Lecture 3/4 Geodesics
Nancy Hingston
IAS WAM 2016

Geodesics are Straight

In physics, a geodesic is the path traveled by a light ray. The light ray travels along a "straight" line; the geometry of space will determine whether the geodesic will shoot off and never be seen again., or if it might come back and hit you in the back of the head as you stare off in the direction in which it disappeared.

Geodesics are Short

If you watch a light ray, and if at time $t = 0$ it is at a point A, and if a short time later it is at the point B. , you will find that it took the (unique!) shortest path from A to B. So geodesics minimize distance, at least over short time periods.

Given a manifold M in $\mathbb{R}^N$, we imagine that we "live" in the "universe" M. That is, we cannot see or travel into the ambient Euclidean space $\mathbb{R}^N$, but only M. In particular, if A and B are points in M, we are only allowed to travel from A to B *on* M. The distance from A to B on M is defined to be the minimal length of a path from A to B measured using the arclength integral from $\mathbb{R}^N$. Example: let $M \subset \mathbb{R}^2$ be a circle S^1 in $\mathbb{R}^2$: $x^2 + y^2 = 1$. If you want to get from $(1, 0)$ to $(0, 1)$ you need to travel along the circle. The distance between these two points is $\frac{\pi}{2}$. If you lived in this (one-dimensional) universe and started off on a walk, you would find that after going a distance 2π you would have returned to your starting place. Example: The distance between two points on the 2-sphere $M = S^2$ is the length of a minimal great circle arc between them.

For those of you who know some topology, the space M inherits a topology and a metric from $\mathbb{R}^N$. The ambient space $\mathbb{R}^N$ contributes topology and metric but is "inaccessible". (You can't go there.)

Let $M \subset \mathbb{R}^N$ be a manifold. We will define geodesics on M using the **Straightness** property: A parameterized curve $\overrightarrow{\mathbf{r}} = \overrightarrow{\mathbf{r}}(t) =< x(t), y(t), ... >: \mathbb{R} \to M$ is *geodesic* if the acceleration vector $\overrightarrow{\mathbf{a}}(t) =< \ddot{x}, \ddot{y}, ... >$ at the point $< x(t), y(t), ... >$ is perpenducular to the tangent space TM:

$$\overrightarrow{\mathbf{a}}(t) \perp T_{\overrightarrow{\mathbf{r}}(t)}M. \quad (*)$$

This means that the curve has the minimum acceleration necessary to stay on M, so the condition $(*)$ is "straightness".

Geodesics exist

Let $P \in M$, and $\mathbf{V} \in T_P M$.We will find a geodesic

$$\overrightarrow{\mathbf{r}} = \overrightarrow{\mathbf{r}}(t) = < x(t), y(t), ... >$$

starting at P with initial velocity $\mathbf{V}$. For simplicty of notation we will assume that M is a 2-dimensional surface in $\mathbb{R}^3$. After a rigid motion of $\mathbb{R}^3$ we can assume that P is the origin, and $T_P M$ is the $x-y$ plane in $\mathbb{R}^3$. By the implicit function theorem, M is defined near the origin by

$$M : z = h(x, y)$$

which leads to

$$\begin{aligned} \dot{z} &= h_x \dot{x} + h_y \dot{y} \\ \ddot{z} &= h_{xx}(\dot{x})^2 + 2h_{xy}\dot{x}\dot{y} + h_{yy}(\dot{y})^2 + h_x \ddot{x} + h_y \ddot{y} \qquad (**) \end{aligned}$$

which say that the path stays on M. Meanwhile note that the vectors

$$< 1, 0, h_x > \text{ and } < 0, 1, h_y >$$

are linearly independent and both perpendicular to the normal vector $< h_x, h_y, -1 >= \nabla(h(x, y) - z)$ to M at P. The condition $(*)$ is thus equivalent to the conditions

$$\begin{aligned} < \quad & \ddot{x}, \ddot{y}, \ddot{z} ... > \cdot < 1, 0, h_x >= 0 \\ < \quad & \ddot{x}, \ddot{y}, \ddot{z} ... > \cdot < 0, 1, h_y >= 0 \end{aligned}$$

which together with $(**)$ give

$$\ddot{x} + h_x(h_{xx}(\dot{x})^2 + 2h_{xy}\dot{x}\dot{y} + h_{yy}(\dot{y})^2 + h_x \ddot{x} + h_y \ddot{y}) = 0$$

$$\ddot{y} + h_y(h_{xx}(\dot{x})^2 + 2h_{xy}\dot{x}\dot{y} + h_{yy}(\dot{y})^2 + h_x \ddot{x} + h_y \ddot{y}) = 0$$

Since $h_x = h_y = 0$ at P (why?) , We can solve to get

$$\begin{pmatrix} \ddot{x} \\ \ddot{y} \end{pmatrix} = - < \overrightarrow{v}, H\overrightarrow{v} > \left[\left(I + \begin{pmatrix} h_x \\ h_y \end{pmatrix} \begin{pmatrix} h_x & h_y \end{pmatrix} \right)^{-1} \begin{pmatrix} h_x \\ h_y \end{pmatrix} \right] \qquad (***)$$

where $\overrightarrow{v} =< \dot{x}, \dot{y} >$ is the horizontal component of the velocity and

$$H = \begin{pmatrix} h_{xx} & h_{yx} \\ h_{xy} & h_{yy} \end{pmatrix}.$$

We now in the usual way turn a 2nd order DEQ in two variables into a 1st order DEQ in 4 variables :

$$\frac{d}{dt}\begin{pmatrix} x \\ y \\ u \\ w \end{pmatrix} = \begin{pmatrix} u \\ v \\ f(x,y,u,v) \\ g(x,y,u,v) \end{pmatrix}$$

where f and g are the components of components of the right-hand side of $(***)$ evaluated at $\overrightarrow{r} =< x, y >$ and $\overrightarrow{v} =< u, v >$. (Note h and H depend on x and y.)

A solution of this system is an integral curve of the vector field in $\mathbb{R}^4$ defined by the right hand side; it is a solution of an equation of the form

$$\frac{d}{dt}\overrightarrow{R}(t) = \overrightarrow{F}(\overrightarrow{R}(t))$$

As the vector field $\overrightarrow{F}$ is Lipschitz continuous, the Picard–Lindelöf theorem implies that there exists a unique flow for small time. The initial condition for this flow is a point $(x_0, y_0, u_0, v_0) \in \mathbb{R}^4$. Thus given a point $P : (x_0, y_0) \in M$, and a vector $\mathbf{V} : (u_0, v_0) \in T_P M$ there exists $\varepsilon > 0$ so that there is a unique a geodesic

$$\overrightarrow{\mathbf{r}} = \overrightarrow{\mathbf{r}}(t) =< x(t), y(t), ... >$$

defined on the interval $[0, \varepsilon)$ starting at P and with $\frac{d\overrightarrow{\mathbf{r}}}{dt}(0) = \mathbf{V}$. *Note we used only the straightness property* $(*)$ *and the fact that* $\overrightarrow{\mathbf{r}}(t)$ *stays on the manifold* M.

It follows from the defining property

$$\overrightarrow{\mathbf{a}}(t) \perp T_{\overrightarrow{\mathbf{r}}(t)}M. \quad (*)$$

and the vector identity

$$\frac{d}{dt}(\overrightarrow{A} \cdot \overrightarrow{B}) = \frac{d}{dt}(\overrightarrow{A}) \cdot \overrightarrow{B} + \overrightarrow{A} \cdot \frac{d}{dt}(\overrightarrow{B})$$

that the speed

$$\left|\frac{d\overrightarrow{\mathbf{r}}}{dt}\right| = \sqrt{\frac{d\overrightarrow{\mathbf{r}}}{dt} \cdot \frac{d\overrightarrow{\mathbf{r}}}{dt}}$$

is constant on any geodesic.

The exponential map

If $\mathbf{V} \in T_P M$, and if $\overrightarrow{\mathbf{r}}(t)$, $0 \le t \le 1$, is a geodesic with $\overrightarrow{\mathbf{r}}(0) = P$, $\frac{d\overrightarrow{\mathbf{r}}}{dt}(0) = \mathbf{V}$, and $\overrightarrow{\mathbf{r}}(1) = Q$, we say

$$\exp_P \mathbf{V} = Q.$$

Note it follows that, if $0 \leq s \leq 1$,

$$\exp_P s\mathbf{V} = \overrightarrow{\mathbf{r}}(s).$$

One can show that for any $P \in M$, there is a neighborhood U of P, and $\varepsilon > 0$ so that $\exp_Q \mathbf{V}$ is defined provided $Q \in U$, $\mathbf{V} \in T_Q M$, and $|\mathbf{V}| < \varepsilon$.

On some (possible smaller) $\varepsilon-$ball, $\exp_P$ is a diffeomorphism onto a neighborhood of P in M. With a little more work one can prove that, given P there is a neighborhood U so that any two points in U are joined by a unique geodesic in U.

We have defined geodesics using the **straightness property** $(*)$. What about the **shortness property?** We follow Milnor's *Morse Theory* to show that geodesics have the local length minimizing property. Outline of proof:

Lemma 1: The geodesics through $P \in M$ are orthogonal to the "spheres"

$$S_P(\varepsilon) = \{\exp_P \mathbf{V} : |\mathbf{V}| = \varepsilon\}$$

It follows from this that : it costs to move "sideways". **Lemma 2**: ....The shortest path between two "spheres" is a radial geodesic. (Follows from Lemma 1. Recall the arclength integral on $\mathbb{R}^N$ is defined as a limit of sums of distances; the distances are computed using Pythagorus so the hypotenuse is always longer than either leg.)

Note that any subarc of a length-minimizing path is also a length minimizing path, and any subarc of a geodesic is also a geodesic; thus the geodesics on M are precisely the paths of the form

$$\{\exp_P t\mathbf{V} \ : t \in J\}$$

for some $P \in M, \mathbf{V} \in T_P M$ and J an interval in $\mathbb{R}$.

Conclusion (more than we have actually proved) **Short** $\Longleftrightarrow$ **Straight** Every point $P \in M$ has a neighborhood U with the property that for any $Q, Q' \in U$, there is a unique shortest path on M from Q to Q', and it is the unique shortest geodesic from Q to Q'. This geodesic and its length (provided the length is nonzero) depend smoothly on the endpoints. The map $\exp_P t\mathbf{V}$: $t \in \mathbb{R} \to M$ preserves distance locally, so that

$$Dist(P, \exp_P \mathbf{V}) = |\mathbf{V}| = Arclength\{\exp_P t\mathbf{V} : 0 \leq t \leq 1\}$$

if $|\mathbf{V}|$ is small. If M is closed and bounded there is an *injectivity radius* $\rho > 0$ with the property that every geodesic of length $< \rho$ is length minimizing.

Note: The characterization of geodesics as "straight" appears to depend on the embedding in $\mathbb{R}^N$, but the "shortest" description clearly only depends on the distance function induced on M .

Here is an **example**: The geodesics on $\mathbb{R}^2$ are straight lines. The map $\mathbb{R}^2 \to \mathbb{R}^3$ given by

$$f(t,u) = (\cos t, \sin t, u)$$

preserves the arclength element $ds^2 = dt^2 + du^2 = dr^2 + r^2 d\theta^2 + dz^2$, and thus any path γ and its image $f \circ \gamma$ have the same length.) Since we define distance on our manifold using arclength, f is a local isometry, that is, the distance from P to Q is equal to the distance from $f(P)$ to $f(Q)$ provided P and Q are close . It follows that f takes geodesics on the plane to geodesics on the cylinder. Thus the (nonconstant) geodesics on the cylinder are horizontal circles, vertical lines, and helicies:

$$\gamma(t) = (\cos \alpha t + \beta, \sin \alpha t + \beta, \alpha' t + \beta')$$

Note that you do not need to stretch a piece of paper to roll it into a cylinder.

Exercise: Assume that $h(x,y) = c$ describes an embedded curve in $\mathbb{R}^2$. Describe what the geodesics on the "cylinder"

$$\mathcal{C} = \{(x,y,z) : h(x,y) = c\} \subset \mathbb{R}^3$$

from the point of view of a being whose whole universe was the surface $\mathcal{C}$.

Example (Clairaut) describes geodesics on a surface of revolution in $\mathbb{R}^3$:

$$r(t) \cos \alpha(t) = \text{Constant} \quad (****)$$

Here r is the distance to the axis of rotation, and α is the angle between a tangent vector to the geodesic at a point P and a tangent vector $\overrightarrow{p}$ to the surface at P in the "parallel" direction. (If you pick $\pi - \alpha$ instead of α, you will just change the sign of the constant.) Remark and easy way to remember/prove the relation: The conserved quantity $(****)$ is the "angular momentum about the axis of revolution", that is, the component of the angular momentum about any point on the axis, in the direction of the axis. We can express the relation as

$$(\overrightarrow{r} \times \overrightarrow{v}) \cdot \overrightarrow{u} = \text{Constant} \quad (*****)$$

where $\overrightarrow{r}$ is the position vector measured from any point (even a moving point) on the axis, $\overrightarrow{v}$ is the velocity vector, and $\overrightarrow{u}$ is a unit vector in the axis direction. (To see that $(****)$ and $(*****)$ are equivalent, use the fact that $|\overrightarrow{v}|$ is constant, and that $\overrightarrow{r}$, $\overrightarrow{u}$, and $\overrightarrow{p}$ are orthogonal; thus $(\overrightarrow{r} \times \overrightarrow{v}) \cdot \overrightarrow{u} = rv \cos \alpha$.) To prove the relation we differentiate:

$$\begin{aligned} & \frac{d}{dt}((\overrightarrow{r} \times \overrightarrow{v}) \cdot \overrightarrow{u}) \\ = & \left(\frac{d}{dt}(\overrightarrow{r}) \times \overrightarrow{v}\right) \cdot \overrightarrow{u} + \left((\overrightarrow{r} \times \frac{d}{dt}\overrightarrow{v}) \cdot \overrightarrow{u}\right) \\ = & (\overrightarrow{v} \times \overrightarrow{v}) \cdot \overrightarrow{u} + ((\overrightarrow{r} \times \overrightarrow{a}) \cdot \overrightarrow{u}) \end{aligned}$$

This is 0 so long as $\overrightarrow{r}(t)$ stays in the plane spanned by $\overrightarrow{a}$ and $\overrightarrow{u}$.

Exercise: (1) Draw some pictures of surfaces of revolution and geodesics on them. The torus of revolution is a nice example.

Theorem of Hopf-Rinow: (1931) If M is closed and path connected, and $P \in M$, then $\exp_P : T_P M \to M$ is surjective.

What does this mean? Imagine that there is a lamp at a point P in M. If M is closed, the light from this lamp will reach every point Q in M. There are no "dark corners" in M.

The theorem is false without the hypothesis that M is closed. Give a counterexample!

Proof: Fix two points Fix $P, Q \in M$ ($P \neq Q$). If M is path connected, there will be a path from P to Q. This looks like a max-min problem! Recall : A continuous function on a compact set has a maximum and a minimum value. Can we use this to find a minimum of the length function on the space of paths from P to Q? Even if P and Q are far apart, such a minimum could only occur at a path whose small subarcs are length minimizing; thus we will have found a geodesic from P to Q. The difficulty is that the space of paths is not compact, not even locally compact. It is possible nevertheless to carry out this argument on the space of H^1 paths (*absolutely continuous paths with finite energy*) where

$$\begin{aligned} E &: H^1(P,Q) \to \mathbb{R} \\ E(\gamma) &= \int_0^1 |\dot{\gamma}|^2 dt < \infty\}. \end{aligned}$$

which has the structure of an infinite-dimensional manifold, and satisfies a "compactness condition" known as Condition C of Palais and Smale: A sequence of paths of bounded energy and with the gradient of the energy going to 0, has a convergent subsequence. The energy turns out to be a better function to work with than the length. Any reparameterization of a geodesic is a critical point of the length function, but a path is a critical point of the energy function if and only if it is geodesic, and therefore parameterized with constant speed . This is a beautiful and elegant theory, and an example of a type of argument that is widespread in mathematics today. However I want to give (or rather assign for homework) a more old-fashioned type of proof of Hopf-Rinow that uses the finite dimensional approximation of Morse. The finite-dimensional approximation is the subset of the space of paths consisting of (continuous) paths $\gamma : [0,1] \to M$ whose restriction to each of the intervals $[0, \frac{1}{N}]$, $[\frac{1}{N}, \frac{2}{N}]$,...,$[\frac{N-1}{N}, 1]$ is a geodesic of length $<\rho$. **FIX** Because each such geodesic is uniquely determined by its endpoints, the set of such paths beginning at P and ending at Q is naturally identified with an open set in the (finite-dimensional!) manifold $M^{N-1} \subset \mathbb{R}^{n(N-1)}$ by the map

$$\gamma \leftrightarrow (P_1, P_2, ..., P_{N-1}) =: (\gamma(\frac{1}{N}), \gamma(\frac{2}{N}), ...\gamma(\frac{N-1}{N}))$$

Finite dim approx

EXERCISE: Let $M \subset \mathbb{R}^m$ be a manifold of dimension n.

(1) Show that the energy of a path γ in the finite dimensional approximation space is

$$E(\gamma) =: E(P_1, P_2, ..., P_{N-1}) = N[d^2(P, P_1) + d^2(P_1, P_2) + d^2(P_2, P_3) + ... + d^2(P_{N-1}, Q)].$$

where d is the minimal geodesic distance on M.

For $a \in \mathbb{R}$ and $N \in \mathbb{N}$ with $\frac{a}{\sqrt{N}}$ less than the injectivity radius of M, let

$$\mathcal{M}_a =: \mathcal{M}_{a,N}^{P,Q} =: \{(P_1, P_2, ..., P_{N-1}) \in M^{N-1} : \ E(P_1, P_2, ..., P_{N-1}) < \frac{a^2}{N}\}$$

(2) Let $\gamma(t)$ be minimal geodesic of length ℓ with $\gamma(0) = R$ and $\gamma(\ell) = Q$, with $\mathbf{U} = \gamma'(0)$ and $\mathbf{V} = \gamma'(\ell)$. (Note $\mathbf{U}$ and $\mathbf{V}$ are unit vectors.) Let $\mathbf{X} \in T_R M$ and $\mathbf{Y} \in T_Q M$, and for $0 \le t \le 1$ let $(R(t), Q(t)) \in M \times M$ be a smooth path starting at $(R(0), Q(0)) = (R, Q)$ with $R'(0) = \mathbf{X}$ and $Q'(0) = \mathbf{Y}$. Show that

$$\frac{d}{dt}|_{t=0} d(R(t), Q(t)) = \mathbf{V} \cdot \mathbf{Y} - \mathbf{U} \cdot \mathbf{X}. \quad (**)$$

(It might be useful to use Lemma 1 above and/or the fact that the geodesic and its length depend smoothly on the enpoints.) It might be useful to use the notation : If $X_i \in T_{P_i} M$, and $f = f(P_1, P_2, ..., P_{N-1})$ is differentiable, then

$$\begin{aligned} X_i f &=: \frac{d}{dt}|_{t=0} f(P_1(t), P_2(t), ..., P_{N-1}(t)) \\ \text{if } \frac{dP_i}{dt} &= X_i \text{ and } \frac{dP_j}{dt} = 0 \text{ if } j \neq i.) \end{aligned}$$

So for example it follows from $(**)$ that

$$\begin{aligned} \mathbf{X}_{i+1} d(P_i, P_{i+1}) &= \mathbf{V}_{i+1} \cdot \mathbf{X}_{i+1} \\ \mathbf{X}_i d(P_i, P_{i+1}) &= -\mathbf{U}_i \cdot \mathbf{X}_i \end{aligned}$$

(What are $\mathbf{U}_i$ and $\mathbf{V}_i$?) If none of the hints is helpful, go on to the next part!

(3) Conclude that the critical points of E on $\mathcal{M}_a$ are precisely the points $(P_1, ..., P_{N-1})$ with the property that $\gamma(P_1, ..., P_{N-1})$ is a geodesic from P to Q. (In other words, all the distances $d(P_i, P_{i+1})$ are the same, **and** the angles at each vertex line up to make a smooth geodesic.) Hint: show that if $\gamma(P_1, ..., P_{N-1})$ is *not* a geodesic (for either reason), then there is a direction at $(P_1, ..., P_{N-1})$ in $\mathcal{M}_a$ in which the energy decreases to first order. (Draw a picture.)

(2) If N is sufficiently large, M is path connected, and there is a path from P to Q of length $< a$, then $\mathcal{M}_a$ is nonempty.

(3) If M is closed, so is $\overline{\mathcal{M}_a} =: \{(P_1, P_2, ..., P_{N-1}) \in M^{N-1} : d^2(P, P_1) + d^2(P_1, P_2) + d^2(P_2, P_3) + ... + d^2(P_{N-1}, P) \leq \frac{a^2}{N}\}$

(4) $\overline{\mathcal{M}_a}$ is bounded.

(5) E is continuous on $\overline{\mathcal{M}_a}$.

Conclude that there is a geodesic on M from P to Q. Note in local coordinates on the manifold M^{N-1} you can use your theorems from multivariable calculus!

Example/exercise: a) Let M be a surface (a 2-dimensional manifold) in $\mathbb{R}^3$. Suppose that a curve γ is the intersection of a plane with M, and that reflection in the plane takes M to M. Suppose a curve γ on M lies on the fixed point set of a reflection $R : \mathbb{R}^3 \to \mathbb{R}^3$ that takes M to M. Show that γ is geodesic.

b) Show that great circles on a standard 2-sphere are closed godesics.

c) Find 3 closed geodesics on the ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1.$$

(When should two (parameterized) paths be considered "different"? When should two (parameterized) closed paths be considered "different"?)

d) Generalize to higher dimensions. Find an infinite number of closed geodesics on a higher dimensional sphere

$$S^k = \{(x_1, ..., x_{k+1}) \in \mathbb{R}^{k+1} : x_1^2 + ... + x_{k+1}^2 = 1.$$

Find $\binom{k+1}{2}$ "different" closed geodesics on the ellipsoid

$$\frac{x_1^2}{a_1^2} + ... + \frac{x_{k+1}^2}{a_{k+1}^2} = 1.$$

Closed Geodesics —How many??

Let M be a manifold. A geodesic $\gamma : \mathbb{R} \to M$ is called *closed* if it has a finite period T, so that $\gamma(t + T) = \gamma(t)$ for all t.

For simplicity I will restrict my discussion to closed, orientable surfaces. The orientability condition is somewhat justifiable in that every unorientable surface has an orientable cover, and closed geodesics on the covering space will descend to closed geodesics on M.

Any constant path on any surfaces is a closed geodesic, but we don't count those. Reparameterizations of a closed curve (different starting point, direction reversed) do not count as "different". Iterates or "multiples" of a geodesic do *not* count separately. A surface of genus ≥ 1 will always have an infinite number of closed geodesics,since there is a closed geodesic in every free homotopy class of closed curves on M. Two closed curves γ and τ are freely homotopic if there is a continuous map of the annulus $S^1 \times I$ into M taking one boundary curve to γ and the other to τ (w correct orientation). To be sure we should make sure that γ and τ are not freely homotopic to *iterates* of the same closed path. I will let you think about how to arrange that. We can prove the existence of a closed geodesic in every free homotopy class of closed curves using the finite dimensional approximation method we used for the Hopf Rinow theorem.

Genus 0

We consider the manifold M gotten by embedding the genus 0 surface

$$S^2 = \{(x, y, z) : x^2 + y^2 + z^2 = 1\}$$

in some Euclidean space $\mathbb{R}^N$, and ask what are the geodesics on the embedded surface.

The above argument does not apply to genus 0 surfaces, since every closed curve on such a surface is freely homotopic to a contant closed path. Since we decided we don't want to count constant loops, the above method produces no closed geodesics at all. Birkhoff in 1927 proved the existence on at least one closed geodesic on a "2-sphere" (or a higher dimensional sphere). He used the "minimax principle" on a 1-paramer family f of closed curves that "cover " the sphere:

He then considers

$$\mu =: \min_{f' \sim f} \max_{\gamma \in im f} length(\gamma)$$

where f' ranges over all "deformations" of f, and shows that $\mu > 0$, and that there is a closed geodesic of length μ on the manifold. With a little topology, the details of this proof can be worked out in the finite dimensional approximation.

Examples. We know that the round sphere has infinitely many closed geodesics. But what happens if we take the round sphere (in $\mathbb{R}^3$ or some higher dimensional space) and deform it a little? Will there still be infinitely many?

Ellipsoids We saw in one of the exercises that every ellipsoid

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$$

has at least 3 closed geodesics, namely the intersections of the ellipsoid with the coordinate planes. Now it has been known for a long time that an ellipsoid always has infinitely many closed geodesics. But Morse showed the following amazing thing: Given $W > 0$ there is $\varepsilon > 0$ so that if

$$1 < a < b < c < 1 + \varepsilon,$$

then the fourth shortest closed geodesic has length at least W. So there really is something special about the number 3.

Embedded closed geodesics on the sphere

A beautiful idea that took a LONG time to have the details worked out.

In 1905 Poincaré conjectured (and gave some arguments supporting the fact that) every orientable genus 0 surface has at least 3 *simple* (nonself-intersecting) closed geodesics. (Note Birkhoff's geodesic is not guaranteed to be embedded.) In 1929, Lusternik and Schnirelmann published a proof of this statement. The idea was beautiful: Start with the round sphere

$$S^2 = \{(x, y, z) : x^2 + y^2 + z^2 = 1\}$$

and the 3-parameter of circles, great and small, on S^2. Embed the sphere, and with it the 3-parameter family of closed curves, in some Euclidean space. Now find a **curve-shortening process** on the space of embedded closed curves. This is to be a continous process that keeps embedded curves embedded , and so that any embedded closed curve that does not shrink to a point converges to a geodesic. The "Lusternik-Schnirelmann category" of the 3-parameter family ensures that the length function must have at least critical points that do not lie in the point loops. (A differentiable function on a closed, bounded set with more than one point must always have at least 2 critical points: a max and a min. point. But for some sets we are guaranteed more by the L-S theory. For example, any differentiable function on a 2-torus must have at least 3 critical points.)

Lusternik and Schnirelmann outlined a curve-shortening process but the details were not correct. Birkhoff had a curve-shortening process that involved replacing small pieces of a curve with geodesic segments **FIG** This process was good at shortening closed curves, but could not guarantee not to introduce self-intersections. (You might say that you don't care whether the closed curves are embedded or not; lets just use the Birkhoff process to find 3 closed geodesics, but in fact the only thing that keeps these three closed geodesics from being multiples of one another is the fact that they are all embedded.) Over the course of the 20th century many curve-shortening processes of this type were proposed, but the arguments always turn out to be very tricky and it seemed to many that the quest for CSP was doomed to failure. I personally refereed many processes and none turned out to be correct. There are a couple of CSP's that I have been told are correct (though I can not personally vouch for them). Ballmann (1978) and Hass-Scott (1994).

Gene Calabi suggested a method that he attributed to Karen Uhlenbeck, a method of flowing a closed curve according to a differential equation as follows: the *geodesic curvature* of a curve on a manifold measures the component of the acceleration that is tangent to the manifold . (So the curvature of a geodesic is 0.) Under the curvature flow, each point on a closed curve evolves according to the curvature vector, which points normal to the curve, and has magitude equal to the geodesic curvature. It is not hard to see that this process shortens curves and does not introduce self-intersections. But it was very difficult to prove that the curvature did not blow up at any point before the curve had a chance to unwind and converge to a point or a geodesic. (By the way this is a lower dimensional analog of the Ricci flow , which is a flow on 2-dimensional surfaces in 3-dimensional manifolds and was used to prove the Geometrization Conjecture and the Poincaré conjecture and which posed similar difficulties.) Grayson in 1989 proved the long-term existence of this flow. From this it follows that any surface of genus 0 surface has at least 3 *simple* closed geodesics.

Further developments.... It is known that a generic metric on a sphere has infinitely many closed geodesics. It is known that any metric on S^2 has infinitely many closed gedesics. This is a very exciting field with lots of plot twists. Look up Ziller's paper on the Katok examples.