The shortest path problem

These notes will locally solve the *shortest path problem* on an n dimensional manifold $M \subset \mathbb{R}^m$. More precisely, given a point p in M, we will show that if q is close enough to p then there is a unique path of shortest length joining p to q. This path is called a *geodesic*. The presentation follows parts of sections 3.2 and 6.1 of [Ta]

1. Manifolds

1.1. **Definition.** We say $M \subset \mathbb{R}^m$ is a C^∞ *n-dimensional manifold* (or *surface*) if, for every $p \in M$ there are open sets $W \subset \mathbb{R}^m$ and $U \subset \mathbb{R}^n$ and a C^∞ function $\varphi\colon U \to W$ such that $p \in W$, $U = \varphi^{-1}(M \cap W)$, $\varphi\colon U \to M \cap W$ is invertible with continuous inverse, and φ' has full rank (i.e. is injective) at all points of U. We call $M \cap W$ a *coordinate patch*, and φ a *coordinate chart*.

1.2. **Examples.**

(1) Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a C^∞ function. Then the graph of M is a C^∞ n-dimensional manifold in $\mathbb{R}^{n+1}$. The whole manifold is a coordinate patch with coordinate chart $\varphi(x) = (x, f(x))$.

(2) Let $a > 0$, and let $M = \varphi(\mathbb{R})$, where $\varphi\colon \mathbb{R} \to \mathbb{R}^2$ be given by

$$\varphi(\alpha) = (a\cos\alpha, a\sin\alpha).$$

Then M is a circle. If we take $U = I$, where I is an interval of length at most 2π, then $\varphi(U)$ is a coordinate patch and the restriction of φ to U is a coordinate chart. If the length of I is greater than 2π, then the restriction of φ is not one-to-one.

(3) The above example works the same way if instead $\varphi\colon \mathbb{R} \to \mathbb{R}^3$ is given by

$$\varphi(\alpha) = (a\cos\alpha, a\sin\alpha, 0).$$

(4) Let $a > b > 0$, and let $M = \varphi(\mathbb{R}^2)$, where $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^3$ is given by

$$\varphi(\alpha,\beta) = (a\cos\alpha, a\sin\alpha, 0) + (b\cos\alpha\cos\beta, b\sin\alpha\cos\beta, b\sin\beta). \tag{1}$$

Then M is a *torus* or *inner tube* as in Figure 3.2.3 of [Ta]. If we take $U = I \times J$, where I and J are both open intervals of length at most 2π, then $\varphi(U)$ is a coordinate patch and the restriction of φ to U is a coordinate chart.

(5) More generally, let $(r(\beta), z(\beta))$ be a parametrized curve $(0,\infty) \times \mathbb{R}$, with r' and z' never simultaneously zero. Define $\varphi(\alpha,\beta) = (r(\beta)\cos\alpha, r(\beta)\sin\alpha, z(\beta))$. Then the image of φ is a 2-manifold in $\mathbb{R}^3$ called a *surface of revolution*.

(6) Another kind of torus is given by $M = \varphi(\mathbb{R}^2)$, where $\varphi\colon \mathbb{R}^2 \to \mathbb{R}^4$ is given by

$$\varphi(\alpha,\beta) = (a\cos\alpha, a\sin\alpha, b\cos\beta, b\sin\beta).$$

Visualize this as the rectangle $[0,a] \times [0,b]$, with opposite edges identified as in an old video game.

1.3. **Exercise.** Check that if φ given by equation (1) then $\varphi'(\alpha,\beta)$ has full rank for any real α and β, and more generally check the same thing about an arbitrary surface of revolution.

Kiril Datchev, February 21, 2024. These are notes are under development, and questions, comments, and corrections are gratefully received at kdatchev@purdue.edu.

1.4. **Curves on manifolds.** Let $\gamma\colon [a,b] \to M$ be a C^∞ curve. Then its length is given by $L = \int_a^b |\gamma'|$. To express this in terms of coordinates, we consider $[a,b]$ short enough that $\gamma([a,b])$ is contained in a coordinate patch, and let $x(t) = \varphi^{-1}(\gamma(t))$. Then $\gamma'(t) = \varphi'(x(t))x'(t)$, and

$$|\gamma'(t)|^2 = |\varphi'(x(t))x'(t)|^2 = \Big(\varphi'(x(t))x'(t)\Big) \cdot \Big(\varphi'(x(t))x'(t)\Big) = \Big(G(x(t))x'(t)\Big) \cdot x'(t),$$

where

$$G(x) = \varphi'(x)^T \varphi'(x).$$

The matrix $G(x)$ is called the *metric tensor*. Its key properties are that it is symmetric (because it is the product of a matrix and its transpose), positive definite (because φ' is injective), and its entries are C^∞ functions of x.

We can also write out the same calculation in terms of components as

$$\begin{aligned}\Big(\varphi'(x(t))x'(t)\Big) \cdot \Big(\varphi'(x(t))x'(t)\Big) &= \sum_{\ell=1}^m \Big(\sum_{j=1}^n \frac{\partial \varphi_\ell(x(t))}{\partial x_j} x_j'(t)\Big)\Big(\sum_{k=1}^n \frac{\partial \varphi_\ell(x(t))}{\partial x_k} x_k'(t)\Big) \\ &= \sum_{j,k=1}^n g_{jk}(x(t)) x_j'(t) x_k'(t),\end{aligned}$$

where

$$g_{jk}(x(t)) = \sum_{\ell=1}^m \frac{\partial \varphi_\ell(x(t))}{\partial x_j} \frac{\partial \varphi_\ell(x(t))}{\partial x_k};$$

in other words, g_{jk} is the dot product of the jth and kth columns of φ'. We abbreviate this further using the *summation convention* that repeated indices are summed over, as

$$|\gamma'(t)|^2 = |\varphi'(x(t))x'(t)|^2 = g_{jk}(x(t))x_j'(t)x_k'(t) = g_{jk}x_j'x_k'.$$

Thus the g_{jk} are the entries of the matrix G.

1.5. **Example.** In Example 1.2 (1), the matrix of φ' is $\binom{I}{f'}$, where I is the $n \times n$ identity matrix and f' is the $1 \times n$ gradient matrix $(\partial_1 f, \dots, \partial_n f)$. Thus $G = (\varphi')^T \varphi' = I + (f')^T f'$, or

$$g_{jk} = \delta_{jk} + \partial_j f \partial_k f,$$

where $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ if $j \neq k$. In other words the metric tensor in these coordinates is the sum of the identity matrix and a scaled orthogonal projection onto the gradient of f.

1.6. **Exercise.** Find g_{jk} for the helicoid[1] parametrized by

$$\varphi\colon (0,1) \times \mathbb{R} \to \mathbb{R}^3, \qquad \varphi(x_1, x_2) = (x_1 \cos x_2, x_1 \sin x_2, x_2),$$

and for the sphere parametrized by

$$\varphi\colon \mathbb{R} \times (0,\pi) \to \mathbb{R}^3, \qquad \varphi(x_1, x_2) = (\cos x_1 \sin x_2, \sin x_1 \sin x_2, \cos x_2).$$

[1] See `https://en.wikipedia.org/wiki/Helicoid`

2. Geodesics

2.1. **Definition of a geodesic.** We now define a geodesic in terms of such coordinates. For $1 \le j,\, k \le n$, let $g_{jk} \in C^\infty(\mathbb{R}^n)$ be such that for all x we have $g_{jk}(x) = g_{kj}(x)$ and $g_{jk}(x)v_jv_k > 0$ for any nonzero $v = (v_1, \dots, v_n) \in \mathbb{R}^n$, where we use the summation convention that repeated indices are summed over, i.e. $g_{jk}v_jv_k = \sum_{j=1}^n \sum_{k=1}^n g_{jk}v_jv_k$. If $x = (x_1, \dots, x_n)\colon [a, b] \to \mathbb{R}^n$ is a C^∞ curve, its *speed* with respect to g at time t is given by $\sqrt{g_{jk}(x(t))x_j'(t)x_k'(t)}$, and its *length* with respect to g is given by

$$L = L_g(x) = \int_a^b \sqrt{g_{jk}x_j'x_k'}dt. = \int_a^b \sqrt{g_{jk}(x(t))x_j'(t)x_k'(t)}dt.$$

We wish to find the curves which minimize this length. We begin by looking at the simpler *energy functional*, which is defined by

$$E = E_g(x) = \frac{1}{2}\int_a^b g_{jk}x_j'x_k'.$$

To look for a minimizer, we take $x(t) = x(t, s)$ such that $x(a, s)$ and $x(b, s)$ are both independent of s, and differentiate with respect to s:

$$\partial_s E = \int_a^b \left(g_{jk}x_j'\partial_s x_k' + \tfrac{1}{2}\partial_\ell g_{jk}x_j'x_k'\partial_s x_\ell'\right)dt.$$

Then we integrate by parts in the first term (removing a t derivative from $\partial_s x_k'$) and swap the k and ℓ indices in the last term to get

$$\partial_s E = \int_a^b \left(-g_{jk}x_j'' - \partial_\ell g_{jk}x_\ell'x_j' + \tfrac{1}{2}\partial_k g_{j\ell}x_j'x_\ell'\right)\partial_s x_k dt.$$

If x_0 is a critical point of the energy functional, then $\partial_s E|_{s=0} = 0$. If this is the case for any variation x_s, then we obtain the system of geodesic equations

$$-g_{jk}x_j'' - \partial_\ell g_{jk}x_\ell'x_j' + \tfrac{1}{2}\partial_k g_{j\ell}x_j'x_\ell' = 0, \qquad \text{for } k = 1, \dots, n. \tag{2}$$

Any solution to (2) is called a *geodesic* with respect to G. Solutions are guaranteed to exist by the existence and uniqueness theorem for ordinary differential equations; more precisely, given any initial conditions p and v in $\mathbb{R}^n$, for some $T > 0$ there is a unique C^∞ function $x\colon (-T, T) \to \mathbb{R}^n$ such that (2) holds and $x(0) = p$ and $\dot{x}(0) = v$. Moreover, as p and v vary smoothly, so does the solution. (We will come back to this big theorem)

We will see shortly that geodesics locally minimize length. For now, observe that differentiating $g_{jk}x_j'x_k'$ with respect to t and plugging in (2) shows that all geodesics have constant speed. Note also that if $x(t)$ is a geodesic, so is $x(\lambda t)$ for any real λ.

2.2. **Examples.** The simplest example is free space, where $g_{jk} \equiv \delta_{jk}$, and the geodesic equations become $x'' = 0$. Thus the geodesics are lines traversed at constant speed; $x(t) = p + vt$.

In the torus example, as in equation (1), we have

$$\varphi' = \begin{pmatrix} -\sin\alpha(a + b\cos\beta) & -b\cos\alpha\sin\beta \\ \cos\alpha(a + b\cos\beta) & -b\sin\alpha\sin\beta \\ 0 & b\cos\beta \end{pmatrix}, \quad G = (\varphi')^T\varphi' = \begin{pmatrix} (a + b\cos\beta)^2 & 0 \\ 0 & b^2 \end{pmatrix}.$$

Thus $g_{11} = (a + b\cos\beta)^2$, $g_{12} = g_{21} = 0$, and $g_{22} = b^2$. The geodesic equations are

$$-(a + b\cos\beta)^2\alpha'' - 2(a + b\cos\beta)(-b\sin\beta)\alpha'\beta' = 0,$$
$$-b^2\beta'' + (a + b\cos\beta)(-b\sin\beta)\alpha'\alpha' = 0.$$

The general solutions of this system do not have a simple form, but the linear ones do, namely either $\alpha(t) = a$ and $\beta(t) = b + ct$, or else $\alpha(t) = a + bt$ and $\beta(t) = m\pi$ for some integer m.

2.3. **Exercises.**

(1) Find the geodesic equations for the ellipsoid parametrized by

$$\varphi\colon \mathbb{R} \times (0, \pi) \to \mathbb{R}^3, \qquad \varphi(x^1, x^2) = (a \cos x^1 \sin x^2, a \sin x^1 \sin x^2, b \cos x^2),$$

where $a > 0$ and $b > 0$ are given constants. Find all (p, v) for which $x(t) = p + vt$ is a geodesic.

(2) Let $R > 0$ be given, and let p and q be two points on a circular cylinder of radius R. Use the parametrization $\varphi(\theta, z) = (R \cos\theta, R \sin\theta, z)$, where $\varphi(0, 0) = p$ and $\varphi(\theta_0, z_0) = q$ for some $(\theta_0, z_0) \in [-\pi, \pi] \times \mathbb{R}$, to find the lengths of all possible geodesics from p to q.

2.4. **Meridians on surfaces of revolution.** Let M be a surface of revolution, with $\varphi(\alpha, \beta) = (r(\beta) \cos\alpha, r(\beta) \sin\alpha, z(\beta))$, and suppose for simplicity that $r'(\beta)^2 + z'(\beta)^2 = 1$; i.e. meridians of the surface a parametrized by arclength (this is also known as the unit speed parametrization). Then, as in the torus example above, we have

$$G(\alpha, \beta) = \begin{pmatrix} r(\beta)^2 & 0 \\ 0 & r'(\beta)^2 + z'(\beta)^2 \end{pmatrix} = \begin{pmatrix} r(\beta)^2 & 0 \\ 0 & 1 \end{pmatrix},$$

and the meridians are geodesics, in the sense that any parametrized curve with $\alpha(t)$ constant and with $\beta(t) = b + ct$ is a geodesic. Let's check that these minimize length.

To that, let p and q be two points on the same meridian, i.e. they are given by $p = (\alpha_0, \beta_0)$ and $q = (\alpha_0, \beta_1)$. Let $(\alpha(t), \beta(t))$ be a curve joining them, i.e. it obeys $\alpha(a) = \alpha_0$, $\alpha(b) = \alpha(0)$, $\beta(a) = \beta_0$, $\beta(b) = \beta_1$. Then its length obeys

$$L = \int_a^b |\gamma'(t)| dt = \int_a^b \sqrt{r(\beta)^2 \alpha'(t)^2 + \beta'(t)^2} dt \geq \int_a^b |\beta'(t)| dt \geq \Big| \int_a^b \beta'(t) dt \Big| = |\beta(b) - \beta(a)|.$$

Note that both inequalities become equalities if the curve is a geodesic, i.e. if $\alpha(t)$ constant and $\beta(t) = b + ct$. Moreover, these are the unique minimizers up to reparametrization, because if α is not constant then the first inequality is strict. Thus we conclude that meridians are solutions to the least path problem on surfaces of revolution. Also note that $|\beta(b) - \beta(a)|$ is a nice formula for the length of the path from p to q.

Of course the above only works if p and q are on the same meridian. But if M is a sphere then this covers the general case, because we can always choose our coordinates so that p and q lie on the same meridian. (The only exception is if p and q are antipodal, but in that case we know the least path problem has no unique solution.)

2.5. **Normal coordinates.** The above discussion of spheres showcases the simplifying power of well-chosen coordinates. To prove that geodesics locally minimize length on a more general manifold, we introduce *normal coordinates* near 0; these are coordinates which give the metric the simplest possible form near 0. We define them in two steps.

We first use the fact that $G(0)$ is positive definite and the spectral theorem to get a basis of eigenvectors of $G(0)$, which are orthonormal with respect to $G(0)$, i.e.

$$\Big[G(0) u_k \Big] \cdot u_\ell = \delta_{k\ell},$$

and let A be a matrix whose columns are the u_j. For example, if $G(0) = ((0, 3), (3, 0))$, then we can use $u_1 = (1, 1)/3\sqrt{2}$ and $u_2 = (1, -1)/3\sqrt{2}$ and $A = ((1, 1), (1, -1))/3\sqrt{2}$. Then put

$$x = A\tilde{x},$$

so that

$$\Big[G(x(t))x'(t)\Big] \cdot x'(t) = \Big[\tilde{G}(\tilde{x}(t))\tilde{x}'(t)\Big] \cdot \tilde{x}'(t)$$

where

$$\tilde{G}(\tilde{x}) = A^T G(A\tilde{x})A.$$

That makes $\tilde{G}(0) = I$, or

$$\tilde{g}_{ij}(0) = \delta_{ij}. \tag{3}$$

Second, let $\mathrm{Exp}(v) = \gamma_v(1)$, where $\tilde{x}(t) = \gamma_v(t)$ is the geodesic such that $\gamma_v(0) = 0$ and $\gamma_v'(0) = v$. This may not be defined for all v (because the geodesic equation may not be solvable up to $t = 1$ for all v), but the existence, uniqueness, and regularity theorem for ordinary differential equations guarantees that $\mathrm{Exp}(v)$ is defined for v small enough and that $\mathrm{Exp}\colon B(c) \to \mathbb{R}^n$ is a C^∞ map for some $c > 0$, where $B(c)$ is the open ball centered at 0 with radius c. The derivative of Exp at 0 is the identity, because

$$\mathrm{Exp}(\lambda v) = \gamma_{\lambda v}(1) = \gamma_v(\lambda) = \gamma_v(0) + \lambda\gamma_v'(0) + r(\lambda, v) = \lambda v + r(\lambda, v),$$

where $|r(\lambda, v)|/\lambda \to 0$ as $\lambda \to 0$. Hence, by the inverse function theorem, after possibly shrinking c, the map $\mathrm{Exp}\colon B(c) \to \mathrm{Exp}(B(c))$ is invertible with C^∞ inverse.

In these coordinates we write

$$\tilde{x} = \mathrm{Exp}(y)$$

and as before

$$\Big[G(x(t))x'(t)\Big] \cdot x'(t) = \Big[\tilde{G}(\tilde{x}(t))\tilde{x}'(t)\Big] \cdot \tilde{x}'(t) = \Big[H(y(t))y'(t)\Big] \cdot y'(t).$$

In these coordinates, if $y(t)$ is a geodesic and $y(0) = 0$, $y'(0) = v$, then $y(t) = tv$ since $\mathrm{Exp}(tv) = \gamma_{tv}(1) = \gamma_v(t)$. Further, since geodesics have constant speed, we have

$$\Big[H(y(t))y'(t)\Big] \cdot y'(t) = \Big[H(y(0))y'(0)\Big] \cdot y'(0) = \Big[H(0)v\Big] \cdot v = |v|^2, \tag{4}$$

where for the last equals we used

$$H(0) = (\mathrm{Exp}'(0))^T \tilde{G}(0)\,\mathrm{Exp}'(0) = I.$$

Thus we know h on radial vectors. The Gauss Lemma says that radial vectors are h-perpendicular to those vectors to which they are Euclidean-perpendicular; i.e.

$$v \cdot w = 0 \quad \Longrightarrow \quad \big[H(v)v\big] \cdot w = h_{jk}(v)v_j w_k = 0. \tag{5}$$

We will prove the Gauss lemma momentarily. To see how it is used, let $y(t)$ be any path in $B(c) \setminus 0$, and let $r(t) = |y(t)|$. Then $y(t) = r(t)\omega(t)$, where $\omega(t) = y(t)/|y(t)|$, and

$$\Big[H(y(t))y'(t)\Big] \cdot y'(t) = \Big[Hy'\Big] \cdot y' = r'^2\Big[H\omega\Big] \cdot \omega + r^2\Big[H\omega'\Big] \cdot \omega' = r'^2 + r^2\Big[H\omega'\Big] \cdot \omega' \geq r'^2.$$

where for the second equality we used the fact that (5) implies $[H\omega] \cdot \omega' = 0$, for the third equality we used (4), and the inequality is strict if and only if $\omega' \neq 0$ because H is positive definite.

Now let $p \in B(c)$, let $y\colon [a, b] \to B(c)$ be any smooth path with $y(a) = 0$ and $y(b) = p$. The length is

$$\int_a^b \sqrt{\Big[H(y(t))y'(t)\Big] \cdot y'(t)}dt \geq \int_a^b |r'| \geq \int_a^b r' = r(b) - r(a) = |p|,$$

with equality if and only if ω is constant and r is monotonic, i.e. if and only if $y(t)$ monotonically parametrizes the line segment from 0 to p. Thus a parametric curve γ from 0 to p always has length $\geq$ the length of the geodesic curve from 0 to p, and the lengths are equal if and only γ monotonically parametrizes the geodesic curve.

It remains to prove the Gauss Lemma. Without loss of generality, $|v| = |w| = 1$. Then put

$$y(t) = y(t,s) = (v\cos s + w\sin s)t \qquad E(y(t)) = \frac{1}{2}\int_0^1 h_{jk}(y(t))y_j'(t)y_k'(t)dt,$$

Then, as before, we get

$$\partial_s E = \int_0^1 \left(h_{jk}y_j'\partial_s y_k' + \tfrac{1}{2}\partial_\ell h_{jk}y_j'y_k'\partial_s y_\ell\right)dt,$$

and integrating by parts, and swapping the k and ℓ indices in the last term gives

$$\partial_s E = h_{jk}(y(1))y_j'(1)\partial_s y_k(1) + \int_0^1 \left(-h_{jk}y_j'' - \partial_\ell h_{jk}y_\ell' y_j' + \tfrac{1}{2}\partial_k h_{j\ell}y_j'y_\ell'\right)\partial_s y_k dt.$$

Combining this with $\partial_s E = 0$ (because y has unit speed for every s) and observing that the integrand vanishes because y solves the geodesic equation, and plugging in $s = 0$, gives (5).

References

[Ta] Michael E. Taylor, *Introduction to Analysis in Several Variables*, AMS Sally Series of Pure and Applied Undergraduate Texts 46, 2020. Preprint available online at `https://mtaylor.web.unc.edu/wp-content/uploads/sites/16915/2018/04/analmv.pdf`.