A GEOMETRIC CRITERION FOR POSITIVE TOPOLOGICAL ENTROPY

KEITH BURNS AND HOWARD WEISS

Northwestern University and Pennsylvania State University

April 25, 1997 10:57pm

Abstract. We prove that a diffeomorphism possessing a homoclinic point with a topological crossing (possibly with infinite order contact) has positive topological entropy, along with an analogous statement for heteroclinic points. We apply these results to study area-preserving perturbations of area-preserving surface diffeomorphisms possessing homoclinic and double heteroclinic connections. In the heteroclinic case, the perturbed map can fail to have positive topological entropy only if the perturbation preserves the double heteroclinic connection or if it creates a homoclinic connection. In the homoclinic case, the perturbed map can fail to have positive topological entropy only if the perturbation preserves the connection. These results significantly simplify the application of the Poincaré-Arnold-Melnikov-Sotomayor method. The results apply even when the contraction and expansion at the near the fixed point is subexponential.

0. Introduction

Let f be a C^1 diffeomorphism. A homoclinic point for f is a point that is both forward and backwards asymptotic to a hyperbolic periodic point of f (and is not the periodic point itself). In other words, q is a homoclinic point if there is a hyperbolic periodic point p such that $q \in W^s(p) \cap W^u(p)$ and $q \neq p$. A homoclinic point is called transverse if the stable and unstable manifolds intersect transversely there. Poincaré [Po] discovered that the existence of a transverse homoclinic point causes very complicated dynamics. Smale [S] proved that a diffeomorphism f that possesses a transverse homoclinic point has a horseshoe. A horseshoe is a set Λ that is invariant under f^{n_0} for some $n_0 \geq 1$ and has the property that f^{n_0} restricted to Λ is topologically conjugate to a nontrivial subshift of finite type, i.e., there exists a *homeomorphism* $\pi : \Lambda \to \Sigma$ such that $\pi \circ f^{n_0} = \sigma \circ \pi$, where Σ is the shift space and $\sigma : \Sigma \to \Sigma$ denotes the shift map. By a nontrivial subshift of finite type, we mean a subshift that is defined by an irreducible matrix that is not a

1991 *Mathematics Subject Classification.* 58F05, 58F15, 58F15, 58F30.

Key words and phrases. entropy, horseshoes, Melnikov method, perturbations.

The first author was partially supported by a Sloan Foundation Fellowship and an N.S.F. grant. The second author was partially supported by a National Science Foundation Postdoctoral Research Fellowship. Both authors would like to thank MSRI for their support during the period that much of this paper was written.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

permutation; the full two shift is the prototypical example. It is easy to show that a map that possesses a horseshoe has positive topological entropy; see Section 1.

It is natural to consider homoclinic points where the stable and unstable manifolds cross, but may not do so transversely. What one means by "cross" is clear in simple contexts, and can be made precise with ideas from elementary topology; a definition using homology is given in Section 4 .

In the case of a surface diffeomorphism with a homoclinic point where the stable and unstable manifolds cross with finite order contact, Conley [Co] showed that there exist transverse homoclinic points in every neighborhood of the nontransverse homoclinic point and hence the map has a horseshoe and positive topological entropy (see [CR2] for a proof). Conley's result should extend to higher dimensions, but we do not know a reference. The case of a crossing with infinite order contact is much more delicate. Various pathologies develop which do not appear in the finite order case. It seems extremely difficult to establish either the existence of a nearby transverse homoclinic point or the presence of a horseshoe.

The difficulty in finding a horseshoe lies in showing that the map $\pi : \Lambda \to \Sigma$ is 1:1. Let us say that a map f has a **horseshoe factor** if some power of f has a nontrivial subshift of finite type as a topological factor, i.e . there is a subset Λ that is invariant under f^{n_0} for some $n_0 \geq 1$ and a continuous map $\pi : \Lambda \to \Sigma$ that is *onto but not necessarily* 1:1 such that $\pi \circ f^{n_0} = \sigma \circ \pi$, where again Σ is the shift space and σ is the shift map. Any diffeomorphism f possessing a homoclinic point at which the stable and unstable manifolds have a topological crossing has a horseshoe factor. We give an elementary proof of this fact in the case of surface diffeomorphisms in Section 2 and generalize our argument to all dimensions in Section 4 (see Theorems 2.1 and 4.3). Gedeon, McCord and Mischaikow have recently proved a generalization of this result [GMM].

The above result has a natural generalization to the case of a heteroclinic cycle, a chain of hyperbolic fixed points linked by heteroclinic points (as illustrated in Figure 0.1). In arbitrary dimensions, we require that each pair of stable and unstable manifolds forming the chain should cross topologically (Theorem 4.8). In dimension two, it is sufficient that one pair should cross toplogically (Theorem 2.2).

Our results also apply to crossings of center manifolds for indifferent fixed points, and furnish examples of maps having positive topological entropy with no *apparent* hyperbolicity.

It follows from the above results that the diffeomorphisms in question have positive topological entropy. In dimension two, but not in higher dimensions, one can then apply a theorem of Katok [K1, K2] to conclude that if f is $C^{1+\alpha}$ and has positive topological entropy, then f possesses a horseshoe that carries most of the entropy of f.

The proofs of the above theorems contain two stages. The first stage is to describe a method by which one proves that a set is being mapped across itself in the appropriate way to produce a horseshoe factor. The crucial second stage is to identify the relevant sets in each case when we wish to apply the general method. For the first stage of the proof, we use only elementary topology in the case of dimension two, and basic homology theory in arbitrary dimensions. Topological methods which can be used to find horseshoe factors are not new, although we believe that ours is more elementary than previous techniques. Our method is similar to an elegant and very general cohomological method of Easton [E].

FIGURE 0.1. A HETEROCLINIC CYCLE

Easton applied his method to a number of examples, but the only discussion of homoclinic points in [E] is in the context of the standard Smale horseshoe. Churchill and Rod [C, CR] have given cohomological methods, which apply to surface diffeomorphisms. The theorem on page 548 of [Ch] is related to Theorem 2.1 of the present paper.

We apply the results of Section 2 to study area-preserving maps of $\mathbb{R}^2$. Let C^1_{AP} be the group of area-preserving C^1 diffeomorphisms of $\mathbb{R}^2$ with the C^1 topology. Suppose $f \in C^1_{AP}$ and f has two hyperbolic fixed points with a double heteroclinic connection (see Section 1). Such a map is illustrated in Figure 0.2. Consider a small perturbation of f in C^1_{AP}. We show that the perturbed map can fail to have positive topological entropy only if the perturbation preserves the double heteroclinic connection or if it creates a homoclinic connection (see Section 1). A perturbation of the latter type is shown in Figure 0.2. We also show that a small perturbation in C^1_{AP} of a map with a homoclinic connection must produce a map with positive topological entropy, unless the perturbation preserves the homoclinic connection. These theorems are formulated precisely in Section 3. They generalize ad hoc arguments used in [KW].

We believe that these results will have numerous applications. There are many papers in which the authors show that the Hamiltonian flow for a small Hamiltonian perturbation of a completely integrable Hamiltonian system with 2 degrees of freedom has positive topological entropy. This amounts to showing that a small area-preserving perturbation of a Poincaré map having a double heteroclinic connection has positive topological entropy. All authors apply a Poincaré-Arnold-Melnikov-Sotomayor (PAMS) type method [A, M, Po (§403), So] to each branch of the connection and need to verify that the PAMS function has a *simple* zero. This implies that the stable and unstable manifolds for the perturbed systems intersect with transversal crossings. This is usually quite difficult to verify, and with few exceptions (e.g. [Z]), has only been established for non-autonomous equations with periodic or quasi-periodic forcing. Obviously, this places a great limitation on the utility of the method. The results in Section 3 obviate the need to show the existence of a simple zero if one wishes to show that the topological entropy is positive; one needs only

FIGURE 0.2. A MAP WITH A DOUBLE HETEROCLINIC CONNECTION AND A PERTURBATION WITH ZERO TOPOLOGICAL ENTROPY

to show that the PAMS function has a zero (that implies crossing) and is not identically zero (that implies that the stable and unstable manifolds do not coincide) to conclude that the perturbed map has positive topological entropy. This remark considerably simplifies the process of showing a perturbed map has positive topological entropy.

In [KW], the authors construct a convex surface whose geodesic flow has positive topological entropy. They start with a triaxial ellipsoid whose geodesic flow possesses a Poincaré map having a double heteroclinic connection. The authors construct an explicit small local conformal perturbation and show that one connection breaks and the perturbed stable and unstable manifolds intersect (not necessarily transversely or with a topological crossing). It follows immediately from the results in Section 3 that the geodesic flow for the perturbed metric has positive topological entropy.

Another application of the results in Section 3 was discovered by Gabriel Paternain [Pa]. It follows from work of Kozlov [Ko], Tatarinov [T], and Ziglin [Z], that one can find a suitable Poincaré section for the geodesic flow on a Poisson sphere such that the Poincaré map has a double heteroclinic connection and that a small perturbation of the geodesic flow by a potential of Hess-Appel'rot type preserves one branch of the connection and splits the other branch such that the stable and unstable manifolds intersect transversely. Theorem 3.2 immediately implies that the perturbed flow has positive topological entropy. Paternain then applies the Maupertuis Principle to obtain an explicit real analytic metric on S^2 with positive curvature whose geodesic flow has positive topological entropy.

1. Preliminaries

Let M be a smooth manifold and $f : M \to M$ a diffeomorphism. Suppose p is a hyperbolic periodic point of f at which the stable and unstable subspaces have dimensions k and l respectively. The sets $W^s(p) = \{x \in M : f^n(x) \to p\}$ and $W^u(p) = \{x \in M : f^n(x) \to p\}$ are immersed smooth manifolds diffeomorphic to $\mathbb{R}^k$ and $\mathbb{R}^l$ respectively. They are called the **stable** and **unstable manifolds** of p respectively. Since we often want to exclude the point p from consideration, it is convenient to define the **deleted stable** and **unstable manifolds** $\widehat{W}^s(p) = W^s(p) \setminus \{p\}$ and $\widehat{W}^u(p) = W^u(p) \setminus \{p\}$. Note that $\widehat{W}^s(p)$ and $\widehat{W}^u(p)$ are connected, unless they are one dimensional, in which case they have two components.

A **homoclinic point** is a point q such that $q \in \widehat{W}^s(p) \cap \widehat{W}^u(p)$ for some hyperbolic periodic point p. A **heteroclinic point** is a point q such that $q \in \widehat{W}^s(p) \cap \widehat{W}^u(p')$ for distinct hyperbolic periodic points p and p'. Homoclinic and heteroclinic points are called **transverse** if they are points of transverse intersection of the relevant stable and unstable manifolds. A **heteroclinic cycle** is a sequence $p_0, \dots, p_{r-1}$ of hyperbolic periodic points such that $\widehat{W}^s(p_i) \cap \widehat{W}^u(p_{i+1}) \neq \emptyset$ for $i = 0, \dots, r-1$. In dealing with heteroclinic cycles, we make the convention that indices denoting which point in the cycle we are considering should be interpreted modulo r.

Consider now hyperbolic periodic points in a two dimensional manifold that have one dimensional stable and unstable manifolds. A periodic point p of this type has a **homoclinic connection** if one of the components of $\widehat{W}^s(p)$ coincides with one of the components of $\widehat{W}^u(p)$. Two distinct periodic points p and p' of this type have a **heteroclinic connection** if a component of the deleted stable manifold of one the points coincides with a component of the deleted unstable manifold of the other point, and have a **double heteroclinic connection** if one component of $\widehat{W}^s(p)$ coincides with one component of $\widehat{W}^u(p')$ and one component of $\widehat{W}^u(p)$ coincides with one component of $\widehat{W}^s(p')$.

The (two-sided) subshift of finite type Σ_A defined by an $n \times n$ matrix A of 0's and 1's is the subset of $\Sigma_n = \{1, \dots, n\}^{\mathbb{Z}}$ consisting of all sequences which satisfy the condition that a symbol $l \in \{1, \dots, n\}$ can follow a symbol $k \in \{1, \dots, n\}$ if and only if $A_{kl} = 1$. The matrix A is **irreducible** if and only if for each i and j there is an $n \geq 1$ such that the ij'th entry of A^n is positive. The Perron-Frobenius theorem [Ga, §13.2] tells us that A has an eigenvalue λ_A that is real and satisfies $\lambda_A \geq |\lambda|$ for all eigenvalues λ of A. The following proposition is well known (see [W, Theorem 7.13(ii)]).

Proposition 1.1. *Let A be an irreducible $n \times n$ matrix of 0's and 1's. Then the topological entropy of the shift on Σ_A is $\log \lambda_A$.*

The **full n shift** is the special case when A is an $n \times n$ matrix with all entries 1 and $\Sigma_A = \Sigma_n$. It has entropy $\log n$.

Lemma 1.2. *Let A be an irreducible $n \times n$ matrix of 0's and 1's. Then $\lambda_A > 1$ unless A is a permutation matrix.*

Proof. If A is a permutation matrix, then $(1, \dots, 1)$ is an eigenvector with eigenvalue 1. Now suppose that A is not a permutation matrix. Let s and S be the minimum and maximum row sums respectively of A. Then $s \leq \lambda \leq S$, cf [Ga, §13.2]. Since A is irreducible, every row contains at least one 1, and hence $\lambda_A \geq s \geq 1$. We also know that every column of A contains at least one 1 and, since A is not a permutation matrix, at least one column of A contains a second 1. The Perron-Frobenius theorem tells us that A has an eigenvector v all of whose entries are positive. Consider the sum of the entries of Av. Each component of v must appear at least once, and at least one component must appear more than once, so this sum is greater than the sum of the entries of v. Hence 1 is not an eigenvalue of A and we must have $\lambda_A > 1$. $\square$

It follows that a nontrivial subshift of finite type (as defined in the introduction) has positive topological entropy.

Lemma 1.3. *Suppose that the diffeomorphism f has a horseshoe factor. Then f has positive topological entropy.*

Proof. We are assuming that there is a subset Λ that is invariant under f^{n_0} for some $n_0 \geq 1$ and has the property that $f^{n_0}|\Lambda$ is semiconjugate to a nontrivial subshift of finite type σ. Now $h_{TOP}(\sigma) > 0$ and it follows from well known properties of topological entropy [W, Theorems 7.2, 7.10] that

$$h_{TOP}(f) \geq \frac{1}{n_0} h_{TOP}(f^{n_0}|\Lambda) \geq \frac{1}{n_0} h_{TOP}(\sigma) > 0.$$

$\square$

We now develop the geometric machinery that will used in the proofs of theorems 2.1 and 4.3. Let $f : M \to M$ be a diffeomorphism with a hyperbolic fixed point p. Let $k = \dim W^s(p)$, $l = \dim W^u(p)$ and $m = k + l = \dim M$. By choosing a suitable basis for T_pM, we may think of $df(p)$ as a linear map $L : \mathbb{R}^m \to \mathbb{R}^m$ which preserves the splitting $\mathbb{R}^m = \mathbb{R}^k \oplus \mathbb{R}^l$, contracts the $\mathbb{R}^k$ factor and expands the $\mathbb{R}^l$ factor. By the Hartman-Grobman theorem, there are a neighbourhood N of p and a homeomorphism h of N into $\mathbb{R}^m$ with $h(p) = (0, 0)$, such that if $x \in N$ and $f(x) \in N$, then $h(f(x)) = L(h(x))$. We may assume that N and h have been chosen so that $h(N) = D^k(1) \times D^l(1)$, where $D^n(r)$ is the closed disc of radius r about the origin in $\mathbb{R}^n$. We may also assume that $D^k(1) \times \{0\}$ and $\{0\} \times D^l(1)$ lie in $W^s(p) \cap N$ and $W^u(p) \cap N$ respectively. In order to simplify our notation, we shall henceforth identify N with $D^k(1) \times D^l(1)$ and suppress the homeomorphism h. Unless otherwise specified, distances in N will be measured with respect to the product of the Euclidean metrics on $D^k(1)$ and $D^l(1)$. We shall call discs in N of the form $\{x\} \times D^l(1)$ and $D^k(1) \times \{y\}$ **vertical** and **horizontal** respectively.

Lemma 1.4. *Let V be a compact subset of $W^u(p) \cap \operatorname{Int} N$. Suppose we are given $\rho \in (0, 1)$ and $\varepsilon > 0$. Then for any large enough n the following hold.*

(1) $f^{-n}(V) \subset \{0\} \times D^l(\rho)$.
(2) *If $(0, y) \in f^{-j}(V)$, then $f^n(D^k(1) \times \{y\}) \subset N$ and has diameter less than ε.*

Proof. We may assume $\varepsilon < \operatorname{dist}(V, \partial N)$.

Choose a Riemannian metric g on the whole manifold M on which the diffeomorphism f acts. The restriction of g to N may not be equivalent to the Euclidean metric that we are using on N, because $h : N \to D^k(1) \times D^l(1)$ may not be differentiable. The two metrics do, however, define the same topology on N. In particular, there is $\alpha > 0$ such that if $z, z' \in N$ and $\operatorname{dist}_g(z, z') < \alpha$ then $\operatorname{dist}(z, z') < \varepsilon$. Also given $\delta > 0$, there is $\beta(\delta) > 0$ such that for any $z, z' \in N$ with $\operatorname{dist}(z, z') \leq \beta(\delta)$ we have $\operatorname{dist}_g(z, z') \leq \delta$.

There is certainly an n_1 such that for any $n \geq n_1$ we have $f^{-n}(V) \subset \{0\} \times D^l(1)$. Set $B = \sup\{\|Df^{n_1}(x)\| : x \in N\}$. For $n \geq 0$, let

$$R_n = \{x \in N : f^k(x) \in D^k(1) \times D^l(1) \text{ for } 0 \leq k \leq n\}.$$

It is clear that R_n is a union of entire horizontal discs, the image under f^n of a horizontal disc in R_n lies in a horizontal disc, and the sets R_n and $f^n R_n$ both decrease with n.

Choose n_2 so that for any $n \geq n_2$, $R_n \subset D^k(1) \times D^l(\rho)$ and the image under f^n of each disc $D^k(1) \times \{y\}$ in R_n lies in a disc $D^k(\beta(\alpha/B)) \times \{y'\}$ for some $y' \in D^l(1)$. Then any $n \geq n_1 + n_2$ has the desired properties. □

2. Topological crossing of stable and unstable manifolds in a surface

It is intuitively evident what it means for two smooth curves in a surface to cross topologically: they should contain arcs with the behaviour shown in the left half of Figure 2.1. This intuitive notion agrees with the two dimensional version of the definition given in Section 4. In the course of this section and the next, we shall need to consider three different ways in which two curves in a surface can intersect: they may coincide; have a topological crossing; or have a topological tangency, in which case neither of the previous possibilities occurs (see the right half of Figure 2.1).

FIGURE 2.1. TOPOLOGICAL CROSSING AND TOPOLOGICAL TANGENCY

Theorem 2.1. *Let $f : M \to M$ be a diffeomorphism of a surface with a hyperbolic periodic point p. Assume that one component of $\widehat{W}^s(p)$ and one component of $\widehat{W}^u(p)$ have a topological crossing. Then some power of f has the full two shift as a topological factor.*

Remarks. We emphasize that the topological crossing need not be transverse and could be of infinite order.

Next suppose that $p_0, \ldots, p_{r-1}$ is a heteroclinic cycle in a surface. Then there are components $V^s(p_i)$ of $\widehat{W}^s(p_i)$ and $V^u(p_i)$ of $\widehat{W}^u(p_i)$, such that $V^u(p_i) \cap V^s(p_{i+1}) \neq \emptyset$ for $i = 0, \ldots, r-1$. (We remind the reader that in this context indices should be interpreted modulo r.)

Theorem 2.2. *Let $p_0, \ldots, p_{r-1}$ be a heteroclinic cycle in a surface M and let $V^s(p_i)$ and $V^u(p_i)$ be defined as above. Suppose that $V^u(p_0)$ and $V^s(p_1)$ cross topologically. Then some power of f has the full two shift as a topological factor and thus $h_{TOP}(f) > 0$.*

We now describe the method that will be used to prove the above theorems. Let $N \subseteq M$ be homeomorphic to $[-1,1] \times [-1,1]$. In the following we shall identify N with $[-1,1] \times [-1,1]$ and suppress the homeomorphism. Let $R = [-1,1] \times [-\rho, \rho]$, where $\rho \in (0,1)$.

We shall call a set $V \subset \operatorname{Int} N$ **essential** if V contains a curve joining the two components of $N \setminus R$.

A set $S \subset R$ will be called a **horizontal strip** if

(1) S is closed and path connected;
(2) S contains a curve joining the left edge $\{-1\} \times [-\rho, \rho]$ and the right edge $\{1\} \times [-\rho, \rho]$ of R;
(3) ∂S is a Jordan curve curve which is the union of a finite number of arcs all of whose endpoints lie on the left edge or the right edge of R.

It is easily seen that ∂S contains exactly two curves joining the left edge $\{-1\} \times [-\rho, \rho]$ and the right edge $\{1\} \times [-\rho, \rho]$ of R, and S lies in the region of R bounded by these curves; see Figure 2.2. We shall call the curve on which the second coordinate is larger c_{upper} and the other curve c_{lower}.

Figure 2.2. A Horizontal Strip

Definition 2.3. Let n be a positive integer and S a horizontal strip. We shall say that f^n **stretches** S **across** R if $f^n S \subset \operatorname{Int} N$, $f^n(\partial S \cap \operatorname{Int} R) \subset N \setminus R$, and f^n maps c_{upper} and c_{lower} into opposite components of $N \setminus R$.

Theorem 2.4. *Suppose N contains two disjoint closed horizontal strips S_0 and S_1 that are stretched across R by f^{n_0} for some $n_0 \geq 1$. Then f^{n_0} has the full two shift as a topological factor.*

Proof. Let $\Lambda = \bigcap_{k=-\infty}^{\infty} f^{kn_0}(S_0 \cup S_1)$. Define $\pi : \Lambda \to \Sigma_2$ so that for any $z \in \Lambda$ the k'th term of the sequence $\pi(z)$ is 0 if $f^{kn_0}z \in S_0$ and 1 if $f^{kn_0}z \in S_1$. The continuity of f and f^{-1} implies that for any $K > 0$ there is a neighbourhood $N(z)$ of z in R such that if $z' \in N(z) \cap \Lambda$, then $(\pi(z))_k = (\pi(z'))_k$ for $-K \le k \le K$. Thus π is continuous. It is easily verified that $\pi \circ f^{n_0}|\Lambda = \pi \circ \sigma$, where σ is the shift on Σ_2. It remains to show that π is surjective.

Key Lemma 2.5. *Let S be a horizontal strip and V an essential subset of N. Suppose that f^n stretches S across R. Then $f^n(S \cap V)$ is essential.*

Proof. It is obvious that V contains an arc $\alpha : [0,1] \to R$ such that $\alpha(0)$ lies on c_{upper} and $\alpha(1)$ lies on c_{lower} and $\alpha(t) \in \operatorname{Int} R$ for $0 < t < 1$. Let us call a component of $\partial S \cap \operatorname{Int} R$ positive if f^n maps it into the same component of $N \setminus R$ as c_{upper} and negative otherwise. Let $t_0 \in [0,1)$ be the last time that $\alpha(t)$ is in a positive component of $\partial S \cap \operatorname{Int} R$ and $t_1 \in (t_0, 1]$ the next time that $\alpha(t)$ is in ∂S. Observe that if α exits S through a given component of $\partial S \cap \operatorname{Int} R$, then α must reenter S through the *same* component of $\partial S \cap \operatorname{Int} R$. It is clear from this that the arc of α between $\alpha(t_0)$ and $\alpha(t_1)$ must lie in S. Its image under f^n is a curve in $\operatorname{Int} N$ that joins the two components of $N \setminus R$. Hence $f^n(S \cap V)$ is essential. □

Lemma 2.6. *For each $r \ge 1$, the set $f^{n_0}S_{\alpha_{-1}} \cap f^{2n_0}S_{\alpha_{-2}} \cap \cdots \cap f^{rn_0}S_{\alpha_{-r}}$ is essential for every $\alpha \in \Sigma_2$.*

Proof. Use induction on r. The case $r = 1$ is trivial. Suppose the lemma holds for r. Then $V = f^{n_0}S_{\alpha_{-2}} \cap \cdots \cap f^{rn_0}S_{\alpha_{-r-1}}$ is essential for any $\alpha \in \Sigma_2$. It follows from the Key Lemma that

$$f^{n_0}(S_{\alpha_1} \cap V) = f^{n_0}S_{\alpha_1} \cap f^{2n_0}S_{\alpha_{-2}} \cap \cdots \cap f^{(r+1)n_0}S_{\alpha_{-r-1}}$$

is essential. □

Corollary 2.7. *For each $\alpha \in \Sigma_2$ and each $r \ge 0$, the set $\Lambda_r(\alpha) = \bigcap_{k=-r}^{r} f^{kn_0}S_{\alpha_{-k}} \neq \emptyset$.*

Proof. Let $\beta = \sigma^{r+1}\alpha$, where σ is the shift on Σ_2. Then

$$\Lambda_r(\alpha) = f^{-(r+1)n_0}(S_{\beta_{-1}} \cap \cdots \cap f^{2rn_0}S_{\beta_{-2r-1}}),$$

which is nonempty by the previous lemma. □

Since the sets $\Lambda_r(\alpha)$ are compact and $\Lambda_0(\alpha) \supset \Lambda_1(\alpha) \supset \ldots$, there is a point $z \in \bigcap_{r=0}^{\infty} \Lambda_r(\alpha)$ for each α. It is obvious that $\pi(z) = \alpha$. Thus π is surjective and the proof of Theorem 2.4 is complete. □

We now apply Theorem 2.4 to prove Theorems 2.1 and 2.2.

Proof of Theorem 2.1. After possibly replacing f by a power of f, we may assume that p is a fixed point of f. We use the notation introduced at the end of Section 1.

The first observation is that $W^u(p) \cap N$ contains two arcs α_1 and α_2 that cross $[-1,1] \times \{0\}$ topologically. We choose α_1 to be $[-1,1] \times \{0\}$. Now let α be any arc that lies in $\widehat{W}^s(p)$ and crosses $\widehat{W}^u(p)$ topologically. After replacing α by its image under a large enough power of f, we may assume that the intersection of α with $\widehat{W}^s(p)$ is contained in $[-1,1] \times \{0\}$. One of the components of $\alpha \cap N$ can now be chosen to be α_2.

Choose $\rho > 0$ such that the endpoints of α_0 and α_1 lie outside $R \stackrel{\text{def}}{=} [-1,1] \times [-\rho, \rho]$.

Lemma 2.8. *For any large enough n there are disjoint closed horizontal strips $S_0(n) \supset f^{-n}\alpha_0$ and $S_1(n) \supset f^{-n}\alpha_1$ that are stretched across R by f^n.*

Proof. Choose $\varepsilon > 0$ so that the closed ε-neighbourhoods of the end points of α_0 and α_1 lie in $\operatorname{Int} N \setminus R$ and the closed ε-neighbourhoods of α_0 and α_1 are disjoint and lie in $\operatorname{Int} N$. Let n be large enough so that the conclusions of Lemma 1.4 hold with $V = \alpha_0$ or α_1. For $i = 0, 1$, let

$$S_i(n) = [-1,1] \times \{y \in [-\rho,\rho] : (0,y) \in f^{-n}(\alpha_i)\}.$$

Each of the sets S_i is a rectangle of the form $[-1,1]\times[y_{i0}, y_{i1}]$. These are certainly horizontal strips; in S_i the curve c_{upper} will be $[-1,1] \times \{y_{i1}\}$ and c_{lower} will be $[-1,1] \times \{y_{i0}\}$. It is clear from our construction that $S_0(n)$ and $S_1(n)$ are closed and disjoint, and that α_i lies in $f^n S_i(n)$ for $i = 0, 1$. Each segment $[-1,1] \times \{y\}$ in one of the S_i is mapped by f^n to a curve of length less than ε which contains a point on one of the α_i, and consequently lies in $\operatorname{Int} N$. The edges $[-1,1] \times \{y_{i0}\}$ and $[-1,1] \times \{y_{i1}\}$ of S_i map into the ε-neighbourhoods of the endpoints of α_i, and thus map into different components of $N \setminus R$. □

Theorem 2.1 follows immediately from the above lemma and Theorem 2.4. □

Proof of Theorem 2.2. After replacing f by one of its powers, and the original cycle by a possibly longer one containing the orbits of the periodic points involved in the original cycle, we may assume that all of the periodic points $p_0, \ldots, p_{r-1}$ are fixed. We shall assume that the surface is oriented and the diffeomorphism f is orientation preserving. If the surface is not orientable, we can replace f by a lift to the double cover of the surface; and if f is not orientation preserving we can replace f by f^2.

Choose a smooth embedding of $[-1,1]\times[-1,1]$ whose image N is as shown in Figure 2.3. The embedding is chosen so that $\{-1\} \times [-1,1]$ maps into $V^u(p_0)$ and there is $\rho \in (0,1)$ such that

(1) $[-1,1] \times \{-\rho\}$ and $[-1,1] \times \{\rho\}$ map into $V^s(p_1)$.
(2) The image R of $[-1,1] \times [-\rho,\rho]$ lies on the same side of $V^u(p_0)$ as $V^s(p_0)$ and on the same side of $V^s(p_1)$ as $V^u(p_1)$.

Our strategy is to show that any sufficiently large iterate f^n maps R across itself twice as shown in Figure 2.3. We can then apply the argument in the proof of Theorem 1. Let α and β be the **vertical sides** of R, namely $\{-1\} \times [-\rho,\rho]$ and $\{-1\} \times [-\rho,\rho]$ respectively. We first show that α and β map across R in the proper way.

In the following discussion all curves are embedded curves. Pick a number $\sigma \in (\rho,1)$ such that $f^n\alpha \cup f^n\beta$ is transverse to $[-1,1] \times [-\sigma,\sigma]$ for all n.

Definition 2.9. Let $\gamma : [0,1] \to M$ be a curve. We shall call γ **good** if for all sufficiently large n, there is $t_n \in [0,1]$ such that

(1) $f^n \circ \gamma(0) \notin N$;
(2) $f^n \circ \gamma([0,t_n])$ does not intersect any part of the boundary of N except $(-1,1)\times\{-1\}$;
(3) $f^n \circ \gamma(t_n) \in (-1,1) \times (\sigma,1)$.

This is obviously an open property with respect to the Hausdorff topology on compact curves.

Figure 2.3. Proof of Theorem 2.2

Definition 2.10. Let η be a compact curve and γ a curve. We say that $f^n\gamma$ accumulates on η if there exist compact curves $\gamma_n \subset \gamma$ such that $f^n\gamma_n$ converges to η in the Hausdorff topology on compact curves.

We shall use the following obvious lemma.

Lemma 2.11. *Suppose γ is a curve that accumulates on a compact curve η. If η is good, then γ is good.*

Proposition 2.12. *Any curve which has one endpoint on $V^s(p_1)$ and lies on the same side of $V^s(p_1)$ as $V^u(p_1)$ is good.*

Proof. Divide the integers modulo r into blocks of consecutive integers so that i and j are in the same block if and only if $V^u(p_k) = V^s(p_{k+1})$ for $i \leq k < j$. Observe that 1 is the first element of a block.

Let $i_1 = 1, \cdots, i_k$ be the first elements of the blocks. Observe that if γ is curve a with an endpoint in $V^s(p_{i_j})$ that lies on the same side of $V^s(p_{i_j})$ as $V^s(p_{i_j})$, then $f^n\gamma$ accumulates onto every compact curve contained in $\{p_{i_j}\} \cup V^u(p_{i_j}) \cup \cdots \{p_{i_{j+1}-1}\} \cup V^u(p_{i_{j+1}-1})$.

We now prove by a descending induction that if $1 \leq j \leq k$, then any curve with an endpoint on $V^s(p_{i_j})$ that lies on the same side of $V^u(p_{i_j})$ as $V^s(p_{i_j})$ is good. The case $j = k$ is obvious. The inductive step is contained in the following lemma. $\square$

Lemma 2.13. *Suppose γ is a curve that has an endpoint on $V^s(p_{i_j})$ and lies on the same side of $V^s(p_{i_j})$ as $V^u(p_{i_j})$. Then at least one of the following occurs:*

1. *$f^n\gamma$ accumulates on a curve β that has an endpoint on $V^s(p_{i_{j+1}})$ and lies on the same side of $V^s(p_{i_{j+1}})$ as $V^u(p_{i_{j+1}})$.*
2. *$f^n\gamma$ contains a curve α that has an endpoint on $V^s(p_{i_{j+1}})$ and lies on the same side of $V^s(p_{i_{j+1}})$ as $V^u(p_{i_{j+1}})$.*

Proof. There are ways (illustrated in Figure 2.4) in which $V^u(p_{i_{j+1}-1})$ and $V^s(p_{i_{j+1}})$ can intersect. In Case I they have with a topological crossing; both phenomena (1) and (2) occur. In case II they have a topological tangency and $V^u(p_{i_{j+1}-1})$ lies on the same side of $V^s(p_{i_{j+1}})$ as $V^u(p_{i_{j+1}})$; only phenomenon (1) occurs. In case III they have a topological tangency and $V^u(p_{i_{j+1}-1})$ lies on the opposite side of $V^s(p_{i_{j+1}})$ as $V^u(p_{i_{j+1}})$; only phenomenon (2) occurs. $\square$

Consider curves which are arcs of $f^n\alpha \cup f^n\beta$, join $[-1,1] \times \{\sigma\}$ to $[-1,1] \times \{-\sigma\}$, lie in the interior of N, and do not contain any shorter curves with these properties. We shall call such curves **connecting curves**. Two connecting curves cannot intersect because they lie in $f^n\alpha \cup f^n\beta$ which is the homeomorphic image of two intervals. There are only finitely many connecting curves, because any horizontal line $[-1,1] \times \{y\}$ that is transverse to $f^n\alpha \cup f^n\beta$ contains only finitely many points that can lie on connecting curves. The first coordinate in N induces an ordering on the connecting curves in an obvious way.

Lemma 2.14. *Two connecting curves are adjacent in this ordering if and only they are joined by a curve in $[-1,1] \times (-\sigma, \sigma)$ that does not intersect $f^n\alpha \cup f^n\beta$.*

Proof. It is obvious that two connecting curves are adjacent if they are joined by a curve in $[-1,1] \times (-\sigma, \sigma)$ that does not intersect $f^n\alpha \cup f^n\beta$.

Figure 2.4. The Three Cases in Lemma 2.13

Conversely, suppose that c_1 and c_2 are adjacent curves. Choose a segment γ of the form $[x_1, x_2] \times \{y_1\}$ that joins c_1 to c_2 and is transverse to $f^n\alpha \cup f^n\beta$. Starting at (x_1, y_1), follow γ until it intersects $f^n\alpha \cup f^n\beta$. On one side of γ or the other we can now move along $f^n\alpha$ or $f^n\beta$ and intersect γ before hitting the lines $[1, 1] \times \{\pm 1\}$. We then continue along γ to the next intersection with $f^n\alpha \cup f^n\beta$. After finitely many steps one reaches (x_2, y). $\square$

Lemma 2.15. *Suppose γ lies in N, joins $f^n\alpha$ and $f^n\beta$ and does not contain any curve that joins $f^n\alpha$ and $f^n\beta$. Then $f^{-n}\gamma$ lies in R.*

Proof. Clearly $f^{-n}\gamma$ cannot cross α or β. Also $f^{-n}\gamma$ cannot cross the piece δ of $V^s(p_1)$ marked in Figure 2.5, for $f^n\delta$ lies near p_1 and is disjoint from N. $\square$

FIGURE 2.5. PROOF OF LEMMA 2.15

We divide the connecting curves into two classes — those of type α that lie in $f^n\alpha$ and those of type β that lie in $f^n\beta$.

Lemma 2.16. *There are connecting curves $\widehat{\alpha}_1, \widehat{\alpha}_2$ of type α and $\widehat{\beta}_1, \widehat{\beta}_2$ of type β such that each pair $\widehat{\alpha}_i, \widehat{\beta}_i$, $i = 1, 2$, is adjacent with respect to the ordering described above.*

Proof. Let x and y be the points where the two arcs of $f^n\alpha$ that emanate from $W^s(p_1)$ first intersect $[-1, 1] \times \{\sigma\}$. These arcs, the piece of $W^s(p_1)$ that join the initial points, and the segment of $[-1, 1] \times \{\sigma\}$ form a Jordan curve. The two arcs β_1, β_2 of $f^n\beta$ that emanate from $W^s(p_1)$ and end at the first intersection with $[-1, 1] \times \{\sigma\}$ enter the interior of the Jordan curve described above and do not cross it. Let α_i^* (resp. β_i^*) be the connecting curves of type α (resp. β) contained in α_i (resp. β_i). It is clear that β_1^* and β_2^* lie in between α_1^* and α_2^*; with the right choice of numbering we may assume that they occur in the order α_1^*, β_1^*, β_2^*, and α_2^*. There must be connecting curves $\widehat{\alpha}_i$ and $\widehat{\beta}_i$ between each pair α_i^* and β_i^* which have the desired properties. $\square$

Lemma 2.17. *Suppose that $\widehat{\alpha}$ and $\widehat{\beta}$ are adjacent connecting curves Then the region in $[-1,1]\times[-\sigma,\sigma]$ between $\widehat{\alpha}$ and $\widehat{\beta}$ contains the image under f^n of a horizontal strip that is stretched across R by f^n .*

Proof. By Lemma 2.14, there is a curve $\gamma \subset [-1,1]\times(-\sigma,\sigma)$ that joins $\widehat{\alpha}$ to $\widehat{\beta}$ and does not intersect $f^n\alpha \cup f^n\beta$. Let C be the component of $[-1,1]\times(-\sigma,\sigma)\setminus(f^n\alpha\cup f^n\beta)$ that contains γ. Let $S = f^{-n}C$. Note that $(\alpha\cup\beta)\cap S = \emptyset$. The other sides of R, namely $\{-1\}\times[-\rho,\rho]$ and $\{1\}\times[-\rho,\rho]$ lie on $W^s(p_1)$ and do not map into N under f^n. Hence $f^n(\partial R)\cap C = \emptyset$ and $S\cap\partial R = \emptyset$. But S is connected and contains $f^{-n}\gamma$, which lies in R by Lemma 2.15. Hence $S \subset R$.

Since $f^n(\alpha\cup\beta)$ is transverse to $[-1,1]\times\{-\sigma,\sigma\}$, we see that ∂C is a Jordan curve consisting of a finite number of arcs belonging to $f^n\alpha$, $f^n\beta$, $[-1,1]\times\{-\sigma\}$ and $[-1,1]\times\{\sigma\}$. It follows that ∂S is a Jordan curve consisting of a finite number of arcs belonging to α, β, $f^{in}([-1,1]\times\{-\sigma\})$ and $f^{-n}([-1,1]\times\{\sigma\})$. The endpoints of these arcs must lie on $\alpha\cup\beta$. Furthermore S is obviously closed and path connected and contains the curve $f^{-n}\gamma$ that joins α to β. Thus S is a horizontal strip.

We now show that f^n stretches $\overline{S}$ across R. It is clear that f^n maps S into $\operatorname{Int} N$ and maps $\partial S\cap\operatorname{Int} R$ into $[-1,1]\times\{-\sigma,\sigma\}\subset N\setminus R$, so we have only to show that c_{upper} and c_{lower} map into opposite components of $N\setminus R$. The curves c_{upper} and c_{lower} do not belong to $\alpha\cup\beta$, so f^n must map them into $[-1,1]\times\{-\sigma,\sigma\}$. Since c_{upper} and c_{lower} lie on opposite sides of $f^{-n}\gamma$ in R, their images under f^n lie on opposite sides of γ in the region of $[-1,1]\times[-\sigma,\sigma]$ that lies between $\widehat{\alpha}$ and $\widehat{\beta}$. Hence c_{upper} and c_{lower} map into opposite components of $N\setminus R$. ☐

We see from the two previous lemmas that the hypotheses of Theorem 2.4 are satisfied. This completes the proof of Theorem 2.2. ☐

Remarks.

(i) In the proofs of the above theorems we have not used hyperbolicity of the fixed points: it would suffice for them to have stable manifolds that contract under f, and unstable manifolds that contract under f^{-1} at a subexponential rate. It is not difficult to devise explicit examples that have fixed points with these properties. There are also natural examples in celestial mechanics.

(ii) The hypothesis in Theorem 2.1 that there be a topological crossing is essential. Gavrilov and Silnikov [GS] have constructed a surface diffeomorphism with zero topological entropy which has a homoclinic point at which the stable and unstable manifolds have a topological tangency.

3. Perturbations of area-preserving maps of $\mathbb{R}^2$

Let C^1_{AP} denote the space of C^1 area-preserving diffeomorphisms of $\mathbb{R}^2$. Suppose $f\in C^1_{AP}$ has a hyperbolic fixed point p with a homoclinic connection. The structural stability of hyperbolic fixed points ensures that any diffeomorphism g that is close enough to f in the C^1 topology will have a fixed point p_g that is close to p.

Suppose that $V^s(p)$ and $V^u(p)$ are the components of $\widehat{W}^s(p)$ and $\widehat{W}^u(p)$ which coincide. Let $V^s_g(p_g)$ and $V^u_g(p_g)$ be the components of $\widehat{W}^s_g(p_g)$ and $\widehat{W}^u_g(p_g)$ that are close to $V^s(p)$ and $V^u(p)$ respectively.

Theorem 3.1. *There exists a neighborhood $\mathcal{U}$ of f in C^1_{AP} such that if $g \in \mathcal{U}$ and $V^s_g(p_g) \neq V^u_g(p_g)$, then $h_{TOP}(g) > 0$.*

Proof. In view of Theorem 2.1, we have only to show that p_g must have a homoclinic point if g is close enough to f. It is a general fact, which goes back to Poincaré, that homoclinic points persist under small small perturbations in C^1_{AP} (see [MM] for an exposition of the proof).

We now give a proof that any g close enough to f has a homoclinic point, which we will modify in the proof of the next theorem.

Let Γ^s and Γ^u denote $V^s_g(p_g)$ and $V^u_g(p_g)$ respectively. Pick nearby points $x_s \in \Gamma^s$ and $x_u \in \Gamma^u$. Let Δ_s and Δ_u be the arcs of Γ^s and Γ^u respectively joining x_s, x_u to $g(x_s), g(x_u)$. Then Δ_s and Δ_u are C^1 close and we can find a short arc α from x_s to x_u such that

$$\alpha \cap (g^{-1}(\Delta_s) \cup \Delta_s) = \{x_s\} \quad \text{and} \quad \alpha \cap (g^{-1}(\Delta_u) \cup \Delta_u) = \{x_u\}.$$

Suppose now that p_g does not have a homoclinic point. Then $\alpha \cup \Delta_s \cup g(\alpha) \cup \Delta_u$ forms a Jordan curve that bounds a compact region R_0 with positive area.

Let R be the compact region bounded by the arc of Γ^s from p_g to x_s, α, Δ_u and the arc of Γ^u from $g(x_u)$ to p_g. Then $g(R)$ is the compact region bounded by the arc of Γ^s from p_g to $g(x_s), \Delta_s$, $g \circ \alpha$, and the arc of Γ^u from $g(x_u)$ to p_g. One sees that either $R \subset gR$ or $gR \subset R$, and that in either case their areas differ by area(R_0). This is impossible if g is area-preserving. □

Now suppose $f \in C^1_{AP}$ has a a double heteroclinic connection between hyperbolic fixed points p and p'. We describe natural and easily verifiable geometric conditions that will ensure that a nearby area-preserving map will have positive topological entropy. As explained in the introduction the essence of the result is that a small area-preserving perturbation of f can fail to produce positive topological entropy only if it preserves the double heteroclinic connection or if it creates a homoclinic connection.

The structural stability of hyperbolic fixed points ensures that any diffeomorphism g that is close enough to f in the C^1 topology will have fixed points p_g and p'_g that are close to p and p' respectively. Suppose that $V^s(p)$, $V^u(p)$, $V^s(p')$, and $V^u(p')$ are the components of $\widehat{W}^s(p)$, $\widehat{W}^u(p)$, $\widehat{W}^s(p')$, and $\widehat{W}^u(p')$ which are involved in the double heteroclinic connection described above, so that $V^s(p) = V^u(p')$ and $V^u(p) = V^s(p')$. Let $V^s_g(p_g)$ be the component of $\widehat{W}^s_g(p_g)$ that is close to $V^s(p)$, and define $V^u_g(p_g)$, $V^s_g(p'_g)$, and $V^u_g(p'_g)$ analogously.

Theorem 3.2. *There exists a neighborhood $\mathcal{U}$ of f in C^1_{AP} such that if $g \in \mathcal{U}$ and*

(1) $V^s_g(p_g) \neq V^u_g(p'_g)$ *or* $V^u_g(p_g) \neq V^s_g(p'_g)$ *and*
(2) $V^s_g(p_g) \neq V^u_g(p_g)$ *and* $V^s_g(p'_g) \neq V^u_g(p'_g)$,

then $h_{TOP}(g) > 0$.

The following corollary, which contains the result used in [KW], is an easy consequence of Theorem 3.2.

Corollary 3.3. *There exists a neighborhood $\mathcal{U}$ of f in C^1_{AP} such that if $g \in \mathcal{U}$, $V^s_g(p_g) \neq V^u_g(p'_g)$ and $V^s_g(p_g) \cap V^u_g(p'_g) \neq 0$, then $h_{TOP}(g) > 0$.*

Remarks.

(1) The above results include the case where one of the connections is not broken.
(2) The area-preserving serving hypothesis is easily seen to be essential in these results.
(3) We emphasize that we are considering a small perturbation of f.

In our proof of Theorem 3.2 we shall use the fact that g can be linearized near p_g by a C^1 change of coordinates that depends continuously on g. This C^1 linearization is not essential for the proof of this theorem but it significantly simplifies the exposition.

Linearization Theorem 3.4. *Let f be a diffeomorphism of $\mathbb{R}^2$ with a hyperbolic fixed point p. For g close to f in the C^1 topology, let p_g be the hyperbolic fixed point of g that is close to p. Let λ_g and μ_g be the eigenvalues of $D_{p_g}g$ on the stable and unstable subspaces at p_g respectively. Let A_g be the linear map of $\mathbb{R}^2$ defined by the matrix*

$$\begin{pmatrix} \lambda_g & 0 \\ 0 & \mu_g \end{pmatrix}.$$

Then there exists a neighborhood U of p diffeomorphic to the open unit ball B in $\mathbb{R}^2$, a neighborhood $\mathcal{N} \subset Diff^1(\mathbb{R}^2)$ of f, and a continuous mapping $H : \mathcal{N} \to Diff^1(U, B)$ such that $A_g \circ H(g) = H(g) \circ g$ on $U \cap g^{-1}U$ for each $g \in \mathcal{N}$.

Remarks. This theorem is due to Hartman [Ha]. It is false in higher dimensions and is false if one asks for the linearization to be C^2 or smoother. A sketch of his theorem, in modern terms, is to take a local diffeomorphism f having 0 as a hyperbolic fixed point and extend it to a global hyperbolic diffeomorphism of $\mathbb{R}^2$ which is linear outside of a ball around 0. One can easily show that the stable and unstable manifolds are C^1 ([HPS]). One then applies the one-dimensional Sternberg Linearization Theorem [St] to the local stable and unstable manifolds of 0. The theorem immediately follows by observing the continuous dependence of the conjugating map in Sternberg's proof in the C^1 topology. We thank A. Katok for helpful remarks about this proof.

Proof of Theorem 3.2. Choose U and the linearizing map H as above. We may assume that $H(f)$ maps the region inside the double heteroclinic connection for f into the first quadrant of $\mathbb{R}^2$.

Define the coordinates x_g and y_g on U by $H(g)(q) = (x_g(q), y_g(q))$. Let $\mathcal{F}$ and $\mathcal{F}^\perp$ be the foliations of $U \setminus \{p_g\}$ that are the level curves of the functions $Q(q) = x_g(q)y_g(q)$ and $q \to x_g(q)^2 - y_g(q)^2$ respectively. We say that q_1 **is above** q_2 if q_1 and q_2 lie on the same leaf of $\mathcal{F}^\perp$ and $Q(q_1) > Q(q_2)$. Suppose that α_1 and α_2 are C^1 arcs which are transverse to $\mathcal{F}^\perp$ and have the same endpoints. We say that α_1 and α_2 are **equivalent** if the regions $R_1 = \{q : q$ lies above a point of α_1 and below a point of $\alpha_2\}$ and $R_2 = \{q :$ q lies above a point of α_2 and below a point of $\alpha_1\}$ have the same area. It is easily seen that p_g cannot lie between two equivalent curves. Furthermore two equivalent curves must either cross toplogically or coincide.

Now suppose that W is a connected C^1 curve in U that is part of a curve that is invariant under g. We say that W **is carried by** a leaf L of $\mathcal{F}$ if there is a point $q \in W \cap L$ such that the arc of W from q to $g(q)$ lies in U, is transverse to $\mathcal{F}^\perp$ and is equivalent to the arc of L from q to $g(q)$. See Figure 3.1.

FIGURE 3.1. W IS CARRIED BY L

An easy continuity argument proves the following lemma.

Lemma 3.5. *Let W be a connected C^1 curve in U that is part of a curve that is invariant under g. Suppose that for some $q \in W$ the arc of W from $g^{-1}(q)$ to $g(q)$ lies in $U \setminus \{p_g\}$ and is transverse to $\mathcal{F}^\perp$. Then there is a unique leaf L of $\mathcal{F}$ that carries W.*

Choose a compact neighborhood U_0 of p_g with the property that $\bigcup_{n=-100}^{100} g^n(U_0) \subset U$. Let q_s be the first point on $V_g^s(p_g')$ (as one moves along $V_g^s(p_g')$ starting from p_g') that lies in U_0. Let W^s be the path component of $U \cap V_g^s(p_g')$ that contains q_s. Define W^u analogously. Let L^s and L^u be the leaves that carry W^s and W^u respectively.

If $L^s = L^u$, then p_g will have a homoclinic point in U, and it follows from Theorem 3.1 that either g has positive topological entropy or there exists a homoclinic connection for p_g'. Henceforth we will assume that $L^s \neq L^u$.

Let Γ^s be the curve that starts at p_g', follows $V_g^s(p_g')$ until the first intersection with L^s and then follows L^s (in the direction from q towards $g(q)$, where q is the point of changeover from $V_g^s(p_g')$ to L^s). Define Γ^u analogously. Suppose that we can choose points x_s on Γ^s and x_u on Γ^u such that the arcs from x_s to $g(x_s)$ and from x_u to $g(x_u)$ are C^1 close. Then we can use the argument in the proof of Theorem 3.1 to see that area preservation will be violated unless p_g has a homoclinic connection. Even though Γ^s and Γ^u are no longer invariant, their construction ensures that the area-preservation argument still applies.

It is clear that we can choose x_s and x_u with the desired properties if L^s or L^u lies in the interior of the first quadrant, as illustrated in Figure 3.2.

FIGURE 3.2.

If neither L^s nor L^u lies in the interior of the first quadrant, $V_g^s(p_g)$ and $V_g^u(p_g)$ lie "inside" Γ^s and Γ^u. If we interchange the roles of p_g and p_g' in the above argument, we will then find that the L^s and L^u constructed from $V_g^s(p_g)$ and $V_g^u(p_g)$ lie in the closure of the first quadrant in the linearizing coordinates around p_g'; see Figure 3.3. Moreover, if one of the original L^s and L^u lay outside the closure of the first quadrant (in the coordinates near p_g), one of the new L^s and L^u will lie in the interior of the first quadrant (in the coordinates near p_g'), and we are done.

FIGURE 3.3.

The only remaining case is when L^s and L^u form the boundary of the first quadrant; see Figure 3.4. But this is only possible if p_g and p_g' form a heteroclinic cycle of length 2.

Moreover each of the pairs $V_g^s(p_g)$, $V_g^u(p'_g)$ and $V_g^u(p_g)$, $V_g^s(p'_g)$ will either coincide or cross topologically. It follows from Theorem 2.2 that g has positive topological entropy or a double heteroclinic connection between p_g and p'_g. □

FIGURE 3.4.

4. Topological crossing of higher dimensional stable and unstable manifolds

Suppose that Y and Z are compact oriented manifolds with boundary, smoothly embedded in an oriented manifold X. Suppose further that $\partial Y \cap Z = \emptyset = \partial Z \cap Y$ and $\dim Y + \dim Z = \dim X$. We wish to define the (oriented) **intersection number of Y with Z**, $\#(Y, Z)$.

Let us recall the definition in the case when Y and Z are without boundary (see e.g. [Hi]). The orientations of X, Y and Z allow one to assign an intersection number of 1 or -1 to each transversal intersection of Y and Z. If Y and Z are transverse, there are finitely many intersection points and $\#(Y, Z)$ is the sum of the intersection numbers at all intersection points. It is easily shown that if Y' is homotopic to Y, Z' is homotopic to Z, Y is transverse to Z, and Y' is transverse to Z', then $\#(Y', Z') = \#(Y, Z)$. One defines the intersection number in the general case by first performing a homotopy to make the submanifolds transversal, and then using the definition in the transversal case.

This procedure can be carried over to our situation, with one caveat. The invariance of intersection number under homotopy breaks down if the boundary of one of the submanifolds is allowed to intersect the other submanifold. One can, however, make the submanifolds transverse with a homotopy that moves points less than any prescribed distance. Thus we simply add to the above definition the proviso that the homotopy used to make Y and Z transverse should move points by less than $\varepsilon/2$, where ε is chosen so that

$$(*) \qquad 0 < \varepsilon < \min(\operatorname{dist}(\partial Y, Z), \operatorname{dist}(\partial Z, Y)).$$

Now suppose that W and W' are smoothly immersed oriented submanifolds of X with complementary dimension. It is natural to say that W and W' cross topologically if there are compact embedded submanifolds with boundary, $V \subset W$ and $V' \subset W'$, such that $\dim V = \dim W$ and $\dim V' = \dim W'$, $\partial V \cap V' = \emptyset = V \cap \partial V'$ and $\#(V, V') \neq 0$. Note that it would be no loss of generality to require V and V' to be connected, for if $\#(V, V') \neq 0$ we can choose components $V_0 \subset V$ and $V_0' \subset V'$ with $\#(V_0, V_0') \neq 0$.

It is possible to generalize these ideas so that they can be applied to nonorientable manifolds. Suppose that Y and Z are compact orientable embedded submanifolds with boundary such that $\dim Y + \dim Z = \dim X$; we no longer require X to be oriented or even orientable. Let U be an orientable open set containing $Y \cap Z$ and $\mathcal{O}$ a choice of orientations for Y, Z and U. Then we can define $\#_{\mathcal{O}}(Y, Z; U)$ by following the above procedure (with the orientation of U replacing that of X) and requiring homotopies of Y and Z to move points by less than $\varepsilon/2$ where, in addition to $(*)$, ε also satisfies

$$\varepsilon < \min(\operatorname{dist}(Y \setminus U, Z), \operatorname{dist}(Z \setminus U, Y)).$$

This condition ensures that no intersection points appear outside U. We point out that in general both the sign and the absolute value of $\#_{\mathcal{O}}(Y, Z; U)$ can change with $\mathcal{O}$.

Definition 4.1. Let W and W' be smoothly immersed submanifolds with complementary dimension in a manifold M. Then W and W' **cross topologically** if there are an orientable open subset U of M and compact orientable embedded submanifolds with boundary, $V \subset W$ and $V' \subset W$ such that

1. $\dim V = \dim W$ and $\dim V' = \dim W'$,
2. $\partial V \cap V' = \emptyset = V \cap \partial V'$,
3. $V \cap V' \subset U$,

and there is a choice $\mathcal{O}$ of orientations on V, V' and U such that $\#_{\mathcal{O}}(V, V'; U) \neq 0$.

It is convenient to give another version of this definition. Observe that if Y, Z and U are connected, then only the sign of $\#_{\mathcal{O}}(Y, Z; U)$ can change with $\mathcal{O}$. Moreover if U' is another connected orientable open set containing $Y \cap Z$, and $\mathcal{O}'$ is the choice of orientations for Y, Z and U' that is compatible with $\mathcal{O}$, then

$$\#_{\mathcal{O}}(Y, Z; U) = \#_{\mathcal{O}'}(Y, Z; U).$$

Thus if Y and Z are connected compact orientable embedded submanifolds with boundary that have complementary dimension, we can define

$$\#_{+}(Y, Z) = |\#_{\mathcal{O}}(Y, Z; U)|,$$

where U is any connected orientable open set containing $Y \cap Z$ and $\mathcal{O}$ is a choice of orientations for Y, Z and U. We hasten to point out that there *are* such open sets, for example the open ε-neighbourhood of Y for small enough ε.

Observe that V,V' and U can be chosen as in Definition 4.1 if and only if there are components V_0 of V and V_0' of V' such that $\#_{+}(V_0, V_0') \neq 0$. Thus Definition 4.1 is equivalent to:

Definition 4.2. Let W and W' be smoothly immersed submanifolds with complementary dimension in a manifold M. Then W and W' **cross topologically** if there are compact connected orientable embedded submanifolds with boundary, $V \subset W$ and $V' \subset W$, such that $\#_+(V, V') \neq 0$.

We call V and V' a **good pair** for W and W' if they satisfy the above properties.

We are now ready to state the main theorem.

Theorem 4.3. *Let $f : M \to M$ be a diffeomorphism with a hyperbolic fixed point p. Assume that $\widehat{W}^s(p)$ and $\widehat{W}^u(p)$ have a topological crossing. Then some power of f has the full two shift as a topological factor.*

Remarks. We emphasize that the topological crossing need not be transverse and could be of infinite order. It follows from Lemma 1.3 that $h_{TOP}(f) > 0$. As in the two dimensional case, we only require contraction of $W^s(p)$ by f and of $W^u(p)$ by f^{-1}, not exponential contraction. Gedeon, McCord and Mischaikow have recently proved a generalization of this result [GMM].

Before proving Theorem 4.3, we develop the topological machinery that will be needed. The ideas are analogous to those in Section 2, but we have to use homology in place of elementary two dimensional topology. Now N will be a subset of M that is homeomorphic to (and identified with) $D^k(1) \times D^l(1)$ where $k + l = \dim M$ and $R = D^k(1) \times D^l(\rho)$ for some $\rho \in (0, 1)$.

We call a set $V \subset N$ **essential** if $inc_* : H_l(V, V \setminus R) \to H_l(N, N \setminus R)$ is nonzero.

If $S \subset N$, the **interior of S as a subset of** N is

$$\mathrm{Int}_N S = \{x \in S : B_\varepsilon(x) \cap N \subset S \text{ for some } \varepsilon > 0\}.$$

We shall call a set $S \subset N$ a **horizontal strip** if S is connected, $S \subset R$, and S can be decomposed into two disjoint sets S^0 and S^∂ such that:

(1) $\overline{S^0} \subset \mathrm{Int}_N S$;
(2) $i = inc_* : H_l(N, N \setminus R) \to H_l(N, N \setminus S^0)$ is nonzero;

Remark. There are only two types of horizontal strip that arise in the results of this paper. The first are those described in Section 2, for which we can choose S^∂ to be a small neighbourhood in S of $c_{lower} \cup c_{upper}$. The second have the form $D^k(1) \times Y$ where $Y \subset D^l(\rho)$; in this case we choose S^∂ to be $D^k(1) \times Y^\partial$, where Y^∂ is a small neighbourhood of ∂Y (see Lemma 4.7 below).

¿From (1) above and the excision theorem, we obtain:

Lemma 4.4. *$i_X = inc_* : H_l(X \cap S, X \cap S^\partial) \to H_l(X, X \setminus S^0)$ is an isomorphism for every set $X \subset N$, in particular when $X = N$.*

Since $H_l(N, N \setminus R)$ is one dimensional, we see from property (2) of a horizontal strip that there is a unique one dimensional subspace $L(S) \subset H_l(S, S^\partial)$ such that

$$i(H_l(N, N \setminus R)) \subset i_N(L(S)).$$

Let n be a positive integer and $S = S^0 \cup S^\partial$ a horizontal strip. We shall say that f^n **stretches S across R** if $f^n S \subset N$, $f^n S^\partial \subset N \setminus R$ and

$$f^n_* : L(S) \to H_l(N, N \setminus R) \text{ is nonzero.}$$

Theorem 4.5. *Suppose N contains two disjoint closed horizontal strips S_0 and S_1 that are stretched across R by f^{n_0} for some $n_0 \geq 1$. Then f^{n_0} has the full two shift as a topological factor.*

Proof. Except for the proof of the Key Lemma, the argument is identical to that in the two dimensional case. In the present context, the Key Lemma says:

Let $S = S^\partial \cup S^0$ be a horizontal strip and V an essential subset of N. Suppose that f^n stretches S across R. Then $f^n(S \cap V)$ is essential.

The proof of this claim is very similar to that of Lemma 3.4 in [E]. Consider the commuting diagram

$$\begin{tikzcd}
H_l(V \cap S, V \cap S^\partial) \arrow[r, "j"] \arrow[d, "f_*^n"'] & H_l(S, S^\partial) \arrow[d, "f_*^n"'] \\
H_l(f^n(V \cap S), f^n(V \cap S) \setminus R) \arrow[r, "k"] & H_l(N, N \setminus R)
\end{tikzcd}$$

in which j and k are induced by inclusions. We want to show that k is a nonzero map. This will be true if $f_*^n \circ j$ is a nonzero map. Since f^n stretches S across R, it suffices to show that j is a nonzero map whose image lies in $L(S)$. To that end, consider the commuting diagram

$$\begin{tikzcd}
H_l(V, V \setminus R) \arrow[r, "i_1"] \arrow[d, "i_2"'] & H_l(N, N \setminus R) \arrow[r, "i"] & H_l(N, N \setminus S^0) \\
H_l(V, V \setminus S^0) \arrow[r, "i_V^{-1}"] & H_l(V \cap S, V \cap S^\partial) \arrow[r, "j"] & H_l(S, S^\partial) \arrow[u, "i_N"']
\end{tikzcd}$$

in which all maps are induced by inclusions except for the map i_V^{-i}, which is defined by Lemma 4.4. We see immediately that the image of j lies in $L(S)$. Moreover i_1 is nonzero because V is essential, and i is nonzero because S is a horizontal strip. Since $H_l(N, N \setminus R)$ is one dimensional, it follows that $i \circ i_1$ is nonzero. Hence j is nonzero. $\square$

Proof of Theorem 4.3. As in the corresponding theorem for surfaces, we may assume that p is a fixed point. We use the notation and geometric structures introduced at the end of Section 1.

Lemma 4.6. *$W^u(p) \cap \operatorname{Int} N$ contains two compact embedded l-dimensional manifolds with boundary, V_0 and V_1, such that*

(1) $V_0 \cap V_1 = \emptyset$;
(2) $\partial V_0 \cup \partial V_1 \subset N \setminus (D^k(1) \times \{0\})$;
(3) $inc_* : H_l(V_i, \partial V_i) \to H_l(N, N \setminus (D^k(1) \times \{0\}))$ *is nonzero for* $i = 0, 1$.

Proof. Let $V_0 = \{0\} \times D^l(1/2)$. Now we find V_1. Let V^s and V^u be a good pair for $\widehat{W}^s(p)$ and $\widehat{W}^u(p)$ according to Definition 4.2. Then $f^n V^s$ and $f^n V^u$ are a good pair for any n. Choose n large enough so that $f^n V^s \subset \operatorname{Int} N$. Choose an open set X such that $f^n V^s \subset X$,

$\overline{X} \subset \operatorname{Int} N$ and ∂X is a smooth submanifold transverse to V^u. Then $f^n V^u \cap \overline{X}$ is a smooth manifold with boundary. There is a component V_1 of $f^n V^u \cap \overline{X}$ such that $f^n V^s$ and V_1 are a good pair for $\widehat{W}^s(p)$ and $\widehat{W}^u(p)$. It is easily seen that V_0 and V_1 have the desired properties. $\square$

Choose $\rho > 0$ such that $\partial V_0 \cup \partial V_1 \subset N \setminus (D^k(1) \times D^l(\rho))$, and set $R = D^k(1) \times D^l(\rho)$. Our choice of ρ makes the sets V_0 and V_1 that were defined in the previous lemma essential.

Lemma 4.7. *For any large enough n there are disjoint closed horizontal strips $S_0(n) \supset f^{-n}V_0$ and $S_1(n) \supset f^{-n}V_1$ that are stretched across R by f^n.*

Proof. Choose $\varepsilon > 0$ so that the closed ε-neighbourhoods of ∂V_0 and ∂V_1 lie in $\operatorname{Int} N \setminus R$. Let n be large enough so that the conclusions of Lemma 1.4 hold with $V = V_0$ or V_1. For $i = 0, 1$, let

$$Y_i(n) = \{y \in D^l(\rho) : (0, y) \in f^{-n}(V_i)\} \quad \text{and} \quad S_i(n) = D^k(1) \times Y_i(n).$$

Our choice of ε and (2) of Lemma 1.4 ensure that we can split $Y_i(n)$ into complementary subsets $Y_i^\partial(n)$ and $Y_i^0(n)$ such that $\overline{Y_i^0} \subset \operatorname{Int} Y_i(n)$ and f^n maps $D^k(1) \times Y_i^\partial(n)$ into $N \setminus R$. Set $S_i^\partial(n) = D^k(1) \times Y_i^\partial(n)$ and $S_i^0(n) = D^k(1) \times Y_i^0(n)$. It is clear from our construction that $S_0(n)$ and $S_1(n)$ are horizontal strips and are closed and disjoint. Moreover $(f^{-n}V_i, f^{-n}\partial V_i) \subset (S_i(n), S_i^\partial(n))$ for $i = 0, 1$. We have the following commuting diagram.

$$\begin{array}{ccc}
H_l(f^{-n}V_i, f^{-n}\partial V_i) & \xrightarrow{f_*^n} & H_l(V_i, \partial V_i) \\
{\scriptstyle inc_*}\big\downarrow & & {\scriptstyle inc_*}\big\downarrow \\
H_l(S_i(n), S_i^\partial(n)) & \xrightarrow{f_*^n} & H_l(N, N \setminus R)
\end{array}$$

The top arrow is an isomorphism and the right arrow is a nonzero map by Lemma 4.6 and the subsequent definition of R. Hence the bottom arrow must also be a nonzero map. Since $H_l(S_i(n), S_i^\partial(n))$ is one dimensional, we have $L(S_i(n)) = H_l(S_i(n), S_i^\partial(n))$. It follows that f^n stretches $S_i(n)$ across R. $\square$

Theorem 4.3 follows immediately from the above lemma and Theorem 4.5. $\square$

As we pointed out in the introduction, Theorem 4.3 has an extension to a heteroclinic cycle in which all the appropriate deleted stable and unstable manifolds have topological crossings.

Theorem 4.8. *Let $f : M \to M$ be a diffeomorphism with a heteroclinic cycle $p_0, \ldots, p_{r-1}$ in which each pair $\widehat{W}^s(p_i)$ and $\widehat{W}^u(p_{i+1})$ has a topological crossing. Then $h_{TOP}(f) > 0$ and some power of f has the subshift of finite type Σ_A as a topological factor, where A is the $2r \times 2r$ matrix*

$$\begin{pmatrix}
1 & 1 & 0 & 0 & & & & \\
0 & 0 & 1 & 1 & & & & \\
 & & 1 & 1 & 0 & 0 & & \\
 & & 0 & 0 & 1 & 1 & & \\
 & & & & \ddots & \ddots & & \\
 & & & & & & & \\
0 & 0 & & & & & 1 & 1 \\
1 & 1 & & & & & 0 & 0
\end{pmatrix}.$$

Proof. Let k be the common dimension of the $\widehat{W}^s(p_i)$ and l the common dimension of the $\widehat{W}^u(p_i)$. Choose neighbourhood N_i of the p_i that are pairwise disjoint and have the same properties as N in the previous arguments. Given $\rho \in (0,1)$, let R_i be the subset of N_i that is identified with $D^k(1) \times D^l(\rho)$. We make the obvious definitions of an essential subset of N_i, a horizontal strip in R_i, and what it means for f^n to map a horizontal strip in R_i across R_j. Each $W^u(p_i)$ contains compact submanifolds with boundary V_0^i and V_1^i such that, for a suitable choice of ρ, V_0^i is an essential subset of N_i and V_1^i is an essential subset of N_{i+1}. For a large enough n_0 there are horizontal strips $S_0^i \supset f^{-n_0}V_0^i$ and $S_1^i \supset f^{-n_0}V_1^i$ in R_i such that f^{n_0} stretches S_0^i across R_i and S_1^i across R_{i+1}. Now let

$$\Lambda = \{x : f^{kn_0}x \in \bigcup_{i=0}^{k-1} \bigcup_{j=0}^{1} S_j^i \text{ for all } k \in \mathbb{Z}\},$$

and define

$$\pi : \Lambda \to \{(0,0),(0,1),\ldots,(k-1,0),(k-1,1)\}^{\mathbb{Z}}$$

so that the m'th term of the sequence $\pi(x)$ is (i,j) if $f^{mn_0}x \in S_j^i$. The image of the map π is Σ_A. $\square$

References

[A] V. Arnold, *Instability of dynamical systems with several degrees of freedom*, Soviet Math. Dokl. **5** (1964), 581–585.

[Bo] R. Bowen, *On axiom A diffeomorphisms*, CBMS Lecture Series #35, American Math. Society, 1977.

[Ch] R. C. Churchill, *Invariant sets which carry cohomology*, J. Diff. Equations **13** (1973), 39–65.

[Co] C. Conley, *Topological Dynamics*, Joseph Auslander, ed., Benjamin, New York, 1968, pp. 129–153.

[CR1] R. C. Churchill and D. L. Rod, *Pathology in dynamical systems I: general theory*, J. Diff. Equations **21** (1976), 39–65.

[CR2] R. C. Churchill and D. L. Rod, *Pathology in dynamical systems. III — analytic Hamiltonians*, J. Diff. Equations **37** (1980), 23–38.

[E] R. W. Easton, *Isolating blocks and symbolic dynamics*, J. Diff. Equations **17** (1975), 23–38.

[Ga] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York.

[GMM] T. Gedeon, C. McCord and K. Mischaikow, Electronic communication.

[GS] N. Gavrilov and L. Silnikov, *On 3-dimensional dynamical systems close to systems with a structurally unstable homoclinic curve, II*, Math. USSR Sb. **90** (1973), 139–156.

[Ha] P. Hartman, *On local homeomorphisms of Euclidean spaces*, Bol. Soc. Mat. Mexicana **5** (1960), 220–241.

[Hi] M. W. Hirsch, *Differential Topology*, Graduate Texts in Mathematics 33, Springer, New York-Heidelberg-Berlin, 1976.

[HPS] M. Hirsch, C. Pugh and M. Shub, *Invariant manifolds*, Springer Lecture Notes in Mathematics, vol. 583, Springer Verlag, New York-Heidelberg-Berlin, 1977.

[K1] A. Katok, *Lyapunov exponents, entropy and periodic orbits for diffeomorphisms*, Publ. Math. Hautes Études Sci **51** (1980), 137–173.

[K2] A. Katok, *Nonuniform hyperbolicity and structure of smooth dynamical systems*, Proc. International Congress of Mathematicians Warszawa 1983 **2**, 1245–1254.

[Ko] V. V. Kozlov, *Integrability and non-integrability in Hamiltonian mechanics*, Russian Math. Surveys **38** (1983), 1–76.

[KW] G. Knieper and H. Weiss, *A surface with positive curvature and positive topological entropy*, to appear, J. Differential Geometry.

[M] V. K. Melnikov, *On the stability of the center for time periodic perturbations.*, Trans. Moscow Math. Society **12** (1963), 1–57.

[N] S. Newhouse, *Dynamical Systems*, C.I.M.E. Lectures, Birkhäuser, Boston-Basel-Stuttgart, 1978, pp. 1–114.

[Pa] G. Paternain, *Real analytic convex surfaces with positive topological entropy and rigid body dynamics*, preprint (1991).

[Po] H. Poincaré, *Les méthodes nouvelles de la mécanique céleste*, Dover Publications (reprint), New York, 1957.

[S] S. Smale, *Diffeomorphisms with many periodic points*, Differential and Combinatorical Topology (edited by S.S. Cairnes) (1965), Princeton University Press, 63-80.

[So] G. Sotomayor, *Structural stability of first order and Banach varieties*, Univ. Nac. Ingen. Inst. Mat. Puras Apl. Notas Mat. **4** (1966), 11–52.

[Sp] E. H. Spanier, *Algebraic Toploogy*, McGraw-Hill, 1966.

[St] Sternberg.

[T] Ya. V. Tatarinov, *A global view of the dynamics of a rigid body: Passage to the quotient system*, Moscow Univ. Mech. Bull. **33** (1978), 17–21.

[W] P. Walters, *An introduction to Ergodic Theory*, Graduate Texts in Mathematics 79, Springer Verlag, Berlin - New York, 1982.

[Z] S. L. Ziglin, *Splitting of separatrices, branching of solutions and non-existence of an integral in the dynamics of the rigid body*, Trans. Moscow Math. Soc. **41** (1982), 283-298.

Keith Burns
Department of Mathematics
Northwestern University
Evanston, IL 60208
U.S.A.
Email: burns@math.nwu.edu

Howard Weiss
Department of Mathematics
Pennsylvania State University
University Park, PA 16802
U.S.A.
Email: weiss@math.psu.edu