

# ROTATION VECTORS AND FIXED POINTS OF AREA PRESERVING SURFACE DIFFEOMORPHISMS

JOHN FRANKS

**ABSTRACT.** We consider the (homological) rotation vectors for area preserving diffeomorphisms of compact surfaces which are homotopic to the identity. There are two main results. The first is that if 0 is in the interior of the convex hull of the rotation vectors for such a diffeomorphism then  $f$  has a fixed point of positive index. The second result asserts that if  $f$  has a vanishing mean rotation vector then  $f$  has a fixed point of positive index. There are several applications including a new proof of the Arnold conjecture for area preserving diffeomorphisms of compact surfaces.

In this paper we investigate area preserving diffeomorphisms of compact surfaces. We are concerned with “rotation vectors” defined in terms of homology which can be associated to the points of the surface. There are two main results. The first is that if  $f$  is an area preserving diffeomorphism of a compact surface which is homotopic to the identity and 0 is in the interior of the convex hull of the rotation vectors of  $f$  then  $f$  has a fixed point of positive index. The second result concerns the mean rotation vector of  $f$ . It asserts that if this vector vanishes then  $f$  has a fixed point of positive index.

A special case of this second result – when the surface in question is a surface of genus zero – was proved in [F5].

In §6 we give several applications including a new proof of the Arnold conjecture for area preserving diffeomorphisms of compact surfaces.

## 1. CHAIN RECURRENCE

We briefly recall the definition of chain recurrence due to Charles Conley in [C]. In the following  $f : X \rightarrow X$  will denote a homeomorphism of a metric space  $X$ .

**(1.1) Definition.** An  $\varepsilon$ -chain for  $f$ , from  $x$  to  $y$ , is a sequence  $x = x_1, x_2, \dots, x_n = y$  in  $X$  such that

$$d(f(x_i), x_{i+1}) < \varepsilon \quad \text{for } 1 \leq i \leq n-1.$$

A point  $x \in X$  is called *chain recurrent* if for every  $\varepsilon > 0$  there is an  $\varepsilon$ -chain from  $x$  to itself. The set  $\mathbf{R}(f)$  of chain recurrent points is called the *chain recurrent set* of  $f$ .

It is easily seen that if the metric space  $X$  is compact then the chain recurrent set  $\mathbf{R}(f)$  is compact and invariant under  $f$ . Moreover it is independent of the choice

---

Received by the editors September 20, 1994 and, in revised form, March 31, 1995.

1991 *Mathematics Subject Classification.* Primary 58C30; Secondary 58F11.

of metric on  $X$ , depending only on  $f$  and the topology of  $X$ . If  $X$  is not compact then  $\mathbf{R}(f)$  is closed and invariant but it depends on the metric defined on  $X$  rather than just the topology.

The following simple result is well known and is valid whether or not  $X$  is compact.

**(1.2) Proposition.** *If  $\mathbf{R}(f) = X$  (in particular if  $f$  preserves a finite measure whose support is all of  $X$ ) and if  $X$  is connected, then for any  $\varepsilon > 0$  and any  $x, y \in X$  there is an  $\varepsilon$ -chain from  $x$  to  $y$ .*

*Proof.* Note that if  $f$  preserves a finite measure whose support is all of  $X$  then by Poincaré recurrence almost every point of  $X$  is recurrent, from which it follows that  $\mathbf{R}(f) = X$ .

We define a relation  $\sim$  on  $\mathbf{R}(f)$  by  $x \sim y$  if and only if for every  $\varepsilon > 0$  there is an  $\varepsilon$ -chain from  $x$  to  $y$  and another from  $y$  to  $x$ . It is clear that  $\sim$  is an equivalence relation. From the definition it is easy to see that each equivalence class for the equivalence relation  $\sim$  is open. Since equivalence classes are pairwise disjoint it follows that the complement of an equivalence class is open. Since  $X$  is connected there can be only one equivalence class. ■

By a *surface of finite type* we mean a smooth two dimensional manifold obtained by taking a smooth compact surface (possibly with boundary) and deleting finitely many points from its interior. Equivalently it is a surface obtained by taking a smooth compact surface with boundary and removing some, or all, of the boundary components

Suppose  $f : M \rightarrow M$  is homotopic to the identity and  $M$  has negative Euler characteristic. If  $\pi : \tilde{M} \rightarrow M$  is the universal covering space of  $M$  there is a canonical lift of  $f$  to a map  $F : \tilde{M} \rightarrow \tilde{M}$ ; namely,  $F$  is that lift obtained by lifting the homotopy  $f_t$  from the identity to  $f$  to form a homotopy of  $\tilde{M}$  starting at the identity on  $\tilde{M}$ .

**(1.4) Lemma.** *Suppose  $N$  is a surface of finite type and  $f_0 : N \rightarrow N$  is a homeomorphism homotopic to the identity whose canonical lift to its universal covering space is  $F : \tilde{N} \rightarrow \tilde{N}$ . If  $F$  is fixed point free and  $f_0$  has no interior fixed points, then there is a complete Riemannian metric on  $N$  which when lifted to  $\tilde{N}$  has a distance function  $d(\cdot, \cdot)$  satisfying  $d(F(x), x) > 1$  for all  $x \in \tilde{N}$ .*

*Proof.* The proof is slightly technical, but it is based on a simple idea. On any compact set it is easy to alter the metric to satisfy the conclusion – in fact one can realize the desired metric by embedding in a Euclidean space. The general case uses the same idea.

Let  $N_1$  be the compact surface obtained from  $N$  by replacing the deleted punctures. We assume that  $N_1$  has been embedded in  $\mathbb{R}^4$ . We next embed  $N$  as a closed set of  $\mathbb{R}^5$ . Let  $g : N \rightarrow (0, \infty)$  be a smooth function which satisfies  $g(x) < \text{dist}(x, N_1 - N)$  for all  $x \in N$  and define  $i_0 : N \rightarrow \mathbb{R}^5$  by  $i_0(x) = (x, 1/g(x))$ . It is easy to see that  $i_0$  is an embedding of  $N$  onto a closed submanifold of  $\mathbb{R}^5$ . Let  $f : i_0(N) \rightarrow i_0(N)$  be defined by setting  $f(i_0(x)) = i_0(f_0(x))$ .

Let  $U$  be a compact tubular neighborhood of the boundary of  $N$  which contains no punctures and let  $N_0$  denote the surface of finite type  $N \setminus \text{int}(U)$ . There are no fixed points of  $f_0$  in  $N_0$ . In  $\mathbb{R}^5$ , for  $r > 1$ , let  $N_r$  denote the compact set

$\{x \in i_0(N_0) \mid \|x\| \leq r + 1\}$ . Choose a smooth function  $\phi : [0, \infty) \rightarrow \mathbb{R}$  such that  $\phi(t) > 1$  for all  $t$ ,  $\phi'(t) \geq 0$  for  $t \geq 0$ , and so large that

$$\phi(r) > \max_{x \in N_r} \frac{1}{\|x - f(x)\|}.$$

(It is not difficult to find a smooth function with positive derivative which is greater than any given continuous monotonic increasing function defined on  $[0, \infty)$ .) It follows that

$$\frac{1}{\phi(r)} < \min_{x \in N_r} \|x - f(x)\|.$$

Define a diffeomorphism  $h : \mathbb{R}^5 \rightarrow \mathbb{R}^5$  by  $h(r, v) = (r\phi(r), v)$ , where  $r$  and  $v$  are spherical coordinates, i.e.  $v \in \mathbb{R}^5$  is a unit vector,  $r > 0$  and  $x = rv$ . Let  $\{v_1, v_2, \dots, v_5\}$  be an orthonormal basis of the tangent space to the unit sphere in  $\mathbb{R}^5$  at the point  $v$ . The matrix for  $Dh_{(r,v)}$  with respect to the orthonormal bases  $\{\frac{\partial}{\partial r}, v_1, v_2, \dots, v_5\}$  at  $(r, v)$  and  $\{\frac{\partial}{\partial r}, v_1, v_2, \dots, v_5\}$  at  $h(r, v)$  has the form

$$\begin{pmatrix} \phi(r) + r\phi'(r) & 0 \\ 0 & \phi(r)I \end{pmatrix}.$$

It follows that

$$(1) \quad \|Dh_{h(r,v)}^{-1}\| \leq \frac{1}{\phi(r)}.$$

We will show that  $\|h(f(x)) - h(x)\| \geq 1$  for all  $x \in i_0(N_0)$ . If this is not the case there exists  $x_0 \in i_0(N_0)$  with  $\|h(f(x_0)) - h(x_0)\| < 1$ . By the mean value theorem applied to the diffeomorphism  $h^{-1}$

$$\|f(x_0) - x_0\| \leq \|Dh_p^{-1}\| \|h(f(x_0)) - h(x_0)\| < \|Dh_p^{-1}\|,$$

where  $p$  is a point of the straight line segment joining  $h(x_0)$  and  $h(f(x_0))$ . Hence by (1) above

$$(2) \quad \|f(x_0) - x_0\| < \frac{1}{\phi(r)} < \min_{x \in N_r} \|x - f(x)\|,$$

where  $r = \|h^{-1}(p)\|$ . But  $h^{-1}$  is a contraction, so  $\|x_0 - h^{-1}(p)\| \leq \|h(x_0) - p\| < 1$ . Thus  $x_0 \in N_r$  when  $r = \|h^{-1}(p)\|$ . This contradicts (2).

It follows that  $\|h(f(x)) - h(x)\| \geq 1$  for all  $x \in i_0(N_0)$ . Letting  $i = h \circ i_0$  gives an embedding such that  $\|i(f(x)) - i(x)\| \geq 1$  for all  $x \in N_0$ . We consider the Riemannian metric on  $N$  obtained by pulling back to  $N_0$  the restriction of the Euclidean metric to  $i(N)$ . In this metric  $d(x, f_0(x)) > 1$  for any  $x \in N_0$ , so if this metric is lifted to  $\widetilde{M}$ , we have  $d(y, F(y)) > 1$  for any  $y \in \pi^{-1}(N_0)$ . But also since  $U$  is compact and  $F$  is fixed point free, there is a constant  $\delta > 0$  such that  $d(y, F(y)) > \delta$  for all  $y \in \pi^{-1}(U)$ . Scaling the metric by  $1/\delta$  we have  $d(y, F(y)) > 1$  for all  $y \in \widetilde{M}$ . ■

## 2. HOMOLOGICAL ROTATION VECTORS

In this section we define homological rotation vectors for surface diffeomorphisms isotopic to the identity map on compact surfaces and investigate their properties. Homological rotation vectors for surfaces of genus zero were considered in [F5]. We also consider the “mean rotation vector” for an invariant measure  $\mu$  which can be expressed as the integral of the rotation vectors of points with respect to  $\mu$  (see (2.3) below).

We are interested in investigating the existence of periodic orbits for area preserving diffeomorphisms on a surface. We will consider a compact surface  $M$  and focus on the case where the diffeomorphism  $f : M \rightarrow M$  is isotopic to the identity, and  $M$  has negative Euler characteristic. We begin by formulating the definition of “homological rotation vector” for such a diffeomorphism  $f : M \rightarrow M$ .

We fix a metric on  $M$  of constant negative curvature. We assume that each boundary component is a geodesic. Even more, we assume that one can form  $M$  by taking a convex geodesic polygon in hyperbolic space (some vertices of which may be points at infinity) and making identifications of some of the edges.

Pick a base point  $b_0$  in the interior of the polygon whose sides are identified to form  $M$ . We want to define a function  $\gamma$  which assigns to each  $x \in M$  a geodesic segment  $\gamma_x$  in  $M$  from  $b_0$  to  $x$ , in such a way that the correspondence  $x \rightarrow \gamma_x$  is measurable. We do this using the fact that  $M$  is a convex geodesic polygon in hyperbolic space with some edges identified, and with  $b_0$  in its interior. We then let  $\gamma_x$  to be the unique geodesic segment from  $b_0$  to  $x$  if  $x$  is in the interior of this polygon. For each pair of edges which are identified we pick one and choose  $\gamma_x$  to be the unique geodesic segment from  $b_0$  to  $x$  which back in the polygon ends on the chosen edge.

Let  $f_t(x)$  be a homotopy from  $f_0 = id : M \rightarrow M$  to  $f_1 = f$ . Because the Euler characteristic of  $M$  is negative  $f_t$  is unique up to homotopy. This means that if  $g_t$  is another homotopy with  $g_0 = id$  and  $g_1 = f$  then there is a homotopy from  $f_t$  to  $g_t$ , i.e. a map  $H : M \times [0, 1] \times [0, 1] \rightarrow M$  such that  $H(x, t, 0) = f_t(x)$  and  $H(x, t, 1) = g_t(x)$ .

For any point  $x \in M$  we want to construct a path in  $M$  from  $x$  to  $f^n(x)$  and then form a loop with the segments  $\gamma_x$  and  $\gamma_{f^n(x)}$ . To do this we observe that if  $\pi : \widetilde{M} \rightarrow M$  is the universal covering space of  $M$  there is a canonical lift of  $f$  to a diffeomorphism  $F : \widetilde{M} \rightarrow \widetilde{M}$ ; namely,  $F$  is that lift obtained by lifting the homotopy  $f_t$  from the identity to  $f$  to form a homotopy of  $\widetilde{M}$  starting at the identity on  $\widetilde{M}$ . The other end of this homotopy is then defined to be  $F$ . The uniqueness of  $f_t$  up to homotopy implies that  $F$  does not depend on the choice of homotopy from the identity to  $f$ .  $F$  is the unique lift whose extension to the ideal points at infinity of  $\widetilde{M}$  has all those points as fixed points.

Consider the path  $\alpha(n, x)$  from  $x$  to  $f^n(x)$  in  $M$  which is given by

$$\alpha(n, x)(t) = f_t^n(x).$$

Again the homotopy class of this path relative to its endpoints is independent of the choice of the homotopy  $f_t$  because of the uniqueness (up to homotopy) of this homotopy.

For each  $x \in M$  let  $h_n(x, f)$  be the closed loop based at  $b_0$  formed by the concatenation of  $\gamma_x$ , the path  $\alpha(n, x)$  in  $M$  from  $x$  to  $f^n(x)$  and  $\gamma_{f^n(x)}$  traversed

backwards. If the diffeomorphism  $f$  is clear from the context we will abbreviate  $h_n(x, f)$  to  $h_n(x)$ .

Note that if  $*$  denotes concatenation of based loops then  $h_n(x) * h_m(f^n(x))$  is homotopic to  $h_{n+m}(x)$ . We will denote by  $[h_n(x)]$  the homology class in  $H_1(M, \mathbb{R})$  of the loop  $h_n(x)$ . Note that  $[h_{n+m}(x)] = [h_n(x)] + [h_m(f^n(x))]$ . We can now formulate the definition of homology rotation vector.

**(2.1) Definition.** Let  $M$  be a compact surface which has negative Euler characteristic and may have non-empty boundary. Suppose  $f : M \rightarrow M$  is a diffeomorphism which is isotopic to the identity map. The *homological rotation vector* of  $x \in M$  is an element of  $H_1(M, \mathbb{R})$  denoted  $\mathcal{R}(x, f)$ , and is defined by

$$\mathcal{R}(x, f) = \lim_{n \rightarrow \infty} \frac{[h_n(x)]}{n}$$

if this limit exists.

If the limit in Definition (2.1) above does not exist then  $\mathcal{R}(x, f)$  is undefined and we say  $x$  has no homological rotation vector.

Let  $\mu$  be a smooth  $f$  invariant measure on  $M$ . The homology classes  $[h_1(x)] \in H_1(M, \mathbb{R})$  depend measurably on  $x$ . In fact there is a closed set of measure zero in  $M$  (consisting of the “edges” of the polygon and their inverse images under  $f$ ) on the complement of which the function  $[h_1(x)]$  is locally constant.

**(2.2) Lemma.** *If  $f : M \rightarrow M$  is as in (2.1) and  $f$  preserves the smooth measure  $\mu$  then the function  $[h_1(x)]$  defined on  $M$  with values in  $H_1(M, \mathbb{R})$  is bounded and  $\mu$  measurable, hence integrable.*

*Proof.* We already remarked that  $[h_1(x)]$  is measurable since it is locally constant on the complement of a closed set of measure zero. We need to show that it is bounded. This is the main place our hypothesis of compactness of  $M$  is used.

Using the fact that  $M$  is a compact surface it is clear the length of the loop  $h_1(x)$  is uniformly bounded in  $x$ . Also only finitely many elements of  $H_1(M, \mathbb{Z})$  can be represented by loops of a given length. It follows that  $[h_1(x)] \in H_1(M, \mathbb{R})$  is a bounded function of  $x$ .

We observe that if  $M$  were not compact, the function  $h_1(x)$  might not be bounded. In fact it is easy to construct a diffeomorphism of the disk punctured at its center which has the property that as one approaches the central puncture the diffeomorphism rotates around that puncture an arbitrarily large amount. One can construct such a diffeomorphism for which  $[h_1(x)]$  is not integrable. Of course this diffeomorphism is not smoothly extendable to the whole disk.

In our setting, however, the function  $[h_1(x)]$  is integrable. Since  $[h_{n+m}(x)] = [h_n(x)] + [h_m(f^n(x))]$ ,

$$\sum_{i=0}^{n-1} [h_1(f^i(x))] = [h_n(x)].$$

Hence by the Birkhoff ergodic theorem, for  $\mu$ -almost all  $x$  the limit

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} [h_1(f^i(x))] = \lim_{n \rightarrow \infty} \frac{[h_n(x)]}{n} = \mathcal{R}(x, f)$$

exists.

**(2.3) Definition.** *The mean rotation vector of  $f$  is an element of  $H_1(M, \mathbb{R})$  denoted  $\mathcal{R}_\mu(f)$ , and is defined by*

$$\mathcal{R}_\mu(f) = \int [h_1(x)] d\mu.$$

Another consequence of the Birkhoff ergodic theorem is that

$$\mathcal{R}_\mu(f) = \int \mathcal{R}(x, f) d\mu.$$

A key property of the mean rotation vector is that it is a homomorphism from the the group of  $\mu$  invariant diffeomorphisms isotopic to the identity to  $H_1(M, \mathbb{R})$ . The proof of this fact is really just the change of variables formula for integration. We give it here for purposes of exposition.

**(2.4) Proposition.** *Suppose  $f$  and  $g$  are diffeomorphisms of the compact surface  $M$  which are isotopic to the identity and preserve a smooth measure  $\mu$ . Then*

$$\mathcal{R}_\mu(f \circ g) = \mathcal{R}_\mu(f) + \mathcal{R}_\mu(g).$$

*Proof.* The loop  $h_1(x, f \circ g)$  is homotopic to the concatenation of the loops  $h_1(x, f)$  and  $h_1(f(x), g)$ , so  $[h_1(x, f \circ g)] = [h_1(x, f)] + [h_1(f(x), g)]$ . Thus

$$\int [h_1(x, f \circ g)] d\mu = \int [h_1(x, f)] d\mu + \int [h_1(f(x), g)] d\mu.$$

Since  $f$  preserves  $\mu$ , we have that  $\int [h_1(f(x), g)] d\mu = \int [h_1(x, g)] d\mu$ . Hence

$$\mathcal{R}_\mu(f \circ g) = \mathcal{R}_\mu(f) + \mathcal{R}_\mu(g). \quad \blacksquare$$

The Poincaré recurrence theorem implies that the set of recurrent points in  $M$  has full measure. Thus the set of points for which  $\mathcal{R}(x, f)$  exists and which are recurrent under  $f$  has full measure. Moreover, any rotation vector is essentially a convex combination of rotation vectors of a recurrent points as the following lemma shows.

**(2.5) Proposition.** *Let  $x$  be a point for which the limit  $\mathcal{R}(x, f) \in H_1(M, \mathbb{R})$  exists and let  $\varepsilon > 0$  be arbitrary. Then there are points  $y_i$ ,  $0 \leq i \leq n$  in the set of  $f$  recurrent points with the property that  $\mathcal{R}(y_i, f)$  exists and there are rationals  $c_i$  such that*

$$\left\| \mathcal{R}(x, f) - \sum_{i=0}^n c_i \mathcal{R}(y_i, f) \right\| < \varepsilon.$$

*Proof.* Consider the sequence of measures  $\{\nu_n\}$  where  $\nu_n$  is supported on the finite set  $\{x, f(x), f^2(x), \dots, f^{n-1}(x)\}$  and assigns measure  $1/n$  to each of these points. Let  $\nu$  be a weak limit of these measures. Then it is invariant under  $f$ . Let  $\phi(x) = h_1(x)$  be the function defined in the beginning of §2.

One can show easily that

$$\lim_{n \rightarrow \infty} \int \phi d\nu_n = \mathcal{R}(x, f).$$

We will show that also

$$\mathcal{R}_\nu(f) = \lim_{n \rightarrow \infty} \int \phi d\nu_n,$$

so  $\mathcal{R}_\nu(f) = \mathcal{R}(x, f)$ . Note that if the function  $\phi(x) = h_1(x)$  defined in the beginning of §2 were continuous then it would be immediate that

$$\mathcal{R}_\nu(f) = \int_M \phi d\nu = \lim_{n \rightarrow \infty} \int_M \phi d\nu_n.$$

By slightly changing the polygon used to define the arcs  $\gamma_x$  in the beginning of §2, we can arrange that the discontinuities of the function  $h_1(x)$  (which are the points on the boundary of this polygon) have measure zero with respect to  $\nu$ . We can choose a decreasing sequence of compact neighborhoods  $V_m$  of this set which satisfy  $\nu_n(V_m) < 1/m$ . Moreover, we can construct continuous functions  $\phi_m : M \rightarrow H_1(M, \mathbb{R})$  which agree with  $\phi$  outside of  $V_m$  and are uniformly bounded (say by  $K$  which we choose also to be a bound for  $\phi$ ). It is not difficult to show that for fixed  $m$  if  $n$  is sufficiently large then  $\nu_n(V_m) < 1/m$ .

Thus for fixed  $m$  and for an arbitrarily large  $n$  (recall that only a subsequence of  $\{\nu_n\}$  converges to  $\nu$ ) we now have

$$\begin{aligned} \left\| \int_M \phi d\nu - \int_M \phi d\nu_n \right\| &= \left\| \int_M (\phi - \phi_m) d\nu \right\| \\ &+ \left\| \int_M \phi_m d\nu - \int_M \phi_m d\nu_n \right\| \\ &+ \left\| \int_M \phi_m d\nu_n - \int_M \phi d\nu_n \right\| \end{aligned}$$

so that

$$\begin{aligned} \left\| \int_M \phi d\nu - \int_M \phi d\nu_n \right\| &< 2K\nu(V_m) + K/m + 2K\nu_n(V_m) \\ &< 5K/m. \end{aligned}$$

Clearly this implies that a subsequence of  $\int \phi d\nu_n$  converges to  $\int \phi d\nu$ , but since  $\lim_{n \rightarrow \infty} \int \phi d\nu_n$  exists, it follows that  $\int \phi d\nu = \lim_{n \rightarrow \infty} \int \phi d\nu_n$ .

If the measure  $\nu$  should happen to be ergodic then for  $\nu$  almost any point in its support (and hence for a recurrent point  $y$  by Poincaré recurrence) we have  $\mathcal{R}(y, f) = \mathcal{R}(x, f)$ . In the general case we can apply the ergodic decomposition theorem (see [Ma], for example) which says roughly that  $\nu$  is a convex combination of ergodic Borel probability measures. In particular, there is a finite set of ergodic probability measures  $\{\mu_i\}$  with the property that

$$\left\| \int_M \mathcal{R}(x, f) d\nu - \sum_{i=0}^n c_i \int_M \mathcal{R}(x, f) d\mu_i \right\| < \varepsilon.$$

But since each  $\mu_i$  is ergodic, there are points  $y_i$  in the support of  $\mu_i$  with  $\mathcal{R}(y_i, f) = \int_M \mathcal{R}(y_i, f) d\mu_i$ . ■

## 3. HANDEL'S FIXED POINT THEOREM

Suppose that  $h : D \rightarrow D$  is an orientation preserving homeomorphism of the disk  $D = \mathbb{D}^2$ . We are interested in the orbits of points whose alpha and omega limit sets each consists of single (but distinct) points in  $\partial D$ . Given a two such points  $x, y \in D$  we will say that their orbits,  $orb(x)$  and  $orb(y)$  are *linked* provided the points  $\alpha(y)$  and  $\omega(y)$  separate the points  $\alpha(x), \omega(x)$  in  $\partial D$ , or equivalently if the straight line segments from  $\alpha(x)$  to  $\omega(x)$  and from  $\alpha(y)$  to  $\omega(y)$  intersect in a single interior point of  $D$ .

We say that orbits  $orb(x_i)$ ,  $1 \leq i \leq n$ , form an *oriented cycle of links* if there is an oriented polygon in the interior of  $D$  whose  $i^{th}$  side is a segment of the line segment from  $\alpha(x_i)$  to  $\omega(x_i)$  and the orientation of this side is consistent with the orientation of this line segment from  $\alpha(x_i)$  to  $\omega(x_i)$ .

We can now state a very powerful fixed point theorem of Michael Handel (see [H1] and [H2]).

**(3.1) Theorem [H1].** *Suppose  $h$  is a homeomorphism of  $\mathbb{D}^2$ . If there are points  $x_i \in \mathbb{D}^2$  whose orbits form an oriented cycle of links, then  $h$  has a fixed point in the interior of  $\mathbb{D}^2$ .*

It seems likely that a slightly stronger conclusion to this theorem is valid, namely that there exists an interior fixed point of positive index. This stronger conclusion is important for some of the applications we wish to make. While this conclusion is not known to hold in general, it does hold in the special setting where we wish to apply it. This result is also due to Handel (personal communication) and we present his proof here.

There are two reasons for presenting the proof. First we need only a special case of the general theorem stated above and Handel has shown in a personal communication how much of the deep and difficult work for the general theorem can be avoided in this case. Secondly, presenting this exposition makes this paper relatively self-contained. The version of Handel's theorem we need can certainly be extracted from his paper [H1], but piecing together parts of his work to achieve the form we need would result in a far from optimal exposition. The author would like to emphasize, however, that this section is only an exposition of the work of Michael Handel.

We begin by describing the special setting for which we shall give the proof. Suppose  $f : M \rightarrow M$  is a diffeomorphism of a compact surface with negative Euler characteristic and  $f$  is isotopic to the identity. It is possible and convenient to provide  $M$  with a metric of constant negative curvature with each boundary component being a geodesic. Let  $\tilde{M}$  denote the universal covering space of  $M$  and let  $C(\tilde{M})$  denote the circle or Cantor set of "points at infinity" of  $\tilde{M}$  which can be added to  $\tilde{M}$  to compactify it. We think of  $\tilde{M}$  as a convex subset of the hyperbolic disk and  $C(\tilde{M})$  as a Cantor set in the boundary circle of the hyperbolic disk or the entire boundary if  $M$  has no boundary components.

If  $F : \tilde{M} \rightarrow \tilde{M}$  is the lift of  $f$  obtained from lifting the homotopy of  $f$  to the identity then  $F$  extends to  $C(\tilde{M})$  by setting  $F(z) = z$  for all  $z \in C(\tilde{M})$ . If  $x \in \tilde{M}$  is not a periodic point of  $F$ , but  $\pi(x) \in M$  is periodic then it is not difficult to see that  $\alpha(x, F)$  and  $\omega(x, F)$  are each a single point in  $C(\tilde{M})$  and these points are distinct.



Indeed, one can find these points as follows: If  $\pi(x)$  has period  $n$  let  $\gamma$  be the loop from  $\pi(x)$  to itself traced by  $\pi(x)$  under the homotopy of  $f^n$  to the identity which is the  $n$ -th iterate of the homotopy from  $f$  to the identity. The curve  $\gamma$  cannot be null homotopic since that would imply that  $x$  is periodic under  $F$ . Let  $\gamma_0$  be the closed geodesic in  $M$  freely homotopic to  $\gamma$  and given the same orientation as  $\gamma$ . If  $\Gamma$  is the lift of  $\gamma$  which starts at  $x$  then there is a lift  $\Gamma_0$  of  $\gamma_0$  which is a uniformly bounded distance from  $\Gamma$ . The curve  $\Gamma_0$  is then a geodesic in the hyperbolic plane which has its ends at  $\alpha(x, F)$  and  $\omega(x, F)$  in  $C(\widetilde{M})$ .

The setting in which we will apply Handel's fixed point theorem is this: The disk  $D$  we consider is  $\widetilde{M} \cup C(\widetilde{M})$  and the points  $x_i$  whose orbits form the oriented cycle of links will all have the property that  $\pi(x_i)$  is a periodic point of  $f$ .

Given a disk diffeomorphism  $F : D \rightarrow D$  and an orbit  $orb(x)$  with  $\omega(x)$  a single point  $w$  in  $\partial D$ , we will say that the orbit has a *simple periodic end* at  $w$  provided that there is some chart  $V$  centered at  $w$  on which there are coordinates  $(x, y)$ ,  $-\varepsilon < x < \varepsilon$ ,  $0 \leq y < \varepsilon$ ,  $\varepsilon > 0$ , in which the map  $F$  is given by

$$F(x, y) = (x/2, y/2).$$

We say that the orbit has a *homotopy periodic end* at  $w$  provided that for any neighborhood  $W$  of  $w$  in  $D$ , with  $x \notin W$ , there is a diffeomorphism  $g$  of  $D$  supported in  $W$  such that  $g^{-1} \circ F \circ g$  is isotopic relative to  $orb(x)$ , by an isotopy supported in  $W$ , to a new diffeomorphism  $F_0$  which has a simple end at  $w$  for the orbit of  $x$ . In other words, if we allow ourselves to change  $F$  by a conjugacy supported in a neighborhood of  $w$  and then by an isotopy relative to the orbit of  $x$  (under the conjugate diffeomorphism), we can create a simple periodic end at  $w$ .

A key lemma is the following one due to M. Handel. The short proof here was suggested by the referee.

**(3.2) Lemma.** *Suppose  $F$  is a diffeomorphism of  $\widetilde{M}$  which is a lift of  $f : M \rightarrow M$  as above, but with  $M$  having empty boundary (punctures are allowed). If  $x \in \widetilde{M}$  is not periodic under  $F$ , but  $\pi(x) \in M$  is periodic then for  $n$  sufficiently large, the orbit of  $x$  under the iterate  $F^n$  has homotopy periodic ends at points of  $C(\widetilde{M})$ .*

*Proof.* Replacing  $f$  and  $F$  with iterates we may assume that  $x \in \widetilde{M}$  is fixed by  $f$ . Let  $y \in \pi^{-1}(x)$ . The hypothesis then implies that there is a non-trivial covering translation  $\alpha : \widetilde{M} \rightarrow \widetilde{M}$  such that  $\alpha \circ F = F \circ \alpha$  and  $\alpha(y) = F(y)$ .

Note that  $\alpha$  is an isometry of the hyperbolic plane  $\widetilde{M}$  and both  $\alpha$  and  $F$  extend to  $C(\widetilde{M})$ , the "circle at infinity" of this hyperbolic plane. In the compactified hyperbolic plane,  $D = \widetilde{M} \cup C(\widetilde{M})$ , we have

$$\lim_{k \rightarrow \infty} \alpha^k(y) = w.$$

Clearly the orbit of  $y$  under  $\beta$  has a simple periodic end at  $w$ .

Let  $D_0$  denote  $D$  with  $w$  and the other fixed point of  $\alpha$  deleted. The quotient space of  $D_0$  under the action generated by  $\alpha$  is a closed annulus  $A$ . Since  $F$  and  $\alpha$  commute  $F$  induces a homeomorphism  $f_0 : A \rightarrow A$  which is isotopic to the identity. If  $y_0 \in A$  is the image of  $y$  under the quotient map then  $f_0(y_0) = y_0$ .

It is a well known and easy fact that if  $g$  is a diffeomorphism of the annulus which is isotopic to the identity and has a fixed point  $p$  then  $g$  is also isotopic to

the identity relative to  $p$ . Applying this to  $f_0$  with fixed point  $y_0$  we obtain an isotopy of  $f_0$  to the identity relative to  $y_0$ . If this isotopy is lifted to  $\widetilde{M}$  so that  $F$  is the lift of  $f_0$  we obtain an isotopy  $H_t$  from  $F$  to  $\alpha$  on all of  $D$  with the property that  $H_t(w) = w$  for  $0 \leq t \leq 1$ .

Without loss of generality we can assume that  $W$  is a neighborhood of  $w$  whose boundary is a segment  $\Gamma$  of a circle perpendicular to  $\partial D$  and joining two points of  $\partial D$  which are near  $w$ , i.e.  $\Gamma$  is a geodesic in  $\widetilde{M}$  together with its endpoints in  $C(\widetilde{M})$ . If  $m$  is large then  $W_0 = \alpha^m(W)$  is a small neighborhood of  $w$  and if  $m$  is sufficiently large then  $\alpha^{-1}(H_t(W_0))$  is disjoint from  $X = D \setminus \text{int}(W)$  for all  $t \in [0, 1]$ .

Consider the isotopy  $K_t(z)$  defined for  $z \in X \cup W_0 \cup P$  by  $K_t(z) = z$  for  $z \in X \cup P$  and  $K_t(z) = \alpha^{-1} \circ H_t(z)$  for  $z \in W_0$ . By the isotopy extension theorem (see (5.8) of [M]) we can extend  $K_t$  to an isotopy of all of  $D$  relative to  $P$  with  $K_0 = \text{id}$ . Then  $F \circ K_t^{-1}$  is an isotopy relative to  $P$  from  $F \circ K_0^{-1} = F$  to  $F \circ K_1^{-1}$  which on a neighborhood of  $w$  is equal to  $F \circ (\alpha^{-1} \circ F)^{-1} = \alpha$ . Thus  $F$  has a homotopy periodic end at  $w$  for the orbit of  $x$ . ■

**(3.3) Lemma.** *Suppose  $F$  is an orientation preserving  $C^1$  diffeomorphism of  $\mathbb{R}^2$  and  $p$  is a fixed point which is isolated in the set  $\text{Fix}(F^n)$  for all  $n > 0$ . If the index  $I(p, F) \leq 0$  then  $I(p, F^n) = I(p, F)$  for all  $n > 0$ .*

*Proof.* Let  $A$  be the matrix of the determinant  $Df_p$ . Since  $F$  preserves orientation  $\det(A) > 0$ . If  $(I - A)$  is non-singular then  $I(p, F) = \text{sgn } \det(I - A)$ . It follows that the eigenvalues of  $A$  cannot be negative or complex since then  $I(p, F) > 0$ . Thus the eigenvalues must be real and positive. In this case it follows from Theorem (2.2) of [CMY] that  $I(p, F^j)$  is independent of  $j$ , so the result follows. ■

**(3.4) Proposition (Handel).** *Suppose  $f : M \rightarrow M$  is a diffeomorphism homotopic to the identity where  $M$  is a surface of finite type with negative Euler characteristic. Let  $F$  be the canonical lift of  $f$  to the universal covering space  $\widetilde{M}$ . Suppose that there are points  $x_1, x_2, \dots, x_k \in \widetilde{M}$  whose orbits form an oriented cycle of links for  $F$  and such that each  $\pi(x_i)$  is a periodic point of  $f$  in the interior of  $M$ . Then  $F$  has a fixed point in the interior of  $\widetilde{M}$ . If the fixed points of  $F$  project to a finite set in  $M$  then  $F$  has a fixed point of positive index.*

*Sketch of Proof.* We first note that the case when  $M$  has some boundary components can be reduced to the case when it has none. We do this by simply deleting the boundary and applying the result to the restriction of  $f$  to the interior of  $M$ . Thus we will assume that  $M$  has no boundary and  $\widetilde{M}$  is the hyperbolic plane.

Let  $D = \widetilde{M} \cup C(\widetilde{M})$  be the compactification of  $\widetilde{M}$  obtained by adding ideal points on the boundary. Then  $F$  extends to all of  $D$  and is the identity on the boundary.

We note that to prove the result it suffices to find a fixed point for  $F^n$  for some  $n > 0$  and to show that if the fixed points of  $F$  are isolated then  $F^n$  has a fixed point of positive index. This is because of the fact that a homeomorphism of the plane which has a periodic point of period greater than one must also have a fixed point and a fixed point of positive index if the fixed points are isolated (see [BF], for example). Thus if we know that  $F^n$  has a fixed point, then  $F$  has a fixed point.

To get a positive index fixed point for  $F$ , assuming that  $F$  has isolated fixed points and there is a fixed point of positive index for  $F^n$ , we make the following observations. If  $F$  has a periodic point of period greater than one and its fixed

points are isolated then it has a fixed point of positive index by the result of [BF] mentioned above. The only other possibility is that  $F$  has fixed points of non-positive index and no other periodic points. But then Lemma (3.3) above asserts that for any such fixed point the index  $I(x, F^j)$  must be constant as  $j$  varies which would mean there is no fixed point of positive index for  $F^n$ . Thus this possibility cannot happen.

The advantage of proving our result only for  $F^n$  is that it allows us to use Lemma (3.2) and hence assume that the orbit of each  $x_i$  has a periodic end (for both forward and backward iterates). We will henceforth make this assumption and refer to the diffeomorphism as  $F$  rather than a power of  $F$ .

We proceed with a sketch of the proof in this case. Let  $P = \bigcup \text{orb}(x_i)$ . By Lemma (3.2)  $F : D \rightarrow D$  is isotopic relative to  $P$  to  $F_0$  which has a simple periodic end at each of the points  $w_i = \omega(x_i, F)$  and at each of the points  $a_i = \alpha(x_i, F)$ .

There is a hyperbolic metric (of constant curvature  $-1$ ) on the space  $X = D \setminus (P \cup \partial D)$ . One can see this by considering  $X$  as a countable union of “pairs of pants” glued together along geodesic boundary components, but with the “pairs of pants” having one “leg” a cusp rather than a geodesic boundary component. (Think of starting with a surface with two cusps and one geodesic boundary component. Add to it a surface with one cusp and two geodesic boundary components by gluing along a geodesic. Repeat this *ad infinitum*.)

Just as with surfaces with finitely many cusps every non-contractible closed curve in  $X$  is isotopic to a unique closed geodesic and every properly embedded line in  $X$ , with the property that neither component of its complement is contractible, is properly isotopic to a unique geodesic. These facts are relatively easy to establish using the fact that  $X$  is an increasing union of hyperbolic surfaces of finite type. It is also easy to see that if two properly embedded lines are disjoint then the geodesics isotopic to them are also disjoint. This can be seen by considering lifts of the curves to the universal covering space of  $X$  which is the hyperbolic plane. While we generally picture  $X$  as the disk  $D$  with punctures, it is important to note that boundary of  $D$  is not part of  $X$  and there is not a natural compactification of the hyperbolic manifold  $X$  which would add a circle of “ideal points” corresponding to this boundary.

Returning to  $D$  we let  $\hat{A}_i$  and  $\hat{W}_i$  denote small neighborhoods of the points  $a_i$  and  $w_i$  respectively. We choose these neighborhoods in the regions where  $F_0$  has periodic ends and so that  $F_0(\hat{W}_i) \subset \text{int}(\hat{W}_i)$  and  $\hat{A}_i \subset \text{int}(F_0(\hat{A}_i))$ , (see Figure 1). Let  $\rho_i$  denote the geodesic in  $X$  properly isotopic to  $\partial \hat{W}_i$  and denote by  $\rho_i(n)$  the geodesic properly isotopic to  $F_0^n(\partial \hat{W}_i)$ . Since the families of curves  $\{\rho_i(n)\}$  and  $\{F_0^n(\partial \hat{W}_i)\}$  have the same combinatorial pattern, relative to  $P$ , there is a diffeomorphism  $h : \text{int}(D) \rightarrow \text{int}(D)$  which is the identity on  $P$ , properly isotopic to the identity on  $X$ , and has the property that  $h(F_0^n(\partial \hat{W}_i)) = \rho_i(n)$  for all  $n > 0$ . Let  $G = h^{-1} \circ F_0 \circ h$  and let  $W_i = h(\hat{W}_i)$ ,  $A_i = h(\hat{A}_i)$ .

Since  $G$  is conjugate to  $F_0$  which has a periodic end at  $a_1$  we can choose an arc  $J_0 \subset \text{int}(A_1)$ , joining two consecutive points on the orbit of  $x_1$ , with the property that  $J_0$  and  $G(J_0)$  intersect only at their common endpoint. Clearly  $G^n(J_0) \cap J_0 = \emptyset$ , for all  $n > 1$ . Each of the arcs  $J_i = G^i(J_0)$ , if it has its endpoints removed, is a properly embedded line in  $X$  joining two cusps. Let  $K_i$  denote the geodesic in  $X$  isotopic to the interior of  $J_i$ . The disjointness of the  $J_i$  implies that  $K_p \cap K_q = \emptyset$  for all  $p \neq q$ .

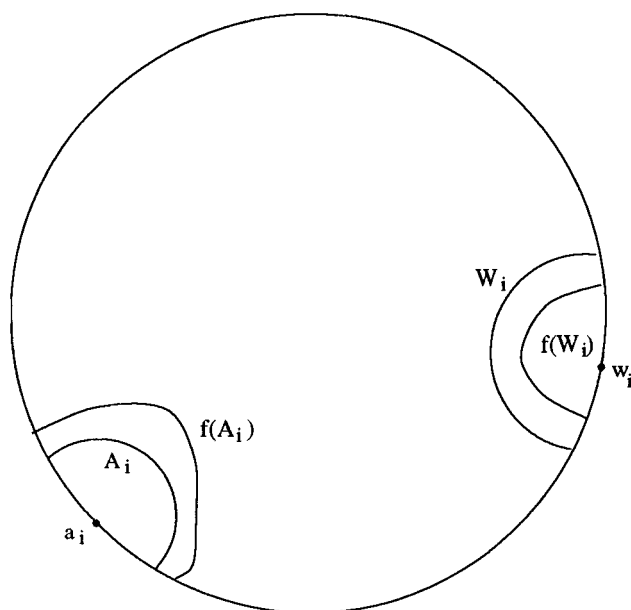


FIGURE 1

We next wish to show that for  $i$  sufficiently large  $J_i$  and  $K_i$  have non-empty intersection with  $W_2$ . To do this we let

$$B = \bigcup_{k=-\infty}^m G^k(J_0)$$

where  $m$  is chosen so large that the ends of  $G^m(J_0)$  are in  $W_1$ . Thus  $B$  is a properly embedded half line in  $D$  splitting  $A_1$  in two components and ending in a point of  $W_1$ . Notice that by construction  $B \cup A_2 = \emptyset$ . Similarly if  $Q$  is an arc joining two consecutive points of the orbit of  $x_2$  in the interior of  $W_2$ , we can form the properly embedded half line

$$C = \bigcup_{k=-r}^{\infty} G^k(Q)$$

splitting  $W_2$  and ending in a point of  $A_2$  with the property that  $C \cap W_1 = \emptyset$ . From the configuration of these arcs in  $D$ , which we know because of the hypothesis that the orbits of  $x_1$  and  $x_2$  link, it is clear that there must be a point  $z \in B \cap C$ . Thus if  $k$  is sufficiently large  $G^k(z) \in \text{int}(W_2)$  and  $G^k(z) \in J_i$  for some  $i$ . In fact this argument shows that for some  $i$  there must be a non-zero intersection number of  $J_i$  with one of the arcs in  $\text{int}(W_2)$  joining two consecutive points on the orbit of  $x_2$ . Since this intersection number will not change under isotopy relative to  $P$ , these two arcs can be replaced by the geodesics (in  $X$ ) isotopic to them. It follows that  $K_i$  must intersect a geodesic which joins two cusps and which lies entirely in  $W_2$ . Thus we have shown that for  $i$  sufficiently large  $J_i$  and  $K_i$  have non-empty intersection with  $W_2$ .

We want to consider subarcs of  $K_i$  which lie entirely in some  $W_j$  and others which lie entirely in the complement of  $\bigcup W_i$  in  $X$ . If  $T$  is a subarc of  $K_i$  which

joins points of the boundaries of  $W_p$  and  $W_q$  but otherwise lies in the complement of  $\bigcup W_i$  we want to think of isotoping  $G(T)$  to a segment of  $K_{i+1}$  keeping the endpoints in  $\bigcup W_i$  (recall that  $G(K_i)$  is isotopic to the geodesic  $K_{i+1}$  in  $X$ ). The problem is that there is no well defined “endpoint” of the arc as we carry out the isotopy from  $G(T)$  to a segment of  $K_{i+1}$ . There is, however, a well defined ending segment. We now make this precise.

Fix a large  $i$  and consider  $\Gamma_i$  the collection of segments of  $K_i \cap (X \setminus \bigcup \text{int}(W_j))$  which connect two different components of  $\bigcup \partial W_j$ . This set is not empty. In fact the argument above which showed that  $K_i \cap W_2 \neq \emptyset$  implies that there are at least two distinct segments in  $\Gamma_i$  which have one endpoint in  $\partial W_1$  and two (perhaps coinciding with the previous two) with one endpoint in  $\partial W_2$ . If  $\gamma \in \Gamma_i$  then it is a geodesic segment with endpoints in  $\bigcup \partial W_j = \bigcup \rho_j$ , say it ends in  $\rho_p$  and  $\rho_q$ . The arc  $G(\gamma)$  has endpoints in the geodesics  $\rho_p(1) = G(\rho_p)$  and  $\rho_q(1) = G(\rho_q)$ . The proper isotopy from  $G(K_i)$  to  $K_{i+1}$  can be done in such a way that the finite set of points of intersection of  $G(K_i)$  with the geodesics  $\{\rho_i(n) \mid n > 0\}$  varies continuously with no new points of intersection created and none destroyed. This can be proved, for example, by considering “curve shortening” and using the fact that the properly embedded line  $G(K_i)$  and the geodesic  $\rho_j(n)$  are the images under the diffeomorphism  $G$  of the geodesic  $K_i$  and the geodesics  $\rho_j(n-1)$ .

Let  $L_1(\gamma)$  be the segment of  $K_{i+1}$  with endpoints in  $\rho_p(1)$  and  $\rho_q(1)$  which is isotopic to  $G(\gamma)$  by the isotopy which keeps endpoints in  $\rho_p(1)$  and  $\rho_q(1)$  as just described. And let  $L(\gamma)$  denote the subset of  $\Gamma_{i+1}$  consisting of subarcs of  $L_1(\gamma)$ . Thus  $L(\gamma)$  is the finite set of all those subarcs of the arc  $L_1(\gamma) \subset K_{i+1}$  which lie in  $X \setminus \bigcup \text{int}(W_m)$  with their endpoints in distinct components of  $\bigcup \partial W_j$ . If we inductively define

$$L^{n+1}(\gamma) = \bigcup_{\beta \in L^n(\gamma)} L(\beta),$$

then one can show inductively that  $L^n(\gamma)$  is the set of all those subarcs of the arc  $L_n(\gamma) \subset K_{i+n}$  which lie in  $X \setminus \bigcup \text{int}(W_m)$  with their endpoints in distinct components of  $\bigcup \partial W_j$ , where  $L_n(\gamma)$  is the segment of  $K_{i+n}$  with endpoints in  $\rho_p(n)$  and  $\rho_q(n)$  which is isotopic to  $G^n(\gamma)$  by an isotopy which keeps endpoints in  $\rho_p(n)$  and  $\rho_q(n)$ .

Finally let

$$\Gamma = \bigcup_j \Gamma_j,$$

where the union is over all  $j \geq i$ . Since the segments in  $\Gamma$  are all subarcs of the pairwise disjoint geodesics  $\{K_j\}$ , they are pairwise disjoint arcs embedded in  $X \setminus \bigcup \text{int}(W_m)$  and it follows that there are finitely many equivalence classes  $[\gamma]$ ,  $\gamma \in \Gamma$ , where  $[\gamma]$  denotes the isotopy class in  $X \setminus \bigcup \text{int}(W_m)$  by an isotopy which keeps the endpoints in  $\bigcup \partial W_m$ .

Notice that if  $[\gamma_0] = [\gamma_1]$  then the sets  $L(\gamma_0)$  and  $L(\gamma_1)$  have the same number of elements in each isotopy class. We form a finite graph with one vertex for each equivalence class  $[\gamma]$  and an oriented edge from  $[\gamma_j]$  to  $[\gamma_k]$  if  $\gamma_k \in L(\gamma_j)$ .

We wish to show that this graph has a vertex with multiple oriented paths of the same length from that vertex to itself. This follows from an argument like the argument above which showed that  $K_n$  intersected  $W_2$  for large  $n$ . In fact if  $\gamma$  is an arc from  $\partial W_j$  to  $\partial W_k$  then our hypothesis of an oriented cycle of links

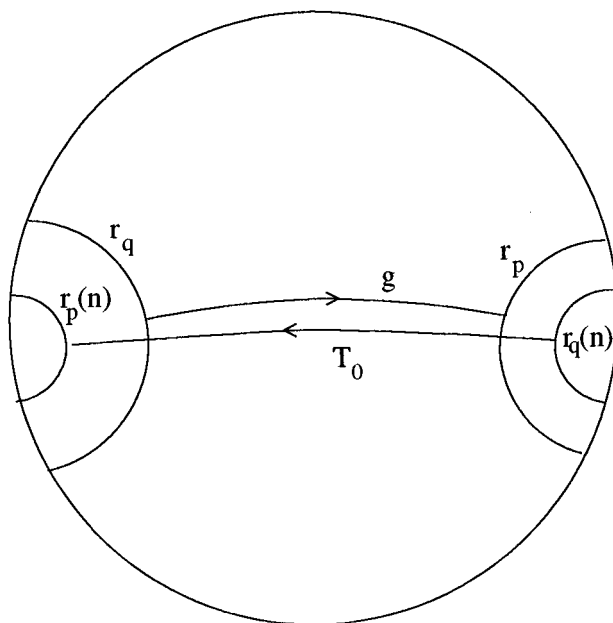


FIGURE 2

implies that for some  $m$ , the points  $w_j$  and  $w_k$  separate  $a_m$  and  $w_m$  in  $\partial D$ . Hence an argument like the one just referred to implies that for some  $N > 0$  the arcs  $G^n(\gamma)$ ,  $n > N$  run from a point of  $W_j$  to a point of  $W_k$  but pass through  $W_m$  in a way that is homotopically essential relative to  $P$ . From this we can conclude that  $L^n(\gamma)$  contains at least two arcs in distinct isotopy classes. This means that on our graph there are paths of length  $n$  from  $[\gamma]$  to at least two distinct vertices. But  $\gamma$  was arbitrary and  $n$  was any sufficiently large integer. It follows that there is one  $n$  with the property that from any vertex of the graph there are paths of length  $n$  to two distinct vertices. From this it is easy to see that there must be a vertex contained in two distinct cycles. If the two cycles are of different length then using multiples of each of them will give two cycles of the same length.

Consequently there is a  $\gamma \in \Gamma$  and an  $n$  such that  $L^n(\gamma)$  contains two subarcs of  $K_{i+n}$  which are isotopic to  $\gamma$  keeping their endpoints in  $\bigcup \partial W_j$ . Of course there may be more than two such subarcs. If all such subarcs are given an orientation which is the restriction of an orientation on  $K_{i+n}$  then there must be two with opposite orientations. In fact it is not difficult to see using the Jordan curve theorem that two adjacent arcs from such a collection must have opposite orientations.

Hence the arc  $G^n(\gamma)$  is isotopic to a subarc  $L_n(\gamma)$  of  $K_{i+n}$  by an isotopy which keeps the endpoints in  $\bigcup G^n(\partial W_j)$ . Moreover  $L_n(\gamma)$  contains a subarc  $\tau$  such that  $[\tau] = [\gamma]$  but their orientations are reversed. Let  $T = L_n(\gamma)$ .

Consider the universal covering space  $\tilde{X}$  of  $X$  and let  $g$  be a lift of  $\gamma$  to  $\tilde{X}$  joining  $r_p$  and  $r_q$ , lifts of  $\rho_p$  and  $\rho_q$  respectively (see Figure 2). Let  $T_0$  be the lift of  $T$  as shown in Figure 2 joining lifts  $r_p(n)$  and  $r_q(n)$  of  $\rho_p(n)$  and  $\rho_q(n)$  respectively. Now  $T_0$  can be isotoped, keeping its endpoints in  $r_q(n)$  of  $\rho_p(n)$ , to a lift of  $G^n(\gamma)$ . Let  $H : \tilde{X} \rightarrow \tilde{X}$  be the lift of  $G^n$  which maps  $g$  to this lift of  $G^n(\gamma)$  and extend  $H$  to  $S_\infty$ , the circle at infinity of the hyperbolic plane  $\tilde{X}$ .

Let  $I_0$  and  $I_1$  be disjoint intervals in  $S_\infty$  which join the endpoints of  $r_p$  and of  $r_q$  respectively. Notice that  $r_p(n) = H(r_p)$  and  $r_p$  lie on opposite sides of  $r_q$ . Likewise  $r_q(n) = H(r_q)$  and  $r_q$  lie on opposite sides of  $r_p$ . This means that  $H(I_0) \subset I_1$  and  $H(I_1) \subset I_0$ . It follows that  $H$  has a point of period 2 on  $S_\infty$ ; in particular, it has no fixed points there. But  $H$  is a homeomorphism of the disk  $\tilde{X} \cup S_\infty$  and so must have a fixed point. Since  $G$  is properly isotopic to  $F$  as maps of  $X$  there is a proper isotopy from  $H$ , a lift of  $G^n$ , to  $H_1$ , a lift of  $F^n$ , obtained by lifting the isotopy from  $G^n$  to  $F^n$ . Since  $H$  and  $H_1$  are lifts of properly isotopic diffeomorphisms on  $X$  they induce the same map on  $S_\infty$ . Thus  $H_1$  must have a fixed point. If the fixed points of  $H_1$  are isolated, it must have a point of positive index. If the fixed points of  $H_1$  are not isolated, then either  $F$  has a periodic point or the fixed points of  $F$  are not isolated.

It follows that  $F^n$  has a fixed point and if  $F$  has isolated fixed points then it has fixed point of positive index. ■

#### 4. THE CONVEX HULL OF THE ROTATION SET

**(4.1) Theorem.** *Let  $M$  be a compact orientable surface with negative Euler characteristic. Suppose  $f : M \rightarrow M$  is a diffeomorphism of the surface  $M$  which is isotopic to the identity map and preserves a smooth probability measure  $\mu$  and has at most finitely many interior fixed points. If  $0$  is in the interior (in  $H_1(M, \mathbb{R})$ ) of the convex hull of the rotation set  $\mathcal{R}(f) = \bigcup \mathcal{R}(x, f)$ , then  $f$  has an interior fixed point of positive index. In fact the canonical lift of  $f$  to its universal cover  $F : \tilde{M} \rightarrow \tilde{M}$  has an interior fixed point which is of positive index.*

*Proof.* According to Steinitz's theorem (see [HDK], page 44) any point in the interior of the convex hull of a set in  $\mathbb{R}^m$  is in fact in the interior of the convex hull of  $2m$  points in that set. Thus if  $m = \dim H_1(M, \mathbb{R})$ , there are  $2m$  vectors  $\{v_i\}$  in the rotation set,  $\mathcal{R}(f)$ , such that  $0$  is in the interior of their convex hull.

Our strategy is to perturb  $f$  slightly to a new map  $g$  which we can prove has a fixed point of positive index and which we can show has precisely the same fixed points, with the same index, as  $f$ .

For each  $i$  there is a point  $x_i$  with  $v_i = \mathcal{R}(x_i, f)$ . We would like to choose  $x_i$  to be an interior periodic point of  $f$ . If there is no such  $x_i$ , e.g. if  $v_i$  has irrational coordinates, we will perturb  $f$  to  $g_0$  in a small neighborhood of  $x_i$  so that there is a periodic point  $y_i$  with  $w_i = \mathcal{R}(y_i, g_0)$  with  $w_i$  so close to  $v_i$  that  $0$  is in the interior of the convex hull of  $\{w_i\}$ . We proceed with the details.

We form a new surface  $N$  by puncturing  $M$  at the finitely many fixed points and by deleting any boundary components of  $M$  which contain a point with a lift in  $\tilde{M}$  which is fixed. We let  $\tilde{N} = \pi^{-1}(N) \subset \tilde{M}$  where  $\pi : \tilde{M} \rightarrow M$  is the universal covering. By (1.4) there is a complete Riemannian metric on  $N$  which induces a metric on  $\tilde{N}$  with the property that  $d(F(y), y) > 1$  for every  $y \in \tilde{N}$ .

Given a periodic point  $y$  for a diffeomorphism  $f$  isotopic to the identity we denote by  $L(y, f)$  the closed loop formed by joining the arcs from  $f^j(y)$  to  $f^{j+1}(y)$  traced out by the isotopy from  $f$  to the identity (here the index  $j$  varies from 0 to one less than the period of  $y$ ). We orient this arc in a direction from  $f^j(y)$  to  $f^{j+1}(y)$  so that the loop formed has a distinguished orientation.

**(4.2) Lemma.** *There is an isotopy of  $f$  to  $f_0$  such that*

- i) *The support of the isotopy is on a compact subset of  $N$ , i.e.  $f_0$  and  $f$  agree outside of a compact subset of  $N$ .*

- ii)  $d(F_0(y), F(y)) < 1/2$  for all  $x \in \tilde{N}$ , where  $F_0$  is the lift of  $f_0$  to  $\tilde{M}$ , obtained by lifting the isotopy.
- iii) There are periodic points  $y_i$  of  $f_0$  in the interior of  $M$  such that either  $L(y_i, f_0)$  is null homotopic for some  $i$ , or if  $\gamma_i$  denotes the geodesic (in a hyperbolic metric on  $M$ ) which is freely homotopic to  $L(y_i, f_0)$ , then the complement of  $\bigcup \gamma_i$  in  $M$  has only components which are either contractible or have genus zero and contain a single puncture or a boundary component of  $M$ .
- iv) There are positive integers  $c_i$  such that  $\sum c_i[\gamma_i] = 0 \in H_1(M, \mathbb{Z})$ .

We postpone the proof of this lemma and complete the proof of (4.1). Note that it suffices to find a fixed point of positive index for  $F_0 : \tilde{M} \rightarrow \tilde{M}$ . This is because  $\tilde{f}$  and  $f_0$  agree except on a compact set  $K$  of  $N$ , so  $F = F_0$  except on  $\pi^{-1}(K) \subset \tilde{N}$ , and the fact that  $F$  has no fixed points in  $\tilde{N}$  together with ii) above implies that  $F_0$  has no fixed points in  $\tilde{N}$  either. Thus any fixed point of  $F_0$  is also a fixed point of  $F$  and  $F = F_0$  on a neighborhood of this fixed point. Thus we will have completed the proof of (4.1) (modulo the proof of (4.2)) if we prove the following.

**(4.3) Lemma.** *Suppose that  $f_0$  and its lift  $F_0$  are as in Lemma (4.2). Then  $F_0$  has a fixed point of positive index.*

*Proof.* If in part iii) of Lemma (4.2) it is the case that  $L(y_i, f_0)$  is null homotopic for some  $i$ , then a lift of  $y_i$  to  $\tilde{M}$  is a periodic point of  $F_0$ . But since  $\tilde{M}$  is homeomorphic to the plane it follows from the assumption that  $F_0$  has a discrete fixed point set and the Brouwer plane translation theorem that in this case  $F_0$  has a fixed point of positive index (see [BF], for example).

In the other case we will show that  $F_0$  satisfies the hypothesis of Theorem (3.4) and hence has the desired fixed point. In fact the points whose orbits form an oriented cycle of links are lifts of the periodic points  $y_i$  we obtained in (4.2) above.

A lift of the orbit of  $y_i$  will be asymptotic to a lift of the oriented geodesic  $\gamma_i$ , and conversely. Thus to show that there are lifts of the points  $y_i$  whose orbits form an oriented cycle of links we need only show there are lifts of the oriented geodesics  $\{\gamma_i\}$  which form an oriented cycle of links. For this it suffices to find a contractible polygon in  $M$  whose boundary is made of subarcs of the geodesics in  $\{\gamma_i\}$  and with the property that the orientations of the boundary segments are all compatible, i.e. form a cyclic ordering on the boundary. If we find such a polygon and lift it to  $\tilde{M}$  and extend its geodesic sides infinitely then the collection of geodesics so obtained will be lifts of geodesics in  $\{\gamma_i\}$  and will form the desired oriented cycle of links.

Consider the components of  $W = M \setminus \bigcup \gamma_i$ . We want to construct an arc in  $M$  which runs from component to component but always crosses between components in such a way that it intersects the  $\gamma_i$  with an intersection number of the same sign. If  $M$  has punctures or boundary components we start our arc, call it  $\alpha$ , in a component of  $W$  containing a puncture or boundary component; otherwise any component will do. From there we cross into a neighboring component, crossing with intersection number 1 or  $-1$ . This determines the sign of the intersection of all crossings we will allow.

We continue this procedure until one of three conditions is met: 1) we reach a different component of  $W$  which contains a puncture or boundary component, 2) we re-enter a component we have already visited, or 3) we enter a component with no



boundary or puncture whose boundary is consistently oriented. To see that we can always arrive at one of these conditions note that whenever we enter a component of  $W$  whose boundary segments are inconsistently oriented, it is possible to leave this region with a crossing having the same intersection number as the crossing by which we entered. Thus since there are finitely many components we must terminate with one of the three conditions satisfied.

Of course, 3) is the one we want. If this occurs the component of  $W$  in which the arc ends is a polygon whose boundary is consistently oriented and the theorem is proved as we remarked above. Thus we wish to show that 1) and 2) above are impossible.

Consider 2) first. In this case we can take the part of our arc which leaves the final component and returns to it and close up this segment to form a closed loop  $\beta$  in  $M$  which can be oriented to have positive intersection number with every geodesic  $\gamma_i$ . But this would mean that there is a positive intersection pairing of  $\beta$  with  $\sum c_i[\gamma_i]$  where  $c_i$  are the positive integers given in iv) of Lemma (4.2) above. On the other hand  $\sum c_i[\gamma_i] = 0$  so this is a contradiction.

Case 1) is very similar. In this case there is an embedded arc running from one cusp (puncture) or boundary component to another. This arc can be oriented to have positive intersection number with each  $\gamma_i$ . It follows that its intersection number with  $\sum c_i[\gamma_i]$  is positive, which is again a contradiction. In this case the intersection pairing we use is a bilinear map from  $H_1(M)$  and  $H_1(M, X)$  to the integers, where  $X$  is a small neighborhood of the punctures and boundary of  $M$  and the arc we constructed determines an element of  $H_1(M, X)$ . This completes the proof of (4.3) (and hence (4.1)) except for the proof of Lemma (4.2) which we now address. ■

We begin the proof of (4.2) with the observation that if  $\varepsilon > 0$  is sufficiently small then any  $\varepsilon$ -chain for  $f$  has a unique lift to an  $\varepsilon$ -chain for  $F : \widetilde{M} \rightarrow \widetilde{M}$ . It follows that given any  $\varepsilon$ -chain for  $f$  from a point  $b_0$  to itself (with  $\varepsilon$  small) and given a point  $b \in \widetilde{M}$  with  $\pi(b) = b_0$ , we can associate with them a free homotopy class of loops in  $M$ . This class is just the homotopy class of the projection to  $M$  of an arc joining the two ends of a lift of the periodic  $\varepsilon$ -chain. We will consider a hyperbolic metric on  $M$  and represent this free homotopy class by the (unique) closed geodesic contained in it. If the  $\varepsilon$ -chain is  $\mathcal{D}$ , then the corresponding geodesic will be denoted  $\gamma_{\mathcal{D}}$ .

Recall that we formed the surface  $N$  by puncturing  $M$  at the finitely many fixed points of  $f$  and by deleting any boundary components of  $M$  which contain a point with a lift in  $\widetilde{M}$  which is fixed and we defined a complete Riemannian metric on  $N$  which induces a metric on the covering space  $\widetilde{N}$  with the property that  $d(F(y), y) > 1$  for every  $y \in \widetilde{N}$ .

**(4.4) Lemma.** *Suppose that  $f$  satisfies the hypothesis of (4.1). Then for any  $\varepsilon > 0$ , and any element  $v \in H_1(M, \mathbb{Z})$  there exist an  $\varepsilon$ -chain  $\mathcal{D}$  from some point  $b_0$  to itself and a positive integer  $a$  with  $av = [\gamma_{\mathcal{D}}] \in H_1(M, \mathbb{Z})$ . Moreover this  $\varepsilon$ -chain is an  $\varepsilon$ -chain for  $f$  restricted to  $N$  in the metric  $d(\cdot, \cdot)$  on  $N$ .*

*Proof.* Recall that 0 is in the interior of the convex hull of  $\{v_i = \mathcal{R}(x_i, f)\}$ . Clearly we can assume each  $v_i$  is non-zero. Moreover, because of Proposition (2.5) we can replace each point  $x_i$  by a finite set  $\{y_{ij}\}$  of recurrent points with the property that

$v_i$  is very close to a convex combination of the rotation vectors  $\{\mathcal{R}(y_{ij}, f)\}$ . If these convex combinations are sufficiently good approximations to the vectors  $v_i$ , then 0 will be in the interior of the convex hull of these approximations and hence in the interior of the convex hull of  $\{\mathcal{R}(y_{ij}, f)\}$ . To put this another way, it is possible to choose the set  $x_i$  (namely by choosing it to be the collection of all  $y_{ij}$ ) so that each  $x_i$  is recurrent and 0 is in the interior of the convex hull of  $\{\mathcal{R}(x_i, f)\}$ . We assume this to have been done and continue to refer to  $\mathcal{R}(x_i, f)$  as  $v_i$ .

We note that in  $M$  the recurrent points form a set of full measure. Hence the recurrent points of  $f$  restricted to  $N$  are dense. It now follows from Proposition (1.2) that for any  $\varepsilon > 0$  there is an  $\varepsilon$ -chain from a point  $b_0 \in N$  to  $x_i$  and one from  $x_i$  to  $b_0$  for each  $i$ . These  $\varepsilon$ -chains are with respect to the metric on  $N$  which induces the distance function  $d(\cdot, \cdot)$  described above, whose existence follows from Lemma (1.4). If  $\varepsilon$  is sufficiently small, then any  $\varepsilon$ -chain  $p_1, p_2, \dots, p_n$  for  $f$  lifts to an  $\varepsilon$ -chain for  $F$  in  $M$  and it is unique once its first point in  $\pi^{-1}(p_1)$  is specified.

For each  $i$  we form an  $\varepsilon$ -chain from  $b_0$  to  $x_i$ , then along a long stretch of the orbit of  $x_i$  to  $x_i$  again and then back to  $b_0$ . This is done by concatenating the chain from  $b_0$  to  $x_i$ , a long stretch of orbit and then the chain from  $x_i$  to  $b_0$ . If this chain is lifted to  $\widetilde{M}$  it will go from  $z_0 \in \pi^{-1}(b_0)$  to  $\alpha(z_0)$  for some deck translation  $\alpha$ . Let  $w_i$  denote the homology class of a loop in  $M$  formed by projecting a path in  $\widetilde{M}$  from  $z_0$  to  $\alpha(z_0)$ . If  $k$  is the number of points in the  $\varepsilon$ -chain then it is clear that by choosing the segment of actual orbit of  $x_i$  that is used in this construction sufficiently large we can make  $w_i/k$  as close as we wish to  $v_i$ . This is because the beginning and ending segments of the chain (from  $b_0$  to  $x_i$  and back) have fixed length but the orbit segment from  $x_i$  to a point very near  $x_i$  can be chosen arbitrarily long and  $\mathcal{R}(x_i) = v_i$ . We do this construction so that 0 is in the interior of the convex hull of  $\{w_i\}$ . Call these  $\varepsilon$ -chains  $\mathcal{C}_i$ .

Now given any element  $v \in H_1(M, \mathbb{Z})$  we can construct an  $\varepsilon$ -chain  $\mathcal{D}$  from  $b_0$  to  $b_0$  which represents a positive multiple of  $v$  in  $H_1(M, \mathbb{R})$ . This is because zero is in the interior of the convex hull of  $\{w_i\}$  so there is a solution of

$$\sum a_i w_i = av$$

with  $a > 0$  and  $a_i \geq 0$ . Since each  $w_i$  is a rational vector, if the vector  $v$  is rational the numbers  $a_i$  can be chosen to be positive rationals. Multiplying both sides by a positive integer, we can assume that  $a$  and each of the  $a_i$  are integers. Now if we let  $\mathcal{D}$  be the concatenation of  $a_1$  copies of  $\mathcal{C}_1$ , with  $a_2$  copies of  $\mathcal{C}_2$ , and  $a_3$  copies of  $\mathcal{C}_3$  etc. we have the desired  $\mathcal{D}$ . ■

The following lemma essentially says that Lemma (4.2) is valid without the necessity of perturbing  $f$  if instead of periodic orbits we are satisfied with periodic  $\varepsilon$ -chains.

Just as a periodic orbit determines a free homotopy class of loops, as described above, so does a periodic  $\varepsilon$ -chain  $\{y_i\}$  for  $f$  with  $\varepsilon$  sufficiently small. We will denote by  $L_0(\{y_i\}, f)$  the closed loop formed by joining the arcs from  $y_i$  to  $f(y_i)$  traced out by the isotopy from  $f$  to the identity and short arcs from  $f(y_i)$  to  $y_{i+1}$ .

**(4.5) Lemma.** *If  $f : M \rightarrow M$  is as in (4.1), then*

- i) *There are periodic  $\varepsilon$ -chains  $\{y_i\}$  for  $f$  in the interior of  $M$  such that either  $L_0(\{y_i\}, f)$  is null homotopic for some  $i$ , or if  $\gamma_i$  denotes the geodesic (in a*

hyperbolic metric on  $M$ ) which is freely homotopic to  $L_0(\{y_i\}, f)$ , then the complement of  $\bigcup \gamma_i$  in  $M$  has only components which are either contractible or have genus zero and contain a single puncture or a boundary component of  $M$ .

- ii) There are positive integers  $c_i$  such that  $\sum c_i[\gamma_i] = 0 \in H_1(M, \mathbb{Z})$ .

*Proof.* We need first to select a set of homology classes which can represent loops as above which are freely homotopic to geodesics that will divide the surface  $M$  into pieces with the desired properties. Recall that these geodesics are with respect to a hyperbolic metric on  $M$ .

Assume first that  $M$  is of genus greater than zero. Then we can choose a homology class  $r_1 \in H_1(M, \mathbb{Z})$  which has non-zero intersection pairing with some other homology class in  $H_1(M, \mathbb{Z})$ . This, of course, means that  $r_1$  is non-zero and that it cannot be represented by a peripheral loop. A peripheral loop is one which can be homotoped to either a boundary component of  $M$  or an arbitrarily small neighborhood of a puncture. We construct the  $\varepsilon$ -chain  $\mathcal{D}_1$  as above and let  $\gamma_1$  be the geodesic freely homotopic to the loop determined by  $\mathcal{D}_1$ . This geodesic will likely have multiplicity greater than one, i.e. it will run several times around the same primitive geodesic. Also the homology class  $[\gamma_1]$  determined by  $\gamma_1$  will be  $a_1 r_1$  for some positive integer  $a_1$ .

Next we select a finite set  $\{r_i\}$  of non-trivial homology classes in  $H_1(M, \mathbb{Z})$  satisfying certain properties. We require that the intersection pairing of  $r_1$  and  $r_i$  be non-zero when  $i > 1$ . Also we let  $X$  be a small neighborhood of the union of the boundary and punctures of  $M$  and require that for any non-trivial  $u \in H_1(M, X, \mathbb{Z})$  there is an  $r_j$  which has non-zero intersection pairing with  $u$ . It is not difficult to see that this is possible with a finite collection  $\{r_i\}$ .

As before, for each  $r_i$ , we construct the  $\varepsilon$ -chain  $\mathcal{D}_i$  as above and let  $\gamma_i$  be the geodesic freely homotopic to the loop determined by  $\mathcal{D}_i$ . The homology class  $[\gamma_i]$  will be  $a_i r_i$  for some positive integer  $a_i$ .

We consider the complement of  $\bigcup \gamma_i$  in  $M$  which consists of a finite number of components. Consider one of these components  $S$ . It cannot contain more than one boundary component or puncture of  $M$  since then it would contain an arc joining two components of  $X$  which would represent a non-zero element  $u \in H_1(M, X, \mathbb{Z})$  which has zero intersection pairing with every  $r_i$  and this would contradict our choice of  $\{r_i\}$ .

We will show that  $S$  must have genus zero. If it were topologically a surface (with boundary) of positive genus it would contain two simple closed curves with non-zero intersection number. But this is impossible since either one of these curves would represent a non-zero  $u \in H_1(M, X, \mathbb{Z})$  which has zero intersection pairing with every  $r_i$ . Therefore,  $S$  has genus zero.

Finally suppose  $S$  contains no boundary component or puncture of  $M$ . We will show that it is contractible. If not, it contains a simple closed curve  $\beta$  which is non-trivial in  $H_1(S)$ . This curve cannot separate  $M$  since both components of the complement would contain points of  $\bigcup \gamma_i$  and this set is connected (each  $\gamma_i$  intersects  $\gamma_1$ ). But the only other possibility is that  $\beta$  has non-zero intersection number with a loop in  $M$  formed by taking a small transversal to  $\beta$  and joining its ends in the complement of  $\beta$ . Hence the homology class of  $\beta$  is non-zero in  $H_1(M, X, \mathbb{Z})$  which also contradicts the construction of  $\{r_i\}$ . We conclude that  $S$  must be contractible or contain a single puncture or boundary component. Thus we have shown that  $S$

either has genus zero and contains a single puncture or boundary component or  $S$  is contractible.

It follows that  $\{\gamma_i\}$  will satisfy i) of the lemma. To arrange for ii) to be satisfied we must add one more geodesic  $\gamma_0$ . We consider the homology class  $v = -\sum_{i>0}[\gamma_i]$  and construct the corresponding  $\varepsilon$ -chain  $\mathcal{D}_0$  and geodesic  $\gamma_0$ . Since a multiple of  $[\gamma_0]$  is equal to  $v$ , and both these classes are integral, it is clear we can find integers  $c_i$  such that

$$\sum_{i=0} c_i [\gamma_i] = 0.$$

Clearly the collection of geodesics enlarged by adding  $\gamma_0$  still satisfies i).

To achieve the same result in case that  $M$  is of genus zero the argument is somewhat different. Since we can find an  $\varepsilon$ -chain from  $b_0$  to itself representing any a positive multiple of any integer homology class we can, in particular, do this for the zero class. In this case there will be a single periodic  $\varepsilon$ -chain for  $f$ . The loop corresponding to this  $\varepsilon$ -chain might be null homotopic – this is the first possibility listed in i). If it is not, let  $\gamma$  be the geodesic freely homotopic to it. Then since  $M$  has genus zero every component of the complement of  $\gamma$  must either be contractible or contain either a boundary component or puncture. Clearly ii) is satisfied in this case with  $\gamma_0 = \gamma$  and  $c_0 = 1$ . ■

We are now ready to complete the proof of Lemma (4.2). Lemma (4.5) gives us a set of geodesics with properties iii) and iv) of Lemma (4.2), but they do not correspond to orbits of periodic points, but only to periodic  $\varepsilon$ -chains of  $f$ . We will perturb  $f$  to  $f_0$  in such a way that each of these chains becomes a periodic orbit for  $f_0$  and so the geodesics corresponding to these periodic orbits are precisely the same as those for the periodic  $\varepsilon$ -chains of  $f$ .

By making arbitrarily small changes in the points of these  $\varepsilon$ -chains we can arrange that, except for the first and last point of each periodic chain (which are equal), all points of the chains as well as their images under  $f$  are distinct from one another. We assume this to have been done.

Let  $x_1, x_2, \dots, x_n$  be the points of all the periodic  $\varepsilon$ -chains  $\{\mathcal{D}_j\}$ . By assumption, each  $x_k$  has a “successor”  $x'_k$  in this set such that  $d(f(x_k), x'_k) < \varepsilon$ , so for each  $k$  there is a smooth curve  $\beta_k : [0, 1] \rightarrow N$  of length  $< \varepsilon$  joining  $f(x_k)$  and  $x'_k$ . We choose a parameterization of the curves  $\beta_k$  proportional to arc length and perturb their embeddings if necessary so that whenever they cross they do so at different values of the parameter, i.e.  $i \neq j$  implies  $\beta_i(s) \neq \beta_j(s)$  for every  $s \in [0, 1]$ .

Let  $\delta = \min d(\beta_i(s), \beta_j(s))$  where the minimum is over all  $s \in [0, 1]$  and all pairs  $i, j$  with  $i \neq j$ . Choose an integer  $\mathcal{N}$  such that  $1/\mathcal{N} < \delta/2$ . Then each of the smooth arcs  $\beta_k([(i-1)/\mathcal{N}, i/\mathcal{N}])$ ,  $1 \leq i \leq \mathcal{N}$ ,  $1 \leq k \leq n$  has a length  $< \varepsilon/\mathcal{N}$ . Also for each fixed  $i$  the arcs  $\beta_k([(i-1)/\mathcal{N}, i/\mathcal{N}])$ ,  $1 \leq k \leq n$  are pairwise disjoint.

For each integer  $j$  between 1 and  $\mathcal{N}$  we construct an isotopy  $h_t(j) : M \rightarrow M$  with the following properties:

- i) The support of  $h_t(j)$  is contained in the union of small pairwise disjoint neighborhoods of the arcs  $\beta_k([(j-1)/\mathcal{N}, j/\mathcal{N}])$ ,
- ii)  $d(h_t(j)(x), x) < \varepsilon/\mathcal{N}$  for all  $x \in N$ ,
- iii)  $h_1(j)(\beta_k((j-1)/\mathcal{N})) = \beta_k(j/\mathcal{N})$  for  $1 \leq j < \mathcal{N}$ .

Let  $h : M \rightarrow M$  be defined by  $h_t = h_t(\mathcal{N}) \circ h_t(\mathcal{N}-1) \circ \dots \circ h_t(1)$ . Then clearly  $d(h_t(x), x) < \varepsilon$  for all  $x \in N$  so part ii) of our lemma is satisfied. Also

$h_1(f(x_i)) = x'_i$  and thus the homeomorphism  $f_0 = h_1 \circ f$  has each of the periodic  $\varepsilon$ -chains  $\mathcal{D}_i$  as a periodic orbit so part iii) is satisfied. Clearly by construction i) is satisfied. ■

## 5. FIXED POINTS AND THE MEAN ROTATION VECTOR

If  $\widetilde{M}$  is the universal covering space of  $M$  we denote by  $F : \widetilde{M} \rightarrow \widetilde{M}$  the canonical lift of  $f$  which fixes the “points at infinity” of  $\widetilde{M}$ , i.e. the lift obtained by starting with the identity map of  $\widetilde{M}$  and lifting the homotopy on  $M$  from the identity to  $f$ . We will identify  $\Pi_1(M)$  with the group of covering transformations for the universal cover. For any element  $\alpha \in \Pi_1(M)$  we will denote the homology class it determines by  $[\alpha]$ .

**(5.1) Theorem.** *Suppose  $M$  is an oriented compact surface with negative Euler characteristic and  $f : M \rightarrow M$  is a diffeomorphism isotopic to the identity which preserves the smooth measure  $\mu$  and has  $\mathcal{R}_\mu(f) = 0$ . Then there is an interior fixed point of  $f$  which has positive index and which lifts to a fixed point of  $F$  the canonical lift of  $f$  to the universal covering space  $\widetilde{M}$ .*

*Proof.* Choose a norm  $\| \cdot \|$  on  $H_1(M, \mathbb{Z})$  and pick a basis of it represented by a family  $\{C_i\}$  of oriented embedded closed curves in  $M$ . Let  $\{A_i\}$  be a family of annuli with each  $A_i$  a compact tubular neighborhood of the corresponding  $C_i$ .

Fix  $i$  and construct a  $\mu$  preserving flow  $\phi_t$  on  $M$  supported in the interior of  $A_i$  and with each non-trivial orbit a circle “parallel” to  $C_i$  and oriented the same as  $C_i$ . The existence of such a flow is a consequence of the von Neumann, Oxtoby-Ulam theorem (see [OU]) which asserts that up to homeomorphism  $\mu$  on  $A_i$  is equivalent to Lebesgue measure on a standard annulus of the same area. It is then clear from the definition that  $\mathcal{R}_\mu(\phi_t) \in H_1(M, \mathbb{R})$  is equal to  $tk[C_i]$  for some positive constant  $k$ .

Choose  $s > 0$  sufficiently small that  $d(\phi_s(x), x) < \varepsilon/3$  for all  $x \in M$ . Note that  $\mathcal{R}_\mu(\phi_s) = sk[C_i]$ , and so by Proposition (2.4) if  $g_i = \phi_s \circ f$ , then  $\mathcal{R}_\mu(g_i) = sk[C_i] = a_i[C_i]$  where  $a_i = sk$ . Also

$$(1) \quad d(g_i(x), f(x)) = d(\phi_s(f(x)), f(x)) < \varepsilon/3.$$

In other words any orbit segment for  $g_i$  is an  $\varepsilon$ -chain for  $f$ .

As mentioned in §2 the Birkhoff ergodic theorem implies that  $\mathcal{R}(x, g_i)$  exists for  $\mu$  almost all  $x \in M$  and

$$\int \mathcal{R}(x, g_i) d\mu = a_i[C_i].$$

This integral (like any integral) can be approximated by a weighted average of values of the integrand. That is, given any  $\delta > 0$  there exist  $x_1, x_2, \dots, x_m$  in the interior of  $M$  and positive constants  $b_j$  such that

$$\left\| a_i[C_i] - \sum_{j=1}^m b_j \mathcal{R}(x_j, g_i) \right\| < \frac{\delta}{a_1},$$

and hence

$$\left\| [C_i] - \sum_{j=1}^m \frac{b_j}{a_i} cR(x_j, g_i) \right\| < \delta.$$

Moreover, we can assume that each  $x_i$  is in fact recurrent under  $g_i$  since by the Poincaré recurrence theorem this is true for a set of full measure in  $M$ .

Clearly we can do the same construction for the homology class  $-[C_i]$  and find a diffeomorphism  $g'_i$  and points  $x'_j$  which are recurrent and for which a homology class within  $\delta$  of  $-[C_i]$  is a positive linear combination of the set  $\{\mathcal{R}(x'_j, g'_i)\}$ .

If  $\delta$  is sufficiently small then 0 will be in the convex hull of any set of  $\delta$  approximations to  $\{[C_i], -[C_i]\}$ . Hence it follows we have found a finite set of diffeomorphisms  $\{g_i\} \cup \{g'_i\}$  and a finite set of points for each of them such that 0 is in the interior of the convex hull of all the vectors of the form  $\mathcal{R}(x_j, g_i)$  or  $\mathcal{R}(x_j, g'_i)$ .

Notice that if it were the case that  $g_i = g'_i = f$  for all  $i$  then the hypothesis of Theorem (4.1) would be satisfied and our result would follow. While  $g_i$  and  $g'_i$  are not equal to  $f$  they are close enough for the proof to work. In Theorem (4.1) the finite set of recurrent points with the property that 0 is in the interior of their convex hull, was used in only one way: an arbitrarily long orbit segment was used as part of a periodic  $\varepsilon$ -chain which corresponded to a loop whose homology class is nearly a multiple of the rotation vector of the point. In our current setting we can use the orbit of the point under a lift of  $g_i$  or  $g'_i$  in the same way. Equation (1) above implies that this orbit for a perturbation of  $f$  is an  $\varepsilon$ -chain for  $f$ . The remainder of the proof is identical to the proof of (4.1). ■

There is a single theorem which generalizes both (4.1) and (5.1) and which will be of use for applications considered in §6.

**(5.2) Theorem.** *Suppose  $M$  is an oriented compact surface with negative Euler characteristic and  $f : M \rightarrow M$  is a diffeomorphism isotopic to the identity which preserves the smooth measure  $\mu$ . Suppose in addition that for every positive  $\varepsilon$ , 0 is in the interior (in  $H_1(M, \mathbb{R})$ ) of the convex hull of the set  $\mathcal{R}(f) \cup B_\varepsilon(\mathcal{R}_\mu(f))$ , the rotation set of  $f$  together with the ball of radius  $\varepsilon$  centered at  $\mathcal{R}_\mu(f)$ . Then  $f$  has an interior fixed point of positive index. In fact the canonical lift of  $f$  to its universal cover  $F : \widetilde{M} \rightarrow \widetilde{M}$  has an interior fixed point which is of positive index.*

*Proof.* In the proof of (5.1), for any  $\varepsilon > 0$  we constructed a finite set of diffeomorphisms  $\{g_i\}$ , which are  $\varepsilon$  close to  $f$  in the  $C^0$  topology and a finite set of points for each of them such that 0 is in the interior of the convex hull of all the vectors of the form  $\mathcal{R}(x_j, g_i)$ . The same construction can be used without the hypothesis that  $\mathcal{R}_\mu(f) = 0$  to produce a finite set of diffeomorphisms  $\{g_i\}$ , which are  $\varepsilon$  close to  $f$  in the  $C^0$  topology and a finite set of points for each of them such that  $\mathcal{R}_\mu$  is in the interior of the convex hull of all the vectors of the form  $\mathcal{R}(x_j, g_i)$ .

As in (5.1) the orbits of the points  $x_j$  under  $g_i$  are  $\varepsilon$ -chains for  $f$ . Clearly our hypothesis guarantees that 0 is in the interior of the convex hull of  $\mathcal{R}(f) \cup (\bigcup \mathcal{R}(x_j, g_i))$ . Thus once again the remainder of the proof is identical to the proof of (4.1). ■

## 6. APPLICATIONS

A heuristic guiding the study of area preserving diffeomorphisms, especially exact symplectic diffeomorphisms, is that they should have properties like the time one map of the flow associated with an autonomous Hamiltonian vector field. In particular we might hope to find estimates on the number and nature of fixed points

based on the analogous results for critical points of smooth functions (the Hamiltonian function). The following result, sometimes called the Arnold conjecture, was proved by A. Floer [Flo] and J.-C. Sikorav [S].

**(6.1) Theorem.** *Suppose  $M$  is a compact oriented surface with non-positive Euler characteristic and without boundary and suppose  $f : M \rightarrow M$  is a diffeomorphism isotopic to the identity which preserves the smooth measure  $\mu$  and has  $\mathcal{R}_\mu(f) = 0$ . Then  $f$  has at least three distinct fixed points. If  $f$  has finitely many fixed points then at least two of them have positive index. If all fixed points of  $f$  have index equal to  $\pm 1$  then  $f$  has at least  $2g + 2$  fixed points, where  $g$  is the genus of  $M$ .*

*Proof.* Clearly we need only consider the case when  $f$  has finitely many fixed points. If the Euler characteristic of  $M$  is negative by Theorem (5.1) above there is a fixed point  $z$  of positive index with a lift  $\hat{z} \in \widetilde{M}$  which is fixed under the canonical lift  $F : \widetilde{M} \rightarrow \widetilde{M}$  of  $f$  to its universal covering space. This implies that  $f$  is isotopic to the identity relative to  $z$ .

If  $M = \mathbb{T}^2$  there is also a fixed point  $z$  of positive index by a result of M. Flucher [Flu] and there is a choice of lift  $F$  which fixes lifts of  $z$ . Moreover  $F$  is a lift of an isotopy from  $f$  to the identity. So in this case also  $f$  is isotopic to the identity relative to  $z$ .

Let  $N$  denote  $M$  with the fixed point  $z$  blown up. Then there is a natural identification of  $H_1(N, \mathbb{R})$  with  $H_1(M, \mathbb{R})$  induced by the obvious map from  $N$  to  $M$ . Clearly, if  $f_0 : N \rightarrow N$  is the blown up version of  $f$  then for any  $x \in \text{int}(N)$  for which  $\mathcal{R}(x, f)$  is defined we have  $\mathcal{R}(x, f) = \mathcal{R}(x, f_0)$ . Thus  $\mathcal{R}_\mu(f_0) = 0$  and  $f_0$  satisfies the hypothesis of (5.1). It follows that  $f_0$  has an interior fixed point of positive index and hence  $f$  has two fixed points of positive index. Since the Euler characteristic of  $M$  is less than or equal to zero, there must also be a fixed point of  $f$  of negative index, making at least three fixed points.

If all fixed points of  $f$  have index equal to  $\pm 1$  then  $f$  has at least two fixed points of index one and the sum of the indices of all fixed points is the Euler characteristic  $2 - 2g$ . Thus  $f$  must have at least  $2g$  fixed points of index  $-1$  making a total of at least  $2g + 2$  fixed points. ■

We can also obtain a result which might be called an Arnold conjecture for diffeomorphisms of surfaces with boundary.

**(6.2) Theorem.** *Suppose  $M$  is a compact oriented surface with negative Euler characteristic and non-empty boundary and  $f : M \rightarrow M$  is a diffeomorphism isotopic to the identity which preserves the smooth measure  $\mu$ . Let  $\{v_i\}$  be the set of rotation vectors of points in the boundary of  $M$ . If there are numbers  $c_i \geq 0$  such that  $-\mathcal{R}_\mu(f) = \sum c_i v_i$  then  $f$  has an interior fixed point. If the interior fixed points of  $f$  are isolated then one of them has positive index. In fact, in this case, the canonical lift of  $f$  to its universal cover  $F : \widetilde{M} \rightarrow \widetilde{M}$  has an interior fixed point which is of positive index.*

*Proof.* The case that  $\mathcal{R}_\mu(f) = 0$  is handled by (5.1). Otherwise, discarding any  $v_i$  for which  $c_i = 0$ , we have

$$\mathcal{R}_\mu(f) + \sum_{i=1}^n c_i v_i = 0,$$

with all  $c_i > 0$ . It follows that for any  $\varepsilon > 0$  the interior of the convex hull of  $B_\varepsilon(\mathcal{R}_\mu(f)) \cup \{v_i\}$  contains 0 and we can apply (5.2). ■

Another way in which the heuristic mentioned above guides us is by suggesting that if a diffeomorphism of a surface has some number of fixed points which are like minima for a Hamiltonian function, then there must be another fixed point which is like a maximum. In some circumstances this is the case. Suppose we are given an area preserving diffeomorphism  $f : M \rightarrow M$  and a fixed point  $p$  such that  $f$  is homotopic to the identity on  $M \setminus \{p\}$ . If the Euler characteristic of  $M \setminus \{p\}$  is negative then and we “blow up”  $p$  replacing it with a boundary component  $C_p$ , then the homological rotation number is defined for each point in the boundary of the resulting surface.

In particular if this rotation vector is not zero, we have a well defined notion of “which way”  $C_p$  is rotating compared with the orientation that boundary component inherits from  $M$ . We will say that a fixed point  $p$  is of *minimum type* (respectively *maximum type*) provided  $\mathcal{R}(z, f)$  is a positive (respectively negative) multiple of  $[C_p]$  for  $z \in C_p$  where the orientation of  $C_p$  is that induced from an orientation of  $M$ . There are two weaknesses of this definition. First a fixed point of saddle type whose unstable manifold has its orientation reversed has index +1 and hence will be either of maximum or minimum type. Secondly if the derivative  $df_p$  of  $f$  at a fixed point  $p$  is the identity then that point has neither maximum nor minimum type. Other than this, any fixed point of positive index will have either maximum or minimum type.

**(6.3) Theorem.** *Suppose  $f : M \rightarrow M$  is an area preserving diffeomorphism of a closed oriented surface  $M \neq S^2$ . Suppose  $f$  is isotopic to the identity and satisfies  $\mathcal{R}_\mu(f) = 0$ . If  $f$  has a finite fixed point set then there is a set  $P$  of fixed points each of which is of positive index with the property that  $f$  is isotopic to the identity relative to  $P$  and not all the points of  $P$  are of the same type (maximum or minimum). If for every fixed point  $p$ ,  $df_p \neq \text{id}$ , then  $P$  contains at least one point of maximum type and at least one of minimum type.*

*Proof.* There is one fixed point of positive index by Theorem (5.1). Without loss of generality suppose it is of minimum type. If we blow up this fixed point we obtain a diffeomorphism isotopic to the identity on a surface  $M_1$  of negative Euler characteristic. It will still have  $\mathcal{R}_\mu(f) = 0$  so we can apply (5.1) again to find another fixed point of positive index. Inductively suppose we have constructed  $f_n : M_n \rightarrow M_n$  by blowing up  $n$  fixed points in this way. We will show that either we find a fixed point with positive index which is not of minimum type or we can blow up the fixed point of minimum type and continue the process. Since there are finitely many fixed points this means we must eventually find a positive index fixed point which is not of minimum type.

Actually the desired induction hypothesis is that  $f : M_n \rightarrow M_n$  is isotopic to the identity and that there are constants  $c_i$  such that  $-\mathcal{R}_\mu(f) = \sum c_i v_i$ , where  $v_i = \mathcal{R}(x_i, f)$  for  $x_i$  in the  $i$ -th boundary component of  $M_n$ , so we can apply Theorem (6.2). This clearly holds for  $n = 1$ . Assuming it holds for  $n$  we observe that (6.2) implies the existence of a fixed point  $z$  of positive index. If  $z$  is of minimal type we blow it up to obtain  $f_{n+1} : M_{n+1} \rightarrow M_{n+1}$  which is still isotopic to the identity. This is because  $z$  is the projection of a fixed point of the canonical lift  $F_n : \tilde{M}_n \rightarrow \tilde{M}_n$  and hence  $f_n$  is isotopic to the identity relative to  $z$ .



Let  $C_i$ ,  $1 \leq i \leq n$ , be the boundary components of  $M_n$  stage oriented consistently with the orientation of  $M_n$ . All that remains to be shown is that

$$(1) \quad -\mathcal{R}_\mu(f_{n+1}) = \sum_{i=1}^{n+1} a_i [C_i],$$

where  $[C_i]$  is the homology class of  $C_i$  and  $a_i \geq 0$ .

Let  $p : M_{n+1} \rightarrow M_n$  be the map that collapses  $C_{n+1}$  to the point  $z$  which was blown up to form it. It is then straightforward to see that  $p_*(\mathcal{R}_\mu(f_{n+1})) = \mathcal{R}_\mu(f_n)$  and that the kernel of  $p_*$  is spanned by  $[C_{n+1}]$ . By the induction hypothesis

$$(2) \quad -\mathcal{R}_\mu(f_n) = \sum_{i=1}^n a_i [C_i],$$

for some  $a_i \geq 0$ . Hence  $\mathcal{R}_\mu(f_{n+1}) + \sum_{i=1}^n a_i [C_i]$  is in the kernel of  $p_*$  and thus a multiple of  $[C_{n+1}]$ . It follows that equation (1) above holds where  $a_i$ ,  $1 \leq i \leq n$  are as in equation (2) and  $a_{n+1}$  is some real number which may not be non-negative as desired.

However, the curves  $\{C_i\}$  form the boundary of the surface  $M_{n+1}$  and hence we have

$$(3) \quad 0 = \sum_{i=1}^{n+1} [C_i].$$

Thus if we add  $k$  times equation (3) to equation (2) we obtain

$$(4) \quad -\mathcal{R}_\mu(f_{n+1}) = \sum_{i=1}^{n+1} b_i [C_i],$$

where  $b_i = k + a_i$ . Clearly if  $k$  is sufficiently large,  $b_{n+1} > 0$  and we have shown the induction hypothesis for  $n + 1$ .

Hence there is an additional fixed point of positive index and eventually one which is not of minimum type. If the new fixed point of positive index is generic and not of minimum type it must be of maximum type. There must also be a fixed point of negative index since the sum of the indices is at most 0. ■

There is a similar theorem for  $S^2$ , where no hypothesis on the mean rotation is needed but we require that there be at least three fixed points of positive index.

**(6.4) Theorem.** *Suppose  $f : S^2 \rightarrow S^2$  is an area preserving diffeomorphism isotopic to the identity with a finite fixed point set. If  $f$  has at least three fixed points of positive index then there is a set  $P$  of fixed points each of which is of positive index with the property that  $f$  is isotopic to the identity relative to  $P$  and not all the points of  $P$  are of the same type (maximum or minimum). If for every fixed point  $p$ ,  $df_p \neq \text{id}$ , then  $P$  contains at least one point of maximum type and at least one of minimum type.*

*Proof.* Pick a set  $P_0$  of three fixed points of positive index. Then  $f$  is isotopic to the identity relative to  $P_0$  since this is true for any diffeomorphism of  $S^2$  isotopic to the identity and any set of three fixed points.

If the points of  $P_0$  are not all of the same type we are done. If they are of the same type we blow them up forming  $f_3 : M_3 \rightarrow M_3$ , a diffeomorphism isotopic to the identity on a surface which is a sphere with three holes and hence has Euler characteristic  $-1$ .

Let  $[C_1], [C_2], [C_3]$  be the homology classes of the boundary components. Then any two of these classes form a basis of  $H_1(M_3, \mathbb{R})$  and  $[C_1] + [C_2] + [C_3] = 0$ . From this it follows that any element of  $H_1(M_3, \mathbb{R})$  can be written as  $a_1[C_1] + a_2[C_2] + a_3[C_3]$  with  $a_i \geq 0$ . In particular

$$-\mathcal{R}_\mu(f_3) = \sum_{i=1}^3 a_i [C_i],$$

for some non-negative  $a_i$ .

It follows that just as in (6.3) we can now repeatedly apply (6.2) to find new fixed points of positive index until we eventually find one which is not of the same type as the points of  $P_0$ . ■

#### REFERENCES

- [BF] M. Barge and J. Franks, *From Topology to Computation: Proceedings of the Smalefest* (M.W. Hirsch, J.E. Marsden, M. Shub, eds.), Springer-Verlag, 1993, pp. 186–195. MR **94m**:54095
- [C] C. Conley, *Isolated Invariant Sets and the Morse index*, vol. 38, C.B.M.S. Regional Conference Series in Math. Amer. Math. Soc., Providence, RI, 1978. MR **80c**:58009
- [Flo] A. Floer, *Proof of the Arnold Conjecture for surfaces and generalizations to certain Kähler manifolds*, Duke Math. Jour. **51** (1986), 1–32. MR **87h**:58188
- [Flu] Martin Flucher, *Fixed points of measure preserving torus homeomorphisms*, Manuscripta Mathematica **68** (1990), 271–293. MR **91j**:58129
- [F1] J. Franks, *Recurrence and Fixed Points of Surface Homeomorphisms*, Ergodic Theory and Dyn. Systems **8** (1988), 99–107. MR **90d**:58124
- [F2] J. Franks, *Generalizations of the Poincaré–Birkhoff Theorem*, Annals of Math. **128** (1988), 139–151. MR **89m**:54052
- [F3] J. Franks, *A New Proof of the Brouwer Plane Translation Theorem*, Ergodic Theory and Dynamical Systems **12** (1992), 217–226. MR **93m**:58059
- [F4] J. Franks, *Rotation Numbers for Area Preserving Homeomorphisms of the Open Annulus*, World Science, 1991. MR **93e**:58153
- [F5] J. Franks, *Geodesics on  $S^2$  and Periodic Points of Annulus Homeomorphisms*, Inventiones Math. **108** (1992), 403–418. MR **93f**:58192
- [CMY] S.-N. Chow, J. M. Mallet-Paret, and J. A. Yorke, *A periodic orbit index which is a bifurcation invariant*, vol. 1007, Springer Lecture Notes in Math., 1983, pp. 109–131. MR **85d**:58058
- [G] Matthew Grayson, *Shortening embedded curves*, Annals of Math. **129** (1989), 71–111. MR **90a**:53050
- [H1] Michael Handel, *A Fixed Point Theorem for Planar Homeomorphisms*, Preprint.
- [H2] Michael Handel, *Zero Entropy Surface Diffeomorphisms*, Preprint.
- [HDK] H. Hadwiger, H. Debrunner, and V. Klee, *Combinatorial Geometry in the Plane*, Holt Rinehart and Winston, New York, 1964. MR **29**:2363
- [M] John Milnor, *Lectures on the h-cobordism theorem*, Princeton Univ. Press, 1965. MR **32**:8352
- [OU] J. Oxtoby and S. Ulam, *Measure preserving homeomorphisms and metrical transitivity*, Annals of Mathematics **42** (1941), 874–920.
- [S] J.-C. Sikorav, *Points fixes d’une application symplectique homologue à l’identité*, Jour. Diff. Geom. **22** (1985), 49–79. MR **88g**:58066

DEPARTMENT OF MATHEMATICS, NORTHWESTERN UNIVERSITY, 2033 SHERIDAN ROAD, EVANSTON, ILLINOIS 60208-2730

E-mail address: john@math.nwu.edu