Newton's method to work are indeed fulfilled.

In case $f$ is a polynomial, MATLAB has an efficient command "polyder" for computing the derivative of $f$. Otherwise we replace the derivative by the slope of the secant, which requires the input of two initial values, call them $x_-$ and $x_c$ and replaces the derivative in Newton's method by

$$f'_{app}(x_c) = \frac{f(x_c) - f(x_-)}{x_c - x_-}.$$

So at each stage of the Newton iteration we carry along two values of $x$, the "current value" denoted say by "xc" and the "old value" denoted by "x_". We also carry along two values of $f$, the value of $f$ at xc denoted by fc and the value of $f$ at x_ denoted by f_. So the Newton iteration will look like

```
fpc=(fc-f_)/(xc-x_);
xnew=xc-fc/fpc;
x_-=xc; f_=fc;
xc=xnew; fc=feval(fname,xc);
```

In the last line, the command feval is the MATLAB evaluation of a function command: if fname is a "script" (that is an expression enclosed in ' ') giving

the name of a function, then feval(fname,x) evaluates the function at the point x.

The second issue - that of deciding whether Newton's method should be used at all - is handled as follows: If the zero in question is a critical point, so that $f'(z) = 0$, there is no chance of Newton's method working. So let us assume that $f'(z) \neq 0$, which means that $f$ changes sign at $z$, a fact that we can verify by looking at the graph of $f$. So assume that we have found an interval $[a, b]$ containing the zero we are looking for, and such that $f$ takes on opposite signs at the end-points:

$$f(a)f(b) < 0.$$

A sure but slow method on narrowing in on a zero of $f$ contained in this interval is the "bisection method": evaluate $f$ at the midpoint $\frac{1}{2}(a + b)$. If this value has a sign opposite to that of $f(a)$ replace $b$ by $\frac{1}{2}(a + b)$. Otherwise replace $a$ by $\frac{1}{2}(a + b)$. This produces an interval of half the length of $[a, b]$ containing a zero.

The idea now is to check at each stage whether Newton's method leaves us in the interval, in which case we apply it, or else we apply the bisection method.

We now turn to the more difficult existence problem.

### 1.2.4 The existence theorem.

For the purposes of the proof, in order to simplify the notation, let us assume that we have "shifted our coordinates" so as to take $x_0 = 0$. Also let

$$B = \{x : |x| \leq 1\}.$$

We need to assume that $P'(x)$ is nowhere zero, and that $P''(x)$ is bounded. In fact, we assume that there is a constant $K$ such that

$$|P'(x)^{-1}| \leq K, \quad |P''(x)| \leq K, \quad \forall x \in B. \tag{1.13}$$

**Proposition 1.2.1** *Let $\tau = \frac{3}{2}$ and choose the $K$ in (1.13) so that*

$$K \geq 2^{3/4}.$$

*Let*

$$c = \frac{8}{3} \ln K.$$

*Then if*

$$P(0) \leq K^{-5} \tag{1.14}$$

*the recursion (1.5) starting with $x_0 = 0$ satisfies*

$$x_n \in B \quad \forall n \tag{1.15}$$

*and*

$$|x_n - x_{n-1}| \leq e^{-c\tau^n}. \tag{1.16}$$

*In particular, the sequence $\{x_n\}$ converges to a zero of $P$.*

**Proof.** In fact, we will prove a somewhat more general result. So we will let $\tau$ be any real number satisfying

$$1 < \tau < 2$$

and we will choose $c$ in terms of $K$ and $\tau$ to make the proof work. First of all we notice that (1.15) is a consequence of (1.16) if $c$ is sufficiently large. In fact,

$$x_j = (x_j - x_{j-1}) + \cdots + (x_1 - x_0)$$

so

$$|x_j| \leq |x_j - x_{j-1}| + \cdots + |x_1 - x_0|.$$

Using (1.16) for each term on the right gives

$$|x_j| \leq \sum_1^j e^{-c\tau^n} < \sum_1^\infty e^{-c\tau^n} < \sum_1^\infty e^{-cn(\tau-1)} = \frac{e^{-c(\tau-1)}}{1 - e^{-c(\tau-1)}}.$$

Here the third inequality follows from writing

$$\tau = 1 + (\tau - 1)$$

so by the binomial formula

$$\tau^n = 1 + n(\tau - 1) + \cdots > n(\tau - 1)$$

since $\tau > 1$. The equality is obtained by summing the geometric series. So if we choose $c$ sufficiently large that

$$\frac{e^{-c(\tau-1)}}{1 - e^{-c(\tau-1)}} \leq 1 \tag{1.17}$$

(1.15) follows from (1.16). This choice of $c$ is conditioned by our choice of $\tau$. But at least we now know that if we can arrange that (1.16) holds, then by choosing a possibly smaller value of $c$ (so that (1.16) continues to hold) we can guarantee that the algorithm keeps going.

So let us try to prove (1.16) by induction. If we assume it is true for $n$, we may write

$$|x_{n+1} - x_n| = |S_n P(x_n)|$$

where we set

$$S_n = P'(x_n)^{-1}. \tag{1.18}$$

We use the first inequality in (1.13) and the definition (1.5) for the case $n-1$ (which says that $x_n = x_{n-1} - S_{n-1}P(x_{n-1})$ to get

$$|S_n P(x_n)| \leq K|P(x_{n-1} - S_{n-1}P(x_{n-1}))|. \tag{1.19}$$

Taylor's formula with remainder says that for any twice continuously differentiable function $f$,

$$f(y+h) = f(y) + f'(y)h + R(y,h) \quad \text{where } |R(y,h)| \leq \frac{1}{2}\sup_z |f''(z)|h^2$$

where the supremum is taken over the interval between $y$ and $y+h$. If we use Taylor's formula with remainder with

$$f = P,\ y = P(x_{n-1}), \text{and } h = S_{n-1}P(x_{n-1}) = x_n - x_{n-1}$$

and the second inequality in (1.13) to estimate the second derivative, we obtain

$$|P(x_{n-1} - S_{n-1}P(x_{n-1}))| \leq |P(x_{n-1}) - P'(x_{n-1})S_{n-1}P(x_{n-1})| + K|x_n - x_{n-1}|^2.$$

Substituting this inequality into (1.19), we get

$$|x_{n+1} - x_n| \leq K|P(x_{n-1}) - P'(x_{n-1})S_{n-1}P(x_{n-1})| + K^2|x_n - x_{n-1}|^2. \quad (1.20)$$

Now since $S_{n-1} = P'(x_{n-1})^{-1}$ the first term on the right vanishes and we get

$$|x_{n+1} - x_n| \leq K^2|x_n - x_{n-1}|^2 \leq K^2 e^{-2c\tau^n}.$$

So in order to pass from $n$ to $n+1$ in (1.16) we must have

$$K^2 e^{-2c\tau^n} \leq e^{-c\tau^{n+1}}$$

or

$$K^2 \leq e^{c(2-\tau)\tau}. \quad (1.21)$$

Since $\tau < 2$ we can arrange for this last inequality to hold if we choose $c$ sufficiently large. To get started, we must verify (1.16) for $n = 1$ This says

$$S_0 P(0) \leq e^{-c\tau}$$

or

$$|P(0)| \leq \frac{e^{-c\tau}}{K}. \quad (1.22)$$

So we have proved:

**Theorem 1.2.1** *Suppose that* (1.13) *holds and we have chosen* $K$ *and* $c$ *so that* (1.17) *and* (1.21) *hold. Then if* $P(0)$ *satisfies* (1.22) *the Newton iteration scheme converges exponentially in the sense that* (1.16) *holds.*

If we choose $\tau = \frac{3}{2}$ as in the proposition, let $c$ be given by $K^2 = e^{3c/4}$ so that (1.21) just holds. This is our choice in the proposition. The inequality $K \geq 2^{3/4}$ implies that $e^{3c/4} \geq 4^{3/4}$ or

$$e^c \geq 4.$$

This implies that

$$e^{-c/2} \leq \frac{1}{2}$$

so (1.17) holds. Then

$$e^{-c\tau} = e^{-3c/2} = K^{-4}$$

so (1.22) becomes $|P(0)| \leq K^{-5}$ completing the proof of the proposition.

We have put in all the gory details, but it is worth reviewing the guts of the argument, and seeing how things differ from the special case of finding the square root. Our algorithm is

$$x_{n+1} = x_n - S_n[P(x_n)] \tag{1.23}$$

where $S_n$ is chosen as (1.18). Taylor's formula gave (1.20) and with the choice (1.18) we get

$$|x_{n+1} - x_n| \leq K^2 |x_n - x_{n-1}|^2. \tag{1.24}$$

In contrast to (1.4) we do not know that $K \leq 1$ so, once we get going, we can't quite conclude that the error vanishes as

$$r^{\tau^n}$$

with $\tau = 2$. But we can arrange that we eventually have such exponential convergence with any $\tau < 2$.

### 1.2.5 Basins of attraction.

The more decisive difference has to do with the "basins of attraction" of the solutions. For the square root, starting with any positive number ends us up with the positive square root. This was the effect of the $e_{n+1} < \frac{1}{2} e_n$ argument which eventually gets us to the region where the exponential convergence takes over. Every negative number leads us to the negative square root. So the "basin of attraction" of the positive square root is the entire positive half axis, and the "basin of attraction" of the negative square root is the entire negative half axis. The only "bad" point belonging to no basin of attraction is the point 0.

Even for cubic polynomials the *global* behavior of Newton's method is extraordinarily complicated. For example, consider the polynomial

$$P(x) = x^3 - x,$$

with roots at 0 and $\pm 1$. We have

$$x - \frac{P(x)}{P'(x)} = x - \frac{x^3 - x}{3x^2 - 1} = \frac{2x^3}{3x^2 - 1}$$

so Newton's method in this case says to set

$$x_{n+1} = \frac{2x_n^3}{3x_n^2 - 1}. \tag{1.25}$$

There are obvious "bad" points where we can't get started, due to the vanishing of the denominator, $P'(x)$. These are the points $x = \pm\sqrt{1/3}$. These two points are the analogues of the point 0 in the square root algorithm.

We know from the general theory, that any point sufficiently close to 1 will converge to 1 under Newton's method and similarly for the other two roots, 0 and -1.

If $x > 1$, then

$$2x^3 > 3x^2 - 1$$

since both sides agree at $x = 1$ and the left side is increasing faster, as its derivative is $6x^2$ while the derivative of the right hand side is only $6x$. This implies that if we start to the right of $x = 1$ we will stay to the right. The same argument shows that

$$2x^3 < 3x^3 - x$$

for $x > 1$. This is the same as

$$\frac{2x^3}{3x^2 - 1} < x,$$

which implies that if we start with $x_0 > 1$ we have $x_0 > x_1 > x_2 > \cdots$ and eventually we will reach the region where the exponential convergence takes over. So every point to the right of $x = 1$ is in the basin of attraction of the root $x = 1$. By symmetry, every point to the left of $x = -1$ will converge to $-1$.

But let us examine what happens in the interval $-1 < x_0 < 1$. For example, suppose we start with $x_0 = -\frac{1}{2}$. Then one application of Newton's method gives

$$x_1 = \frac{-.25}{3 \times .25 - 1} = 1.$$

In other words, one application of Newton's method lands us on the root $x = 1$, right on the nose. Notice that although $-.5$ is halfway between the roots $-1$ and 0, we land on the farther root $x = 1$. In fact, by continuity, if we start with $x_0$ close to $-.5$, then $x_1$ must be close to 1. So all points, $x_0$, sufficiently close to $-.5$ will have $x_1$ in the region where exponential convergence to $x = 1$ takes over. In other words, the basin of attraction of $x = 1$ will include points to the immediate left of $-.5$, even though $-1$ is the closest root.

Suppose we have a point $x$ which satisfies

$$\frac{2x^3}{3x^2 - 1} = -x.$$

So one application of Newton's method lands us at $-x$, and a second lands us back at $x$. The above equation is the same as

$$0 = 5x^3 - x = x(5x^2 - 1)$$

which has roots, $x = 0, \pm\sqrt{1/5}$. So the points $\pm\sqrt{1/5}$ form a cycle of order two: Newton's method cycles between these two points and hence does not converge to any root.

In fact, in the interval $(-1, 1)$ there are infinitely many points that don't converge to any root. We will return to a description of this complicated type of phenomenon later. If we apply Newton's method to cubic or higher degree polynomials and to complex numbers instead of real numbers, the results are even more spectacular. This phenomenon was first discovered by Cayley, and was published in an article which appeared in the second issue of the American Journal of Mathematics in 1879. This paper of Cayley's was the starting point for many future investigations.

## 1.3 The implicit function theorem

Let us return to the positive aspect of Newton's method. You might ask, how can we ever guarantee in advance that an inequality such as (1.14) holds? The answer comes from considering not a single function, $P$, but rather a parameterized family of functions: Suppose that $u$ ranges over some interval, or more generally, over some region in a vector space. To fix the notation, suppose that this region contains the origin, $\mathbf{0}$. Suppose that $P$ is a function of $u$ and $x$, and depends continuously on $(u, x)$. Suppose that as a function of $x$, the function $P$ is twice differentiable and satisfies (1.13) for all values of $u$ (with the same fixed $K$). Finally, suppose that

$$P(\mathbf{0}, 0) = 0. \tag{1.26}$$

Then the continuity of $P$ guarantees that for $|u|$ and $|x_0|$ sufficiently small, the condition (1.14) holds, that is

$$|P(u, x_0)| < r$$

where $r$ is small enough to guarantee that $x_0$ is in the basin of attraction of a zero of the function $P(u, \cdot)$. In particular, this means that for $|u|$ sufficiently small, we can find an $\epsilon > 0$ such that all $x_0$ satisfying $|x_0| < \epsilon$ are in the basin of attraction of the same zero of $P(u, \cdot)$. By choosing a smaller neighborhood, given say by $|u| < \delta$, starting with $x_0 = 0$ and applying Newton's method to $P(u, \cdot)$, we obtain a sequence of $x$ values which converges exponentially to a solution of

$$P(u, x) = 0. \tag{1.27}$$

satisfying

$$|x| < \epsilon.$$

Furthermore, starting with any $x_0$ satisfying $|x_0| < \epsilon$ we also get exponential convergence to the same zero. In particular, there can not be two distinct solutions to (1.27) satisfying $|x| < \epsilon$, since starting Newton's method at a zero gives (inductively) $x_n = x_0$ for all $n$. Thus we have constructed a unique function

$$x = g(u)$$

satisfying

$$P(u, g(u)) \equiv 0. \tag{1.28}$$

This is the guts of the implicit function theorem. We have proved it under assumptions which involve the second derivative of $P$ which are not necessary for the truth of the theorem. (We will remedy this in Chapter**??**.) However these stronger assumptions that we have made do guarantee exponential convergence of our algorithm.

For the sake of completeness, we discuss the basic properties of the function $g$: its continuity, differentiability, and the computation of its derivative.

1.*Uniqueness implies continuity.* We wish to prove that $g$ is continuous at any point $u$ in a neighborhood of $\mathbf{0}$. This means: given $\beta > 0$ we can find $\alpha > 0$ such that

$$|h| < \alpha \Rightarrow |g(u+h) - g(u)| < \beta. \tag{1.29}$$

We know that this is true at $u = 0$, where we could choose any $\epsilon' > 0$ at will, and then conclude that there is a $\delta' > 0$ with $|g(u)| < \epsilon'$ if $|u| < \delta'$. To prove (1.29) at a general point, just choose $(u, g(u))$ instead of $(\mathbf{0}, 0)$ as the origin of our coordinates. and apply the preceding results to this new data. We obtain a solution $f$ to the equation $P(u+h, f(u+h)) = 0$ with $f(u) = g(u)$ which is continuous at $h = 0$. In particular, for $|h|$ sufficiently small, we will have $|u+h| \leq \delta$, and $|f(u+h)| < \epsilon$, our original $\epsilon$ and $\delta$ in the definition of $g$. The uniqueness of the solution to our original equation then implies that $f(u+h) = g(u+h)$, proving (1.29).

2.*Differentiability.* Suppose that $P$ is continuously differentiable with respect to all variables. We have

$$0 \equiv P(u+h, g(u+h)) - P(u, g(u)$$

so, by the definition of the derivative,

$$0 = \frac{\partial P}{\partial u}h + \frac{\partial P}{\partial x}[g(u+h) - g(u)] + o(h) + o[g(u+h) - g(u)].$$

If $u$ is a vector variable, say $\in \mathbf{R}^n$, then $\frac{\partial P}{\partial u}$ is a matrix. The terminology $o(s)$ means some expression which approaches zero so that $o(s)/s \to 0$. So

$$g(u+h) - g(u) = -\left[\frac{\partial P}{\partial x}\right]^{-1}\left[\frac{\partial P}{\partial u}\right]h - o(h) - \left[\frac{\partial P}{\partial x}\right]^{-1} o[g(u+h) - g(u)]. \tag{1.30}$$

As a first pass through this equation, observe that by the continuity that we have already proved, we know that $[g(u+h) - g(u)] \to 0$ as $h \to 0$. The expression $o([g(u+h) - g(u)])$ is, by definition of $o$, smaller than any constant times $|g(u+h) - g(u)|$ provided that $|g(u+h) - g(u)|$ itself is sufficiently small. This means that for sufficiently small $[g(u+h) - g(u)]$ we have

$$|o[g(u+h) - g(u)]| \leq \frac{1}{2K}|g(u+h) - g(u)|$$

where we may choose $K$ so that $|\left[\frac{\partial P}{\partial x}\right]^{-1}| \leq K$. So bringing the last term over to the other side gives

$$|g(u+h) - g(u)| - \frac{1}{2}|g(u+h) - g(u)| \leq |\left[\frac{\partial P}{\partial x}\right]^{-1}\left[\frac{\partial P}{\partial u}\right]h| + o(|h|),$$

and we get an estimate of the form

$$|g(u+h) - g(u)| \leq M|h|$$

for some suitable constant, $M$. But then the term $o[g(u+h)-g(u)]$ becomes $o(h)$. Plugging this back into our equation (1.30) shows that $g$ is differentiable with

$$\frac{\partial g}{\partial u} = -\left[\frac{\partial P}{\partial x}\right]^{-1}\left[\frac{\partial P}{\partial u}\right]. \tag{1.31}$$

To summarize, the implicit function theorem says:

**Theorem 1.3.1 The implicit function theorem.** *Let $P = P(u,x)$ be a differentiable function with $P(0,0) = 0$ and $\left[\frac{\partial P}{\partial x}\right](0,0)$ invertible. Then there exist $\delta > 0$ and $\epsilon > 0$ such that $P(u,x) = 0$ has a unique solution with $|x| < \epsilon$ for each $|u| < \delta$. This defines the function $x = g(u)$. The function $g$ is differentiable and its derivative is given by (1.31).*

We have proved the theorem under more stringent hypotheses in order to get an exponential rate of convergence to the solution. We will provide the details of the more general version, as a consequence of the contraction fixed point theorem, later on. We should point out now, however, that nothing in our discussion of Newton's method or the implicit function theorem depended on $x$ being a single real variable. The entire discussion goes through unchanged if $x$ is a vector variable. Then $\partial P/\partial x$ is a matrix, and (1.31) must be understood as matrix multiplication. Similarly, the condition on the second derivative of $p$ must be understood in terms of matrix norms. We will return to these points later.

## 1.4 Attractors and repellers

We introduce some notation which we will be using for the next few chapters. Let $F : X \to X$ be a differentiable map where $X$ is an interval on the real line. A point $p \in X$ is called a *fixed point* if

$$F(p) = p.$$

A fixed point $a$ is called an *attractor* or an *attractive* fixed point or a *stable* fixed point if

$$|F'(a)| < 1. \tag{1.32}$$

Points sufficiently close to an attractive fixed point, $a$, converge to $a$ geometrically upon iteration. Indeed,

$$F(x) - a = F(x) - F(a) = F'(a)(x-a) + o(x-a)$$

by the definition of the derivative. Hence taking $b < 1$ to be any number larger than $|F'(a)|$ then for $|x-a|$ sufficiently small, $|F(x)-a| \le b|x-a|$. So starting with $x_0 = x$ and iterating $x_{n+1} = F(x_n)$ gives a sequence of points with $|x_n - a| \le b^n |x-a|$.

The *basin of attraction* of an attractive fixed point is the set of all $x$ such that the sequence $\{x_n\}$ converges to $a$ where $x_0 = x$ and $x_{n+1} = F(x_n)$. thus

the basin of attraction of an attractive fixed point $a$ will always include a neighborhood of $a$, but it may also include points far away, and may be a very complicated set as we saw in the example of Newton's method applied to a cubic.

A fixed point, $r$, is called a *repeller* or a *repelling* or an *unstable* fixed point if

$$|F'(r)| > 1. \tag{1.33}$$

Points near a repelling fixed point (as in the case of our renormalization group example, in the next section) are pushed away upon iteration.

An attractive fixed point $s$ with

$$F'(s) = 0 \tag{1.34}$$

is called *superattractive* or *superstable*. Near a superstable fixed point, $s$, (as in the case of Newton's method) the iterates converge exponentially to $s$.

The notation $F^{\circ n}$ will mean the n-fold composition,

$$F^{\circ n} = F \circ F \circ \cdots \circ F \quad (n\text{times}).$$

A fixed point of $F^{\circ n}$ is called a *periodic* point of *period* $n$ . If $p$ is a periodic point of period $n$, then so are each of the points

$$p,\ F(p),\ F^{\circ 2}(p),\ \ldots, F^{\circ(n-1)}(p)$$

and the chain rule says that at each of these points the derivative of $F^{\circ n}$ is the same and is given by

$$(F^{\circ n})'(p) = F'(p)F'(F(p)) \cdots F'(F^{\circ(n-1)}(p)).$$

If any one of these points is an attractive fixed point for $F^n$ then so are all the others. We speak of an *attractive periodic orbit*. Similarly for repelling.

A periodic point will be superattractive for $F^{\circ n}$ if and only if at least one of the points $p, F(p), \ldots F^{\circ(n-1)}(p)$ satisfies $F'(q) = 0$.

## 1.5 Renormalization group

We illustrate these notions in an example: consider a hexagonal lattice in the plane. This means that each lattice point has six nearest neighbors. Let each site be occupied or not independently of the others with a common probability $0 \leq p \leq 1$ for occupation. In *percolation theory* the problem is to determine whether or not there is a positive probability for an infinitely large *cluster* of occupied sites. (By a cluster we mean a connected set of occupied sites.) We plot some figures with $p = .2,\ .5$, and .8 respectively. For problems such as this there is a *critical probability* $p_c$: for $p < p_c$ the probability of of an infinite cluster is zero, while it is positive for for $p > p_c$. One of the problems in percolation theory is to determine $p_c$ for a given lattice.

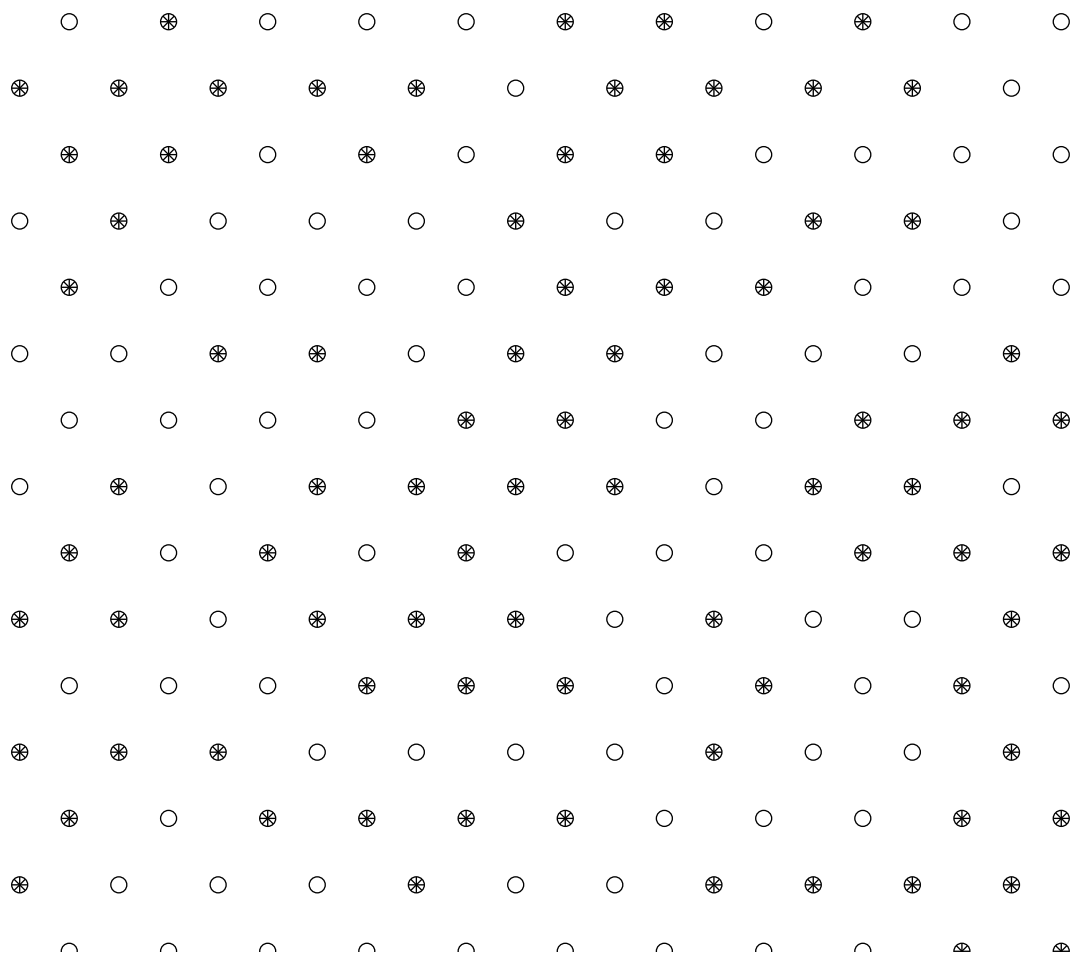

Figure 1.1: p=.2

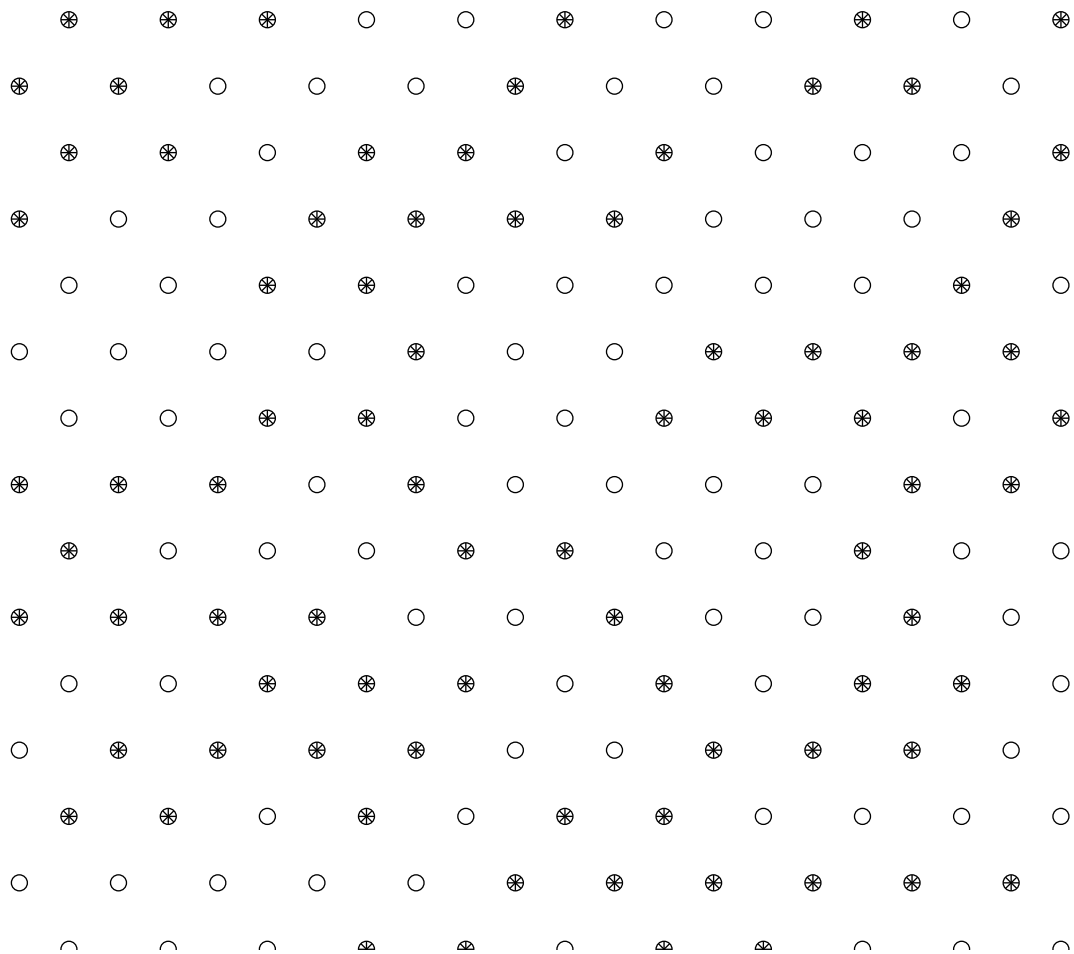

Figure 1.2: p=.5

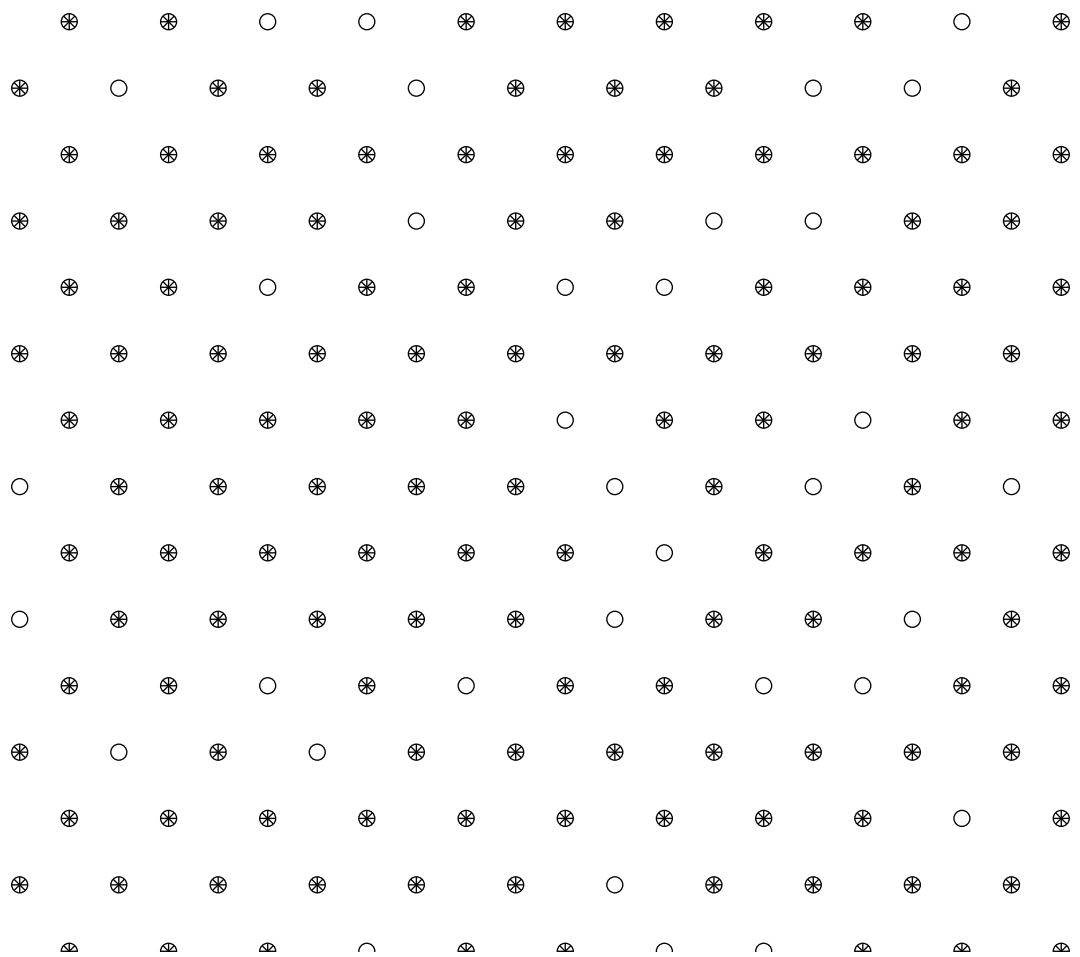

Figure 1.3: p=.8

For the case of the hexagonal lattice in the plane, it turns out that $p_c = \frac{1}{2}$. We won't prove that here, but arrive at the value $\frac{1}{2}$ as the solution to a problem which seems to be related to the critical probability problem in many cases. The idea of the *renormalization group* method is that many systems exhibit a similar behavior at different scales, a property known as *self similarity*. Understanding the transformation properties of this self similarity yields important information about the system. This is the goal of the renormalization group method. Rather than attempt a general definition, we use the hexagonal lattice as a first and elementary illustration. Replace the original hexagonal lattice by a coarser hexagonal lattice as follows: pick three adjacent vertices on the original hexagonal lattice which form an equilateral triangle. This then organizes the lattice into a union of disjoint equilateral triangles, all pointing in the same direction, where, alternately, two adjacent lattice points on a row form a base of a triangle and the third lattice point is a vertex of a triangle from an adjacent row . The center of these triangles form a new (coarser) hexagonal lattice, in fact one where the distance between sites has been increased by a factor of three. See the figures.

Each point on our new hexagonal lattice is associated with exactly three points on our original lattice. Now assign a probability, $p'$ to each point of our new lattice by the principle of majority rule: a new lattice point will be declared occupied if a majority of the associated points of the old lattice are occupied. Since our triangles are disjoint, these probabilities are independent. We can achieve a majority if all three sites are occupied (which occurs with probability $p^3$) or if two out of the three are occupied (which occurs with probability $p^2(1-p)$ with three choices as to which two sites are occupied). Thus

$$p' = p^3 + 3p^2(1-p). \tag{1.35}$$

This has three fixed points: $0, 1, \frac{1}{2}$. The derivative at $\frac{1}{2}$ is $\frac{3}{2} > 1$, so it is repelling. The points 0 and 1 are superattracting. So starting with any $p > \frac{1}{2}$, iteration leads rapidly towards the state where all sites are occupied, while starting with $p < \frac{1}{2}$ leads rapidly under iteration towards the totally empty state. The point $\frac{1}{2}$ is an unstable fixed point for the renormalization transformation.

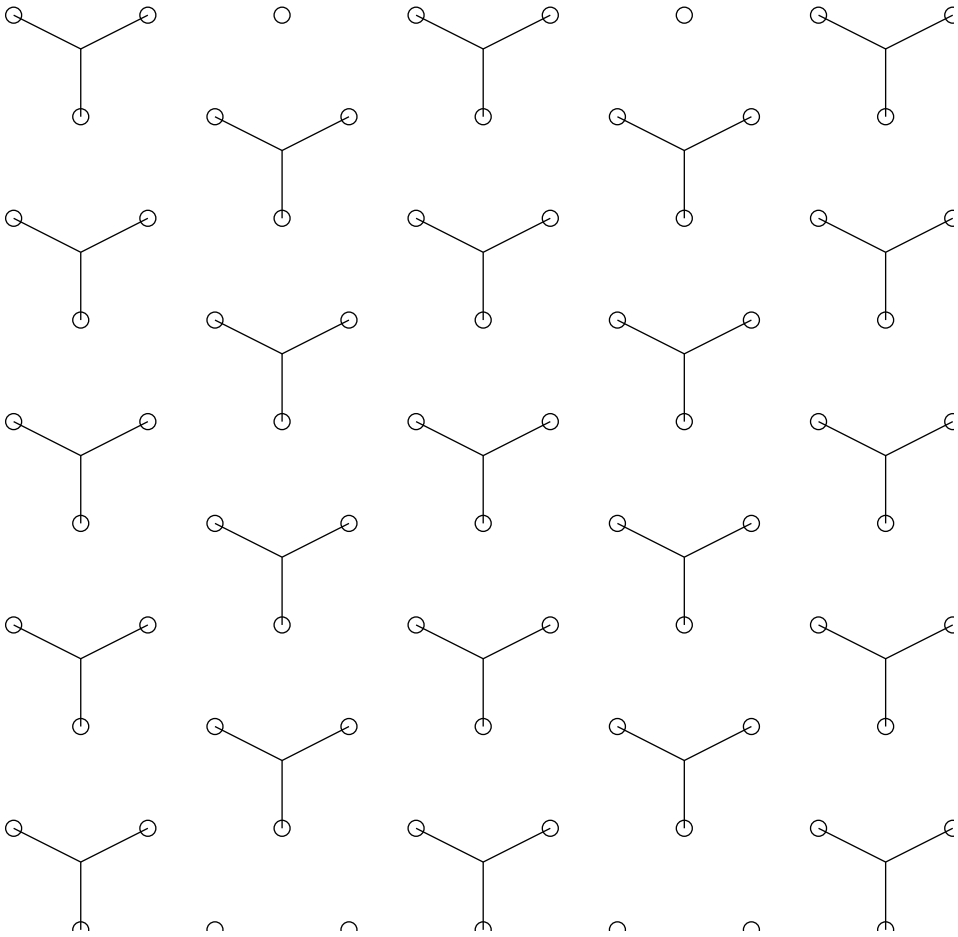

Figure 1.4: The original hexagonal lattice organized into groups of three adjacent vertices.

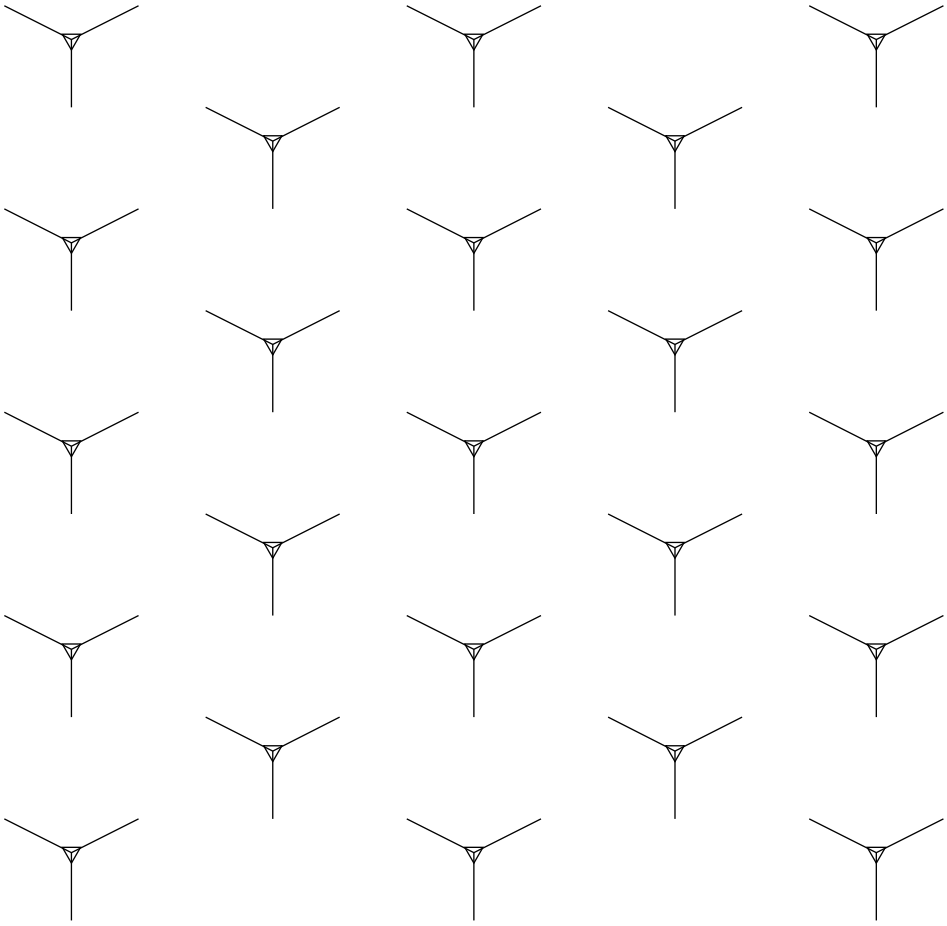

Figure 1.5: The new hexagonal lattice with edges emanating from each vertex, indicating the input for calculating $p'$ from $p$.

# Chapter 2

# Bifurcations

## 2.1 The logistic family.

In population biology one considers iteration of the "logistic function"

$$L_\mu(x) = \mu x(1-x). \tag{2.1}$$

Here $0 < \mu$ is a real parameter. The fixed points of $L_\mu$ are 0 and $1 - \frac{1}{\mu}$. Since $L'_\mu(x) = \mu - 2\mu x$,

$$L'_\mu(0) = \mu, \quad L'_\mu(1 - \frac{1}{\mu}) = 2 - \mu.$$

As $x$ represents a proportion of a population, we are mainly interested only in $0 \le x \le 1$. The maximum of $L_\mu$ is always achieved at $x = \frac{1}{2}$, and the maximum value is $\frac{\mu}{4}$. So for $0 < \mu \le 4$, $L_\mu$ maps $[0,1]$ into itself.

For $\mu > 4$, portions of $[0,1]$ are mapped into the range $x > 1$. A second operation of $L_\mu$ maps these points to the range $x < 0$ and then are swept off to $-\infty$ under successive applications of $L_\mu$.

We now examine the behavior of $L_\mu$ more closely for varying ranges of $\mu$.

### 2.1.1 $0 < \mu \le 1$

For $0 < \mu < 1$, 0 is the only fixed point of $L_\mu$ on $[0,1]$ since the other fixed point is negative. On this range of $\mu$, the point 0 is an attracting fixed point since $0 < L'_\mu(0) < 1$. Under iteration, all points of $[0,1]$ tend to 0 under the iteration. The population "dies out".

For $\mu = 1$ we have

$$L_1(x) = x(1-x) < x, \quad \forall x > 0.$$

Each successive application of $L_1$ to an $x \in (0,1]$ decreases its value. The limit of the successive iterates can not be positive since 0 is the only fixed point. So all points in $(0,1]$ tend to 0 under iteration, but ever so slowly, since $L'_1(0) = 1$.

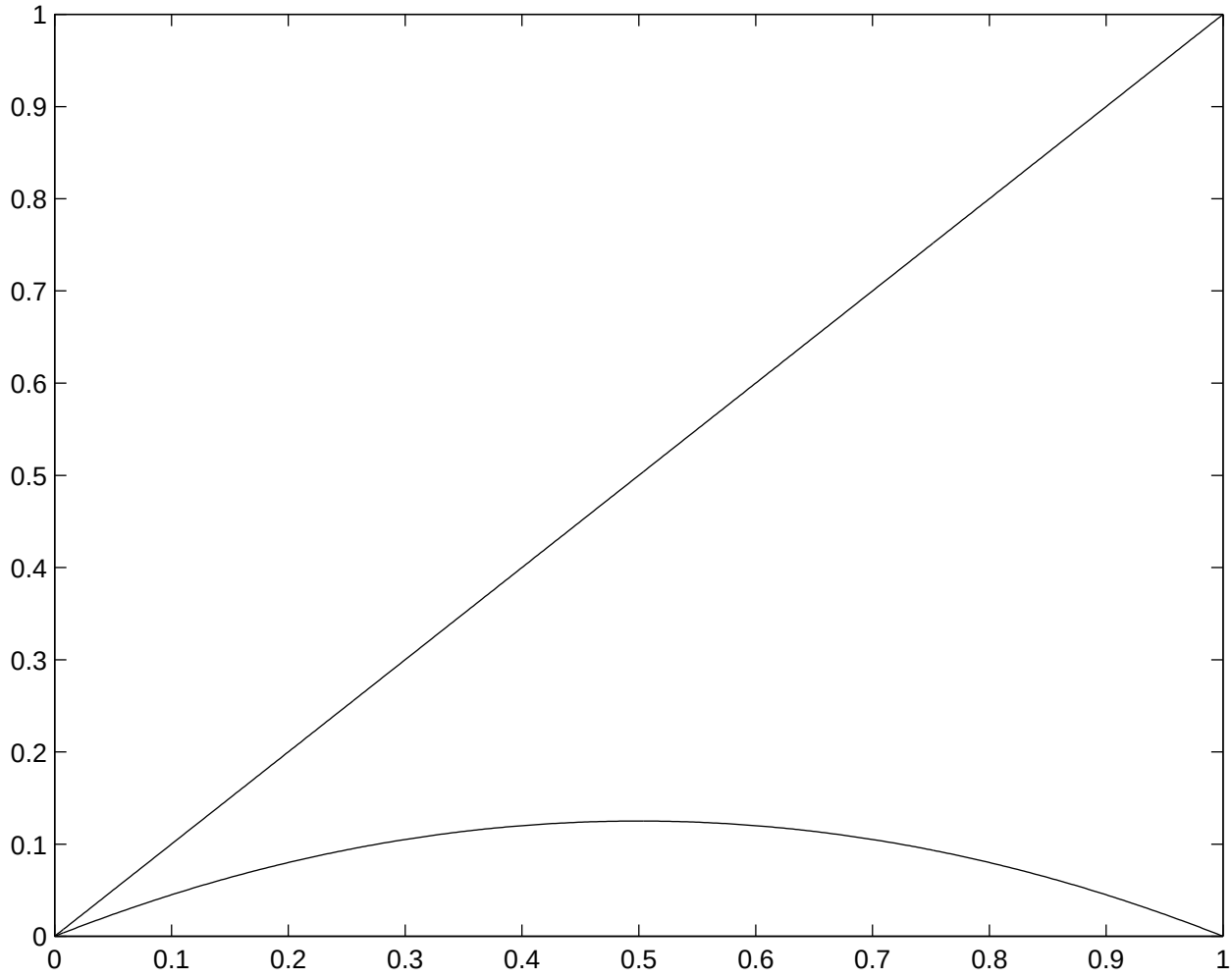


Figure 2.1: $\mu = .5$

In fact, for $x < 0$, the iterates drift off to more negative values and then tend to $-\infty$.

For all $\mu > 1$, the fixed point, 0, is repelling, and the unique other fixed point, $1 - \frac{1}{\mu}$, lies in $[0, 1]$. For $1 < \mu < 3$ we have

$$|L'_\mu(1 - \frac{1}{\mu})| = |2 - \mu| < 1,$$

so the non-zero fixed point is attractive.

We will see that the basin of attraction of $1 - \frac{1}{\mu}$ is the entire open interval $(0, 1)$, but the behavior is slightly different for the two domains, $1 < \mu \leq 2$ and $2 < \mu < 3$:

In the first of these ranges there is a steady approach toward the fixed point from one side or the other; in the second, the iterates bounce back and forth from one side to the other as they converge in towards the fixed point. The graphical iteration spirals in. Here are the details:

### 2.1.2 $1 < \mu \leq 2$.

For $1 < \mu < 2$ the non-zero fixed point lies between 0 and $\frac{1}{2}$ and the derivative at this fixed point is $2 - \mu$ and so lies between 1 and 0.

Suppose that $x$ lies between 0 and $1 - \frac{1}{\mu}$. For this range of $x$ we have

$$\frac{1}{\mu} < 1 - x$$

so, multiplying by $\mu x$ we get

$$x < \mu x(1 - x) = L_\mu(x).$$

Thus the iterates steadily increase toward $1 - \frac{1}{\mu}$, eventually converging geometrically with a rate close to $2 - \mu$. If

$$1 - \frac{1}{\mu} < x$$

then

$$L_\mu(x) < x.$$

If, in addition,

$$x \leq \frac{1}{\mu}$$

then

$$L_\mu(x) \geq 1 - \frac{1}{\mu}.$$

To see this observe that the function $L_\mu$ has only one critical point, and that is a maximum. Since $L_\mu(1 - \frac{1}{\mu}) = L_\mu(\frac{1}{\mu}) = 1 - \frac{1}{\mu}$, we conclude that the minimum value is achieved at the end points of the interval $[1 - \frac{1}{\mu}, \frac{1}{\mu}]$.

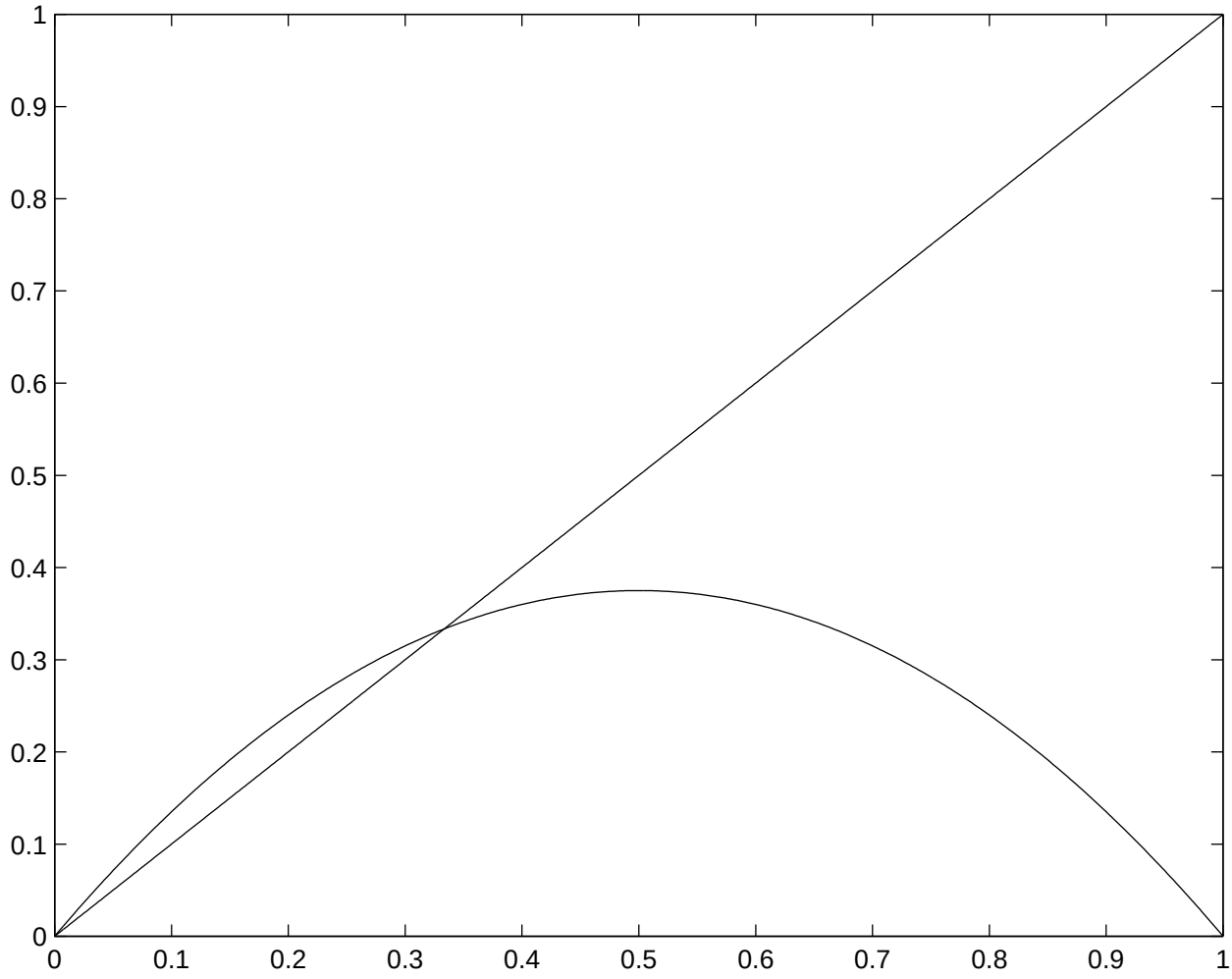


Figure 2.2: $\mu = 1.5$

Finally, for

$$\frac{1}{\mu} < x \leq 1, \quad L_\mu(x) < 1 - \frac{1}{\mu}.$$

So on the range $1 < \mu < 2$ the behavior of $L_\mu$ is as follows: All points $0 < x < 1 - \frac{1}{\mu}$ steadily increase toward the fixed point, $1 - \frac{1}{\mu}$. All points satisfying $1 - \frac{1}{\mu} < x < \frac{1}{\mu}$ steadily decrease toward the fixed point. The point $\frac{1}{\mu}$ satisfies $L_\mu(\frac{1}{\mu}) = 1 - \frac{1}{\mu}$ and so lands on the non-zero fixed point after one application. The points satisfying $\frac{1}{\mu} < x < 1$ get mapped by $L_\mu$ into the interval $0 < x < 1 - \frac{1}{\mu}$, In other words, they overshoot the mark, but then steadily increase towards the non-zero fixed point. Of course $L_\mu(1) = 0$ which is always true.

When $\mu = 2$, the points $\frac{1}{\mu}$ and $1 - \frac{1}{\mu}$ coincide and equal $\frac{1}{2}$ with $L_2'(\frac{1}{2}) = 0$. There is no "steadily decreasing" region, and the fixed point, $\frac{1}{2}$ is superattractive - the iterates zoom into the fixed point faster than any geometrical rate.

### 2.1.3 $2 < \mu < 3$.

Here the fixed point $1 - \frac{1}{\mu} > \frac{1}{2}$ while $\frac{1}{\mu} < \frac{1}{2}$. The derivative at this fixed point is negative:

$$L_\mu'(1 - \frac{1}{\mu}) = 2 - \mu < 0.$$

So the fixed point $1 - \frac{1}{\mu}$ is an attractor, but as the iterates converge to the fixed points, they oscillate about it, alternating from one side to the other. The entire interval $(0, 1)$ is in the basin of attraction of the fixed point. To see this, we may argue as follows:

The graph of $L_\mu$ lies entirely above the line $y = x$ on the interval $(0, 1 - \frac{1}{\mu}]$. In particular, it lies above the line $y = x$ on the subinterval $[\frac{1}{\mu}, 1 - \frac{1}{\mu}]$ and takes its maximum at $\frac{1}{2}$. So $\frac{\mu}{4} = L_\mu(\frac{1}{2}) > L_\mu(1 - \frac{1}{\mu}) = 1 - \frac{1}{\mu}$. Hence $L_\mu$ maps the interval $[\frac{1}{\mu}, 1 - \frac{1}{\mu}]$ onto the interval $[1 - \frac{1}{\mu}, \frac{\mu}{4}]$. The map $L_\mu$ is decreasing to the right of $\frac{1}{2}$, so it is certainly decreasing to the right of $1 - \frac{1}{\mu}$. Hence it maps the interval $[1 - \frac{1}{\mu}, \frac{\mu}{4}]$ into an interval whose right hand end point is $1 - \frac{1}{\mu}$ and whose left hand end point is $L_\mu(\frac{\mu}{4})$. We claim that

$$L_\mu(\frac{\mu}{4}) > \frac{1}{2}.$$

This amounts to showing that

$$\frac{\mu^2(4 - \mu)}{16} > \frac{1}{2}$$

or that

$$\mu^2(4 - \mu) > 8.$$

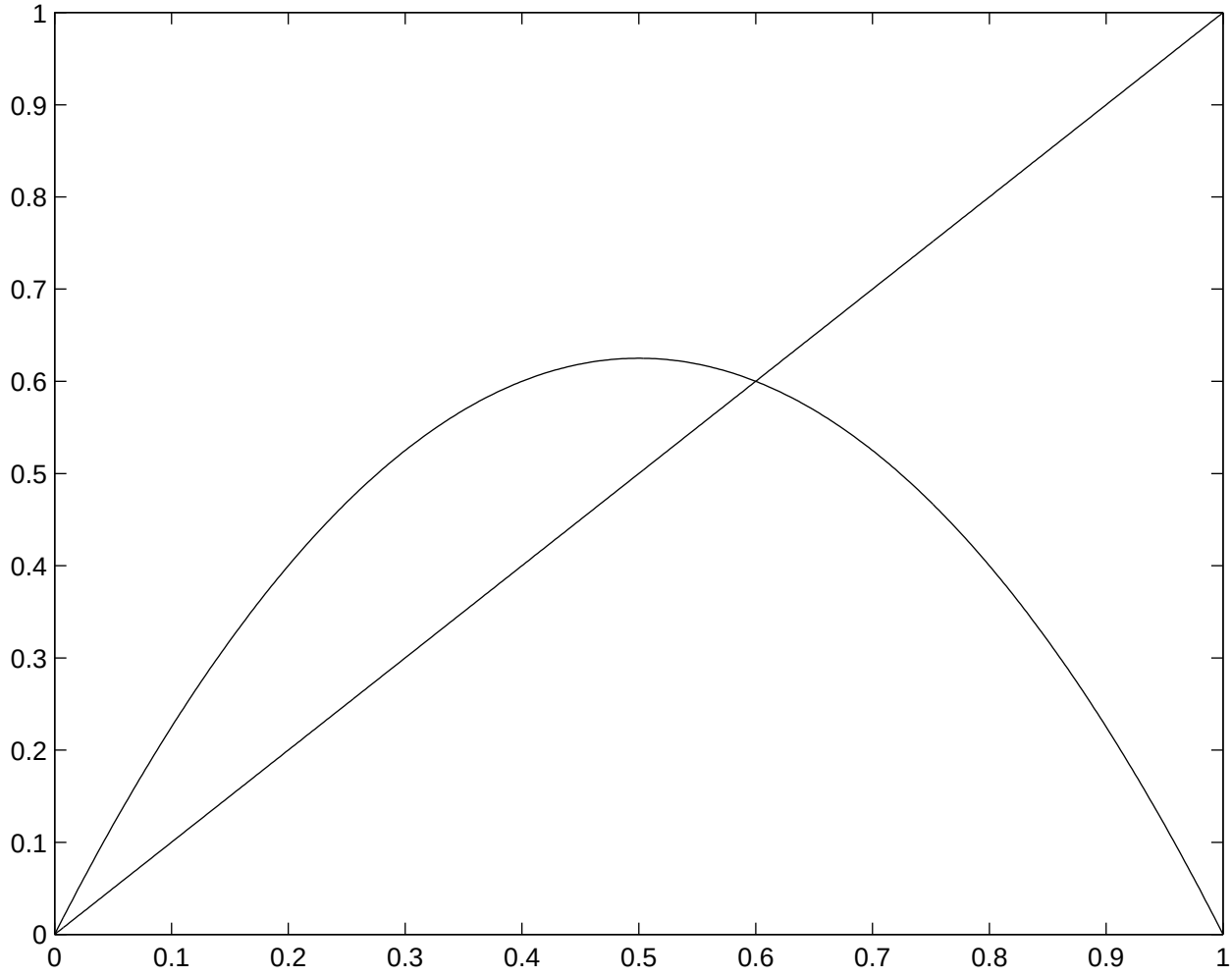


Figure 2.3: $\mu = 2.5,\ \frac{1}{\mu} = .4,\ 1 - \frac{1}{\mu} = .6$

Now the critical points of $\mu^2(4-\mu)$ are 0 and $\frac{8}{3}$ and the second derivative at $\frac{8}{3}$ is negative, so it is a local maximum. So we need only check the values of $\mu^2(4-\mu)$ at the end points, 2 and 3, of the range of $\mu$ we are considering, where the values are 8 and 9.

The image of $[\frac{1}{\mu}, 1-\frac{1}{\mu}]$ is the same as the image of $[\frac{1}{2}, 1-\frac{1}{\mu}]$ and is $[1-\frac{1}{\mu}, \frac{\mu}{4}]$. The image of this interval is the interval $[L_\mu(\frac{\mu}{4}), 1-\frac{1}{\mu}]$, with $\frac{1}{2} < L_\mu(\frac{\mu}{4})$. If we apply $L_\mu$ to this interval, we get an interval to the right of $1-\frac{1}{\mu}$ with right end point $L_\mu^2(\frac{\mu}{4}) < L_\mu(\frac{1}{2}) = \frac{\mu}{4}$. The image of the interval $[1-\frac{1}{\mu}, L_\mu^2(\frac{\mu}{4})]$ must be strictly contained in the image of the interval $[1-\frac{1}{\mu}, \frac{\mu}{4}]$, and hence we conclude that

$$L_\mu^3(\frac{\mu}{4}) > L_\mu(\frac{\mu}{4}).$$

Continuing in this way we see that under even powers, the image of $[\frac{1}{2}, 1-\frac{1}{\mu}]$ is a sequence of nested intervals whose right hand end point is $1-\frac{1}{\mu}$ and whose left hand end points are

$$\frac{1}{2} < L_\mu(\frac{\mu}{4}) < L_\mu^3(\frac{\mu}{4}) < \cdots .$$

We claim that this sequence of points converges to the fixed point, $1-\frac{1}{\mu}$. If not, it would have to converge to a fixed point of $L_\mu^2$ different from 0 and $1-\frac{1}{\mu}$. We shall show that there are no such points. Indeed, a fixed point of $L_\mu^2$ is a zero of

$$L_\mu^2(x) - x = \mu L_\mu(x)(1-L_\mu(x)) = \mu[\mu x(1-x)][1-\mu x(1-x)] - x.$$

We know in advance two roots of this quartic polynomial, namely the fixed points of $L_\mu$,which are 0 and $1-\frac{1}{\mu}$. So we know that the quartic polynomial factors into a quadratic polynomial times $\mu x(x-1+\frac{1}{\mu})$. A direct check shows that this quadratic polynomial is

$$-\mu^2 x^2 + (\mu^2+\mu)x - \mu - 1. \tag{2.2}$$

The $b^2-4ac$ for this quadratic function is

$$\mu^2(\mu^2-2\mu-3) = \mu^2(\mu+1)(\mu-3)$$

which is negative for $-1 < \mu < 3$ and so (2.2) has no real roots.

We thus conclude that the iterates of any point in $(\frac{1}{\mu}, \frac{\mu}{4}]$ oscillate about the fixed point, $1-\frac{1}{\mu}$ and converge in towards it, eventually with the geometric rate of convergence a bit less than $\mu-2$. The graph of $L_\mu$ is strictly above the line $y=x$ on the interval $(0, \frac{1}{\mu}]$ and hence the iterates of $L_\mu$ are strictly increasing so long as they remain in this interval. Furthermore they can't stay there, for this would imply the existence of a fixed point in the interval and we know that there is none. Thus they eventually get mapped into the interval $[\frac{1}{\mu}, 1-\frac{1}{\mu}]$ and the oscillatory convergence takes over. Finally, since $L_\mu$ is decreasing on

$[1-\frac{1}{\mu},1]$, any point in $[1-\frac{1}{\mu},1)$ is mapped into $(0,1-\frac{1}{\mu}]$ and so converges to the non-zero fixed point.

In short, every point in $(0,1)$ is in the basin of attraction of the non-zero fixed point and (except for the points $\frac{1}{\mu}$ and the fixed point itself) eventually converge toward it in a "spiral" fashion.

### 2.1.4 $\mu = 3$.

Much of the analysis of the preceding case applies here. The differences are: the quadratic equation (2.2) now has a (double) root. But this root is $\frac{2}{3} = 1-\frac{1}{\mu}$. So there is still no point of period two other than the fixed points. The iterates continue to spiral in, but now ever so slowly since $L'_\mu(\frac{2}{3}) = -1$.

For $\mu > 3$ we have

$$L'_\mu(1-\frac{1}{\mu}) = 2-\mu < -1$$

so both fixed points, 0 and $1-\frac{1}{\mu}$ are repelling. But now (2.2) has two real roots which are

$$p_{2\pm} = \frac{1}{2}+\frac{1}{2\mu} \pm \frac{1}{2\mu}\sqrt{(\mu+1)(\mu-3)}.$$

Both roots lie in $(0,1)$ and give a period two cycle for $L_\mu$. The derivative of $L_\mu^2$ at these periodic points is given by

$$\begin{aligned}(L_\mu^2)'(p_{2\pm}) &= L'_\mu(p_{2+})L'_\mu(p_{2-}) \\ &= (\mu-2\mu p_{2+})(\mu-2\mu p_{2-}) \\ &= \mu^2-2\mu^2(p_{2+}+p_{2-})+4\mu^2 p_{2+}p_{2-} \\ &= \mu^2-2\mu^2(1+\frac{1}{\mu})+4\mu^2\times\frac{1}{\mu^2}(\mu+1) \\ &= -\mu^2+2\mu+4.\end{aligned}$$

This last expression equals 1 when $\mu = 3$ as we already know. It decreases as $\mu$ increases reaching the value $-1$ when $\mu = 1+\sqrt{6}$.

### 2.1.5 $3 < \mu < 1+\sqrt{6}$.

In this range the fixed points are repelling and both period two points are attracting. There will be points whose images end up, after a finite number of iterations, on the non-zero fixed point. All other points in $(0,1)$ are attracted to the period two cycle. We omit the proof.

Notice also that there is a unique value of $\mu$ in this range where

$$p_{2+}(\mu) = \frac{1}{2}.$$

Indeed, looking at the formula for $p_{2+}$ we see that this amounts to the condition that $\sqrt{(\mu+1)(\mu-3)} = 1$ or

$$\mu^2-2\mu-4 = 0.$$

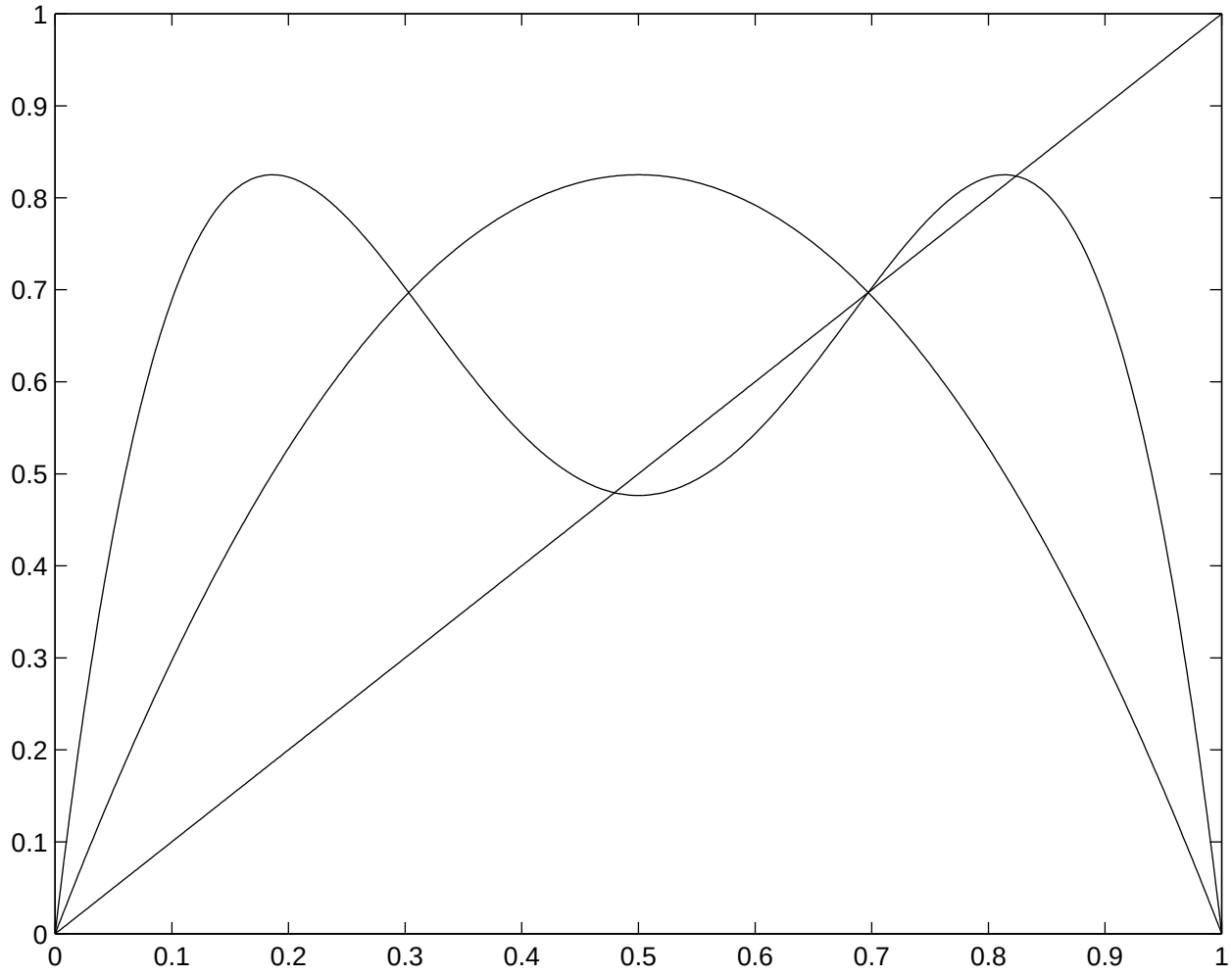


Figure 2.4: $\mu = 3.3$, graphs of $y = x, y = L_\mu(x),\ y = L_\mu^2(x)$.

The positive solution to this equation is given by $\mu = s_2$ where

$$s_2 = 1 + \sqrt{5}.$$

At $s_2$, the period two points are superattracting, since one of them coincides with $\frac{1}{2}$ which is the maximum of $L_{s_2}$.

### 2.1.6 $3.449499... < \mu < 3.569946....$

Once $\mu$ passes $1 + \sqrt{6} = 3.449499...$ the points of period two become unstable and (stable) points of period four appear. Initially these are stable, but as $\mu$ increases they become unstable (at the value $\mu = 3.544090...$) and bifurcate into period eight points, initially stable.

### 2.1.7 Reprise.

The total scenario so far, as $\mu$ increases from 0 to about 3.55, is as follows: For $\mu < b_1 := 1$, there is no non-zero fixed point. Past the first bifurcation point, $b_1 = 1$, the non-zero fixed point has appeared close to zero. When $\mu$ reaches the first superattractive value , $s_1 := 2$, the fixed point is at .5 and is superattractive. As $\mu$ increases, the fixed point continues to move to the right. Just after the second bifurcation point, $b_2 := 3$, the fixed point has become unstable and two stable points of period two appear, one to the right and one to the left of .5. The leftmost period two point moves to the right as we increase $\mu$, and at $\mu = s_2 := 1 + \sqrt{5} = 3.23606797...$ the point .5 is a period two point, and so the period two points are superattractive. When $\mu$ passes the second bifurcation value $b_2 = 1 + \sqrt{6} = 3.449..$ the period two points have become repelling and attracting period four points appear.

In fact, this scenario continues. The period $2^{n-1}$ points appear at bifurcation values $b_n$. They are initially attracting, and become superattracting at $s_n > b_n$ and become unstable past the next bifurcation value $b_{n+1} > s_n$ when the period $2^n$ points appear. The (numerically computed) bifurcation points and superstable points are tabulated as

| $n$ | $b_n$ | $s_n$ |
|---|---|---|
| 1 | 1.000000 | 2.000000 |
| 2 | 3.000000 | 3.236068 |
| 3 | 3.449499 | 3.498562 |
| 4 | 3.544090 | 3.554641 |
| 5 | 3.564407 | 3.566667 |
| 6 | 3.568759 | 3.569244 |
| 7 | 3.569692 | 3.569793 |
| 8 | 3.569891 | 3.569913 |
| 9 | 3.569934 | 3.569946 |
| $\infty$ | 3.569946 | 3.569946 |

The values of the $b_n$ are obtained by numerical experiment. We shall describe a method for computing the $s_n$ using Newton's method. We should point out

that this is still just the beginning of the story. For example, an attractive period three cycle appears at about 3.83. We shall come back to all of these points, but first discuss theoretical problems associated to bifurcations.

## 2.2 Local bifurcations.

We will be studying the iteration (in $x$) of a function, $F$, of two real variables $x$ and $\mu$ . We will need to make various hypothesis concerning the differentiability of $F$. We will always assume usually it is at least $C^2$ (has continuous partial derivatives up to the second order). We may also need $C^3$ in which case we explicitly state this hypothesis. We write

$$F_\mu(x) = F(x, \mu)$$

and are interested in the change of behavior of $F_\mu$ as $\mu$ varies.

Before embarking on the study of bifurcations let us observe that if $p$ is a fixed point of $F_\mu$ and $F'_\mu(p) \neq 1$, then for $\nu$ close to $\mu$, the transformation $F_\nu$ has a unique fixed point close to $p$. Indeed, the implicit function theorem applies to the function

$$P(x, \nu) := F(x, \nu) - x$$

since

$$\frac{\partial P}{\partial x}(p, \mu) \neq 0$$

by hypothesis. We conclude that there is a curve of fixed points $x(\nu)$ with $x(\mu) = p$.

### 2.2.1 The fold.

The first type of bifurcation we study is the *fold bifurcation* where there is no (local) fixed point on one side of the bifurcation value, $b$, where a fixed point $p$ appears at $\mu = b$ with $F'_\mu(p) = 1$, and at the other side of $b$ the map $F_\mu$ has two fixed points, one attracting and the other repelling.

As an example consider the quadratic family

$$Q(x, \mu) = Q_\mu(x) := x^2 + \mu.$$

Fixed points must be solutions of the quadratic equation

$$x^2 - x + \mu = 0,$$

whose roots are

$$p_\pm = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - 4\mu}.$$

For

$$\mu > b = \frac{1}{4}$$

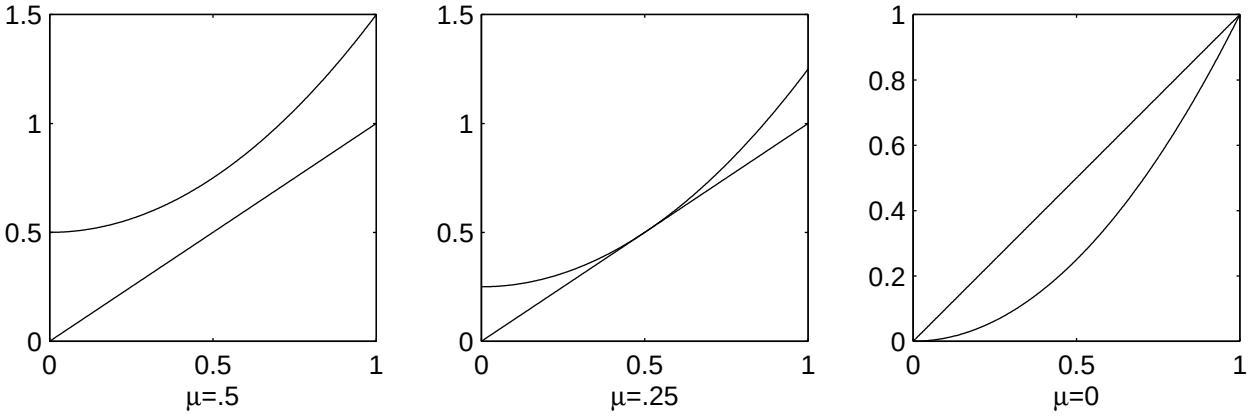


Figure 2.5: $y = x^2 + \mu$ for $\mu = .5,\ .25$ and $0$.

these roots are not real. The parabola $x^2 + \mu$ lies entirely above the line $y = x$ and there are no fixed points.

At $\mu = \frac{1}{4}$ the parabola just touches the line $y = x$ at the point $(\frac{1}{2}, \frac{1}{2})$ and so

$$p = \frac{1}{2}$$

is a fixed point, with $Q'_\mu(p) = 2p = 1$.

For $\mu < \frac{1}{4}$ the points $p_\pm$ are fixed points, with $Q'_\mu(p_+) > 1$ so it is repelling, and $Q'_\mu(p_-) < 1$. We will have $Q'_\mu(p_-) > -1$ so long as $\mu > -\frac{3}{4}$, so on the range $-\frac{3}{4} < \mu < \frac{1}{4}$ we have two fixed points, one repelling and one attracting.

We will now discuss the general phenomenon. In order not to clutter up the notation, we assume that coordinates have been chosen so that $b = 0$ and $p = 0$. So we make the standing assumption that $p = 0$ is a fixed point at $\mu = 0$, i.e. that

$$F(0, 0) = 0.$$

**Proposition 2.2.1 (Fold bifurcation).** *Suppose that at the point* $(0, 0)$ *we have*

$$\text{(a)}\ \ \frac{\partial F}{\partial x}(0,0) = 1, \qquad \text{(b)}\ \ \frac{\partial^2 F}{\partial x^2}(0,0) > 0, \qquad \text{(c)}\ \ \frac{\partial F}{\partial \mu}(0,0) > 0.$$

*Then there are non-empty intervals* $(\mu_1, 0)$ *and* $(0, \mu_2)$ *and* $\epsilon > 0$ *so that*

(i) *If* $\mu \in (\mu_1, 0)$ *then* $F_\mu$ *has two fixed points in* $(-\epsilon, \epsilon)$.
*One is attracting and the other repelling.*

(ii) *If* $\mu \in (0, \mu_2)$ *then* $F_\mu$ *has no fixed points in* $(-\epsilon, \epsilon)$.

**Proof.** All the proofs in this section will be applications of the implicit function theorem. For our current proposition, set

$$P(x, \mu) := F(x, \mu) - x.$$

Then by our standing hypothesis we have

$$P(0, 0) = 0$$

and condition (c) gives

$$\frac{\partial P}{\partial \mu}(0, 0) > 0.$$

The implicit function theorem gives a unique function $\mu(x)$ with $\mu(0) = 0$ and

$$P(x, \mu(x)) \equiv 0.$$

The formula for the derivative in the implicit function theorem gives

$$\mu'(x) = -\frac{\partial P/\partial x}{\partial P/\partial \mu}$$

which vanishes at the origin by assumption (a). We then may compute the second derivative, $\mu''$, via the chain rule; using the fact that $\mu'(0) = 0$ we obtain

$$\mu''(0) = -\frac{\partial^2 P/\partial x^2}{\partial P/\partial \mu}(0, 0).$$

This is negative by assumptions (b) and (c). In other words,

$$\mu'(0) = 0, \text{ and } \mu''(0) < 0$$

so $\mu(x)$ has a maximum at $x = 0$, and this maximum value is 0. In the $(x, \mu)$ plane, the graph of $\mu(x)$ looks locally like a parabola pointing in the lower half plane with its apex at the origin. If we rotate this picture clockwise by ninety degrees, this says that there are no points on this curve sitting over positive $\mu$ values, i.e. no fixed points for positive $\mu$, and two fixed points for $\mu < 0$.

Now consider the function $\frac{\partial F}{\partial x}(x, \mu(x))$. The derivative of this function with respect to $x$ is

$$\frac{\partial^2 F}{\partial x^2}(x, \mu(x)) + \frac{\partial^2 F}{\partial x \partial \mu}(x, \mu(x))\mu'(x).$$

By assumption (b) and $\mu'(0) = 0$, this expression is positive at $x = 0$ and so $\frac{\partial F}{\partial x}(x, \mu(x))$ is an increasing function in a neighborhood of the origin while $\frac{\partial F}{\partial x}(0, 0) = 1$. But this says that

$$F'_\mu(x) < 1$$

on the lower fixed point and

$$F'_\mu(x) > 1$$

at the upper fixed point, completing the proof of the proposition. We should point out that changing the sign in (b) or (c) interchanges the role of the two intervals.

### 2.2.2 Period doubling.

We now turn to the *period doubling bifurcation*. This is what happens when we pass through a bifurcation value with

$$\frac{\partial F}{\partial x}(0,0) = -1. \tag{2.3}$$

We saw examples in the preceding section.

To visualize the phenomenon we plot the function $L_\mu^{\circ 2}$ for the values $\mu = 2.9$ and $\mu = 3.3$ in Figure 2.6. For $\mu = 2.9$ the curve crosses the diagonal at a single point, which is in fact a fixed point of $\mathrm{L}_\mu$ and hence of $L_\mu^{\circ 2}$. This fixed point is stable. For $\mu = 3.3$ there are three crossings. The fixed point of $L_\mu$ has derivative smaller than $-1$, and hence the corresponding fixed point of $L_\mu^{\circ 2}$ has derivative greater than one. The two other crossings correspond to the stable period two orbit.

We now turn to the general theory: Notice that the partial derivative of $F(x,\mu) - x$ with respect to $x$ is $-2$ at the origin. In particular it does not vanish, so we can now solve for $x$ as a function of $\mu$; there is a unique branch of fixed points, $x(\mu)$, passing through the origin. Let $\lambda(\mu)$ denote the derivative of $F_\mu$ with respect to $x$ at the fixed point, $x(\mu)$, i.e. define

$$\lambda(\mu) := \frac{\partial F}{\partial x}(x(\mu),\mu).$$

As notation, let us set

$$F_\mu^{\circ 2} := F_\mu \circ F_\mu$$

and define

$$F^{\circ 2}(x,\mu) := F_\mu^{\circ 2}(x).$$

Notice that

$$(F_\mu^{\circ 2})'(x) = F_\mu'(F_\mu(x))F_\mu'(x)$$

by the chain rule so

$$(F_0^{\circ 2})'(0) = (F_0'(0))^2 = 1.$$

Hence

$$(F_\mu^{\circ 2})''(x) = F_\mu''(F_\mu(x))F_\mu'(x)^2 + F_\mu'(F_\mu(x))F_\mu''(x) \tag{2.4}$$

which vanishes at $x = 0, \mu = 0$. In other words,

$$\frac{\partial^2 F^{\circ 2}}{\partial x^2}(0,0) = 0. \tag{2.5}$$

Let us absorb the import of this equation. One might think that if we set $G_\mu = F_\mu^{\circ 2}$, then $G_\mu'(0) = 1$, so all we need to do is apply Proposition 1 to $G_\mu$. But (2.5) shows that the key condition (b) of Proposition 1 is violated, and hence we must make some alternative hypotheses. The hypotheses that we will make will involve the second and the third partial derivatives of $F$, and also that $\lambda(\mu)$ really passes through $-1$, i.e. $\frac{d\lambda}{d\mu}(0) \neq 0$.

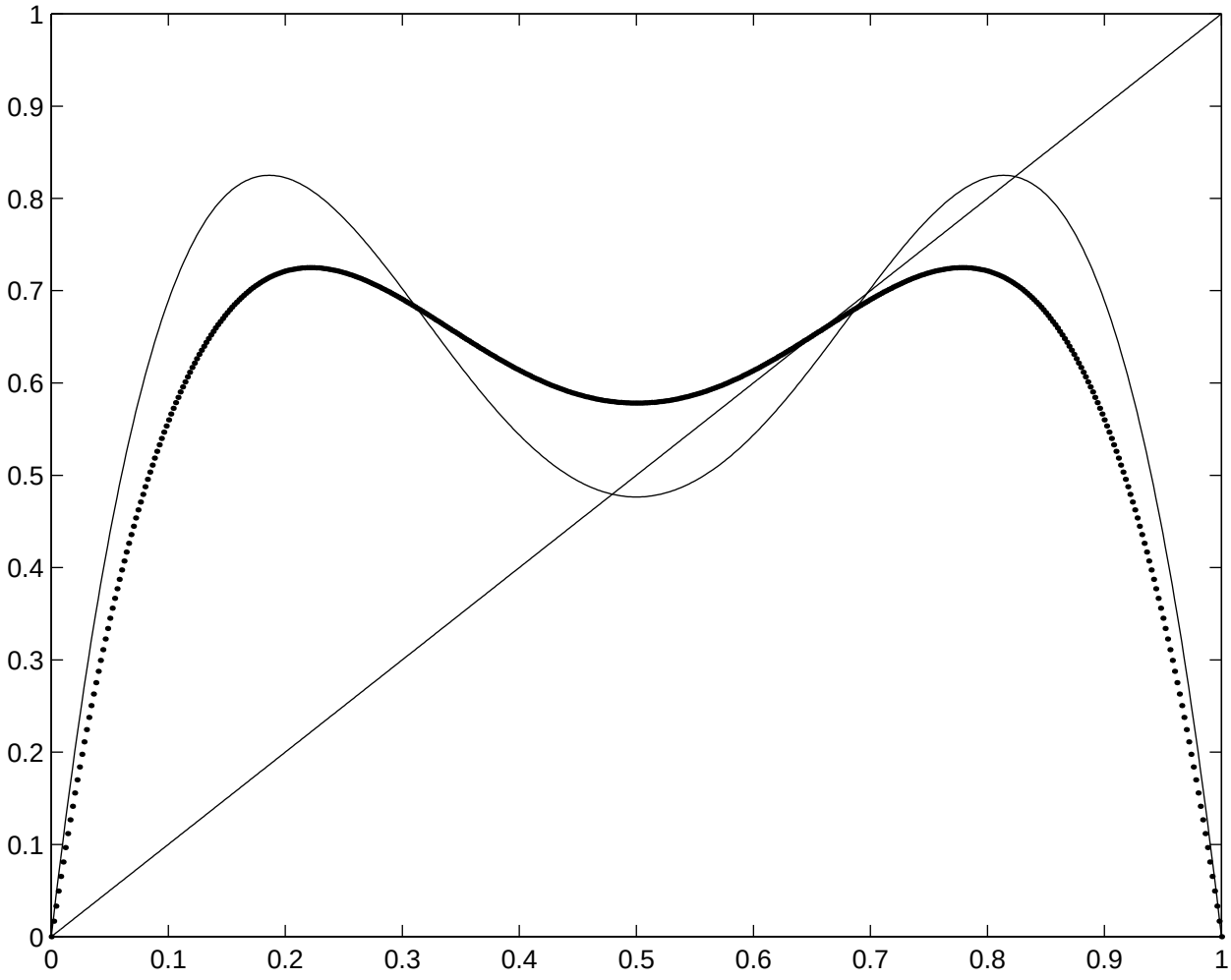


Figure 2.6: Plots of $L_\mu^{\circ 2}$ for $\mu = 2.9$ (dotted curve) and $\mu = 3.3$.

To understand the hypothesis about involving the partial derivatives of $F$, let us differentiate (2.4) once more with respect to $x$ to obtain

$$(F_\mu^{\circ 2})'''(x) =$$

$$F'''_\mu(F_\mu(x))F'_\mu(x)^3+2F''_\mu(F_\mu(x))F''_\mu(x)F'_\mu(x)+F''_\mu(F_\mu(x))F'_\mu(x)F''_\mu(x)+F'_\mu(F_\mu(x))F'''_\mu(x).$$

At $(x,\mu)=(0,0)$ this simplifies to

$$-\left[2\frac{\partial^3 F}{\partial x^3}(0,0)+3\frac{\partial^2 F}{\partial x^2}(0,0)\right]. \tag{2.6}$$

**Proposition 2.2.2 (Period doubling bifurcation).** *Suppose that $F$ is $C^3$, that*

$$\text{(d)}F'_0(0)=-1 \quad \text{(e)} \quad \frac{d\lambda}{d\mu}(0)>0, \quad \textit{and} \quad \text{(f)} \quad 2\frac{\partial^3 F}{\partial x^3}(0,0)+3\frac{\partial^2 F}{\partial x^2}(0,0)>0.$$

*Then there are non-empty intervals $(\mu_1,0)$ and $(0,\mu_2)$ and $\epsilon>0$ so that*

(i) *If $\mu\in(\mu_1,0)$ then $F_\mu$ has one repelling fixed point and one attracting orbit of period two in $(-\epsilon,\epsilon)$*

(ii) *If $\mu\in(0,\mu_2)$ then $F_\mu^{\circ 2}$ has a single fixed point in $(-\epsilon,\epsilon)$ which is in fact an attracting fixed point of $F_\mu$.*

The statement of the theorem is summarized in Figure 2.7:

**Proof.** Let

$$H(x,\mu):=F^{\circ 2}(x,\mu)-x.$$

Then by the remarks before the proposition, $H$ vanishes at the origin together with its first two partial derivatives with respect to $x$. The expression (2.6) (which used condition (d)) together with conditions (f) gives

$$\frac{\partial^3 H}{\partial x^3}(0,0)<0.$$

One of the zeros of $H$ corresponds to the fixed point, let us factor this out: Define $P(x,\mu)$ by

$$H(x,\mu)=(x-x(\mu))P(x,\mu). \tag{2.7}$$

Then

$$\begin{aligned}\frac{\partial H}{\partial x} &= P+(x-\mu)\frac{\partial P}{\partial x}\\ \frac{\partial^2 H}{\partial x^2} &= 2\frac{\partial P}{\partial x}+(x-x(\mu))\frac{\partial^2 P}{\partial x^2}\\ \frac{\partial^3 H}{\partial x^3} &= 3\frac{\partial^2 P}{\partial x^2}+(x-x(\mu))\frac{\partial^3 H}{\partial x^3}.\end{aligned}$$

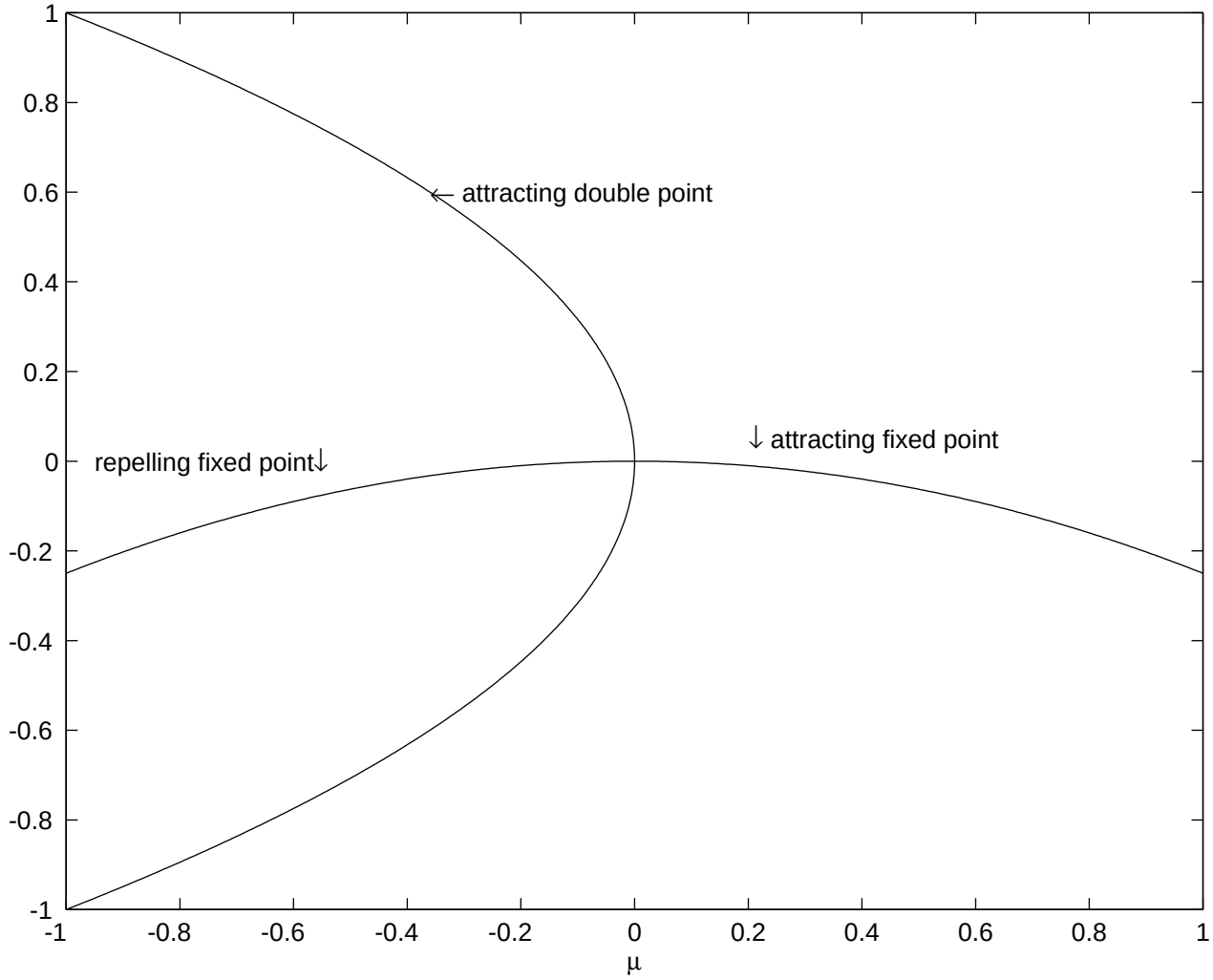


Figure 2.7: Period doubling bifurcation.

So $P$ vanishes at the origin together with its first partial derivative with respect to $x$, while

$$\frac{\partial^3 H}{\partial x^3}(0,0) = 3\frac{\partial^2 P}{\partial x^2}(0,0)$$

so

$$\frac{\partial^2 P}{\partial x^2}(0,0) < 0. \tag{2.8}$$

We claim that

$$\frac{\partial P}{\partial \mu}(0,0) < 0, \tag{2.9}$$

so that we can apply the implicit function theorem to $P(x,\mu) = 0$ to solve for $\mu$ as a function of $x$. This will allow us to determine the fixed points of $F_\mu^{\circ 2}$ which are *not* fixed points of $F_\mu$, i.e. the points of period two. To prove (2.9) we compute $\frac{\partial H}{\partial x}$ both from its definition $H(x,\mu) = F^{\circ 2}(x,\mu) - x$ and from (2.7) to obtain

$$\begin{aligned} \frac{\partial H}{\partial x} &= \frac{\partial F}{\partial x}(F(x,\mu),\mu)\frac{\partial F}{\partial x}(x,\mu) - 1 \\ &= P(x,\mu) + (x - x(\mu))\frac{\partial P}{\partial x}(x,\mu). \end{aligned}$$

Recall that $x(\mu)$ is the fixed point of $F_\mu$ and that $\lambda(\mu) = \frac{\partial F}{\partial x}(x(\mu),\mu)$. So substituting $x = x(\mu)$ into the preceding equation gives

$$\lambda(\mu)^2 - 1 = P(x,\mu).$$

Differentiating with respect to $\mu$ and setting $\mu = 0$ gives

$$\frac{\partial P}{\partial \mu}(0,0) = 2\lambda(0)\lambda'(0) = -2\lambda'(0)$$

which is $< 0$ by (e).

By the implicit function theorem, (2.9) implies that there is a $C^2$ function $\nu(x)$ defined near zero as the unique solution of $P(x,\nu(x)) \equiv 0$. Recall that $P$ and its first derivative with respect to $x$ vanish at $(0,0)$. We now repeat the arguments of the last subsection: We have

$$\nu'(x) = -\frac{\partial P/\partial x}{\partial P/\partial \mu}$$

so

$$\nu'(0) = 0$$

and

$$\nu''(0) = -\frac{\partial^2 P/\partial x^2}{\partial P/\partial x}(0,0) < 0$$

since this time both numerator and denominator are negative. So the curve $\nu$ has the same form as in the proof of the preceding proposition. This establishes

the existence of the (strictly) period two points for $\mu < 0$ and their absence for $\mu > 0$.

We now turn to the question of the stability of the fixed points and the period two points. Condition (e) implies that $\lambda(\mu) < -1$ for $\mu < 0$ and $\lambda(\mu) > -1$ for $\mu > 0$ so the fixed point is repelling to the left and attracting to the right of the origin. As for the period two points, we wish to show that

$$\frac{\partial F^{\circ 2}}{\partial x}(x, \nu(x)) < 1$$

for $x < 0$. But (2.5) and $\nu'(0) = 0$ imply that imply that 0 is a critical point for this function, and the value at this critical point is $\lambda(0)^2 = 1$. To complete the proof we must show that this critical point is a local maximum. So we must compute the second derivative at the origin. Calling this function $\phi$ we have

$$\begin{aligned}
\phi(x) &:= \frac{\partial F^{\circ 2}}{\partial x}(x, \nu(x)) \\
\phi'(x) &= \frac{\partial^2 F^{\circ 2}}{\partial x^2}(x, \nu(x)) + \frac{\partial^2 F^{\circ 2}}{\partial x \partial \mu}(x, \nu(x))\nu'(x) \\
\phi''(x) &= \frac{\partial^3 F^{\circ 2}}{\partial x^3}(x, \nu(x)) + 2\frac{\partial^3 F^{\circ 2}}{\partial x^2 \partial \mu}(x, \nu(x)\nu'(x)) + \frac{\partial^3 F^{\circ 2}}{\partial x \partial \mu^2}(x, \nu(x))(\nu'(x))^2 + \\
&\qquad \frac{\partial^2 F^{\circ 2}}{\partial x \partial \mu}(x, \nu(x))\nu''(x).
\end{aligned}$$

The middle two terms vanish at 0 since $\nu'(0) = 0$. The last term becomes

$$\frac{d\lambda}{d\mu}(0)\nu''(0) < 0$$

by condition (e) and the fact that $\nu''(0) < 0$. In computing the third partial derivative with respect to $x$ of $F^{\circ 2}$ by the chain rule and by Leibniz's rule, all terms involving the second partial derivative vanish at $(0, 0)$ by (2.5) and we are left with

$$2\frac{\partial^3 F}{\partial x^3}(0,0)\left(\frac{\partial F}{\partial x}(0,0)\right)^3$$

which is negative by assumptions (d) and (f). This completes the proof of the proposition.

There are obvious variants on the theorem which involve changing signs in hypotheses (e) and or (f). Thus we may have an attractive fixed point merging with two repelling points of period two to produce a repelling fixed point, and/or the direction of the bifurcation may be reversed.

## 2.3 Newton's method and Feigenbaum's constant

Although the values of $b_n$ for the logistic family are hard to compute except by numerical experiment, the superattractive values can be found by applying

Newton's method to find the solution, $s_n$, of the equation

$$L_\mu^{\circ 2^{n-1}}(\frac{1}{2}) = \frac{1}{2}, \quad L_\mu(x) = \mu x(1-x). \tag{2.10}$$

This is the equation for $\mu$ which says that $\frac{1}{2}$ is a point of period $2^{n-1}$ of $L_\mu$. Of course we want to look for solutions for which $\frac{1}{2}$ does not have lower period.

So we set

$$P(\mu) = L_\mu^{2^{n-1}}(\frac{1}{2}) - \frac{1}{2}$$

and apply the Newton algorithm

$$\mu_{k+1} = \mathcal{N}(\mu_k), \quad \mathcal{N}(\mu) = \mu - \frac{P(\mu)}{P'(\mu)}.$$

with $'$ now denoting differentiation with respect to $\mu$. As a first step, must compute $P$ and $P'$. For this we define the functions $x_k(\mu)$ recursively by

$$x_0 = \frac{1}{2}, \quad x_1(\mu) = \mu\frac{1}{2}(1-\frac{1}{2}), \quad x_{k+1} = L_\mu(x_k),$$

so, we have

$$\begin{aligned} x'_{k+1} &= [\mu x_k(1-x_k))]' \\ &= x_k(1-x_k) + \mu x'_k(1-x_k) - \mu x_k x'_k \\ &= x_k(1-x_k) + \mu(1-2x_k)x'_k. \end{aligned}$$

Let

$$N = 2^{n-1}$$

so that

$$P(\mu) = x_N - \frac{1}{2}, \quad P'(\mu) = x'_N(\mu).$$

Thus, at each stage of the iteration in Newton's method we compute $P(\mu)$ and $P'(\mu)$ by running the iteration scheme

$$\begin{array}{rclcrcl} x_{k+1} &=& \mu x_k(1-x_k) & & x_0 &=& \frac{1}{2} \\ x'_{k+1} &=& x_k(1-x_k) + \mu(1-2x_k)x'_x & & x'_0 &=& 0 \end{array}$$

for $k = 0, \ldots, N-1$. We substitute this into Newton's method, get the next value of $\mu$, run the iteration to get the next value of $P(\mu)$ and $P'(\mu)$ etc.

Suppose we have found $s_1, s_2, ...., s_n$. What should we take as the initial value of $\mu$? Define the numbers $\delta_n$, $n \geq 2$ recursively by $\delta_2 = 4$ and

$$\delta_n = \frac{s_{n-1} - s_{n-2}}{s_n - s_{n-1}}, \quad n \geq 3. \tag{2.11}$$

We have already computed

$$s_1 = 2, \quad s_2 = 1 + \sqrt{5} = 3.23606797\ldots.$$

We take as our initial value in Newton's method for finding $s_{n+1}$ the value

$$\mu_{n+1} = s_n + \frac{s_n - s_{n-1}}{\delta_n}.$$

The following facts are observed:

For each $n = 3, 4, \ldots, 15$, Newton's method converges very rapidly, with no changes in the first nineteen digits after six applications of Newton's method for finding $s_3$, after only one application of Newton's method for $s_4$ and $s_5$, and at most four applications of Newton's method for the computation of each of the remaining values.

Suppose we stop our calculations for each $s_n$ when there is no further change in the first 19 digits, and take the computed values as our $s_n$. These values are strictly increasing. In particular this implies that the $s_n$ we have computed do not yield $\frac{1}{2}$ as a point of lower period.

The $s_n$ approach a limiting value, 3.569945671205296863.

The $\delta_n$ approach a limiting value,

$$\delta = 4.6692016148.$$

This value is known as *Feigenbaum's constant*. While the limiting value of the $s_n$ is particular to the logistic family, $\delta$ is "universal" in the sense that it applies to a whole class of one dimensional iteration families. We shall go into this point in the next section, where we will see that this is a renormalization group phenomenon.

## 2.4 Feigenbaum renormalization.

We have already remarked that the rate of convergence to the limiting value of the superstable points in the period doubling bifurcation, Feigenbaum's constant, is universal, i.e. not restricted to the logistic family. That is, if we let

$$\delta = 4.6692....$$

denote Feigenbaum's constant, then the superstable values $s_r$ in the period doubling scenario satisfy

$$s_r = s_\infty - B\delta^{-r} + o(\delta^{-r})$$

where $s_\infty$ and $B$ depend on the specifics of the family, but $\delta$ applies to a large class of such families.

There is another "universal" parameter in the story. Suppose that our family $f_\mu$ consists of maps with a single maximum, $X_m$, so that $X_m$ must be one of the points on any superstable periodic orbit. (In the case of the logistic family $X_m = \frac{1}{2}$.) Let $d_r$ denote the difference between $X_m$ an the next nearest point on the superstable $2^r$ orbit; more precisely, define

$$d_r = f_{s_r}^{2^{r-1}}(X_m) - X_m.$$

Then

$$d_r \sim D(-\alpha)^r$$

where

$$\alpha \doteq 2.5029...$$

is again universal. This would appear to be a scale parameter (in $x$) associated with the period doubling scenario. To understand this scale parameter, examine the central portion of Fig 2.4 and observe that the graph of $L_\mu^{\circ 2}$ looks like an (inverted and) rescaled version of $\mathrm{L}_\mu$, especially if we allow a change in the parameter $\mu$. The rescaling is centered at the maximum, so in order to avoid notational complexity, let us shift this maximum (for the logistic family) to the origin by replacing $x$ by $y = x - \frac{1}{2}$. In the new coordinates the logistic map is given by

$$y \mapsto L_\mu(y + \frac{1}{2}) - \frac{1}{2} = \mu(\frac{1}{4} - y^2) - \frac{1}{2}.$$

Let $\mathcal{R}$ denote the operator on functions given by

$$\mathcal{R}(h)(y) := -\alpha h(h(y/\alpha)). \tag{2.12}$$

In other words, $\mathcal{R}$ sends a map $h$ into its iterate $h \circ h$ followed by a rescaling. We are going to not only apply the operator $\mathcal{R}$, but also shift the parameter $\mu$ in the maps

$$h_\mu(y) = \mu(\frac{1}{2} - y^2) - \frac{1}{2}$$

from one supercritical value to the next. So for each $k = 0, 1, 2, \ldots$ we set

$$g_{k0} := h_{s_k}$$

and then define

$$\begin{aligned} g_{k,1} &= \mathcal{R}g_{k0} \\ g_{k2} &= \mathcal{R}g_{k1} \\ g_{k3} &= \mathcal{R}g_{k2} \\ \vdots &\quad \vdots \end{aligned}$$

It is observed (numerically) that for each $k$ the functions $g_{kr}$ appear to be approaching a limit, $g_k$ i.e.

$$g_{kr} \to g_k.$$

So

$$g_k(y) = \lim(-\alpha)^r g_{s_{k+r}}^{2^r}(y/(-\alpha)^r).$$

Hence

$$\mathcal{R}g_k = \lim(-\alpha)^{r+1} 2^{r+1} g_{s_{k+r}}(y/(-\alpha)^{r+1}) = g_{k-1}.$$

It is also observed that these limit functions $g_k$ themselves are approaching a limit:

$$g_k \to g.$$

Since $\mathcal{R}g_k = g_{k-1}$ we conclude that

$$\mathcal{R}g = g,$$

i.e. $g$ is a fixed point for the Feigenbaum renormalization operator $\mathcal{R}$. Notice that rescaling commutes with $\mathcal{R}$: If $S$ denotes the operator $(Sf)(y) = cf(y/c)$ then

$$\mathcal{R}(Sf)(y) = -\alpha(c(f(cf(y/(c\alpha))/c) = S(\mathcal{R})f(y).$$

So if $g$ is a fixed point, so is $Sg$. We may thus fix the scale in $g$ by requiring that

$$g(0) = 1.$$

The hope was then that there would be a unique function $g$ (within an appropriate class of functions) satisfying

$$\mathcal{R}g = g, \quad g(0) = 1,$$

or, spelling this out,

$$g(y) = -\alpha g^{\circ 2}(-y/\alpha), \quad g(0) = 1. \tag{2.13}$$

Notice that if we knew the function $g$, then setting $y = 0$ in (2.13) gives

$$1 = -\alpha g(1)$$

or

$$\alpha = -1/g(1).$$

In other words, assuming that we were able to establish all these facts and also knew the function $g$, then the universal rescaling factor $\alpha$ would be determined by $g$ itself. Feigenbaum assumed that $g$ has a power series expansion in $x^2$ took the first seven terms in this expansion and substituted in (2.13). He obtained a collection of algebraic equations which he solved and then derived $\alpha$ close to the observed "experimental" value. Indeed, if we truncate (2.13) we will get a collection of algebraic equations. But these equations are not recursive, so that at each stage of truncation modification is made in all the coefficients, and also the nature of the solutions of these equations is not transparent. So theoretically, if we could establish the existence of a unique solution to (2.13) within a given class of functions the value of $\alpha$ is determined. But the numerical evaluation of $\alpha$ is achieved by the renormalization property itself, rather than from $g(1)$ which is not known explicitly.

The other universal constant associated with the period doubling scenario, the constant $\delta$ was also conjectured by Feigenbaum to be associated to the fixed point $g$ of the renormalization operator; this time with the linearized map $J$, i.e. the derivative of the renormalization operator at its fixed point. Later on we will see that in finite dimensions, if the derivative $J$ of a non-linear transformation $R$ at a fixed point has $k$ eigenvalues $> 1$ in absolute value, and the rest $< 1$ in absolute value, then there exists a $k$-dimensional $R$ invariant surface tangent

at the fixed point to the subspace corresponding to the $k$ eigenvalues whose absolute value is $> 1$. On this invariant manifold, the map $R$ is expanding. Feigenbaum conjectured that for the operator $\mathcal{R}$ (acting on the appropriate infinite dimensional space of functions) there is a one dimensional "expanding" submanifold, and that $\delta$ is the single eigenvalue of $J$ with absolute value greater than 1.

In the course of the past twenty years, these conjectures of Feigenbaum have been verified using high powered techniques from complex analysis, thanks to the combined effort of such mathematicians as Douady, Hubbard, Sullivan, McMullen, and....

## 2.5 Period 3 implies all periods

Throughout the following $f$ will denote a continuous function on the reals whose domain of definition is assumed to include the given intervals in the various statements.

**Lemma 2.5.1** *If $I = [a, b]$ is a compact interval and $I \subset f(I)$ then $f$ has a fixed point in $I$.*

**Proof.** For some $c, d \in I$ we have $f(c) = a, f(d) = b$. So $f(c) \leq c, f(d) \geq d$. So $f(x) - x$ changes sign from $c$ to $d$ hence has a zero in between. QED.

**Lemma 2.5.2** *If $J$ and $K = [a, b]$ are compact intervals with $K \subset f(J)$ then there is a compact subinterval $L \subset J$ such that $f(L) = K$.*

**Proof.** Let $c$ be the greatest point in $J$ with $f(c) = a$. If $f(x) = b$ for some $x > c, x \in J$ let $d$ be the least. Then we may take $L = [c, d]$. If not, $f(x) = b$ for some $x < c, x \in J$. Let $c'$ be the largest. Let $d'$ be the the smallest $x$ satisfying $x > c'$ with $f(x) = a$. Notice that $d' \leq c$. We then take $L = [c', d']$. QED

**Notation.** If $I$ is a closed interval with end points $a$ and $b$ we write

$$I =< a, b >$$

when we do not want to specify which of the two end points is the larger.

**Theorem 2.5.1 Sarkovsky** *Period three implies all periods.*

Suppose that $f$ has a 3-cycle

$$a \mapsto b \mapsto c \mapsto a \mapsto \cdots .$$

Let $a$ denote the leftmost of the three, and let us assume that

$$a < b < c.$$

(Reversing left and right (i.e. changing direction on the real line) and cycling through the points makes this assumption harmless.) Let

$$I_0 = [a, b], \quad I_1 = [b, c]$$

so we have

$$f(I_0) \supset I_1, \quad f(I_1) \supset I_0 \cup I_1.$$

By Lemma 2 the fact that $f(I_1) \supset I_1$ implies that there is a compact interval $A_1 \subset I_1$ with $f(A_1) = I_1$. Since $f(A_1) = I_1 \supset A_1$ there is a compact subinterval $A_2 \subset A_1$ with $f(A_2) = A_1$. So

$$A_2 \subset A_1 \subset I, \quad f^{\circ 2}(A_2) = I_1.$$

By induction proceed to find compact intervals with

$$A_{n-2} \subset A_{n-3} \subset \cdots \subset A_2 \subset A_1 \subset I_1$$

with

$$f^{\circ(n-2)}(A_{n-2}) = I_1.$$

Since $f(I_0) \supset I_1 \supset A_{n-2}$ there is an interval $A_{n-1} \subset I_0$ with $f(A_{n-1}) = A_{n-2}$. Finally, since $f(I_1) \supset I_0$ there is a compact interval $A_n \subset I_1$ with $f(A_n) = A_{n-1}$. So we have

$$A_n \to A_{n-1} \to \cdots \to A_1 \to I_1$$

where each interval maps onto the next and $A_n \subset I_1$. By Lemma 1, $f^n$ has a fixed point, $x$, in $A_n$. But $f(x)$ lies in $I_0$ and all the higher iterates up to $n$ lie in $I_1$ so the period can not be smaller than $n$. So there is a periodic point of any period $n \geq 3$.

Since $f(I_1) \supset I_1$ there is a fixed point in $I_1$, and since $f(I_0) \supset I_1, f(I_1) \supset I_0$ there is a point of period two in $I_0$ which is not a fixed point of $f$. QED

A more refined analysis which we will omit shows that period 5 implies the existence of all periods greater than 5 and period 2 and 4 (but not period 3). In general any odd period implies the existence of periods of all higher order (and all smaller even order). It is easy to graph the third iterate of the logistic map to see that it crosses the diagonal for $\mu > 1 + \sqrt{(}8)$. In fact, one can prove that that at $\mu = 1 + \sqrt{(}8)$ the graph of $L_\mu^{\circ 3}$ just touches the diagonal and strictly crosses it for $\mu > 1 + \sqrt{(}8) = 3.8284\ldots$. Hence in this range there are periodic points of all periods.

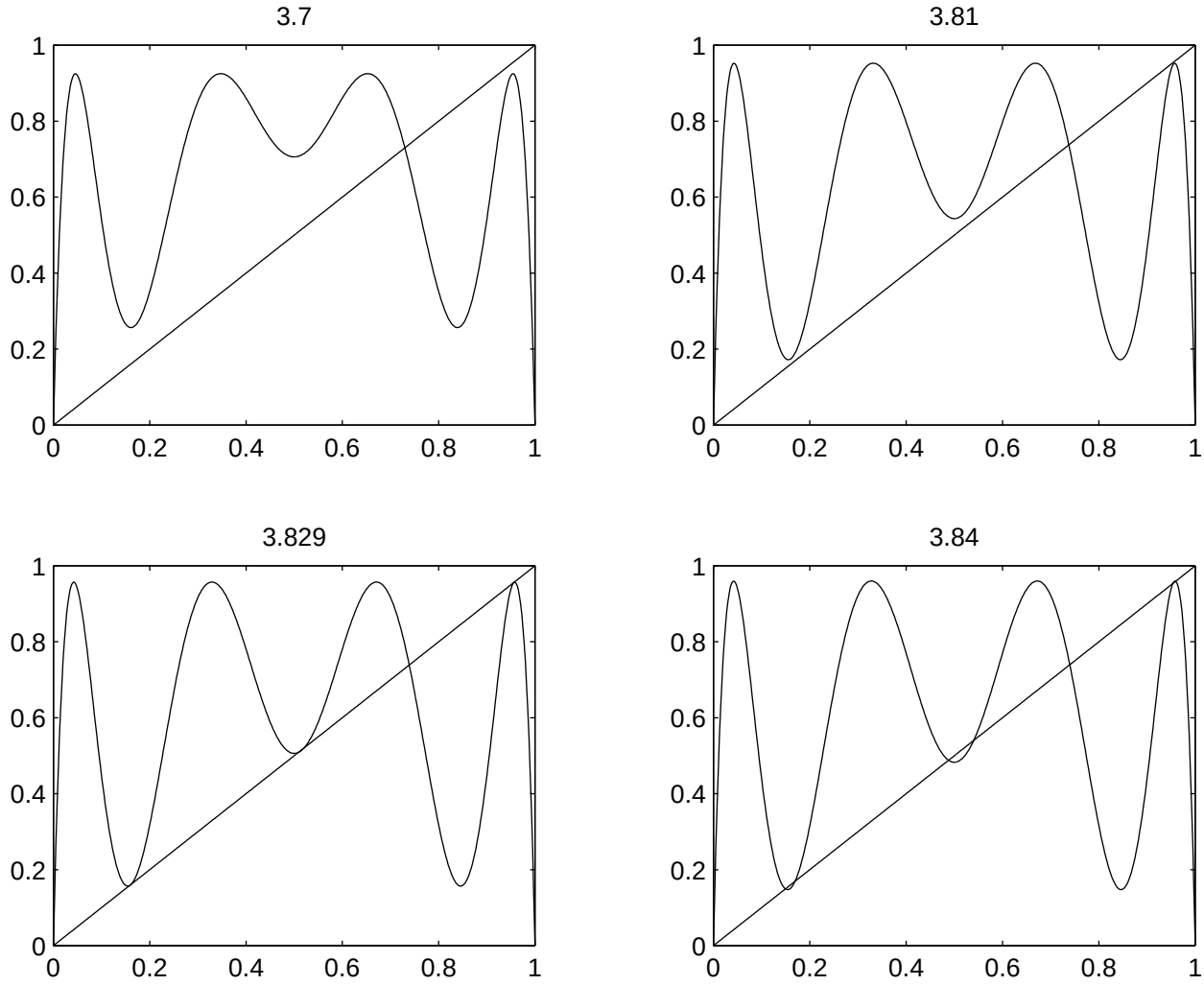


Figure 2.8: Plots of $L_\mu$ for $\mu = 3.7, 3.81, 3.83, 3.84$

## 2.6 Intermittency.

In this section we describe what happens to the period three orbits as we decrease $\mu$ from slightly above the critical value $1+\sqrt{8}$ to slightly below it. For $\mu = 1+\sqrt{8}+.002$ the roots of $P(x)-x$ where $P := L_\mu^{\circ 3}$ and the values of $P'$ at these roots are given by

| $x =$ roots of $P(x)-x$ | $P'(x)$ |
|---|---|
| 0 | 56.20068544683054 |
| 0.95756178779471 | 0.24278522730018 |
| 0.95516891475013 | 1.73457935568109 |
| 0.73893250871724 | −6.13277919589328 |
| 0.52522791460709 | 1.73457935766594 |
| 0.50342728916956 | 0.24278522531345 |
| 0.16402371217410 | 1.73457935778151 |
| 0.15565787278717 | 0.24278522521922 |

We see that there is a stable period three orbit consisting of the points

$$0.1556\ldots, .5034\ldots, .9575\ldots.$$

If we choose our initial value of $x$ close to $.5034\ldots$ and plot the successive 199 iterates of $L_\mu$ applied to $x$ we obtain the upper graph in Fig. 2.9. The lower graph gives $x(j+3)-x(j)$ for $j = 1$ to 197.

We will now decrease the parameter $\mu$ by .002 so that $\mu = 1+\sqrt{8}$ is the parameter giving the onset of period three. For this value of the parameter, the graph of $P = L_\mu^{\circ 3}$ just touches the line $y = x$ at the three double roots of $P(x)-x$ which are at $0.1599288\ldots$, $0.514355\ldots$, $0.9563180\ldots$. (Of course, the eighth degree polynomial $P(x)-x$ has two additional roots which correspond to the two (unstable) fixed points of $L_\mu$; these are not of interest to us.) Since the graph of $P$ is tangent to the diagonal at the double roots, $P'(x) = 1$ at these points, so the period three orbit is not strictly speaking stable. But using the same initial seed as above, we do get slow convergence to the period three orbit, as is indicated by Fig. 2.10.

Most interesting is what happens just before the onset of the period three cycle. Then $P(x)-x$ has only two real roots corresponding to the fixed points of $L_\mu$. The remaining six roots are complex. Nevertheless, if $\mu$ is close to $1+\sqrt{8}$ the effects of these complex roots can be felt. In Fig. 2.11 we have taken $\mu = 1+\sqrt{8}-.002$ and used the same initial seed $x = .5034$ and again plotted the successive 199 iterates of $L_\mu$ applied to $x$ in the upper graph. Notice that there are portions of this graph where the behavior is almost as if we were at a point of period three, followed by some random looking behavior, then almost period three again and so on. This is seen more clearly in the bottom graph of $x(j+3)-x(j)$. Thus the bottom graphs indicates that the deviation from period three is small on the $j$ intervals $j = [1,20]$, $[41,65]$, $[96,108]$, $[119,124]$, $[148,159]$, $[190,?]$.

This phenomenon is known as *intermittency*. We can understand how it works by iterating $P = L_\mu^{\circ 3}$. As we pass close to a minimum of $P$ lying just

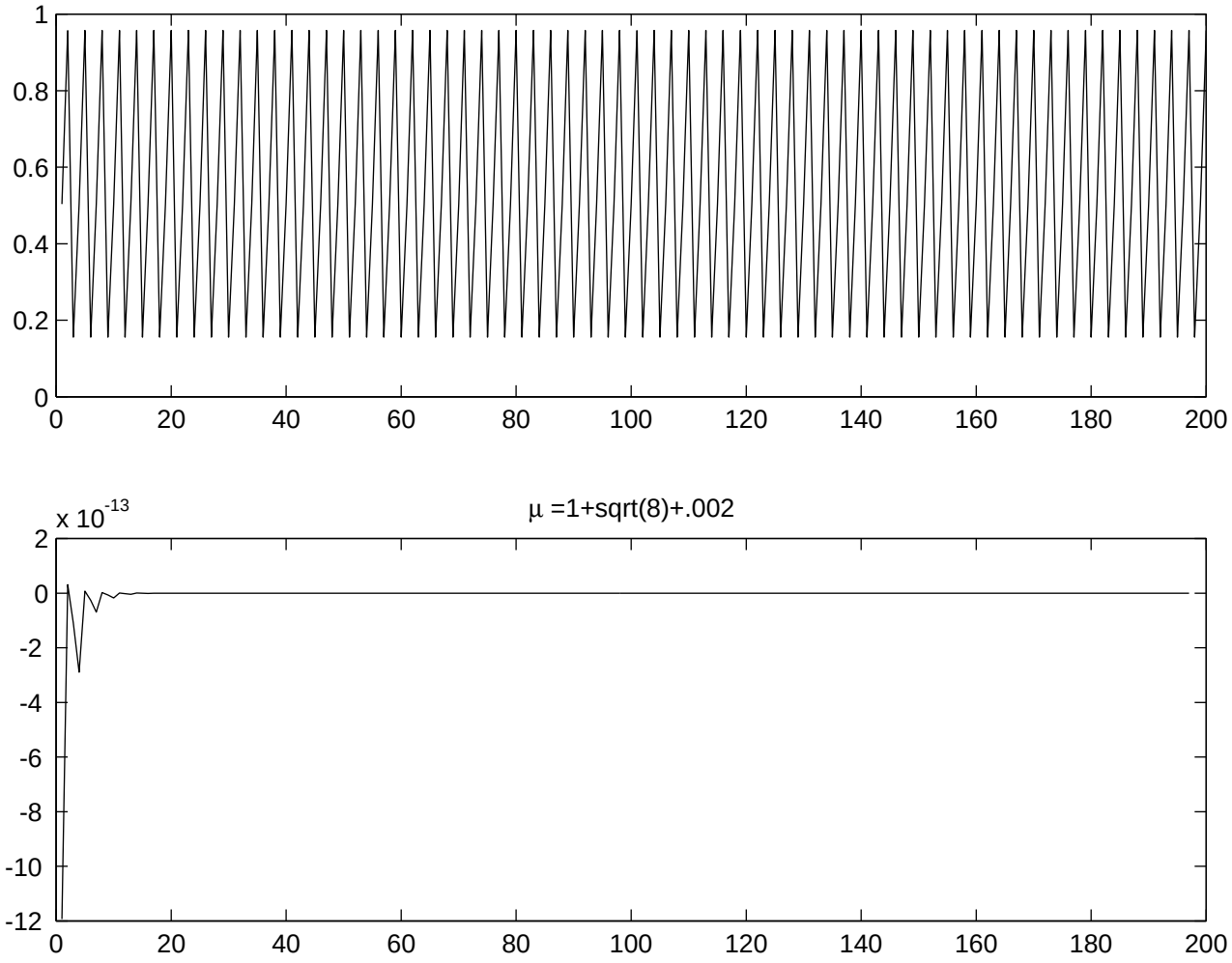


Figure 2.9: $\mu = 1 + \sqrt{8} + .002$

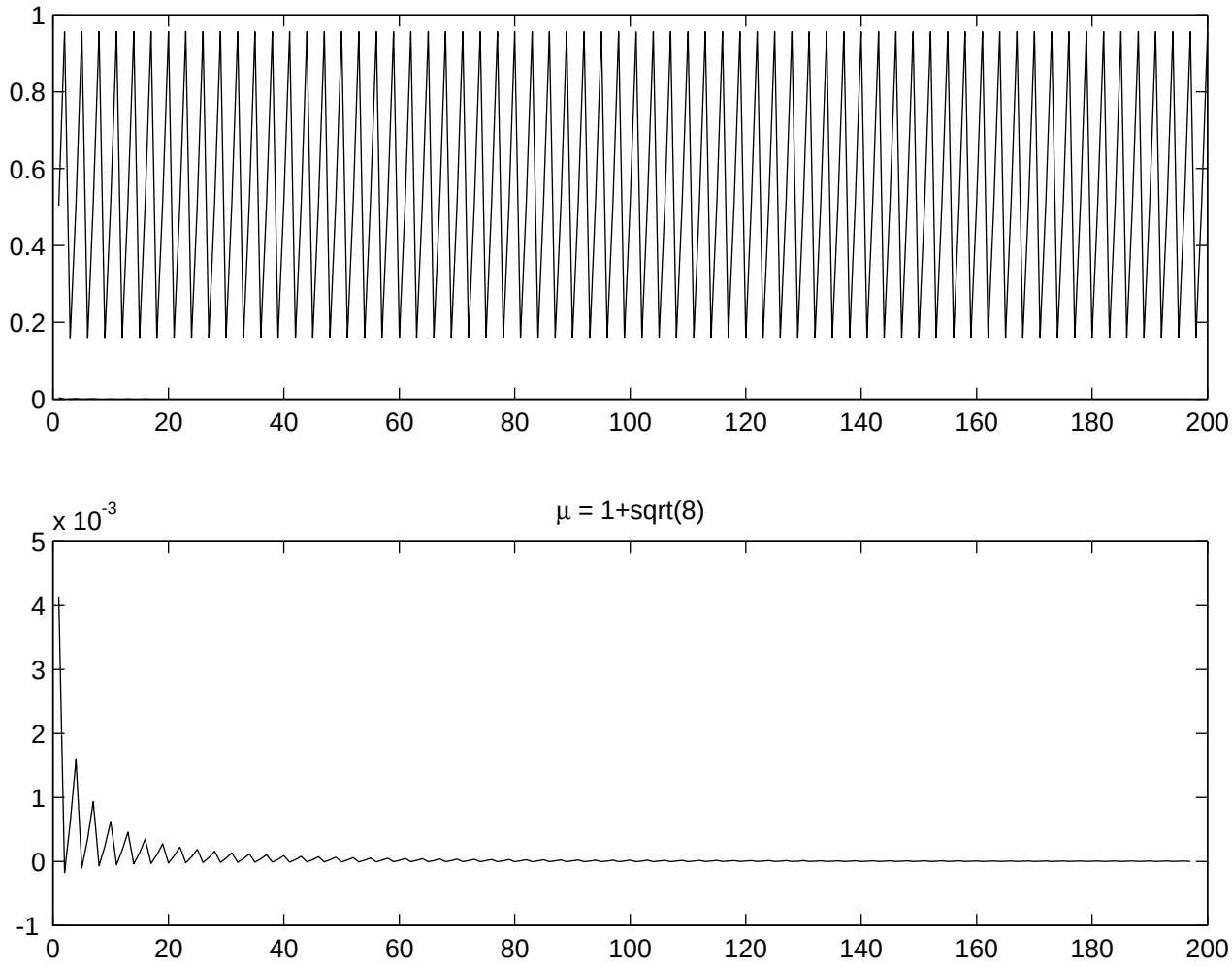


Figure 2.10: $\mu = 1 + \sqrt{(8)}$

above the diagonal, or to a maximum of $P$ lying just below the diagonal it will take many iterative steps to move away from this region - known as a *bottleneck*. Each such step corresponds to an almost period three cycle. After moving away from these bottlenecks, the steps will be large, eventually hitting a bottleneck once again. See Figures 2.12 and 2.13.

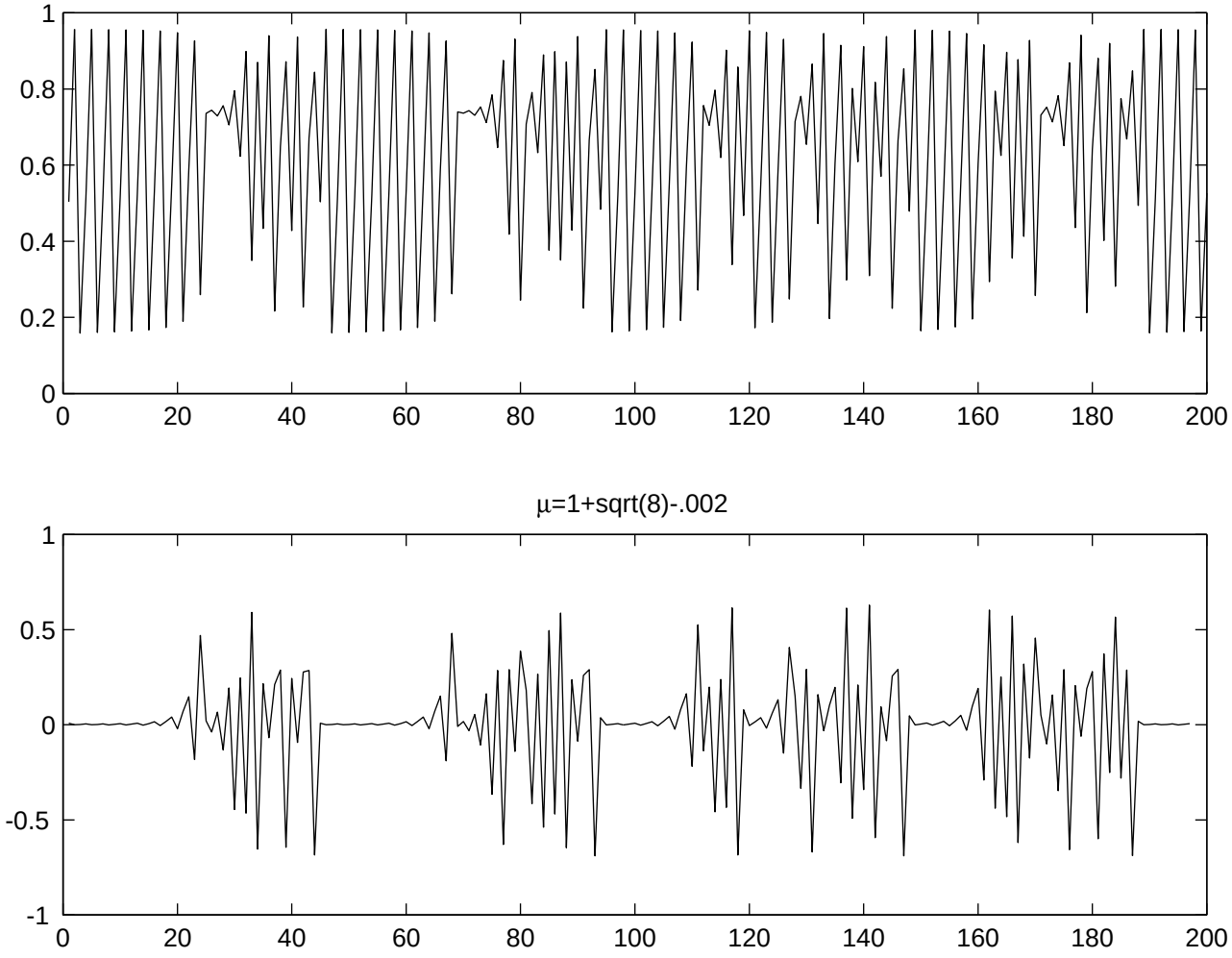


Figure 2.11: $\mu = 1 + \sqrt{(}8) - .002$

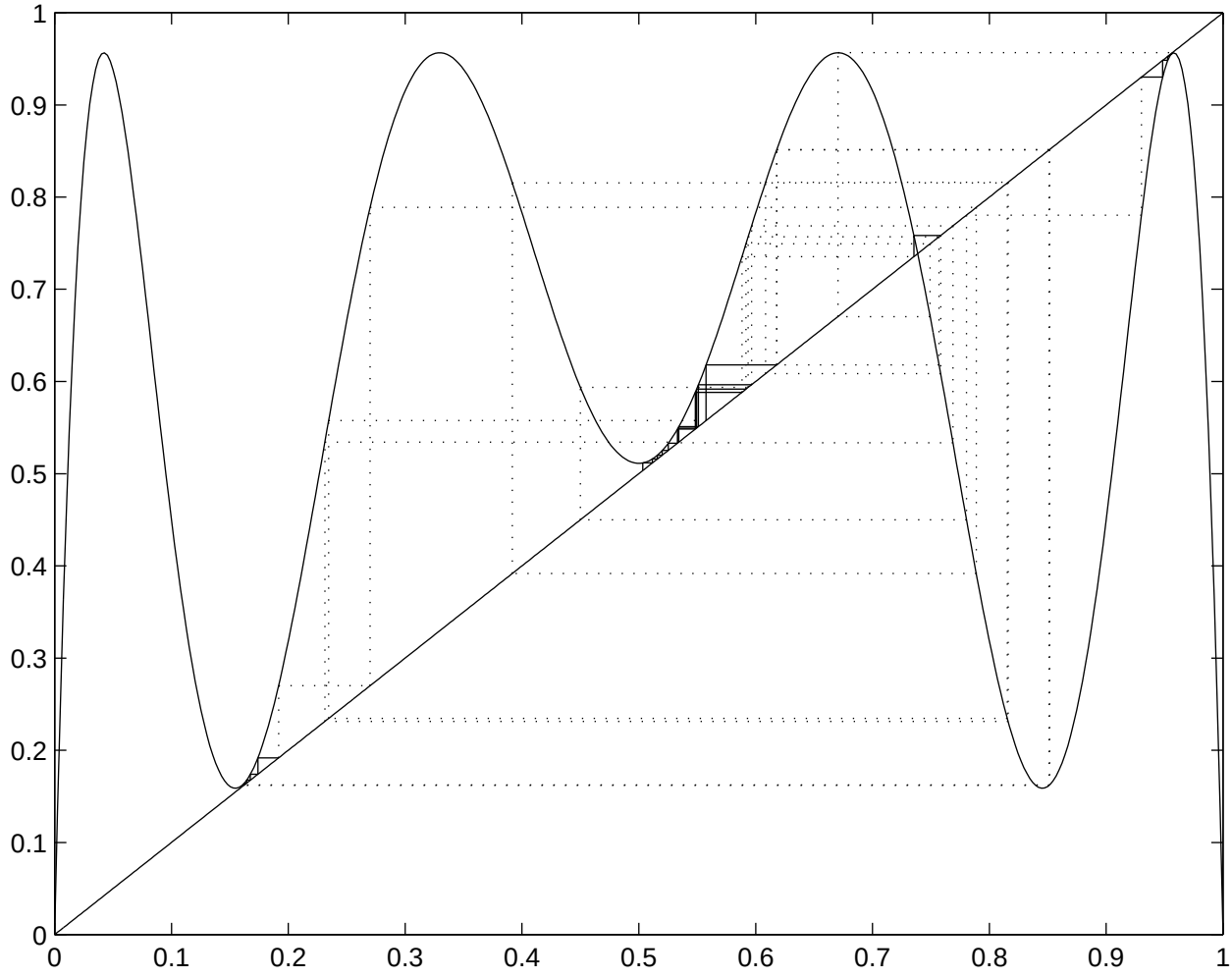


Figure 2.12: Graphical iteration of $P = L_\mu^{\circ 3}$ with $\mu = 1 + \sqrt{8} - .002$ and initial point .5034. The solid lines are iteration steps of size less than .07 representing bottleneck steps. The dotted lines are the longer steps.

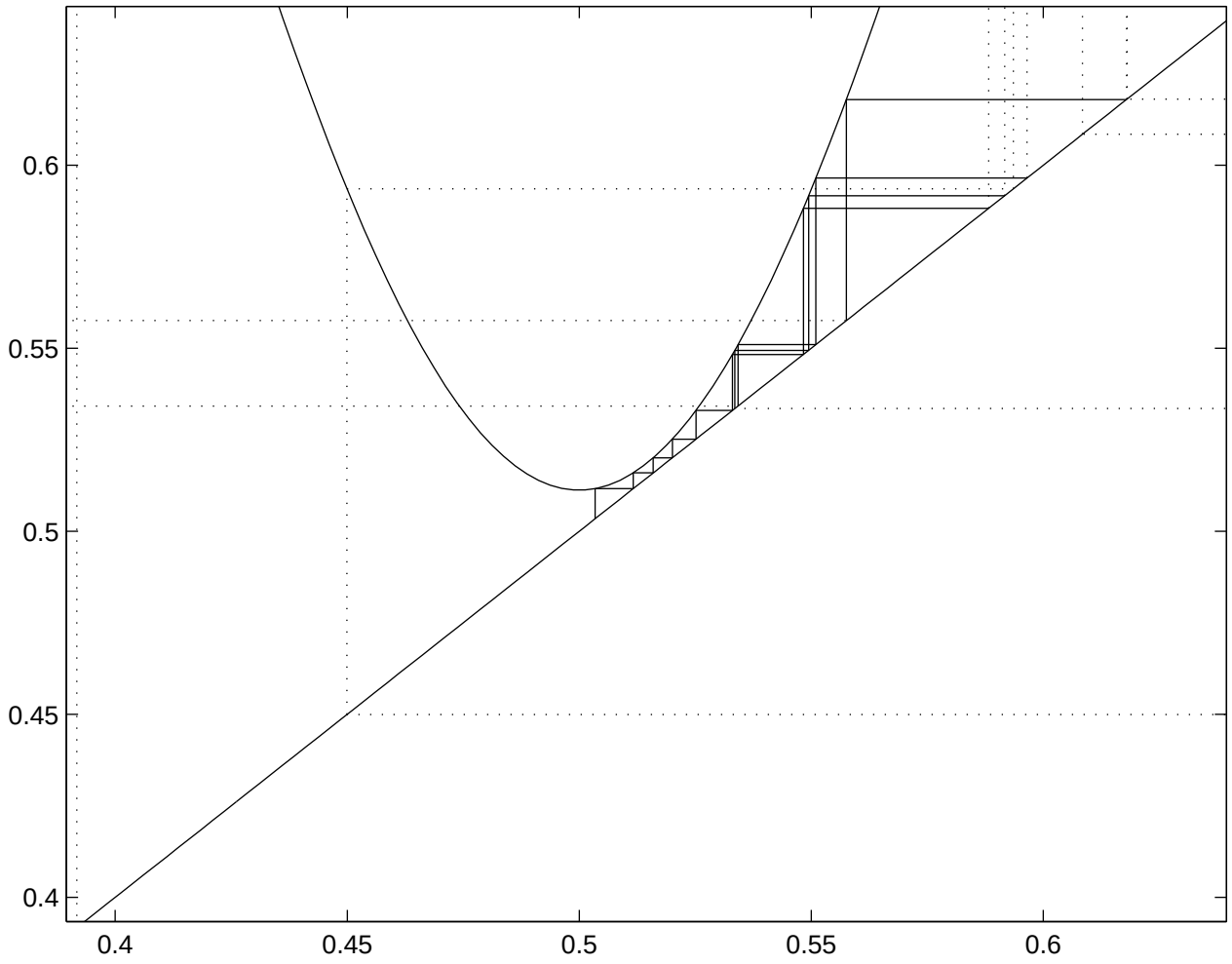


Figure 2.13: Zooming in on the central portion of the preceding figure.

# Chapter 3

# Conjugacy

## 3.1 Affine equivalence

An *affine transformation* of the real line is a transformation of the form

$$x \mapsto h(x) = Ax + B$$

where $A$ and $B$ are real constants with $A \neq 0$. So an affine transformation consists of a change of scale (and possibly direction if $A < 0$) given by the factor $A$, followed by a shift of the origin given by $B$. In the study of linear phenomena, we expect that the essentials of an object be invariant under a change of scale and a shift of the origin of our coordinate system.

For example, consider the logistic transformation, $L_\mu(x) = \mu x(1-x)$ and the affine transformation

$$h_\mu(x) = -\mu x + \frac{\mu}{2}.$$

We claim that

$$h_\mu \circ L_\mu \circ h_\mu^{-1} = Q_c \tag{3.1}$$

where

$$Q_c(x) = x^2 + c \tag{3.2}$$

and where $c$ is related to $\mu$ by the equation

$$c = -\frac{\mu^2}{4} + \frac{\mu}{2}. \tag{3.3}$$

In other words, we are claiming that if $c$ and $\mu$ are related by (3.3) then we have

$$h_\mu(L_\mu(x)) = Q_c(h_\mu(x)).$$

To check this, the left hand side expands out to be

$$-\mu[\mu x(1-x)] + \frac{\mu}{2} = \mu^2 x^2 - \mu^2 x + \frac{\mu}{2},$$

while the right hand side expands out as

$$(-\mu x + \frac{\mu}{2})^2 - \frac{\mu^2}{4} + \frac{\mu}{2} = \mu^2 x^2 - \mu^2 x + \frac{\mu}{2}$$

giving the same result as before, proving (3.1).

We say that the transformations $L_\mu$ and $Q_c, c = -\frac{\mu^2}{4} + \frac{\mu}{2}$ are *conjugate* by the affine transformation, $h_\mu$.

More generally, let $f : X \to X$ and $g : Y \to Y$ be maps of the sets $X$ and $Y$ to themselves, and let $h : X \to Y$ be a one to one map of $X$ onto $Y$. We say that $h$ conjugates $f$ into $g$ if

$$h \circ f \circ h^{-1} = g,$$

or, what amounts to the same thing, if

$$h \circ f = g \circ h.$$

We shall frequently write this equation in the form of a *commutative diagram*

$$\begin{array}{ccc} X & \xrightarrow{f} & X \\ {\scriptstyle h}\downarrow & & \downarrow{\scriptstyle h} \\ Y & \xrightarrow[g]{} & Y \end{array}$$

The statement that the diagram is commutative means that going along the upper right hand path (so applying $h \circ f$) is equal to traversing the left lower path (which is $g \circ f$).

Notice that if $h \circ f \circ h^{-1} = g$, then

$$g^{circn} = h \circ f^{\circ n} \circ h^{-1}.$$

So the problem of studying the iterates of $g$ is the same (up to the transformation $h$) as that of $f$, *providing* that the properties we are interested in studying are not destroyed by $h$.

Certainly affine transformations will always be allowed. Let us generalize the preceding computation by showing that *any* quadratic transformation (with non-vanishing leading term) is conjugate (by an affine transformation) to a transformation of the form $Q_c$ for suitable $c$. More precisely:

**Proposition 3.1.1** *Let $f = ax^2 + bx + d$ then $f$ is conjugate to $Q_c$ by the affine map $h(x) = Ax + B$ where*

$$A = a, \quad B = \frac{b}{2}, \text{ and } \quad c = ad + \frac{b}{2} - \frac{b^2}{4}.$$

**Proof.** Direct verification.

Let us understand the importance of this result. The general quadratic transformation $f$ depends on three parameters $a, b$ and $d$. But if we are interested in the qualitative behavior of the iterates of $f$, it suffices to examine the one parameter family $C_c$. Any quadratic transformation (with non-vanishing leading term) has the same behavior (in terms of its iterates) as one of the $Q_c$. The family of possible behaviors under iteration is one dimensional, depending on a single parameter $c$. We may say that the family $Q_c$ (or for that matter the family $L_\mu$) is *universal* with respect to quadratic maps as far as iteration is concerned.

## 3.2 Conjugacy of $T$ and $L_4$

Let $T : [0,1] \to [0,1]$ be the map defined by

$$T(x) = 2x,\ 0 \le x \le \frac{1}{2}, \quad T(x) = -2x + 2,\ \frac{1}{2} \le x \le 1.$$

So the graph of $T$ looks like a tent, hence its name. It consists of the straight line segment of slope 2 joining $x = 0, y = 0$ to $x = \frac{1}{2}, y = 1$ followed by the segment of slope $-2$ joining $x = \frac{1}{2}, y = 1$ to $x = 1, y = 0$.

Of course, here $L_4$ is our old friend, $L_4(x) = 4x(1-x)$. We wish to show that

$$L_4 \circ h = h \circ T$$

where

$$h(x) = \sin^2\left(\frac{\pi x}{2}\right).$$

In other words, we claim that the diagram of section 1 commutes when $f = T,\ g = L_4$ and $h$ is as above. The function $\sin\theta$ increases monotonically from 0 to 1 as $\theta$ increases from 0 to $\pi/2$. So, setting

$$\theta = \frac{\pi x}{2},$$

we see that $h(x)$ increases monotonically from 0 to 1 as $x$ increases from 0 to 1. It therefore is a one to one continuous map of $[0,1]$ onto itself, and thus has a continuous inverse. It is differentiable everywhere with $h(x) > 0$ for $0 < x < 1$. But $h'(0) = h'(1) = 0$. So $h^{-1}$ is not differentiable at the end points, but is differentiable for $0 < x < 1$.

To verify our claim, we substitute

$$\begin{aligned} L_4(h(x)) &= 4\sin^2\theta(1 - \sin^2\theta) \\ &= 4\sin^2\theta\cos^2\theta \\ &= \sin^2 2\theta \\ &= \sin^2 \pi x. \end{aligned}$$

So for $0 \le x \le \frac{1}{2}$ we have verified that

$$L_4(h(x)) = h(2x) = h(T(x)).$$

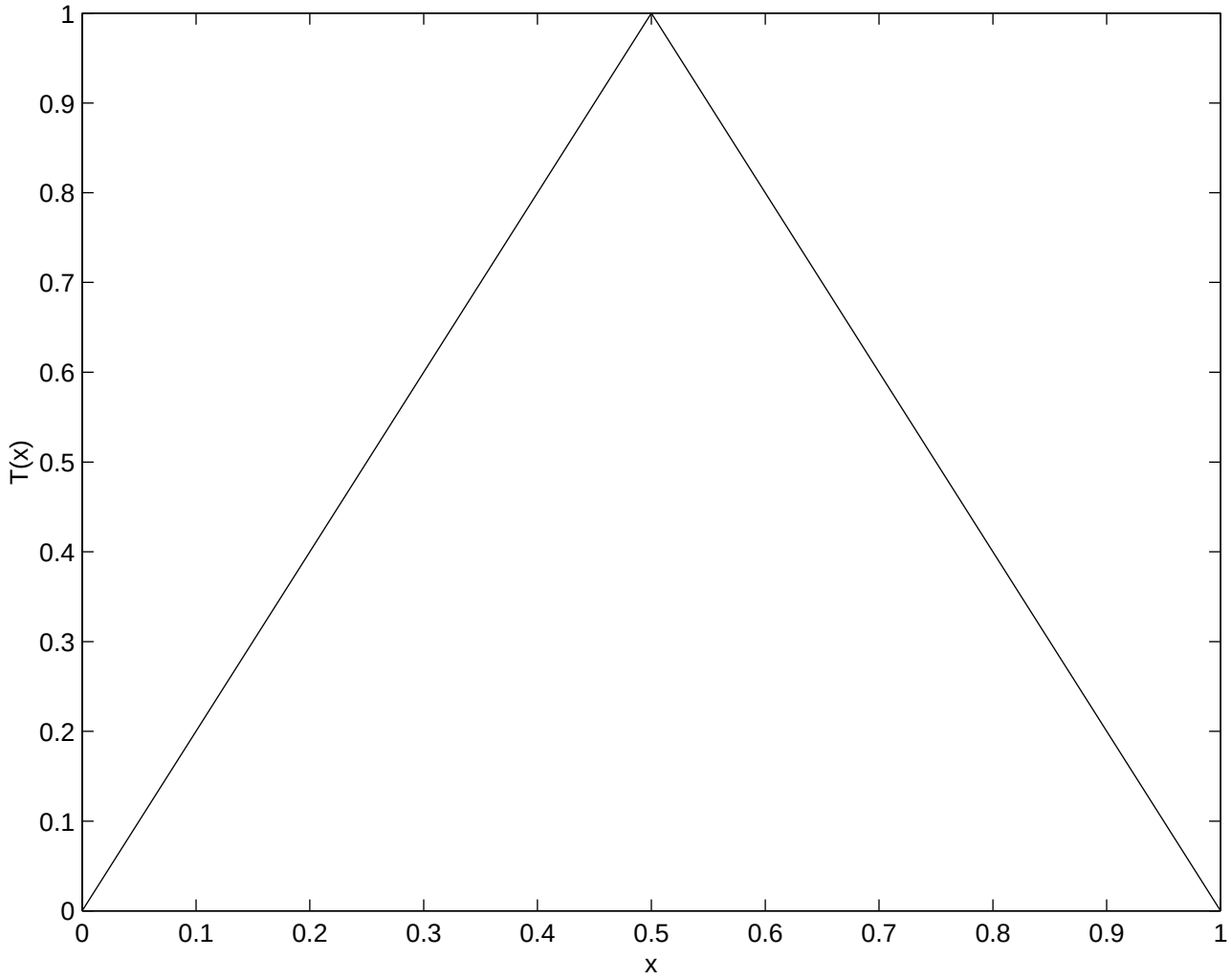


Figure 3.1: The tent map.

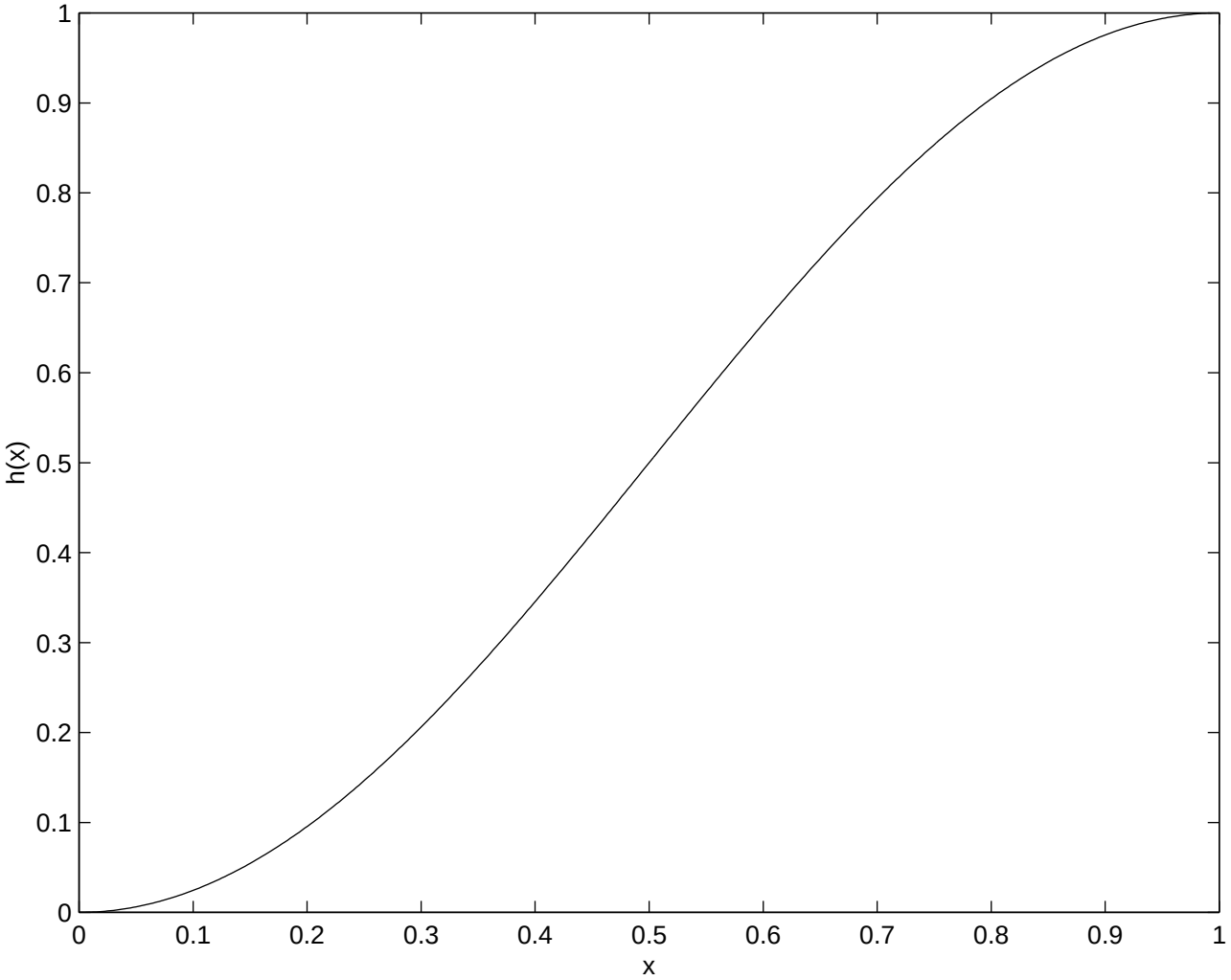


Figure 3.2: $h(x) = \sin^2\left(\frac{\pi x}{2}\right)$.

For $\frac{1}{2} < x \leq 1$ we have

$$\begin{aligned} h(T(x)) &= h(2-2x) \\ &= \sin^2(\pi - \pi x) \\ &= \sin^2 \pi x \\ &= \sin^2 2\theta \\ &= 4\sin^2\theta(1-\sin^2\theta) \\ &= L_4(h(x)) \end{aligned}$$

where we have used the fact that $\sin(\pi - \alpha) = \sin\alpha$ to pass from the second line to the third. So we have verified our claim in all cases.

Many interesting properties of a transformation are preserved under conjugation by a homeomorphism. (A *homeomorphism* is a bijective continuous map with continuous inverse.) For example, if $p$ is a periodic point of period $n$ of $f$, so that $f^{\circ n}(p) = p$, then

$$g^{\circ n}(h(p)) = h \circ f^{\circ n}(p) = h(p)$$

if $h \circ f = g \circ h$. So periodic points are carried into periodic points of the same period under a conjugacy. We will consider several other important properties of a transformation as we go along, and will prove that they are invariant under conjugacy. So what our result means is that if we prove these properties for $T$, we conclude that they are true for $L_\mu$. Since we have verified that $L_4$ is conjugate to $Q_{-2}$, we conclude that they hold for $Q_{-2}$ as well.

Here is another example of a conjugacy, this time an affine conjugacy. Consider

$$V(x) = 2|x| - 2.$$

$V$ is a map of the interval $[-2, 2]$ into itself. Consider

$$h_2(x) = 2 - 4x.$$

So $h_2(0) = 2, h_2(1) = -2$. In other words, $h_2$ maps the interval $[0, 1]$ in a one to one fashion onto the interval $[-2, 2]$. We claim that

$$V \circ h_2 = h_2 \circ T.$$

Indeed,

$$V(h_2(x)) = 2|2 - 4x| - 2.$$

For $0 \leq x \leq \frac{1}{2}$ this equals $2(2-4x) - 2 = 2 - 8x = 2 - 4(2x) = h_2(Tx)$. For $\frac{1}{2} \leq x \leq 1$ we have $V(h_2(x)) = 8x - 6 = 2 - 4(2 - 2x) = h_2(Tx)$. So we have verified the required equation in all cases. The effect of the affine transformation, $h_2$ is to enlarge the graph of $T$, shift it, and turn it upside down. But as far as iterations are concerned, these changes do not effect the essential behavior.

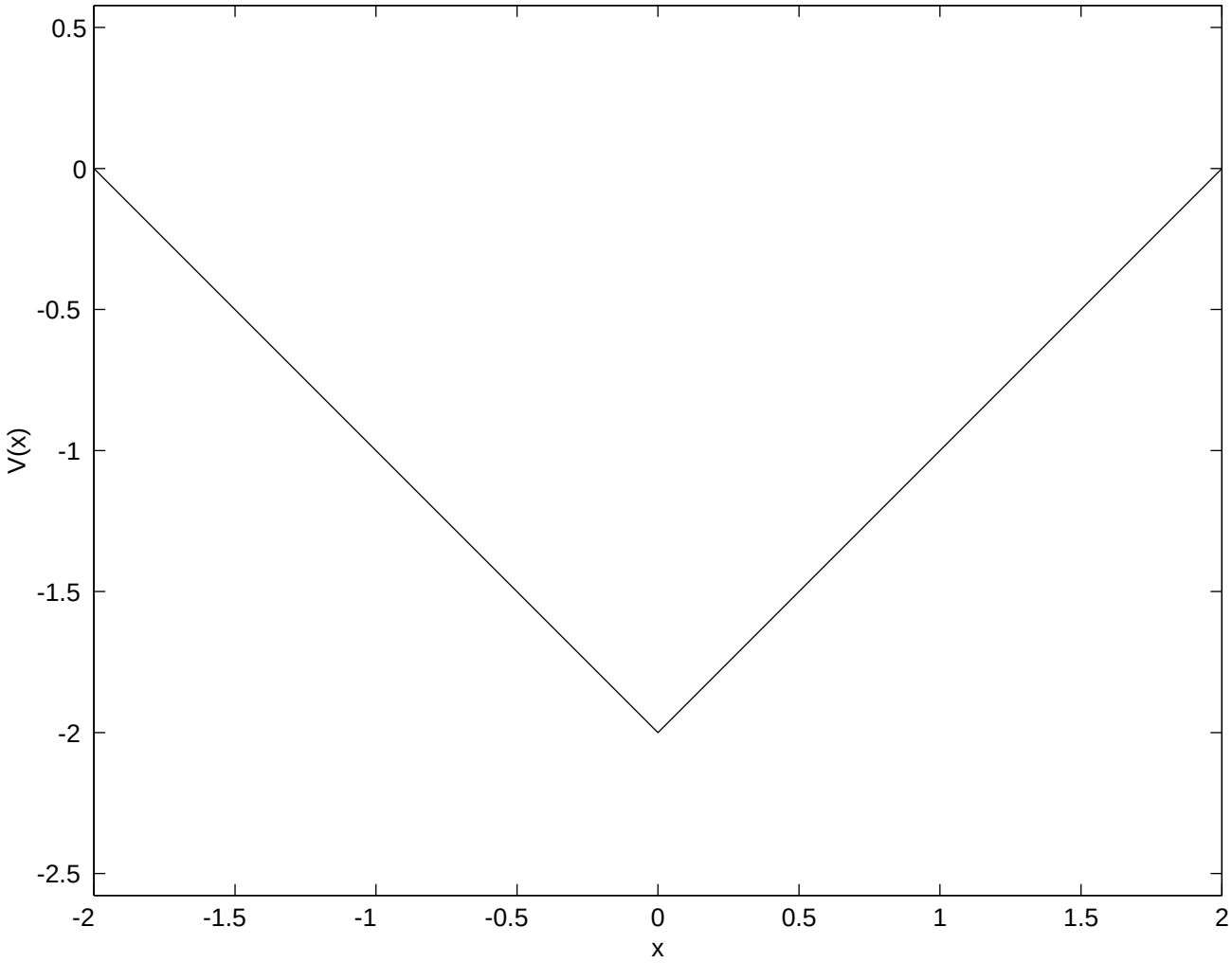


Figure 3.3: $V(x) = 2|x| - 2$.

## 3.3 Chaos

A transformation $F$ is called (topologically) *transitive* if for any two open (non empty) intervals, $I$ and $J$, one can find initial values in $I$ which, when iterated, will eventually take values in $J$. In other words, we can find an $x \in I$ and an integer $n$ so that $F^n(x) \in J$.

For example, consider the tent transformation, $T$. Notice that $T$ maps the interval $[0, \frac{1}{2}]$ onto the entire interval $[0, 1]$, and also maps the interval $[\frac{1}{2}, 1]$ onto the entire interval, $[0, 1]$. So $T^{\circ 2}$ maps each of the intervals $[0, \frac{1}{4}], [\frac{1}{4}, \frac{1}{2}], [\frac{1}{2}, \frac{3}{4}]$ and $[\frac{3}{4}]$ onto the entire interval $[0, 1]$. More generally, $T^{\circ n}$ maps each of the $2^n$ intervals $[\frac{k}{2^n}, \frac{k+1}{2^n}]$, $0 \leq k \leq 2^n - 1$ onto the entire interval $[0, 1]$. But any open interval $I$ contains some interval of the form $[\frac{k}{2^n}, \frac{k+1}{2^n}]$ if we choose $n$ sufficiently large. For example it is enough to choose $n$ so large that $\frac{3}{2^n}$ is less than the length of $I$. So for this value on $n$, $T^{\circ n}$ maps $I$ onto the entire interval $[0, 1]$, and so, in particular, there will be points, $x$, in $I$ with $F(x) \in J$.

**Proposition 3.3.1** *Suppose that $g \circ h = h \circ f$ where $h$ is continuous and surjective, and suppose that $f$ is transitive. Then $g$ is transitive.*

**Proof.** We are given non-empty open $I$ and $J$ and wish to find an $n$ and an $x \in I$ so that $g^{\circ n}(x) \in J$. To say $h$ is continuous means that $h^{-1}(J)$ is a union of open intervals. To say that $h$ is surjective implies that $h^{-1}(J)$ is not empty. Let $L$ be one of the intervals constituting $h^{-1}(J)$. Similarly, $h^{-1}(I)$ is a union of open intervals. Let $K$ be one of them. By the transitivity of $f$ we can find an $n$ and a $y \in K$ with $f^{\circ n}(y) \in L$. Let $x = h(y)$. Then $x \in I$ and $g^{\circ n}(x) = g^{\circ n}(h(y)) = h(f^{\circ n}(y)) \in h(L) \subset J$. QED.

As a corollary we conclude that if $f$ is conjugate to $g$, then $f$ is transitive if and only if $g$ is transitive. (Just apply the proposition twice, once with the roles of $f$ and $g$ interchanged.) But in the proposition we did not make the hypothesis that $h$ was bijective or that it had a continuous inverse. We will make use of this more general assertion.

A set $S$ of points is called *dense* if every non-empty open interval, $I$, contains a point of $S$. The behavior of density under continuous surjective maps is also very simple:

**Proposition 3.3.2** *If $h : X \to Y$ is a continuous surjective map, and if $D$ is a dense subset of $X$ then $h(D)$ is a dense subset of $Y$.*

**Proof.** Let $I \subset Y$ be a non-empty open interval. Then $h^{-1}(I)$ is a union of open intervals. Pick one of them, $K$ and then a point $y \in D \cap K$ which exists since $D$ is dense. But then $f(y) \in f(D) \cap I$. QED

We define $\mathrm{PER}(f)$ to be the set of periodic points of the map $f$. If $h \circ f = g \circ h$, then $f^{\circ n}(p) = p$ implies that $g^{\circ n}(h(p)) = h(f^{\circ n}(p)) = h(p)$ so

$$h[\mathrm{PER}(f)] \subset \mathrm{PER}(g).$$

In particular, if $h$ is continuous and surjective, and if PER($f$) is dense, then so is PER($g$).

Following Devaney and recent work (1992) by J. Banks et.al. *Amer. Math. Monthly* **99** (1992) 332-334, let us call $f$ **chaotic** if $f$ is transitive and PER($f$) is dense. It follows from the above discussion that

**Proposition 3.3.3** *If $h : X \to Y$ is surjective and continuous, if $f : X \to X$ is chaotic, and if $h \circ f = g \circ h$, then $g$ is chaotic.*

We have already verified that the tent transformation, $T$, is transitive. We claim that PER($T$) is dense on $[0,1]$ and hence that $T$ is chaotic. To see this, observe that $T^n$ maps the interval $[\frac{k}{2^n}, \frac{k+1}{2^n}]$ onto $[0,1]$. In particular, there is a point $x \in [\frac{k}{2^n}, \frac{k+1}{2^n}]$ which is mapped into itself. In other words, every interval $[\frac{k}{2^n}, \frac{k+1}{2^n}]$ contains a point of period $n$ for $T$. But any non-empty open interval $I$ contains an interval of the type $[\frac{k}{2^n}, \frac{k+1}{2^n}]$ for sufficiently large $n$. Hence $T$ is chaotic.

From the above propositions it follows that $L_4, Q_{-2}$, and $V$ are all chaotic.

## 3.4 The saw-tooth transformation and the shift

Define the function $S$ by

$$S(x) = 2x, \quad 0 \le x < \frac{1}{2}, \qquad S(x) = 2x - 1, \quad \frac{1}{2} \le x \le 1. \tag{3.4}$$

The map $S$ is discontinuous at $x = .5$. However, we can find a continuous, surjective map, $h$, such that $h \circ S = T \circ h$. In fact, we can take $h$ to be $T$ itself! In other words we claim that

$$\begin{array}{ccc} I & \xrightarrow{S} & I \\ {\scriptstyle T}\downarrow & & \downarrow{\scriptstyle T} \\ I & \xrightarrow[T]{} & I \end{array}$$

commutes where $I = [0,1]$. To verify this, we successively compute both $T \circ T$ and $T \circ S$ on each of the quarter intervals:

$$\begin{array}{lcccll} T(T(x)) & = & T(2x) & = & 4x & \text{for } 0 \le x \le 0.25 \\ T(S(x)) & = & T(2x) & = & 4x & \text{for } 0 \le x \le 0.25 \\ T(T(x)) & = & T(2x) & = & -4x+2 & \text{for } 0.25 < x < 0.5 \\ T(S(x)) & = & T(2x) & = & -4x+2 & \text{for } 0.25 \le x < 0.5 \\ T(T(x)) & = & T(-2x+2) & = & 4x-2 & \text{for } 0.5 \le x \le 0.75 \\ T(S(x)) & = & T(2x-1) & = & 4x-2 & \text{for } 0.5 \le x \le 0.75 \\ T(T(x)) & = & T(-2x+2) & = & -4x+4 & \text{for } 0.75 < x \le 1 \\ T(S(x)) & = & T(2x-1) & = & -4x+4 & \text{for } 0.75 < x \le 1 \end{array}$$

The $h$ that we are using (namely $h = T$) is not one to one. That is why our diagram can commute even though $T$ is continuous and $S$ is not.

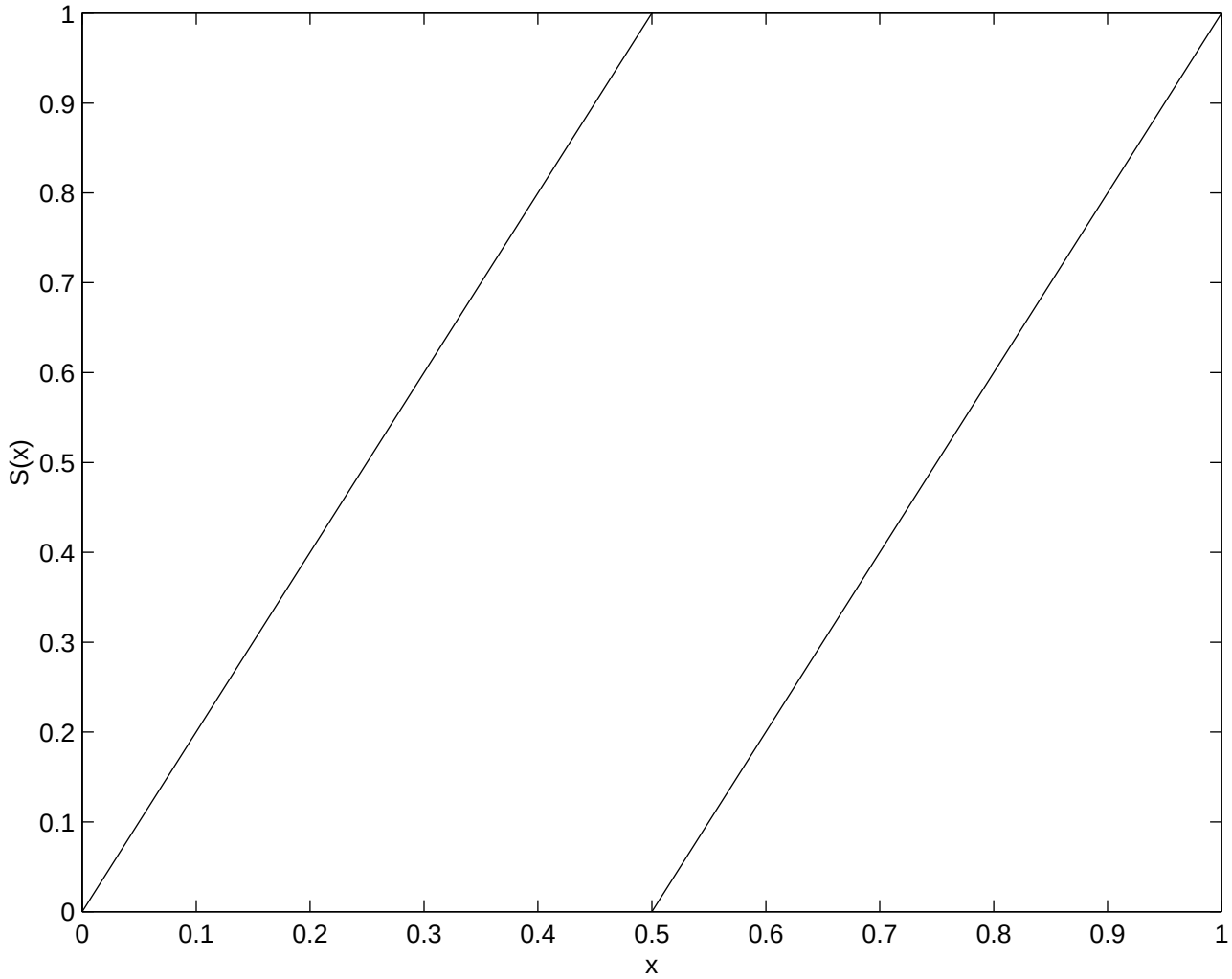


Figure 3.4: The discontinuous function $S$.

We now give an alternative description of the saw-tooth function which makes it clear that it is chaotic. Let $X$ be the set of infinite (one sided) sequences of zeros and ones. So a point of $X$ is a sequence $\{a_1a_2a_3\dots\}$ where each each $a_i$ is either 0 or 1. However we exclude all points with a tail consisting of infinite repeating $1'$s. So a sequence such as $\{00111111111\dots\}$ is excluded. We will identify $X$ with the half open interval $[0,1)$ by assigning to each point $x \in [0,1)$ its binary expansion, and by assigning to each sequence $a = \{a_1a_2a_3\dots\}$ the number

$$h(a) = \sum \frac{a_i}{2^i}.$$

The map $h : X \to [0,1)$ just defined is clear. The inverse map, assigning to each real number between 0 and 1 its binary expansion deserves a little more discussion: Take a point $x \in [0,1)$. If $x < \frac{1}{2}$ the first entry in its binary expansion is 0. If $\frac{1}{2} \leq x$ then the first entry in the binary expansion of $x$ is 1. Now apply $S$. If $S(x) < \frac{1}{2}$ (which means that either $0 \leq x < \frac{1}{4}$ or $\frac{1}{2} \leq x < \frac{3}{4}$) then the second entry of the binary expansion of $x$ is 0, while if $\frac{1}{2} \leq S(x) < 1$ then the second entry in the binary expansion of $x$ is 1. Thus the operator $S$ provides the algorithm for the computation of the binary expansion of $x$. Let us consider, for example, $x = \frac{7}{16}$. Then the sequence $\{S^k(x)\}, k = 0, 1, 2, 3, \dots$ is

$$\frac{7}{16}, \frac{7}{8}, \frac{3}{4}, \frac{1}{2}, 0, 0, 0, \dots.$$

In general it is clear that for any number of the form $\frac{k}{2^n}$, after $n-1$ iterations of the operator $S$ the result will be either 0 or $\frac{1}{2}$. So all $S^k(x) = 0, k \geq n$. In particular, no infinite sequence with a tail of repeating 1's can arise. We see that the binary expansion of $h(a)$ gives us $a$ back, so we may (and shall) identify $X$ with $[0,1)$. Notice that we did not start with any independent notion of topology or metric on $X$. But now that we have identified $X$ with $[0,1)$, we can use use standard notions of distance on the unit interval but expressed in terms of properties of the sequences. For example, if the binary expansions of $x$ and $y$ agree up to the $k$th position, then

$$|x - y| < 2^{-k}.$$

So we define the distance between to sequences $a$ and $b$ to be $2^{-k}$ where $k$ is the first place they do not agree. (Of course we define the distance from an $a$ to itself to be zero.)

The expression of $S$ in terms of the binary representation is very simple:

$$S : .a_1a_2a_3a_4\dots \mapsto .a_2a_3a_4a_5\dots.$$

It consists of throwing away the first digit and then shifting the entire sequence one unit to the left.

From this description it is clear that PER($S$) consists of points with eventually repeating binary expansions, these are the rational numbers. They are dense. We can see that $S$ is transitive as follows: We are given intervals $I$ and $J$.

Let $y = .b_1b_2b_3\ldots$ be a point of $J$, and let $z = .a_1a_2a_3\ldots$ be a point of $I$ which is at a distance greater than $2^{-n}$ from the boundary of $I$. We can always find such a point if $n$ is sufficiently large. Indeed, if we choose $n$ so that the length of $I$ is greater than $\frac{1}{2^{(n-1)}}$, the midpoint of $I$ has this property. In particular, any point whose binary expansion agrees with $z$ up to the $n-$th position lies in $I$. Take $x$ to be the point whose first $n$ terms in the binary expansion are those of $z$, followed by the binary expansion of $y$, so

$$x = 0.a_1a_2a_3\ldots a_nb_1b_2b_3b_4\ldots.$$

The point $x$ lies in $I$ and $S^n(x) = y$. Not only is $S$ transitive, we can hit *any* point of $J$ by applying $S^n$ (with $n$ fixed, depending only on $I$) to a suitable point of $I$. This is much more than is demanded by transitivity. Thus $S$ is chaotic on $[0, 1)$.

Of course, once we know that $S$ is chaotic on the open interval $[0, 1)$, we know that it is chaotic on the closed interval $[0, 1]$ since the addition on one extra point (which gets mapped to 0 by $S$) does not change the definitions.

Now consider the map $t \mapsto e^{2\pi it}$ of $[0, 1]$ onto the unit circle, $S^1$. Another way of writing this map is to describe a point on the unit circle by $e^{i\theta}$ where $\theta$ is an angular variable, that is $\theta$ and $\theta + 2\pi$ are identified. Then the map is $t \mapsto 2\pi t$. This map, $h$, is surjective and continuous and is one to one except at the end points: 0 and 1 are mapped into the same point on $S^1$. Clearly

$$h \circ S = D \circ h$$

where

$$D(\theta) = 2\theta.$$

Or, if we write $z = e^{i\theta}$, then in terms of $z$, the map $D$ sends

$$z \mapsto z^2.$$

So $D$ is called the doubling map or the squaring map. We have proved that it is chaotic. We can use the fact that $D$ is chaotic to give an alternative proof of the fact that $Q_{-2}$ is chaotic. Indeed, consider the map $h : S^1 \to [-2, 2]$

$$h(\theta) = 2\cos\theta.$$

It is clearly surjective and continuous. We claim that

$$h \circ D = Q_{-2} \circ h.$$

Indeed,

$$h(D(\theta)) = 2\cos 2\theta = 2(2\cos^2\theta - 1) = (2\cos\theta)^2 - 2 = Q_{-2}(h(\theta)).$$

This gives an alternative proof that $Q_{-2}$ (and hence $L_4$ and $T$) are chaotic.

## 3.5 Sensitivity to initial conditions

In this section we prove that if $f$ is chaotic, then $f$ is sensitive to initial conditions in the sense of the following

**Proposition 3.5.1 (Sensitivity.)** *Let $f : X \to X$ be a chaotic transformation. Then there is a $d > 0$ such that for any $x \in X$ and any open set $J$ containing $x$ there is a point $y \in J$ and an integer, $n$ with*

$$|f^{\circ n}(x) - f^{\circ n}(y)| > d. \tag{3.5}$$

In other words, we can find points arbitrarily close to $x$ which move a distance at least $d$ away. This for any $x \in X$. We begin with a lemma.

**Lemma 3.5.1** *There is a $c > 0$ with the property that for any $x \in X$ there is a periodic point $p$ such that*

$$|x - f^{\circ k}(p)| > c, \quad \forall k.$$

**Proof of lemma.** Choose two periodic points, $r$ and $s$ with distinct orbits, so that $|f^{\circ k}(r) - f^{\circ l}(s)| > 0$ for all $k$ and $l$. Choose $c$ so that $2c < \min |f^{\circ k}(r) - f^{\circ l}(s)|$. Then for all $k$ and $l$ we have

$$\begin{aligned} 2c &< |f^{\circ k}(r) - f^{\circ l}(s)| \\ &= |f^{\circ k}(r) - x + x - f^{\circ l}(s)| \\ &\leq |f^{\circ k}(r) - x| + |f^{\circ l}(s) - x|. \end{aligned}$$

If $x$ is within distance $c$ to *any* of the points $f^{\circ} l(s)$ then it must be at a greater distance than $c$ from *all* of the points $f^{\circ k}(r)$ and vice versa. So one of the two, $r$ or $s$ will work as the $p$ for $x$.

**Proof of proposition with $d = c/4$.** Let $x$ be any point of $X$ and $J$ any open set containing $x$. Since the periodic points of $f$ are dense, we can find a periodic point $q$ of $f$ in

$$U = J \cap B_d(x),$$

where $B_d(x)$ denotes the open interval of length $d$ centered at $x$,

$$B_d(x) = (x - d, x + d).$$

Let $n$ be the period of $q$. Let $p$ be a periodic point whose orbit is of distance greater than $4d$ from $x$, and set

$$W_i = B_d(f^{\circ i}(p)) \cap X.$$

Since $f^{\circ i}(p) \in W_i$, i.e. $p \in f^{-i}(W_i) = (f^{\circ i})^{-1}(W_i)$ for all $i$, we see that the open set

$$V = f^{-1}(W_1) \cap f^{-2}(W_2) \cap \cdots \cap f^{-n}(W_n)$$

is not empty.

Now we use the transitivity property of $f$ applied to the open sets $U$ and $V$. By assumption, we can find a $z \in U$ and a positive integer $k$ such that $f^k(z) \in V$. Let $j$ be the smallest integer so that $k < nj$. In other words,

$$1 \leq nj - k \leq n.$$

So

$$f^{nj}(z) = f^{nj-k}(f^k(z)) \in f^{nj-k}(V).$$

But

$$\begin{aligned} f^{nj-k}(V) &= f^{nj-k}\left(f^{-1}(W_1) \cap f^{-2}(W_2) \cap \cdots \cap f^{-n}(W_n)\right) \\ &\subset f^{nj-k}(f^{-(nj-k)}W_{nj-k}) \\ &= W_{nj-k}. \end{aligned}$$

In other words,

$$|f^{nj}(z) - f^{nj-k}(p)| < d.$$

On the other hand, $f^{nj}(q) = q$, since $n$ is the period of $q$. Thus

$$\begin{aligned} |f^{nj}(q) - f^{nj}(z)| &= |q - f^{nj}(z)| \\ &= |x - f^{nj-k}(p) + f^{nj-k}(p) - f^{nj}(z) + q - x| \\ &\geq |x - f^{nj-k}(p)| - |f^{nj-k}(p) - f^{nj}(z)| - |q - x| \\ &\geq 4d - d - d = 2d. \end{aligned}$$

But this last inequality implies that either

$$|f^{nj}(x) - f^{nj}(z)| > d$$

or

$$|f^{nj}(x) - f^{nj}(q)| > d$$

for if $x$ were within distance $d$ from both of these points, they would have to be within distance $2d$ from each other, contradicting the preceding inequality. So one of the two, $z$ or $q$ will serve as the $y$ in the proposition with $m = nj$.

## 3.6 Conjugacy for monotone maps

We begin this section by showing that if $f$ and $g$ are continuous strictly monotone maps of the unit interval $I = [0, 1]$ onto itself, and if their graphs are both strictly below (or both strictly above) the line $y = x$ in the interior of $I$, then they are conjugate by a homeomorphism. Here is the precise statement:

**Proposition 3.6.1** *Let $f$ and $g$ be two continuous monotone strictly increasing functions defined on $[0,1]$ and satisfying*

$$\begin{aligned} f(0) &= 0 \\ g(0) &= 0 \\ f(1) &= 1 \\ g(1) &= 1 \\ f(x) &< x \quad \forall x \neq 0, 1 \\ g(x) &< x \quad \forall x \neq 0, 1. \end{aligned}$$

*Then there exists a continuous, monotone increasing function $h$ defined on $[0,1]$ with*

$$h(0) = 0, \qquad h(1) = 1,$$

*and*

$$h \circ f = g \circ h.$$

**Proof.** Choose any point $(x_0, y_0)$ in the open square

$$0 < x < 1, \quad 0 < y < 1.$$

If $(x_0, y_0)$ is to be a point on the curve $y = h(x)$, then the equation $h \circ f = g \circ h$ implies that the point $(x_1, y_1)$ also lies on this curve, where

$$x_1 = f(x_0), \quad y_1 = g(y_0).$$

By induction so will the points $(x_n, y_n)$ where

$$x_n = f^n(x_0), \quad y_n = g^n(y_0).$$

By hypothesis

$$x_0 > x_1 > x_2 > ...,$$

and since there is no solution to $f(x) = x$ for $0 < x < 1$ the limit of the $x_n$, $n \to \infty$ must be zero. Also for the $y_n$. So the sequence of points $(x_n, y_n)$ approaches $(0,0)$ as $n \to +\infty$. Similarly, as $n \to -\infty$ the points $(x_n, y_n)$ approach $(1,1)$. Now choose any continuous, strictly monotone function

$$y = h(x),$$

defined on

$$x_1 \leq x \leq x_0$$

with

$$h(x_1) = y_1, \quad h(x_0) = y_0.$$

Extend its definition to the interval $x_2 \leq x \leq x_1$ by setting

$$h(x) = g(h(f^{-1}(x))), \qquad x_2 \leq x \leq x_1.$$

Notice that at $x_1$ we have

$$g(h(f^{-1}(x_1))) = g(h(x_0)) = g(y_0) = y_1,$$

so the definitions of $h$ at the point $x_1$ are consistent. Since $f$ and $g$ are monotone and continuous, and since $h$ was chosen to be monotone on $x_1 \leq x \leq x_0$, we conclude that $h$ is monotone on $x_2 \leq x \leq x_1$ and hence continuous and monotone on all of $x_2 \leq x \leq x_0$. Continuing in this way, we define $h$ on the interval $x_{n+1} \leq x \leq x_n,\ n \geq 0$ by

$$h = g^n \circ h \circ f^{-n}.$$

Setting $h(0) = 0$, we get a continuous and monotone increasing function defined on $0 \leq x \leq x_0$. Similarly, we extend the definition of $h$ to the right of $x_0$ up to $x = 1$. By its very construction, the map $h$ conjugates $f$ into $g$, proving the proposition.

Notice that as a corollary of the method of proof, we can conclude

**Proposition 3.6.2** *Let $f$ and $g$ be two monotone increasing functions defined in some neighborhood of the origin and satisfying*

$$f(0) = g(0) = 0, \quad |f(x)| < |x|,\ |g(x)| < |x|,\ \forall x \neq 0.$$

*Then there exists a homeomorphism, $h$ defined in some neighborhood of the origin with $h(0) = 0$ and*

$$h \circ f = g \circ h.$$

Indeed, just apply the method (for $n \geq 0$) to construct $h$ to the right of the origin, and do an analogous procedure to construct $h$ to the left of the origin. As a special case we obtain

**Proposition 3.6.3** *Let $f$ and $g$ be differentiable functions with $f(0) = g(0) = 0$ and*

$$0 < f'(0) < 1, \quad 0 < g'(0) < 1. \tag{3.6}$$

*Then there exists a homeomorphism $h$ defined in some neighborhood of the origin with $h(0) = 0$ and which conjugates $f$ into $g$.*

The mean value theorem guarantees that the hypotheses of the preceding proposition are satisfied.

Also, it is clear that we can replace (3.6) by any of the conditions

$$\begin{array}{ll} 1 < f'(0), & 1 < g'(0) \\ 0 > f'(0) > -1, & 0 > g'(0) > -1 \\ -1 > f'(0), & -1 > g'(0), \end{array}$$

and the conclusion of the proposition still holds.

It is important to observe that if $f'(0) \neq g'(0)$, then the homeomorphism, $h$, can not be a diffeomorphism. That is, $h$ can not be differentiable with

a differentiable inverse. In fact, $h$ can not have a non-zero derivative at the origin. Indeed, differentiating the equation $g \circ h = h \circ f$ at the origin gives

$$g'(0)h'(0) = h'(0)f'(0),$$

and if $h'(0) \neq 0$ we can cancel it form both sides of the equation so as to obtain

$$f'(0) = g'(0). \tag{3.7}$$

What is true is that if (3.7) holds, and if

$$|f'(0)| \neq 1, \tag{3.8}$$

then we can find a differentiable $h$ with a differentiable inverse which conjugates $f$ into $g$.

We postpone the proof of this result until we have developed enough machinery to deal with the $n$-dimensional result. These theorems are among my earliest mathematical theorems. A complete characterization of transformations of $\mathbf{R}$ near a fixed point together with the conjugacy by smooth maps if (3.7) and (3.8) hold, were obtained and submitted for publication in 1955 and published in the Duke Mathematical Journal. The discussion of equivalence under homeomorphism or diffeomorphism in $n$-dimensions was treated for the case of contractions in 1957 and in the general case in 1958, both papers appearing in the American Journal of Mathematics. We will return to these matters in Chapter ??.

## 3.7 Sequence space and symbolic dynamics.

In this section we will illustrate a powerful method for studying dynamical systems by examining the quadratic transformation

$$Q_c : x \mapsto x^2 + c$$

for values of $c < -2$.

For any value of $c$, the two possible fixed points of $Q_c$ are

$$p_-(c) = \frac{1}{2}(1 - \sqrt{1 - 4c}), \quad p_+(c) = \frac{1}{2}(1 + \sqrt{1 - 4c})$$

by the quadratic formula. These roots are real with $p_-(c) < p_+(c)$ for $c < 1/4$. The graph of $Q_c$ lies above the diagonal for $x > p_+(c)$, hence the iterates of any $x > p_+(c)$ tend to $+\infty$. If $x_0 < -p_+(c)$, then $x_1 = Q_c(x_0) > p_+(c)$, and so the further iterates also tend to $+\infty$. Hence all the interesting action takes place in the interval $[-p_+, p_+]$. The function $Q_c$ takes its minimum value, $c$, at $x = 0$, and

$$c = -p_+(c) = -\frac{1}{2}(1 + \sqrt{1 - 4c})$$

when $c = -2$. For $-2 \le c \le 1/4$, the iterate of any point in $[-p_+, p_+]$ remains in the interval $[-p_+, p_+]$. But for $c < -2$ some points will escape, and it is this latter case that we want to study.

To visualize the what is going on, draw the square whose vertices are at $(\pm p_+, \pm p_+)$ and the graph of $Q_c$ over the interval $[-p_+, p_+]$. The bottom of the graph will protrude below the bottom of the square. Let $A_1$ denote the open interval on the $x$-axis (centered about the origin) which corresponds to this protrusion. So

$$A_1 = \{x | Q_c(x) < -p_+(c)\}.$$

Every point of $A_1$ escapes from the interval $[-p_+, p_+]$ after one iteration.

Let

$$A_2 = Q_c^{-1}(A_1).$$

Since every point of $[-p_+, p_+]$ has exactly two pre-images under $Q_c$, we see that $A_2$ is the union of two open intervals. To fix notation, let

$$I = [-p_+, p_+]$$

and write

$$I \backslash A_1 = I_0 \cup I_1$$

where $I_0$ is the closed interval to the left of $A_1$ and $I_1$ is the closed interval to the right of $A_1$.Thus $A_2$ is the union of two open intervals, one contained in $I_0$ and the other contained in $I_1$. Notice that a point of $A_2$ escapes from $[-p_+, p_+]$ in exactly two iterations: one application of $Q_c$ moves it into $A_1$ and another application moves it out of $[-p_+, p_+]$.

Conversely, suppose that a point $x$ escapes from $[-p_+, p_+]$ in exactly two iterations. After one iteration it must lie in $A_1$, since these are exactly the points that escape in one iteration. Hence it must lie in $A_2$.

In general, let

$$A_{n+1} = Q_c^{-\circ n}(A_1).$$

Then $A_{n+1}$ is the union of $2^n$ open intervals and consists of those points which escape from $[-p_+, p_+]$ in exactly $n + 1$ iterations. If the iterates of a point $x$ eventually escape from $[-p_+, p_+]$, there must be some $n \ge 1$ so that $x \in A_n$. In other words,

$$\bigcup_{n \ge 1} A_n$$

is the set of points which eventually escape. The remaining points, lying in the set

$$\Lambda := I \backslash \bigcup_{n \ge 1} A_n m,$$

are the points whose iterates remain in $[-p_+, p_+]$ forever. The thrust of this section is to study $\Lambda$ and the action of $Q_c$ on it.

Since $\Lambda$ is defined as the complement of an open set, we see that $\Lambda$ *is closed*. Let us show that $\Lambda$ *is not empty*. Indeed, the fixed points, $p_\pm$ certainly belong to $\Lambda$ and hence so do all of their inverse images, $Q_c^{-n}(p_\pm)$. Next we will prove

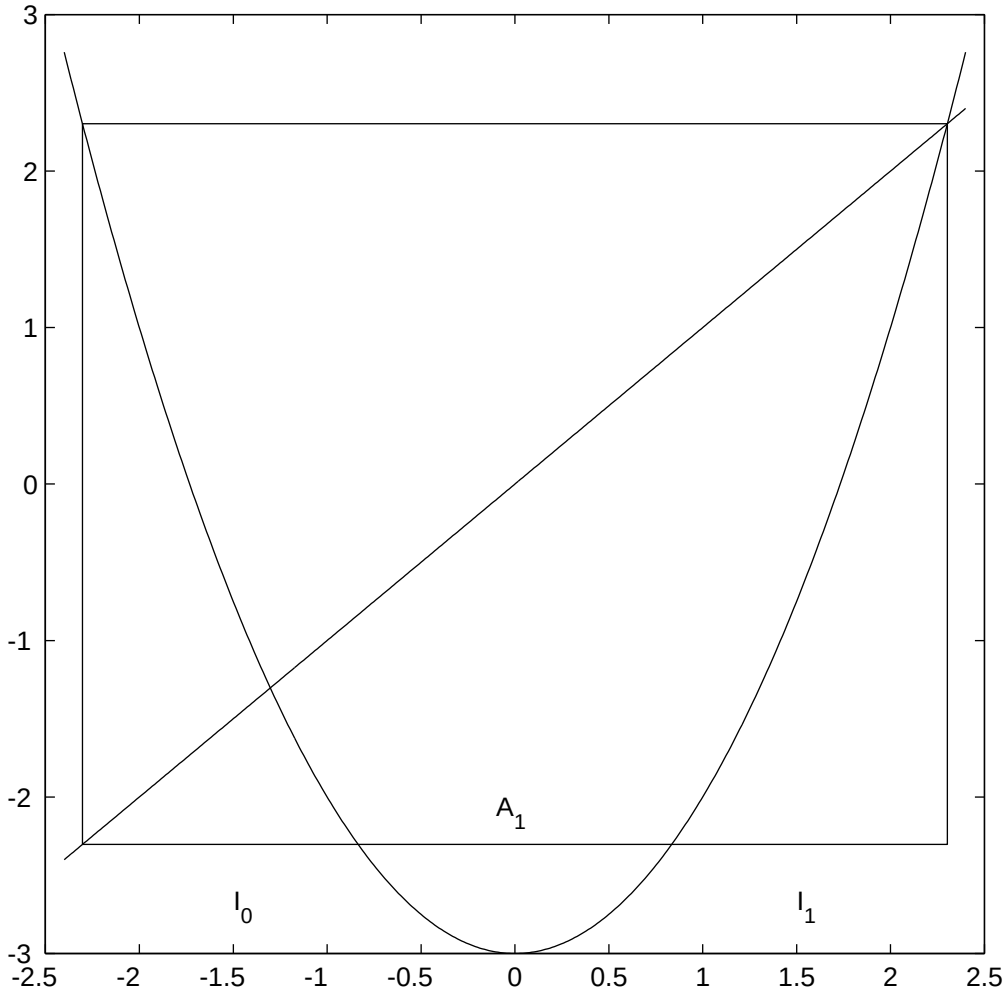


Figure 3.5: $Q_3$.

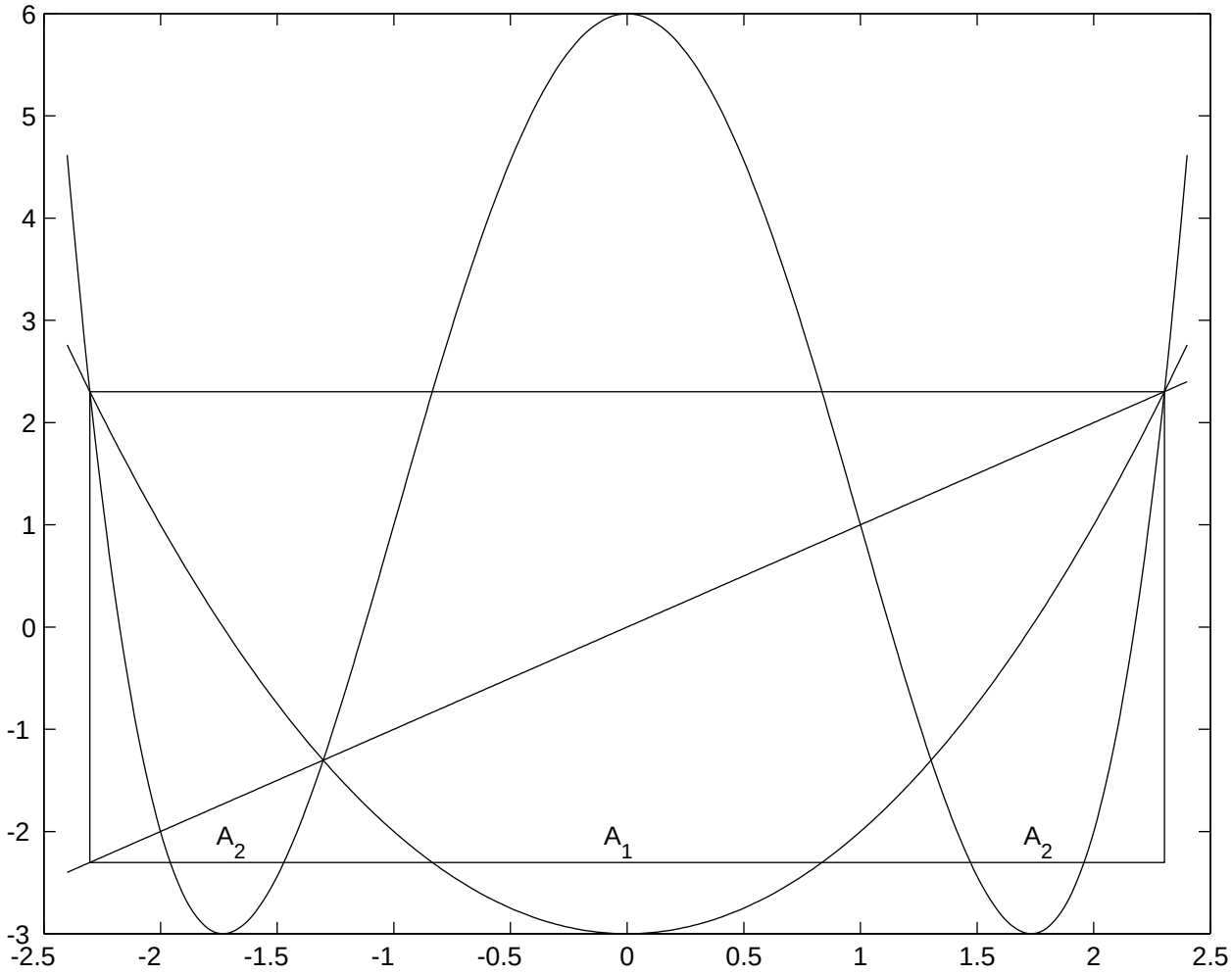


Figure 3.6: $Q_3$ and $Q_3^{\circ 2}$ .

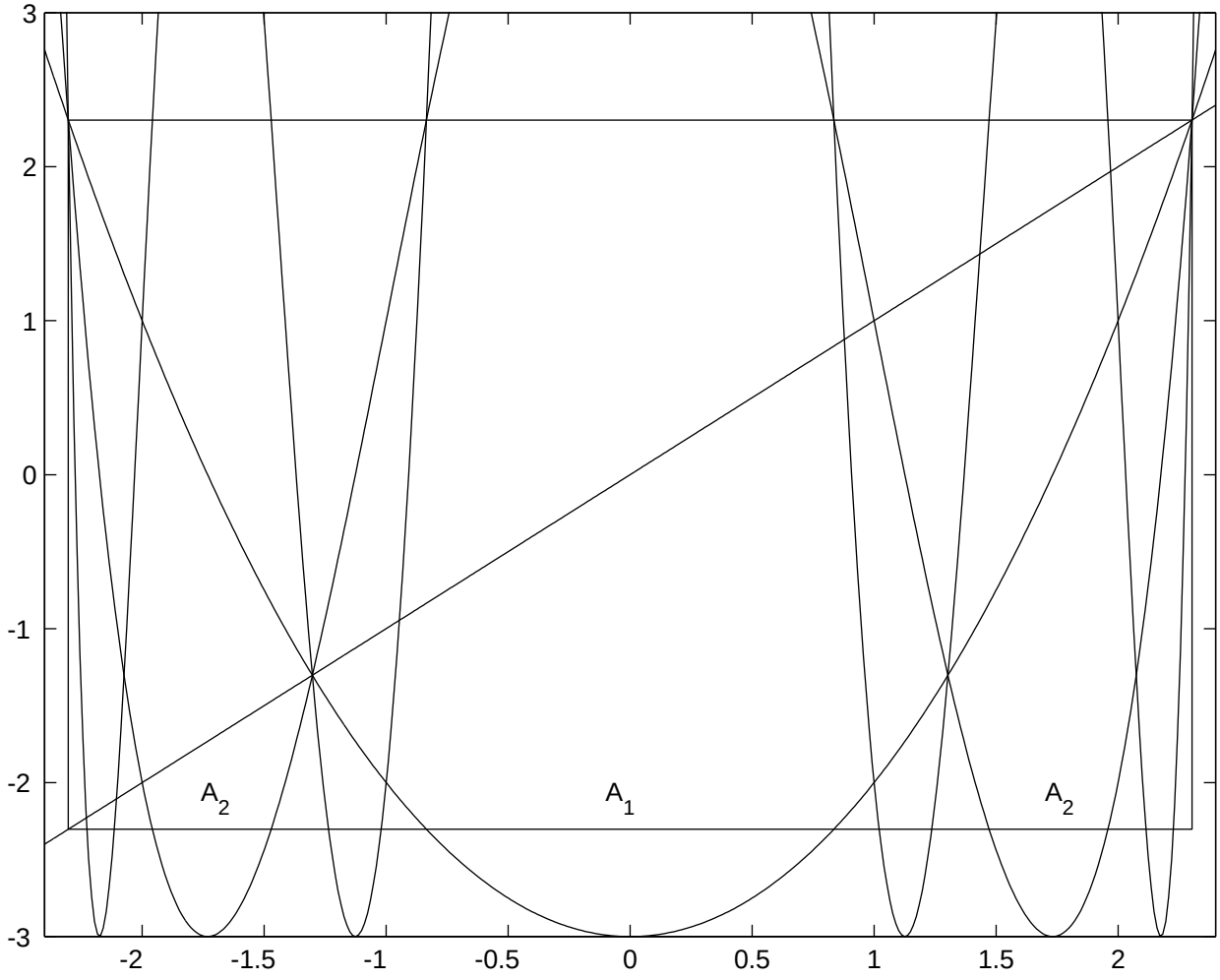


Figure 3.7: $Q_3$, $Q_3^{\circ 2}$, and $Q_3^{\circ 3}$ .

**Proposition 3.7.1** *If*

$$c < -\frac{5+2\sqrt{5}}{4} \doteq -2.368\ldots \tag{3.9}$$

*then $\Lambda$ is totally disconnected, that is, it contains no interval.*

In fact, the proposition is true for all $c < -2$ but, following Devaney [**?**] we will only present the simpler proof when we assume (3.9). For this we use

**Lemma 3.7.1** *If (3.9) holds then there is a constant $\lambda > 1$ such that*

$$|Q_c'(x)| > \lambda > 1, \quad \forall x \in I \backslash A_1. \tag{3.10}$$

**Proof of Lemma.** We have $|Q_c'(x)| = |2x| > \lambda > 1$ if $|x| > \frac{1}{2}\lambda$ for all $x \in I\backslash A_1$. So we need to arrange that $A_1$ contains the interval $[-\frac{1}{2}, \frac{1}{2}]$ in its interior. In other words, we need to be sure that

$$Q_c(\frac{1}{2}) < -p_+.$$

The equality

$$Q_c(\frac{1}{2}) = -p_+$$

translates to

$$\frac{1}{4} + c = -\frac{1+\sqrt{1-4c}}{2}.$$

Solving the quadratic equation gives

$$c = -\frac{5+2\sqrt{5}}{4}$$

as the lower root. Hence if (3.9) holds, $Q_c(\frac{1}{2}) < -p_+$.

**Proof of Prop. 3.7.1.** Suppose that there is an interval, $J$, contained in $\Lambda$. Then $J$ is contained either in $I_0$ or $I_1$. In either event the map $Q_c$ is one to one on $J$ and maps it onto an interval. For any pair of points, $x$ and $y$ in $J$, the mean value theorem implies that

$$|Q_c(x) - Q_c(y)| > \lambda |x - y|.$$

Hence if $d$ denotes the length of $J$, then $Q_c(J)$ is an interval of length at least $\lambda d$ contained in $\Lambda$. By induction we conclude that $\Lambda$ contains an interval of length $\lambda^n d$ which is ridiculous, since eventually $\lambda^n d > 2p_+$ which is the length of $I$. QED.

Now consider a point $x \in \Lambda$. Either it lies in $I_0$ or it lies in $I_1$. Let us define

$$s_0(x) = 0 \quad \forall x \in I_0$$

and

$$s_0(x) = 1 \quad \forall x \in I_1.$$

Since all points $Q_c^{\circ n}(x)$ are in $\Lambda$, we can define $s_n(x)$ to be 0 or 1 according to whether $Q_c^{\circ n}(x)$ belongs to $I_0$ or $I_1$. In other words, we define

$$s_n(x) := \begin{cases} 0 & \text{if} \quad Q_c^{\circ n}(x) \in I_o \\ \\ 1 & \text{if} \quad Q_c^{\circ n}(x) \in I_1 \end{cases} . \tag{3.11}$$

higher iterates of $Q_c$.

So let us introduce the **sequence space**, $\Sigma$, defined as

$$\Sigma = \{(s_0 s_1 s_2 \dots) \mid s_j = 0 \text{ or } 1\}.$$

Notice that in contrast to the space $X$ we introduced in Section 3.4, we are not excluding any sequences. Define the notion of *distance* or *metric* on $\Sigma$ by defining the distance between two points

$$\mathbf{s} = (s_0 s_1 s_2 \dots)$$

and

$$\mathbf{t} = (t_0 t_1 t_2 \dots)$$

to be

$$d(\mathbf{s}, \mathbf{t}) \stackrel{def}{=} \sum_{i=0}^{\infty} \frac{|s_i - t_i|}{2^i}.$$

It is immediate to check that $d$ satisfies all the requirements for a metric: It is clear that $d(\mathbf{s}, \mathbf{t}) \geq 0$ and $d(\mathbf{s}, \mathbf{t}) = 0$ implies that $|s_i - t_i| = 0$ for all $i$, and hence that $\mathbf{s} = \mathbf{t}$. The definition is clearly symmetric in $\mathbf{s}$ and $\mathbf{t}$. And the usual triangle inequality

$$|s_i - u_i| \leq |s_i - t_i| + |t_i - u_i|$$

for each $i$ implies the triangle inequality

$$d(\mathbf{s}, \mathbf{u}) \leq d(\mathbf{s}, \mathbf{t}) + d(\mathbf{t}, \mathbf{u}).$$

Notice that if $s_i = t_i$ for $i = 0, 1, \dots, n$ then

$$d(\mathbf{s}, \mathbf{t}) = \sum_{j=n+1}^{\infty} \frac{|s_j - t_j|}{2^j} \leq \sum_{j=n+1}^{\infty} \frac{1}{2^j} = \frac{1}{2^n}.$$

Conversely, if $s_i \neq t_i$ for some $i \leq n$ then

$$d(\mathbf{s}, \mathbf{t}) \geq \frac{1}{2^j} \geq \frac{1}{2^n}.$$

So if

$$d(\mathrm{s}, \mathrm{t}) < \frac{1}{2^n}$$

then $s_i = t_i$ for all $i \leq n$.

Getting back to $\Lambda$, define the map

$$\iota : \Lambda \to \Sigma$$

by

$$\iota(x) = (s_0(x)s_1(x)s_2(x)s_3(x)\dots) \tag{3.12}$$

where the $s_i(x)$ are defined by (3.11).

The point $\iota(x)$ is called the *itinerary* of the point $x$. For example, the fixed point, $p_+$ lies in $I_1$ and hence do all of its images under $Q_c^n$ since they all coincide with $p_+$. Hence its itinerary is

$$\iota(p_+) = (111111\dots).$$

The point $-p_+$ is carried into $p_+$ under one application of $Q_c$ and then stays there forever. Hence its itinerary is

$$\iota(-p_+) = (01111111\dots).$$

It follows from the very definition that

$$\iota(Q_c(x)) = S(\iota(x))$$

where $S$ is our old friend, the shift map,

$$S : (s_0s_1s_2s_3\dots) \mapsto (s_1s_2s_3s_4\dots)$$

applied to the space $\Sigma$. In other words,

$$\iota \circ Q_c = S \circ \iota.$$

The map $\iota$ conjugates $Q_c$, acting on $\Lambda$ into the shift map, acting on $\Sigma$. To show that this is a legitimate conjugacy, we must prove that $\iota$ is a homeomorphism. That is, we must show that $\iota$ is one-to one, that it is onto, that it is continuous, and that its inverse is continuous:

**One-to one:** Suppose that $\iota(x) = \iota(y)$ for $x, y \in \Lambda$. This means that $Q_c^n(x)$ and $Q^n(y)$ always lie in the same interval, $I_0$ or $I_1$. Thus the interval $[x, y]$ lies entirely in either $I_0$ or $I_1$ and hence $Q_c$ maps it in one to one fashion onto an interval contained in either $I_0$ or $I_1$. Applying $Q_c$ once more, we conclude that $Q_c^2$ is one-to-one on $[x, y]$. Continuing, we conclude that $Q_c^n$ is one-to-one on the interval $[x, y]$, and we also know that (3.9) implies that the length of $[x, y]$ is increased by a factor of $\lambda^n$. This is impossible unless the length of $[x, y]$ is zero, i.e. $x = y$.

**Onto.** We start with a point $\mathbf{s} = (s_0s_1s_2\dots) \in \Sigma$. We are looking for a point $x$ with $\iota() = \mathbf{s}$. Consider the set of $y \in \Lambda$ such that

$$d(\mathrm{s}, \iota(y)) \leq \frac{1}{2^n}.$$

This is the same as requiring that $y$ belong to

$$\Lambda \cap I_{s_0 s_1 \dots s_n}$$

where $I_{s_0 s_1 \dots s_n}$ is the interval

$$I_{s_0 s_1 \dots s_n} = \{y \in I |\ y \in I_{s_0}, Q_c(y) \in I_{s_1}, \dots Q_c^{\circ n}(y) \in I_{s_n}\}.$$

So

$$\begin{aligned} I_{s_0 s_1 \dots s_n} &= I_{s_0} \cap Q_c^{-1}(I_{s_1}) \cap \dots \cap Q_c^{-n}(I_{s_n}) \\ &= I_{s_0} \cap Q_c^{-1}(I_{s_1} \cap \dots \cap Q_c^{-(n-1)}(I_{s_n})) \\ &= I_{s_0} \cap Q_c^{-1}(I_{s_1 \dots s_n}) \qquad (3.13) \\ &= I_{s_0 s_1 \dots s_{n-1}} \cap Q_c^{-n}(I_{s_n}) \subset I_{s_0 \dots s_{n-1}}. \qquad (3.14) \end{aligned}$$

The inverse image of any interval, $J$ under $Q_c$ consists of two intervals, one lying in $I_0$ and the other lying in $I_1$. For $n = 0$, $I_{s_0}$ is either $I_0$ or $I_1$ and hence is an interval. By induction, it follows from (3.13) that $I_{s_0 s_1 \dots s_n}$ is an interval. By (3.14), these intervals are nested. By construction these nested intervals are closed. Since every sequence of closed nested intervals on the real line has a non-empty intersection, there is a point $x$ which belongs to all of these intervals. Hence all the iterates of $x$ lie in $I$, so $x \in \Lambda$ and $\iota(x) = \mathbf{s}$.

**Continuity.** The above argument shows that the interiors of the intervals $I_{s_0 s_1 \dots s_n}$ (intersected with $\Lambda$) form neighborhoods of $x$ that map into small neighborhoods of $\iota(x)$.

**Continuity of $\iota^{-1}$.** Conversely, any small neighborhood of $x$ in $\Lambda$ will contain one of the intervals $I_{s_0 \dots s_n}$ and hence all of the points t whose first $n$ coordinates agree with $\mathrm{s} = \iota(x)$ will be mapped by $\iota^{-1}$ into the given neighborhood of $x$.

To summarize: we have proved

**Theorem 3.7.1** *Suppose that c satisfies (3.9). Let $\Lambda \subset [-p_+, p_+]$ consist of those points whose images under $Q_c^n$ lie in $[-p,p_+]$ for all $n \geq 0$. Then $\Lambda$ is a closed, non-empty, disconnected set. The itinerary map $\iota$ is a homeomorphism of $\Lambda$ onto the sequence space, $\Sigma$, and conjugates $Q_c$ to the shift map, $S$.*

Just as in the case of the space $X$ in section 3.4, the periodic points for $S$ are precisely the periodic or "repeating" sequences. Thus we can conclude from the theorem that there are exactly $2^n$ points of period (at most) $n$ for $Q_c$. Also, the same argument as in section 3.4 shows that the periodic points for $S$ are dense in $\Sigma$, and hence the periodic points for $Q_c$ are dense in $\Lambda$. Finally, the same argument as in section 3.4 shows that $S$ is transitive on $\Sigma$. Hence, the restriction of $Q_c$ to $\Lambda$ is chaotic.

# Chapter 4

# Space and time averages

## 4.1 histograms and invariant densities

Let us consider a map, $F : [0,1] \to [0,1]$, pick an initial seed, $x_0$, and compute its iterates, $x_0, x_1, x_2, \ldots, x_m$ under $F$. We would like to see which parts of the unit interval are visited by these iterates, and how often. For this purpose let us divide the unit interval up into $N$ subintervals of size $1/N$ given by

$$I_k = \left[\frac{k-1}{N}, \frac{k}{N}\right), \ k = 1, \ldots, N-1, \quad I_N = \left[\frac{N-1}{N}, 1\right].$$

We count how many of the iterates $x_0, x_1, \ldots, x_m$ lie in $I_k$. Call this number $n_k$. There are $m+1$ iterates (starting with, and counting, $x_0$) so the numbers

$$p_k = \frac{n_k}{m+1}$$

add up to one:

$$p_1 + \cdots + p_N = 1.$$

We would like to think of these numbers as "probabilities" - the number $p_k$ representing the "probability" that an iterate belongs to $I_k$. Strictly speaking, we should write $p_k(m)$. In fact, we should write $p_k(m, x_0)$ since the procedure depends on the initial seed, $x_0$. But the hope is that as $m$ gets large the $p_k(m)$ tend to a limiting value which we denote by $p_k$, and that this limiting value will be independent of $x_0$ if $x_0$ is chosen "generically". We will continue in this vague, intuitive vein a while longer before passing to a precise mathematical formulation. If $U$ is a union of some of the $I_k$, then we can write

$$p(U) = \sum_{I_k \subset U} p_k$$

and think of $p(U)$ as representing the "probability" that an iterate of $x_0$ belongs to $U$. If $N$ is large, so the intervals $I_k$ are small, every open set $U$ can be

closely approximated by a union of the $I_k$'s, so we can imagine that the "probabilities", $p(U)$, are defined for all open sets, $U$. If we buy all of this, then we can write down an equation which has some chance of determining what these "probabilities", $p(U)$, actually are: A point $y = F(x)$ belongs to $U$ if and only if $x \in F^{-1}(U)$. Thus the number of points among the $x_1, dots, x_{m+1}$ which belong to $U$ is the same as the number of points among the $x_0, \ldots, x_m$ which belong to $F^{-1}(U)$. Since our limiting probability is unaffected by this shift from 0 to 1 or from $m$ to $m+1$ we get the equation

$$p(U) = p(F^{-1}(U)). \tag{4.1}$$

To understand this equation, let us put it in a more general context. Suppose that we have a "measure", $\mu$, which assigns a size, $\mu(A)$, to every open set, $A$. Let $F$ be a continuous transformation. We then define the **push forward** measure, $F_*\mu$ by

$$(F_*\mu)(A) = \mu(F^{-1}(A)). \tag{4.2}$$

Without developing the language of measure theory, which is really necessary for a full understanding, we will try to describe some of the issues involved in the study of equations (4.2) and (4.1) from a more naive viewpoint. Consider, for example, $F = L_\mu, 1 < \mu < 3$. If we start with any initial seed other than $x_0 = 0$, it is clear that the limiting probability is

$$p(I_k) = 1,$$

if the fixed point,$1 - \frac{1}{\mu} \in I_k$ and

$$p(I_k) = 0$$

otherwise. Similarly, if $3 < \mu < 1 + \sqrt{6}$, and we start with any $x_0$ other than 0 or the fixed point, $1 - \frac{1}{\mu}$ then clearly the limiting probability will be $p(I) = 1$ if both points of period two belong to $I$, $p(I) = \frac{1}{2}$ if $I$ contains exactly one of the two period two points, and $p(I) = 0$ otherwise. These are all examples of **discrete** measures in the sense that there is a finite (or countable) set of points, $\{z_k\}$, each assigned a positive number, $m(z_k)$ and

$$\mu(I) = \sum_{z_k \in I} m(z_k).$$

We are making the implicit assumption that this series converges for every bounded interval. The **integral** of a function, $\phi$, with respect to the discrete measure, $\mu$, denoted by $\langle \phi, \mu \rangle$ or by $\int \phi\mu$ is defined as

$$\int \phi\mu = \sum \phi(x_k) m(x_k).$$

This definition makes sense under the assumption that the series on the right hand side is absolutely convergent. The rule for computing the push forward,

$F_*\mu$ (when defined) is very simple. Indeed, let $\{y_l\}$ be the set of points of the form $y_l = F(x_k)$ for some $k$, and set

$$n(y_l) = \sum_{F(x_k)=y_l} m(x_k).$$

Notice that there is some problem with this definition if there are infinitely many points $x_k$ which map to the same $y_l$. Once again we must make some convergence assumption. For example, if the map $F$ is everywhere finite-to-one, there will be no problem. Thus the push forward of a discrete measure is a discrete measure given by the above formula.

At the other extreme, a measure is called **absolutely continuous** (with respect to Lebesgue measure) if there is an integrable function, $\rho$, called the **density** so that

$$\mu(I) = \int_I \rho(x)dx.$$

For any continuous function, $\phi$ we define the integral of $\phi$ with respect to $\mu$ as

$$\langle \phi, \mu \rangle = \int \phi\mu = \int \phi(x)\rho(x)dx$$

if the integral is absolutely convergent. Suppose that the map $F$ is piecewise differentiable and in fact satisfies $|F'(x)| \neq 0$ except at a finite number of points. These points are called *critical points* for the map $F$ and their images are called *critical values*. Suppose that $A$ is an interval containing no critical values, and to fix the ideas, suppose that $F^{-1}(A)$ is the union of finitely many intervals, $J_l$ each of which is mapped monotonically (either strictly increasing or decreasing) onto $A$. The change of variables formula from ordinary calculus says that for any function $g = g(y)$ we have

$$\int_A g(y)dy = \int_{J_k} g(F(x))|F'(x)|dx,$$

where $y = F(x)$. So if we set $g(y) = \rho(x)|1/F'(x)|$ we get

$$\int \rho(x)\frac{1}{|F'(x)|}dy = \int_{J_k} \rho(x)dx = \mu(J_k).$$

Summing over $k$ and using the definition (4.2) we see that $F_*\mu$ has the density

$$\sigma(y) = \sum_{F(x_k)=y} \frac{\rho(x)}{|F'(x)|}. \tag{4.3}$$

Equation (4.3) is sometimes known as the Perron Frobenius equation, and the transformation $\rho \mapsto \sigma$ as the Perron Frobenius operator.

Getting back to our histogram, if we expect the limit measure to be of the absolutely continuous type, so

$$p(I_k) \approx \rho(x) \times \frac{1}{N}, \quad x \in I_k$$

then we expect that

$$\rho(x) \approx \lim_{m\to\infty} \frac{n_k N}{m+1}, \quad x \in I_k$$

as the formula for the limiting density.

## 4.2 the histogram of $L_4$

We wish to prove the following assertions:
(i) *The measure, $\mu$, with density*

$$\sigma(x) = \frac{1}{\pi\sqrt{x(1-x)}} \tag{4.4}$$

*is invariant under $L_4$.* In other words it satisfies

$$L_{4*}\mu = \mu.$$

(ii) *Up to a multiplicative constant, (4.4) is the only continuous density invariant under $L_4$*
(iii) *If we pick the initial seed generically, then the normalized histogram converges to (4.4).*

We give two proofs of (i). The first is a direct verification of the Perron Frobenius formula (4.3) with $y = F(x) = 4x(1-x)$ so $|F'(x)| = |F'(1-x)| = 4|1-2x|$. Notice that the $\sigma$ given by (4.4) satisfies $\sigma(x) = \sigma(1-x)$ so (4.3) becomes

$$\frac{1}{\pi\sqrt{4x(1-x)(1-4x(1-x))}} = \frac{2}{\pi 4|1-2x|\sqrt{x(1-x)}}.$$

But this follows immediately from the identity

$$1 - 4x(1-x) = (2x-1)^2.$$

For our second proof, consider the tent transformation, $T$. For any interval, $I$ contained in $[0,1]$, $T^{-1}(I)$ consists of the union of two intervals, each of half the length of $I$. In other words the ordinary Lebesgue measure is preserved by the tent transformation: $T_*\nu = \nu$ where $\nu$ has density $\rho(x) \equiv 1$. Put another way, the function $\rho(x) \equiv 1$ is the solution of the Perron Frobenius equation

$$\rho(Tx) = \frac{\rho(x)}{2} + \frac{\rho(1-x)}{2}. \tag{4.5}$$

It follows immediately from the definitions, that

$$(F \circ G)_*\mu = F_*(G_*\mu),$$

where $F$ and $G$ are two transformations , and $\mu$ is a measure. In particular, since $h \circ T = L_4 \circ h$ where

$$h(x) = \sin^2 \frac{\pi x}{2},$$

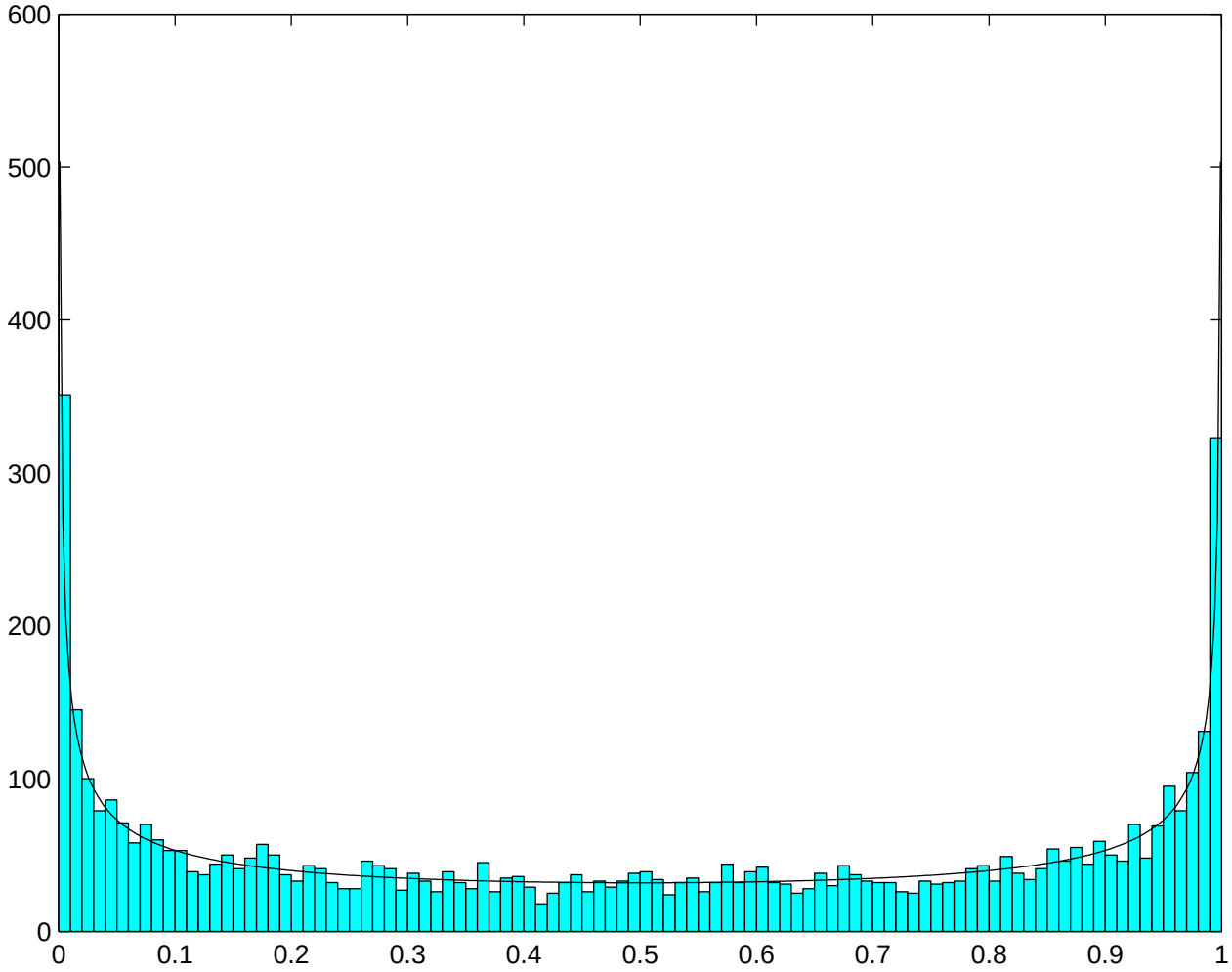


Figure 4.1: The histogram of iterates of $L_4$ compared with $\sigma$.

it follows that if $T_*\nu = \nu$, then $T_*(h_*\nu) = h_*\nu$. So to solve $L_{4*\mu} = \mu$, we must merely compute $h_*\nu$. According to (4.3) this is the measure with density

$$\sigma(y) = \frac{1}{|h'(x)|} = \frac{1}{\pi \sin \frac{\pi x}{2} \cos \frac{\pi x}{2}}.$$

But since $y = \sin^2 \frac{\pi x}{2}$ this becomes

$$\sigma(y) = \frac{1}{\pi\sqrt{y(1-y)}}$$

as desired.

To prove (ii), it is enough to prove the corresponding result for the tent transformation: that $\rho = \text{const.}$ is the only continuous function satisfying (4.5). To see this, let us consider the binary representation of $T$: Let

$$x = 0.a_1a_2a_3\dots$$

be the binary expansion of $x$. If $0 \le x < \frac{1}{2}$, so $a_1 = 0$, then $Tx = 2x$ or

$$T(0.0a_2a_3a_4\dots) = 0.a_2a_3a_4\dots.$$

If $x \ge \frac{1}{2}$, so $a_1 = 1$, then

$$T(x) = -2x + 2 = 1 - (2x - 1) = 1 - S(x) = 1 - 0.a_2a_3a_4\dots.$$

Introducing the notation

$$\bar{0} = 1, \quad \bar{1} = 0,$$

we have

$$0.a_2a_3a_4\cdots + 0.\bar{a_1}\bar{a_2}\bar{a_3}\cdots = 0.1111\cdots = 1$$

so

$$T(0.1a_2a_3a_4\dots) = 0.\bar{a_2}\bar{a_3}\bar{a_4}\dots.$$

In particular, $T^{-1}(0.a_1a_2a_3\dots)$ consists of the two points

$$0.0a_1a_2a_3\dots \quad \text{and} \quad 0.1\bar{a_1}\bar{a_2}\bar{a_3}\dots.$$

Now let us iterate (4.5) with $\rho$ replaced by $f$, and show that the only solution is $f = \text{constant}$. Using the notation $x = 0.a_1a_2\cdots = 0.a$ repeated application of (4.5) gives:

$$\begin{aligned} f(x) &= \frac{1}{2}[f(.0a) + f(.1\bar{a})] \\ &= \frac{1}{4}[f(.00a) + f(.01\bar{a}) + f(.10a) + f(.11\bar{a})] \\ &= \frac{1}{8}[f(.000a) + f(.001\bar{a}) + f(.010a) + f(.011\bar{a}) + f(.100a) + \cdots] \\ &\to \int f(t)dt. \end{aligned}$$

But this integral is a constant, independent of $x$. QED.

The third statement, (iii), about the limiting histogram for "generic" initial seed, $x_0$, demands a more careful formulation. What do we mean by the phrase "generic"? The precise formulation requires a dose of measure theory, the word "generic" should be taken to mean "outside of a set of measure zero with respect to $\mu$". The usual phrase for this is "for almost all $x_0$". Then assertion (iii) becomes a special case of the famous Birkhoff ergodic theorem. This theorem asserts that for almost all points, $p$, the "time average"

$$\lim \frac{1}{n} \sum_{k=0}^{n-1} \phi(L_4^k p)$$

equals the "space average"

$$\int \phi \mu$$

for any integrable function, $\phi$. Rather than proving this theorem, we will explain a simpler theorem, von Neumann's mean ergodic theorem, which motivated Birkhoff to prove his theorem.

Let $F$ be a transformation with an invariant measure, $\mu$. By this we mean that $F_* \mu = \mu$. We let $H$ denote the Hilbert space of all square integrable functions with respect to $\mu$, so the scalar product of $f, g \in H$ is given by

$$(f, g) = \int f \bar{g} \mu.$$

The map $F$ induces a transformation $U : H \to H$ by

$$Uf = f \circ F$$

and

$$(Uf, Ug) = \int (f \circ F)(\overline{g \circ F})\mu = \int f \bar{g} \mu = (f, g).$$

In other words, $U$ is an isometry of $H$. The mean ergodic theorem asserts that the limit of

$$\frac{1}{n} \sum_{0}^{n-1} U^k f$$

exists in the Hilbert space sense, "convergence in mean", rather than the almost everywhere pointwise convergence of the Birkhoff ergodic theorem. Practically by its definition, this limiting element $\hat{f}$ is invariant, i.e. satisfies $U\hat{f} = \hat{f}$. Indeed, applying $U$ to the above sum gives an expression which differs from that sum by only two terms, $f$ and $U^n f$ and dividing by $n$ sends these terms to zero as $n \to \infty$. If, as in our example, we know what the possible invariant elements are, then we know the possible limiting values $\hat{f}$

The mean ergodic theorem can be regarded as a smeared out version of the Birkhoff theorem. Due to inevitable computer error, the mean ergodic theorem may actually be the version that we want.

## 4.3 The mean ergodic theorem

The purpose of this section is to prove

**Theorem 4.3.1** von Neumann's mean ergodic theorem. *Let $U : H \to H$ be an isometry of a Hilbert space, $H$. Then for any $f \in H$ , the limit*

$$\lim \frac{1}{n} \sum U^k f = \hat{f} \tag{4.6}$$

*exists in the Hilbert space sense, and the limiting element $\hat{f}$ is invariant, i.e. $U\hat{f} = \hat{f}$.*

*Proof.* The limit, if it exists, is invariant as we have seen. If $U$ were a unitary operator on a finite dimensional Hilbert space, $H$, then we could diagonalize $U$, and hence reduce the theorem to the one dimensional case. A unitary operator on a one dimensional space is just multiplication by a complex number of the form $e^{i\alpha}$. If $e^{i\alpha} \neq 1$, then

$$\frac{1}{n}(1 + e^{i\alpha} + \cdots + e^{(n-1)i\alpha}) = \frac{1}{n} \frac{1 - e^{in\alpha}}{1 - e^{i\alpha}} \to 0.$$

On the other hand, if $e^{i\alpha} = 1$, the expression on the left is identically one. This proves the theorem for finite dimensional unitary operators. For an infinite dimensional Hilbert space, we could apply the spectral theorem of Stone (discovered shortly before the proof of the ergodic theorem) and this was von Neumann's original method of proof.

Actually, we can proceed as follows:

**Lemma 4.3.1** *The orthogonal complement of the set, $D$, of all elements of the form $Ug - g$, consists of invariant elements.*

*Proof.* If $f$ is orthogonal to all elements in $D$, then,in particular, $f$ is orthogonal to $Uf - f$, so

$$0 = (f, Uf - f)$$

and

$$(Uf, Uf - f) = (Uf, Uf) - (Uf, f) = (f, f) - (Uf, f)$$

since $U$ is an isometry. So

$$(Uf, Uf - f) = (f - Uf, f) = 0.$$

So

$$(Uf - f, Uf - f) = (Uf, Uf - f) - (f, Uf - f) = 0,$$

or

$$Uf - f = 0$$

which says that $f$ is invariant.

So what we have shown, in fact, is

**Lemma 4.3.2** *The union of the set $D$ with the set, $I$, of the invariant functions is dense in $H$.*

In fact, if $f$ is orthogonal to $D$, then it must be invariant, and if it is orthogonal to all invariant functions it must be orthogonal to itself, and so must be zero. So $(D \cup I)^{\perp} = 0$, so $D \cup I$ is dense in $H$.

Now if $f$ is invariant, then clearly the limit(4.6) exists and equals $f$. If $f = Ug - g$, then the expression on the left in (4.6) telescopes into

$$\frac{1}{n}(U^n g - g)$$

which clearly tends to zero. Hence, as a corollary we obtain

**Lemma 4.3.3** *The set of elements for which the limit in (4.6) exists is dense in $H$.*

Hence the mean ergodic theorem will be proved, once we prove

**Lemma 4.3.4** *The set of elements for which the limit in (4.6) exists is closed.*

*Proof.* If

$$\frac{1}{n}\sum U^k g_i \to \hat{g}_i, \quad \frac{1}{n}\sum U^k g_j \to \hat{g}_j,$$

and

$$\| g_i - g_j \| < \epsilon,$$

then

$$\| \frac{1}{n}\sum U^k g_i - \frac{1}{n}\sum U^k g_j \| < \epsilon,$$

so

$$\| \hat{g}_i - \hat{g}_j \| < \epsilon.$$

So if $\{g_i\}$ is a sequence of elements converging to $f$, we conclude that $\{\hat{g}_i\}$ converges to some element, call it $\hat{f}$. If we choose $i$ sufficiently large so that $\| g_i - f \| < \epsilon$, then

$$\| \frac{1}{n}\sum U^k f - \hat{f} \| \leq \| \frac{1}{n}\sum U^k (f - g_i) \| + \| \frac{1}{n}\sum U^k g_i - \hat{g}_i \| + \| \hat{g}_i - \hat{f} \| \leq 3\epsilon,$$

proving the lemma and hence proving the mean ergodic theorem.

## 4.4 the arc sine law

The probability distribution with density

$$\sigma(x) = \frac{1}{\pi\sqrt{x(1-x)}}$$

is called the *arc sine law* in probability theory because, if $I$ is the interval $I = [0, u]$ then

$$\text{Prob } x \in I = \text{Prob } 0 \leq x \leq u = \int_0^u \frac{1}{\pi\sqrt{x(1-x)}} = \frac{2}{\pi} \arcsin \sqrt{u}. \tag{4.7}$$

We have already verified this integration because $I = h(J)$ where

$$h(t) = \sin^2 \frac{\pi t}{2}, \quad J = [0, v], \ h(v) = u,$$

and the probability measure we are studying is the push forward of the uniform distribution. So

$$\text{Prob } h(t) \in I = \text{Prob } t \in J = v.$$

The arc sine law plays a crucial role in the theory of fluctuations in random walks. As a cultural diversion we explain some of the key ideas, following the treatment in Feller [**?**] very closely.

Suppose that there is an ideal coin tossing game in which each player wins or loses a unit amount with (independent) probability $\frac{1}{2}$ at each throw. Let $S_0 = 0, S_1, S_2, \ldots$ denote the successive cumulative gains (or losses) of the first player. We can think of the values of these cumulative gains as being marked off on a vertical $s$-axis, and representing the position of a particle which moves up or down with probability $\frac{1}{2}$ at each (discrete) time unit . Let

$$\alpha_{2k,2n}$$

denote the *probability that up to and including time* $2n$*, the last visit to the origin occurred at time* $2k$. Let

$$u_{2\nu} = \left( \begin{array}{c} 2\nu \\ \nu \end{array} \right) 2^{-2\nu}. \tag{4.8}$$

So $u_{2\nu}$ represents the probability that exactly $\nu$ out of the first $2\nu$ steps were in the positive direction, and the rest in the negative direction. In other words, $u_{2\nu}$ is the probability that the particle has returned to the origin at time $2\nu$. We can find a simple approximation to $u_{2\nu}$ using Stirling's formula for an approximation to the factorial:

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$$

where the $\sim$ signifies that the ratio of the two sides tends to one as $n$ tends to infinity. For a proof of Stirling's formula cf. [**?**]. Then

$$\begin{aligned} u_{2\nu} &= 2^{-2\nu} \frac{(2\nu)!}{(\nu!)^2} \\ &\sim 2^{-2\nu} \frac{\sqrt{2\pi}(2\nu)^{2\nu+\frac{1}{2}} e^{-2\nu}}{2\pi\nu^{2\nu+1} e^{-2\nu}} \\ &= \frac{1}{\sqrt{\pi\nu}}. \end{aligned}$$

The results we wish to prove in this section are

**Proposition 4.4.1** *We have*

$$\alpha_{2k,2n} = u_{2k}u_{2n-2k}, \tag{4.9}$$

*so we have the asymptotic approximation*

$$\alpha_{2k,2n} \sim \frac{1}{\pi\sqrt{k(n-k)}}. \tag{4.10}$$

*If we set*

$$x_k = \frac{k}{n}$$

*then we can write*

$$\alpha_{2k,2n} \sim \frac{1}{n}\sigma(x_k). \tag{4.11}$$

*Thus, for fixed* $0 < x < 1$ *and* $n$ *sufficiently large*

$$\sum_{k<xn} \alpha_{2k,2n} \doteq \frac{2}{\pi}\arcsin\sqrt{x}. \tag{4.12}$$

**Proposition 4.4.2** *The probability that in the time interval from* 0 *to* $2n$ *the particle spends* $2k$ *time units on the positive side and* $2n-2k$ *time units on the negative side equals* $\alpha_{2k,2n}$. *In particular, if* $0 < x < 1$ *the probability that the fraction* $k/n$ *of time units spent on the positive be less than* $x$ *tends to* $\frac{2}{\pi}\arcsin\sqrt{x}$ *as* $n \to \infty$.

Let us call the value of $S_{2n}$ for any given realization of the random walk, the *terminal point*. Of course, the particle may well have visited this terminal point earlier in the walk, and we can ask when it first reaches its terminal point.

**Proposition 4.4.3** *The probability that the first visit to the terminal point occurs at time* $2k$ *is given by* $\alpha_{2k,2n}$.

We can also ask for the first time that the particle reaches its maximum value: We say that the *first maximum occurs at time* $l$ if

$$S_0 < S_l, S_1 < S_l, \ldots S_{l-1} < S_l, \qquad S_{l+1} \le S_l, S_{l+2} \le S_l, \ldots S_{2n} \le S_l. \tag{4.13}$$

**Proposition 4.4.4** *If* $0 < l < 2n$ *the probability that the first maximum occurs at* $l = 2k$ *or* $l = 2k+1$ *is given by* $\frac{1}{2}\alpha_{2k,2n}$. *For* $l = 0$ *this probability is given by* $u_{2n}$ *and if* $l = 2n$ *it is given by* $\frac{1}{2}u_{2n}$.

Before proving these four propositions, let us discuss a few of their implications which some people find counterintuitive. For example, because of the shape of the density, $\sigma$, the last proposition implies that the maximal accumulated gain is much more likely to occur very near to the beginning or to the end of a coin tossing game rather than somewhere in the middle. The third proposition implies that the probability that the first visit to the terminal point occurs at time $2k$ is that same as the probability that it occurs at time $2n-2k$

and that very early first visits and very late first visits are much more probable than first visits some time in the middle.

In order to get a better feeling for the assertion of the first two propositions let us tabulate the values of $\frac{2}{\pi}\arcsin\sqrt{x}$ for $0 \leq x \leq \frac{1}{2}$.

| $x$ | $\frac{2}{\pi}\arcsin\sqrt{x}$ | $x$ | $\frac{2}{\pi}\arcsin\sqrt{x}$ |
|---|---|---|---|
| 0.05 | 0.144 | 0.30 | 0.369 |
| 0.10 | 0.205 | 0.35 | 0.403 |
| 0.15 | 0.253 | 0.40 | 0.236 |
| 0.20 | 0.295 | 0.45 | 0.468 |
| 0.25 | 0.333 | 0.50 | 0.500 |

This table, in conjunction with Prop. 4.4.1 says that if a great many coin tossing games are conducted every second, day and night for a hundred days, then in about 14.4 percent of the cases,the lead will not change after day five.

The proof of all four propositions hinges on three lemmas. Let us graph (by a polygonal path) the walk of a particle. So a "path" is a broken line segment made up of segments of slope $\pm 1$ joining integral points to integral points in the plane (with the time or $t-$axis horizontal and the $s-$axis vertical). If $A = (a, \alpha)$ is a point, we let $A' = (a, -\alpha)$ denote its image under reflection in the $t-$axis.

**Lemma 4.4.1 The reflection principle.** *Let $A = (a, \alpha), B = (b, \beta)$ be points in the first quadrant with $b > a \geq 0, \alpha > 0, \beta > 0$. The number of paths from $A$ to $B$ which touch or cross the $t-$ axis equals the number of all paths from $A'$ to $B$.*

**Proof.** For any path from $A$ to $B$ which touches the horizontal axis, let $t$ be the abscissa of the first point of contact. Reflect the portion of the path from $A$ to $T = (t, 0)$ relative to the horizontal axis. This reflected portion is a path from $A'$ to $T$, and continues to give a path from $A'$ to $B$. This procedure assigns to each path from $A$ to $B$ which touches the axis, a path from $A'$ to $B$. This assignment is bijective: Any path from $A'$ to $B$ must cross the $t-$axis. Reflect the portion up to the first crossing to get a touching path from $A$ to $B$. This is the inverse assignment. QED

A path with $n$ steps will join $(0, 0)$ to $(n, x)$ if an only if it has $p$ steps of slope $+1$ and $q$ steps of slope $-1$ where

$$p + q = n, \quad p - q = x.$$

The number of such paths is the number of ways of picking the positions of the $p$ steps of positive slope and so the number of paths joining $(0, 0)$ to $(n, x)$ is

$$N_{n,x} = \left( \begin{array}{c} p+q \\ p \end{array} \right) = \left( \begin{array}{c} n \\ \frac{n+x}{2} \end{array} \right).$$

It is understood that this formula means that $N_{n,x} = 0$ when there are no paths joining the origin to $(n, x)$.

**Lemma 4.4.2 The ballot theorem.** *Let $n$ and $x$ be positive integers. There are exactly*

$$\frac{x}{n}N_{n,x}$$

*paths which lie strictly above the $t$ axis for $t > 0$ and join $(0,0)$ to $(n,x)$.*

**Proof.** There are as many such paths as there are paths joining $(1,1)$ to $(n,x)$ which do *not* touch or cross the $t-$axis. This is the same as the total number of paths which join $(1,1)$ to $(n,x)$ less the number of paths which do touch or cross. By the preceding lemma, the number of paths which do cross is the same as the number of paths joining $1.-1)$ to $(n,x)$ which is $N_{n-1,x+1}$. Thus, with $p$ and $q$ as above, the number of paths which lie strictly above the $t$ axis for $t > 0$ and which joint $(0,0)$ to $(n,x)$ is

$$\begin{aligned} N_{n-1,x-1} - N_{n-1,x+1} &= \left( \begin{array}{c} p+q-1 \\ p-1 \end{array} \right) - \left( \begin{array}{c} p+q-1 \\ p \end{array} \right) \\ &= \frac{(p+q-1)!}{(p-1)!(q-1)!}\left[\frac{1}{q} - \frac{1}{p}\right] \\ &= \frac{p-q}{p+q} \times \frac{(p+q)!}{p!q!} \\ &= \frac{x}{n}N_{n,x} \quad \text{QED} \end{aligned}$$

The reason that this lemma is called the Ballot Theorem is that it asserts that if candidate P gets $p$ votes, and candidate Q gets $q$ votes in an election where the probability of each vote is independently $\frac{1}{2}$, then the probability that throughout the counting there are more votes for P than for Q is given by

$$\frac{p-q}{p+q}.$$

Here is our last lemma:

**Lemma 4.4.3** *The probability that from time 1 to time $2n$ the particle stays strictly positive is given by $\frac{1}{2}u_{2n}$. In symbols,*

$$\text{Prob } \{S_1 > 0, \ldots, S_{2n} > 0\} = \frac{1}{2}u_{2n}. \tag{4.14}$$

*So*

$$\text{Prob } \{S_1 \neq 0, \ldots, S_{2n} \neq 0\} = u_{2n}. \tag{4.15}$$

*Also*

$$\text{Prob } \{S_1 \geq 0, \ldots, S_{2n} \geq 0\} = u_{2n}. \tag{4.16}$$

**Proof.** By considering the possible positive values of $S_{2n}$ which can range from

2 to $2n$ we have

$$\begin{aligned}
\text{Prob}\,\{S_1 > 0, \ldots, S_{2n} > 0\} &= \sum_{r=1}^{n} \text{Prob}\,\{S_1 > 0, \ldots, S_{2n} = 2r\} \\
&= 2^{-2n} \sum_{r=1}^{n} (N_{2n-1,2r-1} - N_{2n-1,2r+1}) \\
&= 2^{-2n} \left(N_{2n-1,1} - N_{2n-1,3} + N_{2n-1,3} - N_{2n-1,5} + \cdots\right) \\
&= 2^{-2n} N_{2n-1,1} \\
&= \frac{1}{2} p_{2n-1,1} \\
&= \frac{1}{2} u_{2n}.
\end{aligned}$$

The passage from the first line to the second is the reflection principle,as in our proof of the Ballot Theorem, from the third to the fourth is because the sum telescopes. The $p_{2n-1,1}$ on the next to the last line is the probability of ending up at $(2n-1, 1)$ starting from $(0, 0)$. The last equality is simply the assertion that to reach zero at time $2n$ we must be at $\pm 1$ at time $2n-1$ (each of these has equal probability, $p_{2n-1,1}$) and for each alternative there is a 50 percent chance of getting to zero on the next step. This proves (4.14). Since a path which never touches the $t-$axis must be always above or always below the $t-$axis, (4.15) follows immediately from (4.14). As for (4.16), observe that a path which is strictly above the axis from time 1 on, must pass through the point $(1, 1)$ and then stay above the horizontal line $s = 1$. The probability of going to the point $(1, 1)$ at the first step is $\frac{1}{2}$, and then the probability of remaining above the new horizontal axis is Prob $\{S_1 \geq 0, \ldots, S_{2n-1} \geq 0\}$. But since $2n-1$ is odd, if $S_{2n-1} \geq 0$ then $S_{2n} \geq 0$. So, by (4.14) we have

$$\begin{aligned}
\frac{1}{2} u_{2n} &= \text{Prob}\,\{S_1 > 0, \ldots, S_{2n} > 0\} \\
&= \frac{1}{2} \text{Prob}\,\{S_1 \geq 0, \ldots, S_{2n-1} \geq 0\} \\
&= \frac{1}{2} \text{Prob}\,\{S_1 \geq 0, \ldots, S_{2n-1} \geq 0, S_{2n} \geq 0\},
\end{aligned}$$

completing the proof of the lemma.

We can now turn to the proofs of the propositions.

**Proof of Prop.4.4.1** To say that the last visit to the origin occurred at time $2k$ means that

$$S_{2k} = 0$$

and

$$S_j \neq 0, \quad j = 2k+1, \ldots, 2n.$$

By definition, the first $2k$ positions can be chosen in $2^{2k} u_{2k}$ ways to satisfy the first of these conditions. Taking the point $(2k, 0)$ as our new origin, (4.15) says that there are $2^{2n-2k} u_{2n-2k}$ ways of choosing the last $2n-2k$ steps so as to

satisfy the second condition. Multiplying and then dividing the result by $2^{2n}$ proves Prop.4.4.1.

**Proof of Prop.4.4.2.** We consider paths of $2n$ steps and let $b_{2k,2n}$ denote the probability that exactly $2k$ sides lie above the $t-$axis. Prop. 4.4.2 asserts that

$$b_{2k,2n} = \alpha_{2k,2n}.$$

For the case $k = n$ we have $\alpha_{2n,2n} = u_0 u_{2n} = u_{2n}$ and $b_{2n,2n}$ is the probability that the path lies entirely above the axis. So our assertion reduces to (4.16) which we have already proved. By symmetry, the probability of the path lying entirely below the the axis is the same as the probability of the path lying entirely above it, so $b_{0,2n} = \alpha_{0,2n}$ as well. So we need prove our assertion for $1 \leq k \leq n-1$. In this situation, a return to the origin must occur. Suppose that the first return to the origin occurs at time $2r$. There are then two possibilities: the entire path from the origin to $(2r, 0)$ is either above the axis or below the axis. If it is above the axis, then $r \leq k \leq n-1$, and the section of path beyond $(2r, 0)$ has $2k - 2r$ edges above the $t-$axis. The number of such paths is

$$\frac{1}{2} 2^{2r} f_{2r} 2^{2n-2r} b_{2k-2r,2n-2r}$$

where $f_{2r}$ denotes the *probability of first return* at time $2r$:

$$f_{2r} = \text{Prob } \{S_1 \neq 0, \ldots, S_{2r-1} \neq 0, S_{2r} = 0\}.$$

If the first portion of the path up to $2r$ is spent below the axis, the the remaining path has exactly $2k$ edges above the axis, so $n - r \geq k$ and the number of such paths is

$$\frac{1}{2} 2^{2r} f_{2r} 2^{2n-2r} b_{2k,2n-2r}.$$

So we get the recursion relation

$$b_{2k,2n} = \frac{1}{2} \sum_{r=1}^{k} f_{2r} b_{2k-2r,2n-2r} + \frac{1}{2} \sum_{r=1}^{n-k} f_{2r} b_{2k,2n-2r} \quad 1 \leq k \leq n-1. \tag{4.17}$$

Now we proceed by induction on $n$. We know that $b_{2k,2n} = u_{2k} u_{2n-2k} = \frac{1}{2}$ when $n = 1$. Assuming the result up through $n-1$, the recursion formula (4.17) becomes

$$b_{2k,2n} = \frac{1}{2} u_{2n-2k} \sum_{r=1}^{k} f_{2r} u_{2k-2r} + \frac{1}{2} u_{2k} \sum_{r=1}^{n-k} f_{2r} u_{2n-2k-2r}. \tag{4.18}$$

But we claim that the probabilities of return and the probabilities of first return are related by

$$u_{2n} = f_2 u_{2n-2} + f_4 u_{2n-4} + \cdots + f_{2n} u_0. \tag{4.19}$$

Indeed, if a return occurs at time $2n$, then there must be a first return at some time $2r \leq 2n$ and then a return in $2n-2r$ units of time, and the sum in (4.19) is

over the possible times of first return. If we substitute (4.19) into the first sum in (4.18) it becomes $u_{2k}$ while substituting (4.19) into the second term yields $u_{2n-2k}$. Thus (4.18) becomes

$$b_{2k,2n} = u_{2k}u_{2n-2k}$$

which is our desired result.

**Proof of Prop.4.4.3.** This follows from Prop.4.4.1 because of the symmetry of the whole picture under rotation through 180° and a shift: The probability in the lemma is the probability that $S_{2k} = S_{2n}$ but $S_j \neq S_{2n}$ for $j < 2k$. Reading the path rotated through 180° about the end point, and with the endpoint shifted to the origin, this is clearly the same as the probability that $2n - 2k$ is the last visit to the origin. QED

**Proof of Prop 4.4.4.** The probability that the maximum is achieved at 0 is the probability that $S_1 \leq 0, \ldots, S_{2n} \leq 0$ which is $u_{2n}$ by (4.16). The probability that the maximum is first obtained at the terminal point, is, after rotation and translation, the same as the probability that $S_1 > 0, \ldots, S_{2n} > 0$ which is $\frac{1}{2}u_{2n}$ by (4.14). If the maximum occurs first at some time $l$ in the middle, we combine these results for the two portions of the path - before and after time $\ell$ - together with (4.9) to complete the proof. QED

## 4.5 The Beta distributions.

The arc sine law is the special case $a = b = \frac{1}{2}$ of the Beta distribution with parameters $a, b$ which has probability density proportional to

$$t^{a-1}(1-t)^{b-1}.$$

So long as $a > 0$ and $b > 0$ the integral

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

converges, and was evaluated by Euler to be

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

where $\Gamma$ is Euler's Gamma function. So the Beta distributions with $A > 0, b > 0$ are given by

$$\frac{1}{B(a,b)}t^{a-1}(1-t)^{b-1}.$$

We characterized the arc sine law ($a = b = \frac{1}{2}$) as being the unique probability density invariant under $L_4$. The case $a = b = 0$, where the integral does not converge, also has an interesting characterization as an invariant density. Consider transformations of the form

$$t \mapsto \frac{at+b}{ct+d}$$

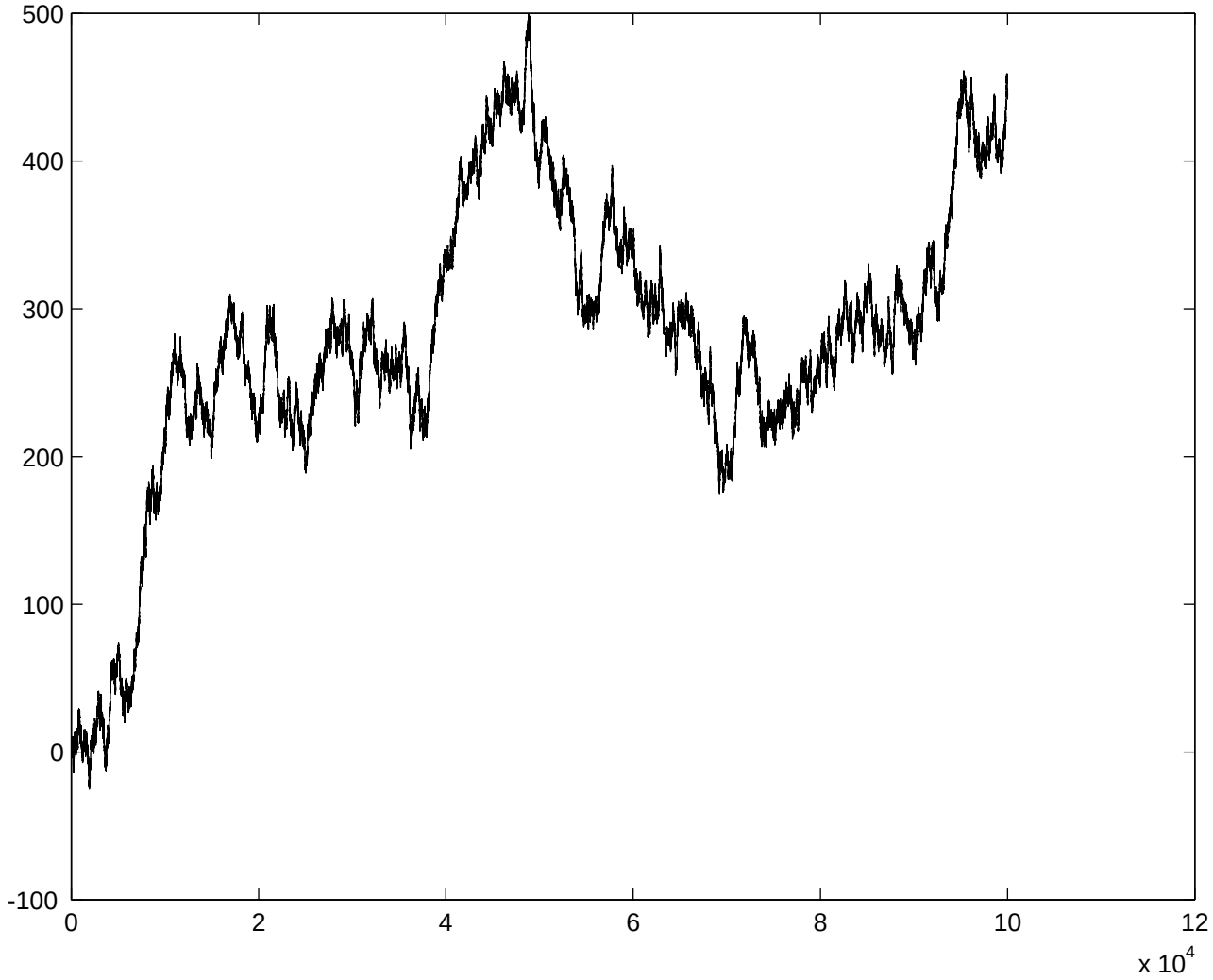


Figure 4.2: A random walk with 100,000 steps. The last zero is at time 3783. For the remaining 96,217 steps the path is positive. According to the arc sine law, with probability 1/5, the particle will spend about 97.6 percent of its time on one side of the origin.

where the matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

is invertible. Suppose we require that the transformation preserve the origin and the point $t = 1$. Preserving the origin requires that $b = 0$, while preserving the ' point $t = 1$ requires that $a = c + d$. Since $b = 0$ we must have $ad \neq 0$ for the matrix to be invertible. Since multi[lying all the entries of the matrix by the same non-zero scalar does not change the transformation, we may as well assume that $d = 1$, and hence the family transformations we are looking at are

$$\phi_a : \quad t \mapsto \frac{at}{(a-1)t+1}, \qquad a \neq 0.$$

Notice that

$$\phi_a \circ \phi_b = \phi_{ab}.$$

Our claim is that, up to scalar multiple, the density

$$\rho(t) = \frac{1}{t(1-t)}$$

is the unique density such that the measure

$$\rho(t)dt$$

is invariant under all the transformations $\phi_a$. Indeed,

$$\phi_a'(t) = \frac{a}{[1-t+at]^2}$$

so the condition of invariance is

$$\frac{a}{[1-t+at]^2}\rho(\phi_a(t)) = \rho(t).$$

Let us normalize $\rho$ by

$$\rho\left(\frac{1}{2}\right) = 4.$$

Then

$$s = \phi_a\left(\frac{1}{2}\right) \Leftrightarrow s = \frac{a}{1+a} \Leftrightarrow a = \frac{s}{1-s}.$$

So taking $t = \frac{1}{2}$ in the condition for invariance and $a$ as above, we get

$$\rho(s) = 4((1-s)/s)[\frac{1}{2} + \frac{1}{2}\frac{s}{1-s}]^2 = \frac{1}{s(1-s)}.$$

This elementary geometrical fact - that $1/t(1-t)$ is the unique density (up to scalar multiple) which is invariant under all the $\phi_a$ - was given a deep philosophical interpretation by Jaynes, [**?**]:

Suppose we have a possible event which may or may not occur, and we have a population of individuals each of whom has a clear opinion (based on ingrained prejudice, say, from reading the newspapers or watching television) of the probability of the event being true. So Mr. A assigns probability p(A) to the event E being true and (1-p(A)) as the probability of its not being rue, while Mr. B assigns probability P(B) to its being true and (1-p(B)) to its not being true and so on.

Suppose an additional piece of information comes in, which would have a (conditional) probability $x$ of being generated if E were true and $y$ of this information being generated if $E$ were not true. We assume that both $x$ and $y$ are positive, and that every individual thinks rationally in the sense that on the advent of this new information he changes his probability estimate in accordance with Bayes' law, which says that the posterior probability $p'$ is given in terms of the prior probability $p$ by

$$p' = \frac{px}{px + (1-p)y} \quad = \phi_a(p) \quad \text{where} \quad a := \frac{x}{y}.$$

We might say that the population as a whole has been invariantly prejudiced if any such additional evidence does not change the proportion of people within the population whose belief lies in a small interval. Then the density describing this state of knowledge (or rather of ignorance) must be the density

$$\rho(p) = \frac{1}{p(1-p)}.$$

According to this reasoning of Jaynes, we take the above density to describe the prior probability an individual (thought of as a population of subprocessors in his brain) would assign to the probability of an outcome of a given experiment. If a series of experiments then yielded $M$ successes and $N$ failures,Bayes' theorem (in its continuous version) would then yield the posterior distribution of probability assignments as being proportional to

$$p^{M-1}(1-p)^{N-1}$$

the Beta distribution with parameters $M, N$.

# Chapter 5

# The contraction fixed point theorem

## 5.1 Metric spaces

Until now we have used the notion of metric quite informally. It is time for a formal definition. For any set $X$, we let $X \times X$ (called the Cartesian product of $X$ with itself) denote the set of all ordered pairs of elements of $X$. (More generally, if $X$ and $Y$ are sets, we let $X \times Y$ denote the set of all pairs $(x, y)$ with $x \in$ and $y \in Y$, and is called the Cartesian product of $X$ with $Y$.)

A **metric** for a set $X$ is a function $d$ from $X$ to the real numbers $\mathbf{R}$,

$$d : X \times X \to \mathbf{R}$$

such that for all $x, y, z \in X$

1. $d(x, y) = d(y, x)$
2. $d(x, z) \leq d(x, y) + d(y, z)$
3. $d(x, x) = 0$
4. If $d(x, y) = 0$ then $x = y$.

The inequality in 2) is known as the **triangle inequality** since if $X$ is the plane and $d$ the usual notion of distance, it says that the length of an edge of a triangle is at most the sum of the lengths of the two other edges. (In the plane, the inequality is strict unless the three points lie on a line.)

Condition 4) is in many ways inessential, and it is often convenient to drop it, especially for the purposes of some proofs. For example, we might want to consider the decimal expansions $.49999\ldots$ and $.50000\ldots$ as different, but as having zero distance from one another. Or we might want to "identify" these two decimal expansions as representing the same point.

A function $d$ which satisfies only conditions 1) - 3) is called a **pseudo-metric**.

A **metric space** is a pair $(X, d)$ where $X$ is a set and $d$ is a metric on $X$. Almost always, when $d$ is understood, we engage in the abuse of language and speak of "the metric space $X$".

Similarly for the notion of a **pseudo-metric space**.

In like fashion, we call $d(x, y)$ the **distance** between $x$ and $y$, the function $d$ being understood.

If $r$ is a positive number and $x \in X$, the (open) **ball of radius** $r$ about $x$ is defined to be the set of points at distance less than $r$ from $x$ and is denoted by $B_r(x)$. In symbols,

$$B_r(x) := \{y|\ d(x, y) < r\}.$$

If $r$ and $s$ are positive real numbers and if $x$ and $z$ are points of a pseudo-metric space $X$, it is possible that $B_r(x) \cap B_s(z) = \emptyset$. This will certainly be the case if $d(x, z) > r + s$ by virtue of the triangle inequality. Suppose that this intersection is not empty and that

$$w \in B_r(x) \cap B_s(z).$$

If $y \in X$ is such that $d(y, w) < \min[r - d(x, w), s - d(z, w)]$ then the triangle inequality implies that $y \in B_r(x) \cap B_s(z)$. Put another way, if we set $t := \min[r - d(x, w), s - d(z, w)]$ then

$$B_t(w) \subset B_r(x) \cap B_s(z).$$

Put still another way, this says that the intersection of two (open) balls is either empty or is a union of open balls. So if we call a set in $X$ **open** if either it is empty, or is a union of open balls, we conclude that the intersection of any finite number of open sets is open, as is the union of any number of open sets. In technical language, we say that the open balls form a base for a topology on $X$.

A map $f : X \to Y$ from one pseudo-metric space to another is called **continuous** if the inverse image under $f$ of any open set in $Y$ is an open set in $X$. Since an open set is a union of balls, this amounts to the condition that the inverse image of an open ball in $Y$ is a union of open balls in $X$, or, to use the familiar $\epsilon, \delta$ language, that if $f(x) = y$ then for every $\epsilon > 0$ there exists a $\delta = \delta(x, \epsilon) > 0$ such that

$$f(B_\delta(x)) \subset B_\epsilon(y).$$

Notice that in this definition $\delta$ is allowed to depend both on $x$ and on $\epsilon$. The map is called uniformly continuous if we can choose the $\delta$ independently of $x$.

An even stronger condition on a map from one pseudo-metric space to another is the **Lipschitz condition**. A map $f : X \to Y$ from a pseudo-metric space $(X, d_X)$ to a pseudo-metric space $(Y, d_Y)$ is called a **Lipschitz map** with **Lipschitz constant** $C$ if

$$d_Y(f(x_1), f(x_2)) \leq C d_X(x_1, x_2) \quad \forall x_1, x_2 \in X.$$

Clearly a Lipschitz map is uniformly continuous.

For example, suppose that $A$ is a fixed subset of a pseudo-metric space $X$. Define the function $d(A, \cdot)$ from $X$ to $\mathbf{R}$ by

$$d(A, x) := \inf\{d(x, w),\ w \in A\}.$$

The triangle inequality says that

$$d(x, w) \leq d(x, y) + d(y, w)$$

for all $w$, in particular for $w \in A$, and hence taking lower bounds we conclude that

$$d(A, x) \leq d(x, y) + d(A, y).$$

or

$$d(A, x) - d(A, y) \leq d(x, y).$$

Reversing the roles of $x$ and $y$ then gives

$$|d(A, x) - d(A, y)| \leq d(x, y).$$

Using the standard metric on the real numbers where the distance between $a$ and $b$ is $|a - b|$ this last inequality says that $d(A, \cdot)$ is a Lipschitz map from $X$ to $\mathbf{R}$ with $C = 1$.

A closed set is defined to be a set whose complement is open. Since the inverse image of the complement of a set (under a map $f$) is the complement of the inverse image, we conclude that the inverse image of a closed set under a continuous map is again closed.

For example, the set consisting of a single point in $\mathbf{R}$ is closed. Since the map $d(A, \cdot)$ is continuous, we conclude that the set

$$\{x | d(A, x) = 0\}$$

consisting of all point at zero distance from $A$ is a closed set. It clearly is a closed set which contains $A$. Suppose that $S$ is some closed set containing $A$, and $y \notin S$. Then there is some $r > 0$ such that $B_r(y)$ is contained in the complement of $C$, which implies that $d(y, w) \geq r$ for all $w \in S$. Thus $\{x | d(A, x) = 0\} \subset S$. In short $\{x | d(A, x) = 0\}$ is a closed set containing $A$ which is contained in all closed sets containing $A$. This is the definition of the **closure** of a set, which is denoted by $\overline{A}$. We have proved that

$$\overline{A} = \{x | d(A, x) = 0\}.$$

In particular, the closure of the one point set $\{x\}$ consists of all points $u$ such that $d(u, x) = 0$.

Now the relation $d(x, y) = 0$ is an equivalence relation, call it $R$. (Transitivity being a consequence of the triangle inequality.) This then divides the space $X$ into equivalence classes, where each equivalence class is of the form $\overline{\{x\}}$, the closure of a one point set. If $u \in \overline{\{x\}}$ and $v \in \overline{\{y\}}$ then

$$d(u, v) \leq d(u, x) + d(x, y) + d(y, v) = d(x, y).$$

since $x \in \overline{\{u\}}$ and $y \in \overline{\{v\}}$ we obtain the reverse inequality, and so

$$d(u, v) = d(x, y).$$

In other words, we may define the distance function on the quotient space $X/R$, i.e. on the space of equivalence classes by

$$d(\overline{\{x\}}, \overline{\{y\}}) := d(u, v), \quad u \in \overline{\{x\}}, v \in \overline{\{y\}}$$

and this does not depend on the choice of $u$ and $v$. Axioms 1)-3) for a metric space continue to hold, but now

$$d(\overline{\{x\}}, \overline{\{y\}}) = 0 \Rightarrow \overline{\{x\}} = \overline{\{y\}}.$$

In other words, $X/R$ is a *metric* space. Clearly the projection map $x \mapsto \overline{\{x\}}$ is an isometry of $X$ onto $X/R$. (An isometry is a map which preserves distances.) In particular it is continuous. It is also open.

In short, we have provided a canonical way of passing (via an isometry) from a pseudo-metric space to a metric space by identifying points which are at zero distance from one another.

A subset $A$ of a pseudo-metric space $X$ is called *dense* if its closure is the whole space. From the above construction, the image $A/R$ of $A$ in the quotient space $X/R$ is again dense. We will use this fact in the next section in the following form:

*If $f : Y \to X$ is an isometry of $Y$ such that $f(Y)$ is a dense set of $X$, then $f$ descends to a map $F$ of $Y$ onto a dense set in the metric space $X/R$.*

## 5.2 Completeness and completion.

The usual notion of convergence and Cauchy sequence go over unchanged to metric spaces or pseudo-metric spaces $Y$. A sequence $\{y_n\}$ is said to **converge** to the point $y$ if for every $\epsilon > 0$ there exists an $N = N(\epsilon)$ such that

$$d(y_n, y) < \epsilon \quad \forall\ n > N.$$

A sequence $\{y_n\}$ is said to be **Cauchy** if for any $\epsilon > 0$ there exists an $N = N(\epsilon$ such that

$$d(y_n, y_m) < \epsilon \quad \forall\ m, n > N.$$

The triangle inequality implies that every convergent sequence is Cauchy. But not every Cauchy sequence is convergent. For example, we can have a sequence of rational numbers which converge to an irrational number, as in the approximation to the square root of 2. So if we look at the set of rational numbers as a metric space $R$ in its own right, not every Cauchy sequence of rational numbers converges in $R$. We must "complete" the rational numbers to obtain $\mathbf{R}$, the set of real numbers. We want to discuss this phenomenon in general.

So we say that a (pseudo-)metric space is **complete** if every Cauchy sequence converges. The key result of this section is that we can always "complete" a metric or pseudo-metric space. More precisely, we claim that

*Any metric (or pseudo-metric) space can be mapped by a one to one isometry onto a dense subset of a complete metric (or pseudo-metric) space.*

By the italicized statement of the preceding section, it is enough to prove this for a pseudo-metric spaces $X$. Let $X_{seq}$ denote the set of Cauchy sequences in $X$, and define the distance between the Cauchy sequences $\{x_n\}$ and $\{y_n\}$ to be

$$d(\{x_n\},\{y_n\}) := \lim_{n\to\infty} d(x_n, y_n).$$

It is easy to check that $d$ defines a pseudo-metric on $X_{seq}$. Let $f : X \to X_{seq}$ be the map sending $x$ to the sequence all of whose elements are $x$;

$$f(x) = (x, x, x, x, \cdots).$$

It is clear that $f$ is one to one and is an isometry. The image is dense since by definition

$$\lim d(f(x_n), \{x_n\}) = 0.$$

Now since $f(X)$ is dense in $X_{seq}$, it suffices to show that any Cauchy sequence of points of the form $f(x_n)$ converges to a limit. But such a sequence converges to the element $\{x_n\}$. QED

Of special interest are vector spaces which have metric which is compatible with the vector space properties and which is complete: Let $V$ be a vector space over the real numbers. A **norm** is a real valued function

$$v \mapsto \|v\|$$

on $V$ which satisfies

1. $\|v\| \geq 0$ and $> 0$ if $v \neq 0$,

2. $\|rv\| = |r|\|v\|$ for any real number $r$, and

3. $\|v + w\| \leq \|v\| + \|w\| \ \forall \ \ v, w \in V$.

Then $d(v, w) := \|v-w\|$ is a metric on $V$, which satisfies $d(v+u, w+u) = d(v, w)$ for all $v, w, u \in V$. The ball of radius $r$ about the origin is then the set of all $v$ such that $\|v\| < r$. A vector space equipped with a norm is called a **normed vector space** and if it is complete relative to the metric it is called a **Banach space**.

## 5.3 The contraction fixed point theorem.

Let $X$ and $Y$ be metric spaces. Recall that a map $f : X \to Y$ is called a **Lipschitz map** or is said to be "Lipschitz continuous", if there is a constant $C$ such that

$$d_Y(f(x_1), f(x_2)) \leq C d_X(x_1, x_2), \quad \forall\ x_1, x_2 \in X.$$

If $f$ is a Lipschitz map, we may take the greatest lower bound of the set of all $C$ for which the previous inequality holds. The inequality will continue to hold for this value of $C$ which is known as the Lipschitz constant of $f$ and denoted by $\mathrm{Lip}(f)$.

A map $K : X \to Y$ is called a **contraction** if it is Lipschitz, and its Lipschitz constant satisfies $\mathrm{Lip}(K) < 1$. Suppose $K : X \to X$ is a contraction, and suppose that $Kx_1 = x_1$ and $Kx_2 = x_2$. Then

$$d(x_1, x_2) = d(Kx_1, Kx_2) \leq \mathrm{Lip}(K) d(x_1, x_2)$$

which is only possible if $d(x_1, x_2) = 0$, i.e. $x_1 = x_2$. So a contraction can have at most one fixed point. The contraction fixed point theorem asserts that if the metric space $X$ is complete (and non-empty) then such a fixed point exists.

**Theorem 5.3.1** *Let $X$ be a non-empty complete metric space and $K : X \to X$ a contraction. Then $K$ has a unique fixed point.*

**Proof.** Choose any point $x_0 \in X$ and define

$$x_n := K^n x_0$$

so that

$$x_{n+1} = Kx_n, \quad x_n = Kx_{n-1}$$

and therefore

$$d(x_{n+1}, x_n) \leq C d(x_n, x_{n-1}), \quad 0 \leq C < 1$$

implying that

$$d(x_{n+1}, x_n) \leq C^n d(x_1, x_0).$$

Thus for any $m > n$ we have

$$d(x_m, x_n) \leq \sum_n^{m-1} d(x_{i+1}, x_i) \leq \left(C^n + C^{n+1} + \cdots + C^{m-1}\right) d(x_1, x_0) \leq C^n \frac{d(x_1, x_0)}{1 - C}.$$

This says that the sequence $\{x_n\}$ is Cauchy. Since $X$ is complete, it must converge to a limit $x$, and $Kx = \lim Kx_n = \lim x_{n+1} = x$ so $x$ is a fixed point. We already know that this fixed point is unique. QED

We often encounter mappings which are contractions only near a particular point $p$. If $K$ does not move $p$ too much we can still conclude the existence of a fixed point, as in the following:

**Proposition 5.3.1** *Let $D$ be a closed ball of radius $r$ centered at a point $p$ in a complete metric space $X$, and suppose $K : D \to X$ is a contraction with Lipschitz constant $C < 1$. Suppose that*

$$d(p, Kp) \leq (1 - C)r.$$

*Then $K$ has a unique fixed point in $D$.*

**Proof.** We simply check that $K : D \to D$ and then apply the preceding theorem with $X$ replaced by $D$: For any $x \in D$, we have

$$d(Kx, p) \leq d(Kx, Kp) + d(Kp, p) \leq Cd(x, p) + (1-C)r \leq Cr + (1-C)r = r \quad \text{QED.}$$

**Proposition 5.3.2** *Let $B$ be an open ball or radius $r$ centered at $p$ in a complete metric space $X$ and let $K : B \to X$ be a contraction with Lipschitz constant $C < 1$. Suppose that*

$$d(p, Kp) < (1 - C)r.$$

*Then $K$ has a unique fixed point in $B$.*

**Proof.** Restrict $K$ to any slightly smaller closed ball centered at $p$ and apply Prop. 5.3.1. QED

**Proposition 5.3.3** *Let $K : X \to X$ be a contraction with Lipschitz constant $C$ of a complete metric space. Let $x$ be its (unique) fixed point. Then for any $y \in X$ we have*

$$d(y, x) \leq \frac{d(y, Ky)}{1 - C}.$$

**Proof.** We may take $x_0 = y$ and follow the proof of Theorem 5.3.1. Alternatively, we may apply Prop. 5.3.1 to the closed ball of radius $d(y, Ky)/(1 - C)$ centered at $y$. Prop. 5.3.1 implies that the fixed point lies in the ball of radius $r$ centered at $y$. QED

Prop. 5.3.3 will be of use to us in proving continuous dependence on a parameter in the next section. In the section on iterative function systems for the construction of fractal images, Prop. 5.3.3 becomes the "collage theorem". We might call Prop. 5.3.3 the "abstract collage theorem".

## 5.4 Dependence on a parameter.

Suppose that the contraction "depends on a parameter $s$". More precisely, suppose that $S$ is some other metric space and that

$$K : S \times X \to X$$

with

$$d_X(K(s, x_1), K(s, x_2)) \leq Cd_X(x_1, x_2), \quad 0 \leq C < 1, \ \forall s \in S, \ x_1, x_2 \in X. \ (5.1)$$

(We are assuming that the $C$ in this inequality does not depend on $s$.) If we hold $s \in S$ fixed, we get a contraction

$$K_s : X \to X, \quad K_s(x) := K(s, x).$$

This contraction has a unique fixed point, call it $p_s$. We thus obtain a map

$$S \to X, \quad s \mapsto p_s$$

sending each $s \in S$ into the fixed point of $K_s$.

**Proposition 5.4.1** *Suppose that for each fixed $x \in X$, the map*

$$s \mapsto K(s, x)$$

*of $S \to X$ is continuous. Then the map*

$$s \mapsto p_s$$

*is continuous.*

**Proof.** Fix a $t \in S$ and an $\epsilon > 0$. We must find a $\delta > 0$ such that $d_X(p_s, p_t) < \epsilon$ if $d_S(s, t) < \delta$. Our continuity assumption says that we can find a $\delta > 0$ such that

$$d_X(K(s, p_t), p_t) = d_X(K(s, p_t), K(t, p_t) \leq (1 - C)\epsilon$$

if $d_S(s, t) < \delta$. This says that $K_s$ moves $p_t$ a distance at most $(1 - C)\epsilon$. But then the 'abstract collage theorem", Prop. 5.3.3, says that

$$d_X(p_t, p_s) \leq \epsilon. \quad \text{QED}$$

It is useful to combine Proposition 5.3.1 and 5.4.1 into a theorem:

**Theorem 5.4.1** *Let $B$ be an open ball of radius $r$ centered at a point $q$ in a complete metric space. Suppose that $K : S \times B \to X$ (where $S$ is some other metric space) is continuous, satisfies* (5.1) *and*

$$d_X(K(s, q), q) < (1 - C)r, \quad \forall\ s \in S.$$

*Then for each $s \in S$ there is a unique $p_s \in B$ such that $K(s, p_s) = p_s$, and the map $s \mapsto p_s$ is continuous.*

.

## 5.5 The Lipschitz implicit function theorem

In this section we follow the treatment in [**?**]. We begin with the inverse function theorem which contains the guts of the argument. We will consider a map $F : B_r(0) \to E$ where $B_r(0)$ is the open ball of radius $r$ about the origin in a Banach space, $E$, and where $F(0) = 0$. We wish to conclude the existence of an inverse to $F$, defined on a possible smaller ball by means of the contraction fixed point theorem.

**Proposition 5.5.1** *Let $F : B_r(0) \to E$ satisfy $F(0) = 0$ and*

$$\mathrm{Lip}[F - \mathrm{id}] = \lambda < 1. \tag{5.2}$$

*Then the ball $B_s(0)$ is contained in the image of $F$ where*

$$s = (1 - \lambda)r \tag{5.3}$$

*and $F$ has an inverse, $G$ defined on $B_s(0)$ with*

$$\mathrm{Lip}[G - \mathrm{id}] \leq \frac{\lambda}{1 - \lambda}. \tag{5.4}$$

**Proof.** Let us set $F = \mathrm{id} + v$ so

$$\mathrm{id} + v : B_r(0) \to E, \quad v(0) = 0, \quad \mathrm{Lip}[v] < \lambda < 1.$$

We want to find a $w : B_s(0) \to E$ with

$$w(0) = 0$$

and

$$(\mathrm{id} + v) \circ (\mathrm{id} + w) = \mathrm{id}.$$

This equation is the same as

$$w = -v \circ (\mathrm{id} + w).$$

Let $X$ be the space of continuous maps of $\overline{B_s(0)} \to E$ satisfying

$$u(0) = 0$$

and

$$\mathrm{Lip}[u] \leq \frac{\lambda}{1 - \lambda}.$$

Then $X$ is a complete metric space relative to the sup norm, and, for $x \in \overline{B_s(0)}$ and $u \in X$ we have

$$\|u(x)\| = \|u(x) - u(0)\| \leq \frac{\lambda}{1 - \lambda}\|x\| \leq r.$$

Thus, if $u \in X$ then

$$u : \overline{B_s} \to \overline{B_r}.$$

If $w_1, w_2 \in X$,

$$\| -v \circ (\mathrm{id} + w_1) + v \circ (\mathrm{id} + w_2) \| \leq \lambda \| (\mathrm{id} + w_1) - (\mathrm{id} + w_2) \| = \lambda \| w_1 - w_2 \| .$$

So the map $K : X \to X$

$$K(u) = -v \circ (\mathrm{id} + u)$$

is a contraction. Hence there is a unique fixed point. This proves the proposition.

Now let $f$ be a homeomorphism from an open subset, $U$ of a Banach space, $E_1$ to an open subset, $V$ of a Banach space, $E_2$ whose inverse is a Lipschitz map. Suppose that $h : U \to E_2$ is a Lipschitz map satisfying

$$\text{Lip}[h]\text{Lip}[f^{-1}] < 1. \tag{5.5}$$

Let

$$g = f + h. \tag{5.6}$$

We claim that $g$ is open. That is, we claim that if $y = g(x)$, then the image of a neighborhood of $x$ under $g$ contains a neighborhood of $g(x)$. Since $f$ is a homeomorphism, it suffices to establish this for $g \circ f^{-1} = \text{id} + h \circ f^{-1}$. Composing by translations if necessary, we may apply the proposition. QED

We now want to conclude that $g$ is a homeomorphism = continuous with continuous inverse, and, in fact, is Lipschitz:

**Proposition 5.5.2** *Let $f$ and $g$ be two continuous maps from a metric space $X$ to a Banach space $E$. Suppose that $f$ is injective and $f^{-1}$ is injective with Lipschitz constant* $\text{Lip}[f^{-1}]$. *Suppose that $g$ satisfies*

$$\text{Lip}[g - f] < \frac{1}{\text{Lip}[f^{-1}]}.$$

*Then $g$ is injective and*

$$\text{Lip}[g^{-1}] \leq \frac{1}{\frac{1}{\text{Lip}[f^{-1}]} - \text{Lip}[g - f]} = \frac{\text{Lip}[f^{-1}]}{1 - \text{Lip}[g - f]\text{Lip}[f^{-1}]}. \tag{5.7}$$

**Proof.** By definition,

$$d(x, y) \leq \text{Lip}[f^{-1}]\|f(x) - f(y)\|$$

where $d$ denotes the distance in $X$ and $\| \ \|$ denotes the norm in $E$. We can write this as

$$\|f(x) - f(y)\| \geq \frac{d(x, y)}{\text{Lip}[f^{-1}]}.$$

So

$$\begin{aligned} \|g(x) - g(y)\| &\geq \|f(x) - f(y)\| - \|(g - f)(x) - (g - f)(y)\| \\ &\geq \left(\frac{1}{\text{Lip}[f^{-1}]} - \text{Lip}[g - f]\right) d(x, y). \end{aligned}$$

Dividing by the expression in parenthesis gives the proposition. QED

We can now be a little more precise as to the range covered by $g$:

**Proposition 5.5.3** *Let $U$ be an open subset of a Banach space $E_1$, and $g$ be a homeomorphism of $U$ onto an open subset of a Banach space, $E_2$. Let $x \in U$ and suppose that*

$$\overline{B_r(x)} \subset U.$$

*If $g^{-1}$ is Lipschitz with*

$$\mathrm{Lip}[g^{-1}] < c$$

*then*

$$\overline{B_{\frac{r}{c}}(g(x))} \subset g\left(\overline{B_r(x)}\right).$$

**Proof.** By composing with translations, we may assume that $x = 0,\ g(x) = 0$. Let

$$v \in B_{\frac{r}{c}}(0).$$

Let

$$T = T(v) = \sup\{t | [0,t]v \subset g\left(\overline{B_r(0)}\right)\}.$$

We wish to show that $T = 1$. Since $g(B_r(0))$ contains a neighborhood of 0, we know that $T(v) > 0$, and, by definition,

$$(0,T)v \subset g\left(\overline{B_r(0)}\right).$$

By the Lipschitz estimate for $g^{-1}$ we have

$$\|g^{-1}(tv) - g^{-1}(sv)\| \leq c|t-s|\|v\|.$$

This implies that the limit $\lim_{t\to T} g^{-1}(tv)$ exists and

$$\lim_{t\to T} g^{-1}(tv) \in \overline{B_r(0)}.$$

So

$$Tv \in (g\left(\overline{B_r(0)}\right).$$

If $T = T(v) < 1$ we would have

$$\begin{aligned} \|g^{-1}(Tv)\| &\leq \|g^{-1}(Tv) - g^{-1}(0)\| \\ &\leq c\|Tv\| \\ &= cT\|v\| \\ &< c\|v\| \\ &\leq r. \end{aligned}$$

This says that

$$Tv \in g\left(B_r(0)\right).$$

But since $g$ is open, we could find an $\epsilon > 0$ such that $[T, T+\epsilon]v \subset g\left(B_r(0)\right)$ contradicting the definition of $T$. QED

The above three propositions, taken together, constitute what we might call the inverse function theorem for Lipschitz maps. But the contraction fixed point

theorem allows for continuous dependence on parameters, and gives the fixed point as a continuous function of the parameters. So this then yields the implicit function theorem.

The differentiability of the solution, in the case that the implicit function is assumed to be continuously differentiable follows as in Chapter 1.

# Chapter 6

# Hutchinson's theorem and fractal images.

## 6.1 The Hausdorff metric and Hutchinson's theorem.

Let $X$ be a complete metric space. Let $\mathcal{H}(X)$ denote the space of non-empty compact subsets of $X$. For any $A \in \mathcal{H}(X)$ and any positive number $\epsilon$, let

$$A_\epsilon = \{x \in X | d(x, y) \leq \epsilon, \text{for some } y \in A\}.$$

We call $A_\epsilon$ the $\epsilon$-*collar* of $A$. Recall that we defined

$$d(x, A) = \inf_{y \in A} d(x, y)$$

to be the distance from any $x \in X$ to $A$, then we can write the definition of the $\epsilon$-collar as

$$A_\epsilon = \{x | d(x, A) \leq \epsilon\}.$$

Notice that the infimum in the definition of $d(x, A)$ is actually achieved, that is, there is some point $y \in A$ such that

$$d(x, A) = d(x, y).$$

This is because $A$ is compact. For a pair of non-empty compact sets, $A$ and $B$, define

$$d(A, B) = \max_{x \in A} d(x, B).$$

So

$$d(A, B) \leq \epsilon \ \ iff \ \ A \subset B_\epsilon.$$

Notice that this condition is not symmetric in $A$ and $B$. So Hausdorff introduced

$$h(A, B) = \max\{d(A, B), d(B, A)\} \tag{6.1}$$

$$= \inf\{\epsilon \mid A \subset B_\epsilon \text{ and } B \subset A_\epsilon\}. \tag{6.2}$$

as a distance on $\mathcal{H}(X)$. He proved

**Proposition 6.1.1** *The function $h$ on $\mathcal{H}(X) \times \mathcal{H}(X)$ satsifies the axioms for a metric and makes $\mathcal{H}(X)$ into a complete metric space. Furthermore, if*

$$A, B, C, D \in \mathcal{H}(X)$$

*then*

$$h(A \cup B, C \cup D) \leq \max\{h(A, C), h(B, D)\}. \tag{6.3}$$

**Proof.** We begin with (6.3). If $\epsilon$ is such that $A \subset C_\epsilon$ and $B \subset D_\epsilon$ then clearly $A \cup B \subset C_\epsilon \cup D_\epsilon = (C \cup D)_\epsilon$. Repeating this argument with the roles of $A, C$ and $B, D$ interchanged proves (6.3).

We prove that $h$ is a metric: $h$ is symmetric, by definition. Also, $h(A, A) = 0$, and if $h(A, B) = 0$, then every point of $A$ is within zero distance of $B$,and hence must belong to $B$ since $B$ is compact, so $A \subset B$ and similarly $B \subset A$. So $h(A, B) = 0$ implies that $A = B$.

We must prove the triangle inequality. For this it is enough to prove that

$$d(A, B) \leq d(A, C) + d(C, B),$$

because interchanging the role of $A$ and $B$ gives the desired result. Now for any $a \in A$ we have

$$\begin{aligned} d(a, B) &= \min_{b \in B} d(a, b) \\ &\leq \min_{b \in B} \left(d(a, c) + d(c, b)\right) \ \forall c \in C \\ &= d(a, c) + \min_{b \in B} d(c, b) \ \forall c \in C \\ &= d(a, c) + d(c, B) \ \forall c \in C \\ &\leq d(a, c) + d(C, B) \ \forall c \in C. \end{aligned}$$

The second term in the last expression does not depend on $c$, so minimizing over $c$ gives

$$d(a, B) \leq d(a, C) + d(C, B).$$

Maximizing over $a$ on the right gives

$$d(a, B) \leq d(A < C) + d(C, B).$$

Maximizing on the left gives the desired

$$d(A, B) \leq d(A, C) + d(C, A).$$

We sketch the proof of completeness. Let $A_n$ be a sequence of compact non-empty subsets of $X$ which is Cauchy in the Hausdorff metric. Define the set $A$ to be the set of all $x \in X$ with the property that there exists a sequence of points $x_n \in A_n$ with $x_n \to x$. It is straighforward to prove that $A$ is compact and non-empty and is the limit of the $A_n$ in the Hausdorff metric.

Suppose that $K : X \to X$ is a contraction. Then $K$ defines a transformation on the space of subsets of $X$ (which we continue to denote by K):

$$K(A) = \{Kx | x \in A\}.$$

Since $K$ continuous, it carries $\mathcal{H}(X)$ into itself. Let $c$ be the Lipschitz constant of $K$. Then

$$\begin{aligned} d(K(A), K(B)) &= \max_{a\in A}[\min_{b\in B} d(K(a), K(b))] \\ &\leq \max_{a\in A}[\min_{b\in B} cd(a, b)] \\ &= cd(A, B). \end{aligned}$$

Similarly, $d(K(B), K(A)) \leq c\, d(B, A)$ and hence

$$h(K(A), K(B)) \leq c\, h(A, B). \tag{6.4}$$

In other words, a contraction on $X$ induces a contraction on $\mathcal{H}(X)$.

The previous remark together with the following observation is the key to Hutchinson's remarkable construction of fractals:

**Proposition 6.1.2** *Let $T_1, \ldots, T_n$ be a collection of contractions on $\mathcal{H}(X)$ with Lipschitz constants $c_1, \ldots, c_n$, and let $c = \max c_i$. Define the transformation $T$ on $\mathcal{H}(X)$ by*

$$T(A) = T_1(A) \cup T_2(A) \cup \cdots \cup T_n(A).$$

*Then $T$ is a contraction with Lipschitz constant $c$.*

**Proof.** By induction, it is enough to prove this for the case $n = 2$. By (6.3)

$$\begin{aligned} h(T(A), T(B)) &= h(T_1(A) \cup T_2(A), T_1(B) \cup T_2(B)) \\ &\leq \max\{h(T_1(A), h(T_1(B)), h(T_2(A), T_2(B))\} \\ &\leq \max\{c_1 h(A, B), c_2 h(A, B)\} \\ &= h(A, B) \max\{c_1, c_2\} = c \cdot h(A, B) \end{aligned}$$

.

Putting the previous facts together we get Hutchinson's theorem;

**Theorem 6.1.1** *Let $K_1, \ldots, K_n$ be contractions on a complete metric space and let $c$ be the maximum of their Lifschitz contants. Define the Hutchinson operator, $K$, on $\mathcal{H}(X)$ by*

$$K(A) = K_1(A) \cup \cdots \cup K_n(a).$$

*Then $K$ is a contraction with Lipschtz constant $c$.*

## 6.2 Affine examples

We describe several examples in which $X$ is a subset of a vector space and each of the $T_i$ in Hutchinson's theorem are affine transformations of the form

$$T_i : x \mapsto A_i x + b_i$$

where $b_i \in X$ and $A_i$ is a linear transformation.

### 6.2.1 The classical Cantor set.

Take $X = [0, 1]$, the unit interval. Take

$$T_1 : x \mapsto \frac{x}{3}, \quad T_2 : x \mapsto \frac{x}{2} + \frac{2}{3}.$$

These are both contractions, so by Hutchinson's theorem there exists a unique closed fixed set $C$. This is the Cantor set.

To relate it to Cantor's original construction, let us go back to the proof of the contraction fixed point theorem applied to $T$ acting on $\mathcal{H}(X)$. It says that if we start with any non-empty compact subset $A_0$ and keep applying $T$ to it, i.e. set $A_n = T^n A_0$ then $A_b \to C$ in the Hausdorff metric, $h$. Suppose we take the interval $I$ itself as our $A_0$. Then

$$A_1 = T(I) = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1].$$

in other words, applying the Hutchinson operator $T$ to the interval $[0, 1]$ has the effect of deleting the "middle third" open interval $(\frac{1}{3}, \frac{2}{3})$. Applying $T$ once more gives

$$A_2 = T^2[0, 1] = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1].$$

In other words, $A_2$ is obtained from $A_1$ by deleting the middle thirds of each of the two intervals of $A_1$ and so on. This was Cantor's original construction. Since $A_{n+1} \subset A_n$ for this choice of initial set, the Hausdorff limit coincides with the intersection.

But of course Hutchinson's theorem (and the proof of the contractions fixed point theorem) says that we can start with *any* non-empty closed set as our initial "seed" and then keep applying $T$. For example, suppose we start with the one point set $B_0 = \{0\}$. Then $B_1 = TB_0$ is the two point set

$$B_1 = \{0, \frac{2}{3}\},$$

$B_2$ consists of the four point set

$$B_2 = \{0, \frac{2}{9}, \frac{2}{3}, \frac{8}{9}\}$$

and so on. We then must take the Hausdorff limit of this increasing colletion of sets. To describe the limiting set $c$ from this point of view, it is useful to use triadic expansions of points in $[0,1]$. Thus

$$\begin{aligned} 0 &= .0000000\cdots \\ 2/3 &= .2000000\cdots \\ 2/9 &= .0200000\cdots \\ 8/9 &= .2200000\cdots \end{aligned}$$

and so on. Thus the set $B_n$ will consist of points whose triadic exapnsion has only zeros or twos in the first $n$ positions followed by a string of all zeros. Thus a point will lie in $C$ (be the limit of such points) if and only if it has a triadic expansion consisting entirely of zeros or twos. This includes the possibility of an infinite string of all twos at the tail of the expansion. for example, the point 1 which belongs to the Cantor set has a triadic expansion $1 = .222222\cdots$. Simialrly the point $\frac{2}{3}$ has the triadic expansion $\frac{2}{3} = .0222222\cdots$ and so is in the limit of the sets $B_n$. But a point such as $.101\cdots$ is not in the limit of the $B_n$ and hence not in $C$. This description of $C$ is also due to Cantor. Notice that for any point $a$ with triadic expansion $a = .a_1a_2a_2\cdots$

$$T_1a = .0a_1a_2a_3\cdots, \quad \text{while} \quad T_2a = .2a_1a_2a_3\cdots.$$

Thus if all the entries in the expansion of $a$ are either zero or two, this will also be true for $T_1$ and $T_2a$. This shows that the $C$ (given by this second Cantor description) satisfies $TC \subset C$. On the other hand,

$$T_1(.a_2a_3\cdots) = .0a_2a_3\cdots, \quad T_2(.a_2a_3\cdots) = .2a_2a_3\cdots$$

which shows that $.a_1a_2a_3\cdots$ is in the image of $T_1$ if $a_1 = 0$ or in the image of $T_2$ if $a_1 = 2$. This shows that $TC = C$. Since $C$ (according to Cantor's second description) is closed, the uniqueness part of the fixed point theorem guarantees that the second description coincides iwth the first.

The statement that $TC = C$ implies that $C$ is "self-similar".

### 6.2.2 The Sierpinski Gasket

Consider the three affine transformations of the plane:

$$T_1 : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}, \quad T_2 : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

$$T_3 : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The fixed point of the Hutchinson operator for this choice of $T_1, T_2, T_3$ is called the Sierpinski gasket, $S$. If we take our initial set $A_0$ to be the right triangle with vertices at

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \text{and} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

then each of the $T_i A_0$ is a similar right triangle whose linear dimensions are one-half as large, and which shares one common vertex with the original triangle. In other words,

$$A_1 = TA_0$$

is obtained from our original triangle be deleting the interior of the (reversed) right triangle whose vertices are the midpoints of our orignal triangle. Just as in the case of the Cantor set, successive applications of $T$ to this choice of original set amounts to sussive deletions of the "middle" and the Hausdorff limit is the intersection of all them: $S = \bigcap A_i$.

We can also start with the one element set

$$B_0 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}$$

Using a binary expansion for the $x$ and $y$ coordinates, application of $T$ to $B_0$ gives the three element set

$$\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} .1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ .1 \end{pmatrix} \right\}.$$

The set $B_2 = TB_1$ will contain nine points, whose binary expansion is obtained from the above three by shifting the $x$ and $y$ exapnsions one unit to the right and either inserting a 0 before both expansions (the effect of $T_1$), insert a 1 before the expansion of $x$ and a zero before teh $y$ or vice versa. Proceding in this fashion, we see that $B_n$ consists of $3^n$ points which have all 0 in the binary expansion of the $x$ and $y$ coordinates, past the $n$-th position, and which are further constrained by the condition that at no earler point do we have both $x_i = 1$ and $y_1 = 1$. Passing to the limit shows that $S$ consists of all points for which we can find (possible inifinite) binary expansions of the $x$ and $y$ coordinates so that $x_i = 1 = y_i$ never occurs. (For example $x = \frac{1}{2}, y = \frac{1}{2}$ belongs to $S$ because we can write $x = .10000\cdots, y = .011111\ldots$). Again, from this (second) description of $S$ in terms of binary expansions it is clear that $TS = S$.

# Chapter 7

# Hyperbolicity.

## 7.1 $C^0$ linearization near a hyperbolic point

Let $E$ be a Banach space. A linear map

$$A : E \to E$$

is called *hyperbolic* if we can find closed subspaces $S$ and $U$ of $E$ which are invariant under $A$ such that we have the direct sum decomposition

$$E = S \oplus U \tag{7.1}$$

and a positive constant $a < 1$ so that the estimates

$$\|A_s\| \leq a < 1, \quad A_s = A_{|S} \tag{7.2}$$

and

$$\|A_u^{-1}\| \leq a < 1, \quad A_u = A_{|U} \tag{7.3}$$

hold. (Here, as part of hypothesis (7.3), it is assumed that the restriction of $A$ to $U$ is an isomorphism so that $A_u^{-1}$ is defined.)

If $p$ is a fixed point of a diffeomorphism $f$, then it is called a *hyperbolic* fixed point if the linear transformation $df_p$ is hyperbolic.

The main purpose of this section is prove that any diffeomorphism, $f$ is conjugate via a local *homeomorphism* to its derivative, $df_p$ near a hyperbolic fixed point. A more detailed statement will be given below. We discussed the one dimensional version of this in Chapter 3.

**Proposition 7.1.1** *Let $A$ be a hyperbolic isomorphism (so that $A^{-1}$ is bounded) and let*

$$\epsilon < \frac{1-a}{\|A^{-1}\|}. \tag{7.4}$$

*If $\phi$ and $\psi$ are bounded Lipschitz maps of $E$ into itself with*

$$\mathrm{Lip}[\phi] < \epsilon, \quad \mathrm{Lip}[\psi] < \epsilon$$

*then there is a unique solution to the equation*

$$(\mathrm{id} + u) \circ (A + \phi) = (A + \psi) \circ (\mathrm{id} + u) \tag{7.5}$$

*in the space, $X$ of bounded continuous maps of $E$ into itself. If $\phi(0) = \psi(0) = 0$ then $u(0) = 0$.*

**Proof.** If we expand out both sides of (7.5) we get the equation

$$Au - u(A + \phi) = \phi - \psi(\mathrm{id} + u).$$

Let us define the linear operator, $L$, on the space $X$ by

$$L(u) = Au - u \circ (\mathrm{id} + \phi).$$

So we wish to solve the equation

$$L(u) = \phi - \psi(A + u).$$

We shall show that $L$ is invertible with

$$\|L^{-1}\| \leq \frac{\|A^{-1}\|}{(1 - a)}. \tag{7.6}$$

Assume, for the moment that we have proved (7.6). We are then looking for a solution of

$$u = K(u)$$

where

$$K(u) = L^{-1}[\phi - \psi(\mathrm{id} + u)].$$

But

$$\begin{aligned} \|K(u_1) - K(u_2)\| &= \|L^{-1}[\phi - \psi(\mathrm{id} + u_1) - \phi + \psi(\mathrm{id} + u_2)]\| \\ &= \|L^{-1}[\psi(\mathrm{id} + u_2) - \psi(\mathrm{id} + u_1)]\| \\ &\leq \|L^{-1}\|\mathrm{Lip}[\psi]\|u_2 - u_1\| \\ &< c\|u_2 - u_1\|, \quad c < 1 \end{aligned}$$

if we combine (7.6) with (7.4). Thus $K$ is a contraction and we may apply the contraction fixed point theorem to conclude the existence and uniqueness of the solution to (7.5). So we turn our attention to the proof that $L$ is invertible and of the estimate (7.6). Let us write

$$Lu = A(Mu)$$

where

$$Mu = u - A^{-1}u \circ (A + \phi).$$

Composition with $A$ is an invertible operator and the norm of its inverse is $\|A^{-1}\|$. So we are reduced to proving that $M$ is invertible and that we have the estimate

$$\|M^{-1}\| \leq \frac{1}{1 - a}. \tag{7.7}$$

Let us write

$$u = f \oplus g, \quad f : E \to S, \quad g : E \to U$$

in accordance with the decomposition (7.1). So if we let $Y$ denote the space of bounded continuous maps from $E$ to $S$, and let $Z$ denote the space of bounded continuous maps from $E$ to $U$, we have

$$X = Y \oplus Z$$

and the operator $M$ sends each of the spaces $Y$ and $Z$ into themselves since $A^{-1}$ preserves $S$ and $U$. We let $M_s$ denote the restriction of $M$ to $Y$, and let $M_u$ denote the restriction of $M$ to $Z$. It will be enough for us to prove that each of the operators $M_s$ and $M_u$ is invertible with a bounds (7.7) with $M$ replaced by $M_s$ and by $M_u$. For $f \in Y$ let us write

$$M_s f = f - Nf, \quad Nf = A^{-1} f \circ (A + \phi).$$

We will prove

**Lemma 7.1.1** *The map $N$ is invertible and we have*

$$\|N^{-1}\| \leq a.$$

**Proof.**We claim that he map $A + \phi$ a homeomorphism with Lipschitz inverse. Indeed

$$\|Ax\| \geq \frac{1}{\|A^{-1}\|}\|x\|$$

so

$$\begin{aligned} \|Ax + \phi(x) - Ay - \phi(y)\| &\geq \left[\frac{1}{\|A^{-1}\|} - \mathrm{Lip}[\phi]\right] \|x - y\| \\ &\geq \frac{a}{\|A^{-1}\|}\|x - y\| \end{aligned}$$

by (7.4). This shows that $A + \phi$ is one to one. Furthermore, to solve

$$Ax + \phi(x) = y$$

for $x$, we apply the contraction fixed point theorem to the map

$$x \mapsto A^{-1}(y - \phi(x)).$$

The estimate (7.4) shows that this map is a contraction. Hence $A + \phi$ is also surjective.

Thus the map $N$ is invertible, with

$$N^{-1} f = A_s f \circ (A + \phi)^{-1}.$$

Since $\|A_s\| \leq a$, we have

$$\|N^{-1} f\| \leq a\|f\|.$$

(This is in terms of the sup norm on $Y$.) In other words, in terms of operator norms,

$$\|N^{-1}\| \leq a.$$

We can now find $M_s^{-1}$ by the geometric series

$$\begin{aligned} M_s^{-1} &= (I-N)^{-1} \\ &= [(-N)(I-N^{-1})]^{-1} \\ &= (-N)^{-1}[I+N^{-1}+N^{-2}+N^{-3}+\cdots] \end{aligned}$$

and so on $Y$ we have the estimate

$$\|M_s^{-1}\| \leq \frac{a}{1-a}.$$

The restriction, $M_u$, of $M$ to $Z$ is

$$M_u g = g - Qg$$

with

$$\|Qg\| \leq a\|g\|$$

so we have the simpler series

$$M_u^{-1} = I + Q + Q^2 + \cdots$$

giving the estimate

$$\|M_u\| \leq \frac{1}{1-a}.$$

Since

$$\frac{a}{1-a} < \frac{1}{1-a}$$

the two pieces together give the desired estimate

$$\|M\| \leq \frac{1}{1-a},$$

completing the proof of the first part of the proposition. Since evaluation at zero is a continuous function on $X$, to prove the last statement of the proposition it is enough to observe that if we start with an initial approximation satisfying $u(0) = 0$ (for example $u \equiv 0$) $Ku$ will also satisfy this condition and hence so will $K^n u$ and therefor so will the unique fixed point.

Now let $f$ be a differentiable, hyperbolic transformation defined in some neighborhood of 0 with $f(0) = 0$ and $df_0 = A$. We may write

$$f = A + \phi$$

where

$$\phi(0) = 0, \quad d\phi_0 = 0.$$

We wish to prove

**Theorem 7.1.1** *There exists neighborhoods $U$ and $V$ of $0$ and a homeomorphism $h : U \to V$ such that*

$$h \circ A = f \circ h. \tag{7.8}$$

We prove this theorem by modifying $\phi$ outside a sufficiently small neighborhood of 0 in such a way that the new $\phi$ is globally defined and has Lipschitz constant less than $\epsilon$ where $\epsilon$ satisfies condition (7.4). We can then apply the proposition to find a global $h$ which conjugates the modified $f$ to $A$, and $h(0) = 0$. But since we will not have modified $f$ near the origin, this will prove the local assertion of the theorem. For this purpose, choose some function $\rho : \mathbf{R} \to \mathbf{R}$ with

$$\begin{aligned} \rho(t) &= 0 \quad \forall \quad t \geq 1 \\ \rho(t) &= 1 \quad \forall \quad t \leq \frac{1}{2} \\ |\rho'(t)| &< K \quad \forall t \end{aligned}$$

where $K$ is some number,

$$K > 2.$$

For a fixed $\epsilon$ let $r$ be sufficiently small so that the on the ball, $B_r(0)$ we have the estimate

$$\|d\phi_x\| < \frac{\epsilon}{2K},$$

which is possible since $d\phi_0 = 0$ and $d\phi$ is continuous. Now define

$$\psi(x) = \rho(\frac{\|x\|}{r})\phi(x),$$

and continuously extend to

$$\psi(x) = 0, \quad \|x\| \geq r.$$

Notice that

$$\psi(x) = \phi(x), \quad \|x\| \leq \frac{r}{2}.$$

Let us now check the Lipschitz constant of $\psi$. There are three alternatives: If $x_1$ and $x_2$ both belong to $B_r(0)$ we have

$$\begin{aligned} \|\psi(x_1) - \psi(x_2)\| &= \|\rho(\frac{\|x_1\|}{r})\phi(x_1) - \rho(\frac{\|x_2\|}{r})\phi(x_2) \\ &\leq |\rho(\frac{\|x_1\|}{r}) - \rho(\frac{\|x_2\|}{r})\|\phi(x_1)\| + \rho(\frac{\|x_2\|}{r})\|\phi(x_1) - \phi(x_2)\| \\ &\leq (K\|x_1 - x_2\|/r) \times \|x_1\| \times (\epsilon/2K) + (\epsilon/2K) \times \|x_1 - x_2\| \\ &\leq \epsilon\|x_1 - x_2|. \end{aligned}$$

If $x_1 \in B_r(0),\ x_2 \notin B_r(0)$, then the second term in the expression on the second line above vanishes and the first term is at most $(\epsilon/2)\|x_1 - x_2\|$. If neither $x_1$ nor $x_2$ belong to $B_r(0)$ then $\psi(x_1) - \psi(x_2) = 0 - 0 = 0$. We have verified that $\mathrm{Lip}[\psi] < \epsilon$ and so have proved the theorem.

## 7.2 invariant manifolds

Let $p$ be a hyperbolic fixed point of a diffeomorphism, $f$. The *stable manifold* of $f$ at $p$ is defined as the set

$$W^s(p) = W^s(p, f) = \{x | \lim_{n\to\infty} f^n(x) = p\}. \tag{7.9}$$

Similarly, the *unstable manifold* of $f$ at $p$ is defined as

$$W^u(p) = W^u(p, f) = \{x | \lim_{n\to\infty} f^{-n}(x) = p\}. \tag{7.10}$$

We have defined $W^s$ and $W^u$ as sets. We shall see later on in this section that in fact they are submanifolds, of the same degree of smoothness as $f$. The terminology, while standard, is unfortunate. A point which is not exactly on $W^s(p)$ is swept away under iterates of $f$ from any small neighborhood of $p$. This is the content of our first proposition below. So it is a very *unstable* property to lie on $W^s$. Better terminology would be "contracting" and "expanding" submanifolds. But the usage is standard, and we will abide by it. In any event, the sets $W^s(p)$ and $W^u(p)$ are, by their very definition, invariant under $f$.

In the case that $f = A$ is a hyperbolic *linear* transformation on a Banach space $E = S \oplus U$, then $W^s(0) = S$ and $W^u(0) = U$ as follows immediately from the definitions. The main result of this section will be to prove that in the general case, the stable manifold of $f$ at $p$ will be a submanifold whose tangent at $p$ is the stable subspace of the linear transformation $df_p$.

Notice that for a hyperbolic fixed point, replacing $f$ by $f^{-1}$ interchanges the roles of $W^s$ and $W^u$. So in much of what follows we will formulate and prove theorems for either $W^s$ or for $W^u$. The corresponding results for $W^u$ or for $W^s$ then follow automatically.

Let $A$ be a hyperbolic linear transformation on a Banach space $E = S \oplus U$, and consider any ball, $B_r = B_r(0)$ of radius $r$ about the origin. If $x \in B_r$ does *not* lie on $S \cap B_r$, this means that if we write $x = x_s \oplus x_u$ with $x_s \in S$ and $x_u \in U$ then $x_u \neq 0$. Then

$$\begin{aligned} \|A^n x\| &= \|A^n x_s\| + \|A^n x_u\| \\ &\geq \|A^n x_u\| \\ &\geq c^n \|x_u\|. \end{aligned}$$

If we choose $n$ large enough, we will have $c^n\|x_u\| > r$. So eventually, $A^n x \notin B_r$. Put contrapositively,

$$S \cap B_r = \{x \in B_r | A^n x \in B_r \forall n \geq 0\}.$$

Now consider the case of a hyperbolic fixed point, $p$, of a diffeomorphism, $f$. We may introduce coordinates so that $p = 0$, and let us take $A = df_0$. By the $C^0$ conjugacy theorem, we can find a neighborhood, $V$ of 0 and homeomorphism

$$h : B_r \to V$$

with

$$h \circ f = A \circ h.$$

Then

$$f^n(x) = h^{-1} \circ A^n \circ h\ (x)$$

will lie in $U$ for all $n \geq 0$ if and only if $h(x) \in S(A)$ if and only if $A^n h(x) \to 0$. This last condition implies that $f^n(x) \to p$. We have thus proved

**Proposition 7.2.1** *Let $p$ be a hyperbolic fixed point of a diffeomorphism, $f$. For any ball, $B_r(p)$ of radius $r$ about $p$, let*

$$B_r^s(p) = \{x \in B_r(p) | f^n(x) \in B_r^s(p) \forall n \geq 0\}. \tag{7.11}$$

*Then for sufficiently small $r$, we have*

$$B_r^s(p) \subset W^s(p).$$

Furthermore, our proof shows that for sufficiently small $r$ the set $B_r^s(p)$ is a topological submanifold in the sense that every point of $B_r^s(p)$ has a neighborhood (in $B_r^s(p)$) which is the image of a neighborhood, $V$ in a Banach space under a homeomorphism, $H$. Indeed, the restriction of $h$ to $S$ gives the desired homeomorphism.

*Remark.* In the general case we can not say that $B_r^s(p) = B_r(p) \cap W^s(p)$ because a point may escape from $B_r(p)$, wander around for a while, and then be drawn towards $p$.

But the proposition does assert that $B_r^s(p) \subset W^s(p)$ and hence, since $W^s$ is invariant under $f^{-1}$, we have

$$f^{-n}[B_r^s(p)] \subset W^s(p)$$

for all $n$, and hence

$$\bigcup_{n \geq 0} f^{-n}[B_r^s(p)] \subset W^s(p).$$

On the other hand, if $x \in W^s(p)$, which means that $f^n(x) \to p$, eventually $f^n(x)$ arrives and stays in any neighborhood of $p$. Hence $p \in f^{-n}[B_r^s(p)]$ for some $n$. We have thus proved that for sufficiently small $r$ we have

$$W^s(p) = \bigcup_{n \geq 0} f^{-n}[B_r^s(p)]. \tag{7.12}$$

We will prove that $B_r^s(p)$ is a submanifold. It will then follow from (7.12) that $W^s(p)$ is a submanifold. The global disposition of $W^s(p)$, and in particular its relation to the stable and unstable manifolds of other fixed points, is a key ingredient in the study of the long term behavior of dynamical systems. In this section our focus is purely local, to prove the smooth character of the set $B_r^s(p)$. We follow the treatment in [**?**].

We will begin with the hypothesis that $f$ is merely Lipschitz, and give a proof (independent of the $C^0$ linearization theorem) of the existence and Lipschitz

character of the $W^u$. We will work in the following situation: $A$ is a hyperbolic linear isomorphism of a Banach space $E = S \oplus U$ with

$$\|Ax\| \leq a\|x\|,\ x \in S, \quad \|A^{-1}x\| \leq a\|x\|, x \in U.$$

We let $S(r)$ denote the ball of radius $s$ about the origin in $S$, and $U(r)$ the ball of radius $r$ in $U$. We will assume that

$$f : S(r) \times U(r) \to E$$

is a Lipschitz map with

$$\|f(0)\| \leq \delta \tag{7.13}$$

and

$$\text{Lip}[f - A] \leq \epsilon. \tag{7.14}$$

We wish to prove the following

**Theorem 7.2.1** *Let $c < 1$. There exists an $\epsilon = \epsilon(a)$ and a $\delta = \delta(a, \epsilon, r)$ so that if $f$ satisfies (7.13) and (7.14) then there is a map*

$$g : E_u(r) \to E_s(r)$$

*with the following properties:*
*(i) $g$ is Lipschitz with* $\text{Lip}[g] \leq 1$.
*(ii) The restriction of $f^{-1}$ to graph$(g)$ is contracting and hence has a fixed point, $p$, on graph$(g)$.*
*(iii) We have*

$$graph(g) = \bigcap f^n(S(r) \oplus U(r)) = W^u(p) \cap [S(r) \oplus U(p)].$$

The idea of the proof is to apply the contraction fixed point theorem to the space of maps of $U(r)$ to $S(r)$. We want to identify such a map, $v$, with its graph:

$$graph(v) = \{(v(x), x),\ x \in U(r)\}.$$

Now

$$f[graph(v)] = \{f(v(x), x)\} = \{(f_s(v(x), x), f_u(v(x), x))\},$$

where we have introduced the notation

$$f_s = p_s \circ f, \quad f_u = p_u \circ f,$$

where $p_s$ denotes projection onto $S$ and $p_u$ denotes projection onto $U$.

Suppose that the projection of $f[graph(v)]$ onto $U$ is injective and its image contains $U(r)$. This means that for any $y \in U(r)$ there is a unique $x \in U(r)$ with

$$f_u(v(x), x) = y.$$

So we write

$$x = [f_u \circ (v, id)]^{-1}(y)$$

where we think of $(v, id)$ as a map of $U(r) \to E$ and hence of

$$f_u \circ (v, id)$$

as a map of $U(r) \to U$. Then we can write

$$f[graph(v)] = \{(f_s(v([f_u \circ (v, id)]^{-1}(y), y)\} = graph G_[f(v)]$$

where

$$G_f(v) = f_s \circ (v, id) \circ [f_u \circ (v, id)]^{-1}. \tag{7.15}$$

The map $v \mapsto G_f(v)$ is called the *graph transform* (when it is defined). We are going to take

$$X = \mathrm{Lip}_1(U(r), S(r))$$

to consist of all Lipschitz maps from $U(r)$ to $S(r)$ with Lipschitz constant $\leq 1$. The purpose of the next few lemmas is to show that if $\epsilon$ and $\delta$ are sufficiently small then the graph transform, $G_f$ is defined and is a contraction on $X$. The contraction fixed point theorem will then imply that there is a unique $g \in X$ which is fixed under $G_f$, and hence that $graph(g)$ is invariant under $f$. We will then find that $g$ has all the properties stated in the theorem.

In dealing with the graph transform it is convenient to use the box metric, $|\ |$, on $S \oplus U$ where

$$|x_s \oplus x_u| = \max\{\|x_s\|, \|x_u\|\}$$

i.e.

$$|x| = \max\{\|p_s(x)\|, \|p_u(x)\|\}.$$

We begin with

**Lemma 7.2.1** *If $v \in X$ then*

$$\mathrm{Lip}[f_u \circ (v, id) - A_u] \leq \mathrm{Lip}[f - A].$$

**Proof.** Notice that

$$p_u \circ A(v(x), x) = p_u(A_s(v(x)), A_u x) = A_u x$$

so

$$f_u \circ (v, id) - A_u = p_u \circ [f - A] \circ (v, id).$$

We have $\mathrm{Lip}[p_u] \leq 1$ since $p_u$ is a projection, and

$$\mathrm{Lip}(v, id) \leq \max\{\mathrm{Lip}[v], \mathrm{Lip}[id]\} = 1$$

since we are using the box metric. Thus the lemma follows.

**Lemma 7.2.2** *Suppose that $0 < \epsilon < c^{-1}$ and*

$$\mathrm{Lip}[f - A] < \epsilon.$$

*Then for any $v \in X$ the map $f_u \circ (v, id) : E_u(r) \to E_u$ is a homeomorphism whose inverse is a Lipschitz map with*

$$\mathrm{Lip}\left[[f_u \circ (v, id)]^{-1}\right] \leq \frac{1}{c^{-1} - \epsilon}. \tag{7.16}$$

**Proof.**Using the preceding lemma, we have

$$\mathrm{Lip}[f_u - A_u] < \epsilon < c^{-1} < \|A_u^{-1}\|^{-1} = (\mathrm{Lip}[A_u])^{-1}\,.$$

By the Lipschitz implicit function theorem we conclude that $f_u \circ (v, id)$ is a homeomorphism with

$$\mathrm{Lip}\left[[f_u \circ (v, id)]^{-1}\right] \leq \frac{1}{\|A_u^{-1}\|^{-1} - \mathrm{Lip}[f_u \circ (v, id) - A_u]} \leq \frac{1}{c^{-1} - \epsilon}$$

by another application of the preceding lemma. QED. We now wish to show that the image of $f_u \circ (v, id)$ contains $U(r)$ if $\epsilon$ and $\delta$ are sufficiently small: By the proposition in section 5.2 concerning the image of a Lipschitz map, we know that the image of $U(r)$ under $f_u \circ (v, id)$ contains a ball of radius $r/\lambda$ about $[f_u \circ (v, id)](0)$ where $\lambda$ is the Lipschitz constant of $[f_u \circ (v, id)]^{-1}$. By the preceding lemma, $r/\lambda = r(c^{-1} - \epsilon)$. Hence $f_u \circ (v, id)(U(r))$ contains the ball of radius

$$r(c^{-1} - \epsilon) - \|f_u(v(0), 0)\|$$

about the origin. But

$$\begin{aligned}
\|f_u(v(0), 0)\| &\leq \|f_u(0, 0)\| + \|f_u(v(0), 0) - f_u(0, 0)\| \\
&\leq \|f_u(0, 0)\| + \|(f_u - p_u A)(v(0), 0) - (f_u - p_u A)(0, 0)\| \\
&\leq |f(0)| + |(f - A)(v(0), 0) - (f - A)(0, 0)| \\
&\leq |f(0)| + \epsilon r.
\end{aligned}$$

The passage from the second line to the third is because $p_u A(x, y) = A_u y = 0$ if $y = 0$. The passage from the third line to the fourth is because we are using the box norm. So

$$r(c^{-1} - \epsilon) - \|f_u(v(0), 0)\| \geq r(c^{-1} - 2\epsilon) - \delta$$

if (7.13) holds. We would like this expression to be $\geq r$, which will happen if

$$\delta \leq r(c^{-1} - 1 - 2\epsilon). \tag{7.17}$$

We have thus proved

**Proposition 7.2.2** *Let $f$ be a Lipschitz map satisfying (7.13) and (7.14) where $2\epsilon < c^{-1} - 1$ and (7.17) holds. Then for every $v \in X$, the graph transform, $G_f(v)$ is defined and*

$$\mathrm{Lip}[G_f(v)] \leq \frac{c + \epsilon}{c^{-1} - \epsilon}.$$

The estimate on the Lipschitz constant comes from

$$\begin{aligned}
\mathrm{Lip}[G_f(v)] &\leq \mathrm{Lip}[f_s \circ (v, \mathrm{id})]\mathrm{Lip}[(f_u \circ (v, \mathrm{id})] \\
&\leq \mathrm{Lip}[f_s]\mathrm{Lip}[v]\mathrm{Lip} \cdot \frac{1}{c^{-1} - \epsilon} \\
&\leq (\mathrm{Lip}[A_s] + \mathrm{Lip}[p_s \circ (f - A)]) \cdot \frac{1}{c^{-1} - \epsilon} \\
&\leq \frac{c + \epsilon}{c^{-1} - \epsilon}.
\end{aligned}$$

In going from the first line to the second we have used the preceding lemma.

In particular, if

$$2\epsilon < c^{-1} - c \tag{7.18}$$

then

$$\mathrm{Lip}[G_f(v)] \leq 1.$$

Let us now obtain a condition on $\delta$ which will guarantee that

$$G_f(v)(U(r) \subset S(r).$$

Since

$$f_u \circ (v, \mathrm{id})U(r) \supset U(r),$$

we have

$$[f_u \circ (v, \mathrm{id})]^{-1}U(r) \subset U(r).$$

Hence, from the definition of $G_f(v)$, it is enough to arrange that

$$f_s \circ (v, \mathrm{id})[U(r)] \subset S(r).$$

For $x \in U(r)$ we have

$$\begin{aligned} \|f_s(v(x), x)\| &\leq \|p_s \circ (f - A)(v(x), x)\| + \|A_s v(x)\| \\ &\leq |(f - A)(v(x), x)| + c\|v(x)\| \\ &\leq |(f - A)(v(x), x) - (f - A)(0, 0)| + |f(0)| + cr \\ &\leq \epsilon|(v(x), x)| + \delta + cr \\ &\leq \epsilon r + \delta + cr. \end{aligned}$$

So we would like to have

$$(\epsilon + c)r + \delta < r$$

or

$$\delta \leq r(1 - c - \epsilon). \tag{7.19}$$

If this holds, then $G_f$ maps $X$ into $X$.

We now want conditions that guarantee that $G_f$ is a contraction on $X$, where we take the sup norm. Let $(w, x)$ be a point in $S(r) \oplus U(r)$ such that $f_u(w, x) \in U(r)$. Let $v \in X$, and consider

$$|(w, x) - (v(x), x)| = \|w - v(x)\|,$$

which we think of as the distance along $S$ from the point $(w, x)$ to $graph(v)$. Suppose we apply $f$. So we replace $(w, x)$ by $f(w, x) = (f_s(w, x), f_u(w, x))$ and $graph(v)$ by $f(graph(v)) = graph(G_f(v))$. The corresponding distance along $S$ is $\|f_s(w, x) - G_f(v)(f_u(w, x)\|$. We claim that

$$\|f_s(w, x) - G_f(v)(f_u(w, x))\| \leq (c + 2\epsilon)\|w - v(x)\|. \tag{7.20}$$

Indeed,

$$f_s(v(x), x) = G_f(v)(f_u(v(x), x)$$

by the definition of $G_f$, so we have

$$\begin{aligned}\|f_s(w,x) - G_f(v)(f_u(w,x))\| &\leq \|f_s(w,x) - f_s(v(x),x)\| + \\ &\quad + \|G_f(v)(f_u((v(x),x) - G_f(v)(f_u(w,x))\| \\ &\leq \mathrm{Lip}[f_s]|(w,x) - (v(x),x)| + \\ &\quad + \mathrm{Lip}[f_u]|(v(x),x) - (w,x)| \\ &\leq \mathrm{Lip}[f_s - p_sA + p_sA]\|w - v(x)\| + \\ &\quad + \mathrm{Lip}[f_u - p_uA]\|w - v(x)\| \\ &\leq (\epsilon + c + \epsilon)\|w - v(x)\|\end{aligned}$$

which is what was to be proved.

Consider two elements, $v_1$ and $v_2$ of $X$. Let $z$ be any point of $U(r)$, and apply (7.20) to the point

$$(w,x) = (v_1([f_u \circ (v_1, \mathrm{id})]^{-1}](z)), [f_u \circ (v_1, \mathrm{id})]^{-1}](z))$$

which lies on $graph(v_1)$, and where we take $v = v_2$ in (7.20). The image of $(w,x)$ is the point $(G_f(v_1)(z), z)$ which lies on $graph(G_f(v_1))$, and, in particular, $f_u(w,x) = z$. So (7.20) gives

$$\|G_f(v_1)(z) - G_f(v_2)(z)\| \leq (c+2\epsilon)\|v_1([f_u \circ (v_1, \mathrm{id})]^{-1}](z)) - v_2([f_u \circ (v_1, \mathrm{id})]^{-1}](z)\|.$$

Taking the sup over $z$ gives

$$\|G_f(v_1) - G_f(v_2)\|_{\sup} \leq (c + 2\epsilon)\|v_1 - v_2\|_{\sup}. \tag{7.21}$$

Intuitively, what (7.20) is saying is that $G_f$ multiplies the $S$ distance between two graphs by a factor of at most $(c + 2\epsilon)$. So $G_f$ will be a contraction in the sup norm if

$$2\epsilon < 1 - c \tag{7.22}$$

which implies (7.18). To summarize: we have proved that $G_f$ is a contraction in the sup norm on $X$ if (7.17), (7.19) and (7.22) hold, i.e.

$$2\epsilon < 1 - c, \quad \delta < r \min(c^{-1} - 1 - 2\epsilon,\ 1 - c - \epsilon).$$

Notice that since $c < 1$, we have $c^{-1} - 1 > 1 - c$ so both expressions occurring in the min for the estimate on $\delta$ are positive.

Now the uniform limit of continuous functions which all have $\mathrm{Lip}[v] \leq 1$ has Lipschitz constant $\leq 1$. In other words, $X$ is closed in the sup norm as a subset of the space of continuous maps of $U(r)$ into $S(r)$, and so we can apply the contraction fixed point theorem to conclude that there is a unique fixed point, $g \in X$ of $G_f$. Since $g \in X$, condition (i) of the theorem is satisfied. As for (ii), let $(g(x), x)$ be a point on $graph(g)$ which is the image of the point $(g(y), y)$ under $f$, so

$$(g(x), x) = f(g(y), y)$$

which implies that

$$x = [f_u \circ (g, \mathrm{id})](y).$$

We can write this equation as

$$p_u \circ f_{|graph(g)} = [f_u \circ (g, \mathrm{id})] \circ (p_u)_{|graph(g)}.$$

In other words, the projection $p_u$ conjugates the restriction of $f$ to $graph(g)$ into $[f_u \circ (g, \mathrm{id})]$. Hence the restriction of$f^{-1}$ to $graph(g)$ is conjugated by $p_u$ into $[f_u \circ (g, \mathrm{id})]^{-1}$. But, by (7.16), the map $[f_u \circ (g, \mathrm{id})]^{-1}$ is a contraction since

$$c^{-1} - 1 > 1 - c > 2\epsilon$$

so

$$c^{-1} - \epsilon > 1 + \epsilon > 1.$$

The fact that $\mathrm{Lip}[g] \leq 1$ implies that

$$|(g(x), x) - (g(y), y)| = \|x - y\|$$

since we are using the box norm. So the restriction of $p_u$ to $graph(g)$ is an isometry between the (restriction of) the box norm on $graph(g)$and the norm on $U$. So we have proved statement (ii), that the restriction of $f^{-1}$ to $graph(g)$ is a contraction.

We now turn to statement (iii) of the theorem. Suppose that $(w, x)$ is a point in $S(r) \oplus U(r)$ with $f(w, x) \in S(r) \oplus U(r)$. By (7.20) we have

$$\|f_s(w, x) - g(f_u(w, x)\| \leq (c + 2\epsilon)\|w - g(x)\|$$

since $G_f(g) = g$. So if the first $n$ iterates of $f$ applied to $(w, x)$ all lie in $S(r) \oplus U(r)$, and if we write

$$f^n(w, x) = (z, y),$$

we have

$$\|z - g(y)\| \leq (c + 2\epsilon)^n \|w - g(x)\| \leq (c + 2\epsilon)r.$$

So if the point $(z, y)$ is in $\bigcap f^n(S(r) \oplus U(r))$ we must have $z = g(y)$, in other words

$$\bigcap f^n(S(r) \oplus U(r)) \subset graph(g).$$

But

$$graph(g) = f[graph(g)] \cap [S(r) \oplus U(r)]$$

so

$$graph(g) \subset \bigcap f^n(S(r) \oplus U(r)),$$

proving that

$$graph(g) = \bigcap f^n(S(r) \oplus U(r)).$$

We have already seen that the restriction of $f^{-1}$ to $graph(g)$ is a contraction, so all points on $graph(g)$ converge under the iteration of $f^{-1}$ to the fixed point, $p$. So they belong to $W^u(p)$. This completes the proof of the theorem.

Notice that if $f(0) = 0$, then $p = 0$ is the unique fixed point.

# Chapter 8

# Symbolic dynamics.

## 8.1 Symbolic dynamics.

We have already seen several examples where a dynamical system is conjugate to the dynamical system consisting of a "shift" on sequences of symbols. It is time to pin this idea down with some formal definitions.

**Definition.** A **discrete compact dynamical system** $(M, F)$ consists of a compact metric space $M$ together with a continuous map $F : M \to M$. If $F$ is a homeomorphism then $(M, F)$ is said to be an **invertible** dynamical system.

If $(M, F)$ and $(N, G)$ are compact discrete dynamical systems then a map $\phi : M \to N$ is called a **homomorphism** if

- $\phi$ is continuous, and
- $$G \circ \phi = \phi \circ F,$$

  in other words if the diagram

  $$\begin{array}{ccc} M & \xrightarrow{F} & M \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\phi} \\ N & \xrightarrow[G]{} & N \end{array}$$

  commutes.

If the homomorphism $\phi$ is surjective it is called a **factor**. If $\phi$ a homeomorphism then it is called a **conjugacy**.

For the purposes of this chapter, we will only be considering compact discrete situations, so shall drop these two words.

Let $\mathcal{A}$ be a finite set called an "alphabet". The set $\mathcal{A}^{\mathbf{Z}}$ consists of all bi-infinite sequences $x = \cdots x_{-2}, x_{-1}, x_0, x_1, x_2, x_3, \cdots$. On this space let us put the metric (a slight variant of the metric we introduce earlier) $d(x, x) = 0$ and, if $x \neq y$ then

$$d(x, y) = 2^{-k} \text{ where } k = \max_i [x_{-i}, x_i] = [y_{-i}, y_i].$$

Here we use the notation $[x_k, x_\ell]$ to denote the "block"

$$[x_k, x_\ell] = x_k x_{k+1} \cdots x_\ell$$

from $k$ to $\ell$ occurring in $x$. (This makes sense if $k \leq \ell$. If $\ell < k$ we adopt the convention that $[x_k, x_\ell]$ is the empty word.) Thus the elements $x$ and $y$ are close in this metric if they agree on a large central block. So a sequence of points $\{x^n\}$ converges if and only if, given any fixed $k$ and $\ell$, the $[x^n_k, x^n_\ell]$ eventually agree for large $n$. From this characterization of convergence, it is easy to see that the space $\mathcal{A}^{\mathbf{Z}}$ is sequentially compact: Let $x^n$ be a sequence of points of $\mathcal{A}^{\mathbf{Z}}$, We must find a convergent subsequence. The method is Cantor diagonalization: Since $\mathcal{A}$ is finite we may find an infinite subsequence $n_i$ of the $n$ such that all the $x_0^{n_i}$ are equal. Infinitely many elements from this subsequence must also agree at the positions $-1$ and 1 since there are only finitely many possible choices of entries. In other words, we may choose a subsequence $n_{i_j}$ of our subsequence such that all the $[x_{-1}^{n_{i_j}}, x_1^{n_{i_j}}]$ are equal. We then choose an infinite subsequence of this subsubsequence such that all the $[x_{-3}, x_3]$ are equal. And so on. We then pick an element $N_1$ from our first subsequence, an element $N_2 > N_1$ from our subsubsequence, an element $N_3 > N_2$ from our subsubsubsequence etc. By construction we have produced an infinite subsequence which converges.

In the examples we studies, we did not allow all sequences, but rather excluded certain types. Let us formalize this. By a **word** from the alphabet $\mathcal{A}$ we simply mean a finite string of letters of $\mathcal{A}$. Let $\mathcal{F}$ be a set of words. Let

$$X_{\mathcal{F}} = \{x \in \mathcal{A}^{\mathbf{Z}} | [x_k, x_\ell] \notin \mathcal{F}\}$$

for any $k$ and $\ell$. In other words, $X_{\mathcal{F}}$ consists of those sequences $x$ for which no word of $\mathcal{F}$ ever occurs as a block in $x$. From our characterization of convergence (as eventual agreement on any block) it is clear that $X_{\mathcal{F}}$ is a closed subset of $\mathcal{A}^{\mathbf{Z}}$ and hence compact. It is also clear that $X_{\mathcal{F}}$ is mapped into itself by the **shift map**

$$\sigma : \mathcal{A}^{\mathbf{Z}} \to \mathcal{A}^{\mathbf{Z}}, \quad (\sigma x)_k := x_{k+1}.$$

It is also clear that $\sigma$ is continuous. By abuse of language we may continue to denote the restriction of $\sigma$ to $X_{\mathcal{F}}$ by $\sigma$ although we may also use the notation $\sigma_X$ for this restriction. A dynamical system of the form $(X, \sigma_X)$ where $x = X_{\mathcal{F}}$ is called a **shift** dynamical system.

Suppose that $(X, \sigma_X)$ with $X = X_{\mathcal{F}} \subset \mathcal{A}^{\mathbf{Z}}$ and $(Y, \sigma_Y)$ with $Y = Y_{\mathcal{G}} \subset \mathcal{B}^{\mathbf{Z}}$ are shift dynamical systems. What does a homomorphism $\phi : X \to Y$ look like? For each $b \in \mathcal{B}$, let

$$C_0(b) = \{y \in Y | y_0 = b\}.$$

(The letter $C$ is used to denote the word "cylinder" and the subscript 0 denotes that we are constructing the so called cylinder set obtained by specifying that the value of $y$ at the "base" 0.) The sets $C_0(b)$ are closed, hence compact, and distinct. The finitely many sets $\phi^{-1}(C_0(b))$ are therefore also disjoint. Since $\phi$ is continuous by the definition of a homomorphism, each of the sets $\phi^{-1}(C_0)(b)$ is compact, as the inverse i age of a compact set under a continuous map from a compact space is compact. Hence there is a $\delta > 0$ such that the distance between any two different sets $\phi^{-1}(C_0(b))$ is $> \delta$. Choose $n$ with $2^{-n} < \delta$. Let if $x, x' \in X$. Then

$$[x_{-n}, x_n] = [x'_{-n}, x'_n] \Rightarrow \phi(x) = \phi(x')$$

since they are at distance at most $2^{-n}$ and hence must lie in the same $\phi^{-1}(C_0(b)$. In other words, there is a map

$$\Phi : \mathcal{A}^{2n+1} \to \mathcal{B}$$

such that

$$\phi(x)_0 = \Phi([x_{-n}, x_n]).$$

But now the condition that $\sigma_Y \circ \phi = \phi \circ \sigma_X$ implies that

$$\phi(x)_1 = \Phi([x_{-n+1}, x_n + 1])$$

and more generally that

$$\phi(x)_j = \Phi([x_{j-n}, x_{j+n}]). \tag{8.1}$$

Such a map is called a **sliding block code** of block size $2n + 1$ (or "with memory $n$ and anticipation $n$") for obvious reasons. Conversely, suppose that $\phi$ is a sliding block code. It clearly commutes with the shifts. If $x$ and $x'$ agree on a central block of size $2N + 1$, then $\phi(x)$ and $\phi(y)$ agree on a central block of size $2(N - n) + 1$. This shows that $\phi$ is continuous. In short, we have proved

**Proposition 8.1.1** *A map $\phi$ between two shift dynamical systems is a homomorphism if and only if it is a sliding block code.*

The advantage of this proposition is that it converts a topological property, continuity, into a finite type property - the sliding block code. Conversely, we can use some topology of compact sets to derive facts about sliding block codes. For example, it is easy to check that a bijective continuous map $\phi : X \to Y$ between compact metric spaces is a homeomorphism, i.e. that $\phi^{-1}$ is continuous. Indeed, if not, we could find a sequence of points $y_i \in Y$ with $y_n \to y$ and $x_n = \phi^{-1}(y_k) \not\to x = \phi^{-1}(y)$. Since $X$ is compact, we can find a subsequence of the $x_n$ which converge to some point $x' \neq x$. Continuity demands that $\phi(x') = y = \phi(x)$ and this contradicts the bijectivity. Form this we conclude that the inverse of a bijective sliding block code is continuous, hence itself a sliding block code - a fact that is not obvious from the definitions.

## 8.2 Shifts of finite type.

For example, let $M$ be any positive integer and suppose that we map $X_{\mathcal{F}} \subset \mathcal{A}^{\mathbf{Z}}$ into $(\mathcal{A}^M)^{\mathbf{Z}}$ as follows: A "letter" in $\mathcal{A}^M$ is an $M$-tuplet of letters of $\mathcal{A}$. Define the map $\phi : X_{\mathcal{F}} \to (\mathcal{A}^M)^{\mathbf{Z}}$ by letting $\phi(x)_i = [x_i, x_{i+M}]$. For example, if $M = 5$ and we write the 5-tuplets as column vectors, the element $x$ is mapped to

$$\dots, \begin{pmatrix} x_{-1} \\ x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}, \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \begin{pmatrix} x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}, dots.$$

This map is clearly a sliding code, hence continuous, and commutes with shift hence is a homomorphism. On the other hand it is clearly bijective since we can recover $x$ from its image by reading the top row. Hence it is a conjugacy of $X$ onto its image. Call this image $X^M$.

We say that $X$ is of **finite type** if we can choose a finite set $\mathcal{F}$ of forbidden words so that $X = X_{\mathcal{F}}$.

### 8.2.1 One step shifts.

If $w$ is a forbidden word for $X$, then any word which contains $w$ as a substring is also forbidden. If $M + 1$ denotes the largest length of a word in $\mathcal{F}$, we may enlarge all the remaining words by adding all suffixes and prefixes to get words of length $M$. Hence, with no loss of generality, we may assume that all the words of $\mathcal{F}$ have length $M$. So $\mathcal{F} \subset \mathcal{A}^M$. Such a shift is called an $M$-step shift. But if we pass from $X$ to $X^{M+1}$, the elements of $(\mathcal{A})^{\mathcal{M}+\infty}$ are now the alphabet. So excluding the elements of $\mathcal{F}$ means that we have replaced the alphabet $\mathcal{A}^{M+1}$ by the smaller alphabet $\mathcal{E}$, the complement of $\mathcal{F}$ in $\mathcal{A}^{\mathcal{M}+\infty}$. Thus $X^{M+1} \subset \mathcal{E}^{\mathbf{Z}}$. The condition that an element of $\mathcal{B}^{\mathbf{Z}}$ actually belong to $X$ is easy to describe: An $M + 1$-tuplet $y_i$ can be followed by an $M + 1$-tuplet $y_{i+1}$ if and only if the last $M$ entries in $y_i$ coincide with the first $M$ entries in $y_{i+1}$. All words $w = yy'$ which do not satisfy this condition are excluded. but all these words have length two. We have proved that the study of shifts of finite type is the same as the study of one step shifts.

### 8.2.2 Graphs.

We can rephrase the above argument in the language of graphs. For any shift and any positive integer $K$ $X$ we let $\mathcal{W}_K(X)$ denote the set of all admissible words of length $K$. Suppose that $X$ is an $M$-step shift. Let us set

$$\mathcal{V} := \mathcal{W}_M(X),$$

and define

$$\mathcal{E} = \mathcal{W}_{M+1}(X)$$

as before. Define maps

$$i : \mathcal{E} \to \mathcal{V}, \quad t : \mathcal{E} \to \mathcal{V}$$

to be

$$i(a_0 a_1 \cdots a_M) = a_0 a_1 \cdots a_{M-1} \quad t(a_0 a_1 \cdots a_M) = a_1 \cdots a_M.$$

Then a sequence $u = \cdots u_1 u_0 u_1 u_2 \cdots \in \mathcal{E}^{\mathbf{Z}}$, where $u_i \in \mathcal{E}$ lies in $X^{M+1}$ if and only if

$$t(u_j) = i(u_j + 1) \tag{8.2}$$

for all $j$.

So let us define a directed multigraph (**DMG** for short) $G$ to consist of a pair of sets $(\mathcal{V}, \mathcal{E})$ (called the set of vertices and the set of edges) together with a pair of maps

$$i : \mathcal{E} \to \mathcal{V}, \quad t : \mathcal{E} \to \mathcal{V}.$$

We may think the edges as joining one vertex to another, the edge $e$ going from $i(e)$ (the initial vertex) to $t(e)$ the terminal vertex. The edges are "oriented" in the sense each has an initial and a terminal point. We use the phrase multigraph since nothing prevents several edges from joining the same pair of vertices. Also we allow for the possibility that $i(e) = t(e)$, i.e. for "loops".

Starting from any **DMG** $G$, we define $Y_G \subset \mathcal{E}^{\mathbf{Z}}$ to consist of those sequences for which (8.2) holds. This is clearly a step one shift.

We have proved that any shift of finite type is conjugate to $Y_G$ for some **DMG** $G$.

### 8.2.3 The adjacency matrix

suppose we are given $\mathcal{V}$. Up to renaming the edges which merely changes the description of the alphabet, $\mathcal{E}$, we know $G$ once we know how many edges go from $i$ to $j$ for every pair of elements $i, j \in \mathcal{V}$. This is a non-negative integer, and the matrix

$$A = A(G) = (a_{ij})$$

is called the **adjacency matrix** of $G$. All possible information about $G$, and hence about $Y_G$ is encoded in the matrix $A$. Our immediate job will be to extract some examples of very useful properties of $Y_G$ from algebraic or analytic properties of $A$. In any event, we have reduced the study of finite shifts to the study of square matrices with non-negative integer entries.

### 8.2.4 The number of fixed points.

For any dynamical system, $(M, F)$ let $p_n(F)$ denote the number (possibly infinite) of fixed points of $F^n$. These are also called periodic points of period $n$. We shall show that if $A$ is the adjacency matrix of the **DMG** $G$, and $(Y_G, \sigma_Y)$ is the associated shift, then

$$p_n(\sigma_Y) = \operatorname{tr} A^n. \tag{8.3}$$

To see this, observe that for any vertices $i$ and $j$, $a_{ij}$ denotes the number of edges joining $i$ to $j$. Squaring the matrix $A$, the $ij$ component of $A^2$ is

$$\sum_k a_{ik}a_{kj}$$

which is precisely the number of words (or paths) of length two which start at $i$ and end at $j$. By induction, the number of paths of length $n$ which join $i$ to $j$ is the $ij$ component of $A^n$. Hence the $ii$ component of $A^n$ is the number of paths of length $n$ which start and end at $i$. Summing over all vertices, we see that $\operatorname{tr} A^n$ is the number of all cycles of length $n$. But if $c$ is a cycle of length $n$, then the infinite sequence $y = \cdots ccccc \cdots$ is periodic with period $n$ under the shift. Conversely, if $y$ is periodic of period $n$, then $c = [y_0, y_{n-1}]$ is a cycle of length $n$ with $y = \cdots ccccc \cdots$. Thus $p_n(\sigma_Y) =$ the number of cycles of length $n = \operatorname{tr} A^n$. QED

### 8.2.5 The zeta function.

Let $(M, F)$ be a dynamical system for which $p_n(F) < \infty$ for all $n$. A convenient bookkeeping device for storing all the numbers $p_n(F)$ is the **zeta function**

$$\zeta_F(t) := \exp\left(\sum_n p_n(F)\frac{t^n}{n}\right).$$

Let $x$ be a periodic point (of some period) and let $m = m(x)$ be the minimum period of $x$. Let $\gamma = \gamma(x) = \{x, Fx, \ldots, F^{m-1}x\}$ be the orbit of $x$ under $F$ and all its powers. So $m = m(\gamma) = m(x)$ is the number of elements of $\gamma$. The number of elements of period $n$ which correspond to elements of $\gamma$ is $m$ if $m|n$ and zero otherwise. If we denote this number by $p_n(F, \gamma)$ then

$$\exp\left(\sum_n p_n(F,\gamma)\frac{t^n}{n}\right) = \exp\left(\sum_j mj\frac{t^{mj}}{mj}\right) =$$

$$\exp\left(\sum_j \frac{t^m j}{j}\right) = \exp\left(-\log(1-t^m)\right) = \frac{1}{1-t^m}.$$

Now

$$p_n(F) = \sum_\gamma p_n(F,\gamma)$$

since a point of period $n$ must belong to some periodic orbit. Since the exponential of a sum is the product of the exponentials we conclude that

$$\zeta_F(t) = \prod_\gamma \left(\frac{1}{1-t^{m(\gamma)}}\right).$$

Now let us specialize to the case $(Y_G, \sigma_Y)$ for some **DMG**, $G$. We claim that

$$\zeta_\sigma(t) = \frac{1}{\det(I - tA)}. \tag{8.4}$$

Indeed,

$$p_n(\sigma) = \operatorname{tr} A^n = \sum \lambda_i^n$$

where the sum is over all the eigenvalues (counted with multiplicity). Hence

$$\zeta_\sigma(t) = \prod \exp \sum \frac{(\lambda_i t)^n}{n} = \prod \left(\frac{1}{1 - \lambda_i t}\right) = \frac{1}{\det(I - tA)}. \quad \text{QED}$$

## 8.3 Topological entropy.

Let $X$ be a shift space, and let $\mathcal{W}_n(X)$ denote the number of words of length $n$ which appear in $X$. Let $w_n = \#(\mathcal{W}_n(X)$ denote the number of words of length $n$. Clearly $w_n \geq 1$ (as we assume that $X$ is not empty), and

$$w_{m+n} \leq w_m \cdot w_n$$

and hence

$$\log_2(w_{m+n}) \leq \log_2(w_m) + \log_2(w_n).$$

This implies that

$$\lim_{n\to\infty} \frac{1}{n} \log_2 w_n$$

exists on account of the following:

**Lemma 8.3.1** *Let $a_1, a_2 \ldots$ be a sequence of non-negative real numbers satisfying*

$$a_{m+n} \leq a_m + a_n.$$

*Then $\lim_{n\to\infty} \frac{1}{n} a_n$ exists and in fact*

$$\lim_{n\to\infty} \frac{1}{n} a_n = \inf_{n\to\infty} \frac{1}{n} a_n.$$

**Proof.** Set $a := \inf_{n\to\infty} \frac{1}{n} a_n$. For any $\epsilon > 0$ we must show that there exists an $N = N(\epsilon)$ such that

$$\frac{1}{n} a_n \leq a + \epsilon \quad \forall\ n \geq N(\epsilon).$$

Choose some integer $r$ such that

$$a_r < a + \frac{1}{2}\epsilon.$$

Such an $r \geq 1$ exists by the definition of $a$. Using the inequality in the lemma, if $0 \leq j < r$

$$\frac{a_{mr+j}}{mr + j} \leq \frac{a_m r}{mr + j} + \frac{a_j}{mr + j}.$$

Decreasing the denominator the right hand side is $\leq$

$$\frac{a_{mr}}{mr} + \frac{a_j}{mr}.$$

There are only finitely many $a_j$ which occur in the second term, and hence by choosing $m$ large we can arrange that the second term is always $< \frac{1}{2}\epsilon$. Repeated application of the inequality in the lemma gives

$$\frac{a_{mr}}{mr} \leq \frac{ma_r}{mr} = \frac{a_r}{r} < a + \frac{1}{2}\epsilon. \quad \text{QED.}$$

Thus we define

$$h(X) = \lim_{n\to\infty} \frac{1}{n}\log_2 \#(\mathcal{W}_n(X)), \tag{8.5}$$

and call $h(X)$ the **topological entropy** of $X$. (This is a standard but unfortunate terminology, as the topological entropy is only loosely related to the concept of entropy in thermodynamics, statistical mechanics or information theory). To show the it is an invariant of $X$ we prove

**Proposition 8.3.1** *Let $\phi : X \to Y$ be a factor (i.e. a surjective homomorphism). Then $h(Y) \leq h(X)$. In particular, if $h$ is a conjugacy, then $h(X) = h(Y)$.*

**Proof.** We know that $\phi$ is given by a sliding block code, say of size $2m+1$. Then every block in $Y$ of size $n$ is the image of a block in $X$ of size $n+2m+1$, i.e.

$$1n\log_2 \#(\mathcal{W}_n(Y)) \leq 1n\log_2 \#(\mathcal{W}_{n+2m+1}(X)).$$

Hence

$$\frac{1}{n}1n\log_2 \#(\mathcal{W}_n(Y)) \leq \left(\frac{n+2m+1}{n}\right)\frac{1}{n+2m+1}1n\log_2 \#(\mathcal{W}_{n+2m+1}(X)).$$

The expression in parenthesis tends to ! as $n \to \infty$ proving that $h(Y) \leq h(X)$. If $\phi$ is a conjugacy, the reverse inequality applies. *Box*

### 8.3.1 The entropy of $Y_G$ from $A(G)$.

The adjacency matrix of a **DMG** has non-negative integer entries, in particular non-negative entries. If a row consisted entirely of zeros, then no edge would emanate from the corresponding vertex, so this vertex would make no contribution to the corresponding shift. Similarly if column consisted entirely of zeros. So without loss of generality, we may restrict ourselves to graphs whose adjacency matrix contains at least one positive entry in each row and in each column. This implies that if $A^k$ has *all* its entries positive, then so does $A^{k+1}$ and hence all higher powers. A matrix with non-negative entries which has this property is called **primitive**. A matrix is called **irreducible** if In terms of the graph $G$, the condition of being primitive means that for all sufficiently large $n$ any vertices $i$ and $j$ can be joined by a path of length $n$. A slightly weaker condition is that

of **irreducibility** which asserts for any $i$ and $j$ there exist (arbitrarily large) $n = n(i,j)$) and a path of length $n$ joining $i$ and $j$. In terms of the matrix, this says that given $i$ and $j$ there is some power $n$ such that $(A^n)_{ij} > 0$. For example, the matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

is irreducible but not primitive.

The **Perron-Frobenius Theorem** whose proof we will give in the next section asserts every irreducible matrix $A$ has a positive eigenvalue $\lambda_A$ such that $\lambda_A \geq |\mu|$ for any other eigenvalue $\mu$ and also that $Av = \lambda_A v$ for some vector $v$ all of whose entries are positive, and that no other eigenvalue has an eigenvector with all positive entries. We will use this theorem to prove:

**Theorem 8.3.1** *Let $G$ be a* **DMG** *whose adjacency matrix $A(G)$is irreducible. Let $y_G$ be the corresponding shift space. then*

$$h(Y_G) = \lambda_{A(G)}. \tag{8.6}$$

**Proof.** The number of words of length $n$ which join the vertex $i$ to the vertex $j$ is the $ij$ entry of $A^n$ where $A = A(G)$. Hence

$$\#(\mathcal{W}_n(Y_G)) = \sum_{ij} (A^n)_{ij}.$$

Let $v$ be an eigenvector of $A$ with all positive entries, and let $m > 0$ be the minimum of these entries and $M$ the maximum. Also let us write $\lambda$ for $\lambda_A$. We have $A^n v = \lambda^n v$, or written out

$$\sum_j (A^n)_{ij} v_j = \lambda^n v_i.$$

Hence

$$m \sum_j (A^n)_{ij} \leq \lambda^n M.$$

Summing over $i$ gives

$$m\#(\mathcal{W}_n(Y_G)) \leq rM\lambda^n$$

where $r$ is the size of the matrix $A$. Hence

$$\log_2 m + \log_2 \#(\mathcal{W}_n(Y_G)) \leq \log_2(Mr) + n\log_2\lambda.$$

Dividing by $n$ and passing to the limit shows that

$$h(Y_G) \leq \lambda_A.$$

On the other hand, for any $i$ we have

$$m\lambda^n \leq \lambda^n v_i \leq \sum_j (A^n)_{ij} v_j \leq M \sum_j (A^n) ij.$$

Summing over $j$ gives

$$rm\lambda^n \leq M\#(\mathcal{W}_n(Y_G)).$$

Again, taking logarithms and dividing by $n$ proves the reverse inequality $h(Y_G) \geq \lambda_A$. QED

For example, if

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$$

then

$$A2 = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$$

so $A$ is primitive. Its eigenvalues are

$$\frac{1 \pm \sqrt{5}}{2}$$

so that

$$h(Y_G) = \frac{1 + \sqrt{5}}{2}.$$

If $A$ is not irreducible, this means that there will be proper subgraphs from which there is "no escape". One may still apply the Perron Frobenius theorem to the collection of all irreducible subgraphs, and replace the $\lambda_A$ that occurs in the theorem by the maximum of the maximum eigenvalues of each of the irreducible subgraphs. We will not go into the details.

## 8.4 The Perron-Frobenius Theorem.

We say that a real matrix $T$ is **non-negative** (or **positive**) if all the entries of $T$ are non-negative (or positive). We write $T \geq 0$ or $T > 0$. We will use these definitions primarily for square $(n \times n)$ matrices and for column vectors $(n \times 1$ matrices). We let

$$Q := \{x \in \mathbf{R}^n : x \geq 0, \quad x \neq 0\}$$

so $Q$ is the non-negative "orthant" excluding the origin. Also let

$$C := \{x \geq 0 : \|x\| = 1\}.$$

So $C$ is the intersection of the orthant with the unit sphere.

A non-negative matrix square $T$ is called **primitive** if there is a $k$ such that all the entries of $T^k$ are positive. It is called **irreducible** if for any $i, j$ there is a $k = k(i, j)$ such that $(T^k)_{ij} > 0$. If $T$ is irreducible then $I + T$ is primitive. Until further notice in this section we will assume that $T$ is non-negative and irreducible.

**Theorem 8.4.1 Perron-Frobenius,1**. *$T$ has a positive (real) eigenvalue $\lambda_{\max}$ such that all other eigenvalues of $T$ satisfy*

$$|\lambda| \leq \lambda_{\max}.$$

*Furthermore $\lambda_{\max}$ has algebraic and geometric multiplicity one, and has an eigenvector $x$ with $X > 0$. Finally any non-negative eigenvector is a multiple of $x$. More generally, if $y \geq 0,\ y \neq 0$ is a vector and $\mu$ is a number such that*

$$Ty \leq \mu y$$

*then*

$$y > 0, \quad \textit{and} \quad \mu \geq \lambda_{\max}$$

*with $\mu = \lambda_{\max}$ if and only if $y$ is a multiple of $x$.*

*If $0 \leq S \leq T,\ S \neq T$ then every eigenvalue $\sigma$ of $S$ satisfies $|\sigma| < \lambda_{\max}$. In particular, all the diagonal minors $T_i$ obtained from $T$ by deleting the $i$-th row and column have eigenvalues all of which have absolute value $< \lambda_{\max}$.*

**Proof.** Let

$$P := (I + T)^{n-1}$$

and for any $z \in Q$ let

$$L(z) := \max\{s : sz \leq Tz\} = \min_{1 \leq i \leq n, z_i \neq 0} \frac{(Tz)_i}{z_i}.$$

By definition $L(rz) = L(z)$ for any $r > 0$, so $L(z)$ depends only on the ray through $z$. If $z \leq y,\ z \neq y$ we have $Pz < Py$. Also $PT = TP$. So if $sz \leq Tz$ then

$$sPz \leq PTz = TPz$$

so

$$L(Pz) \geq L(z).$$

Furthermore, if $L(z)z \neq Tz$ then $L(z)Pz < TPz$. So $L(Pz) > L(z)$ unless $z$ is an eigenvector of $T$. Consider the image of $C$ under $P$. It is compact (being the image of a compact set under a continuous map) and all of the elements of $P(C)$ have all their components strictly positive (since $P$ is positive). Hence the function $L$ is continuous on $P(C)$. Thus $L$ achieves a maximum value, $L_{\max}$ on $P(C)$. Since $L(z) \leq L(Pz)$ this is in fact the maximum value of $L$ on all of $Q$, and since $L(Pz) > L(z)$ unless $z$ is an eigenvector of $T$, we conclude that $L_{\max}$ is achieved at an eigenvector, call it $x$ of of $T$ and $x > 0$ with $L_{\max}$ the eigenvalue. Since $Tx > 0$ and $Tx = L_{\max}x$ we have $L_{max} > 0$.

We will now show that this is in fact the maximum eigenvalue in the sense of the theorem. So let $y$ be any eigenvector with eigenvalue $\lambda$, and let $|y|$ denote the vector whose components are $|y_j|$, the absolute values of the components of $y$. We have $|y| \in Q$ and from

$$Ty = \lambda y$$

and the triangle inequality we conclude that

$$|\lambda||y| \leq T|y|.$$

Hence $|\lambda| \leq L(|y|) \leq L_{\max}$. So we may use the notation

$$\lambda_{\max} := L_{\max}$$

since we have proved that

$$|\lambda| \leq \lambda_{\max}.$$

Suppose that $0 \leq S \leq T$. Then $sz \leq Sz$ and $Sz \leq Tz$ implies that $sz \leq Tz$ so $L_S(z) \leq L_T(z)$ for all $z$ and hence

$$L_{\max}(S) \leq L_{\max}(T).$$

We may apply the same argument to $T^\dagger$ to conclude that it also has a positive maximum eigenvalue. Let us call it $\eta$. (We shall soon show that $\eta = \lambda_max$.) This means that there is a vector $w > 0$ such that

$$w^\dagger T = \eta w.$$

We have

$$w^\dagger T x = \eta w^\dagger x = \lambda_{\max} w^\dagger x$$

implying that $\eta = \lambda_{max}$ since $w^\dagger x > 0$.

Now suppose that $y \in Q$ and $Ty \leq \mu y$. Then

$$\lambda_{\max} w^\dagger y = w^\dagger T y \leq \mu w^\dagger y$$

implying that $\lambda_{max} \leq \mu$, again using the fact that all the components of $w$ are positive and some component of $y$ is positive so $w^\dagger y > 0$. In particular, if $Ty = \mu y$ then then $\mu = \lambda_{\max}$.

Furthermore, if $y \in Q$ and $Ty \leq \mu y$ then

$$0 < Py = (I + T)^{n-1} y \leq (1 + \mu)^{n-1} y$$

so

$$y > 0.$$

If $\mu = \lambda_{\max}$ then $w^\dagger(Ty - \lambda_{max} y) = 0$ but $Ty - \lambda_{\max} \leq 0$ and so $w^\dagger(Ty - \lambda_{max} y) = 0$ implies that $Ty = \lambda_{\max} y$.

Suppose that $0 \leq S \leq T$ and $Sz = \sigma z,\ z \neq 0$. Then

$$T|z| \geq S|z| \geq |\sigma||z|$$

so

$$|\sigma| \leq L_{\max}(T) = \lambda_{\max},$$

as we have already seen. But if $|\sigma| = \lambda_{\max}$ then $L(|z|) = L_{\max}(T)$ so $|z| > 0$ and $|z$ is also an eigenvector of $T$. with the same eigenvalue. But then $(T - S)|z| = 0$ and this is impossible unless $S = T$ since $|z| > 0$. Replacing the $i$-th row and column of $T$ by zeros give an $S \geq 0$ with $S < T$ since the irreducibility of $T$ precludes all the entries in a row being zero.

Now

$$\frac{d}{d\lambda} \det(\lambda I - T) = \sum_i \det(\lambda I - T_{(i)})$$

and each of the matrices $\lambda_{\max} I - T_{(i)}$ has strictly positive determinant by what we have just proved. This shows that the derivative of the characteristic polynomial of $T$ is not zero at $\lambda_{\max}$, and therefore the algebraic multiplicity and hence the geometric multiplicity of $\lambda_{\max}$ is one. QED

Let us go back to one stage in the proof, where we started with an eigenvector $y$, so $Ty = \lambda y$ and we applied the triangle inequality to get

$$|\lambda||y| \leq T|y|$$

to conclude that $|\lambda| \leq \lambda_{\max}$. When do we have equality? This can happen only if all the entries of $\sum_j t_{ij} y_j$ have the same argument, meaning that all the $y_j$ with $t_{ij} > 0$ have the same argument. If $T$ is primitive, we may apply this same argument to $T^k$ for which all the entries are positive, to conclude that all the entries of $y$ have the same argument. So multiplying by a complex number of arrange value one we can arrange that $y \in Q$ and from $Ty = \lambda y$ that $\lambda > 0$ and hence $\lambda = \lambda_{\max}$ and hence that $y$ is a multiple of $x$. In other words, if $T$ is primitive then we have

$$|\lambda| < \lambda_{\max}$$

for all other eigenvalues.

The matrix of a cyclic permutation has all its eigenvalues on the unit circle, and all its entries zero or one. So without the primitivity condition this result is not true. But this example suggests how to proceed.

For any matrix $S$ let $|S|$ denote the matrix all of whose entries are the absolute values of the entries of $S$. Suppose that $|S| \leq T$ and let $\lambda_{\max} = \lambda_{\max}(A)$, and suppose that $Sy = \sigma y$ for some $y \neq 0$,i.e. that $\sigma$ is an eigenvalue of $S$. Then

$$|\sigma||y| = |\sigma y| = |Sy| \leq |S||y| \leq |S||y|$$

so

$$|\sigma| \leq \lambda_{\max} = L_{\max}(A).$$

Suppose we had equality. Then we conclude from the above proof that $|y| = x$, the eigenvector of $T$ corresponding to $\lambda_{\max}$, and then from the above string of inequalities that $|B|x = Ax$ and since all the entries of $x$ are positive that $|B| = A$. Define the complex numbers of absolute value one

$$e^{i\theta_k} := y_k/|y_k| = y_k/x_k$$

and let $D$ denote the diagonal matrix with these numbers as diagonal entries, so that $y = Dx$. Also write $\sigma = e^{i\phi}\lambda_{\max}$. Then

$$\sigma y = e^{i\phi}\lambda_{\max} Dx = SDx$$

so

$$\lambda_{\max} x = e^{-i\phi} D^{-1} SDx = Tx.$$

Since all the entries of $e^{i\phi} D^{-1} SD$ have absolute values $\leq$ the corresponding entries of $T$, and since all the entries of $X$ are positive, we must have $|e^{i\phi} D^{-1} SD| = T$ and all the rows have a common phase and in fact

$$S = e^{i\phi} DTD^{-1}.$$

In particular, we can apply this argument to $S = T$ to conclude that if $e^{i\phi}\lambda_{\max}$ for some $\phi$ then

$$T = e^{i\phi} DTD^{-1}.$$

Since $DTD^{-1}$ has the same eigenvalues as $T$, this shows that rotation through angle $\phi$ carries *all* the eigenvalues of $A$ into eigenvalues.

The subgroup of rotations in the complex plane with this property is a finite subgroup (hence a finite cyclic group) which acts transitively on the set of eigenvalues satisfying $|\sigma| = \lambda_{\max}$. It also must act faithfully on all non-zero eigenvalues, so the order of this cyclic group must be a divisor of the number of non-zero eigenvalues. If $n$ is a prime and $T$ has no zero eigenvalues then either all the eigenvalues have absolute value $\lambda_{\max}$ or $\lambda_{\max}$ has multiplicity one.

We first define the **period** $p$ of a non-zero non-negative matrix as $T$ follows: For each $i$ consider the set of all positive integers $s$ such that $T^s_{ii} > 0$ and let $p_i$ denote the greatest common denominator of this set. We show that this does not depend on $i$. Indeed, for some other $j$, there is, by irreducibility an integer $M$ such that $T^M_{ij} > 0$ and an integer $N$ such that $T^N_{ji} > 0$. Since $T^{M+N}_{ii} > T^M_{ij} T^N_{ji} > 0$ we conclude that $p_i|(M+N)$ and similarly that $p_j|(M+N)$. Also, if $T^s_{ii} > 0$ then $T^{s+M+N}_{jj} > T^M_{ij} T^s_{ii} T^N_{ji} T^M_{ij} > 0$ so $p_j|s$ and so $p_j|p_i$ and the reverse. Thus $p_i = p_j$, and we call this common value $p$.

Using the arguments above we can be more precise. We claim that $T^s_{ii} = 0$ unless $s$ is a multiple of the order of our cyclic group of rotations, so this order is precisely the period of $T$. Indeed, let $k$ be the order of this cyclic group and $\phi = 2\pi/k$. We have

$$T = e^{i\phi} DTD^{-1}$$

and hence

$$T^s = e^{is\phi} DT^s D^{-1},$$

in particular

$$T^s_{ii} = e^{is\phi} T^s_{ii}.$$

Since $e^{is\phi} \neq 1$ if $s$ is not a multiple of $k$ we conclude that $k = p$. So we can supplement the Perron-Frobenius theorem as

**Theorem 8.4.2 Perron-Frobenius 2.** *If $T$ is primitive, all eigenvalues satisfy $|\sigma| < \lambda_{\max}$. More generally, let $p$ denote the period of $T$ as defined above. Then there are exactly $p$ eigenvalues of $T$ satisfy $|\sigma| = \lambda_{\max}$ and the entire spectrum of $T$ is invariant under the cyclic group of rotations of order $p$.*

## 8.5 Factors of finite shifts.

Suppose that $X$ is a shift of finite type and $\phi : X \to Z$ is a surjective homomorphism, i.e. a factor. Then $Z$ need not be of finite type. Here is an illustrative example. Let $\mathcal{A} = \{0, 1\}$ and let $Z \subset \mathcal{A}^{bfZ}$ consist of all infinite sequences such that there are always an even number of zeros between any two ones. So the excluded words are

$$101,\ 10001,\ 1000001,\ 100000001, \ldots$$

(and all words containing them as substrings). It is clear that this can not be replaced by any finite list, since none of the above words is a substring of any other.

On the other hand, let $G$ be the **DMG** associated with the matrix

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix},$$

and let $Y_G$ be the corresponding shift. We claim that there is a surjective homomorphism $\phi : Y_G \to Z$. To see this, assume that we have labelled the vertices of $G$ as $1, 2$, that we let $a$ denote the edge joining 1 to itself, $b$ the edge joining 1 to 2, and $c$the edge joining 2 to 1. So the alphabet of the graph $Y_G$ and the excluded words are

$$ac\ bb, ba, cc$$

and all words which contain these as substrings. So if $ab$ occurs in an element of $Y_G$ it must be followed by a $c$ and then by a sequence of $bc$'s until the next $a$. Now consider the sliding block code of size 1 given by

$$\Phi :\ a \mapsto 1,\ b \mapsto 0,\ c \mapsto 0.$$

From the above description it is clear that the corresponding homomorphism is surjective.

We can describe the above procedure as assigning "labels" to each of the edges of the graph $G$; we assign the label 1 to the edge $a$ and the label 0 to the edges $b$ and $c$.

It is clear that this procedure is pretty general: a **labeling** of a directed multigraph is a map:$\Phi : \mathcal{E} \to \mathcal{A}$ from the set of edges of $G$ into an alphabet $\mathcal{A}$. It is clear that $\Phi$ induces a homomorphism $\phi$ of $Y_G$ onto some subshift of $Z \subset \mathcal{A}^{\mathbf{Z}}$ which is then, by construction a factor of a shift of finite type.

Conversely, suppose $X$ is a shift of finite type and $\phi : X \to Z$ is a surjective homomorphism. Then $\phi$ comes from some block code. Replacing $X$ by $X^N$ where $N$ is sufficiently large we may assume that $X^N$ is one step and that the block size of $\Phi$ is one. Hence we may assume that $X = Y_G$ for some $G$ and that $\Phi$ corresponds to a labeling of the edges of $G$. We will use the symbol $(G, L)$ to denote a **DMG** together with a labeling of its edges. We shall denote the associated shift space by $Y_{(G,L)}$.

Unfortunately, the term **sofic** is used to describe a shift arising in this way,i.e. a factor of a shift of finite type. (The term is a melange of the modern Hebrew mathematical term *sofi* meaning finite with an English sounding suffix.