Closed Geodesics: Topological Existence Theorems

Peter Kagey

April 26, 2019

A note to begin: I wrote this paper with a particular audience in mind, namely me in February of this year. I tried to capture a few things: (1) intuition for the topic (2) rigorous definitions, (3) early theorems and consequences, (4) modern theorems, and (5) conjectures and open problems.

1 Preliminaries

The intuition in this paper is to think of a creature who lives in a manifold walking in a "straight" line in some sense at a constant (nonzero) speed. In particular, some examples worth keeping in mind are

- the torus $\mathbb{R}^2/\mathbb{Z}^2$, where "straight lines" are inherited from \mathbb{R}^2 ;
- the round sphere $\{(x,y,z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\} = S^2 \subset \mathbb{R}^3$, where "straight lines" are great circles;
- a very "wobbly" sphere embedded in R^3 ; and
- $\mathbb{R}^3 \setminus \{0\}$, where straight lines are ordinary lines in $\mathbb{R}^3 \setminus \{0\}$,

where the last example is included to avoid the dangers of thinking about only so-called "geodesically complete" 2-manifolds.

In the case of \mathbb{R}^n , the notion of acceleration along a path at some point in time follows from the usual calculus definition, since the notion of subtracting velocity vectors at two different points along the curve can be done naively, allowing the second derivative computation to work in the expected way

$$\gamma''(t) = \lim_{h \to 0} \frac{\gamma'(t+h) - \gamma'(t)}{h}.$$

However, one has to be careful when talking about acceleration in a general manifold. In particular, $T_{\gamma(t)}$ and $T_{\gamma(t+h)}$ are generally different tangent spaces, and the notion of subtracting such velocity vectors is decidedly more subtle. In particular, an isometric embedding of a Riemannian manifold into Euclidean space via $f \colon M \to \mathbb{R}^n$, preserves notions of velocity, but "zero acceleration" curves in the sense of geodesics may not be "zero acceleration" in an extrinsic sense (for example, a constant speed particle moving along $S^1 \subset \mathbb{R}^2$ accelerates toward the origin.) We can fix this by considering the projection of the acceleration to the tangent plane, but capturing an *intrinsic* notion of a directional derivative is more subtle.

Definition 1.1. Let $\pi: E \to M$ be a smooth vector bundle over M. Then a **connection** in E is any map that sends a tangent vector field and a section of E to another section of E

$$\nabla \colon \mathfrak{X}(M) \times \Gamma(E) \to \Gamma(E)$$

denoted $\nabla_X Y := \nabla(X, Y)$ and satisfying the following three properties:

- (i) ∇ is $C^{\infty}(M)$ -linear in the first argument: $\nabla_{f_1X_1+f_2X_2}Y = f_1\nabla_{X_1}Y + f_2\nabla_{X_2}Y$
- (ii) ∇ is \mathbb{R} -linear in the second argument: $\nabla_X(a_1Y_1 + a_2Y_2) = a_1\nabla_XY_1 + a_2\nabla_XY_2$, and
- (iii) ∇ satisfies a product rule in the second argument: $\nabla_X(fY) = f\nabla_XY + \underbrace{(Xf)}_{\in C^\infty(M)}Y$.

Note 1.1. In the case where $\Gamma(E) = C^{\infty}(M)$, $\nabla(X, f) := Xf$ satisfies the above properties. In particular, ∇ can be thought of as a prescription for extending $X : C^{\infty}(M) \to C^{\infty}(M)$ to act on other vector bundles.

Example 1.1. Consider the case of $M = \mathbb{R}^3$ with the section of the tangent bundle given by $\langle x, 1, e^y \rangle = X \in \mathfrak{X}(\mathbb{R}^3)$ and a section of a rank-2 vector bundle $\langle x^2, y + z \rangle = Y \in \Gamma(E)$. Then taking the derivative of Y along X yields

$$\nabla_X Y = \underbrace{\begin{bmatrix} \frac{\partial x^2}{\partial x} & \frac{\partial x^2}{\partial y} & \frac{\partial x^2}{\partial z} \\ \frac{\partial (y+z)}{\partial x} & \frac{\partial (y+z)}{\partial y} & \frac{\partial (y+z)}{\partial z} \end{bmatrix}}_{dY(x,y,z)} \underbrace{\begin{bmatrix} x \\ 1 \\ e^y \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} 2x^2 \\ 1 + e^y \end{bmatrix}}_{\in \Gamma(E)}.$$

The matrix multiplication perspective makes it clearer to see that in this case (i) ∇ is $C^{\infty}(\mathbb{R}^3)$ -linear in the first term, because the column vector is $C^{\infty}(\mathbb{R}^3)$ -linear, (ii) ∇ is \mathbb{R} -linear in the second term because both partial derivatives and matrices are \mathbb{R} -linear, and (iii) holds by doing the product rule entrywise in the Jacobian matrix d(fY)(x,y,z).

These conditions capture, in some sense, the "essential qualities" of taking the directional derivative of a vector bundle over a given section of the tangent bundle.

Note 1.2. When on a Riemannian manifold, it is common to use a particular connection called the Levi-Civita connection, which is compatible with the metric in the expected way, and which has some other nice properties.

Definition 1.2. The **Levi-Civita connection** is a connection

$$\nabla \colon \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$$

such that for any tangent vector fields $X, Y \in \mathfrak{X}(M)$, ∇ is

- (a) compatible with the metric: $\nabla_X(q(Y,Z)) = q(\nabla_X Y,Z) + q(Y,\nabla_X Z)$, and
- (b) torsion-free: $\nabla_X Y \nabla_Y X = [X, Y]$.
- Note 1.3. When the Levi-Civita connection exists, it is unique.

Note 1.4. The second condition gets its name from the definition of the torsion, namely

$$T(X,Y) := \nabla_X Y - \nabla_Y X - [X,Y].$$

Definition 1.3. Let ∇ be a connection in tangent bundle TM, and let γ be a curve in M. Then the **covariant derivative along** γ is the (unique) map $D_{\gamma} \colon \mathfrak{X}(\gamma) \to \mathfrak{X}(\gamma)$ satisfying the following three properties

- (i) $D_{\gamma}(aV + bW) = aD_{\gamma}V + bD_{\gamma}W$
- (ii) $D_{\gamma}(fV) = f'V + fD_{\gamma}V$
- (iii) $D_{\gamma}V(t) = \nabla_{\gamma'(t)}\widetilde{V}$ for every extension \widetilde{V} of V.

Note 1.5. This definition is sensible, because (i) and (ii) respectively correspond to (ii) and (iii) in the definition of a connection, and the last condition is roughly says that the derivative along γ should agree with a restriction on a vector field which is equal to γ' along the curve.

Definition 1.4. For every smooth map from an interval to the manifold, $\gamma: I \to M$, define the **acceleration** of γ as the vector field $D_{\gamma}\gamma'$ along γ .

Definition 1.5. A smooth curve γ is called a **geodesic** (with respect to ∇) if it's acceleration is zero: $D_{\gamma}\gamma'\equiv 0$.

Theorem 1.1 (Existence and uniqueness of geodesics; from [5], Theorem 4.27). Let M be a smooth manifold and ∇ a connection. Then for every $p \in M$ and $v \in T_pM$ there exists an interval $I \subset \mathbb{R}$ around 0 and a geodesic $\gamma \colon I \to M$ satisfying $\gamma(0) = p$ and $\gamma'(0) = v$.

Furthermore, any two geodesics with the same initial conditions agree on their common domain.

Proof idea. The proof of the theorem works by moving to local coordinates in a way that respects the connection, and then making an appeal to uniqueness and existence for (second order) ordinary differential equations. \Box

Definition 1.6. Given a geodesic $\gamma: I \to M$, an **extension of** γ **to** $\widetilde{I} \supset I$ is a geodesic curve $\widetilde{\gamma}: \widetilde{I} \to M$ such that the geodesics agree on $I: \widetilde{\gamma}|_{I} = \gamma$.

Note 1.6. A proper extension is an extension in which condition (i) is strengthened such that I is a proper subset of \widetilde{I} , and a geodesic with no proper extension is called a **maximal** geodesic.

2 Properties of geodesics

An arbitrary manifold is not endowed with any "measuring tape" and creatures living on the manifold don't have any notion of their speed when they travel. Of course, it's natural to believe that having one gives the other—that is if our creature has a speedometer, it also has an odometer, and vice versa.

A Riemannian manifold comes equipped with an inner product on the tangent space at each point, and an inner product is equivalent to assigning a magnitude to each vector (by the polarization identity). That is, a Riemannian manifold is special because each tangent vector in TM has a notion of its magnitude or speed. This section will develop the natural way of turning the speed at each point to the length of a curve, and a natural way to define the distance between two points derived from the length of the curves between two points.

Definition 2.1. The length of a vector $v \in TM$ in a Riemannian manifold with inner product g is

$$||v||_g = \sqrt{g(v,v)} = \sqrt{\langle v,v \rangle}.$$

Definition 2.2. The length of a (piecewise differentiable) curve $\gamma: I \to M$ in a Riemannian manifold is

$$L(\gamma) = \int_{\gamma} ||\gamma'(t)||_g dt$$

Theorem 2.1. Let (M,g) be a Riemannian manifold. Then g induces a metric $d: M \times M \to \mathbb{R}$ on M, namely

$$d(a,b) = \inf \{L(\gamma) : \gamma \text{ is a path from } a \text{ to } b\}$$

Proof. In order to show that d is a metric, it is necessary and sufficient to show that

1. d(a,b) = 0 if and only if a = b.

(⇒) (A sketch, as this proof is a bit involved and technical.)

Choose some chart (U, φ) around p but not q, then in any compact set $K \subset U$ around p there exists some $C \in (0, \infty)$ such that for all $x \in K$ and $v \in T_x K$, all vectors $||\varphi_* v|| \leq C||v||_g$, which means that all curves in K have length greater than 1/C times the length of their image under φ .

Then let $R = \varphi^{-1}(B_{\varepsilon}(\varphi(p)))$ be the preimage of a small ball around $\varphi(p)$. By construction, p is inside this region, q is outside the region, and the distance to the boundary of the region is at least ε . Thus $d(p,q) \ge \epsilon$ by the Jordan Curve Theorem.

(\Leftarrow) If a=b, the constant path $\gamma(t)=a=b$ has length 0, and the length of a path is strictly non-negative because the integrand is nonnegative: $||v||_q \ge 0$.

2. d(x,y) = d(y,x)

If $\gamma: (t_0, t_1) \to M$ is a path from a to b, then $\overline{\gamma}: (t_0, t_1) \to M$ defined by $\overline{\gamma}(t) = \gamma(t_1 + t_0 - t)$ is a path from b to a which has the same image and same length.

3. $d(a,b) \le d(a,c) + d(c,b)$

The set of pairs of paths from a to c and c to a is in length preserving bijection (via concatenation) with the set of paths from a to b through c, which is a subset of the set of paths from a to b.

Definition 2.3. A connected Riemannian manifold (M, g) is called **geodesically complete** if every geodesic curve $\gamma: I \to M$ there is an extension of γ with \mathbb{R} as the domain, $\widetilde{\gamma}: \mathbb{R} \to M$.

While the distance d(a, b) function on M is defined as the infimum under all paths from a to b, in the case of a geodesically complete (connected) manifold, it is equivalent to define distance as the *minimum* over all paths instead.

Theorem 2.2. In a geodesically complete Riemannian manifold, there exists a geodesic (with respect to the Levi-Civita connection) γ from (a, b) such that $L(\gamma) = d(a, b)$.

Proof idea. This follows from the Hopf-Rinow Theorem (below) along with some technical applications using the exponential map. \Box

Now is perhaps an appropriate time to introduce the exponential map. The idea is that the exponential map takes initial conditions of a geodesic (position and velocity), and (when possible) returns the position of a particle flowing along such a geodesic after unit time. The exponential map takes a point in the tangent bundle $(p, v) \in TM$ and returns the position of a geodesic starting at p and flowing in the direction of v after unit time.

Definition 2.4. Let

$$\mathcal{D} = \{(p, v) \in TM : \exists \text{ a geodesic curve } \gamma \colon [0, 1] \to M \text{ with } \gamma(0) = p \text{ and } \gamma'(0) = v\}$$

Then the **exponential map** is a map $\exp: \mathcal{D} \to M$ satisfying $\exp((p, v)) = \gamma(1)$ where $\gamma: [0, 1] \to M$ is any geodesic with $\gamma(0) = p$ and $\gamma'(0) = v$.

Note 2.1. A geodesically complete (connected) manifold is equivalently defined as one in which $\mathcal{D} = TM$.

3 Manifolds as metric spaces

It is interesting to ask about whether or not a manifold is geodesically complete, that is whether or not the creature in the manifold can travel forever in a "straight line" starting at any point and heading in any direction.

Clearly the creature does not run into any trouble flying through space (\mathbb{R}^3) or walking on the surface of a sphere (S^2), but the creature clearly *does* run into trouble if it is on the surface of a disc ($D^2 \subset \mathbb{R}^2$) or if it is flying toward the origin, but that point has been removed ($\mathbb{R}^3 \setminus \{0\}$).

The Hopf-Rinow Theorem says that these two examples capture the extent of what can go wrong during the creature's journey. In particular, it says that that properties of geodesics and the induced metric agree with each other and with the topology.

Definition 3.1. A complete metric space is a metric space in which every Cauchy sequence is convergent.

Theorem 3.1 ([4]). Let (M, g) be a Riemannian manifold, and let (M, d) be the induced metric space. Then the following are equivalent:

- (i) (M, d) is metrically complete.
- (ii) (M, g) is geodesically complete.
- (iii) (M, g) is geodesically complete at some point.
- (iv) Every closed, bounded set in M is compact.

And furthermore, between any two points $p, q \in M$, there exists a length-minimizing geodesic connecting these two points.

Proof sketch.

- (i) \Longrightarrow (ii) The idea is to assume that M is not complete and arrive at a contradiction. In particular, this means there is some geodesic curve $\gamma \colon [a,b) \to M$ which can not be extended to $\widetilde{\gamma} \colon [a,b] \to M$ —however this can't happen because we can use the limit of a Cauchy sequence to give a value for $\widetilde{\gamma}(b)$ such that $\widetilde{\gamma}$ is a geodesic.
- (ii) \Longrightarrow (iii) If (M, g) is geodesically complete at every point, then in particular it is geodesically complete at some point.
- (iii) \Longrightarrow (iv) Suppose (M, g) is geodesically complete at p. For any closed and bounded $A \subset M$, we can always find a ball $B_R(p)$ around A. Since the exponential map is continuous, we can see that a closed subset of a compact set is itself compact.
- (iv) \Longrightarrow (i) Let (x_n) be a Cauchy sequence in M. Since $\{x_n\}$ is bounded, its closure is compact and so (x_n) has a convergent subsequence, and because it is Cauchy, it converges to this subsequence.

4 Closed geodesics

Definition 4.1. A geodesic $\gamma: I \to M$ is called **closed** if there exist two points $t_1 \neq t_2 \in I$ such that $\gamma(t_1) = \gamma(t_2)$ and $\gamma'(t_1) = \gamma'(t_2)$.

Note 4.1. Every closed geodesic γ has an all-time extension $\widetilde{\gamma} \colon \mathbb{R} \to M$.

Theorem 4.1. Let $\pi_1(M)$ be the fundamental group of M. Then each equivalence class of loops $[\alpha] \in \pi_1$ has a (closed) geodesic representative, in particular it contains a curve of minimal length which is a geodesic.

Proof idea. [9] summaries the proof as

One way to see why this is true is via a Morse-theory approach, with a full proof being given in Klingenberg - Lectures on Closed Geodesics. The gist of it is to pick a loop close to the infimum on the class and flow it under (minus) the gradient flow of the energy functional in the free loop space (this has an obvious analogy with the finite-dimensional case when we flow via gradient and converge to a critical point).

Note 4.2. In a survey paper, [10] writes

one of the first successes of the calculus of variations was to establish rigorously that such a minimizing procedure is effective and produces a closed geodesic. The situation is subtler if the manifold is simply connected, and the question was answered in the affirmative by Lyusternik and Fet in their celebrated 1951 paper

Theorem 4.2 ([7]). Every compact Riemannian manifold (M, g) has at least one closed geodesic.

Proof. In a report from the International Workshop on Geodesics in August 2010, [3] writes

Lyusternik and Fet considered the energy functional on the loop space ΛM and showed that the topology of ΛM is complicated enough so that the energy functional must have critical points with nonzero energy (which are non trivial closed geodesics).

Conjecture 4.1 ([3]). Every compact Riemannian manifold of dimension greater than 1 contains infinitely many geometrically distinct non-constant closed geodesics.

Note 4.3. This particular conjecture in the above report by [3]:

5

How many closed geodesics must exist on a closed manifold? For surfaces, the answer is known: there are always infinitely many geometrically different closed geodesics. This is easily proved using Birkhoffs first argument when the fundamental group is infinite. The remaining cases of the sphere and the projective plane were settled by Bangert (1993) and Franks (1992). In higher dimensions Rademacher has shown that a closed manifold with a generic Riemannian metric admits infinitely many geometrically different closed geodesics Rademacher (1989).

The report gives a heuristic for why this is hard to prove:

It is important to distinguish geometrically different geodesics from repetitions of the same geodesic. This distinction is difficult to make.

Note 4.4. In 1999 [1] writes

The question of which Riemannian manifolds admit simple closed geodesics is still a mystery. It is not known whether all closed Riemannian manifolds contain simple closed geodesics.

Theorem 4.3 ([8]; [2]). Every Riemannian manifold with the topology of a sphere has at least three simple closed geodesics.

Proof. The history of the proof from [6]:

The existence of three simple periodic geodesics on every Riemannian 2-sphere was first proven by L.A. Lyusternik and L.G. Shnirelman. They considered the space ΠM of non-parametrized curves on a 2-dimensional Riemannian manifold M diffeomorphic to S^2 , as well as its subset $\Pi_0 M$ that consists of all constant curves (and, thus, can be identified with M). [One can then] consider the three relative homology classes of the pair $(\Pi S^2, \Pi_0 S^2)$ with coefficients in \mathbb{Z}_2 , where we regard S^2 as the unit round sphere in \mathbb{R}^3

- 1) The 1-dimensional class represented by the relative 1-cycle z_1 formed by all circles on S^2 in planes parallel to the XZ-plane in the ambient R^3 ;
- 2) The 2-dimensional class represented by the relative cycle z_2 formed by all circles in planes parallel to the Z-axis; and
- 3) The 3-dimensional class represented by the relative 3-cycle z_3 formed by all round circles on the sphere (including points regarded as degenerate circles).

Lyusternik and Shnirelman described a curve shortening flow in ΠM , [and] they observed that this flow gets stuck on critical points representing simple periodic geodesics, when applied to z_1 , z_2 , z_3 , and proved that if two of these three geodesics coincide, then there is a whole critical level with a 1-parametric set of distinct simple periodic geodesics. Some errors in the construction of their curve shortening flow had been later corrected by W. Ballman [and others]. Alternatively, one can prove the existence of three simple periodic geodesics using either the curvature flow and Graysons theorem or an especially simple curve shortening flow constructed by J. Hass and P. Scott.

In particular, a recent theorem shows an even stronger result: the lengths of all three geodesics on a sphere are bounded by the size of the manifold, as measured by the supremal distance between two points.

Definition 4.2. The **diameter** of a Riemannian manifold (M, g) is defined by

$$\operatorname{diam}(M) := \sup_{p,q \in M} d(p,q) \in \mathbb{R}_{\geq 0} \cup \{+\infty\}$$

Example 4.1. Consider the unit circle $S^1 = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ as a sub-manifold of \mathbb{R}^2 with the induced inner product. Then $\operatorname{diam}(S^1) = \pi$, with the distance between two antipodal points, say p = (1,0) and q = (-1,0), being half of the circumference of the circle (in the ordinary sense).

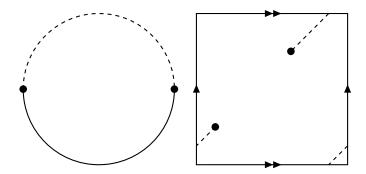


Figure 1: Illustrations of diameters on S^1 and T^2 , and geodesics that achieve them.

Example 4.2. Consider the torus $T^2 = \mathbb{R}^2/\mathbb{Z}^2$ with the inner product inherited from \mathbb{R}^2 . Then the diameter of the torus diam $(T^2) = \sqrt{2}/2$, with this distance achieved by p = (0,0) and $q = (\frac{1}{2}, \frac{1}{2})$.

Theorem 4.4 ([6]). Let M be a Riemannian 2-sphere. Then there exist three distinct simple geodesics with lengths that do not exceed 20d where d is the diameter of M. Furthermore, if no simple closed geodesics of length $\leq 2d$, then there are three distinct simple periodic geodesics on M with lengths $\leq 5d$, 10d, and 20d respectively.

Proof idea. The above authors give their proof idea as follows:

The main idea of the proof of [the above theorem] is to express three homology classes of the space of non-parametrized curves that are used in classical proofs of the Lyusternik-Schnirelmann theorem by cycles that consist of simple closed curves "mainly made" of curves in a meridian-like family that connects two fixed points of M. [...] We attempt to construct such a family where the lengths of all meridians are bounded by const d for an appropriate const. Our repeated attempts can be blocked only by appearance of different short simple periodic geodesics of index 0. So, we either get three short simple periodic geodesics of index 0, or our third attempt to construct a meridianal slicing succeeds. Once one of our attempts succeeds, and we get a slicing of M into short meridians, the original proof of the Lyusternik-Schnirelmann theorem yields the desired upper bounds.

Note 4.5. With these bounds, it is natural to ask if all 2-spheres have a fourth geodesic which has length bounded by the size of the manifold, but [6] writes:

[There is a] classical result of M. Morse, who proved that the fourth periodic [not necessarily simple] geodesic becomes uncontrollably large for ellipsoids with distinct but very close semi-axes.

5 Conclusion

A quick final personal note: I find that the farther I get in my mathematical career, the deeper I have to dig to find new-to-me theorems that spark a sense of wonder, but I've found that this relatively simple topic is full of them, especially those discussed in the previous section.

In particular, the theorem of three simple closed geodesics is quite unbelievable—especially when combined with the result that a fourth (perhaps non-simple) closed geodesic can be arbitrarily large. It's especially remarkable in the context of Conjecture 4.1, which says that it is not known whether or not simple closed geodesics exist on all compact manifolds.

Thanks again for your help in suggesting this as a topic—I found the experience to be immensely rewarding.

References

- [1] Colin Adams, Joel Hass, and Peter Scott. "Simple Closed Geodesics in Hyperbolic 3-Manifolds". In: Bulletin of the London Mathematical Society 31.1 (1999), pp. 81-86. DOI: 10.1112/S0024609398004883. eprint: https://londmathsoc.onlinelibrary.wiley.com/doi/pdf/10.1112/S0024609398004883.
- [2] Werner Ballmann. "Der Satz von Lusternik und Schnirelmann". In: Math. Shriften 102 (1978), pp. 1–25.
- [3] Keith Burns and Vladimir S. Matveev. *Open problems and questions about geodesics*. 2013. eprint: arXiv:1308.5417.
- [4] Heinz Hopf and Willi Rinow. About the concept of the complete differential geometric surface. July July 1931. DOI: https://doi.org/10.1007/BF01601813.
- [5] J.M. Lee. Riemannian Manifolds: An Introduction to Curvature. Graduate Texts in Mathematics. Springer New York, 1997. ISBN: 9780387982717. URL: https://books.google.com/books?id=ZRQgH7FQafgC.
- [6] Yevgeny Liokumovich, Alexander Nabutovsky, and Regina Rotman. Lengths of three simple periodic geodesics on a Riemannian 2-sphere. 2014. eprint: arXiv:1410.8456.
- [7] L.A. Lyusternik and A.I. Fet. "Variational problems on closed manifolds". In: *Dokl. Akad. Nauk SSSR* (N.S.) 81 (Jan. 1951), pp. 17–18.
- [8] Lazar Lyusternik and Lev Schnirelmann. "Sur le problème de trois géodésiques fermées sur les surfaces de genre 0". In: Comptes Rendus de l'Académie des Sciences de Paris 189 (1929), pp. 269–271.
- [9] Aloizio Macedo. Closed geodesic on a non-simply connected Riemannian manifold. URL: https://math.stackexchange.com/a/2844405/121988. (accessed: 04.25.2019).
- [10] Alexandru Oancea. Morse theory, closed geodesics, and the homology of free loop spaces. 2014. eprint: arXiv:1406.3107.