CHAPTER 5

Geodesics and Distance

We are now ready to introduce the important concept of a geodesic. This will help us define and understand Riemannian manifolds as metric spaces. One is led quickly to two types of "completeness". The first is of standard metric completeness, and the other is what we call geodesic completeness, namely, when all geodesics exist for all time. We shall prove the Hopf-Rinow Theorem, which asserts that these types of completeness for a Riemannian manifold are equivalent. Using the metric structure we can define metric distance functions. We shall study when these distance functions are smooth and show the existence of the smooth distance functions we worked with earlier. In the last section we give some metric characterizations of Riemannian isometries and submersions. We also classify complete simply connected manifolds of constant curvature; showing that they are the ones we have already constructed in chapters 1 and 3.

The idea of thinking of a Riemannian manifold as a metric space must be old, but it wasn't until the early 1920s that first Cartan and then later Hopf and Rinow began to understand the relationship between extendability of geodesics and completeness of the metric. Nonetheless, both Gauss and Riemann had a pretty firm grasp on local geometry, as is evidenced by their contributions: Gauss worked with geodesic polar coordinates and also isothermal coordinates, Riemann was able to give a local characterization of Euclidean space as the only manifold whose curvature tensor vanishes. Surprisingly, it wasn't until Klingenberg's work in the 1950s that one got a thorough understanding of the maximal domain on which one has geodesic polar coordinates in side complete manifolds. This work led to the introduction of the two terms *injectivity radius* and *conjugate radius*. Many of our later results will require a detailed analysis of these concepts. The metric characterization of Riemannian isometries wasn't realized until the late 1930s with the work of Myers and Steenrod. Even more surprising is Berestovskii's much more recent metric characterization of Riemannian submersions.

Another important topic that involves geodesics is the variation of arclength and energy. In this chapter we only develop the first variation formula. This is used to show that curves that minimize length must be geodesics if they are parametrized correctly.

We are also finally getting to results where there is going to be a significant difference between the Riemannian setting and the semi-Riemannian setting. Mixed partials and geodesics easily generalize. However, as there is no norm of vectors in the semi-Riemannian setting we do not have arclength or distances. Nevertheless, the energy functional does make sense so we can still obtain a variational characterization of geodesic as critical points for the energy functional.

1. Mixed Partials

So far we have only worked out the calculus for functions on a Riemannian manifold and have seen that defining the gradient and Hessian requires that we use the metric structure. We are now going to study maps into Riemannian manifolds and how to define meaningful derivatives for such maps. The simplest example is to consider a curve $\gamma : I \to M$ on some interval $I \subset \mathbb{R}$. We know how to define the derivative $\dot{\gamma}$, but not how to define the acceleration in such a way that it also gives us a tangent vector to M. A similar but slightly more general problem is that of defining mixed partial derivatives

$$\frac{\partial^2 \gamma}{\partial t^i \partial t^j}$$

for maps γ with several real variables. As we shall see, covariant differentiation plays a crucial role in the definition of these concepts. In this section we only develop a method that covers second partials. In the next chapter we shall explain how to calculate higher order partials as well. This involves a slightly different approach that is not needed for the developments in this chapter.

Let $\gamma : \Omega \to M$, where $\Omega \subset \mathbb{R}^m$. As we usually reserve x^i for coordinates on M we shall use t^i or s, t, u as coordinates on Ω. The first partials

$$\frac{\partial \gamma}{\partial t^i}$$

are simply defined as the velocity field of $t^i \to \gamma\left(t^1, ..., t^i, ..., t^m\right)$ where the remaining coordinates are fixed. We wish to define the second partials so that they also lie TM as opposed to TTM. In addition we shall also require the following two natural properties:

$$\begin{array}{ll} (1) & \frac{\partial^2 \gamma}{\partial t^i \partial t^j} = \frac{\partial^2 \gamma}{\partial t^j \partial t^i}, \\ (2) & \frac{\partial}{\partial t^k} g\left(\frac{\partial \gamma}{\partial t^i}, \frac{\partial \gamma}{\partial t^j}\right) = g\left(\frac{\partial^2 \gamma}{\partial t^k \partial t^i}, \frac{\partial \gamma}{\partial t^j}\right) + g\left(\frac{\partial \gamma}{\partial t^i}, \frac{\partial^2 \gamma}{\partial t^k \partial t^j}\right). \end{array}$$

The first is simply the equality of mixed partials and is similar to assuming that the connection is torsion free. The second is a Leibniz or product rule that is similar to assuming that the connection is metric. Like the Fundamental Theorem of Riemannian Geometry, were we saw that the key properties of the connection in fact also characterized the connection, we can show that these two rules also characterize how we define second partials. More precisely, if we have a way of defining second partials such that these two properties hold, then we claim that there is a Koszul type formula:

$$2g\left(\frac{\partial^2 \gamma}{\partial t^i \partial t^j}, \frac{\partial \gamma}{\partial t^k}\right) = \frac{\partial}{\partial t^i} g\left(\frac{\partial \gamma}{\partial t^j}, \frac{\partial \gamma}{\partial t^k}\right) + \frac{\partial}{\partial t^j} g\left(\frac{\partial \gamma}{\partial t^k}, \frac{\partial \gamma}{\partial t^i}\right) - \frac{\partial}{\partial t^k} g\left(\frac{\partial \gamma}{\partial t^i}, \frac{\partial \gamma}{\partial t^j}\right).$$

This formula is established in the proof of the next lemma.

LEMMA 6. (Uniqueness of mixed partials) *There is at most one way of defining mixed partials so that (1) and (2) hold.*

PROOF. First we show that the Koszul type formula holds if we have a way of defining mixed partials such that (1) and (2) hold:

$$\begin{aligned}
&\frac{\partial}{\partial t^i}g\left(\frac{\partial\gamma}{\partial t^j},\frac{\partial\gamma}{\partial t^k}\right)+\frac{\partial}{\partial t^j}g\left(\frac{\partial\gamma}{\partial t^k},\frac{\partial\gamma}{\partial t^i}\right)-\frac{\partial}{\partial t^k}g\left(\frac{\partial\gamma}{\partial t^i},\frac{\partial\gamma}{\partial t^j}\right)\\
=\;& g\left(\frac{\partial^2\gamma}{\partial t^i\partial t^j},\frac{\partial\gamma}{\partial t^k}\right)+g\left(\frac{\partial\gamma}{\partial t^j},\frac{\partial^2\gamma}{\partial t^i\partial t^k}\right)\\
&+g\left(\frac{\partial^2\gamma}{\partial t^j\partial t^k},\frac{\partial\gamma}{\partial t^i}\right)+g\left(\frac{\partial\gamma}{\partial t^k},\frac{\partial^2\gamma}{\partial t^j\partial t^i}\right)\\
&-g\left(\frac{\partial^2\gamma}{\partial t^k\partial t^i},\frac{\partial\gamma}{\partial t^j}\right)-g\left(\frac{\partial\gamma}{\partial t^i},\frac{\partial^2\gamma}{\partial t^k\partial t^j}\right)\\
=\;& g\left(\frac{\partial^2\gamma}{\partial t^i\partial t^j},\frac{\partial\gamma}{\partial t^k}\right)+g\left(\frac{\partial\gamma}{\partial t^k},\frac{\partial^2\gamma}{\partial t^j\partial t^i}\right)\\
&+g\left(\frac{\partial\gamma}{\partial t^j},\frac{\partial^2\gamma}{\partial t^i\partial t^k}\right)-g\left(\frac{\partial^2\gamma}{\partial t^k\partial t^i},\frac{\partial\gamma}{\partial t^j}\right)\\
&+g\left(\frac{\partial^2\gamma}{\partial t^j\partial t^k},\frac{\partial\gamma}{\partial t^i}\right)-g\left(\frac{\partial\gamma}{\partial t^i},\frac{\partial^2\gamma}{\partial t^k\partial t^j}\right)\\
=\;& 2g\left(\frac{\partial^2\gamma}{\partial t^i\partial t^j},\frac{\partial\gamma}{\partial t^k}\right).
\end{aligned}$$

Next we observe that if we have a map $\gamma:\Omega\to M$, then we can always add an extra parameter t^{n+1} to get a map $\bar{\gamma}:\Omega\times(-\varepsilon,\varepsilon)\to M$ with the property that

$$\frac{\partial\bar{\gamma}}{\partial t^{n+1}}|_p=v\in T_pM,$$

where $v\in T_pM$ is any vector and p is any point in the image of γ. Using $k=n+1$ in the Koszul type formula at p, then shows that $\frac{\partial^2\gamma}{\partial t^i\partial t^j}$ is uniquely defined as our extension is independent of how mixed partials are defined. □

We can now give a local coordinate definition of mixed partials. As long as the definition gives us properties (1) and (2), the above lemma shows that we have a coordinate independent definition.

Note also that if two different maps $\gamma_1,\gamma_2:\Omega\to M$ agree on a neighborhood of a point in the domain, then the right hand side of the Koszul type formula will give the same answer for these two maps. Thus there is no loss of generality in assuming that the image of γ lies in a coordinate system.

THEOREM 9. (Existence of mixed partials) *It is possible to define mixed partials in a coordinate system so that (1) and (2) hold.*

PROOF. We assume that we have $\gamma:\Omega\to U\subset M$ where U is a coordinate neighborhood. Furthermore, assume that the parameters in use are called s and t. This avoids introducing more indices than necessary. Finally write $\gamma=\left(\gamma^1,...,\gamma^n\right)$ using the coordinates. The velocity in the s direction is given by

$$\frac{\partial\gamma}{\partial s}=\frac{\partial\gamma^i}{\partial s}\partial_i$$

so we can make the suggestive calculation

$$\begin{aligned}\frac{\partial}{\partial t}\frac{\partial \gamma}{\partial s} &= \frac{\partial}{\partial t}\left(\frac{\partial \gamma^i}{\partial s}\partial_i\right)\\ &= \frac{\partial}{\partial t}\frac{\partial \gamma^i}{\partial s}\partial_i + \frac{\partial \gamma^i}{\partial s}\frac{\partial}{\partial t}\left(\partial_i\right).\end{aligned}$$

To make sense of $\frac{\partial}{\partial t}\left(\partial_i\right)$ we define

$$\frac{\partial X}{\partial t}|_p = \nabla_{\dot\gamma(t)} X,$$

where $\gamma\left(t\right) = p$ and X is a vector field defined in a neighborhood of p. With that in mind we have

$$\begin{aligned}\frac{\partial}{\partial t}\frac{\partial \gamma}{\partial s} &= \frac{\partial^2 \gamma^k}{\partial t \partial s}\partial_k + \frac{\partial \gamma^i}{\partial s}\nabla_{\frac{\partial \gamma}{\partial t}}\partial_i\\ &= \frac{\partial^2 \gamma^k}{\partial t \partial s}\partial_k + \frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^j}{\partial t}\nabla_{\partial_j}\partial_i\\ &= \frac{\partial^2 \gamma^k}{\partial t \partial s}\partial_k + \frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^j}{\partial t}\Gamma^k_{ji}\partial_k\end{aligned}$$

Thus we define

$$\begin{aligned}\frac{\partial^2 \gamma}{\partial t \partial s} &= \frac{\partial^2 \gamma^k}{\partial t \partial s}\partial_k + \frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^j}{\partial t}\Gamma^k_{ji}\partial_k\\ &= \left(\frac{\partial^2 \gamma^k}{\partial t \partial s} + \frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^j}{\partial t}\Gamma^k_{ji}\right)\partial_k\end{aligned}$$

Since $\frac{\partial^2 \gamma^l}{\partial t \partial s}$ is symmetric in s and t by the usual theorem on equality of mixed partials and the Christoffel symbol Γ^k_{ji} is symmetric in i and j we see that (1) holds.

To check the metric property (2) we use that the Christoffel symbols satisfy the metric property

$$\partial_k g_{ij} = \Gamma_{ki,j} + \Gamma_{kj,i}.$$

With that in mind we calculate

$$\begin{aligned}&\frac{\partial}{\partial t} g\left(\frac{\partial \gamma}{\partial s}, \frac{\partial \gamma}{\partial u}\right)\\ =\ &\frac{\partial}{\partial t}\left(g_{ij}\frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^j}{\partial u}\right)\\ =\ &\frac{\partial g_{ij}}{\partial t}\frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^j}{\partial u} + g_{ij}\frac{\partial^2 \gamma^i}{\partial t \partial s}\frac{\partial \gamma^j}{\partial u} + g_{ij}\frac{\partial \gamma^i}{\partial s}\frac{\partial^2 \gamma^j}{\partial t \partial u}\\ =\ &g_{ij}\left(\frac{\partial^2 \gamma^i}{\partial t \partial s} + \frac{\partial \gamma^k}{\partial s}\frac{\partial \gamma^l}{\partial t}\Gamma^i_{kl}\right)\frac{\partial \gamma^j}{\partial u} + g_{ij}\frac{\partial \gamma^i}{\partial s}\left(\frac{\partial^2 \gamma^j}{\partial t \partial u} + \frac{\partial \gamma^k}{\partial u}\frac{\partial \gamma^l}{\partial t}\Gamma^j_{kl}\right)\end{aligned}$$

$$
\begin{aligned}
&+\frac{\partial g_{ij}}{\partial t}\frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^j}{\partial u} - g_{ij}\frac{\partial \gamma^k}{\partial s}\frac{\partial \gamma^l}{\partial t}\frac{\partial \gamma^j}{\partial u}\Gamma^i_{kl} - g_{ij}\frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^k}{\partial u}\frac{\partial \gamma^l}{\partial t}\Gamma^j_{kl}\\
=\;& g\left(\frac{\partial^2\gamma}{\partial t\partial s},\frac{\partial\gamma}{\partial u}\right)+g\left(\frac{\partial\gamma}{\partial s},\frac{\partial^2\gamma}{\partial t\partial u}\right)\\
&+\frac{\partial g_{ij}}{\partial t}\frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^j}{\partial u} - \frac{\partial \gamma^k}{\partial s}\frac{\partial \gamma^l}{\partial t}\frac{\partial \gamma^j}{\partial u}\Gamma_{kl,j} - \frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^k}{\partial u}\frac{\partial \gamma^l}{\partial t}\Gamma^j_{kl,i}\\
=\;& g\left(\frac{\partial^2\gamma}{\partial t\partial s},\frac{\partial\gamma}{\partial u}\right)+g\left(\frac{\partial\gamma}{\partial s},\frac{\partial^2\gamma}{\partial t\partial u}\right)\\
&+\partial_k g_{ij}\frac{\partial \gamma^k}{\partial t}\frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^j}{\partial u} - \frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^k}{\partial t}\frac{\partial \gamma^j}{\partial u}\Gamma_{ki,j} - \frac{\partial \gamma^i}{\partial s}\frac{\partial \gamma^j}{\partial u}\frac{\partial \gamma^k}{\partial t}\Gamma_{kj,i}\\
=\;& g\left(\frac{\partial^2\gamma}{\partial t\partial s},\frac{\partial\gamma}{\partial u}\right)+g\left(\frac{\partial\gamma}{\partial s},\frac{\partial^2\gamma}{\partial t\partial u}\right).
\end{aligned}
$$

□

In case $M \subset N$ it is often convenient to calculate the mixed partials in N first and then project them onto M. For each $p \in M$ we use the orthogonal projection $\mathrm{proj}_M : T_pN \to T_pM$. The next proposition shows that this is a valid way of calculating mixed partials.

PROPOSITION 16. (Mixed partials in submanifolds) *If $\gamma : \Omega \to M \subset N$ and $\frac{\partial^2\gamma}{\partial t^i\partial t^j} \in T_pN$ is the mixed partial in N, then*

$$\mathrm{proj}_M\left(\frac{\partial^2\gamma}{\partial t^i\partial t^j}\right) \in T_pM$$

is the mixed partial in M.

PROOF. Let $\bar g$ be the Riemannian metric in N and g its restriction to the submanifold M. We know that $\frac{\partial^2\gamma}{\partial t^j\partial t^i} \in TN$ satisfies

$$
\begin{aligned}
2\bar g\left(\frac{\partial^2\gamma}{\partial t^i\partial t^j},\frac{\partial\gamma}{\partial t^k}\right) &= \frac{\partial}{\partial t^i}\bar g\left(\frac{\partial\gamma}{\partial t^j},\frac{\partial\gamma}{\partial t^k}\right)+\frac{\partial}{\partial t^j}\bar g\left(\frac{\partial\gamma}{\partial t^k},\frac{\partial\gamma}{\partial t^i}\right)\\
&\quad -\frac{\partial}{\partial t^k}\bar g\left(\frac{\partial\gamma}{\partial t^i},\frac{\partial\gamma}{\partial t^j}\right).
\end{aligned}
$$

As $\frac{\partial\gamma}{\partial t^i},\frac{\partial\gamma}{\partial t^j},\frac{\partial\gamma}{\partial t^k} \in TM$ this shows that

$$
\begin{aligned}
2\bar g\left(\frac{\partial^2\gamma}{\partial t^i\partial t^j},\frac{\partial\gamma}{\partial t^k}\right) &= \frac{\partial}{\partial t^i} g\left(\frac{\partial\gamma}{\partial t^j},\frac{\partial\gamma}{\partial t^k}\right)+\frac{\partial}{\partial t^j} g\left(\frac{\partial\gamma}{\partial t^k},\frac{\partial\gamma}{\partial t^i}\right)\\
&\quad -\frac{\partial}{\partial t^k} g\left(\frac{\partial\gamma}{\partial t^i},\frac{\partial\gamma}{\partial t^j}\right).
\end{aligned}
$$

Next use that $\frac{\partial\gamma}{\partial t^k} \in TM$ to alter the left hand side to

$$2\bar g\left(\frac{\partial^2\gamma}{\partial t^i\partial t^j},\frac{\partial\gamma}{\partial t^k}\right) = 2g\left(\mathrm{proj}_M\left(\frac{\partial^2\gamma}{\partial t^i\partial t^j}\right),\frac{\partial\gamma}{\partial t^k}\right).$$

This shows that $\mathrm{proj}_M\left(\frac{\partial^2\gamma}{\partial t^i\partial t^j}\right)$ is the correct mixed partial in M. □

We shall use this way of calculating mixed partials in several situations below.

2. Geodesics

We can now define the acceleration of a curve $\gamma : I \to M$ by the formula

$$\ddot{\gamma} = \frac{d^2\gamma}{dt^2}.$$

In local coordinates this becomes

$$\ddot{\gamma} = \frac{d^2\gamma^k}{dt^2}\partial_k + \frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt}\Gamma^k_{ij}\partial_k.$$

A C^∞ curve $\gamma : I \to M$ is called a *geodesic* if $\ddot{\gamma} = 0$. If γ is a geodesic, then the speed $|\dot{\gamma}| = \sqrt{g(\dot{\gamma}, \dot{\gamma})}$ is constant, as

$$\frac{d}{dt} g(\dot{\gamma}, \dot{\gamma}) = 2g(\ddot{\gamma}, \dot{\gamma}) = 0.$$

So a geodesic is a constant-speed curve, or phrased differently, it is parametrized proportionally to arc length. If $|\dot{\gamma}| \equiv 1$, one says that γ is parametrized by arc length.

If $r : U \to \mathbb{R}$ is a distance function, then we know that for $\partial_r = \nabla r$ we have $\nabla_{\partial_r}\partial_r = 0$. The integral curves for $\nabla r = \partial_r$ are therefore geodesics. Below we shall develop a theory for geodesics independently of distance functions and then use this to show the existence of distance functions.

Geodesics are fundamental in the study of the geometry of Riemannian manifolds in the same way that straight lines are fundamental in Euclidean geometry. At first sight, however, it is not even clear that there are going to be any nonconstant geodesics to study on a general Riemannian manifold. In this section we are going to establish that every Riemannian manifold has many non-constant geodesics. Informally speaking, we can find a unique one at each point with a given tangent vector at that point. However, the question of how far it will extend from that point is subtle. To deal with the existence and uniqueness questions, we need to use some information from differential equations.

In local coordinates on $U \subset M$ the equation for a curve to be a geodesic is:

$$\begin{aligned} 0 &= \ddot{\gamma} \\ &= \frac{d^2\gamma^k}{dt^2}\partial_k + \frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt}\Gamma^k_{ij}\partial_k \end{aligned}$$

Thus, the curve $\gamma : I \to U$ is a geodesic if and only if the coordinate components γ^k satisfy:

$$\ddot{\gamma}^k(t) = -\dot{\gamma}^i(t)\dot{\gamma}^j(t)\Gamma^k_{ji}|_{\gamma(t)}$$

for $k = 1, \ldots, n$. Because this is a second-order system of differential equations, we expect an existence and a uniqueness result for the initial value problem of specifying value and first derivative, i.e.,

$$\begin{aligned} \gamma(0) &= q, \\ \dot{\gamma}(0) &= \dot{\gamma}^i(0)\,\partial_i|_q. \end{aligned}$$

But because the system is nonlinear, we are not entitled to expect that solutions will exist for all t.

The precise statements obtained from the theory of ordinary differential equations are a bit of a mouthful, but we might as well go for the whole thing right off the bat, since we shall need it all eventually. Still working in our coordinate situation, we get the following facts:

THEOREM 10. (Local Uniqueness) *Let I_1 and I_2 be intervals with $t_0 \in I_1 \cap I_2$, if $\gamma_1 : I_1 \to U$ and $\gamma_2 : I_2 \to U$ are geodesics with $\gamma_1(t_0) = \gamma_2(t_0)$ and $\dot\gamma_1(t_0) = \dot\gamma_2(t_0)$, then $\gamma_1|_{I_1\cap I_2} = \gamma_2|_{I_1\cap I_2}$.*

THEOREM 11. (Existence) *For each $p \in U$ and $v \in \mathbb{R}^n$, there is a neighborhood V_1 of p, a neighborhood V_2 of v, and an $\varepsilon > 0$ such that for each $q \in V_1$ and $w \in V_2$, there is a geodesic $\gamma_{q,w} : (-\varepsilon, \varepsilon) \to U$ with*

$$\begin{aligned} \gamma(0) &= q, \\ \dot\gamma(0) &= w^i \partial_i|_q. \end{aligned}$$

Moreover, the mapping

$$(q, w, t) \to \gamma_{q,w}(t)$$

is C^∞ on $V_1 \times V_2 \times (-\varepsilon, \varepsilon)$.

It is worthwhile to consider what these assertions become in informal terms. The existence statement includes not only "small-time" existence of a geodesic with given initial point and initial tangent, it also asserts a kind of local uniformity for the interval of existence. If you vary the initial conditions but don't vary them too much, then there is a fixed interval $(-\varepsilon, \varepsilon)$ on which all the geodesics with the various initial conditions are defined. Some or all may be defined on larger intervals, but all are defined at least on $(-\varepsilon, \varepsilon)$.

The uniqueness assertion amounts to saying that geodesics cannot be tangent at one point without coinciding. Just as two straight lines that intersect and have the same tangent (at the point of intersection) must coincide, so two geodesics with a common point and equal tangent at that point must coincide.

Both of the differential equations statements are for geodesics with image in a fixed coordinate chart. By relatively simple covering arguments these statements can be extended to geodesics not necessarily contained in a coordinate chart. Let us begin with the uniqueness question:

LEMMA 7. (Global Uniqueness) *Let I_1 and I_2 be open intervals with $t_0 \in I_1 \cap I_2$, if $\gamma_1 : I_1 \to M$ and $\gamma_2 : I_2 \to M$ are geodesics with $\gamma_1(t_0) = \gamma_2(t_0)$ and $\dot\gamma_1(t_0) = \dot\gamma_2(t_0)$, then $\gamma_1|_{(I_1\cap I_2)} = \gamma_2|_{(I_1\cap I_2)}$.*

PROOF. Define

$$A = \{t \in I_1 \cap I_2 : \gamma_1(t) = \gamma_2(t),\, \dot\gamma_1(t) = \dot\gamma_2(t)\}.$$

Then $t_0 \in A$. Also, A is closed in $I_1 \cap I_2$ by continuity of γ_1, γ_2, $\dot\gamma_1$, and $\dot\gamma_2$. Finally, A is open, by virtue of the local uniqueness statement for geodesics in coordinate charts: if $t_1 \in A$, then choose a coordinate chart U around $\gamma_1(t_1) = \gamma_2(t_1)$. Then $(t_1 - \varepsilon, t_1 + \varepsilon) \subset I_1 \cap I_2$ and $\gamma_i|_{(t_1-\varepsilon,t_1+\varepsilon)}$ both have images contained in U. The coordinate uniqueness result then shows that $\gamma_1|_{(t_1-\varepsilon,t_1+\varepsilon)} = \gamma_2|_{(t_1-\varepsilon,t_1+\varepsilon)}$, so that $(t_1 - \varepsilon, t_1 + \varepsilon) \subset A$. □

The coordinate-free global existence picture is a little more subtle. The first, and easy, step is to notice that if we start with a geodesic, then we can enlarge its interval of definition to be maximal. This follows from the uniqueness assertions: If we look at all geodesics $\gamma : I \to M$, $0 \in I$, $\gamma(0) = p$, $\dot\gamma(0) = v$, p and v fixed, then the union of all their domains of definition is a connected open subset of $\mathbb{R}$ on which such a geodesic is defined. And clearly its domain of definition is maximal.

The next observation, also straightforward, is that if $\widehat{K}$ is a compact subset of TM, then there is an $\varepsilon > 0$ such that for each $(q, v) \in \widehat{K}$, there is a geodesic $\gamma : (-\varepsilon, \varepsilon) \to M$ with $\gamma(0) = q$ and $\dot{\gamma}(0) = v$. This is an immediate application of the local uniformity part of the differential equations existence statement together with the usual covering-of-compact-set argument.

The next point to ponder is what happens when the maximal domain of definition is not all of $\mathbb{R}$. For this, let I be a connected open subset of $\mathbb{R}$ that is bounded above, i.e., I has the form $(-\infty, b)$, $b \in \mathbb{R}$ or (a, b), $a, b \in \mathbb{R}$. Suppose $\gamma : I \to M$ is a maximal geodesic. Then $\gamma(t)$ must have a specific kind of behavior as t approaches b : If K is any compact subset of M, then there is a number $t_K < b$ such that if $t_K < t < b$, then $\gamma(t) \in M - K$. We say that γ *leaves every compact set* as $t \to b$.

To see why γ must leave every compact set, suppose K is a compact set it doesn't leave, i.e., suppose there is a sequence $t_1, t_2, \ldots \in I$ with $\lim t_j = b$ and $\gamma(t_j) \in K$ for each j. Now $|\dot{\gamma}(t_j)|$ is independent of j, since geodesics have constant speed. So $\{\dot{\gamma}(t_j) : j = 1, \ldots\}$ lies in a compact subset of TM, namely,

$$\widehat{K} = \{v_q : q \in K, v \in T_qM, |v| \leq |\dot{\gamma}|\}.$$

Thus there is an $\varepsilon > 0$ such that for each $v_q \in \widehat{K}$, there is a geodesic $\gamma : (-\varepsilon, \varepsilon) \to M$ with $\gamma(0) = q$, $\dot{\gamma}(0) = v$. Now choose t_j such that $b - t_j < \varepsilon/2$. Then $\gamma_{q,v}$ patches together with γ to extend γ: beginning at t_j we can continue γ by ε, which takes us beyond b, since t_j is within $\varepsilon/2$ of b. This contradicts the maximality of I.

One important consequence of these observations is what happens when M itself is compact:

LEMMA 8. *If M is a compact Riemannian manifold, then for each $p \in M$ and $v \in T_pM$, there is a geodesic $\gamma : \mathbb{R} \to M$ with $\gamma(0) = p$, $\dot{\gamma}(0) = v$. In other words, geodesics exist for all time.*

A Riemannian manifold where all geodesics exist for all time is called *geodesically complete*.

A slightly trickier point is the following: Suppose $\gamma : I \to M$ is a geodesic and $0 \in I$, where I is a bounded connected open subset of $\mathbb{R}$. Then we would like to say that for $q \in M$ near enough to $\gamma(0)$ and $v \in T_qM$ near enough to $\dot{\gamma}(0)$ there is a geodesic $\gamma_{q,v}$ with q, v as initial position and tangent, respectively, *and* with $\gamma_{q,v}$ defined on an interval almost as big as I. This is true, and it is worth putting in formal language:

LEMMA 9. *Suppose $\gamma : [a, b] \to M$ is a geodesic on a compact interval. Then there is a neighborhood U in TM of $\dot{\gamma}(0)$ such that if $v \in U$, then there is a geodesic*

$$\gamma_v : [a, b] \to M$$

with $\dot{\gamma}_v(0) = v$.

PROOF. Subdivide the interval $a = b_0 < b_1 < \cdots < b_k = b$ in such a way that we have neighborhoods V_i of $\dot{\gamma}(b_i)$ where any geodesic $\gamma : [b_i, b_i + \varepsilon) \to M$ with $\dot{\gamma}(b_i) \in V_i$ is defined on $[b_i, b_{i+1}]$. Using that the map $(t, v) \to \gamma_v(t)$ is continuous, where γ is the geodesic with $\dot{\gamma}(0) = v$ we can select a new neighborhood $U_0 \subset V_0$ of $\dot{\gamma}(b_0)$ such that $\dot{\gamma}_v(b_1) \in V_1$ for $v \in U_0$. Next select $U_1 \subset U_0$ so that $\dot{\gamma}_v(b_2) \in V_2$ for $v \in U_1$ etc. In this way we get the desired neighborhood $U = U_{k-1}$ in at most k steps. □

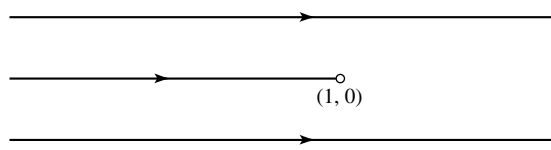

Figure 5.1

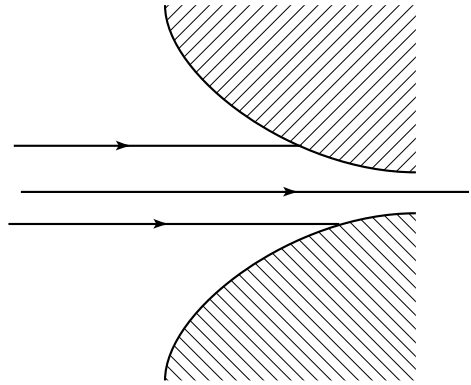

Figure 5.2

All this seems a bit formal, pedantic and perhaps abstract as well, in the absence of explicitly computed examples. First, one can easily check that geodesics in Euclidean space are straight lines. Using this observation it is simple to give examples of the above ideas by taking M to be open subsets of $\mathbb{R}^2$ with its usual metric.

EXAMPLE 28. *In the plane $\mathbb{R}^2$ minus one point, say $\mathbb{R}^2 - \{(1,0)\}$ the unit speed geodesic from $(0,0)$ with tangent $(1,0)$ is defined on $(-\infty, 1)$ only. But nearby geodesics from $(0,0)$ with tangents $(1+\varepsilon_1, \varepsilon_2)$, $\varepsilon_1, \varepsilon_2$ small, $\varepsilon_2 \neq 0$, are defined on $(-\infty, \infty)$. Thus maximal intervals of definition can jump up in size, but, as already noted, not down. See also Figure 5.1.*

EXAMPLE 29. *On the other hand, for the region*

$$\{(x,y) : |xy| < 1\},$$

the curve $t \to (t,0)$ is a geodesic defined on all of $\mathbb{R}$ that is a limit of unit speed geodesics $t \to (t,\varepsilon)$, $\varepsilon \to 0$, each of which is defined only on a finite interval $\left(-\frac{1}{\varepsilon}, \frac{1}{\varepsilon}\right)$. Note that as required, the endpoints of these intervals go to infinity (in both directions). See also Figure 5.2.

The reader should think through these examples and those in the exercises very carefully, since geodesic behavior is a fundamental topic in all that follows.

EXAMPLE 30. *We think of the spheres $(S^n(r), \mathrm{can}) = S^n_{r^{-2}}$ as being in $\mathbb{R}^{n+1}$. The acceleration of a curve $\gamma : I \to S^n(r)$ can be computed as the Euclidean acceleration projected onto $S^n(r)$. Thus γ is a geodesic iff $\ddot{\gamma}$ is normal to $S^n(r)$. This means that $\ddot{\gamma}$ and γ should be proportional as vectors in $\mathbb{R}^{n+1}$. Great circles $\gamma(t) = a\cos(\alpha t) + b\sin(\alpha t)$, where $a, b \in \mathbb{R}^{n+1}$, $|a| = |b| = r$, and $a \perp b$, clearly*

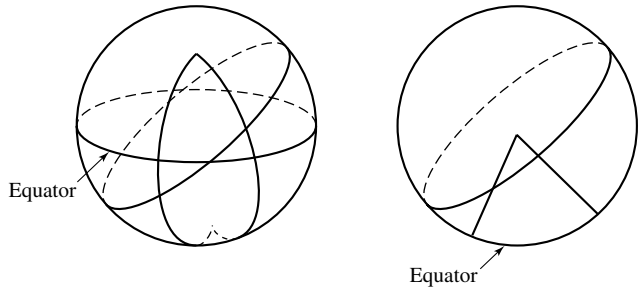

Figure 5.3

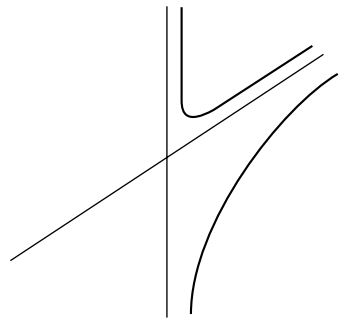

Figure 5.4

have this property. Furthermore, since $\gamma(0) = a \in S^n(r)$ *and* $\dot\gamma(0) = \alpha b \in T_a S^n(r)$, *we see that we have a geodesic for each initial value problem.*

We can easily picture great circles on spheres as depicted in Figure 5.3. Still, it is convenient to have a different way of understanding this. For this we project the sphere orthogonally onto the plane containing the equator. Thus the north and south poles are mapped to the origin. As all geodesics are great circles, they must project down to ellipses that have the origin as center and whose greater axis has length r. Of course, this simply describes exactly the way in which we draw three-dimensional pictures on paper.

EXAMPLE 31. *We think of* $S^n_{-r^{-2}}$ *as a hypersurface in Minkowski space* $\mathbb{R}^{1,n}$. *In this case the acceleration is still the projection of the acceleration in Minkowski space. In Minkowski space the acceleration in the usual coordinates is the same as the Euclidean acceleration. Thus we just have to find the Minkowski projection onto the hypersurface. By analogy with the sphere, one might guess that the hyperbolae* $\gamma(t) = a\cosh(\alpha t) + b\sinh(\alpha t),\ a, b \in \mathbb{R}^{1,n},\ |a|^2 = -r^2,\ |b|^2 = r^2$, *and* $a \perp b$ *all in the Minkowski sense, are our geodesics. And indeed this is true.*

This time the geodesics are hyperbolae. Drawing several of them on the space itself as seen in Minkowski space is not so easy. However, as with the sphere we can resort to the trick of projecting hyperbolic space onto the plane containing the last n coordinates. The geodesics there can then be seen to be hyperbolae whose asymptotes are straight lines through the origin. See also Figure 5.4.

EXAMPLE 32. *On a Lie group G with a left-invariant metric one might suspect that the geodesics are the integral curves for the left-invariant vector fields. This in turn is equivalent to the assertion that $\nabla_X X \equiv 0$ for all left-invariant vector fields. But our Lie group model for the upper half plane does not satisfy this. However, we did show in chapter 3 that $\nabla_X X = \frac{1}{2}[X, X] = 0$ when the metric is bi-invariant and X is left-invariant. Moreover, all compact Lie groups admit bi-invariant metrics (see exercises to chapter 1).*

3. The Metric Structure of a Riemannian Manifold

The positive definite inner product structures on the tangent space of a Riemannian manifold automatically give rise to a concept of lengths of tangent vectors. From this one can obtain an idea of the length of a curve as the integral of the length of its velocity vector field. This is a direct extension of the usual calculus concept of the length of curves in Euclidean space. Indeed, the definition of Riemannian manifolds is motivated from the beginning by lengths of curves. The situation is turned around a bit from that of $\mathbb{R}^n$, though: On Euclidean spaces, we have in advance a concept of distance between points. Thus, the definition of lengths of curves is justified by the fact that the length of a curve should be approximated by sums of distances for a fine subdivision (e.g., a fine polygonal approximation). For Riemannian manifolds, there is no immediate idea of distance between points. Instead, we have a natural idea of (tangent) vector length, hence curve length, and we shall use the length-of-curve idea to define distance between points. The goal of this section is to carry out these constructions in detail.

First, recall that a mapping $\gamma : [a, b] \to M$ is a piecewise C^∞ curve if γ is continuous and if there is a partition $a = a_1 < a_2 < \ldots < a_k = b$ of $[a, b]$ such that $\gamma|_{[a_i, a_{i+1}]}$ is C^∞ for $i = 1, \ldots, k-1$. Occasionally it will be convenient to work with curves that are merely absolutely continuous. A curve $\gamma : [a, b] \to \mathbb{R}^n$ is absolutely continuous if the derivative exists almost everywhere and $\gamma(t) = \gamma(a) + \int_a^t \dot{\gamma}(s)\, ds$. If $F : \mathbb{R}^n \to \mathbb{R}^n$ is a diffeomorphism, then we see that also $F \circ \gamma$ is absolutely continuous. Thus it makes sense to work with absolutely continuous curves in smooth manifolds.

Let $\gamma : [a, b] \to M$ be a piecewise C^∞ (or merely absolutely continuous) curve in a Riemannian manifold. Then the *length* $\ell(\gamma)$ is defined as follows:

$$\ell(\gamma) = \int_a^b |\dot{\gamma}(t)|\, dt = \int_a^b \sqrt{g(\dot{\gamma}(t), \dot{\gamma}(t))} dt.$$

It is clear from the definition that the function $t \to |\dot{\gamma}(t)|$ is integrable in the Riemann (or Lebesgue) integral sense, so $\ell(\gamma)$ is a well-defined finite, nonnegative number. The chain and substitution rules show that $\ell(\gamma)$ is invariant under reparametrization. A curve $\gamma : [a, b] \to M$ is said to be parametrized by arc length if $\ell(\gamma|_{[a,t]}) = t - a$ for all $t \in [a, b]$, or equivalently, if $|\dot{\gamma}(t)| = 1$ at all smooth points $t \in [a, b]$. A curve $\gamma : [a, b] \to M$ such $|\dot{\gamma}(t)| > 0$ wherever it is smooth can be reparametrized by arc length without changing the length of the curve. To see this consider

$$s = \varphi(t) = \int_a^t |\dot{\gamma}(\tau)|\, d\tau.$$

Thus φ is strictly increasing on $[a, b]$, and the curve $\gamma \circ \varphi^{-1} : [0, \ell(\gamma)] \to M$ has tangent vectors of unit length at all points where it is smooth. A slightly stickier,

and often ignored, point is what happens to curves that have stationary points. We can still construct the integral:

$$s = \varphi(t) = \int_a^t |\dot{\gamma}(\tau)|\, d\tau,$$

but we can't find a smooth inverse to φ if $\dot{\gamma}$ is zero somewhere. We can, however, find a curve $\sigma : [0, \ell(\gamma)] \to M$ such that

$$\gamma(t) = (\sigma \circ \varphi)(t) = \sigma(s).$$

To ensure that σ is well-defined we just have to check that $\gamma(t_1) = \gamma(t_2)$ if $\varphi(t_1) = \varphi(t_2)$. The latter equality, however, implies that $|\dot{\gamma}| = 0$ (almost everywhere) on $[t_1, t_2]$ so it does follow that $\gamma(t_1) = \gamma(t_2)$. We now need to check that σ has unit speed. This is straightforward at points where $\dot{\gamma} \neq 0$, but at the stationary points for γ it is not even clear that σ is differentiable. In fact it need not be if γ has a cusp-like singularity. The set of trouble points is the set of critical values for φ so it is at least a set of measure zero (this is simply Sard's theorem for functions $\mathbb{R} \to \mathbb{R}$). This shows that we can still define the length of σ as

$$\ell(\sigma) = \int_0^{\ell(\gamma)} |\dot{\sigma}|\, ds$$

and that σ is parametrized by arclength. In this way we have constructed a generalized reparametrization of γ, that is parametrized by arclength. Note that even if we start with a smooth curve γ the reparametrized curve σ might just be absolutely continuous. It is therefore quite natural to work with the larger class of absolutely continuous curves. Nevertheless, we have chosen to mostly stay with the more mundane piecewise smooth curves as they suffice for developing the theory Riemannian manifolds.

We are now ready to introduce the idea of distance between points. First, for each pair of points $p, q \in M$ we define the path space

$$\Omega_{p,q} = \{\gamma : [0,1] \to M : \gamma \text{ is piecewise } C^\infty \text{ and } \gamma(0) = p, \gamma(1) = q\}.$$

We can then define the distance $d(p,q)$ between points $p, q \in M$ as

$$d(p,q) = \inf\{\ell(\gamma) : \gamma \in \Omega_{p,q}\}.$$

It follows immediately from this condition that $d(p,q) = d(q,p)$ and $d(p,q) \leq d(p,r) + d(r,q)$. The fact that $d(p,q) = 0$ only when $p = q$ will be established below. Thus, $d(\ ,\ )$ satisfies all the properties of a metric.

As for metric spaces, we have various metric balls defined via the metric

$$\begin{aligned} B(p,r) &= \{x \in M : d(p,x) < r\}, \\ \bar{B}(p,r) &= D(p,r) = \{x \in M : d(p,x) \leq r\}. \end{aligned}$$

More generally, we can define the distance between subsets $A, B \subset M$ as

$$d(A,B) = \inf\{d(p,q) : p \in A, q \in B\}.$$

With this we then have

$$\begin{aligned} B(A,r) &= \{x \in M : d(A,x) < r\}, \\ \bar{B}(A,r) &= D(A,r) = \{x \in M : d(A,x) \leq r\}. \end{aligned}$$

The infimum of curve lengths in the definition of $d(p,q)$ can fail to be realized. This is illustrated, for instance, by the "punctured plane" $\mathbb{R}^2 - \{(0,0)\}$

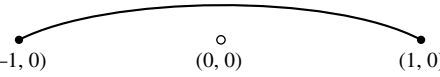

Figure 5.5

with the usual Riemannian metric of $\mathbb{R}^2$ restricted to $\mathbb{R}^2 - \{(0,0)\}$. The distance $d((-1,0),(1,0)) = 2$, but this distance is not realized by any curve, since every curve of length 2 in $\mathbb{R}^2$ from $(-1,0)$ to $(1,0)$ passes through $(0,0)$ (see Figure 5.5). In a sense that we shall explore later, $\mathbb{R}^2 - \{(0,0)\}$ is incomplete. For the moment, we introduce some terminology for the cases where the infimum $d(p,q)$ is realized.

A curve $\sigma \in \Omega(p,q)$ is a *segment* if $\ell(\sigma) = d(p,q)$ and σ is parametrized proportionally to arc length, i.e., $|\dot{\sigma}|$ is constant on the set where σ is smooth.

EXAMPLE 33. *In Euclidean space $\mathbb{R}^n$, segments according to this definition are straight line segments parametrized with constant speed, i.e. curves of the form $t \to p + t \cdot v$. In $\mathbb{R}^n$, each pair of points p, q is joined by a segment $t \to p + t(q - p)$ that is unique up to reparametrization.*

EXAMPLE 34. *In $S^2(1)$ segments are portions of great circles with length $\leq \pi$. (We assume for the moment some basic observations of spherical geometry: these will arise later as special cases of more general results.) Every two points are joined by a segment, but there may be more than one segment joining a given pair if the pair are far enough apart, i.e., each pair of antipodal points is joined by infinitely many distinct segments.*

EXAMPLE 35. *In $\mathbb{R}^2 - \{(0,0)\}$, as already noted, not every pair of points is joined by a segment.*

Later we shall show that segments are always geodesics. Conversely, geodesics are segments if they are short enough; precisely, if γ is a geodesic defined on an open interval containing 0, then $\gamma|_{[0,\varepsilon]}$ is a segment for all sufficiently small $\varepsilon > 0$. Furthermore, we shall show that each pair of points in a Riemannian manifold can be joined by at least one segment provided that the Riemannian manifold is complete as a metric space in the metric just defined. This result explains what is "wrong" with the punctured plane. It also explains why spheres have to have segments between each pair of points: compact spaces are always complete in any metric compatible with the (compact) topology.

Some work needs to be done before we can prove these general statements. To start with, let us dispose of the question of compatibility of topologies.

THEOREM 12. *The metric topology obtained from the distance $d(\ ,\ )$ on a Riemannian manifold is the same as the manifold topology.*

PROOF. Fix $p \in M$ and a coordinate neighborhood U of p such that $x^i(p) = 0$. We assume in addition that $g_{ij}|_p = \delta_{ij}$. On U we have the given Riemannian metric g and also the Euclidean metric g_0 defined by

$$g_0(\partial_i, \partial_j) = \delta_{ij}.$$

Thus g_0 is constant and equal to g at p. Finally we can after possibly shrinking U also assume that

$$\begin{aligned} U &= B^{g_0}(p,\varepsilon) \\ &= \{x \in U : d_{g_0}(p,x) < \varepsilon\} \\ &= \left\{x \in U : \sqrt{(x^1)^2 + \cdots + (x^n)^2} < \varepsilon\right\}. \end{aligned}$$

Thus the Euclidean distance is

$$d_{g_0}(p,x) = \sqrt{(x^1)^2 + \cdots + (x^n)^2}.$$

For $x \in U$ we can compare these two metrics as follows: There are continuous functions: $\lambda, \mu : U \to (0,\infty)$ such that if $v \in T_x M$, then

$$\lambda(x)\,|v|_{g_0} \leq |v|_g \leq \mu(x)\,|v|_{g_0}.$$

Moreover, $\lambda(x), \mu(x) \to 1$ as $x \to p$.

Now let $c : [0,1] \to M$ be a curve from p to $x \in U$.

1: If c is a straight line in the Euclidean metric, then it lies in U and

$$\begin{aligned} d_{g_0}(p,x) &= \ell_{g_0}(c) \\ &= \int_0^1 |\dot{c}|_{g_0}\,dt \\ &\geq \frac{1}{\max \mu(c(t))} \int_0^1 |\dot{c}|_g\,dt \\ &= \frac{1}{\max \mu(c(t))} \ell_g(c) \\ &\geq \frac{1}{\max \mu(c(t))} d_g(p,x). \end{aligned}$$

2: If c lies entirely in U then

$$\begin{aligned} \ell_g(c) &= \int_0^1 |\dot{c}|_g\,dt \\ &\geq (\min \lambda(c(t))) \int_0^1 |\dot{c}|_{g_0}\,dt \\ &\geq (\min \lambda(c(t)))\, d_{g_0}(p,x). \end{aligned}$$

3: If c leaves U, then there will be a smallest t_0 such that $c(t_0) \notin U$, then

$$\begin{aligned} \ell_g(c) &\geq \int_0^{t_0} |\dot{c}|_g\,dt \\ &\geq (\min \lambda(c(t))) \int_0^{t_0} |\dot{c}|_{g_0}\,dt \\ &\geq (\min \lambda(c(t)))\,\varepsilon \\ &\geq (\min \lambda(c(t)))\, d_{g_0}(p,x). \end{aligned}$$

By possibly shrinking U again we can now guarantee that $\min \lambda \geq \lambda_0 > 0$ and $\max \mu \leq \mu_0 < \infty$. We have then proven that

$$d_g(p,x) \leq \mu_0 d_{g_0}(p,x)$$

and

$$\begin{aligned}\lambda_0 d_{g_0}(p,x) &\leq \inf \ell_g(c)\\ &= d_g(p,x).\end{aligned}$$

Thus the Euclidean and Riemannian distances are comparable on a neighborhood of p. This shows that the metric topology and the manifold topology (coming from the Euclidean distance) are equivalent. It also shows that $p = q$ if $d(p,q) = 0$.

Finally note that

$$\lim_{x\to p} \frac{d_g(p,x)}{d_{g_0}(p,x)} = 1$$

since $\lambda(x), \mu(x) \to 1$ as $x \to p$. □

Just as compact Riemannian manifolds are automatically geodesically complete, this theorem also shows that such spaces are metrically complete.

COROLLARY 2. *If M is a compact manifold and g is a Riemannian metric on M, then (M, d_g) is a complete metric space, where d_g is the Riemannian distance function determined by g.*

Let us relate these new concepts to our distance functions from chapter 2.

LEMMA 10. *Suppose $r : U \to \mathbb{R}$ is a smooth distance function and $U \subset (M,g)$ is open, then the integral curves for ∇r are segments in (U,g).*

PROOF. Fix $p, q \in U$ and let $\gamma(t) : [0,b] \to U$ be a curve from p to q. Then

$$\begin{aligned}\ell(\gamma) &= \int_0^b |\dot\gamma| dt\\ &= \int_0^b |\nabla r| \cdot |\dot\gamma| dt\\ &\geq \int_0^b |g(\nabla r, \dot\gamma)|\, dt\\ &\geq \left|\int_0^b d(r\circ\gamma)\, dt\right|\\ &= |r(q) - r(p)|.\end{aligned}$$

Here the first inequality is the Cauchy-Schwarz inequality. This shows that

$$d(p,q) \geq |r(q) - r(p)|.$$

If we choose γ as an integral curve for ∇r, i.e., $\dot\gamma = \nabla r \circ \gamma$, then equality holds in the Cauchy-Schwarz inequality and $d(r\circ\gamma) > 0$. Thus

$$\ell(\gamma) = |r(q) - r(p)|.$$

This shows that integral curves must be segments. Notice that we only considered curves in U, and therefore only established the result for (U,g) and not (M,g). □

EXAMPLE 36. *Let $M = S^1 \times \mathbb{R}$ and $U = \left(S^1 - \{e^{i0}\}\right) \times \mathbb{R}$. On U we have the distance function $u(\theta, x) = \theta$, $\theta \in (0, 2\pi)$. The previous lemma shows that any curve $\gamma(t) = (e^{it}, r_0)$, $t \in I$, where I does not contain 0 is a segment in U. If, however, the length of I is $> \pi$, then such curves can clearly not be segments in M.*

The *functional distance* d_F between points in a manifold is defined as

$$d_F(p,q) = \sup\{|f(p) - f(q)| : f : M \to \mathbb{R} \text{ has } |\nabla f| \leq 1 \text{ on } M\}.$$

This distance is always smaller than the arclength distance. One can, however, show as before that it generates the standard manifold topology. In fact, after we have established the existence of smooth distance functions, it will become clear that the two distances are equal provided p and q are sufficiently close to each other.

4. First Variation of Energy

In this section we shall study the arclength functional

$$\begin{aligned} \ell(\gamma) &= \int_0^1 |\dot{\gamma}|\, dt, \\ \gamma &\in \Omega_{p,q} \end{aligned}$$

in further detail. The minima, if they exist, are pre-segments. That is, they have minimal length but we are not guaranteed that they have the correct parametrization. We also saw that in some cases suitable geodesics minimize this functional. One problem with this functional is that it is invariant under change of parametrization. Minima, if they exist, therefore do not come with a fixed parameter. This problem can be overcome, at the expense of geometric intuition, by considering the energy functional

$$\begin{aligned} E(\gamma) &= \frac{1}{2}\int_0^1 |\dot{\gamma}|^2\, dt, \\ \gamma &\in \Omega_{p,q}. \end{aligned}$$

This functional measures the total kinetic energy of a particle traveling along γ with the speed dictated by γ. We start by showing that these two functionals have the same minima.

PROPOSITION 17. *If $\sigma \in \Omega_{p,q}$ is a constant speed curve that minimizes $\ell : \Omega_{p,q} \to [0,\infty)$, then σ also minimizes $E : \Omega_{p,q} \to [0,\infty)$. Conversely if $\sigma \in \Omega_{p,q}$ minimizes $E : \Omega_{p,q} \to [0,\infty)$, then σ also minimizes $\ell : \Omega_{p,q} \to [0,\infty)$.*

PROOF. The Cauchy-Schwarz inequality for functions tells us that

$$\begin{aligned} \ell(\gamma) &= \int_0^1 |\dot{\gamma}| \cdot 1 dt \\ &\leq \sqrt{\int_0^1 |\dot{\gamma}|^2\, dt}\sqrt{\int_0^1 1^2 dt} \\ &= \sqrt{\int_0^1 |\dot{\gamma}|^2\, dt} \\ &= \sqrt{2E(\gamma)}, \end{aligned}$$

with equality holding iff $|\dot{\gamma}| = c \cdot 1$ for some constant c, i.e., γ has constant speed. In case γ is only absolutely continuous this inequality still holds. Moreover, when equality holds the speed is constant wherever it is defined. Let $\sigma \in \Omega_{p,q}$ be a

constant speed curve that minimizes ℓ and $\gamma \in \Omega_{p,q}$. Then

$$\begin{aligned} E(\sigma) &= \frac{1}{2}(\ell(\sigma))^2 \\ &\leq \frac{1}{2}(\ell(\gamma))^2 \\ &\leq E(\gamma), \end{aligned}$$

so σ also minimizes E.

Conversely let $\sigma \in \Omega_{p,q}$ minimize E and $\gamma \in \Omega_{p,q}$. If γ does not have constant speed we can without changing its length reparametrize it to an absolutely continuous curve $\bar{\gamma}$ that has constant speed almost everywhere. Then

$$\begin{aligned} \ell(\sigma) &\leq \sqrt{2E(\sigma)} \\ &\leq \sqrt{2E(\bar{\gamma})} \\ &= \ell(\bar{\gamma}) \\ &= \ell(\gamma). \end{aligned}$$

□

Our next goal is to show that minima of E must be geodesics. To do this we have to develop the *first variation formula for energy.* A variation of a curve $\gamma : I \to M$ is a family of curves $\bar{\gamma} : (-\varepsilon, \varepsilon) \times [a,b] \to M$, such that $\bar{\gamma}(0,t) = \gamma(t)$ for all $t \in [a,b]$. We say that such a variation is piecewise smooth if it is continuous and we can partition $[a,b]$ in to intervals $[a_i, a_{i+1}]$, $i = 0, ..., m-1$, in such a way that $\bar{\gamma} : (-\varepsilon, \varepsilon) \times [a_i, a_{i+1}] \to M$ is smooth. Thus the curves $t \to \gamma_s(t) = \bar{\gamma}(s,t)$ are all piecewise smooth, while the curves $s \to \bar{\gamma}(s,t)$ are smooth. The velocity field for this variation is the field $\frac{\partial \bar{\gamma}}{\partial t}$ which is well-defined on each interval $[a_i, a_{i+1}]$. At the break points a_i, there are two possible values for this field; a right derivative and a left derivative:

$$\begin{aligned} \frac{\partial \bar{\gamma}}{\partial t^+}|_{(s,a_i)} &= \frac{\partial \bar{\gamma}|_{[a_i,a_{i+1}]}}{\partial t}|_{(s,a_i)}, \\ \frac{\partial \bar{\gamma}}{\partial t^-}|_{(s,a_i)} &= \frac{\partial \bar{\gamma}|_{[a_{i-1},a_i]}}{\partial t}|_{(s,a_i)}. \end{aligned}$$

The *variational field* is defined as $\frac{\partial \bar{\gamma}}{\partial s}$. This field is well-defined everywhere. It is smooth on each $(-\varepsilon, \varepsilon) \times [a_i, a_{i+1}]$ and continuous on $(-\varepsilon, \varepsilon) \times I$. The special case where $a = 0$, $b = 1$, $\bar{\gamma}(s,0) = p$ and $\bar{\gamma}(s,1) = q$ for all s is of special importance as all of the curves $\gamma_s \in \Omega_{p,q}$. Such variations are called *proper variations* of γ.

LEMMA 11. (The First Variation Formula) *Let $\bar{\gamma} : (-\varepsilon, \varepsilon) \times [a,b] \to M$ be a piecewise smooth variation, then*

$$\begin{aligned} \frac{dE(\gamma_s)}{ds} &= -\int_a^b g\left(\frac{\partial^2 \bar{\gamma}}{\partial t^2}, \frac{\partial \bar{\gamma}}{\partial s}\right) dt + g\left(\frac{\partial \bar{\gamma}}{\partial t^-}, \frac{\partial \bar{\gamma}}{\partial s}\right)\Big|_{(s,b)} - g\left(\frac{\partial \bar{\gamma}}{\partial t^+}, \frac{\partial \bar{\gamma}}{\partial s}\right)\Big|_{(s,a)} \\ &\quad + \sum_{i=1}^{m-1} g\left(\frac{\partial \bar{\gamma}}{\partial t^-} - \frac{\partial \bar{\gamma}}{\partial t^+}, \frac{\partial \bar{\gamma}}{\partial s}\right)\Big|_{(s,a_i)}. \end{aligned}$$

PROOF. It suffices to prove the formula for smooth variations as we can otherwise split up the integral into parts that are smooth:

$$\begin{aligned} E(\gamma_s) &= \int_a^b \left|\frac{\partial\bar{\gamma}}{\partial t}\right|^2 dt \\ &= \sum_{i=0}^{m-1} \int_{a_i}^{a_{i+1}} \left|\frac{\partial\bar{\gamma}}{\partial t}\right|^2 dt \end{aligned}$$

and apply the formula to each part of the variation.

For a smooth variation $\bar{\gamma} : (-\varepsilon, \varepsilon) \times [a, b] \to M$ we have

$$\begin{aligned} \frac{dE(\gamma_s)}{ds} &= \frac{d}{ds}\frac{1}{2}\int_a^b g\left(\frac{\partial\bar{\gamma}}{\partial t}, \frac{\partial\bar{\gamma}}{\partial t}\right) dt \\ &= \frac{1}{2}\int_a^b \frac{\partial}{\partial s} g\left(\frac{\partial\bar{\gamma}}{\partial t}, \frac{\partial\bar{\gamma}}{\partial t}\right) dt \\ &= \int_a^b g\left(\frac{\partial^2\bar{\gamma}}{\partial s\partial t}, \frac{\partial\bar{\gamma}}{\partial t}\right) dt \\ &= \int_a^b g\left(\frac{\partial^2\bar{\gamma}}{\partial t\partial s}, \frac{\partial\bar{\gamma}}{\partial t}\right) dt \\ &= \int_a^b \frac{\partial}{\partial t} g\left(\frac{\partial\bar{\gamma}}{\partial s}, \frac{\partial\bar{\gamma}}{\partial t}\right) dt - \int_a^b g\left(\frac{\partial\bar{\gamma}}{\partial s}, \frac{\partial^2\bar{\gamma}}{\partial t^2}\right) dt \\ &= g\left(\frac{\partial\bar{\gamma}}{\partial s}, \frac{\partial\bar{\gamma}}{\partial t}\right)\Big|_a^b - \int_a^b g\left(\frac{\partial\bar{\gamma}}{\partial s}, \frac{\partial^2\bar{\gamma}}{\partial t^2}\right) dt \\ &= -\int_a^b g\left(\frac{\partial\bar{\gamma}}{\partial s}, \frac{\partial^2\bar{\gamma}}{\partial t^2}\right) dt + g\left(\frac{\partial\bar{\gamma}}{\partial s}, \frac{\partial\bar{\gamma}}{\partial t}\right)\Big|_{(s,b)} - g\left(\frac{\partial\bar{\gamma}}{\partial s}, \frac{\partial\bar{\gamma}}{\partial t}\right)\Big|_{(s,a)}. \end{aligned}$$

□

We can now completely characterize the local minima for the energy functional.

THEOREM 13. (Characterization of local minima) *If $\gamma \in \Omega_{p,q}$ is a local minimum for $E : \Omega_{p,q} \to [0, \infty)$, then γ is a smooth geodesic.*

PROOF. The assumption guarantees that

$$\frac{dE(\gamma_s)}{ds} = 0$$

for any proper variation of γ. The trick now is to find appropriate variations. In fact, if $V(t)$ is any vector field along $\gamma(t)$, i.e., $V(t) \in T_{\gamma(t)}M$, then there is a variation so that $V(t) = \frac{\partial\gamma}{\partial s}|_{(0,t)}$. One such variation is gotten by declaring that the variational curves $s \to \gamma(s, t)$ are geodesics with $\frac{\partial\gamma}{\partial s}|_{(0,t)} = V(t)$. As geodesics vary nicely with respect to the initial data this variation will be as smooth as V is. Finally, if $V(a) = 0$ and $V(b) = 0$, then the variation is proper.

Using such a variational field the first variation formula at $s = 0$ only depends on γ itself and the variational field V

$$\begin{aligned}\frac{dE(\gamma_s)}{ds}|_{s=0} &= -\int_a^b g(\ddot{\gamma}, V)\,dt + g\left(\frac{d\gamma}{dt^-}(b), V(b)\right) - g\left(\frac{d\gamma}{dt^+}(a), V(a)\right) \\ &\quad + \sum_{i=1}^{m-1} g\left(\frac{d\gamma}{dt^-}(a_i) - \frac{d\gamma}{dt^+}(a_i), V(a_i)\right) \\ &= -\int_a^b g(\ddot{\gamma}, V)\,dt + \sum_{i=1}^{m-1} g\left(\frac{d\gamma}{dt^-}(a_i) - \frac{d\gamma}{dt^+}(a_i), V(a_i)\right).\end{aligned}$$

We now specify V further. First select $V(t) = \lambda(t)\ddot{\gamma}(t)$, where $\lambda(a_i) = 0$ at the break points a_i where γ might not be smooth, and also $\lambda(a) = \lambda(b) = 0$. Finally assume that $\lambda(t) > 0$ elsewhere. Then

$$\begin{aligned}0 &= \frac{dE(\gamma_s)}{ds}|_{s=0} \\ &= -\int_a^b g(\ddot{\gamma}, \lambda(t)\ddot{\gamma})\,dt \\ &= -\int_a^b \lambda(t)|\ddot{\gamma}|^2\,dt.\end{aligned}$$

Since $\lambda(t) > 0$ where $\ddot{\gamma}$ is defined it must follow that $\ddot{\gamma} = 0$ at those points. Thus γ is a broken geodesic. Next select a general variational field V such that

$$\begin{aligned}V(a_i) &= \frac{d\gamma}{dt^-}(a_i) - \frac{d\gamma}{dt^+}(a_i), \\ V(a) &= V(b) = 0\end{aligned}$$

and otherwise arbitrary, then we obtain

$$\begin{aligned}0 &= \frac{dE(\gamma_s)}{ds}|_{s=0} \\ &= \sum_{i=1}^{m-1} g\left(\frac{d\gamma}{dt^-}(a_i) - \frac{d\gamma}{dt^+}(a_i), V(a_i)\right) \\ &= \sum_{i=1}^{m-1} \left|\frac{d\gamma}{dt^-}(a_i) - \frac{d\gamma}{dt^+}(a_i)\right|^2.\end{aligned}$$

This forces

$$\frac{d\gamma}{dt^-}(a_i) = \frac{d\gamma}{dt^+}(a_i)$$

and hence the broken geodesic has the same velocity from the left and right at the places where it is potentially broken. Uniqueness of geodesics then shows that γ is a smooth geodesic. □

This also shows:

COROLLARY 3. (Characterization of segments) *Any piecewise smooth segment is a geodesic.*

While this result shows precisely what the local minima of the energy functional must be it does not guarantee that geodesics are local minima. In Euclidean space all geodesics are minimal as they are the integral curves for globally defined distance

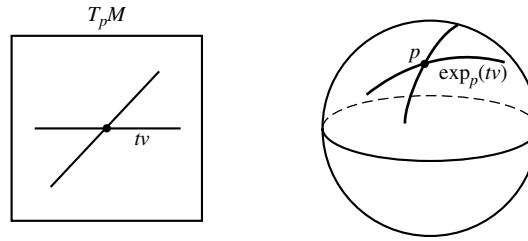

Figure 5.6

functions: $u(x) = v \cdot x$, where v is a unit vector. On the unit sphere, however, no geodesic of length $> \pi$ can be locally minimizing. Since such geodesics always form part of a great circle, we see that the complement of the geodesic in the great circle has length $< \pi$. This shows that the geodesic can't be an absolute minimum. However, we can also easily construct a variation where the nearby curves are all shorter. We shall spend much more time on these issues in the subsequent sections as well as the next chapter. Certainly much more work has to be done before we can say more about when geodesics are minimal. The above proof does, however, tell us that a geodesic $\gamma \in \Omega_{p,q}$ is always a *stationary point* for $E : \Omega_{p,q} \to [0, \infty)$, in the sense that

$$\frac{dE(\gamma_s)}{ds}|_{s=0} = 0$$

for all proper variations of γ.

5. The Exponential Map

For a tangent vector $v \in T_pM$, let γ_v be the unique geodesic with $\gamma(0) = p$ and $\dot{\gamma}(0) = v$, and $[0, \ell_v)$ the nonnegative part of the maximal interval on which γ is defined. Notice that $\gamma_{\alpha v}(t) = \gamma_v(\alpha t)$ for all $\alpha > 0$ and $t < \ell_{\alpha v}$. In particular, $\ell_{\alpha v} = \alpha^{-1}\ell_v$. Let $O_p \subset T_pM$ be the set of vectors v such that $1 < \ell_v$, so that $\gamma_v(t)$ is defined on $[0, 1]$. Then define the *exponential map* at p by

$$\begin{aligned} \exp_p &: O_p \to M \\ \exp_p(v) &= \gamma_v(1). \end{aligned}$$

In the exercises to this chapter we have a problem that elucidates the relationship between the just defined exponential map and the Lie group exponential map introduced earlier. In Figure 5.6 we have shown how radial lines in the tangent space are mapped to radial geodesics in M via the exponential map. The "homogeneity property" $\gamma_v(t) = \gamma_{tv}(1)$ shows that $\exp_p(tv) = \gamma_v(t)$. Given that, it is natural to think of $\exp_p(v)$ in a polar coordinate representation: From p ones goes "distance" $|v|$ in the direction $v/|v|$. This gives the point $\exp_p(v)$, since $\gamma_{v/|v|}(|v|) = \gamma_v(1)$.

The individual $\exp_p$ maps can be combined to form a map $\exp : \bigcup O_p \to M$ by setting $\exp|_{O_p} = \exp_p$. This map exp is also called the *exponential map.*

The standard theory of ordinary differential equations that we have already discussed tells us that the set $O = \bigcup O_p$ is open in TM and that $\exp : O \to M$ is smooth. In addition $O_p \subset T_pM$ is open, and $\exp_p : O_p \to M$ is also smooth. It is an important property that $\exp_p$ is in fact a local diffeomorphism around $0 \in T_pM$. The details of this are given in the following:

PROPOSITION 18. *If $p \in M$, then*
(1)

$$D\exp_p : T_0(T_pM) \to T_pM$$

is nonsingular at the origin of T_pM. Consequently $\exp_p$ is a local diffeomorphism.

(2) Define $E : O \to M \times M$ by $E(v) = (\pi(v), \exp v)$, where $\pi(v)$ is the base point of v, i.e., $v \in T_{\pi(v)}M$. Then for each $p \in M$ and with it the zero vector, $0_p \in T_pM$,

$$DE : T_{(p,0_p)}(TM) \to T_{(p,p)}(M \times M)$$

is nonsingular. Consequently, E is a diffeomorphism from a neighborhood of the zero section of TM onto an open neighborhood of the diagonal in $M \times M$.

PROOF. The proofs of both statements are an immediate application of the inverse function theorem, once a crucial observation has been made. This observation is as follows: Let $I_0 : T_pM \to T_0T_pM$ be the canonical isomorphism, i.e., $I_0(v) = \frac{d}{dt}(tv)|_{t=0}$. Now we recall that if $v \in O_p$, then $\gamma_v(t) = \gamma_{tv}(1)$ for all $t \in [0,1]$. Thus,

$$\begin{aligned} D\exp_p(I_0(v)) &= \frac{d}{dt}\exp_p(tv)|_{t=0} \\ &= \frac{d}{dt}\gamma_{tv}(1)|_{t=0} \\ &= \frac{d}{dt}\gamma_v(t)|_{t=0} \\ &= \dot{\gamma}_v(0) \\ &= v. \end{aligned}$$

In other words $D\exp_p \circ I_0$ is the identity map on T_pM. This shows that $D\exp_p$ is nonsingular. The second statement of (1) follows from the inverse function theorem.

The proof of (2) is again an exercise in unraveling tangent spaces and identifications. The tangent space $T_{(p,p)}(M \times M)$ is naturally identified with $T_pM \times T_pM$. The tangent space $T_{(p,0_p)}(TM)$ is also naturally identified to $T_pM \times T_{0_p}(T_pM) \simeq T_pM \times T_pM$.

We know that E takes (p, v) to $\big(p, \exp_p(v)\big)$. Note that varying p is just the identity in the first coordinate, but something unpredictable in the second. While if we fix p and vary v in T_pM, then the first coordinate is fixed and we simply have $\exp_p(v)$ in the second coordinate. This explains what the differential $DE|_{(p,0_p)}$ is. If we consider it as a linear map $T_pM \times T_pM \to T_pM \times T_pM$, then it is the identity on the first factor to the first factor, identically 0 from the second factor to the first, and the identity from the second fact to the second factor as it is $D\exp_p \circ I_{0_p}$. Thus it looks like

$$\begin{bmatrix} I & 0 \\ * & I \end{bmatrix}$$

which is clearly nonsingular.

Now, the inverse function theorem gives (local) diffeomorphisms via E of neighborhoods of $(p, 0_p) \in TM$ onto neighborhoods of $(p, p) \in M \times M$. Since E maps the zero section of TM diffeomorphically to the diagonal in $M \times M$ and the zero section is a properly embedded submanifold of TM, it is easy to see that these local diffeomorphisms fit together to give a diffeomorphism of a neighborhood of the zero section in TM onto a neighborhood of the diagonal in $M \times M$. □

All this formalism with the exponential maps yields some results with geometric meaning. First, we get a coordinate system around p by identifying T_pM with $\mathbb{R}^n$ via an isomorphism, and using that the exponential map $\exp_p : T_pM \to M$ is a diffeomorphism on a neighborhood of the origin. Such coordinates are called *normal (exponential) coordinates* at p. They are unique up to how we choose to identify T_pM with $\mathbb{R}^n$. Requiring this identification to be a linear isometry gives uniqueness up to an orthogonal transformation of $\mathbb{R}^n$. Later in the chapter we show that they are indeed normal in the sense that the Christoffel symbols vanish at p.

The second item of geometric interest is the following idea: Thinking about S^2 and great circles (which we know are geodesics), it is clear that we cannot say that two points that are close together are joined by a unique geodesic. On S^2 there will be a short geodesic connection, but there will be other, long ones, too. What might be hoped is that points that are close together would have a unique "short" geodesic connecting them. This is exactly what (2) in the proposition says! As long as we keep q_1 and q_2 near p, there is only one way to go from q_1 to q_2 via a geodesic that isn't very long, i.e., has the form $\exp_{q_1} tv$, $v \in T_{q_1}M$, with $|v|$ small. This will be made more useful and clear in the next section, where we show that such short geodesics in fact are segments.

Suppose N is an embedded submanifold of M. The normal bundle of N in M is the vector bundle over N consisting of the orthogonal complements of the tangent spaces $T_pN \subset T_pM$.

$$TN^\perp = \{v \in T_pM : p \in N, v \in (T_pN)^\perp \subset T_pM\}.$$

So for each $p \in N$, $T_pM = T_pN \oplus (T_pN)^\perp$ is an orthogonal direct sum. Define the *normal exponential map* $\exp^\perp$ by restricting exp to $O \cap TN^\perp$ so $\exp^\perp : O \cap TN^\perp \to M$. As in part (2) of the previous proposition, one can show that $D\exp^\perp$ is nonsingular at 0_p, $p \in N$. Then it follows that there is an open neighborhood U of the zero section in $TN^\perp$ on which $\exp^\perp$ is a diffeomorphism onto its image in M. Such an image $\exp^\perp(U)$ is called a *tubular neighborhood* of N in M, because if N is a curve in $\mathbb{R}^3$ it looks like a solid tube around the curve.

6. Why Short Geodesics Are Segments

In the previous section, we saw that points that are close together on a Riemannian manifold are connected by a short geodesic, and by exactly one short geodesic in fact. But so far, we don't have any real evidence that such short geodesics are segments. In this section we shall prove that short geodesics are segments. Incidentally, several different ways of saying that a curve is a segment are in common use: "minimal geodesic," "minimizing curve," "minimizing geodesic," and even "minimizing geodesic segment."

The precise result we want to prove in this section is this:

THEOREM 14. *Suppose M is a Riemannian manifold, $p \in M$, and $\varepsilon > 0$ is such that*

$$\exp_p : B(0, \varepsilon) \to U \subset M$$

is a diffeomorphism onto its image in M. Then $U = B(p, \varepsilon)$ and for each $v \in B(0, \varepsilon)$, the geodesic $\gamma_v : [0, 1] \to M$ defined by

$$\gamma_v(t) = \exp_p(tv)$$

is the unique segment in M from p to $\exp_p v$.

On $U = \exp_p(B(0,\varepsilon))$ we have the function $r(x) = |\exp_p^{-1}(x)|$. That is, r is simply the Euclidean distance function from the origin on $B(0,\varepsilon) \subset T_pM$ in exponential coordinates. We know that $\nabla r = \partial_r = \frac{1}{r}(x^i\partial_i)$ in Cartesian coordinates on T_pM. The goal here is to establish:

LEMMA 12. (The Gauss Lemma) *On (U,g) the function r satisfies $\nabla r = \partial_r$, where $\partial_r = D\exp_p(\partial_r)$.*

Let us see how this implies the Theorem.

PROOF OF THEOREM. First observe that in $B(0,\varepsilon)$ the integral curves for ∂_r are the line segments $\gamma(s) = s \cdot \frac{v}{|v|}$ of unit speed. The integral curves for ∂_r on U are therefore the unit speed geodesics $\gamma(s) = \exp\left(s \cdot \frac{v}{|v|}\right)$. Thus the Lemma implies that r is a distance function on U. This shows that among curves from p to $q = \exp(x)$ in $U - \{p\}$, the geodesic from p to q is the shortest curve, furthermore, it has length $< \varepsilon$. In particular, $U \subset B(p,\varepsilon)$. To see that this geodesic is a segment in M, we must show that any curve that leaves U has length $> \varepsilon$. Suppose we have a curve $\gamma : [0,b] \to M$ from p to q that leaves U. Let $a \in [0,b]$ be the largest value so that $\gamma(a) = p$. Then $\gamma|_{[a,b]}$ is a shorter curve from p to q. Next let $t_0 \in (a,b)$ be the first value for which $\gamma(t_0) \notin U$. Then $\gamma|_{(a,t_0)}$ lies entirely in $U - \{p\}$ and is shorter than the original curve. We now see

$$\begin{aligned}
\ell\left(\gamma|_{(a,t_0)}\right) &= \int_a^{t_0} |\dot\gamma|\, dt \\
&= \int_a^{t_0} |\nabla r| \cdot |\dot\gamma|\, dt \\
&\geq \int_a^{t_0} dr\,(\dot\gamma)\, dt \\
&= r(\gamma(t_0)) - r(\gamma(a)) \\
&= \varepsilon,
\end{aligned}$$

since $r(p) = 0$ and the values of r converge to ε as we approach ∂U. Thus γ is not a segment from p to q.

Finally we have to show that $B(p,\varepsilon) = U$. We already have $U \subset B(p,\varepsilon)$. Conversely if $q \in B(p,\varepsilon)$ then it is joined to p by a curve of length $< \varepsilon$. The above argument then shows that this curve lies in U. Whence $B(p,\varepsilon) \subset U$. □

PROOF OF GAUSS LEMMA. We select an orthonormal basis for T_pM and introduce Cartesian coordinates. These coordinates are then also used on $B(p,\varepsilon)$ via the exponential map. Denote these coordinates by $(x^1,\dots,x^n)$ and the coordinate vector fields by $\partial_1,\dots,\partial_n$. Then

$$\begin{aligned}
r^2 &= (x^1)^2 + \cdots + (x^n)^2, \\
\partial_r &= \frac{1}{r} x^i \partial_i.
\end{aligned}$$

To show that this is the gradient for $r(x)$ on (M,g), we must prove that $dr(v) = g(\partial_r, v)$. We already know that

$$dr = \frac{1}{r}(x^1 dx^1 + \cdots + x^n dx^n),$$

but we have no knowledge of g, since it is just some abstract metric.

We prove that $dr(v) = g(\partial_r, v)$ by using suitable vector fields in place of v. In fact we are going to use Jacobi fields for r. Let us start with $v = \partial_r$. The right hand side is 1 as the integral curves for ∂_r are unit speed geodesics. The left hand side is also quickly computed to be 1. Next we take a rotation vector field $J = -x^i\partial_j + x^j\partial_i,\ i, j = 2, \ldots, n,\ i < j$. In dimension 2 this is simply the angular field ∂_θ. We immediately see that the left hand side vanishes: $dr\,(J) = 0$. For the right hand side we first note that J really is a Jacobi field as $L_{\partial_r} J = [\partial_r, J] = 0$. Using that $\nabla_{\partial_r}\partial_r = 0$ we then get

$$\begin{aligned}
\partial_r g(\partial_r, J) &= g(\nabla_{\partial_r}\partial_r, J) + g(\partial_r, \nabla_{\partial_r} J) \\
&= 0 + g(\partial_r, \nabla_{\partial_r} J) \\
&= -g(\partial_r, \nabla_J \partial_r) \\
&= -\frac{1}{2} D_J g(\partial_r, \partial_r) \\
&= 0.
\end{aligned}$$

Thus $g(\partial_r, J)$ is constant along geodesics emanating from p. Next observe that

$$\begin{aligned}
|g(\partial_r, J)| &\leq |\partial_r||J| \\
&= |J| \\
&\leq \left|x^i\right| |\partial_j| + \left|x^j\right| |\partial_i| \\
&\leq r\,(x)\,(|\partial_i| + |\partial_j|)
\end{aligned}$$

Continuity of $D\exp_p$ shows that ∂_i, ∂_j are bounded on $B\,(p, \varepsilon)$. Thus $|g(\partial_r, J)| \to 0$ as $r \to 0$. This shows that $g(\partial_r, J) = 0$. Finally we observe that any vector v is a linear combination of ∂_r and rotation vector fields. This proves the claim. □

There is an equivalent statement of the Gauss Lemma asserting that

$$\exp_p : B\,(0, \varepsilon) \to B\,(p, \varepsilon)$$

is a *radial isometry*:

$$g\left(D\exp_p(\partial_r), D\exp_p(v)\right) = g_p\,(\partial_r, v)$$

on T_pM. A careful translation process of the previous proof shows that this is exactly what we have proved.

The next corollary is also an immediate consequence of the above theorem and its proof.

COROLLARY 4. *If $x \in M$ and $\varepsilon > 0$ is such that $\exp_x : B\,(0, \varepsilon) \to B\,(p, \varepsilon)$ is defined and a diffeomorphism, then for each $\delta < \varepsilon$,*

$$\exp_x(B(0, \delta)) = B(x, \delta),$$

and

$$\exp_x(\bar{B}(0, \delta)) = \bar{B}(x, \delta).$$

7. Local Geometry in Constant Curvature

Let us restate what we have done in this chapter so far. Given $p \in (M, g)$ we found coordinates near p using the exponential map such that the distance function $r(x) = d(p, x)$ to p has the formula

$$r(x) = \sqrt{(x^1)^2 + \cdots + (x^n)^2}.$$

Furthermore, we showed that $\nabla r = \partial_r$. By calculating each side in the formula in coordinates we get

$$\sum_{i,j} \frac{1}{r} g^{ij} x^j \partial_i = \nabla r = \partial_r = \frac{1}{r} x^i \partial_i.$$

After equating the coefficients of these vector fields we obtain the following curious relationship between the coordinates and the metric coefficients

$$\sum_j g^{ij} x^j = x^i$$

which is equivalent to

$$\sum_j g_{ij} x^j = x^i.$$

This relationship, as we shall see, fixes the behavior of g_{ij} around p up to first order and shows that the coordinates are normal in the sense used in chapter 2.

LEMMA 13.

$$g_{ij} = \delta_{ij} + O\left(r^2\right).$$

PROOF. The fact that $g_{ij}|_p = \delta_{ij}$ follows from taking one partial derivative on both sides of the above relation

$$\begin{aligned} \delta^i_k &= \partial_k x^i \\ &= \partial_k \sum_j g_{ij} x^j \\ &= \left(\partial_k g_{ij}\right) x^j + g_{ij} \partial_k x^j \\ &= \left(\partial_k g_{ij}\right) x^j + g_{ik}. \end{aligned}$$

As $x^j(p) = 0$, the claim follows. The fact that $\partial_k g_{ij}|_p = 0$ comes about by taking two partial derivatives on both sides

$$\begin{aligned} 0 &= \partial_l \partial_k x^i \\ &= \partial_l \left(\left(\partial_k g_{ij}\right) x^j\right) + \partial_l g_{ik} \\ &= \left(\partial_l \partial_k g_{ij}\right) x^j + \partial_k g_{ij} \partial_l x^j + \partial_l g_{ik} \\ &= \left(\partial_l \partial_k g_{ij}\right) x^j + \partial_k g_{il} + \partial_l g_{ik}. \end{aligned}$$

Evaluating at p then gives us

$$\partial_k g_{il}|_p + \partial_l g_{ik}|_p = 0.$$

Now combine this with the relations for the Christoffel symbols of the first kind:

$$\begin{aligned} \partial_i g_{kl} &= \Gamma_{ik,l} + \Gamma_{il,k}, \\ \Gamma_{ij,k} &= \frac{1}{2}\left(\partial_j g_{ik} + \partial_i g_{jk} - \partial_k g_{ji}\right) \end{aligned}$$

to get

$$\begin{aligned}\Gamma_{kl,i}|_p &= \frac{1}{2}\left(\partial_k g_{il}|_p + \partial_l g_{ik}|_p - \partial_i g_{kl}|_p\right)\\ &= -\frac{1}{2}\partial_i g_{kl}|_p\\ &= -\frac{1}{2}\left(\Gamma_{ik,l}|_p + \Gamma_{il,k}|_p\right)\\ &= \frac{1}{4}\left(\partial_l g_{ik}|_p + \partial_k g_{il}|_p\right)\\ &= 0\end{aligned}$$

which is what we wanted to prove. □

In polar coordinates around p any Riemannian metric therefore has the form

$$g = dr^2 + g_r$$

where g_r is a metric on S^{n-1}. The Euclidean metric looks like

$$\delta_{ij} = dr^2 + r^2 ds^2_{n-1},$$

where ds^2_{n-1} is the canonical metric on S^{n-1}. Since these two metrics agree up to first order we obtain

$$\begin{aligned}\lim_{r\to 0} g_r &= \lim_{r\to 0}\left(r^2 ds^2_{n-1}\right) = 0,\\ \lim_{r\to 0}\left(\partial_r g_r - \frac{2}{r}g_r\right) &= \lim_{r\to 0}\left(\partial_r\left(r^2 ds^2_{n-1}\right) - \frac{2}{r}\left(r^2 ds^2_{n-1}\right)\right) = 0.\end{aligned}$$

As

$$\partial_r g_r = 2\mathrm{Hess} r$$

this translates into

$$\lim_{r\to 0}\left(\mathrm{Hess} r - \frac{1}{r}g_r\right) = 0.$$

This can also be seen by computing the Hessian of $\frac{1}{2}r^2$ at p. Just note that this function has a critical point at p. Thus the coordinate formula for the Hessian is independent of the metric and must therefore be the identity map at p.

THEOREM 15. (Riemann, 1854) *If a Riemannian n-manifold (M, g) has constant sectional curvature k, then every point in M has a neighborhood that is isometric to an open subset of the space form S^n_k.*

PROOF. We use polar coordinates around $p \in M$ and the asymptotic behavior of g_r and Hess$r$ near p that was just established. We shall also use the fundamental equations that were introduced in chapter 2. On the same neighborhood we can also introduce a metric of constant curvature k

$$\begin{aligned}\tilde{g} &= dr^2 + \mathrm{sn}_k^2(r)\, ds^2_{n-1},\\ \mathrm{Hess}_{\tilde{g}} r &= \frac{\mathrm{sn}_k'(r)}{\mathrm{sn}_k(r)}\tilde{g}_r.\end{aligned}$$

Since the curvature is k for both of these metrics we see that $\mathrm{Hess}_{\tilde{g}} r$ and $\mathrm{Hess}_g r$ solve the same equation when evaluated on unit parallel fields perpendicular to ∂_r.

Note, however, that g and $\tilde{g}$ most likely have different parallel fields X and $\tilde{X}$:

$$\begin{aligned}\partial_r\left(\mathrm{Hess}_g r\left(X,X\right)\right)-\mathrm{Hess}_g^2 r\left(X,X\right) &= -\sec\left(X,\partial_r\right)=-k,\\ \lim_{r\to 0}\left(\mathrm{Hess}_g r\left(X,X\right)-\frac{1}{r}\right) &= 0,\end{aligned}$$

$$\begin{aligned}\partial_r\left(\mathrm{Hess}_{\tilde{g}} r\left(\tilde{X},\tilde{X}\right)\right)-\mathrm{Hess}_{\tilde{g}}^2 r\left(\tilde{X},\tilde{X}\right) &= -\sec\left(\tilde{X},\partial_r\right)=-k,\\ \lim_{r\to 0}\left(\mathrm{Hess} r_{\tilde{g}}\left(\tilde{X},\tilde{X}\right)-\frac{1}{r}\right) &= 0,\end{aligned}$$

Hence

$$\mathrm{Hess}_g r\left(X,X\right)=\mathrm{Hess}_{\tilde{g}} r\left(\tilde{X},\tilde{X}\right)=\frac{\mathrm{sn}_k'\left(r\right)}{\mathrm{sn}_k\left(r\right)}.$$

This shows that

$$\mathrm{Hess}_g r=\frac{\mathrm{sn}_k'\left(r\right)}{\mathrm{sn}_k\left(r\right)}g_r$$

When evaluating on Jacobi fields instead we see that both g_r and $\mathrm{sn}_k^2\left(r\right)ds_{n-1}^2$ solve the equation

$$\begin{aligned}\partial_r g_r &= 2\frac{\mathrm{sn}_k'\left(r\right)}{\mathrm{sn}_k\left(r\right)}g_r,\\ \lim_{r\to 0} g_r &= 0.\end{aligned}$$

This shows that

$$g=dr^2+\mathrm{sn}_k^2\left(r\right)ds_{n-1}^2.$$

In other words we have found a coordinate system on a neighborhood around $p\in M$ where the metric is the same as the constant curvature metric. □

8. Completeness

One of the foundational centerpieces of Riemannian geometry is the Hopf-Rinow theorem. This theorem states that all concepts of completeness are equivalent. This should not be an unexpected result for those who have played around with open subsets of Euclidean space. For it seems that in these examples, geodesic and metric completeness break down in exactly the same places. As with most foundational theorems, the proof is slightly intricate.

THEOREM 16. (H. Hopf-Rinow, 1931) *The following statements are equivalent:*

(1) M is geodesically complete, i.e., all geodesics are defined for all time.

(2) M is geodesically complete at p, i.e., all geodesics through p are defined for all time.

(3) M satisfies the Heine-Borel property, i.e., every closed bounded set is compact.

(4) M is metrically complete.

PROOF. (1) ⇒ (2), (3) ⇒ (4) are trivial.

(4) ⇒ (1) Recall that every geodesic $\gamma:[0,b)\to M$ defined on a maximal interval must leave every compact set if $b<\infty$. This violates metric completeness as $\gamma(t_i)$, $t_i\to b$ is a Cauchy sequence.

(2) ⇒ (3) Consider $\exp_p:T_pM\to M$. It suffices to show that

$$\exp_p\left(\overline{B}(0,r)\right)=\overline{B}(p,r)$$

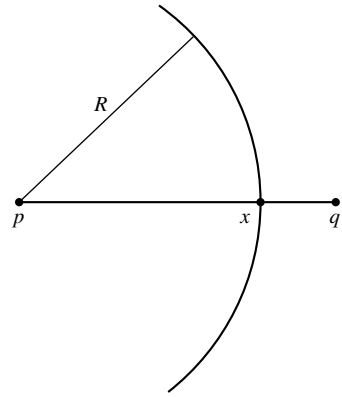

Figure 5.7

for all r (note that $\subset$ always holds). Consider

$$I = \{r : \exp(\overline{B}(0,r) = \overline{B}(p,r)\}.$$

(i) We have already seen that I contains all r close to zero.
(ii) I is closed: Let $r_i \in I$ converge to r and select $q \in \overline{B}(p,r)$ and $q_i \in \overline{B}(p,r_i)$ converging to q. We can find $v_i \in \overline{B}(0,r_i)$ with $q_i = \exp_p(v_i)$. Then (v_i) will subconverge to some $v \in \overline{B}(0,r)$. Continuity of $\exp_p$ then implies that $\exp_p(v) = q$. (You should think about why it is possible to choose the q_i's.)
(iii) I is open: We show that if $R \in I$, then $R+\varepsilon \in I$ for all small ε . First, choose a compact set K that contains $\overline{B}(p,R)$ in its interior. Then fix $\varepsilon > 0$ such that all points in K of distance $\leq \varepsilon$ can be joined by a unique geodesic segment. Given

$$q \in \overline{B}(p, R+\varepsilon) - \overline{B}(p,R)$$

select for each $\delta > 0$ a curve $\gamma_\delta : [0,1] \to M$ with

$$\begin{aligned} \gamma_\delta(0) &= p, \\ \gamma_\delta(1) &= q, \\ L(\gamma_\delta) &\leq d(p,q) + \delta. \end{aligned}$$

Suppose t_δ is the first value such that $\gamma_\delta(t_\delta) \in \partial\overline{B}(p,R)$. If x is an accumulation point for $\gamma_\delta(t_\delta)$, then we must have that

$$R + d(x,q) = d(p,x) + d(x,q) = d(p,q).$$

Now choose a segment from q to x and a segment from p to x of the form $\exp_p(tv)$, see also Figure 5.7. These two geodesics together form a curve from p to q of length $d(p,q)$. Hence, it is a segment. Consequently, it is smooth and by uniqueness of geodesics is the continuation of $\exp_p(tv)$, $0 \leq t \leq 1 + \frac{\varepsilon}{|\dot\gamma|}$. This shows that $q \in \exp_p\left(\overline{B}\,(0, R+\varepsilon)\right)$.

Statements (i), (ii), and (iii) together imply that $I = [0,\infty)$, which is what we wanted to prove. □

From part (ii) of (2) ⇒ (3) we get the additional result:

Corollary 5. *If M is complete in any of the above ways, then any two points in M can be joined by a segment.*

COROLLARY 6. *Suppose M admits a proper (preimages of compact sets are compact) Lipschitz function $f : M \to \mathbb{R}$. Then M is complete.*

PROOF. We establish the Heine-Borel property. Let $C \subset M$ be bounded and closed. Since f is Lipschitz the image $f(C)$ is also bounded. Thus $f(C) \subset [a, b]$ and $C \subset f^{-1}([a, b])$. As f is proper the preimage $f^{-1}([a, b])$ is compact. Since C is closed and a subset of a compact set it must itself be compact. □

This corollary makes it easy to check completeness for all of our examples. In these examples, the distance function can be extended to a proper continuous function on the entire space.

From now on, virtually all Riemannian manifolds will automatically be assumed to be connected and complete.

9. Characterization of Segments

In this section we will try to determine when a geodesic is a segment and then use this to find a maximal domain in T_pM on which the exponential map is an embedding. These issues can be understood through a systematic investigation of when distance functions to points are smooth. All Riemannian manifolds are assumed to be complete in this section.

9.1. The Segment Domain. Fix $p \in (M, g)$ and let $r(x) = d(x, p)$. We know that r is smooth near p and that the integral curves for r are geodesics emanating from p. Since M is complete, these integral curves can be continued indefinitely beyond the places where r is smooth. These geodesics could easily intersect after some time, thus they don't generate a flow on M, but just having them at points where r might not be smooth helps us understand why r is not smooth at these places. We know from chapter 2 that another obstruction to r being smooth is the possibility of conjugate points (we use the notation *conjugate points* instead of focal point for distance functions to a point).

To clarify matters we introduce some terminology: The *segment domain* is

$$\operatorname{seg}(p) = \left\{ v \in T_pM : \exp_p(tv) : [0, 1] \to M \text{ is a segment} \right\}.$$

The Hopf-Rinow Theorem implies that $M = \exp_p(\operatorname{seg}(p))$. We see that $\operatorname{seg}(p)$ is a closed star-shaped subset of T_pM. The star "interior" of $\operatorname{seg}(p)$ is

$$\operatorname{seg}^0(p) = \{sv : s \in [0, 1),\ v \in \operatorname{seg}(p)\}.$$

We shall show below that this set is in fact the interior of $\operatorname{seg}(p)$, but this requires that we know the set is open. We start by proving

PROPOSITION 19. *If $x \in \exp_p(\operatorname{seg}^0(p))$, then it joined to p by a unique segment. In particular $\exp_p$ is injective on $\operatorname{seg}^0(p)$.*

PROOF. To see this note that there is a segment $\sigma : [0, 1) \to M$ with $\sigma(0) = p,\ \sigma(t_0) = x,\ t_0 < 1$. Therefore, if $\hat{\sigma} : [0, t_0] \to M$ is another segment from p to x, we could construct a nonsmooth segment

$$\gamma(s) = \begin{cases} \hat{\sigma}(s), & s \in [0, t_0], \\ \sigma(s), & s \in [t_0, 1], \end{cases}$$

and we know that this is impossible. □

On the image $U_p = \exp_p(\text{seg}^0(p))$ we can define $\partial_r = D\exp_p(\partial_r)$, which is, we hope, the gradient for

$$r(x) = d(x,p) = |\exp_p^{-1}(x)|.$$

From our earlier observations we know that r would be smooth on U_p with gradient ∂_r if we could show that $\exp_p : \text{seg}^0(p) \to U_p$ is a diffeomorphism. This requires in addition to injectivity that the map is nonsingular and $\text{seg}^0(p)$ is open. Nonsingularity is taken care of in the next lemma.

LEMMA 14. *$\exp_p : \text{seg}^0(p) \to U_p$ is nonsingular everywhere, or, in other words, $D\exp_p$ is nonsingular at every point in $\text{seg}^0(p)$.*

PROOF. If $\exp_p$ is singular somewhere, then we can find v such that $\exp_p$ is singular at v and nonsingular at all points tv, $t \in [0,1)$. We claim that $v \notin \text{seg}^0(p)$. As $\gamma(t) = \exp_p(tv)$ is an embedding on $[0,1]$ we can find neighborhoods U around $[0,1)v \subset T_pM$ and V around $\gamma([0,1)) \subset M$ such that $\exp_p : U \to V$ is a diffeomorphism. Note that $v \notin U$ and $\gamma(1) \notin V$. If we take a tangent vector $w \in T_vT_pM$, then we can extend it to a Jacobi field J on T_pM, i.e., $[\partial_r, J] = 0$. Next J can be pushed forward via $\exp_p$ to a vector field, also called J, that also commutes with ∂_r on V. If $D\exp_p|_v w = 0$, then

$$\lim_{t\to 1} J|_{\exp(tv)} = \lim_{t\to 1} D\exp_p(J)|_{\exp(tv)} = 0.$$

In particular, we see that $D\exp_p$ is singular at v iff $\exp_p(v)$ is a conjugate point for r. This characterization of course assumes that r is smooth on a region that has $\exp_p(v)$ as a accumulation point.

The fact that

$$\lim_{t\to 1} g(J,J)|_{\exp(tv)} \searrow 0 \text{ as } t \to 1$$

implies that there must be a sequence of numbers $t_n \to 1$ such that

$$\frac{\partial_r g(J,J)}{g(J,J)}|_{\exp(t_n v)} \to -\infty \text{ as } n \to \infty.$$

Now use the first fundamental equation evaluated on the Jacobi field J

$$\partial_r g(J,J) = 2\text{Hess}r(J,J)$$

to conclude that Hess$r$ satisfies

$$\frac{\text{Hess}r(J,J)}{g(J,J)}|_{\exp(t_n v)} \to -\infty \text{ as } n \to \infty.$$

If we assume that $v \in \text{seg}^0(p)$, then $\gamma(t) = \exp_p(tv)$ is a segment on some interval $[0, 1+\varepsilon]$, $\varepsilon > 0$. Choose ε so small that $\tilde{r}(x) = d(x, \gamma(1+\varepsilon))$ is smooth on a ball $B(\gamma(1+\varepsilon), 2\varepsilon)$ (which contains $\gamma(1)$). Then consider the function

$$e(x) = r(x) + \tilde{r}(x).$$

From the triangle inequality, we know that

$$e(x) \geq 1 + \varepsilon = d(p, \gamma(1+\varepsilon))$$

Furthermore, $e(x) = 1+\varepsilon$ whenever $x = \gamma(t)$, $t \in [0, 1+\varepsilon]$. Thus, e has an absolute minimum along $\gamma(t)$ and must therefore have nonnegative Hessian at all the points $\gamma(t)$. On the other hand,

$$\frac{\text{Hess}e(J,J)}{g(J,J)}|_{\exp(t_n v)} = \frac{\text{Hess}r(J,J)}{g(J,J)}|_{\exp(t_n v)} + \frac{\text{Hess}\tilde{r}(J,J)}{g(J,J)}|_{\exp(t_n v)} \underset{n\to\infty}{\to} -\infty$$

since $\mathrm{Hess}\tilde{r}$ is bounded in a neighborhood of $\gamma(1)$ and the term involving $\mathrm{Hess}r$ goes to $-\infty$ as $n \to \infty$. □

We have now shown that $\exp_p$ is injective and has nonsingular differential on $\mathrm{seg}^0(p)$. Before showing that $\mathrm{seg}^0(p)$ is open we characterize elements in the star "boundary" of $\mathrm{seg}^0(p)$ as points that fail to have one of these properties.

LEMMA 15. *If $v \in \mathrm{seg}(p) - \mathrm{seg}^0(p)$, then either*
(1) $\exists w\,(\neq v) \in \mathrm{seg}(p) : \exp_p(v) = \exp_p(w)$, *or*
(2) $D\exp_p$ *is singular at* v.

PROOF. Let $\gamma(t) = \exp_p(tv)$. For $t > 1$ choose segments

$$\begin{aligned} \sigma_t(s) &: [0,1] \to M, \\ \sigma_t(0) &= p, \\ \sigma_t(1) &= \gamma(t). \end{aligned}$$

Since we have assumed that $\gamma : [0,t]$ is not a segment for $t > 1$ we see that $\dot{\sigma}_t(0)$ is never proportional to $\dot{\gamma}(0)$. Now choose $t_n \to 1$ such that $\dot{\sigma}_{t_n}(0) \to w \in T_pM$. We have that

$$\ell(\sigma_{t_n}) = |\dot{\sigma}_{t_n}(0)| \to \ell(\gamma|_{[0,1]}) = |\dot{\gamma}(0)|,$$

so $|w| = |\dot{\gamma}(0)|$. Now either $w = \dot{\gamma}(0)$ or $w \neq \dot{\gamma}(0)$. In the latter case, we note that w is not a positive multiple of $\dot{\gamma}(0)$ since $|w| = |\dot{\gamma}(0)|$. Therefore, we have found the promised w from (1). If the former happens, we must show that $D\exp_p$ is singular at v. If, in fact, $D\exp_p$ is nonsingular at v, then $\exp_p$ is an embedding near v. Thus,

$$\begin{aligned} \dot{\sigma}_{t_n}(0) &\to v = \dot{\gamma}(0), \\ \exp_p(\dot{\sigma}_{t_n}(0)) &= \exp_p(t_n\dot{\gamma}(0)), \end{aligned}$$

implies $\dot{\sigma}_{t_n}(0) = t_n \cdot v$, showing that γ is a segment on some interval $[0, t_n]$, $t_n > 1$. This, however, contradicts our choice of γ. □

Notice that in the first case the gradient ∂_r on M becomes undefined at $x = \exp_p(v)$, since it would be either $D\exp_p(v)$ or $D\exp_p(w)$, while in the second case the Hessian of r becomes undefined, since it is forced to go to $-\infty$ along certain fields. Finally we show

PROPOSITION 20. $\mathrm{seg}^0(p)$ *is open.*

PROOF. If we fix $v \in \mathrm{seg}^0(p)$, then there is going to be a neighborhood V around v on which $\exp_p$ is a diffeomorphism onto its image. If $v_i \in V$ converge to v, then we know that $D\exp_p$ is also nonsingular at v_i. Now assume that $w_i \in \mathrm{seg}(p)$ satisfy

$$\exp_p(v_i) = \exp_p(w_i).$$

In case w_i has an accumulation point $w \neq v$, we get $v \notin \mathrm{seg}^0(p)$. Hence $w_i \to v$, showing that $w_i \in V$ for large i. As $\exp_p$ is a diffeomorphism on V this implies that $w_i = v_i$. Thus we have shown that $v^i \in \mathrm{seg}^0(p)$. □

All of this implies that $r(x) = d(x,p)$ is smooth on the open and dense subset $U_p - \{p\} \subset M$ and in addition that it is not smooth on $M - U_p$.

The set $\mathrm{seg}(p) - \mathrm{seg}^0(p)$ is called the *cut locus* of p in T_pM. Thus, being inside the cut locus means that we are on the region where r is smooth. Going back to our characterization of segments, we have

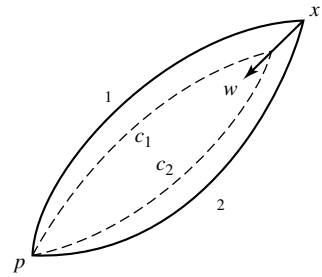

Figure 5.8

COROLLARY 7. *Let $\gamma : [0,\infty) \to M$ be a geodesic with $\gamma(0) = p$. If*

$$\operatorname{cut}(\dot\gamma(0)) = \max\{t : \gamma|_{[0,t]} \text{ is a segment}\},$$

then r is smooth at $\gamma(t)$, $t < \operatorname{cut}(\dot\gamma(0))$, but not smooth at $x = \gamma(\operatorname{cut}(\dot\gamma(0)))$. Furthermore, the failure of r to be smooth at x is because $\exp_p : \operatorname{seg}(p) \to M$ either fails to be one-to-one at x or has x as a critical value.

9.2. The Injectivity Radius. The largest radius ε for which

$$\exp_p : B(0,\varepsilon) \to B(p,\varepsilon)$$

is a diffeomorphism is called the *injectivity radius* $\operatorname{inj}(p)$ at p. If $v \in \operatorname{seg}(p) - \operatorname{seg}^0(p)$ is the closest point to 0 in this set, then we have that $\operatorname{inj}(p) = |v|$. It turns out that such v can be characterized as:

LEMMA 16. (Klingenberg)*: Suppose $v \in \operatorname{seg}(p) - \operatorname{seg}^0(v)$ and that $|v| = \operatorname{inj}(p)$. Then either*

(1) there is precisely one other vector w with

$$\exp_p(w) = \exp_p(v),$$

and it is characterized by

$$\frac{d}{dt}|_{t=1} \exp_p(tw) = -\frac{d}{dt}|_{t=1} \exp_p(tv),$$

or

(2) $x = \exp_p(v)$ is a critical value for $\exp_p : \operatorname{seg}(p) \to M$.

In the first case there are exactly two segments from p to $x = \exp_p(v)$, and they fit smoothly together at x to form a geodesic loop.

PROOF. Suppose x is a regular value for $\exp_p : \operatorname{seg}(p) \to M$ and that $\gamma_1, \gamma_2 : [0,1] \to M$ are segments from p to $x = \exp_p(v)$. If $\dot\gamma_1(1) \neq -\dot\gamma_2(1)$, then we can find $w \in T_xM$ such that $g(w, \dot\gamma_1(1))$, $g(w, \dot\gamma_2(1)) < 0$, i.e., w forms an angle $> \frac{\pi}{2}$ with both $\dot\gamma_1(1)$ and $\dot\gamma_2(1)$. Next select $c(s)$ with $\dot c(0) = w$. As $D\exp_p$ is nonsingular at $\dot\gamma_i(0)$ there are unique curves $v_i(s) \in T_pM$ with $v_i(0) = \dot\gamma_i(0)$ and $D\exp_p(v_i(s)) = c(s)$ (see also Figure 5.8). But then the curves $t \to \exp_p(tv_i(s))$ have length

$$\begin{aligned} |v_i| &= d(p, c(s)) \\ &< d(p,x) \\ &= |v|. \end{aligned}$$

This implies that $\exp_p$ is not one-to-one on $\operatorname{seg}^0(p)$, a contradiction. □

10. Riemannian Isometries

We are now ready to explain the key properties of Riemannian isometries. Much of theory is local, so we shall not necessarily assume that the Riemannian manifolds being investigated are complete. After this thorough discussion of Riemannian isometries we classify all complete simply connected Riemannian manifolds of constant sectional curvature.

10.1. Local Isometries. We say that a map $F : (M, g) \to (N, \bar{g})$ is a *local Riemannian isometry* if for each $p \in M$ the differential $DF_p : T_pM \to T_{F(p)}N$ is a linear isometry. A special and trivial example of such a map is a local coordinate system $\varphi : U \to \Omega \subset \mathbb{R}^n$ where we use the induced metric g on U and its coordinate representation $g_{ij}dx^i dx^j$ on Ω.

PROPOSITION 21. *Let $F : (M, g) \to (N, \bar{g})$ be a local Riemannian isometry.*
(1) F maps geodesics to geodesics.
(2) $F \circ \exp_p(v) = \exp_{F(p)} \circ DF_p(v)$ if $\exp_p(v)$ is defined. In other words

$$\begin{array}{ccc} O_p \subset T_pM & \xrightarrow{DF} & O_{F(p)} \subset T_{F(p)}N \\ \exp_p \downarrow & & \exp_{F(p)} \downarrow \\ M & \xrightarrow{F} & N \end{array}$$

(3) F is distance decreasing.
(4) If F is also a bijection, then it is distance preserving.

PROOF. (1) The first property is completely obvious. We know that geodesics depend only on the metric and not on any given coordinate system. However, a local Riemannian isometry is locally nothing but a change of coordinates.

(2) If $\exp_p(v)$ is defined, then $t \to \exp_p(tv)$ is a geodesic. Thus also $t \to F\left(\exp_p(tv)\right)$ is a geodesic. Since

$$\begin{aligned} \frac{d}{dt} F\left(\exp_p(tv)\right)|_{t=0} &= DF\left(\frac{d}{dt}\exp_p(tv)|_{t=0}\right) \\ &= DF(v), \end{aligned}$$

we have that $F\left(\exp_p(tv)\right) = \exp_{F(p)}(tDF(v))$. Setting $t = 1$ then proves the claim.

(3) This is also obvious as F must preserve the length of curves.

(4) Both F and F^{-1} are distance decreasing so they must both be distance preserving. □

This proposition quickly yields two important results for local Riemannian isometries.

PROPOSITION 22. (Uniqueness of Riemannian Isometries) *Let $F, G : (M, g) \to (N, \bar{g})$ be local Riemannian isometries. If M is connected and $F(p) = G(p)$, $DF_p = DG_p$, then $F = G$ on M.*

PROOF. Let

$$A = \{x \in M : F(x) = G(x),\ DF_x = DG_x\}.$$

We know that $p \in A$ and that A is closed. Property (2) from the above proposition tells us that

$$\begin{aligned} F \circ \exp_x (v) &= \exp_{F(x)} \circ DF_x (v) \\ &= \exp_{G(x)} \circ DG_x (v) \\ &= G \circ \exp_x (v), \end{aligned}$$

if $x \in A$. Since $\exp_x$ maps onto a neighborhood of x it follows that some neighborhood of x also lies in A. This shows that A is open and hence all of M as M is connected. □

PROPOSITION 23. *Let $F : (M, \bar{g}) \to (N, g)$ be a Riemannian covering map. $(M, \bar{g})$ is complete if and only if (N, g) is complete.*

PROOF. Let $\gamma : (-\varepsilon, \varepsilon) \to N$ be a geodesic with $\gamma (0) = p$ and $\dot{\gamma} (0) = v$. For any $\bar{p} \in F^{-1} (p)$ there is a unique lift $\bar{\gamma} : (-\varepsilon, \varepsilon) \to M$, i.e., $F \circ \bar{\gamma} = \gamma$, with $\bar{\gamma} (0) = \bar{p}$. Since F is a local isometry, the inverse is locally defined and also an isometry. Thus $\bar{\gamma}$ is also a geodesic.

If we assume N is complete, then γ and also $\bar{\gamma}$ will exist for all time. As all geodesics in M must be of the form $\bar{\gamma}$ this shows that all geodesics in M exist for all time.

If, conversely, we suppose that M is complete, then $\bar{\gamma}$ can be extended to be defined for all time. Then $F \circ \bar{\gamma}$ is a geodesic defined for all time that extends γ. Thus N is geodesically complete. □

LEMMA 17. *Let $F : (M, g) \to (N, \bar{g})$ be a local Riemannian isometry. If M is complete, then F is a Riemannian covering map.*

PROOF. Fix $q \in N$ and assume that $\exp_q : B (0, \varepsilon) \to B (q, \varepsilon)$ is a diffeomorphism. We claim that $F^{-1} (B (q, \varepsilon))$ is evenly covered by the sets $B (p, \varepsilon)$ where $F (p) = q$. Completeness of M guarantees that $\exp_p : B (0, \varepsilon) \to B (p, \varepsilon)$ is defined and property (2) that

$$F \circ \exp_p (v) = \exp_q \circ DF_p (v)$$

for all $v \in B (0, \varepsilon) \subset T_p M$. As $\exp_q : B (0, \varepsilon) \to B (q, \varepsilon)$ and $DF_p : B (0, \varepsilon) \to B (0, \varepsilon)$ are diffeomorphisms it follows that $F \circ \exp_p : B (0, \varepsilon) \to B (q, \varepsilon)$ is a diffeomorphism. Thus each of the maps $\exp_p : B (0, \varepsilon) \to B (p, \varepsilon)$ and $F : B (p, \varepsilon) \to B (q, \varepsilon)$ are diffeomorphisms as well. Finally we need to make sure that

$$F^{-1} (B (q, \varepsilon)) = \bigcup_{F(p)=q} B (p, \varepsilon).$$

If $x \in F^{-1} (B (q, \varepsilon))$, then we can join q and $F (x)$ by a unique geodesic $\gamma (t) = \exp_q (tv)$, $v \in B (0, \varepsilon)$. Completeness of M again guarantees a geodesic $\sigma : [0, 1] \to M$ with $\sigma (1) = x$ and $DF_x (\dot{\sigma} (1)) = \dot{\gamma} (1)$. Since $F \circ \sigma$ is a geodesic with the same initial values at $t = 1$ as γ we must have $F (\sigma (t)) = \gamma (t)$ for all t. As $q = \gamma (0)$ we have therefore proven that $F (\sigma (0)) = q$ and hence that $x \in B (\sigma (0), \varepsilon)$. □

If $S \subset \mathrm{Iso} (M, g)$ is a set of isometries, then the *fixed point set* of S is defined as those points in M that are fixed by all isometries in S

$$\mathrm{Fix} (S) = \{x \in M : F (x) = x \text{ for all } F \in S\}.$$

While the fixed point set for a general set of diffeomorphisms can be quite complicated, the situation for isometries is much more manageable. A submanifold

$N \subset (M,g)$ is said to be *totally geodesic* if for each $p \in N$ a neighborhood of $0 \in T_pN$ is mapped into N via the exponential map $\exp_p$. This means that geodesics in N are also geodesics in M and conversely that any geodesic in M which is tangent to N at some point must lie in N for a short time.

PROPOSITION 24. *Let $S \subset \mathrm{Iso}(M,g)$ be a set of isometries, then each connected component of the fixed point set is a totally geodesic submanifold.*

PROOF. Let $p \in \mathrm{Fix}(S)$ and consider the subspace $V \subset T_pM$ that is fixed by the linear isometries $DF_p : T_pM \to T_pM$, where $F \in S$. Note that each such F fixes p so we know that $DF_p : T_pM \to T_pM$. If $v \in V$, then $t \to \exp_p(tv)$ must be fixed by each of the isometries in S as the initial position and velocity is fixed by these isometries. Thus $\exp_p(tv) \in \mathrm{Fix}(S)$ as long as it is defined. This shows that $\exp_p : V \to \mathrm{Fix}(S)$.

Next let $\varepsilon > 0$ be chosen so that $\exp_p : B(0,\varepsilon) \to B(p,\varepsilon)$ is a diffeomorphism. If $q \in \mathrm{Fix}(S) \cap B(p,\varepsilon)$, then the unique geodesic $\gamma : [0,1] \to B(p,\varepsilon)$ from p to q has the property that its endpoints are fixed by each $F \in S$. Now $F \circ \gamma$ is also a geodesic from p to q which in addition lies in $B(p,\varepsilon)$ as the length is unchanged. Thus $F \circ \gamma = \gamma$ and hence γ lies in $\mathrm{Fix}(S) \cap B(p,\varepsilon)$.

Thus we have shown that $\exp_p : V \cap B(0,\varepsilon) \to \mathrm{Fix}(S) \cap B(p,\varepsilon)$ is a bijection. This establishes the lemma. □

10.2. Constant Curvature Revisited. We just saw that isometries are uniquely determined by their differential. What about the existence question? Given any linear isometry $L : T_pM \to T_qN$, is there an isometry $F : M \to N$ such that $DF_p = L$? If we let $M = N$, this would in particular mean that if π is a 2-plane in T_pM and $\tilde{\pi}$ a 2-plane in T_qM, then there should be an isometry $F : M \to M$ such that $F(\pi) = \tilde{\pi}$. But this would imply that M has constant sectional curvature. The above problem can therefore not be solved in general. If we go back and inspect our knowledge of $\mathrm{Iso}(S^n_k)$, we see that these spaces have enough isometries so that any linear isometry $L : T_pS^n_k \to T_qS^n_k$ can be extended to a global isometry $F : S^n_k \to S^n_k$ with $DF_p = L$. In some sense these are the only spaces with this property, as we shall see.

THEOREM 17. *Suppose (M,g) is a Riemannian manifold of dimension n and constant curvature k. If M is simply connected and $L : T_pM \to T_qS^n_k$ is a linear isometry, then there is a unique local Riemannian isometry called the* monodromy map *$F : M \to S^n_k$ with $DF_p = L$. Furthermore, this map is a diffeomorphism if (M,g) is complete.*

Before giving the proof, let us look at some examples.

EXAMPLE 37. *Suppose we have an immersion $M^n \to S^n_k$. Then F will be one of the maps described in the theorem if we use the pullback metric on M. Such maps can fold in wild ways when $n \geq 2$ and need not resemble covering maps in any way whatsoever.*

EXAMPLE 38. *If $U \subset S^n_k$ is a contractible bounded open set with ∂U a smooth hypersurface, then one can easily construct a diffeomorphism $F : M = S^n_k - \{pt\} \to S^n_k - U$. Near the missing point in M the metric will necessarily look pretty awful, although it has constant curvature.*

EXAMPLE 39. *If $M = \mathbb{R}P^n$ or $(\mathbb{R}^n - \{0\})/$antipodal map, then M is not simply connected and does not admit an immersion into S_k^n.*

EXAMPLE 40. *If M is the universal covering of the constant curvature sphere with a pair of antipodal point removed $S^2 - \{\pm p\}$, then the monodromy map is not one-to-one. In fact it must be the covering map $M \to S^2 - \{\pm p\}$.*

COROLLARY 8. *If M is a closed simply connected manifold with constant-curvature k, then $k > 0$ and $M = S^n$. Thus, $S^p \times S^q, \mathbb{C}P^n$ do not admit any constant curvature metrics.*

COROLLARY 9. *If M is geodesically complete and noncompact with constant curvature k, then $k \leq 0$ and the universal covering is diffeomorphic to $\mathbb{R}^n$. In particular, $S^2 \times \mathbb{R}^2$ and $S^n \times \mathbb{R}$ do not admit any geodesically complete metrics of constant curvature.*

Now for the proof of the theorem. A different proof of the case where M is complete is developed in the exercises to this chapter.

PROOF. We know that M can be covered by sets U_α such that each U_α admits a Riemannian embedding $F_\alpha : U_\alpha \to S_k^n$. Furthermore, if $q \in U_\alpha$, $\bar{q} \in S_k^n$ and $L : T_q U_\alpha \to T_{\bar{q}} S_k^n$ is a linear isometry, then there is a unique F_α such that $F_\alpha(q) = \bar{q}$ and $DF_\alpha|_p = L$.

The construction of F now proceeds in the same way one does analytic continuation on simply connected domains. We fix base points $p \in M$, $\bar{p} \in S_k^n$ and a linear isometry $L : T_p M \to T_{\bar{p}} S_k^n$. Next let $x \in M$ be an arbitrary point. If $\gamma : [0,1] \to M$ is a curve from p to x, then we can cover it by a string of sets $U_{\alpha_0}, ..., U_{\alpha_k}$, where $p \in U_{\alpha_0}$, $x \in U_{\alpha_k}$, and $\gamma(t_i) \in U_{\alpha_i} \cap U_{\alpha_{i+1}}$. Define F on U_{a_0} so that $F(p) = \bar{p}$ and $DF_p = L$. Then define $F|_{U_{\alpha_{i+1}}}$ successively such that it agrees with $F|_{U_{\alpha_i}}$ and $DF|_{U_{\alpha_i}}$ at $\gamma(t_i)$. This defines F uniquely on all of the sets U_{α_i} and hence also at x. If we covered γ by a different string of sets, then uniqueness of isometries tell us that we have to get the same answer along γ as we assume that $F(p) = \bar{p}$ and $DF_p = L$. If we used a different path $\bar{\gamma}$ which was also covered by the same string of sets U_{α_i} we would clearly also end up with the same answer at x. Finally we use that M is simply connected to connect any two paths γ_0, γ_1 from p to x by a family of paths $H(s,t)$ such that each $\gamma_s(t) = H(s,t)$ is a path from p to x. If F_{γ_s} is the map we obtain near x by using the path γ_s, then we have just seen that $F_{\gamma_s}(x)$ is fixed as long as s is so small that all the curves are covered by the same string of sets. This shows that $s \to F_{\gamma_s}(x)$ is locally constant and hence that $F(x)$ is well-defined by our construction.

If M is complete we know that F has to be a covering map. As S_k^n is simply connected it must be a diffeomorphism. □

We can now give the classification of complete simply connected Riemannian manifolds with constant curvature. This result was actually proven before the issues of completeness were completely understood. Killing first proved the result assuming in effect that the manifold has an $\varepsilon > 0$ such that for all p the map $\exp_p : B(0,\varepsilon) \to B(p,\varepsilon)$ is a diffeomorphism. Hopf then realized that it was sufficient to assume that the manifold was geodesically complete. Since metric completeness immediately implies geodesic completeness this is clearly the best result one could have expected at the time.

COROLLARY 10. (Classification of Constant Curvature Spaces, Killing, 1893 and H. Hopf, 1926) *If (M,g) is a connected, geodesically complete Riemannian manifold with constant curvature k , then the universal covering is isometric to S^n_k.*

This result shows how important the completeness of the metric is. A large number of open manifolds admit immersions into Euclidean space of the same dimension (e.g., $S^n \times \mathbb{R}^k$) and therefore carry incomplete metrics with zero curvature. Carrying a complete Riemannian metric of a certain type, therefore, often implies various topological properties of the underlying manifold. Riemannian geometry at its best tries to understand this interplay between metric and topological properties.

10.3. Metric Characterization of Maps. As promised we shall in this section give some metric characterizations of Riemannian isometries and Riemannian submersions. For a Riemannian manifold (M,g) we let the corresponding metric space be denoted by (M,d_g) or simply (M,d) if only one metric is in play. It is natural to ask whether one can somehow recapture the Riemannian metric g from the distance d_g. If for instance $v,w \in T_pM$, then we would like to be able to compute $g(v,w)$ from knowledge of d_g. One way of doing this is by taking two curves α,β such that $\dot\alpha(0)=v$ and $\dot\beta(0)=w$ and observe that

$$\begin{aligned} |v| &= \lim_{t\to 0}\frac{d\left(\alpha(t),\alpha(0)\right)}{t}, \\ |w| &= \lim_{t\to 0}\frac{d\left(\beta(t),\beta(0)\right)}{t}, \\ \cos\angle(v,w) &= \frac{g\left(v,w\right)}{|v|\,|w|} = \lim_{t\to 0}\frac{d\left(\alpha(t),\beta(t)\right)}{t}. \end{aligned}$$

Thus, g can really be found from d given that we use the differentiable structure of M. It is perhaps then not so surprising that many of the Riemannian maps we consider have synthetic characterizations, that is, characterizations that involve only knowledge of the metric space (M,d).

Before proceeding with our investigations, let us introduce a new type of coordinates. Using geodesics we have already introduced one set of geometric coordinates via the exponential map. We shall now use the distance functions to construct *distance coordinates*. For a point $p \in M$ fix a neighborhood $U \ni p$ such that for each $x \in U$ we have that $B\left(q,\mathrm{inj}(q)\right) \supset U$. Thus, for each $q \in U$ the distance function $r_q(x) = d\left(x,q\right)$ is smooth on $U - \{q\}$. Now choose $q_1,\ldots,q_n \in U - \{p\}$, where $n = \dim M$. If the vectors $\nabla r_{q_1}(p),\ldots,\nabla r_{q_n}(p) \in T_pM$ are linearly independent, the inverse function theorem tells us that $\varphi = (r_{q_1},\ldots,r_{q_n})$ can be used as coordinates on some neighborhood V of p. The size of the neighborhood will depend on how these gradients vary. Thus, an explicit estimate for the size of V can be gotten from bounds on the Hessians of the distance functions. Clearly, one can arrange for the gradients to be linearly independent or even orthogonal at any given point.

We just saw that bijective Riemannian isometries are distance preserving. The next result shows that the converse is also true.

THEOREM 18. (Myers-Steenrod, 1939) *If (M,g) and $(N,\bar g)$ are Riemannian manifolds and $F : M \to N$ a bijection, then F is a Riemannian isometry if F is distance-preserving, i.e., $d_{\bar g}\left(F(p),F(q)\right) = d_g\left(p,q\right)$ for all $p,q \in M$.*

PROOF. Let F be distance-preserving. First let us show that F is differentiable. Fix $p \in M$ and let $q = F(p)$. Near q introduce distance coordinates $(r_{q_1},\ldots,r_{q_n})$

and find p_i such that $F(p_i) = q_i$. Now observe that

$$\begin{aligned} r_{q_i} \circ F(x) &= d(F(x), q_i) \\ &= d(F(x), F(p_i)) \\ &= d(x, p_i). \end{aligned}$$

Since $d(p, p_i) = d(q, q_i)$, we can assume that the q_is and p_is are chosen such that $r_{p_i}(x) = d(x, p_i)$ are smooth at p. Thus, $(r_{q_1}, \ldots, r_{q_n}) \circ F$ is smooth at p, showing that F must be smooth at p.

To show that F is a Riemannian isometry it suffices to check that $|DF(v)| = |v|$ for all tangent vectors $v \in TM$. For a fixed $v \in T_pM$ let $\gamma(t) = \exp_p(tv)$. For small t we know that γ is a constant speed segment. Thus, for small t, s we can conclude

$$|t - s| \cdot |v| = d_g(\gamma(t), \gamma(s)) = d_{\bar{g}}(F \circ \gamma(t), F \circ \gamma(s)),$$

implying

$$\begin{aligned} |DF(v)| &= \left| \frac{d(F \circ \gamma)}{dt} \right|_{t=0} \\ &= \lim_{t \to 0} \frac{d_{\bar{g}}(F \circ \gamma(t), F \circ \gamma(0))}{|t|} \\ &= \lim_{t \to 0} \frac{d_g(\gamma(t), \gamma(0))}{|t|} \\ &= |\dot{\gamma}(0)| \\ &= |v|. \end{aligned}$$

□

Our next goal is to find a characterization of Riemannian submersions. Unfortunately, the description only gives us functions that are C^1, but there doesn't seem to be a better formulation. Let $F : (M, \bar{g}) \to (N, g)$ be a function. We call F a *submetry* if for every $p \in M$ we can find $r > 0$ such that for each $\varepsilon \leq r$ we have $F(B(p, \varepsilon)) = B(F(p), \varepsilon)$. Submetries are locally distance-nonincreasing and therefore also continuous. In addition, we have that the composition of submetries (or Riemannian submersions) are again submetries (or Riemannian submersions). We can now prove

THEOREM 19. (Berestovski, 1995) *If $F : (M, \bar{g}) \to (N, g)$ is a surjective submetry, then F is a C^1 Riemannian submersion.*

PROOF. Fix points $q \in N$ and $p \in M$ with $F(p) = q$. Then select distance coordinates $(r_1, \ldots, r_k)$ around q. Now observe that all of the r_is are Riemannian submersions and therefore also submetries. Then the compositions $r_i \circ F$ are also submetries. Thus, F is C^1 iff all the maps $r_i \circ F$ are C^1. Therefore, it suffices to prove the result in the case of functions $r : (U \subset M, g) \to ((a, b), \text{can})$.

Let $x \in M$. By restricting r to a small convex neighborhood of x, we can assume that the fibers of r are closed and that any two points in the domain are joined by a unique geodesic. We now wish to show that r has a continuous unit gradient field ∇r. We know that the integral curves for ∇r should be exactly the unit speed geodesics that are mapped to unit speed geodesics by r. Since r is distance-nonincreasing, it is clear that any piecewise smooth unit speed curve that is mapped to a unit speed geodesic must be a smooth unit speed geodesic. Thus,

these integral curves are unique and vary continuously to the extent that they exist. To establish the existence of these curves we use the submetry property. First fix $p \in M$ and let $\gamma(t) : [0, r] \to (a, b)$ be the unit speed segment with $\gamma(0) = r(p)$. Let U_t denote the fiber of r above $\gamma(t)$. Now select a unit speed segment $\bar{\gamma} : [0, r] \to M$ with $\bar{\gamma}(0) = p$ and $\gamma(r) \in U_r$. This is possible since $r(B(p, \varepsilon)) = B(\gamma(0), \varepsilon)$. It is now easy to check, again using the submetry property, that $\gamma(t) = r \circ \bar{\gamma}(t)$, as desired. □

11. Further Study

There are many textbooks on Riemannian geometry that treat all of the basic material included in this chapter. Some of the better texts are [**19**], [**20**], [**41**], [**56**] and [**73**]. All of these books, as is usual, emphasize the variational approach as being *the* basic technique used to prove every theorem. To see how the variational approach works the text [**68**] is also highly recommended.

12. Exercises

(1) Assume that (M, g) has the property that all geodesics exist for a fixed time $\varepsilon > 0$. Show that (M, g) is geodesically complete.
(2) A Riemannian manifold is said to be homogeneous if the isometry group acts transitively. Show that homogeneous manifolds are geodesically complete.
(3) Assume that we have coordinates in a Riemannian manifold so that $g_{1i} = \delta_{1i}$. Show that x^1 is a distance function.
(4) Let γ be a geodesic in a Riemannian manifold (M, g). Let g' be another Riemannian metric on M with the properties: $g'(\dot{\gamma}, \dot{\gamma}) = g(\dot{\gamma}, \dot{\gamma})$ and $g'(X, \dot{\gamma}) = 0$ iff $g(X, \dot{\gamma}) = 0$. Show that γ is also a geodesic with respect to g'.
(5) Show that if we have a vector field X on a Riemannian manifold (M, g) that vanishes at $p \in M$, then for any tensor T we have $L_X T = \nabla_X T$ at p. Conclude that the Hessian of a function is independent of the metric at a critical point. Can you find an interpretation of $L_X T$ at p?
(6) Show that any Riemannian manifold (M, g) admits a conformal change $\left(M, \lambda^2 g\right)$, where $\lambda : M \to (0, \infty)$, such that $\left(M, \lambda^2 g\right)$ is complete.
(7) On an open subset $U \subset \mathbb{R}^n$ we have the induced distance from the Riemannian metric, and also the induced distance from $\mathbb{R}^n$. Show that the two can agree even if U isn't convex.
(8) Let $N \subset (M, g)$ be a submanifold. Let ∇^N denote the connection on N that comes from the metric induced by g. Define the second fundamental form of N in M by

$$\mathrm{II}(X, Y) = \nabla^N_X Y - \nabla_X Y$$

   (a) Show that $\mathrm{II}(X, Y)$ is symmetric and hence tensorial in X and Y.
   (b) Show that $\mathrm{II}(X, Y)$ is always normal to N.
   (c) Show that $\mathrm{II} = 0$ on N iff N is totally geodesic.
   (d) If R^N is the curvature tensor for N, then

$$\begin{aligned} g(R(X, Y)Z, W) &= g\left(R^N(X, Y)Z, W\right) \\ &\quad - g(\mathrm{II}(Y, Z), \mathrm{II}(X, W)) + g(\mathrm{II}(X, Z), \mathrm{II}(Y, W)). \end{aligned}$$

(9) Let $f : (M, g) \to \mathbb{R}$ be a smooth function on a Riemannian manifold.
   (a) If $\gamma : (a, b) \to M$ is a geodesic, compute the first and second derivatives of $f \circ \gamma$.
   (b) Use this to show that at a local maximum (or minimum) for f the gradient is zero and the Hessian nonpositive (or nonnegative).
   (c) Show that f has everywhere nonnegative Hessian iff $f \circ \gamma$ is convex for all geodesics γ in (M, g).

(10) Let $N \subset M$ be a submanifold of a Riemannian manifold (M, g).
   (a) The distance from N to $x \in M$ is defined as

$$d(x, N) = \inf \{d(x, p) : p \in N\}.$$

   A unit speed curve $\sigma : [a, b] \to M$ with $\sigma(a) \in N, \sigma(b) = x$, and $\ell(\sigma) = d(x, N)$ is called a segment from x to N. Show that σ is also a segment from N to any $\sigma(t), t < b$. Show that $\dot{\sigma}(a)$ is perpendicular to N.
   (b) Show that if N is a closed subspace of M and (M, g) is complete, then any point in M can be joined to N by a segment.
   (c) Show that in general there is an open neighborhood of N in M where all points are joined to N by segments.
   (d) Show that $d(\cdot, N)$ is smooth on a neighborhood of N and that the integral curves for its gradient are the geodesics that are perpendicular to N.

(11) Compute the cut locus on a square torus $\mathbb{R}^2/\mathbb{Z}^2$.

(12) Compute the cut locus on a sphere and real projective space with the constant curvature metrics.

(13) In a metric space (X, d) one can measure the length of continuous curves $\gamma : [a, b] \to X$ by

$$\ell(\gamma) = \sup \left\{ \sum d(\gamma(t_i), \gamma(t_{i+1})) : a = t_1 \leq t_2 \leq \cdots \leq t_{k-1} \leq t_k = b \right\}.$$

   (a) Show that a curve has finite length iff it is absolutely continuous. Hint: Use the characterization that $\gamma : [a, b] \to X$ is absolutely continuous if and only if for each $\varepsilon > 0$ there is a $\delta > 0$ so that $\sum d(\gamma(s_i), \gamma(s_{i+1})) \leq \varepsilon$ provided $\sum |s_i - s_{i+1}| \leq \delta$.
   (b) Show that this definition gives back our previous definition for smooth curves on Riemannian manifolds.
   (c) Let $\gamma : [a, b] \to M$ be an absolutely continuous curve whose length is $d(\gamma(a), \gamma(b))$. Show that $\gamma = \sigma \circ \varphi$ for some segment σ and reparametrization φ.

(14) Show that in a Riemannian manifold,

$$d\left(\exp_p(tv), \exp_p(tw)\right) = |t| \cdot |v - w| + O\left(t^2\right).$$

(15) Assume that we have coordinates x^i around a point $p \in (M, g)$ such that $x^i(p) = 0$ and $g_{ij} x^j = x^i$. Show that these must be exponential coordinates. Hint: Define

$$r = \sqrt{(x^1)^2 + \cdots + (x^n)^2}$$

and show that it is a smooth distance function away from p, and that the integral curves for the gradient are geodesics emanating from p.

(16) If $N_1, N_2 \subset M$ are totally geodesic submanifolds, show that each component of $N_1 \cap N_2$ is a submanifold which is totally geodesic. Hint: The potential tangent space at $p \in N_1 \cap N_2$ should be $T_pN_1 \cap T_pN_2$.
(17) Show that for a complete manifold the functional distance is the same as the distance. What about incomplete manifolds?
(18) Let $\gamma : [0,1] \to M$ be a geodesic such that $\exp_{\gamma(0)}$ is regular at all $t\dot{\gamma}(0)$, for $t \leq 1$. Show that γ is a local minimum for the energy functional. Hint: Show that the lift of γ via $\exp_{\gamma(0)}$ is a minimizing geodesic in a suitable metric.
(19) Show, using the exercises on Lie groups from chapters 1 and 2, that on a Lie group G with a bi-invariant metric the geodesics through the identity are exactly the homomorphisms $\mathbb{R} \to G$. Conclude that the Lie group exponential map coincides with the exponential map generated by the bi-invariant Riemannian metric. Hint: First show that homomorphisms $\mathbb{R} \to G$ are precisely the integral curves for left invariant vector fields through $e \in G$.
(20) Repeat the previous exercise assuming that the metric is a bi-invariant semi-Riemannian metric. Show that the matrix group $Gl_n(\mathbb{R})$ of invertible $n \times n$ matrices admits a bi-invariant semi-Riemannian metric. Hint: for $X, Y \in T_I Gl_n(\mathbb{R})$ define

$$g(X,Y) = -\mathrm{tr}(XY).$$

(21) Construct a Riemannian metric on the tangent bundle to a Riemannian manifold (M,g) such that $\pi : TM \to M$ is a Riemannian submersion and the metric restricted to the tangent spaces is the given Euclidean metric.
(22) For a Riemannian manifold (M,g) let FM be the frame bundle of M. This is a fiber bundle $\pi : FM \to M$ whose fiber over $p \in M$ consists of orthonormal bases for T_pM. Find a Riemannian metric on FM that makes π into a Riemannian submersion and such that the fibers are isometric to $O(n)$.
(23) Show that a Riemannian submersion is a submetry.
(24) (Hermann) Let $f : (M, \bar{g}) \to (N, g)$ be a Riemannian submersion.
   (a) Show that (N,g) is complete if $(M,\bar{g})$ is complete.
   (b) Show that f is a fibration if $(M,\bar{g})$ is complete i.e., for every $p \in N$ there is a neighborhood $p \in U$ such that $f^{-1}(U)$ is diffeomorphic to $U \times f^{-1}(p)$. Give a counterexample when $(M,\bar{g})$ is not complete.