Math 118, Spring 2,001

Shlomo Sternberg

January 11, 2001

Contents

Chapter 8

Hyperbolicity.

8.1 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

$$E = S \oplus U \tag{8.1}$$

and a positive constant $a < 1$ so that the estimates

$$\|A_s\| \leq a < 1, \quad A_s = A_{|S} \tag{8.2}$$

and

$$\|A_u^{-1}\| \leq a < 1, \quad A_u = A_{|U} \tag{8.3}$$

hold. (Here, as part of hypothesis (8.3), 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 8.1.1 *Let A be a hyperbolic isomorphism (so that A^{-1} is bounded) and let*

$$\epsilon < \frac{1-a}{\|A^{-1}\|}. \tag{8.4}$$

If ϕ and ψ 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

$$(\mathrm{id} + u) \circ (A + \phi) = (A + \psi) \circ (\mathrm{id} + u) \tag{8.5}$$

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 (8.5) 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

$$\|L^{-1}\| \leq \frac{\|A^{-1}\|}{(1 - a)}. \tag{8.6}$$

Assume, for the moment that we have proved (8.6). 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 (8.6) with (8.4). Thus K is a contraction and we may apply the contraction fixed point theorem to conclude the existence and uniqueness of the solution to (8.5). So we turn our attention to the proof that L is invertible and of the estimate (8.6). 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

$$\|M^{-1}\| \leq \frac{1}{1 - a}. \tag{8.7}$$

Let us write

$$u = f \oplus g, \quad f : E \to S, \quad g : E \to U$$

in accordance with the decomposition (8.1). 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 (8.7) 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 8.1.1 *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 (8.4). 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 (8.4) 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 8.1.1 *There exists neighborhoods U and V of 0 and a homeomorphism $h : U \to V$ such that*

$$h \circ A = f \circ h. \tag{8.8}$$

We prove this theorem by modifying ϕ outside a sufficiently small neighborhood of 0 in such a way that the new ϕ is globally defined and has Lipschitz constant less than ϵ where ϵ satisfies condition (8.4). 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 ϵ 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 ψ. 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 $\text{Lip}[\psi] < \epsilon$ and so have proved the theorem.

8.2 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

$$W^s(p) = W^s(p, f) = \{x| \lim_{n\to\infty} f^n(x) = p\}. \tag{8.9}$$

Similarly, the *unstable manifold* of f at p is defined as

$$W^u(p) = W^u(p, f) = \{x| \lim_{n\to\infty} f^{-n}(x) = p\}. \tag{8.10}$$

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 8.2.1 *Let p be a hyperbolic fixed point of a diffeomorphism, f. For any ball, $B_r(p)$ of radius r about p, let*

$$B_r^s(p) = \{x \in B_r(p) | f^n(x) \in B_r^s(p) \forall n \geq 0\}. \tag{8.11}$$

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

$$W^s(p) = \bigcup_{n \geq 0} f^{-n}[B_r^s(p)]. \tag{8.12}$$

We will prove that $B_r^s(p)$ is a submanifold. It will then follow from (8.12) 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

$$\|f(0)\| \leq \delta \tag{8.13}$$

and

$$\mathrm{Lip}[f - A] \leq \epsilon. \tag{8.14}$$

We wish to prove the following

Theorem 8.2.1 *Let $c < 1$. There exists an $\epsilon = \epsilon(a)$ and a $\delta = \delta(a, \epsilon, r)$ so that if f satisfies (8.13) and (8.14) then there is a map*

$$g : E_u(r) \to E_s(r)$$

with the following properties:
(i) g is Lipschitz with $\mathrm{Lip}[g] \leq 1$.
(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

$$f[graph(v)] = \{(f_s(v([f_u \circ (v, id)]^{-1}(y), y)\} = graphG_[f(v)]$$

where

$$G_f(v) = f_s \circ (v, id) \circ [f_u \circ (v, id)]^{-1}. \tag{8.15}$$

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 ≤ 1. The purpose of the next few lemmas is to show that if ϵ and δ 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 8.2.1 *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 8.2.2 *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*

$$\mathrm{Lip}\left[[f_u \circ (v, id)]^{-1}\right] \leq \frac{1}{c^{-1} - \epsilon}. \tag{8.16}$$

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 ϵ and δ 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/λ about $[f_u \circ (v, id)](0)$ where λ 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 (8.13) holds. We would like this expression to be $\geq r$, which will happen if

$$\delta \leq r(c^{-1} - 1 - 2\epsilon). \tag{8.17}$$

We have thus proved

Proposition 8.2.2 *Let f be a Lipschitz map satisfying (8.13) and (8.14) where $2\epsilon < c^{-1} - 1$ and (8.17) 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

$$2\epsilon < c^{-1} - c \tag{8.18}$$

then

$$\mathrm{Lip}[G_f(v)] \leq 1.$$

Let us now obtain a condition on δ 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

$$\delta \leq r(1 - c - \epsilon). \tag{8.19}$$

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

$$\|f_s(w, x) - G_f(v)(f_u(w, x))\| \leq (c + 2\epsilon)\|w - v(x)\|. \tag{8.20}$$

Indeed,

$$f_s(v(x), x) = G_f(v)(f_u(v(x), x)$$

by the definition of G_f, so we have

$$\begin{aligned}
\|f_s(w,x) - G_f(v)(f_u(w,x))\| &\leq \|f_s(w,x) - f_s(v(x),x)\| + \\
&\quad + \|G_f(v)(f_u((v(x),x) - G_f(v)(f_u(w,x))\| \\
&\leq \mathrm{Lip}[f_s]|(w,x) - (v(x),x)| + \\
&\quad + \mathrm{Lip}[f_u]|(v(x),x) - (w,x)| \\
&\leq \mathrm{Lip}[f_s - p_s A + p_s A]\|w - v(x)\| + \\
&\quad + \mathrm{Lip}[f_u - p_u A]\|w - v(x)\| \\
&\leq (\epsilon + c + \epsilon)\|w - v(x)\|
\end{aligned}$$

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 (8.20) 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 (8.20). 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 (8.20) 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

$$\|G_f(v_1) - G_f(v_2)\|_{\sup} \leq (c + 2\epsilon)\|v_1 - v_2\|_{\sup}. \tag{8.21}$$

Intuitively, what (8.20) 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

$$2\epsilon < 1 - c \tag{8.22}$$

which implies (8.18). To summarize: we have proved that G_f is a contraction in the sup norm on X if (8.17), (8.19) and (8.22) 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 δ are positive.

Now the uniform limit of continuous functions which all have $\mathrm{Lip}[v] \leq 1$ has Lipschitz constant ≤ 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 off^{-1} to $graph(g)$ is conjugated by p_u into $[f_u \circ (g, \mathrm{id})]^{-1}$. But, by (8.16), 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 (8.20) 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 9

Symbolic dynamics.

9.1 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

- ϕ is continuous, and
- $$G \circ \phi = \phi \circ F,$$

  in other words if the diagram

  $$\begin{array}{ccc} M & \xrightarrow{F} & M \\ \phi\downarrow & & \downarrow\phi \\ N & \xrightarrow[G]{} & N \end{array}$$

  commutes.

If the homomorphism ϕ is surjective it is called a **factor**. If ϕ 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 ℓ 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 ℓ, 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

$$X_{\mathcal{F}} = \{x \in \mathcal{A}^{\mathbf{Z}} | [x_k, x_\ell] \not\in \mathcal{F}\}$$

for any k and ℓ. 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 σ is continuous. By abuse of language we may continue to denote the restriction of σ to $X_{\mathcal{F}}$ by σ although we may also use the notation σ_X for this restriction. A dynamical system of the form (X, σ_X) where $x = X_{\mathcal{F}}$ is called a **shift** dynamical system.

Suppose that (X, σ_X) with $X = X_{\mathcal{F}} \subset \mathcal{A}^{\mathbf{Z}}$ and (Y, σ_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 ϕ 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

$$\phi(x)_j = \Phi([x_{j-n}, x_{j+n}]). \tag{9.1}$$

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 ϕ 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 ϕ is continuous. In short, we have proved

Proposition 9.1.1 *A map ϕ 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 ϕ^{-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.

9.2 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

$$\ldots, \begin{pmatrix} x_{-1} \\ x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}, \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}, \begin{pmatrix} x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{pmatrix}, dots.$$

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 $X = X_{\mathcal{F}}$.

9.2.1 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.

9.2.2 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

$$t(u_j) = i(u_j + 1) \tag{9.2}$$

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 (9.2) 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.

9.2.3 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.

9.2.4 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, σ_Y) is the associated shift, then

$$p_n(\sigma_Y) = \operatorname{tr} A^n. \tag{9.3}$$

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

9.2.5 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 γ. The number of elements of period n which correspond to elements of γ 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, σ_Y) for some **DMG**, G. We claim that

$$\zeta_\sigma(t) = \frac{1}{\det(I - tA)}. \tag{9.4}$$

Indeed,

$$p_n(\sigma) = \operatorname{tr} A^n = \sum \lambda_i^n$$

where the sum is over all the eigenvalues (counted with multiplicity). Hence

$$\zeta_\sigma(t) = \prod \exp \sum \frac{(\lambda_i t)^n}{n} = \prod \left(\frac{1}{1 - \lambda_i t} \right) = \frac{1}{\det(I - tA)}. \quad \text{QED}$$

9.3 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 9.3.1 *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

$$h(X) = \lim_{n\to\infty} \frac{1}{n}\log_2 \#(\mathcal{W}_n(X)), \tag{9.5}$$

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 9.3.1 *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 ϕ 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 ϕ is a conjugacy, the reverse inequality applies. *Box*

9.3.1 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 λ_A such that $\lambda_A \geq |\mu|$ for any other eigenvalue μ 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 9.3.1 *Let G be a* **DMG** *whose adjacency matrix $A(G)$ is irreducible. Let y_G be the corresponding shift space. then*

$$h(Y_G) = \lambda_{A(G)}. \tag{9.6}$$

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 λ for λ_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 λ_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.

9.4 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 9.4.1 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 μ is a number such that

$$Ty \leq \mu y$$

then

$$y > 0, \quad \text{and} \quad \mu \geq \lambda_{\max}$$

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 σ 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 λ, 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 η. (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

$$w^\dagger Tx = \eta w^\dagger x = \lambda_{\max} w^\dagger x$$

implying that $\eta = \lambda_{max}$ since $w^\dagger x > 0$.

Now suppose that $y \in Q$ and $Ty \leq \mu y$. Then

$$\lambda_{\max} w^\dagger y = w^\dagger Ty \leq \mu w^\dagger y$$

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 σ 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 ϕ then

$$T = e^{i\phi}DTD^{-1}.$$

Since DTD^{-1} has the same eigenvalues as T, this shows that rotation through angle ϕ 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

$$T^s = e^{is\phi}DT^sD^{-1},$$

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 9.4.2 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.*

9.5 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 Φ induces a homomorphism ϕ 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 ϕ 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 Φ is one. Hence we may assume that $X = Y_G$ for some G and that Φ 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.