Theorem of Three Geodesics on S^2 using Curve-Shortening Flow

Marek Antoni Kurczyński

December 8, 2020

Abstract

This paper contains a short, complete, proof that every Riemannian metric on the 2-sphere admits at least 3 simple closed geodesics, using curve shortening flow.

Acknowledgements

This research was made possible by the generous supervision of Jason D. Lotay, Professor of Pure Mathematics at the University of Oxford.

1 Introduction

The theorem of three geodesics states that there exist three simple, closed geodesics on any Riemannian surface with the topology of the sphere S^2. The theorem was first claimed by L. Lusternik and L. Shnirelman [7] in their seminal work in 1929, but was later found to be incomplete. The work of M. Morse [9] and W. Klingenberg [6] introduced new methods of 'calculus of variations in the large' and development of algebraic topology to tackle this theorem and its generalities, but a correct solution continued to be lacking, partly because of mistakes in the algebro-topological constructions used. Another attempt at a solution may be found in I.A. Taimanov's work [10], building upon ideas of J. Jöst [5].

Main Theorem 1.1. *On a surface* (S^2, g)*, g a Riemannian metric, there exist three simple closed geodesics.*

Curve shortening flow (CSF) on surfaces began to be studied in the late 1980s by M.A. Grayson[4], M.E. Gage [3], R. Hamilton and others. In his paper, Grayson suggested an outline of using the powerful machinery of CSF to prove the theorem of the three geodesics, but it is lacking crucial details. We give a novel proof of the theorem of three geodesics using the CSF on surfaces, effectively bypassing arguments of $O(2)$-actions, cohomology and Morse Theory which caused the demise of many previously attempted proofs. It is hoped the methods used here may inspire uses to more advanced problems.

2 Facts of Curve Shortening Flow

The CSF acts on simple closed curves $\gamma(s,t)$, $\gamma : S^1 \times [0,T) \to M$, of a surface M by the non-linear PDE:

$$\partial_t \gamma = kN$$

where k is the curvature of the curve and N is the normal to γ. It is straightforward to find that for L the length of γ,

$$\partial_t L = -\int_{S^1} k^2 ds \tag{1}$$

and hence the CSF is decreasing the length of the curve as fast as possible, only using the local information of the curve. We only care about the case of M being a compact oriented surface, then some important properties of CSF include:

- Embedded curves stay embedded [3].
- If $\gamma(s,0)$ is an embedded curve and $\gamma(s,t)$ exists for infinite time, then there's a subsequence $\gamma(s,t_n)$ that converges to a single covering of an embedded geodesic γ_∞ in the C^∞ sense [3]. If $\gamma(s,t)$ does not exist for infinite time then it must shrink to a point in finite time.
- It may not be that the geodesic limit is unique, but any two limits must have the same length and intersect at least once [4].

3 The Loop Space

Lusternik and Shnirelman's original deformation [7] was problematic because the retraction was only defined locally and then assumed to be able to carry on globally. Later on, correct retractions were eventually found, as cited below, with Taimonov's [10] in particular providing an example for our topology on the loop space. Such a retraction gives three homology classes on the loop space of $M^2 := (S^2, g)$, for which we give explicit representations.

We endow a topology on a loop space with the metric

$$d(p,q) = \inf \max_{s \in S^1} d_g(p(s), q(s)) + |L(p) - L(q)|$$

where the infimum is taken over all parametrisations of the curves, d_g the induced metric on M^2. The CSF is then continuous and the second term ensures the length functional is also continuous (see Milnor [8], Ch. 3).

Definition 3.1. We define $\Lambda S^2 = \{\gamma : S^1 \to S^2 | \gamma \in C^\infty \text{ and is simple}\}$ with the usual identification of free loops. We consider the point curves of S^2 to also be simple and hence inside ΛS^2. Denote $\Lambda_0 S^2$ as the point curves and $(\Lambda S^2, \Lambda_0 S^2)$ is the topological pair.

The smooth simple loops of M^2 are smooth simple loops in (S^2, g), and point curves in M^2 are point curves in (S^2, g), so that finding the homology of $(\Lambda S^2, \Lambda_0 S^2)$ is sufficient.

3.1 Retracting $(\Lambda S^2, \Lambda_0 S^2)$

We will find a finite-dimensional retraction of this space in order to find its homology classes. We take S^2 as a subset of $\mathbb{R}^3$, radius 1, centre the origin. It has the round metric from the embedding. The conditions of a loop being smooth implies the curvature is bounded and so the theorems of Gage and Grayson can be applied on the CSF of these loops. We use the round metric to find deformations using the CSF, but our results are topological and for any metric.

Definition 3.2. Denote ΛM^α as the subset of the loop space of loops with length less than or equal to α. For notational clarity, ΛS^α are loops in S^2 with length less than or equal to α, and $\Lambda S := \Lambda(S^2)$ defined as before.

Lemma 3.3. *There's a strong deformation retraction from ΛS to $\Lambda S^{2\pi+\varepsilon}$, for any $\varepsilon > 0$.*

Proof. By the properties of CSF, denoted ϕ_t, for any $\gamma \in \Lambda S \setminus \Lambda S^{2\pi+\varepsilon}$ after a finite time T_γ, $\phi_{T_\gamma}(\gamma)$ is of length $2\pi+\varepsilon$. The CSF deforms continuously so this gives isotopies for any such length curve. We therefore get a homotopy $H : \Lambda S \times I \to \Lambda S$, in the obvious manner:

$$H(s,t) := \begin{cases} \phi_{tT_\gamma}(\gamma) & \text{for } L(\gamma) > 2\pi, \\ \gamma & \text{otherwise.} \end{cases}$$

□

See [2], [6], or [10] for retractions from $\Lambda S^{2\pi+\varepsilon}$ to the circles of S^2, by a circle we mean any non-empty intersection of S^2 and a plane in the ambient $\mathbb{R}^3$. Denote this space of circles as $\mathscr{C}$. In the appendix we show how $(\mathscr{C}, \Lambda_0 S^2)$ is diffeomorphic to $(\mathbb{R}P^3 - D^3, \partial)$.

3.2 Homology $H_*(\Lambda M, \Lambda_0 M; \mathbb{Z}_2)$

Suppose we have one of the homology functors H_*, and a retraction $r : X \to A$, with inclusion $\iota : A \to X$. Then by functoriality we have $Id_{H_*(A)} = H_*(r \cdot \iota) = H_*(r) \cdot H_*(\iota)$, so $H_*(r)$ is surjective and $H_*(\iota)$ is injective. By the injectivity of $H_*(\iota)$ we therefore have three homology classes, with coefficients in $\mathbb{Z}_2$ (cf. appendix), of dimension 1, 2, and 3 in $(\Lambda S^2, \Lambda_0 S^2)$, not just the retraction, which we'll call h_1, h_2, and h_3 respectively. We give explicit representations of these homology classes in the appendix.

Take the minmax values

$$\kappa(h_i) = \inf_{\Gamma \in h_i} \max_{\gamma \in \Gamma} L(\gamma)$$

where we denote Γ as a cycle in one of the homology classes, and γ as a loop in ΛM. For non-trivial homology class h, we must have $\kappa(h) > 0$ or there would be a sequence of cycles arbitrarily close to the point curves and hence eventually trivial.

A cycle representing a homology class can be made to include any cycles of the previous homology class (c.f. appendix) and by this inclusion we have the inequality:

$$0 < \kappa(h_1) \leq \kappa(h_2) \leq \kappa(h_3). \tag{2}$$

4 Existence of three simple closed geodesics

The arguments in this section are all original, with the proof of Lemma 4.1 influenced by [1].

Lemma 4.1. *There exists a geodesic of length the minmax value $\kappa(h)$ of a homology class h in the loop space.*

Proof. If there's no geodesic length $\kappa(h)$ we claim $\exists \varepsilon > 0$ such that

$$L(\gamma_0) - L(\gamma_1) \geq 2\varepsilon \ \forall \gamma \in \Lambda M^{\kappa(h)+\varepsilon} \setminus \Lambda M^{\kappa(h)-\varepsilon}.$$

Note by equation (1) we have $L(\gamma_0) - L(\gamma_1) = -\int_0^1 \partial_t L \, dt = \int_0^1 \int_{S^1} k^2 \, ds \, dt$. Then take a $\delta > 0$ such that there's no other geodesics of length in the interval $[\kappa(h) - \delta, \kappa(h) + \delta]$. The loops of length in $[\kappa(h) - \delta, \kappa(h) + \delta]$ is a closed set so we cannot have a sequence of curves with curvatures $k_n(s)$ and $\lim_{n \to \infty} \sup_{s \in S^1} |k_n(s)| = 0$, or else there would be a subsequence limiting to a geodesic in the set. So we have an $\varepsilon > 0$ such that $\int_{S^1} k^2 \, ds \geq 2\varepsilon$ for any loop of such length, and we get the required result.

The definition of $\kappa(h)$ gives a cycle C such that $[[C]] = h$ and $C \subset \Lambda M^{\kappa(h)+\varepsilon}$. Now apply the CSF to each loop (i.e. point in ΛM) in the cycle C, note that ΛM and $\Lambda_0 M$ are invariant under the CSF and the flow is continuous so $C(1)$, the cycle C after time 1, is still a cycle of h. But then $C(1) \subset \Lambda M^{\kappa(h)-\varepsilon}$ which contradicts the choice of $\kappa(h)$. □

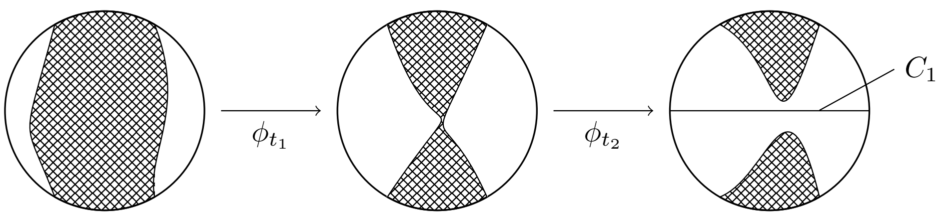
Main Theorem 4.2. *On a surface diffeomorphic to S^2, there exist three simple, closed geodesics.*

Proof. By the above lemma and inequality (2), we are left with the cases $\kappa(h_1) = \kappa(h_2)$ and $\kappa(h_2) = \kappa(h_3)$. Our notation is that C_i is some algebraic cycle in h_i.

Case $\kappa := \kappa(h_1) = \kappa(h_2)$: By the definition of the minmax values we have that $\forall \varepsilon > 0\ \exists C_2$ such that $\max_{\gamma \in C_2} L(\gamma) \leq \kappa + \varepsilon$ and $\exists \gamma \in C_2$ such that $L(\gamma) \geq \kappa$. Likewise for cycles in the other homology classes. We pick some $\varepsilon > 0$ and consider the closed subset

$$A := \{\gamma \in C_2 : \kappa \leq L(\gamma) \leq \kappa + \varepsilon\}.$$

From lemma (4.1) we know that when applying the CSF ϕ_t to C_2, at least one point will go to a geodesic length κ eventually, and $\phi_t(C_2)$ is still a cycle of h_2. In the case of $\kappa(h_1) = \kappa(h_2)$ we note that A has to be non-contractible, or else there's a C_1 cycle inside $\phi_t(C_2)$ that avoids A and hence contradicts the choice of $\kappa(h_1)$. The non-contractibility also means there's a $C_1 \subset A$, and $\phi_t(C_1) \subset \phi_t(A)$. The set $\phi_t(A)$ must remain non-contractible for all time or the same contradiction would occur as above, therefore we get a C_1 that exists for all time, i.e. a 1-cycle of h_1 made up of only geodesics. The main contradiction is illustrated below, with a C_1 path contradicting $\kappa(h_1) = \kappa(h_2)$, the hatched area is A (in $\phi_t(C_2)$):



Case $\kappa := \kappa(h_2) = \kappa(h_3)$: Similarly define $B := \{\gamma \in C_3 : \kappa \leq L(\gamma) \leq \kappa + \varepsilon\}$ and we note $C_3 \setminus B$ cannot contain a 2-cycle or there's a contradiction with $\kappa(h_2)$. Then B must contain a 1-cycle, or else it could be deformation retracted to a point and a 2-cycle in $C \setminus B$ could easily be constructed. Then we go through the same process as in the previous case, by applying CSF on B, to get that there must be a 1-cycle made up of only geodesics. $\square$

Corollary. *If any of the minmax values equal, then there are infinitely many distinct simple, closed geodesics of the same length.*

There are surfaces with only 3 simple closed geodesics, such as an ellipsoid with the 3 axis lengths sufficiently close to 1 and pairwise unequal [9].

A Appendix

A.1 $\mathscr{C}$, $SO(3)$ and $\mathbb{R}P^3$

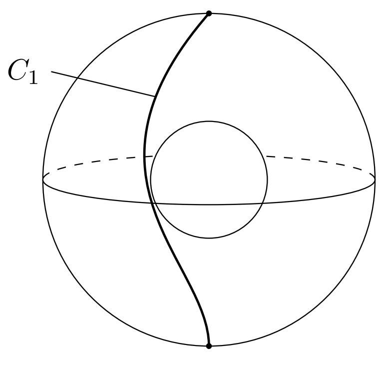
We show that the space $\mathscr{C}$, the circles of S^2, is $\mathbb{R}P^3 - D^3$. Firstly to motivate, we will consider the simpler case of $SO(3)$, the group of rotations in $\mathbb{R}^3$.

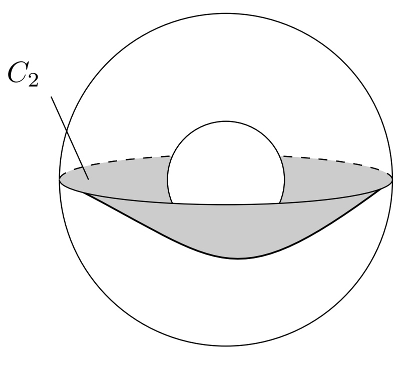
This Lie group may be modelled by the disc radius π, $D^3_\pi \subset \mathbb{R}^3$. The length of a vector $\underline{x} \in D^3_\pi$ gives the angle of rotation and the direction of $\underline{x}$ is the rotation axis, with $-\underline{x}$ having minus the angle of rotation of $\underline{x}$. We note that rotation by $-\pi$ and π give the same element of $SO(3)$ and hence we identify the antipodal points of the boundary ∂D^3_π. This space is diffeomorphic to $\mathbb{R}P^3$, so $SO(3)$ is diffeomorphic to $\mathbb{R}P^3$.

Denote $D^n_r \subset \mathbb{R}^n$ as the n-dimensional ball of radius r, centre origin. Now to model $(\mathscr{C}, \Lambda_0 S^2)$ we consider $D^3_2 - D^3_1$ where $\underline{x} \in D^3_2 - D^3_1$ corresponds to $\Pi \cap S^2 \in \mathscr{C}$ with Π being normal to $\underline{x}$, distance to origin $2 - |\underline{x}|$, thinking of Π as 'attached' at $\underline{x}$. We note that on the boundary ∂D^3_2 antipodal points give the same plane through the origin and therefore we identify these, similarly as for $SO(3)$, giving the space $\mathbb{R}P^3 - D^3$. Note the point curves are the boundary of this space and hence $(\mathscr{C}, \Lambda_0 S^2)$ is diffeomorphic to $(\mathbb{R}P^3 - D^3, \partial)$.

A.2 The Three Homology Classes of $(\mathbb{R}P^3 - D^3, \partial)$ with Coefficients in $\mathbb{Z}_2$

See Klingenberg [6], chapter 5, for a proof that $H_k(\mathbb{R}P^3 - D^3, \partial; \mathbb{Z}_2)$ is non-trivial for $k = 1, 2, 3$. For the purposes of proving the main theorem, it is sufficient to describe the homology classes in the retraction, then CSF may deform a cycle out of the retraction. We use the model of $(\mathbb{R}P^3 - D^3, \partial)$ as $(D^3_2 - D^3_1, \partial)$ with the antipodal points of the boundary of the larger ball identified.





As in the fundamental group, the generator of the first homology is a loop reaching two antipodal points. Then it's clear two such loops are null homologous. A cycle of a second homology generator is a two-parameter surface 'attached' at the boundary. Two copies of such cycles may be homotoped to a sphere with the missing disc inside, which is homotopic to the boundary ∂ and hence null in the relative homology. The third homology is the whole retraction.

Of course we may have additional parts of a cycle that are null homologous, such as a sphere with the missing ball inside, but this does not contribute to minmax value and hence such pathologies may be safely ignored.

References

[1] W. Ballmann, G. Thorbergsson, and W. Ziller. Existence of closed geodesics on positively curved manifolds. *J. Differential Geom.*, 18(2): 221–252, 1983.

[2] Werner Ballmann. Doppelpunktfreie geschlossene Geodätische auf kompakten Flächen. *Mathematische Zeitschrift*, 161(1):41–46, 1978.

[3] Michael E. Gage. Curve shortening on surfaces. *Annales scientifiques de l'Ecole normale supérieure*, 23(2):229–256, 1990.

[4] Matthew A. Grayson. Shortening embedded curves. *Annals of Mathematics*, 129(1):71–111, 1989.

[5] Jürgen Jost. A nonparametric proof of the theorem of Lusternik and Schnirelman. *Archiv der Mathematik*, 53(5):497–509, 1989.

[6] Wilhelm Klingenberg. *Lectures on closed geodesics*, volume 230. Springer Science & Business Media, 2012.

[7] L. Lusternik and L. Schnirelmann. Sur le problème de trois géodésiques fermées sur les surfaces de genre 0. *C.R. Acad. Sci. Paris*, 189:269–271, 1929.

[8] John Milnor. *Morse theory (AM-51)*, volume 51. Princeton University Press, 2016.

[9] H.C. Marston Morse. *The calculus of variations in the large*, volume 18. American Mathematical Soc., 1934.

[10] Iskander A. Taimanov. Closed extremals on two-dimensional manifolds. *Russian Mathematical Surveys*, 47(2):163, 1992.