J. DIFFERENTIAL GEOMETRY
36 (1992) 517-549

SIMPLE CLOSED GEODESICS ON CONVEX SURFACES

EUGENIO CALABI & JIANGUO CAO

Abstract

A geodesic is said to be simple if it does not have any self-intersection point. It will be shown that the shortest closed geodesic is simple on any smooth Riemannian 2-sphere of nonnegative curvature.

We will also derive various estimates for lengths of simple closed geodesics, in terms of the diameter D, total area A, and curvature K of a given surface M^2. In particular, if we let L be the length of the longest simple closed geodesic on a smooth Riemannian sphere of curvature $0 \leq K \leq 1$, then $2D \leq L \leq A/2$. Furthermore, equality $L = A/2$ holds if and only if M^2 is isometric to the unit sphere.

Finally, if M^2 is a Riemannian sphere with nonnegative curvature, then we find that the isoperimetric inequality $A \leq 8D^2/\pi$ is useful.

Introduction

The purpose of this note is to study simple closed geodesics on compact oriented convex surfaces. A geodesic γ is said to be simple if γ has no self-intersections. In what follows, all geodesics are assumed to be nontrivial. Hence, any point curve will not be counted as a closed geodesic. If a Riemannian surface M^2 is homeomorphic to the two-sphere S^2 and if M^2 has nonnegative sectional curvature, then M^2 is called a convex surface.

First, we would like to find out which closed geodesics are simple on a given surface. The following theorem gives a partial answer.

Theorem D. *If g is a C^3 smooth metric on a two-sphere S^2 with nonnegative curvature, then any nontrivial closed geodesic of the shortest length is simple.*

In Theorem D, we only consider the C^3 smooth metric g, since there are examples of nonsmooth metrics on a two-sphere S^2 in which the shortest geodesics are not simple. For instance, the bi-equilateral triangle (two

Received January 26, 1990 and, in revised form, October 23, 1990. The first author was supported in part by National Science Foundation grant DMS 87-02359, and the second author by National Science Foundation grant DMS-8610730 at the Institute for Advanced Study.

equilateral triangles glued along the boundary) has both simple and nonsimple closed geodesics of shortest length (cf. [10, p. 20]). The convexity assumption is needed in Theorem D because when the curvature becomes negative, one can find some examples where the shortest closed geodesic is not simple (cf. [19, p. 31]).

Second, we would like to work on other aspects of closed geodesics. In particular, it might be an interesting problem to study simple closed geodesics more quantitatively. Hence, we are led to estimate lengths of simple closed geodesics, in terms of other geometric data: such as the diameter, total area, and upper bound of curvature of a given surface. The first main estimate in this note is:

Theorem A. *Let M^2 be a C^3 smooth Riemannian surface diffeomorphic to S^2, with curvature $0 \leq K \leq 1$ and area A. Then the length of any simple closed geodesic γ is less than or equal to $A/2$, i.e., $L(\gamma) \leq A/2$. Furthermore, equality holds if and only if M^2 is isometric to the unit sphere.*

Using Theorem A, one can easily derive the following fact.

Corollary A. *Let $\mathscr{M}_1(M)$ be the moduli space of all simple closed geodesics in a smooth Riemannian surface M. If M is a two-sphere with nonnegative curvature, then $\mathscr{M}_1(M)$ is a compact subspace of $C^0(S^1, M)$.*

The conclusion of Theorem A (resp. Corollary A) fails if the hypothesis $K \geq 0$ is weakened to $K \geq -\epsilon^2$ for any $\epsilon > 0$. For example, on a flat torus or a two-sphere of *bone* type, there is no upper bound for lengths of all simple closed geodesics. If M_n^2 is a hyperbolic surface of genus n, then $\text{Area}(M_n^2) = 4\pi(n+1)$. It is known that lengths of primitive simple closed geodesics are related to the moduli space of hyperbolic surfaces with a fixed genus n, thus, they can be arbitrary long (cf. [21]). Therefore, it is reasonable to limit our attention to two-spheres of nonnegative curvature.

There are also examples of nonsmooth metrics on a two-sphere in which the least upper bound for lengths of simple closed geodesics is infinity. For instance, a special tetrahedron (two rectangles glued along its boundary) has a family of simple closed geodesics whose lengths are unbounded.

Our next step is to give a lower bound for the length of longest simple closed geodesics. Hence, we introduce

$$L(M) = \sup\{L(\gamma) | \gamma \text{ is a simple closed geodesic on } M\}$$

During the summer of 1989, the first author found a nonsharp lower bound for $L(M)$ in terms of the diameter and curvature of M. Later on, Chris Croke kindly pointed out to us that there is a sharp estimate for $L(M)$, which only depends on the diameter. We are grateful to C. Croke for permitting us to quote his unpublished result.

Theorem B. *Let M^2 be a Riemannian manifold, diffeomorphic to S^2, of diameter D. Then there is a simple closed geodesic γ with $L(\gamma) \geq 2D$, i.e., $L(M^2) \geq 2D$. Moreover, if equality $L(M^2) = 2D$ holds, then there are infinitely many simple closed geodesics of length $2D$.*

Finally, we wish to point out that the following isoperimetric inequality is useful.

Theorem C. *Let M^2 be a Riemannian two-sphere with nonnegative curvature. Then the following inequality holds:*

$$\text{Area}(M^2) \leq \frac{8}{\pi} D^2 .$$

The best estimate of A/D^2 for convex surfaces could be $\pi/2$, which is achieved by two flat discs glued together along the boundary. Our estimate $8/\pi$ is off the best conjectured constant only by a factor ≤ 1.62.

To illustrate the usefulness of our main theorems, we present some applications.

Corollary 0.1. *Let M^2 be a smooth Riemannian sphere with $0 \leq K \leq 1$, and let D and A be as above. Then the longest simple closed geodesic γ exists and the length of γ satisfies the following inequalities:*

$$\sqrt{A\pi/2} \leq L(\gamma) \leq A/2 , \qquad 2D \leq L(\gamma) \leq 4D^2/\pi .$$

Furthermore, equality $L(\gamma) = A/2$ holds if and only if M^2 is isometric to the unit sphere.

Corollary 0.2. *Let M^2 be a smooth Riemannian sphere of nonnegative curvature, and let γ be the shortest simple closed geodesic on M^2. Then $L(\gamma) \leq 9D$.*

We would like to say a few words about the ratio $L(M)/D^2$, when D becomes infinite. In §3 of Part I, we will construct a one-family of smooth metrics g_i on S^2 whose curvature satisfies $0 \leq K \leq 1$, in which $L(M_i)/D_i^2 \geq (32\pi)^{-1}$ and $D_i \to \infty$ as $i \to \infty$. Hence, the upper bound for the rate of growth $L(M^2)$ has to be quadratic in terms of D in Corollary 0.1.

The least upper bound of $L(M)/D^2$ for all smooth convex surfaces of $K \leq 1$ is still not known. Comparing the unit two-sphere, our estimate is off the best estimate at most by a factor 2. Our estimates $2D \leq L(M) \leq A/2$ are sharp and optimal.

There are several known methods to find simple closed geodesics. On the two-sphere S^2 with an arbitrary smooth Riemannian metric g, there always exist three simple closed geodesics by the Lusternik-Schnirelmann theorem (cf. [2] or [11]). However, those three geodesics may not be

shortest ones. Poincaré also suggested that one could find a simple geodesic on a convex surface M by minimizing the arclength functional over the set $\mathscr{A}$ of all simple smooth closed curves which separate M into two pieces of equal total curvature. His method was carried out correctly by C. Croke in [9].

The proof of Theorem D introduces a completely different approach. We will study the space of one-cycles (also called rimmed domains) instead of the ordinary loop space on a surface. The space of one-cycles has some nice properties. For example, it allows us to *cut and paste* any nonsimple closed geodesic along its intersection points to get a nearby family of one-cycles, in which the lengths of one-cycles become smaller (cf. §2 of Part II). Such perturbation is one of the basic techniques in Part II. A minimax argument for the space of one-cycles is also needed to prove Theorem D. In order to keep the proof short, we will quote some results of Almgren and Pitts, which are related to the geometric measure theory. For the reader who is not familiar with the geometric measure theory, we will also give an independent proof of the minimax principal on the space of one-cycles in the appendix.

In the proof of Theorem A, we will make use of an integral formula, which is due to L. Santaló (cf. §1 of Part I). The higher dimensional analogues of Theorem A are still not known to the authors. Some relevant results may be found in [3]. The proofs of Corollaries 0.1–0.2 will be given in §3 of Part I.

Conventions. When a metric g is continuous up to its third derivatives on a two-dimensional manifold M^2, (M^2, g) is called a C^3 smooth Riemannian surface. In what follows, we always let M^2 be an oriented two-sphere S^2 endowed with a C^3 smooth metric g unless otherwise stated. Given a piecewise smooth curve γ (not necessarily closed), by $L(\gamma)$ we will mean the length of γ. The injectivity radius of the Riemannian surface M^2 is denoted by $\mathrm{inj}(M^2)$. If Ω is an open domain, $A(\Omega)$ is defined to be the area of Ω. K stands for the curvature of (M^2, g). These notations will be used throughout this paper.

Acknowledgment

The second-named author is very indebted to Professor C. Croke, his teacher, for continuous support and encouragement, especially the help in writing §2 of Part I. Both authors would like to thank H. Gluck and W. Ziller for their helpful comments and interest in this work.

Finally, the authors want to thank the referee for pointing out some minor errors and misprints in an earlier version of this paper.

PART I
AREA, DIAMETER, AND LENGTH OF SIMPLE CLOSED GEODESICS

1. Total surface area and the length of simple closed geodesic

In this section, we will derive a sharp upper bound, in a smooth Riemannian sphere of nonnegative curvature, for the length of all simple closed geodesics in terms of the total area and the curvature k. By rescaling the metric, we may assume without loss of generality that $0 \leq K \leq 1$.

We will show that if M is a C^3 smooth Riemannian surface, diffeomorphic to S^2, with curvature $0 \leq K \leq 1$ and area A, then the length of any simple closed geodesic γ is less than or equal to $A/2$. This estimate is sharp, which is achieved by the unit sphere. Inequality $L(\gamma) \leq A/2$ fails if the hypothesis $0 \leq K \leq 1$ is weakened to $-\epsilon^2 \leq K \leq 1$ for any $\epsilon > 0$. The estimate also fails for flat tori and surfaces of higher genus except for the projective plane $\mathbf{R}P^2$.

We begin with an elementary example to demonstrate the main idea of this section. Let Ω be a rectangle with width a and height b on the Euclidean plane E^2. Obviously, one sees that $A(\Omega) = ab$, the area of Ω, and $L(\partial\Omega) = 2(a + b)$., the perimeter of Ω. Our goal is to study the upper bound of the ratio $L(\partial\Omega)/A(\Omega)$. If $a = 1$ and $b \to 0$, then the ratio becomes infinite. However, if we denote the width of Ω by $W_0(\Omega) = \min\{a, b\}$, we can easily verify that $L(\partial\Omega) \leq 2A(\Omega)/W_0(\Omega)$.

Let γ be a simple, closed geodesic on a Riemannian sphere M. Clearly, γ divides M into two simply connected, open components, ay Ω_1 and Ω_2. We will define the width, $W(\Omega_j)$, of each Ω_j, and then estimate $L(\gamma)$ from above in terms of the area $A(\Omega_j)$ and $W(\Omega_j)$ for either of two domains Ω_j. There is a classical formula due to L. Santaló which gives a sharp estimate of $L(\gamma)/A(\Omega_i)$ in terms of the "width" (cf. [20] or [8]). To set the stage for our application of the Santaló formula, we must introduce an analogous notion of $W(\Omega_j)$, the width of any given, open, connected surface Ω_j with boundary $\partial\Omega_j$, as follows.

The new width $W(\Omega_j)$ is defined to be the infimum for lengths of all geodesic chords within Ω_j. More precisely, we let N represent the inwardly pointing unit normal vector field of $\partial\Omega_j$, UM represent the

unit circle bundle of M, and $U\Omega_j = UM|_{\Omega_j}$. The upper semicircle bundle is denoted by

$$U^+\partial\Omega_j = \{v | v \in UM|_{\partial\Omega_j}, \langle v, N_{\Pi(v)}\rangle > 0\}.$$

For any $v \in U^+\partial\Omega_j$ or $v \in U\Omega_j$, we let σ_v be a geodesic with $\sigma'(0) = v$ and $\sigma(0) = \Pi(v)$, where $\Pi: UM \to M$ is the canonical projection map. Naturally, we let $l(v)$ be the smallest value of $t > 0$ (possibly ∞) such that $\sigma_v(t) \in \partial\Omega_j$. Finally, one can define

$$W(\Omega_j) = \inf\{l(v) | v \in U^+\partial\Omega_j\}.$$

The Santaló formula (cf. [10, p. 421]) states that

$$\begin{aligned} 2\pi A(\Omega_j) = \operatorname{Vol}(UM|_{\Omega_j}) &\geq \int_{U^+\partial\Omega_j} l(v)\langle v, N_{\Pi(v)}\rangle \, dv \\ &\geq W(\Omega_j)L(\partial\Omega_j)\int_0^\pi \sin\theta \, d\theta = 2W(\Omega_j)L(\gamma). \end{aligned} \tag{1.1}$$

Hence, we have

$$L(\gamma) \leq \frac{\pi}{W(\Omega_j)} A(\Omega_j), \qquad j = 1, 2. \tag{1.2}$$

In what follows, we want to estimate $W(\Omega_j)$ from below in terms of curvature of a given surface. Our lower bound for $W(\Omega_j)$ will be independent of the choice of any simple closed geodesic γ and Ω_j.

Lemma 1.1. *Let M be a C^3 smooth Riemannian sphere whose curvature K satisfies $0 \leq K \leq 1$, and let γ, Ω_j, and $W(\Omega_j)$ be as above. Then $l(v) \geq \pi$ for all $v \in U^+(\partial\Omega_j)$, i.e.,*

$$W(\Omega_j) \geq \pi. \tag{1.3}$$

Moreover, if $l(v) = \pi$ for all $v \in U^+\partial\Omega$, then $L(\partial\Omega_j) = 2\pi$.

The proof of Lemma 1.1 is quite involved, and we postpone it to the end of this section. At this moment, under the assumption that Lemma 1.1 holds, we would like to prove

Theorem A. *Let M be a C^3 smooth Riemannian surface diffeomorphic to S^2, with curvature $0 \leq K \leq 1$ and area A. Then the length of any simple closed geodesic γ is less than or equal to $A/2$, i.e., $L(\gamma) \leq A/2$. Furthermore, equality holds for some γ if and only if M is isometric to the unit sphere.*

Proof of Theorem A. Let Ω_1 be an open component of $M - \gamma$ with the property that $\operatorname{Area}(\Omega_1) \leq \frac{1}{2}\operatorname{Area}(M)$. It follows from Lemma 1.1 and

(1.2) that

$$L(\gamma) \leq \frac{\pi}{2W(\Omega_1)} A(M) \leq A/2.$$

Furthermore, equality holds if and only if $\text{Area}(\Omega) = \text{Area}(M)/2 = \text{Area}(M \setminus \overline{\Omega})$, $l(v) = \pi$ for all $v \in U^+\partial\Omega$. Using a result of Bangert, we now conclude that Ω is isometric to the unit semisphere (cf. [4]). One can also give an alternative proof as follows: Lemma 1.1 implies that $L(\partial\Omega) = 2\pi$. In this case, by (1.2) we get $A(M) = 4\pi$. It will be shown in Lemma 1.2 that $A(M) \geq 4\pi$, and equality holds if and only if M is the standard unit two-sphere. This completes the proof of Theorem A under the assumption that Lemmas 1.1 and 1.2 are true. q.e.d.

In order to carry out the proof of Lemma 1.1, we need to derive some preliminary facts.

Lemma 1.2. *Let $M = (S^2, g)$ with nonnegative curvature K, $0 \leq K \leq 1$. Then the injectivity radius of M, $\text{inj}(M)$, is greater than or equal to π, i.e.,*

$$\text{inj}(M) \geq \pi. \tag{2.4}$$

Consequently, $\text{Area}(M) \leq 4\pi$ *and* $A(M) = 4\pi$ *if and only if M is isometric to the standard unit two-sphere.*

Remark. Inequality (1.4) clearly fails on certain flat tori. We include a proof of Lemma 1.2 here for the sake of completeness.

Proof. It follows from the Gauss-Bonnet theorem that $\max K(x) > 0$. Hence, the set $G = \{x | x \in M, K(x) > 0\}$ is nonempty. By Corollary 5.7 in [7], we may assume that there is a smooth closed geodesic γ, parametrized by its arc-length, with $\gamma(0) = p$ through p and q such that $L(\gamma) = 2d(p, q) = 2\,\text{inj}(M)$. Let $\overline{G}$ be the closure of G. If $\gamma \cap \overline{G} = \varnothing$, we can move γ within the flat region $M \setminus G$ until γ hits $\overline{G}$. Therefore, we may assume that $\gamma \cap \overline{G} \neq \varnothing$ at the beginning. Let N be the unit normal vector field of γ which is pointing towards G and

$$h_s(t) = \exp_{\gamma(t)}[sN(t)].$$

When $0 < s < \epsilon_0 =$ the focal radius of γ, it follows from the Corollary of Rauch II that $L(h_s) \leq L(\gamma)$; the equality holds if and only if the strip $\Delta_s = \{h_\alpha(t) | 0 \leq \alpha \leq s\}$ is flat (see Lemma 1.4 below, [12] or [7, p. 31]). In our case, $L(h_s) < L(\gamma)$ whenever $s > 0$, since $K(x) \geq 0$ and $K > 0$ somewhere in any strip Δ_s. Now, for the same reason as in [7, p. 99], we conclude that q is conjugate to p. Hence, $\text{inj}(M) = d(p, q) \geq \pi$.

Using the fact that $\text{inj}(M) \geq \pi$ and $K \leq 1$, one can use the standard area comparison theorem or the Berger-Kazdan inequality to conclude that

$A(M) \geq 4\pi$ and the equality holds if and only if M is the standard unit 2-sphere (cf. [12], or [5], [14]). In fact, when $A(M) = 4\pi$ and $0 \leq K \leq 1$, the Gauss-Bonnet formula implies that $K(p) = 1$ for all $p \in M$. Hence, M is a round sphere of curvature 1. q.e.d.

As an application of Lemma 1.2, one has

Corollary 1.3. *Let M be as in Lemma* 1.2, *and let σ_1 and σ_2 be two distinct geodesic segments with the same endpoints. Then*

$$\max\{L(\sigma_1), L(\sigma_2)\} \geq \pi.$$

We also need a special sharp version of the second Rauch comparison theorem (Berger's Lemma) to prove Lemma 1.1. Since it was not clearly stated in any literature, we present it here with a simple proof.

Let M be a two-dimensional Riemannian manifold with nonnegative curvature $K \geq 0$, $\eta: [0, 1] \to M$ be a geodesic with $\|\eta'(0)\| = 1$, and $N(0)$ be a unit normal vector of η at $\eta(0)$.

Lemma 1.4. *Suppose J is a Jacobi field along η with $J(0) = a\eta'(0) + bN(0)$, $b \neq 0$, and $J'(0) = 0$. If η has no focal points of the geodesic σ defined by $\sigma(s) = \exp_{\eta(0)}[sN(0)]$ in $[-1, 1]$, then*

$$\|J(t)\| \leq \|J(0)\| \quad \textit{for all } t \in [0, 1].$$

Moreover, $\|J(s)\| = \|J(0)\|$ for some $s > 0$ if and only if $K(\eta(t)) = 0$ for all $t \in [0, s]$.

Proof. Let $N(t)$ be the unit normal vector field along η. Since $\dim M = 2$, one gets $N'(t)k = 0$. Since $J'(0) = 0$, it follows that J can be decomposed into

$$J(t)k = a\eta'(t) + f(t)N(t),$$

where $f(0) = b$ and $f'(0) = 0$ (cf. [7, p. 19]). Clearly, one sees

$$\|J(t)\|^2 = a^2 + f^2(t).$$

Since σ does not have any focal point on η, f does not change its sign on $[0, 1]$. By the assumption $K \geq 0$, we know that

$$f''(t) = -K(\eta(t))f(t),$$
$$f(t)f''(t) \leq -K(t)f^2(t) \leq 0.$$

If $f(0) = b > 0$, then $f''(t) \leq 0$ as long as $f(t) > 0$. This together with $f'(0) = 0$ implies that $f'(t) \leq 0$ and $0 < f(t) \leq f(0) = b$.

When $f(0) = b < 0$, one can also show that $-f(t) \leq -f(0) = -b$ and $f^2(t) \leq f^2(0)$. Hence,

$$\|J(t)\|62 = a^2 + f^2(t) \leq a^2 + b^2 = \|J(0)\|^2.$$

$\|J(s)\| = \|J(0)\|$ holds for some $s > 0$ if and only if $K(\eta(t)) = 0$ for all $t \in [0, s]$. q.e.d.

Now, we are ready to prove Lemma 1.1.

Proof of Lemma 1.1. Notice that Ω is an open domain. By the definition of $W(\Omega)$, we are interested in all geodesic chords which lie in Ω except for endpoints. In order to give a criterion for geodesics which are transversal to $\partial\Omega$, we need to introduce an intrinsic distance function d_γ of $\partial\Omega = \gamma$.

For any pair of points $\{p, q\}$ on a simple closed geodesic γ, p and q divide γ into two geodesic segments γ_1 and γ_2. We define $d_\gamma(p, q) = \min\{L(\gamma_1), L(\gamma_2)\}$.

Let σ be a geodesic segment σ with endpoints $\{p, q\} \subset \gamma$ and of length $L(\sigma) < \pi$. If σ is different from γ_1 and γ_2, then it follows from Corollary 1.3 that $d_\gamma(p, q) \geq \pi$. Conversely, if η is a geodesic with endpoints $p_0, q_0 \in \partial\Omega$, and if $L(\eta) < d_\gamma(p_0, q_0)$, then η is clearly transversal to $\partial\Omega$. This observation will be used later on.

We will use a contradiction method to finish the proof of Lemma 1.1. It takes several steps to get a contradiction.

Suppose to the contrary that Lemma 1.1 is false. Then there would be a family of geodesics $\sigma_i: (0, L_i) \to \Omega$ with endpoints $\{p_i, q_i\} \subset \gamma$ and of length $L_i < \pi - \epsilon_0$, where $\epsilon_0 = \frac{1}{2}[\pi - W(\Omega)] > 0$ and $L_i \to W(\Omega)$ as $i \to \infty$. Clearly, by the argument above, we have

$$d_\gamma(p_i, q_i) \geq \pi \quad \text{for all } i. \tag{1.5}$$

Since $\overline{\Omega}$ is compact, we may choose a subsequence of $\{\sigma_i\}$ which is convergent to a normal geodesic $\sigma_0: [0, L(\sigma_0)] \to \overline{\Omega}$, with endpoints $\{p_0, q_0\} \subset \partial\Omega$, and of length

$$L(\sigma_0) = W(\Omega) < \pi \leq d_\gamma(p_0, q_0). \tag{1.6}$$

Because of (1.6), we know that $p_0 \neq q_0$ and σ_0 is a nontrivial geodesic which lies in the interior of $\overline{\Omega}$ except for its endpoints.

Let $L_0 = L(\sigma_0)$. Lemma 1.2 tells us that $\text{inj}(M) \geq \pi > L_0$. By the definition of $W(\Omega)$ and the first variational formula, one can show that σ_0 is perpendicular to $\partial\Omega = \gamma$ at points p_0 and q_0. Let N be the unit normal vector field along σ_0 and $G = \{x | x \in \Omega, K(x) > 0\}$. Since Ω is diffeomorphic to a disc and $\partial\Omega$ forms a closed geodesic, one sees that $G \neq \varnothing$. If $\sigma_0 \cap \overline{G} = \varnothing$, we move σ_0 in direction N within the flat region $\Omega \setminus G$ until σ_0 hits $\overline{G}$ somewhere. Therefore, we may assume that $\sigma_0 \cap \overline{G} \neq \varnothing$ and N is pointing towards G. As we did in the proof of

Lemma 1.2, we introduce a family of curves

$$h_s(t) = \exp_{\sigma_0(t)}[sN(t)], \qquad t \in [0, L_0]. \tag{1.7}$$

If δ is the focal radius of σ_0 in Ω, we claim that

$$L(h_s) < L(\sigma_0), \quad \text{whenever } 0 < s < \delta. \tag{1.8}$$

This can be seen as follows. For any fixed s with $0 < s < \delta$, it follows from Lemma 1.4 that

$$L(h_s) \leq L(\sigma_0), \tag{1.9}$$

and the equality holds in (1.9) if and only if the strip $\Delta_s = \{h_u(t) | 0 \leq u \leq s,\ 0 \leq t \leq L_0\}$ is flat. However, our strip is not flat as long as $s > 0$; thus, the strict inequality (1.8) holds.

Now, we fix and $s > 0$ and apply the Birkhoff curve shortening process to h_s with fixed endpoints. From this, we get a geodesic σ with endpoints $h_s(0), h_s(L_0) \in \partial\Omega$, and

$$L(\sigma) \leq L(h_s) < L(\sigma_0) = W(\Omega). \tag{1.10}$$

It remains for us to show that $\sigma \subset \overline{\Omega}$. It is clear that there is an $\epsilon > 0$ such that for all $x, y \in \partial\Omega$, with $d(x, y) < \epsilon$, the minimizing geodesic τ from x to y satisfies $\tau \subset \overline{\Omega}$. In other words, $\partial\Omega$ is convex to Ω. Hence, it follows that $\sigma \subset \overline{\Omega}$ (cf. [10, pp. 3–7]). Furthermore, if s is sufficiently small, we have

$$d_\gamma(\sigma(0), \sigma(L_0)) = d_\gamma(h_s(0), h_s(L_0)) \geq d_\gamma(p_0, q_0) - 2s > \pi - \epsilon_0 > L(\sigma).$$

Therefore, σ is transversal to $\partial\Omega$ and lies in Ω except for its endpoints. This fact together with (1.10) gives a contradiction to the definition of $W(\Omega)$. This finishes the proof of the first part of Lemma 1.1:

$$W(\Omega) \geq \pi. \tag{1.11}$$

The second part of Lemma 1.1 follows immediately from Bangert's result (cf. [4]).

Remark. There is another interesting estimate for the length of any given simple closed geodesic in the two-sphere of nonnegative curvature. Let γ be a simple closed geodesic which bounds a domain Ω. We introduce the notion of $\rho(\Omega)$, the radius of domain Ω, by letting

$$\rho(\Omega) = \max\{d(p, \gamma) | p \in \Omega\}.$$

If $K \geq 0$, then we claim

$$(*) \qquad A(\Omega) \geq \tfrac{1}{2} L(\gamma)\rho(\Omega).$$

This can be seen as follows.

Let $\sigma_s = \{q | d(q, \gamma) = s\}$. Using the co-area formula, one gets

$$A(\Omega) \geq \int_0^{\rho(\Omega)} L(\sigma_s)\, ds.$$

Let $f(s) = L(\sigma_s)$. It is clear that $f'(0) = 0$, since σ_0 is a closed geodesic. The function f is also semicontinuous. When f is smooth at $s = s_0$ and $K \geq 0$, then the second variational formula tells us that $f''(s_0) \leq 0$ (cf. [7, p. 20]). Using this fact together with $f(0) = L(\gamma)$ and $f(\rho) \geq 0$, we conclude that $f(s) \geq [1 - s/\rho(\Omega)]L(\gamma)$. This leads us to

$$A(\Omega) \geq \int_0^{\rho} f(s)\, ds \geq \tfrac{1}{2} L(\gamma)\rho(\Omega).$$

Since the assumption that $K \leq 1$ is not needed here, this estimate is independent of the one given by Theorem A.

2. Diameter of a surface and the length of simple closed geodesics

We shall discuss the relation between the length of a simple closed geodesic and the diameter of a Riemannian sphere M. In this section, the curvature K of M is not necessarily nonnegative. M is allowed to have an arbitrary metric.

Theorem B. *Let M be a Riemannian manifold, diffeomorphic to S^2 of diameter D. Then there is a simple closed geodesic γ of length $L(\gamma) \geq 2D$. Furthermore, if the longest simple closed geodesic γ exists and $L(\gamma) = 2D$ holds, then there are infinitely many simple closed geodesics of length $2D$.*

Recall that $L(M)$ is defined to be $\sup\{L(\gamma) | \gamma$ is a simple closed geodesic$\}$. It is well known that the energy functional E defined on the loop space satisfies the Palais-Smale condition. The critical points of E are closed geodesics. Hence, if $L(M) < \infty$, then the set of all simple closed geodesics is a compact subset of $C^0(S^1, M)$. In particular, the longest simple closed geodesic exists as long as $L(M) < \infty$.

Our new observation in this section is based on the proof of the theorem of Lusternik and Schnirelmann, which is due to Ballmann (cf. [2] or the appendix of [15]) and Grayson [11]. The Lusternik-Schnirelmann theorem asserts that, on any two-dimensional sphere with an arbitrary metric, there exist at least three simple closed geodesics. We denote these three geodesics by γ_1, γ_2, and γ_3 and assume that $L(\gamma_1) \leq L(\gamma_2) \leq L(\gamma_3)$. C. Croke observed that $L(\gamma_2) \geq 2D$. The following proof was suggested by him.

Proof of Theorem B. We will adopt some notation from the appendix of [15]. It was known that the second simple closed geodesic γ_2 was found by a minimax method among some special two-parameter families of closed curves. Denote the set of all these two-parameter closed curves by Γ, where Γ is a subset of $C^0([-1, 1] \times [0, 1]; C^0(S^1, S^2))$ and S^1 is the unit circle. The precise definition of Γ will be given later. At the moment, we want to explain the main idea in the proof of Theorem B. Our strategy is to show that

$$\max\{L(u_{t_1,t_2})| -1 \le t_1 \le 1,\ 0 \le t_2 \le 1\} \ge 2D \tag{2.1}$$

holds for each $u \in \Gamma$. This assertion would imply

$$L(\gamma_2) = \inf_{u\in\Gamma} \max_{\substack{-1\le t_1\le 1\\ 0\le t_2\le 1}} \{L(u_{t_1,t_2})\} \ge 2D.$$

Theorem B would follow immediately.

It will take several steps to verify (2.1). First, let us give the precise definition of Γ as follows.

Let $S^2 = \{(x, y, x)|x^2 + y^2 + z^2 = 1\}$. We first introduce u^0, the generator of Γ. All other elements of Γ will be homotopic to u^0 under some extra conditions (see (2.2)–(2.5) below). Our u^0 will consist of *all small* circles on S^2 that meet the yz-plane orthogonally. For any $(t_1, t_2) \in [-1, 1]\times[0, 1]$, we assign a small circle u_{t_1,t_2} of radius $|\cos \frac{\pi}{2}t_1|$ as follows:

$$\begin{aligned} u^0_{t_1,t_2}(s) &= u^0(t_1, t_2, s) \\ &= \left(\cos\frac{\pi}{2}t_1 \cos 2\pi s, \cos\frac{\pi}{2}t_1 \sin 2\pi s, \sin\frac{\pi}{2}t_1\right) \\ &\quad \cdot \begin{bmatrix} 1, & 0, & 0 \\ 0, & \cos\pi t_2, & \sin\pi t_2 \\ 0, & -\sin\pi t_2, & \cos\pi t_2 \end{bmatrix}. \end{aligned}$$

Notice that

(i) For each $t_2 \in [0, 1]$, $u^0_{\pm 1,t_2}$ is a point curve.

(ii) $u^0(t_1, 0, s) = u^0(-t_1, 1, -s)$.

(iii) For any pair of points $\{p, q\} \subset S^2$, there is a small circle in u^0 that passes through p and q.

Naturally, we require that $u \in \Gamma$ if and only if u satisfies:

(2.2) u induces a continuous map from $[-1, 1] \times [0, 1] \times S^1$ to S^2,

(2.3) For each $t_2 \in [0, 1]$, $u_{\pm 1, t_2}$ is a point curve,

(2.4) $u(t_1, 0, s) = u(-t_1, 1, -s)$ for all t_1 and s,

(2.5) there is a continuous homotopy h^λ between u^0 and u, $\lambda \in [0, 1]$, such that $h^0 = u^0$, $h^1 = u$, and each h^λ satisfies conditions (2.2)–(2.4).

It is sufficient to verify the following assertion:

Claim 1. *For each $u \in \Gamma$ and any given pair of points of $p, q \in S^2$, there is a closed curve u_{t_1, t_2} which passes through p and q for some $(t_1, t_2) \in [-1, 1] \times [0, 1]$.*

If one chooses p, q on the Riemannian sphere $M = (S^2, g)$ satisfying $d(p, q) = D$, then inequality (2.1) follows from Claim 1 immediately. The proof of Claim 1 can be carried out by an elementary topological argument below. (Claim 1 is obviously true for u^0.)

Denote $[-1, 1] \times [0, 1] \times S^1 \times S^1$ by T. Inspired by (2.4), we introduce the following equivalence relations

$$(t_1, 0, s_1, s_2) \sim (-t_1, 1, -s_1, s_2),$$
$$(t_1, t_2, s_1, s_2) \sim (t_1, t_2, s_2, s_1).$$

Using the equivalence relations above, we let $\tilde{T}$ be the quotient space of T with the quotient topology and K^3 be the subspace defined by $\{\pm 1\} \times [0, 1] \times S^1 \times S^1 / \sim$. Clearly, $\tilde{T}$ is a four-dimensional orbifold with boundary K^3. In fact, the four-dimensional Z_2-homology of $(\tilde{T}, K^3)$ is nonzero, i.e.,

$$H_4(\tilde{T}, K^3; Z_2) = Z_2 \neq 0,$$

where $Z_2 = 2/2Z$ and Z is the ring of all integers.

On $S^2 \times S^2$ we also introduce an equivalence relation by letting $(p, q) \sim (q, p)$ for any pair of points $p, q \in S^2$. Let $[p, q]$ be the equivalent class of (p, q), $[S^2 \times S^2]$ be the quotient space of $S^2 \times S^2$, and $\Delta = \{[p, p] | p \in S^2\}$. Obviously $[S^2 \times S^2]$ is a four-dimensional orbifold with boundary Δ. Clearly, one sees that $H_4([S^2 \times S^2], \Delta : Z_2) = Z_2$.

For each $u \in \Gamma$, we assign a continuous map F_u from $\tilde{T}$ to $[S^2 \times S^2]$ as follows:

$$F_u(t_1, t_2, s_1, s_2) = [u(t_1, t_2, s_1), u(t_1, t_2, s_2)]. \tag{2.6}$$

It follows from (2.3) that F_u maps K^3 to Δ. Let $\deg_2(F_u)$ denote the Z_2-degree of the map $F_u\colon (\tilde{T}, K^3) \to ([S^2 \times S^2], \Delta)$. In what follows, we will compute $\deg_2(F_u)$.

When $u = u^0$, F_{u^0} is a one-to-one map on the open set

$$G = \{(t_1, t_2, s_1, s_2)/\sim \mid -1 < t_1 < 1,\ 0 < t_2 < 1,\ 0 < s_1 < s_2 < 1\}.$$

Note that G is dense in $\tilde{T}$. Hence, $\overline{G}$ represented a generator of $H_4(\tilde{T}, K^3; Z_2)$, and so does its image. Therefore, we have demonstrated that F_{u^0} induces an isomorphism

$$(F_{u^0})_*\colon H_4(\tilde{T}, K^3; Z_2) \to H_4([S^2 \times S^2], \Delta; Z_2).$$

In particular, the Z_2-degree of the map $F_{u^0}\colon (\tilde{T}, K^3) \to ([S^2 \times S^2], \Delta)$ is nonzero, i.e., $\deg_2(F_{u^0}) = \pm 1 \neq 0 \pmod 2$. It follows from (2.5) that $\deg_2(F_u) \neq 0$ for all $u \in \Gamma$. Hence, Claim 1 has been verified. This completes the proof of (2.1).

When $L(M) = 2D$ holds, we know

$$2D \leq L(\gamma_2) \leq L(\gamma_3) \leq L(M) = 2D.$$

Hence, $L(\gamma_2) = L(\gamma_3)$ holds in the Lusternik-Schnirelmann theorem, which guarantees the existence of infinitely many simple closed geodesics of length $2D$ (cf. [15]).

3. An isoperimetric inequality and its applications

In this section, we will give a nonsharp isoperimetric inequality and derive various estimates for lengths of simple closed geodesics.

We begin with a very crude observation:

Theorem C. *Let M be a Riemannian surface, diffeomorphic to S^2, with nonnegative curvature $K \geq 0$. Then*

$$\text{Area}(M) \leq \frac{8}{\pi} D^2.$$

Proof. This is an easy consequence of the two eigenvalue estimates which are due to Hirsch and Zhong-Yang respectively. Hirsch's theorem tells us that, on any Riemannian sphere M, the first eigenvalue of M, λ_1, is less than or equal to $8\pi/A$, i.e.,

$$\lambda_1 \leq 8\pi/A. \tag{3.1}$$

(See [13] and [22].)

On the other hand, for any Riemannian manifold with nonnegative Ricci curvature, Zhong and Yang found a lower bound of λ_1 in terms of the diameter D (cf. [24]):

$$\lambda_1 \geq \pi^2/D^2 . \tag{3.2}$$

Combining (3.1) and (3.2), we get the desired result.

Remark. Estimate (3.2) is optimal for certain flat tori, not for round two-spheres (cf. [23, p. 114]).

Using Theorems A, B, and C, one can easily derive:

Corollary 0.1. *Let M be a smooth Riemannian sphere with $0 \leq K \leq 1$. Then the longest simple closed geodesic γ exists and*

$$\sqrt{A\pi}/\sqrt{2} \leq L(\gamma) \leq A/2 , \tag{0.1}$$

$$2D \leq L(\gamma) \leq 4D^2/\pi . \tag{0.2}$$

Furthermore, equality $L(\gamma) = A/2$ holds if and only if M is isometric to the unit sphere.

We would like to say a few words about the ratio L/D^2, when D becomes infinite. The following example shows that there is a family of smooth metrics g_i on S^2 whose curvature is bounded by $0 \leq K \leq 1$, in which diameter $D_i \to \infty$ and $L_i/D_i^2 \geq (40\pi)^{-1}$ as $i \to \infty$.

Example 3.1. Let us take two copies of a square rectangle of width $2\pi\kappa$, and then glue them along the boundary smoothly except for corner points. The resulting surface is a flat, nonsmooth two-sphere $\hat{S}^2$ with four exceptional points. Let us call these four points p_1, p_2, p_3, p_4. Around each point p_j we circle out U_j, a neighborhood of p_j. Replace U_j by a quarter of the unit sphere and make the metric smooth under the curvature restriction $0 \leq K \leq 1$ within a metric ball of radius π centered at p_j. Thus, we get a smooth Riemannian sphere of M_k of diameter D_k satisfying $\kappa\pi < D_k \leq 2\pi\kappa\sqrt{5} + \delta$, where δ is independent of κ.

We claim that there is a simple closed geodesic γ_k of length $L(\gamma_k) \geq k\pi\sqrt{16+\kappa^2}/2$ within the flat region of M_k. In order to see this, one needs to view M_k in a different way. Let $\mathbf{Z}$ be the ring of all integers and let us take a lattice $\Gamma\kappa$ in $\mathbf{R}^2$ given by

$$\Gamma_k = \{(4\kappa\pi i_1 + 2\kappa\pi , 4\kappa\pi i_2 + 2\kappa\pi) | i_1 , i_2 \in \mathbf{Z}\} .$$

Clearly, $\mathbf{R}^2/\Gamma_k$ is a flat torus. Notice that the origin $(0, 0)$ is not an element of Γ_k, when κ is a fixed positive integer. Denote the group of $\{\pm \mathrm{Id}\}$ by $\mathbf{Z}_2$, where $-\mathrm{Id}$ is the antipodal map of $\mathbf{R}^2$. It is not hard to see that the orbit space of the group $\mathbf{Z}_2$ acting on $\mathbf{R}^2/\Gamma_k$, say $\mathbf{Z}_2 \setminus (\mathbf{R}^2/\Gamma_k) = \widehat{S}^2$, is homeomorphic to M_k. A fundamental domain of this orbifold can be chosen to be an open rectangle $Q = (0, 2\pi\kappa) \times (-2\pi\kappa, 2\pi\kappa)$. Let us identify Q with the birectangle given at the beginning. Denote the quotient map from $\mathbf{R}^2$ to $\widehat{S}^2$ by F_k. The preimage of singular points $\{p_j\}$ via F_k is exactly $2\pi\kappa(\mathbf{Z}\oplus\mathbf{Z})$. If $\mathfrak{B}$ denotes the union of all disks of radius π whose centers are elements of $2\pi\kappa(\mathbf{Z}\oplus\mathbf{Z})$, then F_k maps $\mathbf{R}^2 - \mathfrak{B}$ to the flat region of (M_k, g_k), which is locally an isometry. Clearly, there is a straight line $\tilde{\gamma}_k$ within $\mathbf{R}^2 - \mathfrak{B}$ which passes through the point $(0, 3k\pi)$ with slope $\kappa/4$. The projection of $\tilde{\gamma}_k$ is the desired closed geodesic γ_k of length $\kappa\pi\sqrt{16+\kappa^2}/2$. Hence, we conclude that $L(\gamma_k)/D_k^2 > (40\pi)^{-1}$ and $D_k > \kappa\pi$, when κ is sufficiently large.

Example 3.2. In Example 3.1, if we replace Γ_k by an affine lattice $\widehat{\Gamma}_k$ of angle 60° and mesh size $4\kappa\pi$, then it is easy to see that the resulting surface is a tetrahedron, which is homeomorphic to S^2 with four singular points. However, in this case, the fundamental domain for this quotient space $\mathbf{Z}_2 \setminus (\mathbf{R}^2/\widehat{\Gamma}_k)$ can also be chosen to be an equilateral triangle of length $4\pi\kappa$. Smoothing the metric along four vertices of the tetrahedron as in Example 3.1, we get a smooth Riemannian sphere $\widehat{M}_k$ of curvature $0 \leq K \leq 1$. One can easily show that $\kappa\pi < D_k \leq 4\kappa\pi + \delta$, where δ is independent of κ. Obviously, there is an affine map from $\mathbf{R}^2$ to $\mathbf{R}^2$, which takes Γ_k to $\widehat{\Gamma}_k$. Playing the same game as before, one can find a simple closed geodesic γ_k in the flat part of $\widehat{M}_k$ of length $L(\gamma_k) = \kappa\pi\sqrt{\kappa^2+4k+16}/2$. This fact implies that $L(\widehat{M})/D_k^2 > (32\pi)^{-1}$ and $D_k > k\pi$, when $\kappa \gg 1$.

Finally, we wish to point out that there is an estimate for the length of the shortest simple closed geodesic, which grows at most linearly in terms of D.

Corollary 0.2. *Let M be a smooth Riemannian sphere of nonnegative curvature. If γ is the shortest simple closed geodesic on M, then $L(\gamma) \leq 9D$.*

Proof. It will be shown in Part II that the shortest closed geodesic is simple on a smooth Riemannian sphere of nonnegative curvature (cf. Theorem D). On the other hand, $L(\gamma) \leq 9D$ holds for the shortest closed geodesic γ (cf. Theorem 4.1 in [10]).

PART II
THE SHORTEST CLOSED GEODESIC IS SIMPLE ON SMOOTH COMPACT CONVEX SURFACES

1. Birkhoff's ideas and rimmed domains

In order to find the shortest closed geodesic, we will work on a variational problem for the space of all one-cycles in a given closed, smooth Riemannian surface. In order to get started, we want to borrow two major ideas from Birkhoff which are related to the ordinary loop space (cf. [2], [10]). The first one we will use is his method of finding closed geodesics on spheres, which is called the minimax method.

Let M be the two-sphere S^2 endowed with a C^3 smooth metric g. Given a piecewise smooth curve γ (not necessarily closed), by $L(\gamma)$ we will mean the length of γ. We denote the injectivity radius of the Riemannian surface M by $\text{inj}(M)$.

If f_t is a one-parameter family of closed curves starting and ending at a point curve in such a way that the induced map $f: S^2 \to S^2$ (see Figure 1) has nonzero degree, then Birkhoff's argument (or 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 curves in this one-parameter family. We shall extend his method to the space of one-cycles to get an analogous minimax principle (cf. §2 of this part).

The second idea that we will use is the Birkhoff case shortening process, B.C.S.P. (which Birkhoff used in the above mentioned argument). The precise definition of B.C.S.P. is given in [15] and [10]. Since we need to derive some new properties of B.C.S.P., we recall it here.

In general, the B.C.S.P. β^N depends on an integer $N > 2$, where N is chosen so large that $L(\gamma)/N$ is smaller than the injectivity radius, $\text{inj}(M)$, of the surface M. For any piecewise smooth closed curve γ, we will define a new curve $\beta^N(\gamma)$ as well as a homotopy γ_s, $s \in [0, 1]$, from $\gamma = \gamma_0$ to $\beta^N(\gamma) = \gamma_1$. The homotopy γ_s will be defined in such a way that $L(\gamma_{s_1}) \leq L(\gamma_{s_2})$ whenever $s_2 \geq s_1$.

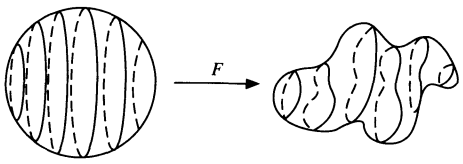

FIGURE 1

We may assume that $\gamma\colon [0, 1] \to M$ is a closed curve parametrized proportional to arc-length; if not, the first part of the homotopy reparametrizes γ 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 $i = 1, 2, \cdots, N$. For $s \in [0, \frac{1}{2}]$, γ_s will be given by

$$\gamma_s(i/N + t) = \tau_i^s(t), \qquad \text{when } 0 \le t \le 2s/N,$$
$$\gamma_s(i/N + t) = \gamma(i/N + t), \qquad \text{when } 2s/N \le t \le 1/N,$$

where τ_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 arc-length. Finally, γ_1 is defined as the unique closed broken geodesic with $\gamma_1(i/N + 1/(2N)) = \gamma_{1/2}(i/N + 1/(2N))$ which is parametrized proportional to arc-length on each interval $[i/N + 1/(2N), (i+1)/N + 1/(2N)]$. We then define γ_s for $s \in [\frac{1}{2}, 1]$ to be a homotopy between $\gamma_{1/2}$ and γ_1 in the same way that γ_s, $s \in [0, \frac{1}{2}]$, homotopes from γ_0 to $\gamma_{1/2}$.

We will apply B.C.S.P.. to some special closed curves which have non-negative geodesic curvatures. More precisely, we let γ be a simple (no self-intersections) closed curve on M which divides M into two components. Let Ω (open) be one of these components.

Definition 1.0. A closed curve γ will be called convex to Ω if there is an $\epsilon > 0$ such that for all $x, y \in \gamma$, with $d(x, y) < \epsilon$, the minimizing geodesic τ from x to y satisfies $\tau \subset \overline{\Omega}$.

The following has been shown in [10].

Lemma 1.1. *Let γ be convex to Ω and have length L, and let $N > L/\operatorname{inj}(M)$ (also $N \gg 2$). Then if we apply B.C.S.P. with N breaks to γ the resulting curves γ_t satisfy*

(1) $\gamma_t \subset \overline{\Omega}$,

(2) *γ_t is simple and convex to $\Omega_t = \Omega - \{x \in \gamma_s | 0 \le s \le t\}$.*

Using Lemma 1.1, we have

Lemma 1.2. *Let γ be convex to Ω and have length L. Then there exists either a simple closed geodesic σ of length $L(\sigma) \le L$, or a homotopy σ_s, $s \in [0, 1]$, which satisfies the following conditions:*

(1.1) *$\sigma_1 = \gamma$, $\sigma_0 =$ a point curve, and $L(\sigma_s) \le L$ for all s,*

(1.2) *$\{\sigma_s | 0 \le s \le 1\}$ gives rise in a natural way to a map F_σ from the two-disk D^2 onto Ω with γ as the boundary, and $_\sigma$ has degree ± 1.*

Proof. Let Λ^L be the space of piecewise smooth closed curves $\alpha\colon [0, 1] \to M$ with length less than or equal to L, and where Λ^L has the C^0-

topology. It is known that the B.C.S.P. $\beta^N: \Lambda^L \to \Lambda^L$ is continuous when $N > (4(\text{inj}(M))^{-1})L$ (cf. [15]). The nontrivial closed geodesics or point curves are the only fixed points of β^N. It can be shown that, for the curve γ given above, the sequence of simple curves $\{\gamma_i\}$ defined by $\gamma_0 = \gamma$ and $\gamma_i = \beta^N(\gamma_{i-1})$ converges to either a simple closed geodesic σ of length $L(\sigma) \leq L$ or a point curve (cf. [10, pp. 4–7]). If the limit is a point curve, then the homotopies $\{\gamma_s\}_{0 \leq s < \infty}$ give rise in a natural way to a map from the two-disk D^2 into Ω with γ as the boundary. In fact, setting

$$D^2 = \{se^{t2\pi i} | (s, t) \in [0, 1] \times [0, 1)\}, \tag{1.3}$$

one can define

$$\sigma_s(t) = \gamma_{(1-s)/s}(t), \quad F_\sigma: D^2 \to \Omega, \quad \text{and} \quad F_\sigma(se^{2\pi i}) = \sigma_s(t).$$

Assertion (1.1) follows from the definition of σ_s. Using Lemma 1.1, we know that F_σ maps D^2 onto Ω with γ as boundary and F_σ has degree ± 1. q.e.d.

In the applications in this paper γ will be a piecewise geodesic curve. In this case, the definition that γ being convex to Ω reduces to the condition that all the angles of γ are convex to Ω.

Let us now consider an interesting example. Suppose $\eta: [0, 1] \to M$ is a nonsimple closed geodesic and M is a sphere. Obviously, η divides M into more than two components, say, $\Omega^1, \Omega^2, \cdots, \Omega^k$, $k \geq 3$. Now, we fix an orientation of M and let $\Omega^1, \Omega^2, \cdots, \Omega^k$ have the inherited orientation from M. For each Ω^i, $1 \leq i \leq k$, it is not hard to see that $\partial\Omega^i$, the boundary of Ω^i, consists of broken geodesics whose angles are convex to Ω^i. This leads us to make the following observation.

Corollary 1.3. *Let η, Ω^i, and $\partial\Omega^i$ be as above. Then there is either a simple closed geodesic σ of length $L(\sigma) < L(\partial\Omega^i)$ or a homotopy α_s, $s \in [0, 1]$, which satisfies*

(1.5) *$\alpha_1 = \partial\Omega^i$, $\alpha_0 =$ a point curve, and $L(\alpha_s) \leq L(\partial\Omega^i)$ for all s.*

(1.6) *$\{\alpha_s\}$ gives rise in a natural way to a map F_α from the two-disk D^2 onto Ω^i with po^i as the boundary of degree ± 1.*

Proof. If $\partial\Omega^i$ has no self-intersection points, Corollary 1.3 follows from Lemma 1.2 immediately. In the case that $\partial\Omega^i$ has self-intersection points $q_1, \cdots, q_n$, $n \geq 1$, and if q_i is one of them, we notice that the angles of $\partial\Omega^i$ at q_i are strictly convex to Ω^i. Without loss of generality, for each q_i, we may assume that there are locally two curves α and β

such that $\max\{L(\alpha), L(\beta)\} < \text{inj}(M)$,

$$\alpha: [a, b] \to \partial\Omega^i, \qquad \beta: [a, b] \to \partial\Omega^i,$$

such that $\alpha(t) \neq \beta(t)$ for $t \neq (a+b)/2$ and $\alpha(a+b)/2 = \beta((a+b)/2) = q_i$. We move α away from β, by the following local homotopy:

$$(1.7)\quad \alpha_s(t)k = \alpha(t), \qquad \text{when } t \notin \left[\frac{a+b-s(b-a)}{2}, \frac{a+b+s(b-a)}{2}\right],$$

$$(1.8)\quad \alpha_s(t) = \tau^s(t), \qquad \text{when } t \in \left[\frac{a+b-s(b-a)}{2}, \frac{a+b+s(b-a)}{2}\right],$$

where τ^s is the minimizing geodesic from $\gamma((a+b-s(b-a))/2)$ to $\gamma((a+b+s(b-a))/2)$ parametrized proportional to the arc-length on the interval $[(a=b-s(b-a))/2, (a+b+s(b-a))/2]$.

Therefore, when $\partial\Omega^i$ has self-intersection points, we choose the first part of the deformation α_s, $s \in [0, \epsilon]$, locally given by (1.7)–(1.8). The resulting new domain Ω^i_ϵ is still connected and the boundary $\partial\Omega^i_\epsilon$ is the union of simple closed curves. Moreover, $\partial\Omega^i_\epsilon$ is convex to Ω^i_ϵ. Let σ be one of those closed curves, $\sigma \subset \partial\Omega^i_\epsilon$. In what follows we shall show that $\partial\Omega^i_\epsilon = \sigma$ if there is no shorter closed geodesics. In fact, Lemma 1.2 tells us that there is either a simple closed geodesic γ with the length $L(\gamma) \leq L(\partial\Omega^i_\epsilon) < L(\partial\Omega^i)$, or there is a homotopy σ_s given in Lemma 1.2. When the second case occurs, one can show that the set $G = \{x \in \sigma_s | 0 \leq s \leq 1\}$ is equal to Ω^i_ϵ by the Jordan curve theorem, since Ω^i_ϵ is path-connected. Therefore, one knows that $\partial\Omega^i_\epsilon = \sigma$. Hence, $\partial\Omega^i$ is connected (by letting $\epsilon \to 0$). This completes the proof of Corollary 1.3.

We remark that the orientation of α_1 in Corollary 1.3 does not necessarily coincide with the original orientation of η.

2. The space of one-cycles and a minimax principle

In this section we will treat nonsimple closed geodesics as one-cycles. In addition, we will discuss a minimax principle (the generalized Birkhoff argument) on the space of one-cycles with integer coefficients over a surface M.

There are many ways to define the topology on the space of one-cycles. At the moment, we use the weak topology which was first introduced by deRham. Roughly speaking, the space of one-cycles can be thought of as a subspace of the "dual" space of one-forms on M.

Let $\Lambda^1(M)$ be the space of all C^∞ smooth one-forms on M. For any oriented one-dimensional rectifiable set T, there is a corresponding linear function defined on $\Lambda^1(M)$ which is

$$T(\varphi) = \int_T \varphi \quad \text{for all } \varphi \in \Lambda^1(M). \tag{2.1}$$

In what follows, T will also be called a rectifiable current.

Furthermore, if a rectifiable current T satisfies

$$T(df) = \int_T df = 0 \tag{2.2}$$

for all smooth functions f, then T is said to be closed, or T is a one-cycle.

The weak topology of the space of one-cycles can be described in the usual way. We only have to define the limit of a sequence of currents $\{T_i\}$. We require that

$$\lim_{i\to\infty} T_i = T \tag{2.3}$$

holds if and only if

$$\lim_{i\to\infty} T_i(\varphi) = T(\varphi) \quad \text{for all } \varphi \in \Lambda^1(M). \tag{2.4}$$

This weak topology is closely related to the "mass" distance $\underline{\mathfrak{M}}$. Recall that g is the metric defined on S^2 which induces a natural pointwise metric on $\Lambda^1(M)$. For any currents T, we introduce

$$\underline{\mathfrak{M}}(T) = \sup\left\{ \int_T \varphi \,\middle|\, |\varphi(x)| \le 1,\ \varphi \in \Lambda^1(M) \text{ and } x \in M \right\}. \tag{2.5}$$

Suppose γ is a piecewise smooth curve and γ has only finitely many self-intersection points. Then one can verify that

$$\mathfrak{M}(\gamma) = L(\gamma) = \text{ the length of } \gamma. \tag{2.6}$$

From now on, we let Z be the ring of all integers and we will work on the set

$$\mathbf{Z}_1(S^2, Z) = \left\{ \sum a_i T_i \,\middle|\, a_i \in Z \text{ is an integer and } T_i \text{ is a rectifiable one-cycle} \right\}. \tag{2.7}$$

$\mathbf{Z}_1(S^2, Z)$ is called a one-dimensional cycle group, or the space of one-cycles with integer coefficients in Z. It has the weak topology mentioned above.

The space of one-cycles $\mathbf{Z}_1(S^2, Z)$ has been studied by Almgren and others. In particular, one of Almgren's results implies the following.

Lemma 2.1. *Let $\Pi_1(\mathbf{Z}_1(S^2, Z), \{0\})$ be the fundamental group of space $\mathbf{Z}_i(S^2, Z)$ with the topology above. Then there is a natural isomorphism*

$$\Psi: \Pi_1(\mathbf{Z}_1(S^2, Z), \{0\}) \to H_2(S^1, Z) = Z \tag{2.8}$$

between $\Pi_1(\mathbf{Z}_1(S^1, Z))$ and $H_2(S^2, Z)$, the two-dimensional homology group of S^2 with integer coefficients in Z.

Proof. This is a special case of the main theorem in [1].

Notice that the isomorphism Ψ above was intuitively taken by using the "evaluation map" F in Figure 1. There are some useful methods to find a one-parameter family of one-cycles σ_t, $t \in [0, 1]$, with the property $\Psi([\sigma]) \neq 0$. For instance, let $f: M \to [0, 1]$ be a Morse function. Then, if we take the sublevel set $D_t = f^{-1}([0, t])$ and the level set $\sigma_t = f^{-1}(t) = \partial D_t$ with the induced orientation from D_t, it is easy to see $\Psi([\sigma])$ is a generator of $H_2(M, Z)$. In some special cases, σ_s and $\psi(\sigma)$ can also be constructed by using broken geodesics. For example, see the following lemma.

Lemma 2.2. *Let η be a nonsimple closed geodesic of length L. Then there exist either a nontrivial closed geodesic σ of length $L(\sigma) < L$, or a one-parameter family of one-cycles σ_t, $t \in [-1, 1]$, satisfying the following conditions:*

$$L(\sigma_t) \leq L \quad \textit{for all } t \in [-1, 1], \tag{2.9}$$

(2.10) *$\{\sigma_t\}$ gives rise in a natural way to a map $F_\sigma: S^2 \to S^2$ of degree ± 1.*

Proof. Since η is a nonsimple closed geodesic, η divides M into k connected components $\Omega^1, \Omega^2, \cdots, \Omega^k$ with $k \geq 3$. Let Ω^i (open) be one of these components. We think of M as a k-legged star fish with legs $\Omega^1, \cdots, \Omega^k$ (see Figure 2).

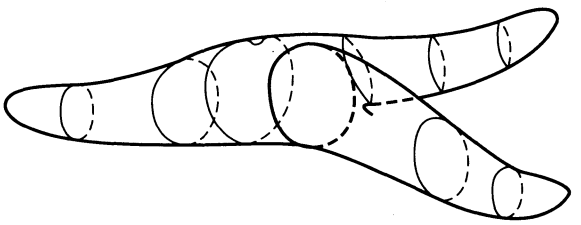

FIGURE 2. 3-LEGGED STAR FISH.

Applying Corollary 1.3 to $\partial\Omega^i$ and Ω^i, we get either a nontrivial closed geodesic σ of length $L(\sigma) < L$, or a homotopy α_s^i which satisfies (1.5)–(1.6). We may assume that there is no nontrivial closed geodesic σ of length $L(\sigma) < L$. The rest of the proof uses $\{\alpha^i\}$, $1 \le i \le k$, to construct σ_t, $t \in [0, 1]$, which satisfy (2.9)–(2.10).

Fix an orientation of M, let Ω^i have the inherited orientation from M, and let Z_2 be the group $Z/2Z$. Since the one-dimensional homology group of M with Z_2 coefficients vanishes, $[\eta] \in H_1(M, Z_2) = 0$, there exist domains $\Omega^{k_1}, \cdots, \Omega^{k_n}$ such that

$$\eta = \sum \partial\Omega^{k_i} \quad \text{in } \mathbf{Z}_1(M, Z_2). \tag{2.11}$$

Let $H_1(M, Z)$ be the first homology group with integral coefficients. By using (2.11), one can also show that

$$\eta = \sum n_i \partial\Omega^{k_i} \quad \text{in } Z_1(S^2, Z) \text{ and } n_i = \pm 1. \tag{2.12}$$

Finally, we choose σ_0 to be

$$\sigma_0 = \sum \partial\Omega^{k_i} \quad \text{in } \mathbf{Z}_1(S^2, Z). \tag{2.13}$$

Applying Corollary 1.3 to σ_0 and Ω^{k_i}, $i = 1, \cdots, n$, one gets half of the homotopy σ_s, $s \in [-1, 0]$. The other half can be derived from the rest of the domains by the same method. We remark that the orientation of σ_0 does not necessarily coincide with the orientation of η, since some of $\{n_i\}$ in (2.12) may be -1. q.e.d.

We are interested in the one-parameter family of one-cycles σ_t given by Lemma 2.2, because there is a minimax principle which works on such families. In general, we shall consider all one-parameter families of one-cycles $\{\sigma_t\}$ which are not homotopic to 0 in the space of one-cycles.

More precisely, let $\sigma\colon [0, 1] \to \mathbf{Z}_1(S^2, Z)$ be a continuous loop in $\mathbf{Z}_1(S^1, Z)$, $\sigma_t = \sigma(t)$, such that

(2.14) σ_t starts and ends at the sum of point curves, hence, $\sigma_0 = \sigma_1 = 0$;

(2.15) $[\sigma] \neq 0$, i.e., σ_t is a nonnull homotopy loop in $\Pi_1(\mathbf{Z}_1(S^2, Z), \{0\})$.

We define a critical value (minimax value) of the mass functional $\underline{\mathfrak{M}}$ to be

$$c_0 = \inf_{[\sigma]\neq 0} \sup_{0\le t\le 1} \{\underline{\mathfrak{M}}(\sigma_t)\} \simeq \inf_{[\sigma]\neq 0} \sup_{0\le t\le 1} \{L(\sigma_t)\}. \tag{2.16}$$

We would like to say a few words about the relations between the length of the nonsimple closed geodesic and the critical value c_0 given in (2.16). Lemma 2.2 tells us that, for any nonsimple closed geodesic γ of $L(\gamma)$,

either γ is not the shortest, or $L(\gamma) \geq c_0$. The next lemma asserts that the equality $L(\gamma) = c_0$ never holds when γ has more than one self-intersection point (counting multiplicities).

Let γ be a closed geodesic. Then γ will be said to have $k-2$ self-intersection points if $M - \gamma$ has k connected components.

Lemma 2.3 (Almgren-Pitts). *Let γ be a nonsimple closed geodesic of length L which has more than one self-intersection. Then there exists either a nontrivial closed geodesic σ of length $L(\sigma) < L$, or a one-parameter family of one-cycles σ_t, $t \in [0, 1]$, satisfying* (2.14)–(2.15) *and* $\max_{0 \leq t \leq 1} L(\sigma_t) < L$. *Furthermore, σ_t can be chosen so that σ_t is made of at most two closed curves for each $t \in [0, 1]$.*

Proof (Pitts). If there is no closed geodesic σ of $L(\sigma) < L$, we will construct a one-family of one-cycles σ_t so that $L(\sigma_t) < L(\gamma)$ and σ_t satisfies (2.14)–(2.15). Such a one-parameter family σ_t could be constructed by the same argument as in the proof of Lemma 2.2 with some modifications.

Let Ω^i, $\partial\Omega^i$, and α^i be as in the proof of Lemma 2.2, $i = 1, \cdots, k$. Since γ has more than one self-intersection point, one knows that $k \geq 4$. Almgren and Pitts, in particular, found the desired σ_t for the case γ has two self-intersections p_1, p_2 and a four-legged star fish $k = 4$ (cf. [19, pp. 35–40]). For $k > 4$, the proof remains the same with little changes. We recall Pitts' argument here, since it will be used in the appendix.

To describe the path σ_t, $t \in [0, 1]$, we subdivide the interval $[0, 1]$ into five subintervals (cf. Figure 3). For $t \in [0, 1/5]$, we bring a cycle up to two of the legs of the star fish; $\sigma(1/5)$ is a figure eight, and so has one self-intersection. From $1/5$ to $2/5$, using (1.4)–(1.5), we "open up" the figure eight $\sigma(1/5)$ by a length-decreasing homotopy. When t is in $[2/5, 3/5)$, σ_t has two components: the curve $\sigma(2/5)$ plus a second cycle moving up the third leg of the star fish, which can be done by using α_s given in Corollary 1.3. At $t = 3/5$, $\sigma(3/5)$ has one self-intersection. From $3/5$ to $4/5$, we once again "open up" the figure eight $\sigma(3/5)$ by a mass-decreasing homotopy; from $t = 4/5$ to $t = 1$, we "slide" the cycle $\sigma(4/5)$ to a point by using Lemma 1.2. It is clear that $\sigma(3/5)$ is the cycle of longest length in the path $\{\sigma_t\}$ and $L(\sigma(3/5)) < L(\gamma) = L$. q.e.d.

The following one-dimensional minimax theorem is due to Almgren and Pitts.

Theorem 2.4 (Almgren-Pits). *Let c_0 be as in* (2.16). *Then there is a nontrivial closed geodesic γ_0 of length $L(\gamma_0) = c_0$. Furthermore, if γ_0 is also the shortest closed geodesic on M, then γ_0 is either simple, or a figure eight.*

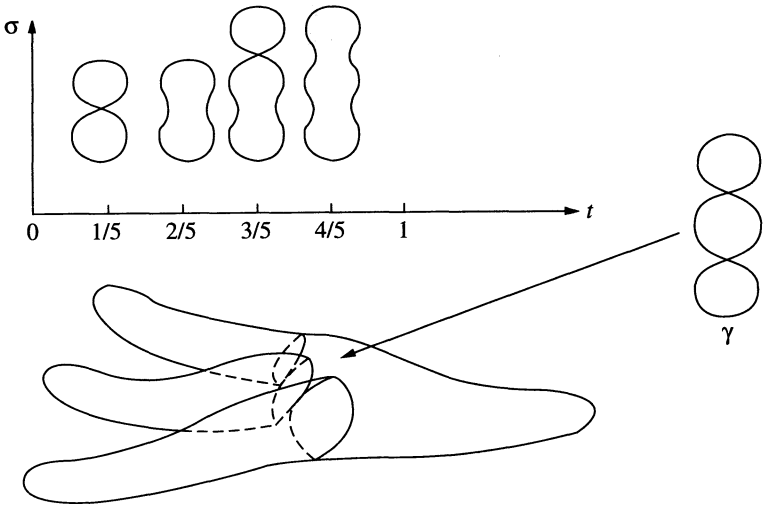

FIGURE 3. 4-LEGGED STAR FISH.

Proof. See [18] and [19, p. 162], or an independent proof in the appendix, which does not use the geometric measure theory.

3. The proof of Theorem D

In this section, we shall show that, on any smooth Riemannian two-sphere of nonnegative curvature, the shortest closed geodesic is simple. In this case, we will also demonstrate that the length of any shortest closed geodesic is equal to the minimax value c_0, given by (2.16).

Notice that, if the curvature is negative somewhere on a given two-sphere, then the length of the shortest closed geodesic is not necessarily equal to the minimax value c_0. Standing as the counterexample is the "dumbbell" (cf. [6]). The "dumbbell" manifold is homeomorphic to S^2, shown in Figure 4 (next page). The pipe connecting the two halves is to be thought of as having fixed length L_0 and varying radius $\epsilon \to 0$. Using the Clairaut's relation, one can verify that the minimax value $c_0 > 2\pi\epsilon$, where $2\pi\epsilon$ is the length of the shortest geodesic.

Now, we are ready to prove

Theorem D. *If g is a C^3-smooth metric on a two-sphere S^2 with nonnegative curvature, then any nontrivial closed geodesic of shortest length is simple.*

Proof. Let γ be a nontrivial closed geodesic of the shortest length. It follows from Lemma 2.3 and Theorem 2.4 that γ is either a figure eight

FIGURE 4. THE DUMBBELL.

or a simple closed geodesic. We will show that γ cannot be a figure eight by a contradiction method. According to the minimax principal (Theorem 2.4), it is sufficient to verify the following assertion:

Claim 1. *If γ were a figure eight, there would be either a shorter closed geodesic σ of $L(\sigma) < L(\gamma)$, or a one-parameter family of one-cycles σ_t, $t \in [-1, 1]$, satisfying*

$$L(\sigma_t) < L(\gamma) \quad \text{for all } t \in [-1, 1]; \tag{3.1}$$

$$\sigma_{-1} = \sigma_1 = 0 \text{ are zeros as one-currents. Moreover, } \sigma_{-1} \text{ and } \sigma_1 \text{ are sums of point curves}; \tag{3.2}$$

$$\sigma_t \text{ is a nonnull homotopy in } \Pi_1(\mathbf{Z}_1(S^2, Z), \{0\}), \tag{3.3}$$

i.e., $[\sigma] \neq 0$.

The construction of σ_t is done in several steps by the same method as in the proof of Lemma 2.2, with some changes. The new ingredient in the construction is to use the second variational formula. First, we need to deform the closed geodesic γ to a new figure eight σ_0 which satisfies (3.1).

Step 1. *A new figure eight σ_0 of length $L(\sigma_0) < L(\gamma)$.* Clearly, the given γ violates restriction (3.1). Hence, we have to replace γ by a "nearby" figure eight with smaller length. The desired σ_0 can be found by using the exponential map along the parallel normal vector field $\vec{n}$ of γ.

By the assumption, the curvature K of the metric g is nonnegative and M is homeomorphic to S^2. It follows from the Gauss-Bonnet formula that there is a nonempty open set G in M on which the curvature is strictly positive, i.e., $G = \{x | K(x) > 0\} \neq \varnothing$. If the metric is flat in a neighborhood of γ, $K = 0$, then one can move γ by a translation to get a family of new closed geodesics of figure eight type. Such a translation will stop until a new geodesic hits $\overline{G}$, the closure of G. Therefore, we may assume that there is a point q on γ such that $q \in \overline{G}$.

Let $\vec{n}(s)$ be the smooth unit normal vector field along $\gamma(t)$ pointing inward to G. Now we define a small perturbation of γ by the exponential map:

$$f_t(s) = \exp_{\gamma(s)}[t\vec{n}(s)]. \tag{3.4}$$

For the same reason as in the proof of Lemma 1.2 in Part I, we see that there is an $\epsilon > 0$ with

$$L(f_t) < L(\gamma) \quad \text{for all } t \text{ with } 0 < t < 2\epsilon. \tag{3.5}$$

For simplicity, we change the parametrization of f_ϵ so that $f_\epsilon : [-1, 1] \to S^2$ and $f_\epsilon(0) = f_\epsilon(-1) = f_\epsilon(1) = p_0$, which is the unique intersection point.

Step 2. *Cut and paste* f_ϵ *at its intersection point* p_0. In order to get a new closed curve σ_0 which is a limit of simple closed curves, we let

$$f^+ = f_\epsilon|_{[0,1]}, \quad f^- = f_\epsilon|_{[-1,0]}, \quad \sigma_0 = f^+ - f^-,$$
$$\eta^- = -f^-|_{[\epsilon-1,-\epsilon]}, \qquad \eta^+ = f^+|_{[\epsilon,1-\epsilon]},$$
$$p_1 = \sigma_0(\epsilon - 1), \quad p_2 = \sigma_0(\epsilon), \quad p_3 = \sigma_0(1-\epsilon), \quad p_4 = \sigma_0(-\epsilon).$$

For any pair of points $\{p, q\}$ with $d(p, q) < \frac{1}{2}\operatorname{inj}(M)$, we denote the minimal geodesic from p to q by $\overrightarrow{pq}$. Finally, we define

$$\sigma_{-\epsilon} = \eta^- + \overrightarrow{p_1p_2} + \eta^+ + \overrightarrow{p_3p_4}, \qquad \sigma_\epsilon = \eta^- + \overrightarrow{p_1p_4} + \eta^+ + \overrightarrow{p_3p_2}.$$

One can find the homotopy between $\sigma_{-\epsilon}$ and σ_ϵ, say σ_t, $t \in [-\epsilon, \epsilon]$, such that $L(\sigma_t) \le L(\sigma_0) < L(\gamma)$ for $t \in [-\epsilon, \epsilon]$. This part of the construction of $\{\sigma_t\}$, $t \in [-\epsilon, \epsilon]$, is straightforward.

Our final step is to shrink σ_ϵ and $\sigma_{-\epsilon}$ to the point curves.

Step 3. *The deformation from* σ_ϵ *and* $\sigma_{-\epsilon}$ *to point curves.* We cannot apply Lemma 1.2 to σ_ϵ or $\sigma_{-\epsilon}$ directly, since $\sigma_{\pm\epsilon}$ may not be convex to the domain it encloses. This leads us to make the following observations on each part of $\sigma_{-\epsilon}$ and σ_ϵ. Let

$$\gamma^- = \gamma|_{[\epsilon-1,-\epsilon]}, \qquad \gamma^+ = \gamma|_{[\epsilon,1-\epsilon]},$$
$$q_1 = \gamma(\epsilon - 1), \quad q_2 = \gamma(\epsilon), \quad q_3 = \gamma(1-\epsilon), \quad q_4 = \gamma(-\epsilon),$$
$$\sigma_{2-\epsilon} = \gamma^{-1} + \overrightarrow{q_1q_2} + \gamma^+ + \overrightarrow{q_3q_4}, \qquad \sigma_{2\epsilon} = \gamma^- + \overrightarrow{q_1q_4} + \gamma^+ + \overrightarrow{q_3q_2}.$$

Obviously, there is a homotopy of σ_t, $t \in [\epsilon, 2\epsilon] \cup [-2\epsilon, -\epsilon]$, such that

$$L(\sigma_t) \le L(\gamma) - \delta < L(\gamma), \tag{3.6}$$

where δ is a positive constant which depends on

$$\min_{1 \le i \le 3} \{d(q_i, \sigma_0(0)) + d(q_{i+1}, \sigma_0(0)) - d(q_i, q_{i+1})\} > 0.$$

Because each component $\sigma_{2\epsilon}$ and $\sigma_{-2\epsilon}$ is a broken closed geodesic which is convex to the domain it encloses, one can use Lemma 1.2 to find either a shorter closed geodesic σ of $L(\sigma) < L(\gamma)$, or the rest of σ_t, $t \in [-1, -2\epsilon] \cup [2\epsilon, 1]$, such that

$$L(\sigma_t) < L(\gamma) \quad \text{for all } t \in [-1, -2\epsilon] \cup [2\epsilon, 1]. \tag{3.1}$$

(3.2) $\sigma_{-1} = \sigma_1 = 0$ are zeros as one-currents. Moreover, they are sums of point curves.

Combining Steps 1–3, we get the $\{\sigma_t\}$, $t \in [-1, 1]$, as claimed in (3.1)–(3.3).

This finishes the proof of Claim 1. Thus, Theorem D follows from Lemma 2.3 and Theorem 2.4 immediately.

We conclude this part by studying the length of the shortest closed geodesic.

Theorem 3.1. *Let g be the metric on the two-sphere S^2 with nonnegative curvature $K \geq 0$ and let γ be a nontrivial closed geodesic of shortest length L. Then L is equal to the minimax value c_0 given in* (2.16).

Proof. It follows from Theorem 2.4 that $L(\gamma) \leq c_0$. Hence, we only have to demonstrate $L(\gamma) \geq c_0$. For a closed geodesic γ of the shortest length, it suffices to construct a one-parameter family of closed curves f_t, $t \in [-1, 1]$, starting and ending at a point curve in such a way that the induced map $F: S^2 \to S^2$ (see Figure 1) has nonzero degree, $L(f_t) \leq L(\gamma)$ and $f_0 = \gamma$.

Theorem D tells us that γ is simple. Therefore, γ divides S^2 into two components Ω^- and Ω^+. We will shrink $\gamma = \partial\Omega^+$ to a point curve with the domain Ω^+ by f_t, $t \in [0, 1]$. One can play the same game on the other "half" of S^2, Ω^-.

The perturbation f_t of γ, for $t \geq 0$, will be carried out by the same method as in Step 1 in the proof of Theorem D. Suppose $\gamma: [0, 1] \to M$ is a parametrization proportional to the arc-length, and let $\vec{n}(s)$, $s \in [0, 1]$, be the unit normal vector field which points inward to Ω^+. Let $G = \{x | x \in \Omega^+ \text{ and } K(x) > 0\}$. For the same reason as above, we may assume that $\gamma \cap \overline{G} \neq \varnothing$ and $\vec{n}$ points inward to G. Let

$$F(s \cdot t) = f_t(s) = \exp_{\gamma(s)}[t\vec{n}(s)], \qquad s \in [0, 1]. \tag{3.4}$$

Using Lemma 1.4 of Part I and $K > 0$ on G, one can show that there is an $\epsilon > 0$ and

$$L(f_t) < L(\gamma) \quad \text{for all } t \text{ with } 0 < t < \epsilon. \tag{3.7}$$

Now applying the Birkhoff curve shortening process to f_ϵ, we get a limiting curve f_1. Since γ is the shortest closed geodesic, f_1 has to be a point curve. Because $\partial\Omega^+$ is convex to Ω^+ (cf. §1), we know that $f_1 \in \Omega^+$. Using the same argument as in the proof of Lemma 1.2, we get a homotopy f_t, $t \in [\epsilon, 1]$, which satisfies $L(f_t) \leq L(f_\epsilon) < L(\gamma)$. This gives the desired f_t for $0 \leq t \leq 1$. The remaining part of f_t, $-1 \leq t \leq 0$, can be produced by the same method. Hence, we get $\{f_t\}_{-1\leq t\leq 1}$ which induces a map $f: S^2 \to S^2$ of degree ± 1, $L(f_t) \leq L(\gamma)$ and $f_0 = \gamma$. This completes the proof of Theorem 3.1.

Appendix

In this appendix, we will give an alternative proof of the minimax principal for the space of one-cycles (cf. Theorem 2.4) without using the geometric measure theory. All notation remain the same as in§2 of Part II unless otherwise specified.

The main tool that we will use is the variational method. However, there are some technical difficulties. Recall that we introduced the "mass" functional $\underline{\mathfrak{M}}$ on the space of one-cycles (cf. (2.5)). The "mass" functional $\underline{\mathfrak{M}}$ is only a semicontinuous function with respect to the weak topology on $\mathbf{Z}_1(S^2, Z)$. For example, let $\{(x, y)\}$ be the geodesic normal coordinates at a given point p, and let σ_ϵ be the ellipsoid given by $(x/c)^2+(y/\epsilon)^2 = 1$ with counter-clockwise orientation, where $c < \frac{1}{2}\operatorname{inj}(M)$ is a constant. One can see that, in such a family of one-cycles, σ_ϵ,

$$\text{(A.1)} \quad \sigma_\epsilon \to 0 \quad \text{in } \tilde{Z}^1(S^2, Z) \qquad \text{and} \qquad \underline{\mathfrak{M}}(\sigma_0) = 0 \quad \text{as } \epsilon \to 0;$$

$$\text{(A.2)} \qquad \underline{\mathfrak{M}}(\sigma_\epsilon) \geq 2c \neq 0 \quad \text{for all } \epsilon > 0.$$

In order to get around the unpleasant situation above, we need to replace the mass functional $\underline{\mathfrak{M}}$ by the length functional L on the space of all parametrized one-cycles. The length functional L has some nice properties. For instance, the sum of the lengths of two curves is independent of the choice of orientations on two curves. We also notice that, for any finite number of paths $\varphi_1, \varphi_2, \cdots, \varphi_n$, $\varphi_i: [0, 1] \to S^2$, with the constraint $\sum[\varphi_i(1) - \varphi_i(0)] = 0$, the sum of these paths gives us a one-cycle $\sum \varphi_i$. For the purpose of this appendix, Lemma 2.3 tells us that it is sufficient to consider all one-cycles which have at most two connected components. Therefore, we are going to work on the set of all pairs of paths $\{(\varphi_1, \varphi_2)\}$ with the constraint

$$\text{(A.3)} \qquad \varphi_1(1) + \varphi_2(1) - \varphi_1(0) - \varphi_2(0) = 0.$$

For simplicity, we denote this set by Γ:

(A.4) $\Gamma = \{\Phi | \Phi = (\varphi_1, \varphi_2),\ \varphi_i : [0, 1] \to S^2$ is a piecewise smooth path and $\sum \varphi_i(1) - \sum \varphi_i(0) = 0$, i.e., Φ satisfies (A.3)$\}$.

Notice that there is a natural map which takes Γ to the space of one-cycles:

(A.5) $$\mathscr{P} : \Gamma \to \mathbf{Z}_1(S^2, Z)$$
$$(\varphi_1, \varphi_2) \to \varphi_1 + \varphi_2 .$$

When $\Phi = (\varphi_1, \varphi_2)$, by $L(\Phi)$ we mean the sum $L(\varphi_1) + L(\varphi_2)$. We would like to define a strong topology on Γ so that $\mathscr{P}$ and L become continuous maps. One good choice is the modification of the Fréchet topology.

We shall now define the modified Fréchet distance $\hat{d}(\Phi, \tilde{\Phi})$ between two elements $\Phi = (\varphi_1, \varphi_2)$ and $\tilde{\Phi} = (\tilde{\varphi}_1, \tilde{\varphi}_2) \in \Gamma$ as follows. For simplicity, let $U_\epsilon = [0, \epsilon] \cup [1 - \epsilon, 1]$. We may also assume that the surface M is isometrically embedded in the Euclidean space R^n by Nash's theorem. The distance $\hat{d}$ consists of two parts. First, we consider the matched pair of arcs:

(A.6) $$\underline{d}_{\epsilon_i}(\Phi, \tilde{\Phi}) = \max_{\epsilon_i \le t \le 1-\epsilon_i} d(\varphi_i(t), \tilde{\varphi}_i(t)) + \int_{\epsilon_i}^{1-\epsilon_i} |\varphi_i'(t) - \tilde{\varphi}_i'(t)|^2 \, dt .$$

The residue part can be measured by

(A.7) $$\underset{\sim}{d}_{\epsilon_i}(\Phi, \tilde{\Phi}) = L(\varphi_i|_{U_{\epsilon_i}}) + L(\tilde{\varphi}_i|_{U_{\epsilon_i}}) .$$

Finally, we define $\hat{d}$ to be the greatest lower bound for every choice of $\epsilon_i \in [0, 1]$, $i = 1, 2$,

(A.8) $$\hat{d}(\Phi, \tilde{\Phi}) = \inf_{0 \le \epsilon_i \le 1} \left\{ \sum_i \underline{d}_{\epsilon_i}(\Phi, \tilde{\Phi}) + \sum_i \underset{\sim}{d}_{\epsilon_i}(\Phi, \tilde{\Phi}) \right\} .$$

From now on, we will simply denote the topological space $(\Gamma, \hat{d})$ by Γ. It follows from the definition of $\hat{d}$ that $\mathscr{P}$ is in fact a continuous map from Γ to $\mathbf{Z}_1(S^2, Z)$. Since Γ is relatively easy to deal with, we will transfer our geometric problem on $\mathbf{Z}_1(S^2, Z)$ to a variational problem on Γ via the map $\mathscr{P}$.

For any given path $\underline{\Phi} : [0, 1] \to \Gamma$, if $\mathscr{P}(\underline{\Phi})$ satisfies conditions (2.14)–(2.15), then $\underline{\Phi}$ is said to be admissible, or briefly $[\underline{\Phi}] \neq 0$.

For the set of all admissible paths in Γ, one can find the following fact about its minimax value.

Lemma A.1. *Let c_0 be the minimax value of the mass function $\underline{\mathfrak{M}}$ on $\mathbf{Z}_1(S^2, Z)$ given in* (2.16). *Then*

$$c_0 = \inf_{[\Phi]\neq 0} \max_{0\leq s\leq 1} L(\Phi^s). \tag{A.9}$$

Proof. Using the same argument as in the proof of Lemma 2.3 of Part II, one can show that c_0 can be taken among all nontrivial paths $\{\sigma^s\}$, $s \in [0, 1]$, in the subspace $\mathscr{P}(\Gamma)$ of $\mathbf{Z}_1(S^2, Z)$, where $\mathscr{P}(\Gamma)$ is the set of all one-cycles which have at most two components.

Recall that, if Φ only has finitely many self-intersections, then, by (2.7), one knows $\underline{\mathfrak{M}}(\mathscr{P}(\Phi)) = L(\Phi)$ .1 The set of all such Φ is a dense subset of Γ. Hence, (A.9) holds.

Proof of Theorem 2.4. In order to carry out the proof, we need to consider the variation of the length functional L on Γ. Let $\mathbf{X}(M)$ be the set of all smooth vector fields on M. For any $X \in \mathbf{X}(M)$, X generates a one-parameter group of diffeomorphism h_t (cf. [16, p. 10]). The derivation of L at $\Phi \in \Gamma$ in direction X is defined by

$$\delta L_\Phi(X) = \frac{d}{dt}(L(h_t \circ \Phi))|_{t=0}. \tag{A.10}$$

If $\delta L_\Phi(X) = 0$ for all $X \in \mathbf{X}(M)$, then Φ is called a critical point of L in Γ and $L(\Phi)$ is called a critical value of L.

Our next step is to show that c_0 is a critical value of L. This can be done by a finite-dimensional approximation of a subspace of $\Gamma^{c_1} = \{\Phi | \Phi \in \Gamma,\ L(\Phi) \leq c_1\}$, where $c_1 > c_0$ is a constant. Such an approximation can be described as follows. Choose an integer $N > 2$ so large that $c_1/N > \text{inj}(M)$. For any $\Phi = (\varphi_1, \varphi_2) \in \Gamma^{c_1}$, applying the first half of B.C.S.P. to each φ_i, $i = 1, 2$, one can get a new element $\Phi^{1/2} \in \Gamma$, as well as a homotopy Φ^s, $s \in [0, 1/2]$, from $\Phi^0 = \Phi$ to $\Phi^{1/2}$ (cf. §1 of Part II). The homotopy was defined in such a way that $L(\Phi^s) \leq L(\Phi^\tau)$ whenever $s \geq \tau$. Hence, if $\mathscr{N}$ is the product of $2N$ copies of M, then Γ^{c_1} is homotopic to a compact quotient space of $\mathscr{N}$. Using the same argument as in [16, pp. 88–100], one can show that c_0 is a critical value of L.

Suppose that $\Phi = (\varphi_1, \varphi_2)$ is a critical point in Γ with $L(\Phi) = c_0$; we want to show that $\varphi_1 \cup \varphi_2$ forms a closed geodesic. When Φ is a critical point, φ_1 and φ_2 have to be geodesic paths. Since $\varphi_1 + \varphi_2$ is a closed one-cycle, endpoints of φ_1 and φ_2 have to meet at at most two distinct points. When the set $\{\varphi_1(0), \varphi_1(1), \varphi_2(0), \varphi_2(1)\}$ has exactly two distinct points, it is easy to verify that φ_1 and φ_2 form a closed geodesic or a union of closed geodesics (or possibly point curves). If endpoints φ_1 and φ_2 meet

at a single point and both are nontrivial geodesic loops, then we let

$$e_1 = \varphi_1'(0)/\|\varphi_1'(0)\|, \qquad e_2 = \varphi_1'(1)/\|\varphi_1'(1)\|,$$
$$e_3 = \varphi_2'(0)/\|\varphi_2'(0)\|, \qquad e_4 = \varphi_2'(1)/\|\varphi_2'(1)\|.$$

Using the fact $\delta L_\Phi(X) = 0$ for all $X \in \mathbf{X}(M)$ and the first variational formula in [7, p. 24], one knows that

$$e_1 - e_2 + e_3 - e_4 = 0. \tag{A.11}$$

Using (A.11) and the fact that $\|e_i\| = 1$, $0 \leq i \leq 4$, one may assume that $e_1 = -e_3$ and $e_2 = -e_4$ after reindexing e_i, $i = 1, 2, 3, 4$, suitably. Therefore, $\varphi_1 \cup \varphi_2$ forms a closed geodesic after changing the orientation of φ_1 appropriately.

Moreover, if there exists the shortest closed geodesic γ_0 of length $L(\gamma_0) = c_0$ on (S^2, g), then using Lemma 2.3 and the argument above, we conclude that γ_0 has at most one self-intersection. This completes the proof of the minimax principle.

Remark. Suppose M is a closed smooth Riemannian surface of genus k and γ is a nontrivial closed geodesic of shortest length among its homology class $[\gamma]$. Then using the argument above, one can estimate the number of self-intersections of γ. In particular, if $[\gamma] = 0$ in $H_1(M, Z)$, then γ has at most $2k + 1$ self-intersections.

References

[1] F. J. Almgren, *The homotopy groups of the integral cycle groups*, Topology **1** (1960) 257–299.

[2] W. Ballmann, *Der Satz von Lusternik und Schnirelmann*, Bonner Math. Schriften **102** (1978), 1–25.

[3] W. Ballmann, G. Thorbergsson & W. Ziller, *Closed geodesics on positively curved manifolds*, Ann. of Math. (2) **116** (1982) 213–247.

[4] V. Bangert, *Manifolds with geodesic chords of constant length*, Math. Ann. **265** (1983) 273–281.

[5] M. Berger, *Blaschke's conjecture for spheres*, Manifolds, All of Whose Geodesics are Closed (A. Besse, ed.), Ergebnisse Math. u. i. Grenzgeb., Vol. 93, Appendices D and E, Springer, Berlin, 1978.

[6] J. Cheeger, *A low bound for the smallest eigenvalue of the Laplacian*, Problems in Analysis, A Symposium in Honor of Salomon Bochner (R. C. Gunning, ed.), Princeton University Press, Princeton, NJ, 1970.

[7] J. Cheeger & D. Ebin, *Comparison theorems in Riemannian geometry*, North-Holland Math. Library, North-Holland, Amsterdam, 1975.

[8] C. Croke, *Some isoperimetric inequality and eigenvalue estimates*, Ann. Sci. École Norm. Sup (4) **13** (1980) 419–435.

[9] ——, *Poincare's problem and the length of shortest closed geodesic on a convex hypersurface*, J. Differential Geometry **17** (1980) 595–634.

[10] ——, *Area and the length of shortest closed geodesic*, J. Differential Geometry **27** (1988) 1–22.
[11] M. Grayson, *Shortening embedded curves*, Ann. of Math. **129** (1989) 71–111.
[12] E. Heintze & H. Karcher, *A general composition theorem with applications to volume estimates for submanifolds*, Ann. Sci. École Norm. Sup. (4) **11** (1978) 451–470.
[13] J. Hersch, *Quarter proprietes isoperimetriques de membrances spheriques homogenes*, C. R. Acad. Sci. Paris Sér. A **270** (1970) 1645–1648.
[14] J. L. Kazdan, *An isoperimetric inequality and Wiedersehen manifolds*, Seminar on Differential Geometry (S. Y. Yau, ed.), Ann. of Math. Studies, No. 102, Princeton University Press, Princeton, NJ, 1982, 525–537.
[15] W. Klingenberg, *Lectures on closed geodesics*, Appendix, Springer, Berlin, 1978, 203–219.
[16] L. A. Lyusternik, *The topology of function spaces and the calculus of variations in the large*, Trudy Mat. Inst. Steklov. **19** (1949); English transl., Transl. Math. Monographs, Vol. 16, Amer. Math. Soc., Providence, RI, 1966.
[17] J. Milnor, *Morse theory*, Annals of Math. Studies, No. 51, Princeton University Press, Princeton, NJ, 1963.
[18] J. Pitts, *Regularity and singularity of one dimensional stational integral varifolds on manifolds arising from variational methods in the large*, Symposia Mathematics, Vol. XIV, Roma, Italy, 1974.
[19] ——, *Existence and regularity of minimal surfaces on Riemannian manifolds*, Math. Notes, No. 27, Princeton University Press, Princeton, NJ, 1981.
[20] L. A. Santalo, *Integral geometry in general spaces*, Proc. Internat. Congr. Math. (Cambridge 1950), Vol. I, Amer. Math. Soc., Providence, RI, 1952, 483–489.
[21] W. P.Thurston, *The topology and geometry of 3-manifolds*, Lecture Notes in Math., preprint, Princeton University, 1979.
[22] P. Yang & S. T. Yau, *Eigenvalues of the Laplacian of compact surfaces and minimal submanifolds*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **7** (1980) 55–63.
[23] S. T. Yau, *Nonlinear analysis in geometry*, Enseignement Math. **33** (1987) 109–138.
[24] J. Q. Zhong & H. C. Yang, *On the estimate of first eigenvalue of a compact Riemannian manifold*, Sci. Sinica, Ser. A **27** (1984) 1252–1265.

UNIVERSITY OF PENNSYLVANIA
CORNELL UNIVERSITY