Pergamon

Topology, Vol. 33, No. 1, pp. 25–43, 1994

0040-9383/94 $6.00 + 0.00

SHORTENING CURVES ON SURFACES

JOEL HASS† and PETER SCOTT‡

(*Received* 15 *August* 1992; *in revised form* 24 *April* 1993)

§0. INTRODUCTION AND PRELIMINARY RESULTS

METHODS of shortening a curve in a manifold have been used to establish the existence of closed geodesics, and in particular of simple closed geodesics on 2-spheres. For this purpose, a curve evolution process should (a) not increase the number of self-intersections of a curve, (b) exist for all time or until a curve collapses to a point, (c) shorten curves sufficiently fast so that curves which exist for all time converge to a geodesic, and (d) depend continuously on the choice of initial curve. Birkhoff originated what is now known as the Birkhoff curve shortening process, where midpoints of polygonal approximations to a curve are successively connected by geodesic segments [4]. This type of shortening has the advantage that (b), (c) and (d) are easy to establish, but the disadvantage that (a) seems difficult to arrange. A process of evolving a curve on a surface by its curvature is perhaps the most natural flow. Short term existence is easy to establish for this flow, but long term existence involves deep questions in PDEs and geometry. This flow has recently been studied with considerable success in a series of papers [7, 8, 9, 1]. All four of the desired properties have been shown to hold for the flow by curvature of an embedded curve on a Riemannian surface. For non-embedded curves in Riemannian surfaces, some open questions remain about the types of singularities which may develop in the curvature flow. In particular, it is not known whether arcs of double points can be created.

In this paper we introduce a new curve shortening flow. Like the Birkhoff process, this flow involves replacing arcs of a curve with geodesic segments. Unlike the Birkhoff process, it picks out its piecewise-geodesic structure purely from the geometry of the image manifold rather than from a parametrization of the curve. This flow, which we call the *disk flow*, is developed in §1.

In §2 we use the flow to solve a purely topological problem concerning intersections of curves on surfaces. Turaev [17] has posed the problem in the following form:

Question. Let s_0 and s_1 be homotopic curves on a surface, each with k double points. Is there a homotopy s_t from s_0 to s_1 with the property that each curve s_t in the homotopy has at most k double points?

The answer is yes, as we show in Theorems 2.1 and 2.2.

As with the curvature flow, we can use the disk flow to study the evolution of families of curves. This is carried out in §3, where we give a new proof of the theorem of Lusternik and Schnirelman establishing the existence of three simple closed geodesics on any 2-sphere. In

†Partially supported by NSF grant DMS9024796 and the Alfred P. Sloan Foundation.
‡Partially supported by NSF grant DMS9003974.

§4 we obtain new results on the existence of simple geodesic arcs on a disk with a convex boundary. Finally in §5 we make some concluding observations. We consider only orientable surfaces, though most results extend to the non-orientable case.

§1. A CURVE SHORTENING FLOW

We give a simple construction of a curve flow that takes a finite collection of curves on a surface and homotops so that:

1. The number of self-intersection points of each curve is non-increasing.
2. The number of intersection points between each pair of curves is non-increasing.
3. Either a curve disappears in a finite time or it eventually lies arbitrarily close to a geodesic.
4. The flow extends continuously over k-parameter families of curves.

This flow seems to have all the benefits of the curvature flow for geometrical applications, but has the advantage of being easy to construct and understand, particularly for singular curves. It also is easy to implement algorithmically, and is well suited for computer modeling of curve flows. Moreover it generalizes to higher dimensions in interesting ways. A precise definition of the flow will be given later, but we first give a rough description.

Let γ be a piecewise-smooth immersed curve on a Riemannian surface F. Cover F with convex disks $D_1, D_2, \ldots D_n$ of radius smaller than the injectivity radius. We choose the disks in general position, so that any point on F meets the boundary of at most two disks, the boundaries of the disks meet γ transversely, and so that disks of half the radius with the same centers still cover F. Such a cover will be called well-positioned relative to γ. Define D_{n+i} cyclically, so that $D_{n+i} = D_i$, $i \geq 1$. Roughly, the disk flow is defined by homotoping each arc of $\gamma \cap D_1$ to the unique geodesic arc with the same endpoints, then repeating for each of $D_2, D_3 \ldots$.

We will explore the properties and convergence of this flow, and show that the number of intersection points is non-increasing.

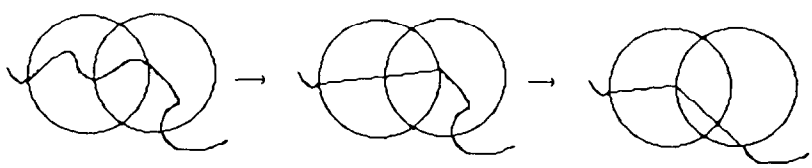

Fig. 1. Two iterations in the construction of the disk flow.

We first prove a few combinatorial lemmas. We say that an immersed curve s on a surface F, contains an embedded 1-gon if there is an embedded subarc a of S^1 with the two endpoints a^+ and a^- of a identified by s, the resulting loop is embedded and bounds a disk D on F. We say that an immersed curve s on a surface F, contains an embedded 2-gon if there is a pair of disjoint embedded sub-arcs a and b of S^1 with $s(a^+) = s(b^+)$, $s(a^-) = s(b^-)$, the loop $s(a) \cup s(b)$ is embedded and bounds a disk D on F. See [11] for a discussion of the existence of such configurations. We say that an embedded 1-gon or 2-gon is innermost if it does not properly contain either an embedded 1-gon or 2-gon.

LEMMA 1.2. Given a triangle with embedded arcs crossing it, any two of which intersect in at most one point and all of which miss one edge, each of the other two edges has an innermost triangle adjacent to it.

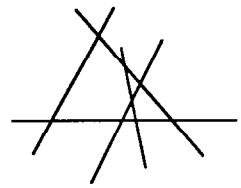

Fig. 2.

Proof. The proof is by induction on the number of line segments n. It is trivial for $n = 1$. Assume the result for $n < k$. Start at the vertex opposite the missed edge e_0 and go along one of the two other edges e_1 until arriving at the first line. This defines a subtriangle, not necessarily innermost, which has no lines across its edge which lies in e_1. By induction this gives an innermost triangle on the opposite edge e_2 since the subtriangle has fewer than k lines crossing it. Repeat the construction with the edge e_2 to get a triangle meeting the edge e_1. General position is not required.

LEMMA 1.4. Given an innermost 2-gon with embedded arcs crossing it, there is an innermost triangle adjacent to each edge of the 2-gon.

Fig. 3.

Proof. Since the 2-gon is innermost, any two arcs inside it meet in at most one point. Consider the triangle formed by moving in from a vertex of the 2-gon along one of the edges to the first arc encountered. Combinatorially, this triangle meets the hypothesis of Lemma 1.2, and so has an innermost triangle on each of its other two edges, and in particular along an edge of the 2-gon.

We call the process of sliding one edge of an innermost triangle across the vertex formed by the other two edges a triangle move. We call the process of replacing an arc of a curve by a homotopic shortest geodesic arc a straightening of the curve. To simplify the statements of the following results we will adopt the convention that a point curve and the empty curve are homotopic simple curves, and that each of these is a trivial example of a geodesic.

LEMMA 1.6. A finite collection of piecewise-smooth, transversely intersecting curves in a convex disk can be homotoped (rel boundary) to a collection of geodesics so that the number of self-intersection points of each curve and the number of intersection points between each pair of curves is non-increasing during the homotopy.

Proof. We induct on the number of excess double points k in the relative homotopy classes of the arcs.

Suppose first that $k = 0$, so that each curve is embedded and any two curves which intersect meet in one point only. If there is a closed curve, find one not containing any other closed curve in its interior. This curve can be homotoped to a point and thus to the empty

curve, introducing no new intersections. So we can assume there are no closed curves. Let $\{a_i\}$ be the collection of arcs, let a_1 be the first arc and let d_1 be the geodesic segment connecting the endpoints of a_1. Suppose some arc a_j meets d_1 in 2 points, forming a 2-gon with a sub-arc of d_1. Find a 2-gon T with one side a sub-arc of d_1 which is innermost among such 2-gons. Then T is innermost among all 2-gons, since no pair of arcs meets twice, and each arc a_i crosses T at most once. Applying Lemma 1.4 we can do triangle moves until the interior of T meets no arc a_i, and then eliminate T by sliding a_j across d_1. Since a triangle always exists meeting the a_j edge of T, we need only move a_j at each stage. Continuing, we eliminate all 2-gons between any arc a_i and $d_1, i > 1$ and have $a_1 \cap d_1 = \partial a_1$. Then d_1 and a_1 cobound a 2-gon and proceeding as before, a_1 and d_1 can be made to coincide by moving a_1 without introducing new double points. We now repeat with a_2 and its corresponding geodesic d_2. Since d_1 meets d_2 in at most one point, it is never moved as we straighten a_2. Continuing we can straighten all the arcs without introducing any new double points.

Suppose now that $k > 0$, and that the lemma holds for configurations with fewer than k intersection points. Then there is a non-embedded curve or two curves which intersect in more than one point. It follows that there is an embedded 1-gon or an embedded 2-gon which is innermost among embedded 1-gons and 2-gons [11]. An innermost embedded 1-gon has nothing inside it at all and can be removed, completing the lemma by induction. An innermost embedded 2-gon which has some arcs crossing its interior has inside it only embedded arcs going from one edge to the other. By Lemma 1.4 an innermost triangle can be found. We can do triangle moves across the edges of the 2-gon until no triangles exist in the 2-gon, and it can then be removed with the elimination of two double points in the disk by sliding one of its boundary arcs across the other. Lemma 1.6 now follows by induction. Note that the arcs are moved by a regular homotopy except when a local 1-gon is shrunk to a point and eliminated, as in Fig. 4 and that the resulting straightened arcs intersect transversely.

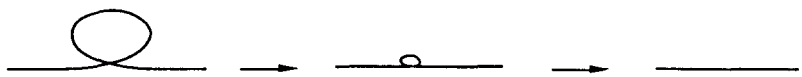

Fig. 4.

We now give the definition of the disk flow. Consider a curve γ_0, not necessarily connected or embedded. Assume that no two arcs of $\gamma_t \cap D_i$ have a common boundary point and that γ_{i-1} is not completely contained in D_i. We define γ_t for $t \in [i-1, i]$, $i = 1, 2, \ldots$ to be the result of performing the continuous straightening process of Lemma 1.6 in the disk D_i on $\gamma_{i-1} \cap D_i$. Although this homotopy may not decrease lengths, the lengths at integral time periods are non-increasing. If γ_{i-1} is completely contained in D_i then the homotopy shrinks it to a point, and then the empty curve.

If γ_0 has transverse self-intersection, small shrinkings of D_i will always suffice to ensure that no two arcs have a common endpoint, since straightened arcs also intersect transversely. If two arcs of $\gamma_j \cap D_i$ have a common boundary point, we first shrink D_i by a factor of λ, where $1-\frac{1}{4^i} < \lambda < 1$, so that there is no common boundary point, and then perform the homotopy of Lemma 1.6. Note that no disk ever shrinks to less than half its original radius, since $\prod_{i=1}^{\infty}(1-\frac{1}{4^i}) > \frac{1}{2}$, so that the disks still cover F. This shrinking process is necessary in order to avoid a situation where two arcs have common boundary points, and their straightenings coincide along an arc, or meet at one point, without crossing. If γ_0 is

embedded, or if we count only transverse intersection points, then we do not need to shrink disks. The disk flow defines a regular homotopy of a curve, except for a finite number of times when a small loop in the curve disappears, as in Fig. 4.

The disk flow is not canonical, since the choice of homotopy in Lemma 1.6 is not uniquely defined. However it has some useful properties, which we introduce some new notation to describe. Define the map Δ on the set of curves on F to be the map obtained by taking a curve $s = s_0$ to the curve s_n obtained by straightening s successively in each of $D_1, \ldots, D_n$. We use the topology on the space of curves on a manifold in which an ε-neighborhood of a curve γ consists of all curves γ' such that γ and γ' admit parametrizations which are ε-close in the C^0-topology. This is the topology induced by the Frechet metric on the space of curves.

THEOREM 1.8. *Let γ, γ' be transversely self-intersecting curves in F. Let $D_1, \ldots, D_n$ be a collection of disks covering F, well positioned relative to $\gamma \cup \gamma'$. Let γ_t, γ_t' be their images under the disk flow.*

(1) *The number of self-intersection points of γ_t is non-increasing with t, $t \in [0, \infty)$.*

(2) *The number of intersection points between γ_t and γ_t' is non-increasing with t, $t \in [0, \infty)$.*

(3) *Either γ_t disappears in finite time, or a subsequence of the curves $\{\gamma_t\}$ converges to a geodesic as $t \to \infty$. In the second case, if U is any open neighborhood of the set of geodesics homotopic to γ_t then there is a $T > 0$ such that γ_t lies in U for $t > T$.*

(4) *If a sequence $\{\gamma_i\}$ converges to a geodesic γ_∞ as $i \to \infty$ then length$(\gamma_i) \to$ length(γ_∞).*

(5) *Length$(\Delta(\gamma)) \leq$ length(γ), with equality if and only if γ is a geodesic or a point.*

Proof. The first and second assertions are a direct consequence of Lemma 1.6. Suppose now that γ_t persists for $t \to \infty$ so that length(γ_t) is bounded from below. The sequence of curves γ_i, $i = 1, 2, \ldots$ is then a sequence of rectifiable curves of non-increasing length on a compact manifold. It follows from Ascoli's Theorem that there is a subsequence $\{\gamma_j\}$ which converges uniformly to a rectifiable curve δ, with length$(\delta) \leq \lim_{j \to \infty}$ length(γ_j).

Claim 1.9. If a rectifiable curve δ is not a geodesic, we can find a subarc δ' of δ, such that either

a) δ' is not a geodesic and $\delta' \cap \partial D_i = \phi$ for all i, or

b) δ' consists of a pair of geodesic segments meeting at an angle at a point q lying on the boundary of some disk D_i.

Proof of Claim 1.9. If δ is not a geodesic it contains subcurves of arbitrarily short length which are not geodesic. If none of these miss the boundaries of the disks D_i, then the arcs of δ not meeting the boundaries must be geodesic segments. In this case either δ is globally a geodesic or (b) holds, proving the claim.

We now prove assertion (3). Assume first that case (a) holds. Then there is an $\varepsilon > 0$ such that straightening an arc of δ containing δ' decreases length(δ) by at least ε. Note that straightening in a disk D_j not containing δ' does not move δ'. Since $\gamma_j \to \delta$, there is a J such that if $j > J$ and D_i contains δ', then replacing $\gamma_j \cap D_i$ by geodesic arcs decreases length(γ_j) by at least $\varepsilon/2$. It follows that length$(\Delta(\gamma_j)) <$ length$(\gamma_j) - \varepsilon/2$. For j sufficiently large, length(γ_j)–length$(\delta) < \varepsilon/2$. This implies length$(\Delta(\gamma_j)) <$ length(δ), a contradiction. So any convergent subsequence of $\{\gamma_i\}$ converges to a geodesic. Suppose now that there is a subsequence $\{\gamma_k\}$ for which no γ_k lies within a neighborhood U of the geodesics homotopic to γ. We could then pass to a convergent subsequence $\{\gamma_k\}$ with the same

property, and the limit would not be a geodesic, a contradiction. This establishes assertion 3 for this case.

Assume now that case (b) of the claim holds. Straightening δ' in the disks whose boundary contain q may now shorten δ while not shortening approximating arcs. However q lies in the interior of a disk D_k, and straightening in D_k uniformly shortens the approximating arcs, again leading to a contradiction. We have now established assertion 3.

Similarly, if a subsequence $\{\gamma_j\}$ converges to a geodesic γ_∞ as $j \to \infty$, and $\text{length}(\gamma_\infty) < \lim_{j\to\infty} \text{length}(\gamma_j)$, then we could find a subarc δ of γ_∞ and approximating arcs with the same property. The approximating arcs can be taken to either miss the boundaries of the disks or to meet a boundary at an angle. In either case assertion (4) of the theorem follows as above. Assertion (5) also follows from the same argument, proving Theorem 1.8.

Note that while a subsequence of $\{\gamma_i\}$ converges to a geodesic, the entire sequence may possibly oscillate between different geodesics. This possibility is also encountered with other curve flows, and it is not known whether it actually occurs. When F has a metric of negative or zero curvature, we can show that γ_t converges to a unique geodesic. The same methods prove that convergence to a single geodesic holds in a generic metric. We say that two geodesics are *parallel* if they cobound a flat annulus.

THEOREM 1.10. *Let F be a negatively curved or flat closed surface. The disk flow applied to a curve γ_0 gives a homotopy γ_t under which γ_t either disappears in finite time or γ_t converges to a unique geodesic γ_∞ as $t \to \infty$.*

Case 1. *F has negative curvature.* We can lift γ to the cover F_γ corresponding to γ, which is topologically a cylinder. We will show that $\gamma_t \to g$ where g is the unique simple closed geodesic in F homotopic to γ_t. We call the lift of a curve to F_γ by the same names as the curve. One possible approach to this case is to explicitly estimate how long two curves which are equidistant from g and surround γ take to converge to g. We observe instead that Theorem 1.8 implies that there is a subsequence $\{\gamma_j\}$ which converges to the unique geodesic γ_∞ homotopic to γ. Once a curve is sufficiently close to γ_∞ under the disk flow, it always stays near γ_∞ and converges to γ_∞ as $t \to \infty$.

Case 2. *F is flat.* Lift to the cover corresponding to $[\gamma]$, which in this case is a flat cylinder. Call the smallest distance between two parallel geodesics containing γ_t the width of γ_t. Since $\{\gamma_j\}$ is converging to a geodesic its width is decreasing to zero. The disk flow on $[i, i+1]$ takes γ_i to a curve γ_{i+1} of no larger width, so the entire sequence has width converging to zero. Moreover if γ_i lies between two geodesics γ_l and γ_r then applying the disk flow to the union of these three curves shows that γ_t also lies between γ_l and γ_r for $t >$ i. Note that we can do the straightening process of the disk flow so that γ_t is straightened before γ_l and γ_r. The flow of γ_t is then unchanged by the addition of the extra two curves. It follows that the entire sequence is converging to a unique geodesic.

We now state a more general result.

THEOREM 1.11. *Let F be a Riemannian 2-manifold. The disk flow applied to a curve γ_0 gives a homotopy γ_t under which γ_t either disappears in finite time or γ_t converges to a unique geodesic γ_∞ as $t \to \infty$ unless there are an infinite number of distinct, non-parallel geodesics with uniformly bounded length in the homotopy class of γ_0.*

Proof. The previous arguments prove convergence in the absence of such a sequence of geodesics.

Finally we state a result, originally proved by Ballman [2] (see [6] for another proof) which is an immediate consequence of Theorem 1.8.

COROLLARY 1.12. *Let F be a closed surface with a Riemannian metric. Every essential simple closed curve is isotopic to a simple geodesic.*

§2. DOUBLE POINTS AND HOMOTOPIES OF CURVES

In this section we apply the disk flow to solve the topological problem of Turaev mentioned in the introduction. By a curve in a surface we mean an immersed 1-dimensional submanifold, and we identify two curves with the same image, so that we are not concerned with curve orientations. In counting double points it is required that the curve s_t be self-transverse, though not necessarily in general position. This property is satisfied by curves in a generic homotopy. A k-tuple point is then counted with multiplicity, counting as $k(k-1)/2$ double points. The question is answered affirmatively through the following two results.

THEOREM 2.1. *Let s_0 and s_1 be homotopic curves in general position on a surface, each minimizing the number of double points in their common homotopy class. Then there is a homotopy s_t from s_0 to s_1 such that s_t is self-transverse for all t and the number of double points of s_t is constant.*

THEOREM 2.2. *Let s_0 be a curve in general position on a surface which does not minimize the number of double points in its homotopy class. Then there is a homotopy s_t from s_0 to a curve s_1 which has minimal self-intersection such that the number of double points of the curve s_t is non-increasing with t. s_t is a regular homotopy except for a finite number of times when a small loop in the curve shrinks to a point.*

COROLLARY 2.3. *Let s_0 and s_1 be homotopic curves on a surface, each with k double points. There is a homotopy s_t from s_0 to s_1 with the property that each curve s_t has at most k double points.*

Proof of Corollary 2.3. Applying Theorem 2.2 to s_0 we can homotop it to a curve s_0' which has minimal self-intersection without increasing the number of double points. Similarly we can homotop s_1 to a curve s_1' which has minimal self-intersection without increasing the number of double points. Theorem 2.1 implies that s_0' and s_1' can be homotoped to one another without increasing the number of double points. Combining the three homotopies proves the corollary.

Example 2.4. The corresponding result for non-connected curves is false, as illustrated by the following pair of homotopic two component curves.

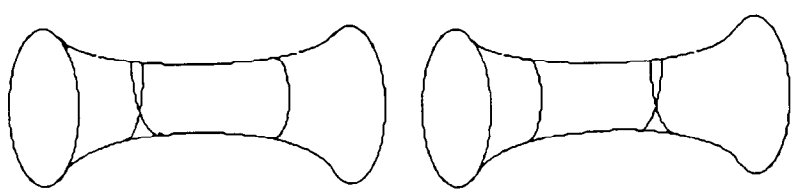

Fig. 5.

The proof of Theorems 2.1 and 2.2 will show however that the obstruction to generalizing the result to non-connected curves is completely illustrated in this example. If no pair of components are powers of parallel curves then the corresponding result still holds.

To simplify the notation, we let $d(f)$ denote the number of self-intersections of a self-transverse map $f\colon S^1 \to F$ and let $d(f, g)$ denote the number of intersections of two transverse maps $f\colon S^1 \to F$ and $g\colon S^1 \to F$. For α and β free homotopy classes of curves on F, let $d(\alpha) = \inf\{d(f)\colon f \in \alpha\}$ and let $d(\alpha, \beta) = \inf\{d(f, g)\colon f \in \alpha,\ g \in \beta\}$.

The following lemma handles an important special case.

LEMMA 2.6. *If a curve a_0 contains an embedded 1-gon or 2-gon then a_0 can be homotoped to a curve a_1 with $d(a_1) < d(a_0)$ by a homotopy a_t with $d(a_t)$ non-increasing.*

Proof. If there exists an embedded 1-gon or 2-gon, we can find an innermost embedded 1-gon or 2-gon. If there is an innermost embedded 1-gon with nothing crossing it, then it can be homotoped away and we are done. If an arc does cross an innermost embedded 1-gon, then this arc must be embedded, and there is an embedded 2-gon inside the 1-gon, contradicting the innermost hypothesis. So we can assume there are no innermost embedded 1-gons. Pick an innermost embedded 2-gon. Using Lemma 1.4, we can do a series of triangle moves to homotop the curve so that nothing crosses the 2-gon. The 2-gon can then be homotoped away to give a new curve a_1 with $d(a)$ reduced by two.

Proof of Theorem 2.1. In what follows we fix a constant curvature metric on F.

Case 1. *s_0 is null homotopic.* Then s_0 is embedded and is isotopic to the empty curve, as is any other embedded null-homotopic curve.

Case 2. *s_0 represents a primitive element α of $\pi_1(F)$.* Let s_2 be a geodesic in the chosen metric on F. It is well known that s_2 has minimal self-intersection, so that $d(s_2) = d(\alpha)$. In fact, it is shown in [6] that a length minimizing primitive curve in any metric minimizes self-intersection. The disk flow applied to s_0 takes it to s_2 without ever introducing a new double point, by Theorem 1.10.

Case 3. *s_0 represents a non-primitive element α of $\pi_1(F)$.* Suppose that s_0 represents the homotopy class $\alpha = \beta^k$ with β primitive. We can apply the argument of case 1 to construct a homotopy s_t to a geodesic s_∞ homotopic to s_0, but s_∞ in this case factors through a cover of a curve r_1, and no longer minimizes its self-intersection. r_1 is a primitive geodesic however, and does minimize its self-intersection. Also, s_t minimizes self-intersection for $0 \leq t < \infty$ and $s_t \to s_\infty$ smoothly as $t \to \infty$. For t close to ∞, s_∞ is the composition of a map $p\colon S^1 \to A$ of the circle into the annulus of degree k and a map $q\colon A \to F$ of the annulus into F, with image a thin regular neighborhood of r. Thus $d(\alpha)$ is greater or equal to $d(p) + kd(r)$. Lemma 1.9 of [11] establishes that $d(p)$ is minimized by a curve p_0 with $(k-1)$ double points, with p_0 unique up to ambient isotopy. Thus $d(\alpha) = k - 1 + kd(r)$. In general there can be many curves realizing this self-intersection, not ambient isotopic to one another in F. We will show that any pair of such curves are related by a regular homotopy through curves of minimal self-intersection, completing the proof of Theorem 2.1. In order to do this we establish in Lemma 2.7 a refinement of Lemma 1.9 of [11]. Lemma 2.7 together with the above argument completes the proof of Theorem 2.1.

Let A be the annulus $S^1 \times I$ with a flat product metric, so that each $S^1 \times \{pt\}$ is a geodesic. Say that a smooth curve is ε-horizontal if the projection of its unit tangent vector

to the I-factor of A has norm less than ε at each point on the curve. A curve is horizontal if it is 0-horizontal in the above sense. ε-vertical and vertical are similarly defined.

LEMMA 2.7. *Let $\varepsilon < 1$ be a small positive constant and let p, p' be general position immersions of degree k of S^1 to the annulus $A = S^1 \times I$ with $(k-1)$ double points. Suppose that p and p' are ε-horizontal. Then there is a regular homotopy from p to p' through ε-horizontal loops each of which has $(k-1)$ double points.*

Proof. In Lemma 1.9 of [11] it is shown that the self-intersection is minimal and that the configuration of the curve is unique up to isotopy in A. An ε-horizontal curve is transverse to each vertical fiber $\{pt\} \times I$ of $S^1 \times I$, since $\varepsilon < 1$. Choose a point q in S^1 so that the vertical fiber through q does not contain a double point of p. Let R denote the rectangle obtained by cutting A open along $q \times I$. As in Lemma 1.6, we can homotop p rel boundary in R to p_1 so that each arc of p_1 is a geodesic, without introducing new double points. Triangle moves can be carried out by sliding any of the three boundary edges of a triangle across the other two edges. When the three edges are each ε-horizontal, sliding the longest edge across the other two can be done keeping it ε-horizontal. Thus the curve remains ε-horizontal at all times. Lemma 1.9 of [11] implies that a curve intersecting R in geodesic segments can be divided into two connected subarcs λ and μ, each of which projects injectively into the I-factor of $S^1 \times I$, with λ strictly increasing and μ strictly decreasing. See Fig. 6 for an illustration of the $k = 4$ case. λ and μ can be straightened to geodesic arcs with the same endpoints as in Lemma 1.6, without introducing any new double points and maintaining the ε-horizontal property. Similarly p' can be regularly homotoped to p_1', which can be divided into two subarcs λ' and μ' with the same property. By composing p_1 and p_1' with a contraction along the I-factor and an appropriate map of the S^1-factor we can regularly homotop p_1 to p_1' so that at each stage each loop is the union of two simple, ε-horizontal geodesic arcs with injective projection to I, and each loop always has exactly $(k-1)$ double points. This proves Lemma 2.7.

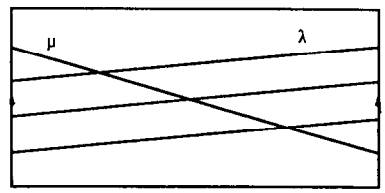

Fig. 6.

LEMMA 2.9. *Let $\varepsilon < 1$ be a positive constant and let p be a general position immersion of degree k of S^1 to the annulus $A = S^1 \times I$. Suppose that p is ε-horizontal. Then there is a regular homotopy from p to an immersion p_1 with exactly $(k-1)$ double points such that p_t is ε-horizontal and the number of double points is non-increasing during the homotopy.*

Proof. If p has $(k-1)$ double points, this is a consequence of Lemma 2.7. Otherwise, Lemma 1.9 of [11] implies that p contains an embedded 1-gon or 2-gon. As p is transverse to the I-fibers of $S^1 \times I$ there cannot be an embedded 1-gon. Embedded 2-gons can be removed as in Lemma 1.4, maintaining the ε-horizontal property.

Proof of Theorem 2.2. Let s_0 be a curve in general position on a surface which does not minimize the number of double points in its homotopy class and let s_1 be a geodesic

homotopic to s_0 if s_0 is essential, or an embedded circle bounding a small disk if s_0 is null-homotopic. We will find a homotopy s_t from s_0 to s_1 such that the number of double points of the curve s_t is non-increasing with t. Note that s_1 has minimal self-intersection if it is primitive. If F is a torus put a flat metric on it, and if genus $(F) > 1$ fix a hyperbolic metric on F.

Case 1. *s_0 is null homotopic.* If F is a 2-sphere and s_0 is not embedded, the result follows from [11], which shows that there exists an embedded 1-gon or 2-gon on F. Otherwise, put a constant curvature metric on F and apply the disk flow to s_0. This gives a homotopy of s_t and Theorem 1.8 implies that s_t disappears after a finite time, since there is no geodesic homotopic to s_0. Thus s_i lies in a disk D_j for i large enough. We can then apply the argument of [11] again to establish the existence of an embedded 1-gon or 2-gon.

Case 2. *s_0 represents a primitive element α of $\pi_1(F)$.* The disk flow takes s_0 to a geodesic s_2, which has minimal self-intersection, without introducing new double points along the way. For the genus 1 case, s_1 is parallel to s_2. In the higher genus case they coincide, since geodesics are unique in a negatively curved surface.

Case 3. *s_0 represents a non-primitive element α of $\pi_1(F)$.* Suppose that s_0 represents the homotopy class $\alpha = \beta^k$ with β primitive. We apply the disk flow to construct a homotopy s_t to a geodesic s_∞ homotopic to s_0, where s_∞ in this case factors through a cover of a curve r_1, and no longer minimizes its self-intersection. r_1 is a primitive geodesic, and does minimize its self-intersection. But s_t introduces no new points of self-intersection for $0 \leq t < \infty$ and $s_t \to s_\infty$ smoothly as $t \to \infty$. Thus for t close to ∞, s_t is the composition of a map $p: S^1 \to A$ of the circle into the annulus of degree k and a map $q: A \to F$ of the annulus into F, with image a thin regular neighborhood of r_1, and $p(S^1)$ is ε-horizontal in A for some small ε. Thus $d(\alpha)$ is greater or equal to $d(p) + kd(r_1)$. Lemma 1.9 of [11] establishes that $d(p)$ is minimized by a curve p_0 with $(k-1)$ double points, with p_0 unique up to ambient isotopy and if p has more than $(k-1)$ double points then there is an embedded 1-gon or 2-gon on the annulus. Since $p(S^1)$ is ε-horizontal, there is no 1-gon and a 2-gon can be eliminated by a homotopy which keeps p ε-horizontal and decreases the number of double points as in Lemma 2.9. Since $d(r_1)$ is minimal already, the resulting curve minimizes the number of double points in its homotopy class. This concludes the proof of Theorem 2.2.

§3. FLOWING FAMILIES OF CURVES

We now consider a version of the flow defined in §1 which will apply to families of curves on a surface, parametrized by some manifold M. We can extend the disk flow to a family of curves parametrized by the unit interval by applying separately to each one the disk flow as defined in §1. It is not now clear that the number of intersection points between two distinct curves γ_t and γ'_t is non-increasing with t, as the straightening process of Lemma 1.6 only applies to finitely many curves. However the number of intersection points between any two curves is non-increasing at integral values of t. Moreover embedded curves stay embedded for all $t > 0$. A new problem arises in case a curve in the family is tangent to the boundary of a disk D_i. We then need to fill in a 'gap' in order to maintain a continuous family of curves.

By a piecewise-smooth k-parameter family of curves in M we mean a piecewise-smooth map from $N^k \times S^1 \to M$ where N is a compact k-dimensional manifold and the map is piecewise smooth on each curve γ_n: $\{n\} \times S^1 \to M$, $n \in N$.

LEMMA 3.1. *Given a piecewise-smooth k-parameter family of curves, the disks $\{D_i\}$ can be perturbed slightly so that no curve has more than k interior tangency points to the boundary of any disk.*

Proof. This lemma can be proved using non-trivial results from the theory of singularities of maps. We use instead an elegant approach due to White [18]. We first consider the case where the family is smooth.

We formulate the problem in the following form: Given a smooth k-parameter family of curves in the plane and a function $h: R^2 \to R$, h can be perturbed slightly so that no curve γ has more than k points of tangency to $h^{-1}(0)$. The lemma follows by letting h be a perturbation of the radial function on each disk in turn.

Let X be the Banach space of $C^{l,\alpha}$ maps from R^2 to R with the $C^{l',\alpha'}$ topology, where $l' > l, \alpha' > \alpha, l + \alpha > 2$. Let $F: R^k \times S^1 \to R^2$ be a fixed k-parameter family of simple curves. Let Z denote the subset of $[S^1]^{k+1}$ consisting of $k+1$ distinct points on S^1. Let Y be the subset of $X \times R^k \times Z$ consisting of points $(h, t_1, t_2 \ldots t_k, \theta_1, \theta_2 \ldots \theta_{k+1})$ such that 0 is a regular value of h, $hF(t_1, t_2 \ldots t_k, \theta_i) = 0$ and $\partial(\boldsymbol{hf})/\partial\boldsymbol{\theta}[t_1, t_2 \ldots t_k, \theta_i] = 0$, $i = 1, 2, \ldots, k+1$. Thus a point is in Y if there are $k+1$ distinct points on the circle whose image under F is tangent to the curve $h^{-1}(0)$ in R^2. Y has codimension $2k+2$ in $X \times R^k \times [S^1]^{k+1}$, and White shows that this implies that its projection to X is closed and nowhere dense. Thus a small perturbation of any map in X will make it miss the projection of Y, implying at most k tangent points of $h^{-1}(0)$ on any circle and proving Lemma 3.1 for the smooth case.

For the piecewise-smooth case, we allow each curve γ_n to flow for a short time under the heat flow to a slightly shorter smooth curve. Short time existence of the heat flow is given by standard theory of parabolic equations. We emphasize here that long term existence of this flow is not standard and requires the extensive arguments of [9, 7, 8]. Since N is compact a short-term length decreasing flow exists for the whole family. The resulting family of curves is smooth, and we can then apply the previous step, concluding Lemma 3.1.

In general the problem of finding a continuous length decreasing deformation of a family of curves which preserves embeddedness is very difficult. However for some special cases it is easy. We say that an arc α in a disk is a pseudo-graph if there is a foliation of the convex hull of α by disjoint geodesic segments, each of which meets α at most once. Note that pseudo-graphs are automatically embedded. Given a curve $\tau_0: [0, 1] \to D$, let τ_t denote the curve obtained from τ_0 by replacing $\tau[0, t]$ by the geodesic from $\tau(0)$ to $\tau(t)$ and call τ_t an initial straightening of τ. More generally, let τ' be a curve obtained from τ by replacing $\tau(t_1, t_2)$ by the geodesic segment connecting $\tau(t_1)$ and $\tau(t_2)$, and call τ' a straightening of τ.

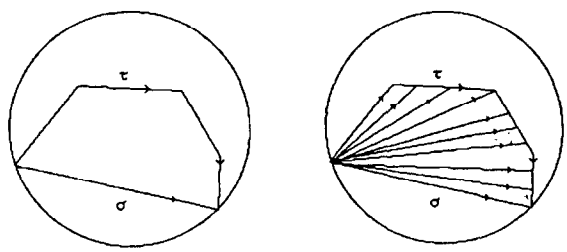

Fig. 7. A family of initial straightenings interpolating between τ and σ

LEMMA 3.3. *A pseudo-graph τ remains a pseudo-graph after a straightening.*

Proof. Let L be any geodesic segment in the foliation which makes τ a pseudo-graph, so that L intersects τ at most once. A straightening of τ will also intersect L at most once. Since

this holds for all L, the straightened curve is embedded, and is a pseudo-graph with respect to the same foliation.

LEMMA 3.4. *Let D denote a convex disk, let $\sigma\colon [0, 1] \to D$ be a proper geodesic in D and let $\Sigma = \{$pseudo-graphs τ in D with $\partial\tau = \partial\sigma\}$. There is a continuous length decreasing deformation retraction of Σ to σ.*

Proof. Perform initial straightening and apply Lemma 3.3.

LEMMA 3.5. *A piecewise geodesic curve in a convex disk with vertices on ∂D is a pseudo-graph.*

Proof. Such curves can be drawn in the unit disk in R^2 to originate at $(-1, 0)$ and have x-coordinate always increasing. They are pseudographs in the unit disk, with respect to the foliation consisting of geodesic segments connecting $(x, -y)$ to (x, y), irrespective of the convex metric.

Example 3.6. The seven possible configurations of such a curve with up to three interior tangencies are shown in Fig. 8, up to homeomorphism of the disk. Each is drawn so as to be a pseudo-graph with respect to the obvious vertical foliation of the disk. If the curve intersects the disk in more than one arc, then combinations of these may occur. However their convex hulls are disjoint so it suffices to consider the connected case.

LEMMA 3.8. *A generic k-parameter family of curves on a Riemannian 2-sphere can be straightened continuously. The straightened family is homotopic to the original family.*

Proof. Consider first a 1-parameter family $g_s = g_{s,0}$ of embedded curves sweeping out the 2-sphere, where $s \in [0, 1]$. Cover the 2-sphere with finitely many disks $D_1, \ldots, D_n$ as before. Flow the family $g_{s,0}$ to get a new family $g_{s,1}$ by the following process:

In the disk D_1 replace each arc of intersection of $g_{s,0} \cap D_1$ by the corresponding geodesic arc. Replace any closed curve $g_{s,0}$ lying completely in D_1 with the empty set. If a gap develops due to a tangency of some curve $g_{s',0}$ with ∂D_1, fill it in with piecewise geodesic arcs to form a new continuous family, with maximal length no larger than before.

Reparametrize the new family by the unit interval, replacing a curve with an interior tangency with a closed interval of curves which fill the gap. This defines a map $f_1\colon I \to I$ with $f_1(s_1) = s_0$, where s_1 is the index of any curve which originated with γ_{s_0}. Repeat for $D_2, D_3, \ldots$ A sequence of parameters $S = (s_0, s_1, s_2, \ldots)$ and maps $f_i\colon I \to I$ with

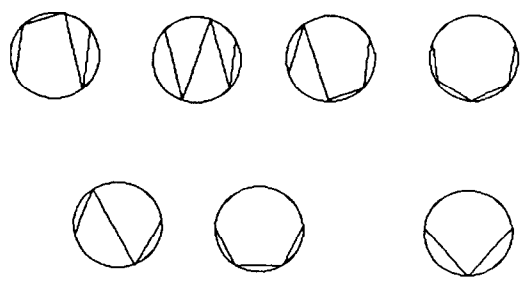

Fig. 8. The possible configurations of a connected arc with one, two and three internal tangencies, up to homeomorphism of the disk

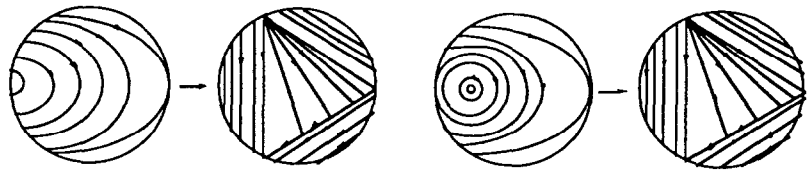

Fig. 9. Two examples of straightened 1-parameter families of curves

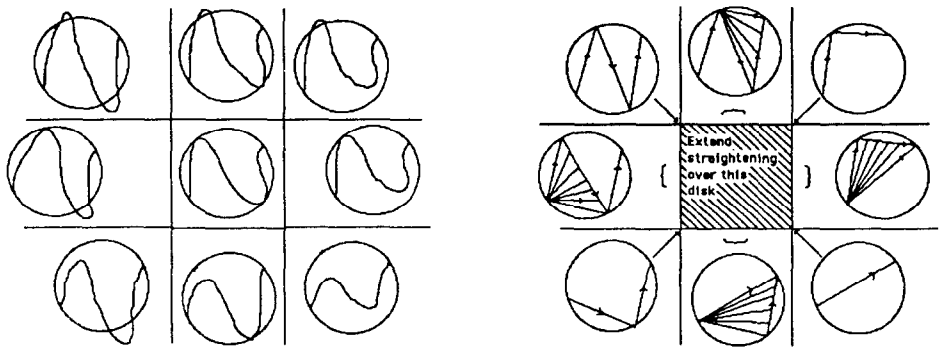

Fig. 10. A typical 2-parameter family of curves near a point where a double tangency occurs, before and after straightening.

$f_i(s_i) = s_{i-1}$ determines the evolution of any given curve at integral times. The parameter space $(s_0, s_1, s_2, \ldots)$ with the inverse limit topology is homeomorphic to the unit interval. Note that in this topology the map Δ, defined in section one, gives a continuous map from $I \times S^1$ to itself, decreasing the length of each non-geodesic curve.

In a two-parameter family we may have isolated parameter values corresponding to a curve which is tangent to the boundary of some disk D_i at two points. We will describe a straightening process which extends over these values. In the complement of these points the straightening process is defined just as before. Expanding each curve having a single point of internal tangency to an interval of curves by initial straightening, and straightening any other curve, the straightening process becomes defined except on the values in the parameter space where there are two internal tangencies in some disk D_i. A neighborhood of such a point is depicted in Fig. 10, where the point has been replaced by a 2-cell. The straightening process is defined continuously on the boundary of a 2-cell by the 1-parameter case. The straightened curves on the boundary of this 2-cell are each pseudo-graphs with common boundary, so Lemma 3.4 states that the straightening process extends continuously over the missing 2-cell and that each resulting curve has length no greater than the curve that it flowed from.

The parameter space is replaced by a new parameter space in which a point is replaced by a 2-cell. In the limit the parametrizing space is the inverse limit. Topologically it is homeomorphic to the original parametrizing space [5]. In practice it suffices to consider the evolution of a family of curves parametrized by a manifold for a finite amount of time, and it is clear that the topology of the parameter space is unchanged when a point is replaced by a disk finitely many times.

For a 3-parameter family the argument is similar. The straightening process is defined and continuous for curves with 2 or fewer tangencies. At a triple tangency various double tangencies converge. It is necessary to extend straightened arcs over a 3-cell. On the

boundary of this cell the family consists of pseudo-graphs with a common pair of endpoints. Lemma 3.4 says that these can be canonically deformed to the corresponding geodesic without increasing length. The case of a k-parameter family also follows similarly. Finally we note that since $\mathrm{Diff}(D^2, \mathrm{rel}\,\partial)$ is contractible, the straightened family is homotopic to the original family, concluding the proof of Lemma 3.8.

We now indicate how to use the disk flow to establish the existence of three simple closed geodesics on a 2-sphere endowed with an arbitrary Riemannian metric, giving a new proof of the famous theorem of Lusternik and Schnirelman [15]. Previous proofs of this result have constructed length or energy decreasing deformations for families of embedded curves. See [14] for a discussion of this problem, and [2, 9, 13] for related results. We describe a process of replacing a generic continuous family of embedded curves on a 2-sphere with a new family such that any curve in the final family is shorter than the corresponding initial curve. During the deformation from the first family to the second, lengths of some curves may increase, but each curve in the first family is shorter than the curve in the second family that it deforms to.

THEOREM 3.11. [15]. *A Riemannian 2-sphere contains three simple closed geodesics.*

Proof. First we give an easy construction of the existence of one simple closed geodesic. Consider a family of embedded curves, parametrized by $I = [0, 1]$ and starting and ending with a point, which together define a degree one map of the 2-sphere. Pick $\delta_0 > 0$ so that any embedded curve of length at most δ_0 lies in one of the disks $\{D_i\}$. Then if all curves in the family have length less than δ_0 we can extend the map of the sphere to a continuous map of the ball, a contradiction to the assumption that the family represents a degree one map of the 2-sphere. So the length of a maximal length curve in such a family can never be less than δ_0. Apply the disk flow to the family described above. We can shrink the disk D_i slightly if necessary, to ensure that $\{\partial D_i\}$ is tangent to any curve in the family in at most one point, and is tangent to only finitely many curves in the family. The set of parameter values for curves that flow to a point and disappear is open, since any such curve eventually lies inside a single disk D_i, and thus so does any nearby curve. Let $S = (s_0, s_1, s_2 \ldots)$ be the smallest parameter value in the lexicographic ordering for which γ_S does not disappear. Then γ_S has a subsequence converging to a closed geodesic.

To establish the existence of three distinct simple closed geodesics we follow the arguments of [14, Appendix A1]. The space of unparametrized embedded piecewise-smooth curves on S^2, with all point curves identified, is homotopy equivalent to RP^3. We denote it by Σ, and by Σ^k the curves in Σ of length at most k. The Z_2-homology of Σ is generated by one cycle in each dimension up to dimension three. The zero-dimension homology corresponds to point curves. Let h_j be the non-trivial Z_2-homology class in the jth homology group of Σ, $j = 1, 2, 3$. Fix a Riemannian metric on S^2 and let z be a j-cycle with $[z] = h_j$. Define the length of a cycle z to be the maximum length of a point in Σ which is in the image of z. Define $k(h_j) = \inf\{\mathrm{length}(z) : [z] = h_j\}$.

Let $U_j \subset \Sigma$ be an open neighborhood of the set of simple closed geodesics of length $k(h_j)$, with the C^0-topology as in §1. If there are only finitely many simple closed geodesics then we can take each component of U_j to be contractible. We will apply the map Δ defined in §1 to the curves in Σ. Note that Δ is not continuous on Σ near curves which have internal tangencies to a disk D_i. To deal with this problem, we perturb the disks D_i slightly so that no simple closed geodesic is tangent to the boundary of a disk D_i. This is easily arranged if there are less than three simple closed geodesics, and otherwise we are done. Curves which are close to the simple closed geodesics have images under Δ which remain close, and thus

there is a smaller neighborhood $U_{j'}$ of the set of closed geodesics of length $k(h_j)$ with $\Delta(U_{j'}) \subset U_j$.

The following lemma is a key step in the proof of the Theorem 3.11.

LEMMA 3.12. *If there are less than three distinct simple closed geodesics on S^2 then there is an $\varepsilon > 0$ and a cycle z_j representing h_j such that every curve in z_j either lies in U_j or has length less than $k(h_j) - \varepsilon$.*

Proof. Fix a homology class h_j. We first show that there is an $\varepsilon > 0$ such that $\Delta(\Sigma^{k(h_j)+\varepsilon}) \subset \Sigma^{k(h_j)-\varepsilon} \cup U_j$. If not, we can find a sequence of curves s_m such that $\text{length}(s_m) < k(h_j) + 1/m$, $\text{length}(\Delta(s_m)) > k(h_j) - 1/m$ and s_m is not in $U_{j'}$, $m = 1, 2, 3, \ldots$. Since the curves s_m are of uniformly bounded length, we can assume, after passing to a subsequence, that they converge to a curve s. We will show that $\text{length}(s) = \text{length}(\Delta s) = k(h_j)$ (this is not immediate, since Δ is not continuous on Σ). First, the arguments of Theorem 1.8 show that $\text{length}(\Delta s) = \text{length}(s)$, as otherwise Δ uniformly decreases the length of the approximating curves s_m. By assertion 5) of Theorem 1.8, s is a geodesic or a point. We need to show that $\text{length}(s) = \lim_{m\to\infty} \text{length}(s_m)$.

Since $s_m \to s$ and s is a geodesic, if $\text{length}(s) < \lim_{m\to\infty} \text{length}(s_m)$ then there is an arc $\alpha \subset s$ and a sequence of arcs $\alpha_m \subset s_m$ converging to α with $\text{length}(\alpha) < \lim_{m\to\infty} \text{length}(\alpha_m)$ and with each α_m contained in $\text{int}(D_k)$ for some disk D_k. Then $\text{length}(\Delta(s_m)) < \text{length}(s_m) - \delta$ for some fixed δ independent of m, a contradiction. So $s_m \subset U_{j'}$ for large m, a contradiction. We conclude that $\Delta(\Sigma^{k(h_j)+\varepsilon}) \subset \Delta(\Sigma^{k(h_j)-\varepsilon}) \cup U_j$ for sufficiently small ε.

To prove Lemma 3.12, take a cycle z representing h_j with every curve in z contained in $\Sigma^{k(h_j)+\varepsilon}$. Then $\Delta(z) \subset \Sigma^{k(h_j)-\varepsilon} \cup U_j$, proving the lemma.

The existence of at least three distinct simple closed geodesics now follows from more or less standard arguments concerning "subordinated homology classes", going back to Lusternik and Schnirelman, which we briefly present. Σ is homotopy equivalent to RP^3. The 1-parameter family we described above is a cycle h_1 generating $H_1(\Sigma)$. Take a 2-parameter family $z_2 : RP^2 \to \Sigma$ representing a generator h_2 of $H_2(\Sigma; Z_2)$. Such a cycle can be constructed on a round sphere by taking all great circles perpendicular to the equator, together with all round curves parallel to one of these circles. Note that an orientation reversing closed curve on this RP^2 represents a 1-parameter family of curves in h_1. For a non-round sphere, a diffeomorphism to the round sphere gives corresponding families.

Flow the 2-parameter family of curves in the cycle z_2. There must be some curve in z_2 at each time which has length $\geq k(h_2)$. Theorem 1.8 implies that a subsequence of these curves will converge to a closed geodesic of length $\geq k(h_2)$. If $k(h_2) > k(h_1)$ then we have obtained two distinct geodesics. If $k(h_2) = k(h_1)$ and there is only one closed geodesic obtained from flowing both cycles, we will deduce a contradiction. Pick z_2 so that every curve in z_2 is either in U_1 (which coincides with U_2) or has length less than $k(h_2) - \varepsilon$, for some fixed ε. Let $W_1 = z_2^{-1}(U_1)$, so that W_1 is an open set in RP^2. Shrinking U_1 slightly, we can arrange that W_1 is the interior of a codimension-zero submanifold of RP^2. W_1 then intersects any orientation reversing closed curve in RP^2, since otherwise there is a representative of h_1 consisting of curves which are all shorter than $k(h_1)$. In that case W_1 contains an embedded curve β representing a generator of the fundamental group of RP^2. $z_2(\beta)$ is then a representative of h_1, but is completely contained in U_1, and since all curves in U_1 lie in a neighborhood of a single simple closed geodesic on S^2 and are homotopic to a generator of the fundamental group of this neighborhood, $z_2(\beta)$ can be homotoped to a constant map

in U_1, a contradiction. The same argument further implies that there must be infinitely many closed geodesics of the same length if $k(h_2) = k(h_1)$.

A similar argument applied to h_3 gives a third closed geodesic if $k(h_3) = k(h_2)$, but we need an additional observation.

Claim 3.13. Changing the domain of z_2 from RP^2 to an arbitrary 2-manifold F^2 does not change $k(h_2)$, so that $k(h_2) = \inf\{\max\{\text{length}(z)\colon z\colon F^2 \to \Sigma,\ F^2 \text{ is any surface, and } [z] = h_2\}$.

Proof. Suppose that there is a map $z\colon F \to \Sigma$ where F is a surface, $[z] = h_2$ and $\text{length}(z) < k(h_2)$. Let $k'(h_2) = \inf\{\max\{\text{length}(s)\colon s \in z,\ z\colon F^2 \to \Sigma,\ F \text{ is any surface, and } [z] = h_2\}$. Applying Lemma 3.12 we can find a map $z'\colon F^2 \to \Sigma$ with $[z'] = h_2$ and with each curve in the image of z' either lying in U_1 or having length $< k'(h_2) - \varepsilon$. Let $W_1 = z'^{-1}(U_1)$. Since U_1 is contractible, if W_1 is not simply connected then F can be surgered in W_1 to produce a new surface F' and map $z''\colon F' \to \Sigma$ with $[z''] = h_2$ and with $W_1' = z''^{-1}(U_1)$ consisting entirely of disks. Now $[z''] = h_2$ so F' contains a loop α with $z''(\alpha)$ representing h_1. If $z''(F')$ is mapped into RP^3 by using the homotopy equivalence of Σ with RP^3, then α is obtained by taking the intersection pairing $F' \cdot F'$. Equivalently, α can be obtained by taking a representative in F' of the homology class $PD(PD(\alpha) \cup PD(\alpha))$, where PD indicates the Poincaré dual in $H^*(RP^3; Z_2)$. Intersection theory implies that α must intersect W_1'. But any curve can be homotoped off a collection of disks, so we have a contradiction, proving the claim.

We now find a third simple closed geodesic. If $k(h_3) > k(h_2)$ then we obtain a third closed simple geodesic of length greater than $k(h_2)$ by flowing a cycle z_3 and taking a convergent sequence of curves with length $\geq k(h_3)$. If $k(h_3) = k(h_2)$ then we can consider the cycle z_3, the image of an RP^3 in Σ corresponding to all the round circles in the standard S^2. By Lemma 3.12 we can arrange that all curves in z_3 either have length less than $k(h_2) - \varepsilon$ or have image under Δ which lies in U_2, a neighborhood of the simple closed geodesic of length $k(h_2)$. Take the set W_2 to be the interior of a codimension-zero submanifold of RP^3 mapping to U_2. Every non-trivial 2-cycle embedded in RP^3 represents h_2, and so must meet W_2 by Claim 3.13. It follows that W_2 contains an embedded loop representing h_1, again giving a contradiction since a loop of curves lying in U_2 all lie close to a single curve and thus can not be non-trivial homologically.

§4. FLOWING ARCS

The disk flow easily adapts to the setting of properly immersed arcs, allowing the flow of such an arc to a geodesic arc. In this setting we get new results on the existence of geodesics, since the analysis involved in the existence of the curvature flow has not been extended to the relative case.

We can prove for example a relative version of the Lusternik–Schnirelman theorem. The arguments are very similar to the ones in Section 4, so we give them more briefly.

We define the process of flowing an arc in exactly the same manner as flowing a curve, except near the endpoints of the arcs. There we can define various types of flow. One choice, corresponding to free boundary conditions, is illustrated in Fig. 11, where we depict a disk D_i meeting ∂F along part of its boundary. On an arc in D_i not meeting ∂F the flow is defined as before. If the arc meets ∂F in one point, then the flow replaces it by the unique geodesic arc with the same endpoint $\partial D_i - \partial F$ and with the other endpoint meeting ∂F perpendicularly. If the arc has both endpoints on ∂F then it disappears, as in Fig. 11(c). Two types of

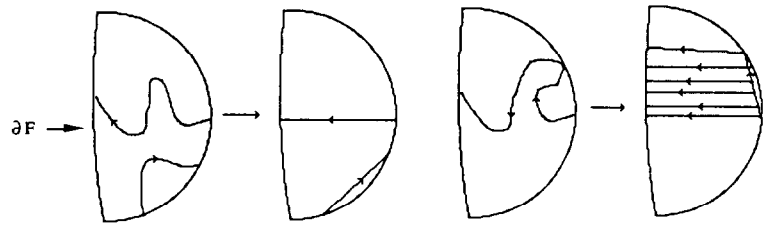

Fig. 11(a).

Fig. 11(b).

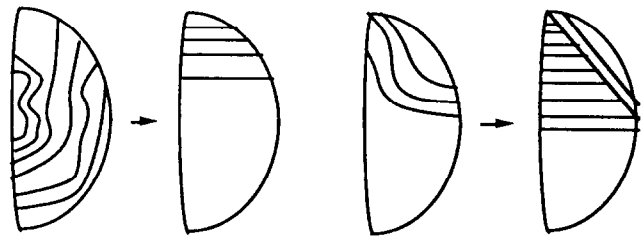

Fig. 11(c).

Fig. 11(d).

gaps can occur for families of arcs under this flow, but as before the gaps can be filled in by arcs which are no longer than the ones which give rise to them. The first type of gap occurs as before when an arc is tangent to ∂D_i, as shown in Fig. 11(b). The second kind of gap occurs when an arc has one endpoint on $\partial D_i \cap \partial F$, as in Fig. 11(d).

THEOREM 4.2. *A Riemannian 2-disk with convex boundary contains at least two distinct simple geodesic arcs with endpoints perpendicular to the boundary.*

Proof. The proof is identical to that of Theorem 3.11, except that in this case the space we work with is the space of simple arcs in a disk together with trivial one point arcs, with all the trivial arcs identified to a point. This space is homotopy equivalent to RP^2 and so has only two Z_2-homology classes instead of three, leading to only two normal geodesic arcs.

§5. CONCLUDING REMARKS

We present here some general remarks and questions concerning various flows on surfaces.

Remark 5.1. *No singularities develop for the curvature flow of a curve γ which lifts to an embedding in some cover of F.*

Proof. Lift γ to $F\gamma$ where it lifts to a generator of the fundamental group. This surface is geodesically convex, so the lift of γ exists for all time under the curvature flow and flows to a geodesic [9]. Its projection to F gives the curvature flow on F.

Remark 5.2. *Curvature flow for a curve that lifts to a simple curve converges to a unique geodesic on a surface on non-positive curvature.*

The proof is similar to that for the disk flow given in §1.

Remark 5.3. *The smooth curvature flow can be used instead of the disk flow to make most of the arguments of* §2.

The smooth curvature flow is more canonical than the disk flow, and can be run simultaneously for all curves. However one needs to restrict to curves that do not develop singularities, so that their curvature flow exists for all times, and even for this class one needs to make arguments more complicated than those used in studying the evolution of embedded curves. Grayson's results [9], together with recent results of Angenent [1], provide the needed techniques for $d(s)$ minimal. When $d(s)$ is not minimal, additional arguments seem necessary.

Question 5.4. *Does the disk flow converge to the curvature flow as the disks are taken smaller?*

From the nature of the heat equation one would expect this to occur if time is decreased as the square of the disk radius. For certain special cases it can be shown that convergence does occur. Note that it suffices to check convergence for piecewise geodesic curves. The disk flow is ideally suited to implementation on a computer, so this question is interesting to determine whether the curvature flow can be approximated through this method. Also note that this could lead to a new proof of Grayson's result [9]. It seems likely that the Birkhoff flow does not converge to the curvature flow, but rather to some type of energy flow.

Remark 5.5. *Higher dimensions.* The disk flow generalizes to higher dimensions in two key cases. The first is for curves in an n-dimensional manifold. For a finite number of curves, the construction of a disk flow is almost identical, with segments of curves being straightened in a k-ball instead of a 2-disk. Of course the intersection properties do not persist in higher dimensions, but it is still true that curves can be flowed to points or to geodesics. Long term existence for the curvature flow is unknown in dimensions larger than two. The Birkhoff process also extends to higher dimensions. The advantage of the disk flow in higher dimensions is that it is more natural from the point of view of computer implementation.

The other generalization is to two-dimensional minimal surfaces in 3-manifolds. Under a suitable hypothesis about the incompressibility of a surface, it can be shown that it flows by a local minimizing process to an embedded minimal surface. For example, an embedded surface isotopic to a totally geodesic surface will flow to that totally geodesic surface under a disk flow process which replaces the intersection of the surface with a ball by least area disks. This flow is described in [12], but in that setting convergence of the flow for a particular initial choice of surface was not clear. In [12] this is overcome by flowing a minimizing sequence of surfaces. In the setting just described, this is unnecessary.

Remark 5.6. *Flows without metrics.* It is not actually necessary to have a metric on a manifold to define the disk flow on curves. It suffices to have a cover of a manifold by disks with unique geodesics defined for each pair of boundary points. This occurs for example if M is an affine manifold. One then cannot hope in general to get convergence to a geodesic, since there are examples of homotopy classes of curves in affine manifolds which do not contain geodesic representatives.

Acknowledgements—Results similar to those of Theorem 2.1 have been independently obtained by Marc Shepard in his Ph.D. thesis [16].

REFERENCES

1. S. ANGENENT: Parabolic equations for curves on surfaces I, *Ann. Math.* **132** (1990), 451–484.
2. W. BALLMAN: Doppelpunkte freie geschlossene Geodatische auf kompacten flachen, *Math. Z.* **161** (1978), 41–46.
3. W. BALLMAN, T. THORBERGSSON and W. ZILLER: Existence of closed geodesics on positively curved manifolds, *J. Diff. Geom.* **18** (1983), 221–252.
4. G. D. BIRKHOFF: Dynamical systems with two degrees of freedom, *Trans. Amer. Math. Soc.* **18** (1917), 199–300.
5. M. BROWN: Some applications of an approximation theorem for inverse limits, *Proc. AMS* **11** (1960), 478–483.
6. M. H. FREEDMAN, J. HASS and G. P. SCOTT: Closed geodesics on surfaces, *Bull. London Math. Soc.* **14** (1982), 385–391.
7. M. GAGE: Curve shortening makes convex curves circular, *Invent. Math.* **76** (1984), 357–364.
8. M. GAGE and R. S. HAMILTON: The heat equation shrinking convex plane curves. *J. Diff. Geom.* **23** (1986), 69–96.
9. M. GRAYSON: Shortening embedded curves, *Ann. Math.* **129** (1989), 71–112.
10. M. GRAYSON: The shape of a figure-eight curve under the curve shortening flow, *Invent. Math.* **96** (1989), 177–180.
11. J. HASS and G. P. SCOTT: Intersections of curves on surfaces, *Israel J. Math.* **51** (1985), 90–120.
12. J. HASS and G. P. SCOTT: The existence of least area surfaces in 3-manifolds, *Trans. Amer. Math. Soc.* **310** (1988), 87–114.
13. J. JOST: A nonparametric proof of the theorem of Lusternik and Schnirelman, *Arch. Math.* **53** (1989), 497–509.
14. W. KLINGENBERG: Lectures on closed geodesics, *Grundlehren Math. Wiss* **230**, Springer, Berlin, (1978).
15. L. LUSTERNIK and L. SCHNIRELMAN: Sur le probleme de trois geodesiques fermees sur les surface de genre 0, *C.R. Acad. Sci. Paris* **189** (1929), 269–271.
16. M. SHEPARD: Ph.D. thesis, U.C. Berkeley, (1991).
17. V. TURAEV: Problem list from workshop on low dimensional topology, Problem 10, Luminy, (1989).
18. B. WHITE: Generic properties of stationary surfaces, *Indiana Math. J.* **36** (1987), 567–602.

Department of Mathematics
University of California
Davis, CA 95616
U.S.A.

and

Department of Mathematics
University of Michigan
Ann Arbor, MI 48109
U.S.A.