

## AREA AND THE LENGTH OF THE SHORTEST CLOSED GEODESIC

CHRISTOPHER B. CROKE

### 1. Introduction

The main purpose of this paper is to prove the following theorem:

**Theorem 4.2.** *For any metric on a two-dimensional sphere  $31\sqrt{A} \geq L$ , where  $A$  represents the area, and  $L$  the length of the shortest nontrivial closed geodesic.*

The constant 31 above is not the best constant. One suspects that the best constant would be  $3^{1/4}2^{1/2} = 1.86121$ . We will discuss this later.

The question, for which the above theorem is the answer in two dimensions, was posed by Gromov for all closed manifolds in [12, p. 135]. The corresponding theorem is known for all other closed surfaces as we will see below. The difficulty with the sphere is that it is simply connected. In fact, all other results relating area (or volume) to the length of closed geodesics concern essential (not null homotopic) geodesics. That is, they concern  $\text{sys}(M)$  (read “systole of  $M$ ”), the length of the shortest essential geodesic.

The first theorem of this type was proved by Loewner in 1949 in an unpublished work (see [3] and [4]). He showed that for any metric on the two torus  $3^{1/4}2^{1/2}\sqrt{A(M)} \geq \text{sys}(M)$  with equality holding if and only if  $M$  is a flat equilateral torus. (The fact that the constant is the same as the conjectured constant for  $S^2$  comes from the fact that both extremal metrics are built out of two flat equilateral triangles.) The proof of the theorem relies on the fact that all metrics are conformal to a flat metric. Using a similar method Pu in 1952 (see [17]) showed that for any metric on  $\mathbf{RP}^2$ ;  $\sqrt{\pi/2}\sqrt{A(M)} \geq \text{sys}(M)$ , with equality holding if and only if  $M$  has constant curvature.

In 1960 Accola [1] and Blatter [6] independently showed that there was a function  $f(g)$  such that for any metric on a surface of genus  $g$ ,  $f(g)\sqrt{A(M)} \geq \text{sys}(M)$ . Unfortunately, as  $g$  tends to  $\infty$  the function  $f(g)$  tends to  $\infty$  while one would expect it to tend to 0. In 1981 Hebda [15] and independently

Burago and Zalgaller [7] improved this result by showing that for all surfaces of genus  $\geq 1$ ,  $\sqrt{2} \sqrt{A(M)} \geq \text{sys}(M)$ .

The major work in this field is the recent work of Gromov [12]. In it are many results. Among them is the result that for a surface of genus  $g$ ,  $\tilde{f}(g) \sqrt{A(M)} \geq \text{sys}(M)$ , where, in this case,  $\tilde{f}(g)$  is a function (given explicitly) which tends to 0 as  $g$  tends to  $\infty$  (see [12, pp. 4–5], Theorem 0.2.A). The main result of [12] is a higher dimensional theorem: There is a constant  $c(n)$  depending only on dimension  $n$  such that for every essential manifold  $M$  of dimension  $n$  we have  $c(n) \sqrt[n]{\text{Vol}(M)} \geq \text{sys}(M)$  ( $c(n)$  can be taken to be  $6(n+1)n \sqrt[n]{(n+1)!}$ ). In the above statement “ $M$  essential” means that there is a map  $f$  from  $M$  to a  $k(\pi, 1)$  such that  $f_*[M] \neq 0$  where  $[M] \in H_n(M)$  is the fundamental class (use  $\mathbf{Z}_2$  coefficients if  $M$  not orientable). As examples of essential manifolds we have  $\mathbf{R}P^n$  and  $\mathbf{T}^n$ .

It is easy to construct examples of nonessential manifolds with  $\text{sys}(M) = 1$  and arbitrarily small volume, for example take product metrics on  $S^1 \times S^2$  where  $S^2$  gets arbitrarily small. However, they may still have short nonessential closed geodesics. The general question of volume versus the length of the shortest closed geodesic is still very much open in higher dimensions. The goal of this paper is to answer the question in two dimensions.

The paper is divided into six sections, the first of which being this introduction. The second section recalls the Birkhoff curve shortening process, the fundamental tool in this paper, and derives some new properties. The third section contains the basic lemmas from which the theorems are proved in §§4 and 5. The key to all the proofs is Lemma 3.1.

In §4 the main theorem is proved. Along the way we also prove

**Theorem 4.1.** *For any metric on  $S^2$  we have  $9D \geq L$  where  $D$  represents the diameter.*

In §5 we consider complete noncompact surfaces of finite area  $A$ . It was shown in [20] and [2] that all such surfaces have closed geodesics (in fact infinitely many). Gromov in [12] showed (as a special case of Theorem 4.4.A) that for most such surfaces we have  $\text{const} \sqrt{A} \geq L$ . We show that the techniques used to prove the main theorem serve to show  $\text{const} \sqrt{A} \geq L$  for all such surfaces. The two main cases not covered by Gromov’s theorem are the plane and the cylinder.

In §6 we consider the case of convex hypersurfaces of  $\mathbf{R}^{n+1}$ . We show

**Theorem 6.1.** *If  $M^n \subset \mathbf{R}^{n+1}$  is a convex hypersurface, then  $c(n) \sqrt[n]{\text{Vol}(M)} \geq L$ .*

The constant  $c(n)$  is discussed and is in some sense only off by a factor of two from the sharp constant. In particular  $c(2) = 2 \cdot 3^{1/4} \cdot 2^{1/2}$ . The argument

leads one to guess as to the optimal metric. In particular the extremal metric in two dimensions should be two copies of an equilateral triangle glued together along the boundary (of course this is a degenerate metric).

Andre Treibergs [19] has independently proved Theorem 6.1 as well as an extension to higher dimensional minimax volumes (yielding in particular upper bounds for areas of minimal surfaces in convex three spheres). We have nevertheless included our proof (which was proved at about the same time—Summer 1983) because it is significantly easier than the proof in [19] and it leads one to a conjecture as to the sharp constant. In [19] a different metric is conjectured as optimal. However, the geodesic used in the calculation in [19] was not the shortest closed geodesic. In fact the extremal metric conjectured above (two equilateral triangles) is better than the one suggested in [19].

Gromov has suggested that using the main Lemma (3.1) along with ideas in [12] one should be able to show that the filling radius (see [12] §1 for a definition) is larger than a constant times the length of the shortest closed geodesic. This, along with the main theorem (1.3.A) of [12], its extension ([12] 4.4.C), and results in [16], would yield an alternative proof of the main results in this paper.

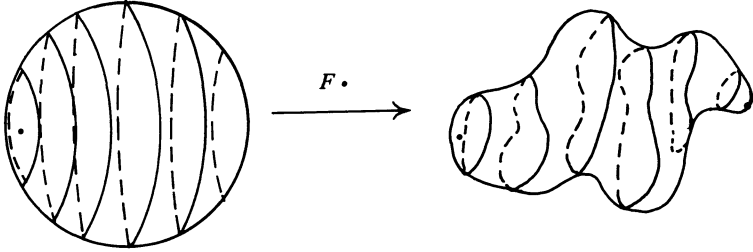
Many thanks are due to M. Gromov for helpful conversations on all aspects of this paper. Thanks are also due to H. Karcher for early discussions on §5, G. Thorbergsson for arousing my interest in the problem, and M. Berger for help in finding references.

The author would also like to thank Max-Planck-Institut Für Mathematik, Institut des Hautes Études Scientifiques, and Mathematical Sciences Research Institute for their hospitality and financial support during the preparation of this paper.

## 2. Birkhoff's Ideas

In this paper we will work in the space  $\Lambda$  of piecewise smooth closed curves  $\gamma: [0, 1] \rightarrow M$ , where  $M$  is a riemannian manifold and  $\Lambda$  has the  $C^0$ -topology. By  $L(\gamma)$  we will mean the length of  $\gamma$ .

We borrow two major ideas from Birkhoff (see [5]). The first idea we will use is his method of finding closed geodesics on spheres. In particular when  $M$  is  $S^2$  we will find a 1-parameter family of curves starting and ending at a point curve in such a way that the induced map  $F: S^2 \rightarrow S^2$  (see figure) has nonzero degree. Birkhoff's argument (or the minimax argument) allows us to conclude that  $M$  has a nontrivial closed geodesic of length less than or equal to the length of the longest curve in the 1-parameter family. We will use this



argument in the proofs of the main Theorems 4.1 and 4.2. We will use the higher dimensional version of this in §6 in discussing convex hypersurfaces. Further, we will even use a modification of this argument in the case of noncompact surfaces (§5).

The second idea that we will use is the Birkhoff curve shortening process, B.C.S.P. (which Birkhoff used in the above mentioned argument). Since we need to derive some new properties of B.C.S.P., we will recall it here.

The B.C.S.P.,  $\beta^N: \Lambda^E \rightarrow \Lambda^E$ , depends on an integer parameter  $N$ , and is a map from  $\Lambda^E$ , the space of curves of energy less than  $E$ , to itself.  $\beta^N$  is called the B.C.S.P. with  $N$  breaks. For fixed  $E$ ,  $N$  is chosen so large such that  $\sqrt{E}/N$  is smaller than the injectivity radius of the manifold,  $\text{inj}(M)$ , or in some cases the injectivity radius of a part of the manifold. In each application we will choose the number of breaks large enough to suit our purposes.

Given a curve  $\gamma \in \Lambda^E$  we will define the new curve  $\beta^N(\gamma) \in \Lambda^E$  as well as a homotopy  $\gamma_s$ ,  $s \in [0, 1]$ , from  $\gamma = \gamma_0$  to  $\beta^N(\gamma) = \gamma_1$ . The homotopy will be defined in such a way that  $L(\gamma_{S_2}) \leq L(\gamma_{S_1})$  whenever  $S_2 \geq S_1$ .

We will assume that  $\gamma$  is parametrized proportional to arclength. If not the first part of the homotopy is to reparametrize  $\gamma$  so that it is. We define  $\gamma_{1/2}$  to be the unique piecewise geodesic closed curve such that  $\gamma_{1/2}(i/N) = \gamma(i/N)$  for all integers  $i = 0, 1, 2, \dots, N$ , and such that  $\gamma_{1/2}|_{[i/N, (i+1)/N]}$  is a minimizing geodesic parametrized proportional to arclength. The uniqueness of  $\gamma_{1/2}$  comes from:

$$d\left(\gamma\left(\frac{i}{N}\right), \gamma\left(\frac{i+1}{N}\right)\right) \leq L(\gamma|_{[i/N, (i+1)/N]}) \leq \frac{L(\gamma)}{N} \leq \frac{\sqrt{E}}{N} < \text{inj}(M).$$

For  $s \in [0, \frac{1}{2}]$  we define  $\gamma_s$  by

$$\gamma_s\left(\frac{i}{N} + t\right) = \begin{cases} \tau_i^s(t) & 0 \leq t \leq \frac{2s}{N}, \\ \gamma\left(\frac{i}{N} + t\right) & \frac{2s}{N} \leq t \leq \frac{1}{N}, \end{cases}$$

where  $\tau_i^s$  is the minimizing geodesic from  $\gamma(i/N)$  to  $\gamma(i/N + 2s/N)$  parametrized on the interval  $[0, 2s/N]$  proportional to arclength. The uniqueness and continuity of the  $\gamma_s$  follows (as before) from the fact that  $L(\tau_i^s) < \text{inj}(M)$  for all  $i$  and  $s$ .

$\gamma_1$  is defined as the unique piecewise geodesic closed curve with  $\gamma_1(i/N + 1/(2N)) = \gamma_{1/2}(i/N + 1/(2N))$  which is parametrized proportional to arclength on each interval  $[i/N + 1/(2N), (i+1)/N + 1/(2N)]$ . We then define  $\gamma_s$  for  $s \in [\frac{1}{2}, 1]$  to homotope between  $\gamma_{1/2}$  and  $\gamma_1$  in the same way that  $\gamma_s$ ,  $s \in [0, \frac{1}{2}]$  homotopes from  $\gamma_0$  to  $\gamma_{1/2}$ . The continuity and uniqueness follow as before.

Birkhoff shows that  $\beta^N: \Lambda^E \rightarrow \Lambda^E$  is continuous if  $N$  is large in terms of  $E$  and the geometry of  $M$  (we will not need this fact directly). The closed geodesics are the only fixed points of  $\beta^N$ . Birkhoff also shows that for any  $\gamma \in \Lambda^E$  the sequence  $\{\gamma_i\}$ , defined by  $\gamma = \gamma_0$  and  $\gamma_{i+1} = \beta^N(\gamma_i)$ , converges to a closed geodesic. However, this closed geodesic may be a point curve. If the limit is a point curve then the homotopies described above give rise in a natural way to a map from the two disk  $D^2$  into  $M$  with  $\gamma$  as the boundary. We should remark that since the first step in B.C.S.P. is to reparametrize proportional to arclength and since  $\beta^N(\gamma)$  will not in general be globally parametrized by arclength parts of the map from  $D^2$  into  $M$  will consist of simply reparametrizing these curves.

For the rest of this section let  $M$  be a complete connected oriented surface (two dimensional). Let  $\gamma \in \Lambda$  be a simple (no self intersections) closed curve on  $M$  which divides  $M$  into two components. Let  $\Omega$  (open) be one of these components. Then  $\gamma$  will be called *convex to*  $\Omega$  if there is an  $\varepsilon > 0$  such that for all  $x, y \in \gamma$ , with  $d(x, y) < \varepsilon$ , the minimizing geodesic  $\tau$  from  $x$  to  $y$  satisfies  $\tau \subset \bar{\Omega}$ . In fact this means that if  $x, y \in \bar{\Omega}$  with  $d(x, y) < \varepsilon$  then  $\tau \subset \bar{\Omega}$ . For if not there would be points  $\bar{x}, \bar{y} \in \gamma \cap \tau$  such that the segment  $\bar{\tau}$  of  $\tau$  from  $\bar{x}$  to  $\bar{y}$  does not lie in  $\bar{\Omega}$ . But  $d(\bar{x}, \bar{y}) = L(\bar{\tau}) < L(\tau) < \varepsilon$ . Hence, by the definition of  $\varepsilon$ ,  $\bar{\tau} \subset \bar{\Omega}$  yielding a contradiction.

In the applications in this paper  $\gamma$  will be a piecewise geodesic curve. In this case, the above definition reduces to the condition that all the angles of  $\gamma$  are convex to  $\Omega$ . In general it means that, in addition to the angles being convex to  $\Omega$ , at the  $C^\infty$  points of  $\gamma$  the curvature vector points toward  $\Omega$ .

For  $x \in M$ , we let  $\text{inj}(x)$  represent the injectivity radius at  $x$  (i.e. the minimum distance from  $x$  to its cut locus). For a compact set  $K$  we let  $\text{inj}(K) = \min\{\text{inj}(x) \mid x \in K\}$ .

For  $\gamma$  convex to  $\Omega$  as above we define  $\varepsilon(\gamma) > 0$  as follows. For each  $x \in \gamma$  we let  $\gamma_x = \gamma|_{[a,b]}$  be the largest connected open segment of  $\gamma$  containing  $x$

such that for all  $y \in \gamma_x$ ,  $d(x, y) < \text{inj}(\gamma)$ . Let  $\varepsilon(x, \gamma) = \frac{1}{2}d(x, \gamma - \gamma_x)$  (if  $\gamma = \gamma_x$  then set  $\varepsilon(x, \gamma) = \frac{1}{2} \text{inj}(\gamma)$ ). We then define

$$\varepsilon(\gamma) = \min\{\varepsilon(x, \gamma) \mid x \in \gamma\}.$$

Note that  $\varepsilon(\gamma) \leq \frac{1}{2} \text{inj}(\gamma)$ .

We now introduce an elementary lemma. The proof is straightforward but is included for completeness.

**Lemma 2.1.** *Let  $\gamma$  be convex to  $\Omega$ . Then*

1) *For  $x \in \gamma$  and  $y \in \gamma_x$  the unique minimizing geodesic  $\tau$  from  $x$  to  $y$  satisfies  $\tau \subset \bar{\Omega}$  and either  $\tau \cap \partial\Omega = \{x, y\}$  or  $\tau \subset \partial\Omega$ .*

2) *for  $x \in \gamma$  and  $y \in \bar{\Omega}$  such that  $d(x, y) < \varepsilon(\gamma)$ , the unique minimizing geodesic  $\tau$  from  $x$  to  $y$  satisfies  $\tau \subset \bar{\Omega}$ .*

3) *Let  $x \in \gamma$  and  $y \in \Omega$ . Then if  $\tau$  is the shortest path in  $\bar{\Omega}$  from  $x$  to  $y$  then  $\tau$  is a geodesic of  $M$ ,  $\tau \cap \partial\Omega = \{x\}$ , and  $\tau'(0)$  is not tangent to  $\gamma = \partial\Omega$ .*

4) *Let  $x, y \in \Omega$  and  $\tau$  the shortest path in  $\bar{\Omega}$  from  $x$  to  $y$ . Then  $\tau \subset \Omega$ .*

*Proof.* We start by noting that if  $\tau$  is a geodesic segment such that  $\tau \subset \bar{\Omega}$  and  $\tau'(x_0)$  is tangent to  $\gamma = \partial\Omega$  for some  $x_0 \in \gamma$  then the convexity of  $\gamma$  implies that  $\tau \subset \partial\Omega = \gamma$ . In fact, if  $x_0$  is not a  $C^\infty$  point of  $\gamma$  then the above holds under the assumption that  $\tau'(x_0)$  is tangent to either of the tangents of  $\gamma$ .

The second part of (1) follows from the first part (i.e.  $\tau \subset \bar{\Omega}$ ) and the above by noting that if an interior point of  $\tau$  intersects  $\gamma$  then it must be tangent to  $\gamma$  at that point since the angles of  $\gamma$  are convex to  $\Omega$ .

To see the first part of (1), we assume for simplicity that  $x = \gamma(0)$ ,  $y = \gamma(a)$  and for all  $0 \leq t \leq a$ ,  $d(x, \gamma(t)) < \text{inj}(\gamma)$ . Let  $\tau_t$  be the unique minimizing geodesic from  $x$  to  $\gamma(t)$ . Since  $d(x, \gamma(t)) < \text{inj}(\gamma)$ ,  $\tau_t$  varies continuously with  $t$ . Let  $\bar{t} = \sup\{t \in [0, a] \mid \tau_s \subset \bar{\Omega} \text{ for } s \leq t\}$ . By the definition of convexity  $\bar{t} > 0$ . By the continuity of  $\tau_t$  we have  $\tau_{\bar{t}} \subset \bar{\Omega}$ . Assume  $\bar{t} \neq a$ . There are two cases. In the first case  $\tau_{\bar{t}} \subset \partial\Omega$ . In this case the convexity of  $\partial\Omega$  makes it clear that  $\tau_{\bar{t}+\delta} \subset \bar{\Omega}$  for small  $\delta$  contradicting the definition of  $\bar{t}$ . In the other case  $\tau_{\bar{t}} \cap \partial\Omega = \{x, \gamma(\bar{t})\}$ . Assume  $\tau_{\bar{t}}: [0, 1] \rightarrow M$  is parametrized proportional to arclength and has length  $l(\bar{t})$ . Let  $\varepsilon > 0$  be less than the  $\varepsilon$  in the definition of the convexity of  $\gamma$  and less than  $l(\bar{t})/3$ . Since  $\tau_{\bar{t}}[\varepsilon/l(\bar{t}), 1 - \varepsilon/l(\bar{t})] \subset \Omega$  we can choose  $\delta > 0$  so that for all  $\bar{t} < t < \bar{t} + \delta$ ,  $\tau_t[\varepsilon/l(t), 1 - \varepsilon/l(t)] \subset \Omega$ . Further for such  $t$   $\tau_t[0, \varepsilon/l(t)] \subset \Omega$  since  $\tau_t(0), \tau_t(\varepsilon/l(t)) \in \bar{\Omega}$ ,  $d(\tau(0), \tau(\varepsilon/l(t))) = \varepsilon$ , and the definition of  $\varepsilon$ . Similarly  $\tau_t[1 - \varepsilon/l(t), 1] \subset \bar{\Omega}$ . Thus we have  $\tau_t \subset \bar{\Omega}$  contradicting the maximality of  $\bar{t}$ . Thus we have shown 2.1.1.

To see 2.1.2, let  $\tau$  be the unique (since  $\varepsilon(\gamma) \leq \frac{1}{2} \text{inj}(\gamma)$ ) geodesic from  $x$  to  $y$ . If  $\tau \not\subset \bar{\Omega}$  then there is an interval  $[t_0, t_1]$  such that  $\tau(t_0, t_1) \subset M - \bar{\Omega}$  while  $\tau(t_0), \tau(t_1) \in \bar{\Omega}$ . But  $L(\tau) < \varepsilon(\gamma) \leq \varepsilon(\tau(t_0), \gamma)$  means  $\tau(t_1) \in \gamma_{\tau(t_0)}$  which contradicts 2.1.1.

For 2.1.3 and 2.1.4 we see that if  $\tau$  is a minimizing curve in  $\bar{\Omega}$  then  $\tau \cap \Omega$  is a geodesic of  $M$ . Further  $\overline{\tau \cap \Omega}$  can only intersect  $\partial\Omega$  at an endpoint of  $\tau$ . To see this we note that  $\tau$  cannot be tangent to  $\partial\Omega$  unless  $\tau \cap \Omega = \emptyset$  (i.e.  $\tau \subset \partial\Omega$ ) as mentioned before, but on the other hand if an angle is made  $\tau$  can be shortened by “cutting the corner”. The convexity of  $\gamma$  allows this “cutting the corner” to happen through curves in  $\bar{\Omega}$  even at non- $C^\infty$  points. 2.1.3 and 2.1.4 now follow easily.

**Lemma 2.2.** *Let  $\gamma$  be convex to  $\Omega$  and have length  $L$ . Assume  $\bar{\Omega}$  is compact and let  $N > L/\text{inj}(\bar{\Omega})$  (also  $N \gg 2$ ). Then if we apply B.C.S.P. with  $N$  breaks to  $\gamma$  the resulting curves  $\gamma_t$  satisfy:*

- (1)  $\gamma_t \subset \bar{\Omega}$ ,
- (2)  $\gamma_t$  is simple and convex to  $\Omega_t \equiv \Omega - \{x \in \gamma_s \mid 0 \leq s \leq t\}$ .

*Proof.* We can assume  $\gamma$  is parametrized proportional to arclength for if not we can homotope the parameter to make it so. Each  $\gamma_t$ ,  $t \in [0, \frac{1}{2}]$  consists of segments of  $\gamma$  and minimizing geodesic segments  $\tau_i^t$  between  $\gamma(i/N)$  and  $\gamma(i/N + 2t/N)$ . Since  $L(\gamma|_{[i/N, i/N+2t/N]}) \leq L/N < \text{inj}(\bar{\Omega}) \leq \text{inj}(\gamma)$ , Lemma 2.1.1 says  $\tau_i^t \subset \bar{\Omega}$ . Further by 2.1.1  $\tau_i^t \cap \gamma = \{\gamma(i/N), \gamma(i/N + 2t/N)\}$  or  $\tau_i^t = \gamma|_{[i/N, i/N+2t/N]}$  (we note that  $\tau$  cannot coincide with the other arc of  $\gamma$  because it is too long since  $N > 2$ ). To see that  $\gamma_t$  is simple we need only see that  $\tau_i^t$  intersects  $\tau_j^t$  only at common endpoints if  $i \neq j$ . Since they are both minimizing geodesics they can intersect at most once. But consider the open set  $\Omega_i^t \subset \Omega$  bounded by  $\tau_i^t \cup (-\gamma|_{[i/N, i/N+2t/N]})$  (this is empty if  $\tau_i^t \subset \gamma = \partial\Omega$ ). If  $\tau_j^t$  intersects  $\tau_i^t$  at interior points of  $\tau_i^t$  and  $\tau_j^t$  then it must intersect transversely and hence enter  $\Omega_i^t$  (in this case  $\Omega_i^t$  of course cannot be empty).  $\tau_j^t$  must thus leave  $\Omega_i^t$  again but since it cannot intersect  $\tau_i^t$  again and does not intersect  $\gamma|_{[i/N, i/N+2t/N]}$  we get a contradiction. Thus  $\gamma_t$  is simple. Since  $\gamma$  is convex to  $\Omega$  and  $\tau_i^t \subset \Omega$  we see that the angles of  $\gamma_t$  are convex to  $\Omega - \bigcup_{s=1}^N \bar{\Omega}_s^t$ . Now  $\bar{\Omega}_i^t = \{x \in \tau_i^s \mid 0 \leq s \leq t\}$  by the convexity of  $\bar{\Omega}_i^t$  and the fact that  $\bar{\Omega}_i^t$  lies inside the injectivity radius of  $\gamma(i/N)$ . (The fact that  $\tau_i^s \subset \bar{\Omega}_i^t$  follows from the proof of Lemma 2.1.1.) Thus we see that  $\gamma_t$  is convex to  $\Omega_t$ .

The proof for  $t \in [\frac{1}{2}, 1]$  follows in exactly the same way since  $\text{inj}(\bar{\Omega}) \leq \text{inj}(\gamma_{1/2})$ .

**Remark.** If  $K$  is a compact set and  $N$  is chosen such that  $N > L/\text{inj}(K)$  then the above lemma holds as long as  $\gamma_t \subset K$  even if  $\Omega$  is not assumed to be compact.

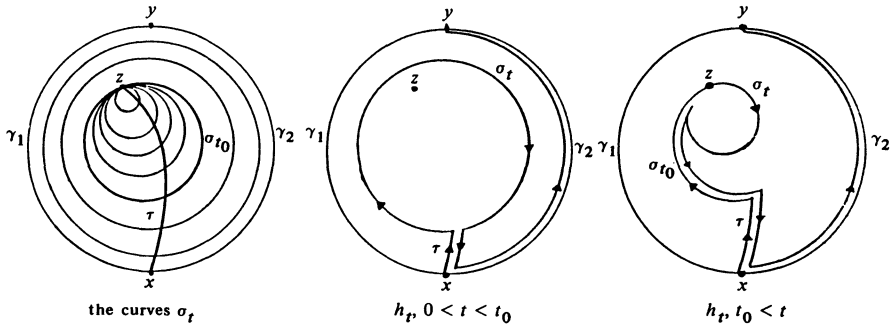
### 3. The Main Lemmas

We now prove the lemma which is the heart of this paper.

**Lemma 3.1.** *Let  $\gamma_1$  and  $\gamma_2$  be two piecewise smooth curves from  $x$  to  $y$  such that  $\gamma_1 \cup -\gamma_2$  forms a simple closed curve which is convex to an open disk  $\Omega$ . Assume further that for every  $z \in \Omega$ ,  $d_{\bar{\Omega}}(x, z) \leq D$ , where  $d_{\bar{\Omega}}$  represents the distance as measured in  $\bar{\Omega}$  and  $D$  is some real number. Then either there is a nontrivial closed geodesic lying in  $\bar{\Omega}$  of length less than or equal to  $L = L(\gamma_1 \cup -\gamma_2)$  or  $\gamma_1$  is homotopic to  $\gamma_2$  through curves from  $x$  to  $y$  lying in  $\bar{\Omega}$  of length  $\leq 3L + 2D$ .*

*Proof.* Assume there is no nontrivial closed geodesic in  $\bar{\Omega}$  of length less than or equal to  $L$ . Applying B.C.S.P. repeatedly to  $\gamma_1 \cup -\gamma_2$  and using Lemma 2.2 we get a homotopy  $\sigma_t$ ,  $t \in [0, 1]$  from  $\sigma_0 = \gamma_1 \cup -\gamma_2$  to a point curve  $\sigma_1$  (say  $\sigma_1 = \{z\}$ ). Further each  $\sigma_t$  is convex to  $\Omega_t$  ( $\Omega_0 = \Omega$ ) with  $\Omega_t \subset \Omega_s$  for  $t > s$ . In particular  $z \in \bar{\Omega}_t$  for all  $t \in [0, 1]$ . Let  $t_0 = \min\{t \mid z \in \partial\Omega_t = \sigma_t\}$ . It is clear that for  $t < t_0$ ,  $z \in \Omega_t$  and for  $t \geq t_0$ ,  $z \in \sigma_t$ . Let  $\tau: [0, 1] \rightarrow M$  be a minimizing path from  $x$  to  $z$  in  $\bar{\Omega}$ .  $\tau$  will be a geodesic of  $M$  if  $t_0 \neq 0$  by Lemma 2.1.3, and we will have no need of  $\tau$  if  $t_0 = 0$ .

We claim that for all  $t < t_0$ ,  $\tau \cap \sigma_t$  is a single point  $z_t$ . The fact that  $x \notin \Omega_t$  and  $z \in \Omega_t$  implies that  $\tau \cap \sigma_t$  is not empty. By Lemma 2.1.3 we need only show that  $\tau(s) \in \sigma_t$  implies  $\tau|_{[s,1]} \subset \bar{\Omega}_t$ . We now fix  $s \in [0, 1]$ . We let  $t_s = \min\{t \mid \tau(s) \in \sigma_t\}$ ,  $t^s = \max\{t \mid \tau(s) \in \sigma_t\}$ , and  $\bar{t} = \sup\{t \mid \tau|_{[s,1]} \subset \bar{\Omega}_t\}$ . To prove the claim we need to show  $\bar{t} \geq \min\{t^s, t_0\}$ . We can assume  $\bar{t} < t_0$ . We see by continuity that  $\tau|_{[s,1]} \subset \bar{\Omega}_{\bar{t}}$  and by the maximality of  $\bar{t}$  that  $\tau|_{[s,1]} \cap \sigma_{\bar{t}} \neq \emptyset$ . By Lemma 2.1.4  $\tau(s) \in \bar{\Omega}_{\bar{t}}$  and hence we see  $t^s \geq \bar{t} \geq t_s$ . Choose  $0 < \varepsilon < \varepsilon(\sigma_{\bar{t}})$  and  $s_0$  such that  $s < s_0 < 1$  and  $L(\tau|_{[s_0,1]}) < \varepsilon$ . Since  $\tau|_{[s_0,1]} \subset \bar{\Omega}_{\bar{t}}$  there is a  $\delta > 0$  such that  $\tau|_{[s_0,1]} \subset \Omega_t$  and  $\varepsilon(\sigma_t) > \varepsilon$  for all  $t$  with  $\bar{t} \leq t < \bar{t} + \delta$ . Thus for all  $t$ ,  $\bar{t} \leq t \leq \min\{t^s, t_0, \bar{t} + \delta\}$  we have  $\tau|_{[s_0,1]} \subset \bar{\Omega}_t$  by Lemma 2.1.2, and hence  $\tau|_{[s,1]} \subset \bar{\Omega}_t$ . Thus the maximality of  $\bar{t}$  forces  $\bar{t} = \min\{t^s, t_0\}$  and the claim follows.





We are now ready to define the homotopy. First homotope  $\gamma_1$  to  $\gamma_1 \cup -\gamma_2 \cup \gamma_2$  through curves of length  $< 2L$ . Now for each  $t \in [0, t_0)$  let  $s(t)$  be the unique value of  $s$  such that  $\tau(s(t)) \in \sigma_t$ . Let  $s(t_0) = \sup\{s(t) \mid t \in [0, t_0)\}$ . We homotope  $\gamma_1 \cup -\gamma_2 \cup \gamma_2$  to  $\tau|_{[0, s(t_0)]} \cup \sigma_{t_0} \cup -\tau|_{[0, s(t_0)]} \cup \gamma_2$  through the curves  $h_t = \tau|_{[0, s(t)]} \cup \sigma_t \cup -\tau|_{[0, s(t)]} \cup \gamma_2$ , where  $\sigma_t$  represents going once around  $\sigma_t$  starting and ending at  $\tau(s(t))$ . The length of  $h_t$  is less than  $D + L + D + L = 2L + 2D$ . Let  $\tau_{t_0} = \tau|_{[0, s(t_0)]}$  and  $\bar{\sigma}_{t_0}$  represents the shortest arc of  $\sigma_0$  from  $\tau(s(t_0))$  to  $z$ . We now homotope  $\tau_{t_0} \cup \sigma_{t_0} \cup -\tau_{t_0} \cup \gamma_2$  to  $\tau_{t_0} \cup \bar{\sigma}_{t_0} \cup \sigma_{t_0} \cup -\bar{\sigma}_{t_0} \cup -\tau_{t_0} \cup \gamma_2$ , where here  $\sigma_{t_0}$  represents going once around  $\sigma_{t_0}$  starting and ending at  $z$ , through curves of length  $< D + L/2 + L + L/2 + D + L = 3L + 2D$ . This curve is in turn homotopic to  $\tau_{t_0} \cup \bar{\sigma}_{t_0} \cup -\bar{\sigma}_{t_0} \cup -\tau_{t_0} \cup \gamma_2$  via the curves  $h_t = \tau_{t_0} \cup \bar{\sigma}_{t_0} \cup \sigma_t \cup -\bar{\sigma}_{t_0} \cup -\tau_{t_0} \cup \gamma_2$  for  $t \in [t_0, 1]$ , whose lengths are less than  $3L + 2D$ . This last curve is clearly homotopic to  $\gamma_2$  through curves of length less than  $2L + 2D$  and the lemma follows.

**Remark 1.**  $3L + 2D$  is not optimal. One could probably improve this to  $\frac{3}{2}L + 2D$  with a little work, but examples show one cannot expect to do much better. As neither estimate leads to sharp answers in the theorems we will not worry about it.

**Remark 2.** One should note (we will use this fact later) that the homotopy defined above defines a map from the disk  $D^2$  to  $\bar{\Omega}$  of local degree  $\pm 1$  since the generic point will have a single preimage.

We now come to the lemma in which the area  $A$  of the manifold enters. It enters in the (coarea) formula

$$A \geq \int_a^b L(S(x, t)) dt$$

where  $\infty \geq b \geq a \geq 0$ ,  $x$  is a fixed point in  $M$ , and  $S(x, t) = \{y \in M \mid d(x, y) = t\}$ , i.e.  $S(x, t)$  is the ‘‘circle’’ of radius  $t$  centered at  $x$ . In general  $S(x, t)$  need not be very nice, but for generic  $t$  (i.e. for all but a closed set of measure 0)  $S(x, t)$  is a piecewise smooth disjoint union of Jordan curves. This was shown by Hartman [14, Proposition 6.1] who generalized results of Fiala [11] to the differentiable category (Fiala considered analytic metrics). The above coarea formula can be found, for example, in equations 6.30 and 6.31 of [14].

**Lemma 3.2.** *Let  $M$  be a complete oriented surface of finite area  $A$ . Let  $x, y, z \in M$  with  $\tau_y^x, \tau_z^y$ , and  $\tau_x^z$  minimizing geodesics connecting the respective points. Then:*

(1) *If  $w \in \tau_y^x$  is such that  $d(w, x) > \sqrt{A}$  and  $d(w, y) > \sqrt{A}$ , then there is a closed curve through  $w$  which is essential in  $M - \{x, y\}$  and has length  $\leq 2\sqrt{A}$ .*

(2) Let  $d_x = d(x, \tau_z^y)$  (similarly for  $d_y$  and  $d_z$ ). If  $d_x > \sqrt{2A}$ ,  $d_y > \sqrt{2A}$  and  $d_z > \sqrt{2A}$ , then there is a nontrivial closed geodesic of length  $\leq \sqrt{8A}$ .

(3) If  $M$  is compact,  $d(x, y) = D(M)$ , the diameter of  $M$ , and  $d_z > 2\sqrt{2A}$ , then there is a nontrivial closed geodesic of length  $\leq \sqrt{8A}$ .

(4) In the case where  $M$  is diffeomorphic to  $S^1 \times \mathbf{R}^1$  and  $\gamma$  is a line in  $M$  ( $\gamma$  minimizes distance between any two of its points) and for some  $z$ ,  $d(z, \gamma) > \sqrt{2A}$  then there is a nontrivial closed geodesic of length  $\leq \sqrt{8A}$ .

*Proof.* We begin by noting that for a piecewise smooth simple closed curve  $\sigma$  on a complete surface  $M$ , in particular for a component of  $S(x, t)$  for fixed  $x$  and generic  $t$ , either  $\sigma$  is essential in  $M$  or  $\sigma$  splits  $M$  into two pieces  $M_1$  and  $M_2$ . In the latter case  $\sigma$  is essential in  $M - \{x_1, x_2\}$  for  $x_1 \in M_1$  and  $x_2 \in M_2$ . One sees this by assuming that  $\sigma$  does not split  $M$ . In which case one can create a closed curve  $\tau$  intersecting  $\sigma$  transversely exactly once. Hence the intersection number modulo 2 of  $\sigma$  with  $\tau$  is nonzero and  $\sigma$  is essential. In the case where  $\sigma$  divides  $M$ , choose  $\tau$  to be a curve from  $x_1$  to  $x_2$  intersecting  $\sigma$  once transversely. The same argument gives  $\sigma$  essential in  $M - \{x_1, x_2\}$ . It is easy to see that if  $M_i$  is noncompact  $x_i$  can be taken to be  $\infty$  (i.e., not included in the removed set).

Let  $\tau: [0, 1] \rightarrow M$  be  $\tau_y^x$  parametrized by arclength with  $\tau(0) = x$ ,  $\tau(L) = y$ , and  $\tau(t_0) = w$ . By assumption  $t_0 > \sqrt{A}$  and  $L - t_0 > \sqrt{A}$ . Since

$$A \geq \int_{t_0 - \sqrt{A}}^{t_0 + \sqrt{A}} L(S(x, t)) dt \quad \text{and} \quad \int_{t_0 - \sqrt{A}}^{t_0 + \sqrt{A}} 2\sqrt{A} - 2|t - t_0| dt = 2A$$

there is a generic  $t \in [t_0 - \sqrt{A}, t_0 + \sqrt{A}]$  such that  $L(S(x, t)) \leq 2\sqrt{A} - 2|t - t_0|$ . Let  $\sigma$  be the component of  $S(x, t)$  through  $\tau(t)$ .  $\sigma$  is a simple closed curve (by the genericity of  $t$ ) with  $L(\sigma) \leq 2\sqrt{A} - 2|t - t_0|$  and  $\sigma$  is essential in  $M - \{x, y\}$ . That  $\sigma$  is essential in  $M - \{x, y\}$  follows since either  $\sigma$  is essential in  $M$  (hence in  $M - \{x, y\}$ ) or it splits  $M$  into two pieces with  $x$  in one and  $y$  in the other since  $\tau$  intersects  $\sigma$  transversely exactly once. Thus the curve  $\tau|_{[t_0, t]} \cup \sigma \cup -\tau|_{[t_0, t]}$  is a closed curve through  $w$  of length  $\leq 2\sqrt{A}$  and essential in  $M - \{x, y\}$ .

The first step in the proof of (2) is to note that  $d_x + d_y > d(x, y) = L(\tau_y^x) \equiv L$ . This is proved by adding the four natural triangle inequalities involving  $d_x$  and  $d_y$  (for example if  $q \in \tau_y^z$  is the closest point to  $x$  on  $\tau_y^z$ , i.e.  $d(x, q) = d_x$ , then two of the triangle inequalities are  $d_x + d(q, z) \geq d(x, z)$  and  $d_x + d(q, y) \geq d(x, y)$ ). You get strict inequality since the four inequalities cannot be simultaneously equalities (since  $z \notin \tau_y^x$ ). On the other hand  $L = d(x, y) \geq \max\{d_x, d_y\} > \sqrt{2A}$ .

Let  $\tau: [0, L] \rightarrow M$  be  $\tau_y^x$  with the arclength parameter  $t$ . Choose  $t_0 \in [0, L]$  such that  $\sqrt{A}/2 < t_0 < d_x$  and  $\sqrt{A}/2 < L - t_0 < d_y$ , which can be done

since  $d_x + d_y > L$  and  $d_x, d_y \geq \sqrt{2A} > \sqrt{A/2}$ . Now  $B(x, t_0) \cap B(y, L - t_0)$  has measure 0 so

$$\begin{aligned} A &\geq \text{Area}(B(x, t_0)) + \text{Area}(B(y, L - t_0)) \\ &\geq \int_0^{\sqrt{A/2}} L(S(x, t_0 - s)) ds + \int_0^{\sqrt{A/2}} L(S(y, L - t_0 - s)) ds. \end{aligned}$$

Hence as before there is a generic  $s \in [0, \sqrt{A/2}]$  such that

$$L(S(x, t_0 - s)) + L(S(y, L - t_0 - s)) \leq 2\sqrt{2}\sqrt{A} - 4s$$

and both  $S(x, t_0 - s)$  and  $S(y, L - t_0 - s)$  are disjoint unions of simple closed curves. Let  $\sigma_1$  be the component of  $S(x, t_0 - s)$  through  $\tau(t_0 - s)$  and  $\sigma_2$  the component of  $S(y, L - t_0 - s)$  through  $\tau(t_0 + s)$ . Since  $t_0 < d_x$ ,  $\sigma_1 \cap \tau_y^z = \emptyset$ . If  $\sigma_1$  does not separate  $M$  then applying B.C.S.P. repeatedly yields the desired nontrivial (in fact essential) closed geodesic of length  $\leq \sqrt{8A}$ . So we can assume  $\sigma_1$  separates  $M$  which must have  $x$  on one side and  $y$  and  $z$  on the other.  $\sigma_1$  intersects both  $\tau_z^x$  and  $\tau_y^x$  transversely (in fact perpendicularly) in one point. Similarly we can assume  $\sigma_2$  intersects  $\tau_z^y$  and  $\tau_y^x$  transversely once. Define  $\sigma$  to be  $\sigma_1 \cup \tau|_{[t_0-s, t_0+s]} \cup \sigma_2 \cup -\tau|_{[t_0-s, t_0+s]}$ . We see that  $L(\sigma) \leq \sqrt{8A}$ . We make sure to choose the orientation of  $\sigma_1$  and  $\sigma_2$  so that  $\sigma$  has the form of a figure 8 around  $x$  and  $y$ , that is so that the oriented intersection number of  $\sigma$  with  $\tau$  (say in  $M - \{x, y\}$ ) is  $+2$  rather than 0. Applying B.C.S.P. repeatedly to  $\sigma$  leads either to a closed geodesic of length  $\leq L(\sigma) \leq \sqrt{8A}$  or to a point curve. But it cannot lead to a point curve for if it did some curve  $\sigma_{s_0}$  in the homotopy would have to pass through a vertex ( $x$ ,  $y$  or  $z$ ) while still intersecting the opposite geodesic ( $\tau_z^y$ ,  $\tau_z^x$ , or  $\tau_y^x$ ) which we see by intersection number arguments. Now the fact that  $L(\sigma_{s_0}) \leq \sqrt{8A}$  and  $d_x$ ,  $d_y$  and  $d_z > \frac{1}{2}\sqrt{8A}$  yields the desired contradiction.

Part (3) follows directly from part (2) and triangle inequalities. Let  $w \in \tau_z^y$  be such that  $d(x, w) = d_x$ . Then  $d_x + d(w, y) \geq d(x, y) = D \geq d(z, y)$  and  $d_x + d(w, z) \geq d(x, z) \geq 2\sqrt{2A}$ . Adding the above gives  $d_x > \sqrt{2A}$ . Similarly  $d_y > \sqrt{2A}$ .

Part (4) also follows from part (2). Choose  $w \in \gamma$  such that  $d(z, w) = d(z, \gamma)$  and choose  $x$  and  $y$  on  $\gamma$  such that  $d(w, y) > d(z, w) + \sqrt{2A}$ ,  $d(w, x) > d(z, w) + \sqrt{2A}$  and  $w$  is between  $x$  and  $y$ , i.e.  $d(x, y) = d(x, w) + d(w, y)$ . We therefore have  $d(x, w) + d(w, y) = d(x, y) \leq d_x + d(z, y) \leq d_x + d(z, w) + d(w, y)$ . Hence  $d_x \geq d(x, w) - d(z, w) > \sqrt{2A}$ . Similarly  $d_y > \sqrt{2A}$ . Since we have by assumption  $d_z > \sqrt{2A}$  part (2) yields part (4).

We now study further the case of Lemma 3.2.1. We will consider the case of a geodesic segment  $\tau$  in  $M$ ; we are interested in three cases in particular. The first is when  $M$  is diffeomorphic to  $S^2$  and  $\tau$  is a minimizing geodesic. The

second is when  $M$  is diffeomorphic to  $\mathbf{R}^2$  and  $\tau$  is a ray (i.e.  $\tau: [0, \infty) \rightarrow M$  and  $\tau$  is a minimizing geodesic between any two points on it). The third is when  $M$  is a cylinder  $S^1 \times \mathbf{R}^1$  and  $\tau$  is a line (i.e.  $\tau: (+\infty, \infty) \rightarrow M$  and it minimizes between any two points on it). We will speak of  $\tau$  as a minimizing geodesic from  $x$  to  $y$  but either one or both of  $x$  and  $y$  will represent  $\infty$  in the second and third cases above.

**Lemma 3.3.** *Let  $M$  be one of the three cases above and  $\tau$  a minimizing geodesic from  $x$  to  $y$  (as discussed above). Let  $w \in \tau$  be such that  $d(w, y) > \sqrt{A}$  and  $d(w, x) > \sqrt{A}$ . Then there is a shortest closed curve  $\gamma$  through  $w$  which is essential in  $M - \{x, y\}$ . If  $\gamma_1$  and  $\gamma_2$  are two such shortest curves we have:*

(1)  $\gamma_i: [0, l_w] \rightarrow M$  is a simple closed geodesic loop (not necessarily smooth at  $w$ ) at  $w$  of length  $l_w \leq 2\sqrt{A}$ .

(2)  $\gamma_i \cap \tau = \{w\}$  and the vectors  $\gamma'_i(0)$  and  $-\gamma'_i(l_w)$  lie on opposite sides of  $\tau$ .

(3)  $\gamma_i \cap \gamma_j = \{w\}$  or  $\gamma_i \equiv \gamma_j$ .

(4) Assume further that both  $\gamma_i$  separate  $M$  into two pieces. Let  $\Omega_i$  be the component of  $M - \gamma_i$  such that  $\gamma_i$  is convex to  $\Omega_i$  (this must be true for at least one component for  $\gamma_i$  has but one angle). Then either  $\Omega_i \cap \Omega_j = \emptyset$ ,  $\Omega_i \subset \Omega_j$ , or  $\Omega_j \subset \Omega_i$ .

*Proof.* By Lemma 3.2.1 there are such short curves of length  $\leq 2\sqrt{A}$ . Since  $w$  is further from  $x$  or  $y$  than  $\sqrt{A}$  a shortest such curve cannot pass through  $x$  or  $y$  hence must be a geodesic loop  $\gamma_w$ . The above holds in all cases even though  $x$  or  $y$  may be  $\infty$ , by choosing points close to  $\infty$  (after fixing  $w$ ) and then applying Lemma 3.2.1. We now see that  $\gamma_w \cap \tau = \{w\}$ , for if not we could replace one arc of  $\gamma_w$  with a segment of  $\tau$  reducing the length (since  $\tau$  minimizes from  $x$  to  $y$ ) and staying essential in  $M - \{x, y\}$  (for the correct choice of arc to replace). Now if both vectors  $\gamma'_w(0)$  and  $-\gamma'_w(l_w)$  lie on the same side of  $\tau$  then the fact that  $\tau \cap \gamma_w = \{w\}$  implies that  $\gamma_w$  can be homotoped to miss  $\tau$ . Since  $M - \tau$  is simply connected and  $M - \{x, y\} \supset M - \tau$  we see  $\gamma_w$  cannot be essential in  $M - \{x, y\}$ . Hence  $\gamma'_w(0)$  and  $-\gamma'_w(l_w)$  lie on opposite sides of  $\tau$  as claimed. Now assume  $\gamma_w$  was not simple. By throwing away part of  $\gamma_w$  one can construct a closed curve  $\bar{\gamma}_w$  through  $w$  which is shorter than  $\gamma_w$  but whose intersection with  $\tau$  is the same as  $\gamma_w$ 's, that is  $\bar{\gamma}_w$  intersects  $\tau$  transversely in one point. Thus  $\bar{\gamma}_w$  is essential in  $M - \{x, y\}$ . This contradicts the minimality of  $\gamma_w$  and hence  $\gamma_w$  is simple. This proves (1) and (2).

Assume that  $\gamma_w$  and  $\bar{\gamma}_w$  are two such loops. Orient them so that  $\gamma'_w(0)$  and  $\bar{\gamma}'_w(0)$  lie on the same side of  $\tau$ . If  $\gamma_w$  and  $\bar{\gamma}_w$  intersect then one can construct two closed curves  $\tau_1$  and  $\tau_2$  through  $w$  as follows:  $\tau_1$  starts at  $w$  follows  $\gamma_w$  to the point of intersection and then follows  $\bar{\gamma}_w$  back to  $w$ .  $\tau_2$  does the opposite.

Both  $\tau_1$  and  $\tau_2$  intersect  $\tau$  only at  $w$  and are transverse there and hence are essential in  $M - \{x, y\}$ . Now  $L(\tau_1) + L(\tau_2) = L(\gamma_w) + L(\bar{\gamma}_w) = 2l_w$  so one of  $\tau_1$  or  $\tau_2$ , say  $\tau_i$ , satisfies  $L(\tau_i) \leq l_w$ . But  $\tau_i$  is not smooth at the intersection point of  $\gamma_w$  with  $\bar{\gamma}_w$  and hence can be shortened contradicting the definition of  $l_w$ . Thus (3) follows.

From (3) we see that  $\gamma_i - \{w\} \subset \Omega_j$  or  $\gamma_i - \{w\} \subset M - \bar{\Omega}_j$ . We can tell which of these two cases happens by looking near  $w$ . Look at the angle between  $\gamma'_j(0)$  and  $-\gamma'_j(l_w)$  (the one that is less than  $\pi$  or in the case of the angle equal to  $\pi$  the one containing  $\Omega_j$ ). If both  $\gamma'_i(0)$  and  $-\gamma'_i(l_w)$  lie in this angle then  $\gamma_i - \{w\} \subset \Omega_j$ . If not, they must both be outside since  $\gamma_i - \{w\} \subset M - \bar{\Omega}_j$ . It is not hard to see that if  $\gamma_i - \{w\} \subset M - \bar{\Omega}_j$  and  $\gamma_j - \{w\} \subset M - \bar{\Omega}_i$  then  $\Omega_i \cap \Omega_j = \emptyset$ . Further if  $\gamma_i - \{w\} \subset \Omega_j$  and  $\gamma_j - \{w\} \subset M - \bar{\Omega}_i$  then  $\Omega_i \subset \Omega_j$ . Thus we need only consider the case  $\gamma_j - \{w\} \subset \Omega_i$  and  $\gamma_i - \{w\} \subset \Omega_j$ . But this cannot happen since it says  $\gamma'_i(0)$  and  $-\gamma'_i(l_w)$  lie between  $\gamma'_j(0)$  and  $-\gamma'_j(l_w)$  and vice versa.

This concludes the proof of the lemma.

#### 4. The Main Theorems

In this section we consider the case where  $M$  is diffeomorphic to  $S^2$ .

**Theorem 4.1.** *Let  $M$  be a riemannian manifold, diffeomorphic to  $S^2$ , of diameter  $D$ . Then  $L \leq 9D$ , where  $L$  is the length of the shortest nontrivial closed geodesic on  $M$ .*

*Proof.* Choose  $x, y \in M$  such that  $d(x, y) = D$ . Let  $\mathfrak{A} = \{\tau \mid \tau \text{ is a minimizing geodesic from } x \text{ to } y\}$ . Berger's lemma (see [8, p. 106]) says that for every  $V \in T_x M$  there is a  $\tau \in \mathfrak{A}$  such that  $\langle V, \tau'(0) \rangle \geq 0$ . Similarly for every  $W \in T_y M$  there is a  $\tau \in \mathfrak{A}$  such that  $\langle W, -\tau'(D) \rangle \geq 0$ . Thus we can pick a finite number of distinct geodesics  $\tau_1, \tau_2, \dots, \tau_n \in \mathfrak{A}$  (in fact it is not hard to see that  $n$  can be taken to satisfy  $2 \leq n \leq 4$ ) such that  $\{\tau_1, \tau_2, \dots, \tau_n\}$  satisfies the same property as  $\mathfrak{A}$ . We can assume  $n \neq 2$  for if so  $\tau_1 \cup -\tau_2$  is a closed geodesic of length  $2D$ . We order them so that the  $\tau'_i(0)$  come in order counterclockwise from  $\tau'_1(0)$ . Since  $\tau_i$  minimizes length  $\tau_i \cap \tau_j = \{x, y\}$  for  $i \neq j$ . Since for every  $V \in T_x M$  there is an  $i$  such that  $\langle V, \tau'_i(0) \rangle \geq 0$  we see that  $\sphericalangle(\tau'_i(0), \tau'_{i+1}(0)) \leq \pi$ , where the angle is measured in the counterclockwise sense (here  $n + 1$  is the same as 1). Similarly  $\sphericalangle(-\tau'_{i+1}(D), -\tau'_i(D)) \leq \pi$ . In particular  $\tau_i \cup -\tau_{i+1}$  is a simple closed curve which is convex to the domain  $\Omega_i$  lying between them (in the obvious way). If  $z \in \Omega_i$  then the minimizing geodesic from  $z$  to  $y$  must lie in  $\bar{\Omega}_i$  by the minimality of  $\tau_i$  and  $\tau_{i+1}$ . Thus by Lemma 3.1 either there is a closed geodesic of length  $\leq 2D$  or  $-\tau_i$  is

homotopic to  $-\tau_{i+1}$  through curves of length  $\leq 3L + 2D = 8D$  lying in  $\Omega_i$ . We now describe a short homotopy from the point curve  $\{x\}$  to the point curve  $\{y\}$ :

$$\begin{aligned} \{x\} &\sim (\tau_1 \cup -\tau_1) \sim (\tau_1 \cup -\tau_2) \sim \cdots \\ &\sim (\tau_1 \cup -\tau_n) \sim (\tau_1 \cup -\tau_1) \sim \{y\} \end{aligned}$$

where the homotopy from  $-\tau_i$  to  $-\tau_{i+1}$  is through curves in  $\bar{\Omega}_i$  as in Lemma 3.1. Remark 2 following Lemma 3.1 shows that the induced map from  $S^2$  to  $S^2$  has degree 1. Hence by Birkhoff's idea there is a nontrivial closed geodesic of length less than the length of the longest curve in the homotopy, i.e., less than  $9D$ .

**Remark.**  $9D$  is clearly not the best constant for this theorem. In fact by improving Lemma 3.1 as suggested in the remark following it one could improve this to  $6D$  but this is also unlikely to be sharp.

**Theorem 4.2.** *Let  $M$  be a riemannian manifold, diffeomorphic to  $S^2$ , of area  $A$ . Then  $L \leq 31\sqrt{A}$ , where  $L$  is the length of the shortest nontrivial closed geodesic.*

*Proof.* By Theorem 4.1 we can assume  $D > \frac{31}{9}\sqrt{A} > 2\sqrt{A}$ . Let  $x, y \in M$  such that  $d(x, y) = D$ . Let  $\tau$  be a minimizing geodesic from  $x$  to  $y$  parametrized by arclength  $\tau: [0, D] \rightarrow M$ . For each  $t \in (\sqrt{A}, D - \sqrt{A})$  there is a simple geodesic loop  $\gamma_t$  (not necessarily unique) through  $\tau(t)$  as in Lemma 3.3.  $\gamma_t$  separates  $M$  into pieces  $\Omega_x, \Omega_y$  with  $x \in \Omega_x$  and  $y \in \Omega_y$ , and, since it is a geodesic loop, is convex to  $\Omega_x$  or  $\Omega_y$ . For each  $t \in (\sqrt{A}, D - \sqrt{A})$  we say  $t \in S_x$  if there is a  $\gamma_t$  as in Lemma 3.3 with  $\gamma_t$  convex to  $\Omega_x$ . Similarly define  $S_y$ . It is easy to see that both  $S_x$  and  $S_y$  are closed subsets of  $(\sqrt{A}, D - \sqrt{A})$ . It follows from the fact that a sequence of geodesic loops has a convergent subsequence to a geodesic loop (the resulting loop can't pass through  $x$  or  $y$  for length reasons). Thus either  $S_x \cap S_y \neq \emptyset$  or one of  $S_x$  or  $S_y$  is empty.

We first consider the case where  $S_x \cap S_y \neq \emptyset$ . Let  $t_0 \in S_x \cap S_y$ . This means there are geodesic loops  $\gamma_x$  and  $\gamma_y$  through  $\tau_y^x(t_0)$  with  $\gamma_x$  convex to  $\Omega_x$  and  $\gamma_y$  convex to  $\Omega_y$ , where  $\Omega_x$  and  $\Omega_y$  are open with  $x \in \Omega_x$  and  $y \in \Omega_y$ . If  $\gamma_x = \gamma_y$  then it is a closed geodesic of length  $\leq 2\sqrt{A}$  (by Lemma 3.3) and the theorem follows. If not then Lemma 3.3 tells us that  $\gamma_x \cap \gamma_y = \{\tau_y^x(t_0)\}$  and  $\Omega_x \cap \Omega_y = \emptyset$  since  $x \notin \Omega_y$  and  $y \notin \Omega_x$ . Let  $\Omega = M - (\bar{\Omega}_x \cup \bar{\Omega}_y)$ . Then  $\partial\Omega = \gamma_x \cup -\gamma_y$ , which is convex to  $\Omega$  by Lemma 3.3.2. Assuming there are no closed geodesics of length  $\leq 2\sqrt{A}$  in  $\Omega_x$  repeated applications of B.C.S.P. and Lemma 2.2 show  $\gamma_x$  is homotopic to a point curve through curves in  $\Omega_x$  of length  $\leq 2\sqrt{A}$ . Similarly  $\gamma_y$  is homotopic to a point curve through curves in  $\Omega_y$  of length  $\leq 2\sqrt{A}$ . Further, if there is no closed geodesic in  $\Omega$  of length

$\leq 4\sqrt{A}$ ,  $\gamma_x$  is homotopic to  $\gamma_y$  through curves in  $\Omega$  of length  $\leq 12\sqrt{A} + 2\bar{D}$  by Lemma 3.1, where  $\bar{D} = \max\{d_\Omega(z, \tau_y^x(t_0)) \mid z \in \Omega\}$ . (As stated one cannot apply Lemma 3.1 directly to  $\gamma_x \cup -\gamma_y$  since it is not strictly speaking a simple curve. But after the first application of B.C.S.P. it becomes simple and the argument carries through.)

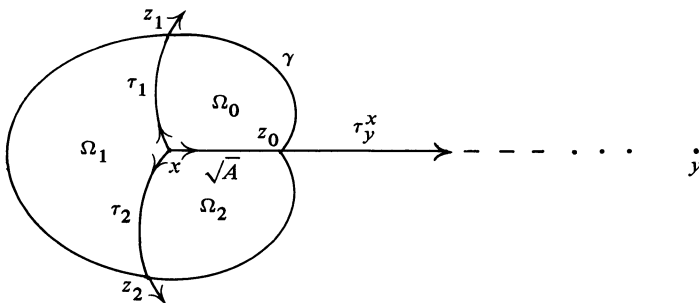
By putting these three homotopies together we get a one parameter family of curves from a point curve to a point curve such that the induced map from  $S^2$  to  $S^2$  has degree 1 (see the remark following 3.1). Thus the Birkhoff method yields a nontrivial closed geodesic of length  $\leq 12\sqrt{A} + 2\bar{D}$ .

We thus need to bound  $\bar{D}$ . Let  $z \in \Omega$ . By Lemma 3.2.3 we may assume that  $d(z, \tau_y^x) \leq 2\sqrt{2A}$ . Since  $\tau_y^x \cap \gamma_x = \{\tau_y^x(t_0)\}$ ,  $\tau_y^x \cap \gamma_y = \{\tau_y^x(t_0)\}$ ,  $x \in \Omega_x$ , and  $y \in \Omega_y$  it is easy to see that  $\tau_y^x \cap \Omega = \emptyset$ . Thus  $d(z, \gamma_x \cup -\gamma_y) \leq 2\sqrt{2A}$  and  $\bar{D} \leq (2\sqrt{2} + 1)\sqrt{A}$ .

Thus if  $S_x \cap S_y \neq \emptyset$  there is a nontrivial closed geodesic of length  $\leq (14 + 4\sqrt{2})\sqrt{A} < 31\sqrt{A}$ .

We now consider the case where one of  $S_x$  or  $S_y$  is empty. We can assume  $S_x = \emptyset$  and  $S_y = (\sqrt{A}, D - \sqrt{A})$ . In fact there will be a simple closed geodesic loop  $\gamma$  at  $\tau(\sqrt{A})$  which is convex to  $\Omega_y$ , is essential in  $M - \{x, y\}$ , and has length  $\leq 2\sqrt{A}$ . We find  $\gamma$  as a limit (of a subsequence if necessary) of minimal loops at  $\tau(t_i)$  for  $t_i \searrow \sqrt{A}$ . All of the properties follow immediately once we see that  $\gamma \subset M - \{x, y\}$ .  $y \notin \gamma$  by length considerations. If  $x \in \gamma$  then  $\gamma$  must have length  $= 2\sqrt{A}$  and both arcs of  $\gamma$  are minimizing geodesics from  $\tau(\sqrt{A})$  to  $x$  but this cannot be as  $\tau|_{[0, \sqrt{A}]}$  is the unique minimizing geodesic from  $\tau(0) = x$  to  $\tau(\sqrt{A})$  (since it minimizes past  $\sqrt{A}$ ).

Since  $x$  is at maximum distance from  $y$  we can use Berger's lemma (see [8, p. 106]) to find minimizing geodesics  $\tau_1$  and  $\tau_2$  from  $x$  to  $y$  such that  $\tau_y^x(0)$ ,  $\tau_1'(0)$ , and  $\tau_2'(0)$  do not lie in an open half plane. (It may happen that there is only one other geodesic  $\tau_1$  in which case  $\tau_1'(0) = -\tau_y^x(0)$ . The arguments that follow work equally well in this case but we will make them in the case where  $\tau_1$  and  $\tau_2$  both exist.)



The geodesic loop  $\gamma$  intersects each of  $\tau_y^x$ ,  $\tau_1$ , and  $\tau_2$  in one point  $z_0 = \tau_y^x\sqrt{A}$ ,  $z_1$ , and  $z_2$  respectively (see figure). This is true since if not we would be able to replace a segment of  $\gamma$  with a segment of  $\tau_i$  decreasing the length and leaving the new curve essential in  $M - \{x, y\}$ . This would contradict the minimality of the length of the geodesic loops which converge to  $\gamma$ .

We will use the notation  $\overline{xz_i}$  and  $\overline{z_i z_j}$  to represent the geodesic segments (in figure) between the corresponding points. Note that  $\overline{z_i z_j}$  is the appropriate segment of  $\gamma$  and not necessarily a minimizing geodesic.

We know  $L(xz_0) = \sqrt{A}$ ,  $L(z_i z_j) \leq 2\sqrt{A}$  and  $L(xz_i) \leq 2\sqrt{A}$  for  $i = 1, 2$  since  $L(\gamma) \leq 2\sqrt{A}$ . The geodesic triangles  $xz_0 z_1$ ,  $xz_1 z_2$ , and  $xz_2 z_0$  are convex to the domain  $\Omega_0$ ,  $\Omega_1$ , and  $\Omega_2$  respectively (see figure). We can assume by Lemma 3.2.3 that for every  $z \in M$ ,  $d(z, \tau_y^x) \leq 2\sqrt{2A}$ . If  $z \in \Omega_i$  then  $d_{\partial\Omega_i}(z, x) \leq (2\sqrt{2} + 2)\sqrt{A}$ , since the minimizing geodesic from  $z$  to  $\tau_y^x$  must hit  $\partial\Omega_i$  and any point on  $\partial\Omega_i$  is connectable to  $x$  along  $\partial\Omega_i$  through curves of total length  $\leq 2\sqrt{A}$ . (The last part of the above can be seen as follows: Starting at  $w \in \gamma$  one can trace along the short loop of  $\gamma$  to  $z_0$  (length  $\leq \sqrt{A}$ ) then follow  $\overline{xz_0}$  back to  $x$  (length  $= \sqrt{A}$ ). This curve may leave  $\partial\Omega_i$  but the curve that starts like this until it hits a  $z_i$  and then runs to  $x$  along  $\tau_i$  must be even shorter. Of course if  $w \in \partial\Omega_i - \gamma$  one simply follows a  $\tau$  to  $x$ .)

We now create a homotopy from the point curve  $\{x\}$  to  $\gamma$  using Lemma 3.1 repeatedly as follows:

$$\begin{aligned} \{x\} &\sim (\overline{xz_0}) \cup (\overline{z_0 x}) \sim (\overline{xz_0}) \cup (\overline{z_0 z_1}) \cup (\overline{z_1 x}) \\ &\sim (\overline{xz_0}) \cup (\overline{z_0 z_1}) \cup (\overline{z_1 z_2}) \cup (\overline{z_2 x}) \\ &\sim (\overline{xz_0}) \cup (\overline{z_0 z_1}) \cup (\overline{z_1 z_2}) \cup (\overline{z_2 z_0}) \cup (\overline{z_0 x}) \sim \gamma. \end{aligned}$$

The longest curves in this homotopy have length  $\leq \sqrt{A} + 2\sqrt{A} + 3(6\sqrt{A}) + 2(2\sqrt{2} + 2)\sqrt{A} = (25 + 4\sqrt{2})\sqrt{A} < 31\sqrt{A}$ . Since  $\gamma$  is convex to  $\Omega_y$ , we may assume as usual that  $\gamma$  is homotopic to a point curve through curves in  $\Omega_y$  of length  $\leq 2\sqrt{A}$ . Combining these homotopies the Birkhoff idea (since once again the induced map from  $S^2$  to  $S^2$  has degree 1) yields a nontrivial closed geodesic of length  $\leq 31\sqrt{A}$ . The theorem follows.

## 5. Noncompact Surfaces of Finite Area

It is known (see [20] and [2]) that every complete surface of finite area has closed geodesics (in fact infinitely many). In this section, using ideas developed in previous sections, we show:



**Theorem 5.1.** *There is a constant  $c$  such that if  $M$  is a complete surface (without boundary) of area  $A$  then  $c\sqrt{A} \geq L$ , where  $L$  is the length of the shortest closed geodesic.*

*Proof.* By taking oriented double covers we may assume  $M$  is orientable. As discussed in the introduction, previous work had reduced the compact case to the case of  $S^2$ . Since Theorem 4.2 takes care of the  $S^2$  case we may assume  $M$  is not compact. By Theorem 4.4A of [12] we may assume that  $M$  is diffeomorphic to  $S^2 - \{\text{points}\}$ . We treat this as three cases. Case 1 is when  $M$  has at least three ends. In Case 2,  $M$  is diffeomorphic to  $S^1 \times \mathbf{R}^1$  and in Case 3,  $M$  is diffeomorphic to  $\mathbf{R}^2$ .

*Case 1.*  $M$  has at least three ends.

Choose  $x_0 \in M$  and  $R_0 > 0$  so large that three of the ends of  $M - B(x_0, R_0)$  are pairwise disconnected in  $M - B(x_0, R_0)$ . Choose  $x_1, x_2$ , and  $x_3$  one in each end such that  $d(x_0, x_i) = 2R_0 + \sqrt{2A}$ . Let  $\tau_{x_i}^{x_j}$  be minimizing geodesics from  $x_i$  to  $x_j$ ,  $i = 1, 2, 3$ . If we show  $d(x_1, \tau_{x_3}^{x_2}) > \sqrt{2A}$ ,  $d(x_2, \tau_{x_3}^{x_1}) > \sqrt{2A}$  and  $d(x_3, \tau_{x_2}^{x_1}) > \sqrt{2A}$  then Lemma 3.2.2 proves the theorem. By the triangle inequality (in fact a sum of two)  $2d(x_1, \tau_{x_3}^{x_2}) + d(x_2, x_3) \geq d(x_1, x_2) + d(x_1, x_3)$ . We also have  $4R_0 + 2\sqrt{2A} \geq d(x_i, x_j) \geq 2R_0 + 2\sqrt{2A}$ . Thus  $2d(x_1, \tau_{x_3}^{x_2}) + 4R_0 + 2\sqrt{2A} \geq 4R_0 + 4\sqrt{2A}$  and hence  $d(x_1, \tau_{x_3}^{x_2}) \geq \sqrt{2A}$ . The same argument for  $x_2$  and  $x_3$  yields the result.

*Case 2.*  $M$  is diffeomorphic to  $S^1 \times \mathbf{R}^1$ .

Let  $\tau(t)$ ,  $t \in (-\infty, \infty)$ , be a line in  $M$ . (You can find  $\tau$  by taking a limit of minimizing geodesics  $\gamma_i$  from  $x_i$  to  $y_i$  where  $x_i \rightarrow -\infty$  and  $y_i \rightarrow +\infty$ .) We now apply Lemma 3.3. Thus for each  $t$  we get a geodesic loop  $\gamma_t$  (not necessarily unique) at  $\tau(t)$ , of length  $L(t) < 2\sqrt{A}$ , satisfying (1), (2) and (3) of Lemma 3.3.  $\gamma_t$  separates since it is simple (Lemma 3.3.1) and essential. Hence Lemma 3.3.4 applies.

If  $\gamma_{t_0}$  is convex to  $+\infty$  (and not a closed geodesic) then there is an  $\varepsilon > 0$  such that  $L(t) < L(t_0)$  for all  $t \in (t_0, t_0 + \varepsilon)$ . One can see this by looking at the curves  $\gamma_{t_0}|_{[\delta, L(t_0) - \delta]} \cup \sigma$ , where  $\sigma$  is the minimizing geodesic from  $\gamma_{t_0}(L(t_0) - \delta)$  to  $\gamma_{t_0}(\delta)$ . Each of these curves is shorter than  $\gamma_{t_0}$  and is essential for small  $\delta$ . The set of points where these curves intersect  $\tau$  (each curve once) includes some interval  $\tau((t_0, t_0 + \varepsilon))$  by the convexity assumption. Similarly if  $\gamma_{t_0}$  is convex to  $-\infty$  there is an  $\varepsilon > 0$  such that  $L(t) < L(t_0)$  for all  $t \in (t_0 - \varepsilon, t_0)$ .

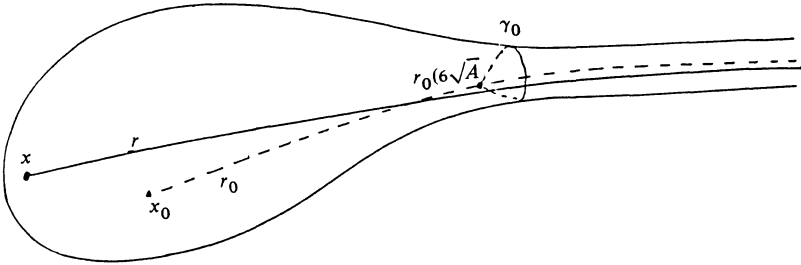
Since the area is finite  $L(t)$  cannot be bounded from below as  $t \rightarrow +\infty$  or as  $t \rightarrow -\infty$ , for if so, then  $L(S(x, t))$  would also be bounded below as  $t \rightarrow \infty$  implying the area is infinite. Thus we see that there is a  $\gamma_{t_0}$  and a  $\gamma_{t_1}$  such that  $\gamma_{t_0}$  is convex to  $-\infty$  and  $\gamma_{t_1}$  is convex to  $+\infty$ . Now by an easy limit argument both  $\{t | \exists \gamma_t \text{ convex to } +\infty\}$  and  $\{t | \exists \gamma_t \text{ convex to } -\infty\}$  are closed. Hence

there is a  $t_2$  in the intersection of these sets giving geodesic loops  $\gamma_+$  and  $\gamma_-$  at  $t_2$  with  $\gamma_-$  convex to  $-\infty$ ,  $\gamma_+$  convex to  $+\infty$ , and satisfying (1), (2), (3) and (4) of Lemma 3.3.

Let  $\Omega$  be the open disk between  $\gamma_+$  and  $\gamma_-$ . We may assume by Lemma 3.2.4 that for all  $z \in \Omega$ ,  $d(z, \tau) < \sqrt{2A}$ . Hence  $d_\Omega(z, \tau(t_2)) < \frac{3}{2}\sqrt{2A}$  (follow the minimizing geodesic going from  $z$  to  $\tau$  until it hits  $\partial\Omega$  then follow a boundary curve back to  $\tau(t_2)$ ). Thus by Lemma 3.1 we may assume  $\gamma_-$  is homotopic to  $\gamma_+$  through curves in  $\Omega$  of length  $\leq (12 + 3\sqrt{2})\sqrt{A}$ .

Now since  $\gamma_-$  and  $\gamma_+$  are convex to  $+\infty$  and  $-\infty$  respectively the argument proceeds as usual except since we are in the noncompact case we need to modify B.C.S.P. slightly. This is done in [2, pp. 87–88].

*Case 3.  $M$  is diffeomorphic to  $\mathbf{R}^2$ .*



Pick  $x_0 \in M$  and let  $r_0: [0, \infty) \rightarrow M$  be a ray from  $x_0$ . By Lemma 3.3 there is a shortest geodesic loop  $\gamma_0$  at  $r_0(6\sqrt{A})$  of length  $\leq 2\sqrt{A}$  essential in  $M - \{x_0\}$ . Let  $K$  be the relatively compact (i.e.  $\bar{K}$  is compact) component of  $M - \gamma_0$ .  $K \neq \emptyset$  since  $x_0 \in K$ . Let  $x \in K$  maximize the distance to  $\gamma_0$ . So, in particular,  $d(x, \gamma_0) \geq d(x_0, \gamma_0) \geq 5\sqrt{A}$ . Let  $r$  be a ray from  $x$ . Again, by Lemma 3.3, for  $t > \sqrt{A}$  there are geodesic loops  $\gamma_t$  through  $r(t)$  (not unique as usual) of length  $\leq 2\sqrt{A}$ , essential in  $M - \{x\}$ , and satisfying Lemma 3.3.

As in the proof of the  $S^1 \times \mathbf{R}$  case there are large  $t$  with  $\gamma_t$  convex to  $\infty$ . The proof, as usual breaks up into two cases. Either all the loops  $\gamma_t$ ,  $t > \sqrt{A}$ , are convex to  $\infty$  or for some  $t_0 > \sqrt{A}$  there are two such loops, one convex to  $x$  and one convex to  $\infty$ .

The case where all  $\gamma_t$  are convex to  $\infty$  is treated as in the  $S^2$  case. Let  $\gamma$  be such a loop at  $r(\sqrt{A})$  and let  $\Omega$  be the relatively compact component of  $M - \gamma$ .  $\Omega$  is contained in  $K$  since  $d(x, \gamma(t)) \leq 2\sqrt{A}$  for all  $t$  while  $d(x, \gamma_0) \geq 5\sqrt{A}$ . Now for  $z \in \Omega$ ,  $d(z, \gamma) + d(\gamma, \gamma_0) \leq d(z, \gamma_0) \leq d(x, \gamma_0) \leq d(x, \gamma) + \sqrt{A} + d(\gamma, \gamma_0) \leq 2\sqrt{A} + d(\gamma, \gamma_0)$ . Hence for all  $z \in \Omega$ ,  $d(z, \gamma) \leq 2\sqrt{A}$ . Since  $x$  is at a local maximum of the distance to  $\gamma_0$ , the proof of Berger's lemma (see [8, p. 106]) yields minimizing geodesics  $\tau_i$  from  $x$  to  $\gamma_0$  such that for all  $V \in T_x M$

there is an  $i$  such that  $\langle V, \tau_i'(0) \rangle \geq 0$ . Thus the same proof as in the  $S^2$  case shows that  $\gamma$  is homotopic through short ( $\leq \text{const}\sqrt{A}$ ) curves in  $\bar{\Omega}$  to the point curve  $\{x\}$ . The rest of the argument is the modified B.C.S.P. since  $\gamma$  is convex to  $\infty$ .

In the other case there are two such geodesic loops,  $\gamma_1$ , and  $\gamma_2$  at  $r(t_0)$  with  $\gamma_1$  convex to  $x$  and  $\gamma_2$  convex to  $\infty$ . Let  $\Omega$  be the disk between them. The proof would follow as before if we could show that for all  $z \in \Omega$ ,  $d(z, \partial\Omega) \leq \text{const}\sqrt{A}$ . In fact, we will show that if for some  $z \in \Omega$ ,  $d(z, \partial\Omega) > 4\sqrt{A}$ , then applying B.C.S.P. to  $\gamma_1 \cup \gamma_2$  yields a nontrivial closed geodesic of length  $\leq 4\sqrt{A}$ . Thus let  $z \in \Omega$  be such that  $d(z, \partial\Omega) > 4\sqrt{A}$ . We see that  $t_0 \geq 3\sqrt{A}$ , for if  $t_0 < 3\sqrt{A}$ ,  $\gamma_2 \subset K$  and  $d(z, \gamma_2) \leq 4\sqrt{A}$  by the arguments for the previous case. Let  $\sigma_1$  be a smooth curve from  $z$  to  $\infty$  lying in the unbounded component of  $\gamma_1$  which intersects  $\gamma_2$  transversely once, and let  $\sigma_2$  be a smooth curve in the bounded component of  $\gamma_2$  intersecting  $\gamma_1$  transversely once from  $z$  to  $x$ . Now  $\gamma_1 \cup \gamma_2$  has intersection number 1 with  $\sigma_1$  and with  $\sigma_2$  while it has intersection number 2 with  $r$ . Hence if  $\gamma_1 \cup \gamma_2$  were to shrink to a point or run off to  $\infty$  under B.C.S.P. then some curve  $h$  in the homotopy must either pass through  $z$  while still intersecting  $r$  or pass through  $x$  while still intersecting  $\sigma_1$ . But  $d(z, r) \geq d(z, \partial\Omega) > 4\sqrt{A}$  and  $d(x, \sigma_1) \geq d(x, \gamma_1) \geq t_0 - \sqrt{A} \geq 2\sqrt{A}$ . In either case  $L(h) \geq 4\sqrt{A}$  contradicting  $L(h) < L(\gamma_1 \cup \gamma_2) \leq 4\sqrt{A}$ .

The theorem follows.

## 6. Convex Hypersurfaces

For  $D^n \subset \mathbf{R}^n$  a convex domain and  $l$  a line through the origin the width of  $D$  in the direction of  $l$  is the distance between the two (parallel) tangent spaces to  $\partial D$  perpendicular to  $l$ . The width of  $D$  is the minimum over all directions  $l$ . The main theorem in this section leads us to consider the constants:

$$c_0(n) = \inf\{\text{Vol}(D) \mid D^n \subset \mathbf{R}^n \text{ is convex of width } 1\}.$$

Unfortunately the value of  $c_0(n)$  is only known in the case  $n = 2$ . In this case the infimum is achieved for  $D$  equal to the equilateral triangle of side  $2/\sqrt{3}$  and  $c_0(2) = 1/\sqrt{3}$ .

It is easy to get a lower bound for  $C(n)$ , however, the exact value is unknown (see [13, Problem 26]). The best known lower bound for  $C_0(n)$  is  $2\sqrt{3}/n!$ , which is due to Firey [10].

**Theorem 6.1.** *Let  $M^n \subset \mathbf{R}^{n+1}$  be a convex hypersurface and  $L$  the length of the shortest nontrivial closed geodesic. Then*

$$2 \cdot \frac{2}{\sqrt[n]{2c_0}} \cdot \sqrt[n]{\text{Vol}(M)} \geq L.$$

**Remarks.** One suspects that the best constant would be  $2/\sqrt[n]{2c_0}$  (in which case the theorem is off by a factor of 2) and that equality would hold (in a generalized sense) for  $M^n =$  two copies of  $D^n$  glued together along  $\partial D$ , where  $D$  is the best domain in the definition of  $c_0(n)$ . In particular for  $n = 2$  (even without the convexity assumption) one suspects the best constant to be achieved only by the bi-equilateral triangle (two equilateral triangles glued along the boundary). It was pointed out by Calabi that this (degenerate) convex two manifold has a simple and a nonsimple closed geodesic of shortest length. Recently Calabi has shown that the shortest closed geodesic on a convex (nondegenerate) surface is in fact simple. In the nonconvex case it is easy to find examples where the shortest closed geodesic is not simple.

To prove the theorem we need

**Lemma 6.2.** *Let  $M^n \subset \mathbf{R}^{n+1}$  be a convex hypersurface. Then*

(1) *If  $P^2$  is a two plane in  $\mathbf{R}^{n+1}$  and  $\Pi_{P^2}: \mathbf{R}^{n+1} \rightarrow P^2$  is the orthogonal projection then  $L(\partial(\Pi_{P^2}(M))) \geq L$ .*

(2) *If  $P^n \subset \mathbf{R}^{n+1}$  is a hypersurface and  $\Pi_{P^n}$  is the orthogonal projection then  $\text{Vol}(M) \geq 2 \text{Vol}(\Pi_{P^n}(M))$ .*

*Proof of Lemma 6.2.* Statement (2) is clear. Statement (1) follows from Birkhoff's idea. By slicing  $M$  with 2 planes parallel to  $P^2$  we get a family of curves for which Birkhoff's idea works (for the details of this see the proof of Lemma 1.6 of [9]). Hence  $L$  is less than or equal to the length of the longest curve in this family. But since each curve in this family projects onto  $P^2$ , in a length preserving way, to a convex curve inside  $\partial(\Pi_{P^2}(M))$ , we see that the length is less than or equal to  $L(\partial(\Pi_{P^2}(M)))$  and the lemma follows.

*Proof of Theorem 6.1.* Let  $K^{n+1}$  be the convex body such that  $\partial K = M$  and let  $\varepsilon$  be the width of  $K$ , which we assume is in the direction of a line  $l_1$ . Let  $P^n$  be the hyperplane perpendicular to  $l_1$  and  $D = \Pi_{P^n}(M)$ . Let  $w$  be the width of  $D$ , which we assume is in the direction of a line  $l_2$ . Let  $P^2$  be the plane determined by  $l_1$  and  $l_2$ . Applying Lemma 6.2 to  $P^2$  and  $P^n$  yields:

$$L \leq L(\partial(\Pi_{P^2}(M))) \leq 2\varepsilon + 2w \leq 4w$$

and

$$\text{Vol}(M) \geq 2 \text{Vol}(\Pi_{P^n}(M)) \geq 2c_0(n)w^n.$$

Combining these two inequalities yields the theorem.

## References

- [1] R. D. M. Accola, *Differential and extremal lengths on Riemannian surfaces*, Proc. Math. Acad. Sci. USA **46** (1960) 540–543.
- [2] V. Bangert, *Closed geodesics on complete surfaces*, Math. Ann. **251** (1980) 83–96.

- [3] M. Berger, *Du côté de chez Pu*, Ann. Sci. École Norm. Sup. **4** (1972) No. 1, 1–44.
- [4] ———, *A l'ombre de Loewner*, Ann. Sci. École Norm. Sup. **4** (1972) 5, No. 2, 241–260.
- [5] G. D. Birkhoff, *Dynamical systems*, Amer. Math. Soc. Colloq. Publ. Vol. 9, Providence, RI, 1927.
- [6] C. Blatter, *Über Extremallängen auf geschlossenen Flächen*, Comment Math. Helv. **35** (1961) No. 3, 153–168.
- [7] Y. Burago & V. Zalgaller, *Geometric inequalities*, “Nauka” (1980), Russian.
- [8] J. Cheeger & D. Ebin, *Comparison theorems in Riemannian geometry*, North-Holland Math. Libr., North-Holland, Amsterdam, 1975.
- [9] C. Croke, *Poincaré's problem and the length of the shortest closed geodesic on a convex hypersurface*, J. Differential Geometry **17** (1982) 595–634.
- [10] W. Firey, *Lower bounds for volumes of convex bodies*, Arch. Math. **16** (1965) 69–74.
- [11] F. Fiala, *Le problème der isopérimètres sur les surfaces ouverts à courbure positive*, Comment Math. Helv. **13** (1940–41), 293–346.
- [12] M. Gromov, *Filling Riemannian manifolds*, J. Differential Geometry **18** (1983) 1–147.
- [13] P. M. Gruber & Schneider, *Problems in geometric convexity*, Contributions to Geometry, Proc. Geometry Sympos. in Liegen, 1978, edited by Jürgen Tölke and Jörg M. Wills, Birkhäuser, Basel, 1979, 255–278.
- [14] P. Hartman, *Geodesic parallel coordinates in the large*, Amer. J. Math. **86** (1964) 705–727.
- [15] J. Hebda, *Some lower bounds for the area of surfaces*, Invent. Math. **65** (1982) 485–491.
- [16] M. Katz, *The filling radius of two-point homogeneous spaces*, J. Differential Geometry **18** (1983) 505–511.
- [17] P. Pu, *Some inequalities in certain non-orientable Riemannian manifolds*, Pacific J. Math. **2** (1952) 55–71.
- [18] L. A. Santaló, *Integral geometry and geometric probability*, Encyclopedia. Math. Appl., Addison-Wesley, London, 1976.
- [19] A. Treibergs, *Estimates of volume by the length of shortest closed geodesics on a convex hypersurface*, Invent. Math. **80** (1985) 481–488.
- [20] G. Thorbergsson, *Closed geodesics on non-compact Riemannian manifolds*, Math. Z. **159** (1978) 249–258.

UNIVERSITY OF PENNSYLVANIA

