Chapter Five

Matrix Manifolds: Second-Order Geometry

Many optimization algorithms make use of second-order information about the cost function. The archetypal second-order optimization algorithm is Newton's method. This method is an iterative method that seeks a critical point of the cost function f (i.e., a zero of $\operatorname{grad} f$) by selecting the update vector at x_k as the vector along which the directional derivative of $\operatorname{grad} f$ is equal to $-\operatorname{grad} f(x_k)$. The second-order information on the cost function is incorporated through the directional derivative of the gradient.

For a quadratic cost function in $\mathbb{R}^n$, Newton's method identifies a zero of the gradient in one step. For general cost functions, the method is not expected to converge in one step and may not even converge at all. However, the use of second-order information ensures that algorithms based on the Newton step display superlinear convergence (when they do converge) compared to the linear convergence obtained for algorithms that use only first-order information (see Section 4.5).

A Newton method on Riemannian manifolds will be defined and analyzed in Chapter 6. However, to provide motivation for the somewhat abstract theory that follows in this chapter, we begin by briefly recapping Newton's method in $\mathbb{R}^n$ and identify the blocks to generalizing the iteration to a manifold setting. An important step in the development is to provide a meaningful definition of the derivative of the gradient and, more generally, of vector fields; this issue is addressed in Section 5.2 by introducing the notion of an affine connection. An affine connection also makes it possible to define parallel translation, geodesics, and exponentials (Section 5.4). These tools are not mandatory in defining a Newton method on a manifold, but they are fundamental objects of Riemannian geometry, and we will make use of them in later chapters. On a Riemannian manifold, there is one preferred affine connection, termed the *Riemannian connection*, that admits elegant specialization to Riemannian submanifolds and Riemannian quotient manifolds (Section 5.3). The chapter concludes with a discussion of the concept of a Hessian on a manifold (Sections 5.5 and 5.6).

5.1 NEWTON'S METHOD IN $\mathbb{R}^N$

In its simplest formulation, Newton's method is an iterative method for finding a solution of an equation in one unknown. Let F be a smooth function from $\mathbb{R}$ to $\mathbb{R}$ and let x_* be a *zero* (or *root*) of F, i.e., $F(x_*) = 0$. From an initial point x_0 in $\mathbb{R}$, Newton's method constructs a sequence of iterates according to

$$x_{k+1} = x_k - \frac{F(x_k)}{F'(x_k)}, \tag{5.1}$$

where F' denotes the derivative of F. Graphically, x_{k+1} corresponds to the intersection of the tangent to the graph of F at x_k with the horizontal axis (see Figure 5.1). In other words, x_{k+1} is the zero of the first-order Taylor expansion of F around x_k. This is clearly seen when (5.1) is rewritten as

$$F(x_k) + F'(x_k)(x_{k+1} - x_k) = 0. \tag{5.2}$$

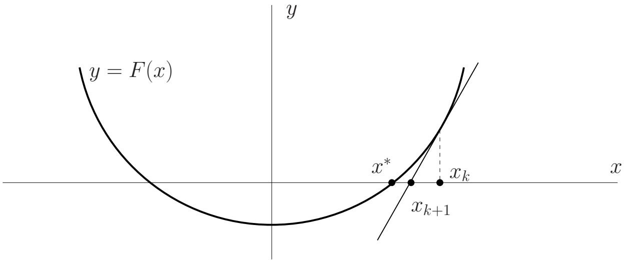

Figure 5.1 Newton's method in $\mathbb{R}$.

Let $G : \mathbb{R}^n \to \mathbb{R}^n : G(x) := x - F(x)/F'(x)$ be the iteration map from (5.1) and note that x_* is a fixed point of G. For a generic fixed point where $F(x_*) = 0$ and $F'(x_*) \neq 0$, the derivative

$$G'(x_*) = 1 - \frac{F'(x_*)}{F'(x_*)} + \frac{F(x_*)F''(x_*)}{(F'(x_*))^2} = 0,$$

and it follows that Newton's method is locally quadratically convergent to x_* (see Theorem 4.5.3).

Newton's method can be generalized to functions F from $\mathbb{R}^n$ to $\mathbb{R}^n$. Equation (5.2) becomes

$$F(x_k) + \mathrm{D}F\,(x_k)\,[x_{k+1} - x_k] = 0, \tag{5.3}$$

where $\mathrm{D}F\,(x)\,[z]$ denotes the *directional derivative* of F along z, defined by

$$\mathrm{D}F\,(x)\,[z] := \lim_{t \to 0} \frac{1}{t}(F(x + tz) - F(x)).$$

A generalization of the argument given above shows that Newton's method locally quadratically converges to isolated roots of F for which $DF(x_*)$ is full rank.

Newton's method is readily adapted to the problem of computing a critical point of a cost function f on $\mathbb{R}^n$. Simply take $F := \operatorname{grad} f$, where

$$\operatorname{grad} f(x) = (\partial_1 f(x), \ldots, \partial_n f(x))^T$$

is the Euclidean gradient of f. The iterates of Newton's method then converge locally quadratically to the isolated zeros of $\operatorname{grad} f$, which are the isolated critical points of f. Newton's equation then reads

$$\operatorname{grad} f(x_k) + \mathrm{D}(\operatorname{grad} f)\,(x_k)\,[x_{k+1} - x_k] = 0.$$

To generalize this approach to manifolds, we must find geometric analogs to the various components of the formula that defines the Newton iterate on $\mathbb{R}^n$. When f is a cost function an abstract Riemannian manifold, the Euclidean gradient naturally becomes the Riemannian gradient $\operatorname{grad} f$ defined in Section 3.6. The zeros of $\operatorname{grad} f$ are still the critical points of f. The difference $x_{k+1} - x_k$, which is no longer defined since the iterates x_{k+1} and x_k belong to the abstract manifold, is replaced by a tangent vector η_{x_k} in the tangent space at x_k. The new iterate x_{k+1} is obtained from η_{x_k} as $x_{k+1} = R_{x_k}(\eta_{x_k})$, where R is a retraction; see Section 4.1 for the notion of retraction. It remains to provide a meaningful definition for "$\mathrm{D}(\operatorname{grad} f)(x_k)[\eta_{x_k}]$".

More generally, for finding a zero of a tangent vector field ξ on a manifold, Newton's method takes the form

$$\begin{aligned} \xi_{x_k} + \text{“}\mathrm{D}\xi(x_k)[\eta_{x_k}]\text{”} &= 0, \\ x_{k+1} &= R_{x_k}(\eta_{x_k}). \end{aligned}$$

The only remaining task is to provide a geometric analog of the directional derivative of a vector field.

Recall that tangent vectors are defined as derivations of real functions: given a scalar function f and a tangent vector η at x, the real $\mathrm{D}f\,(x)\,[\eta]$ is defined as $\left.\frac{\mathrm{d}(f(\gamma(t)))}{\mathrm{d}t}\right|_{t=0}$, where γ is a curve representing η; see Section 3.5. If we try to apply the same concept to vector fields instead of scalar fields, we obtain

$$\left.\frac{\mathrm{d}\,\xi_{\gamma(t)}}{\mathrm{d}t}\right|_{t=0} = \lim_{t\to 0} \frac{\xi_{\gamma(t)} - \xi_{\gamma(0)}}{t}.$$

The catch is that the two vectors $\xi_{\gamma(t)}$ and $\xi_{\gamma(0)}$ belong to two different vector spaces $T_{\gamma(t)}\mathcal{M}$ and $T_{\gamma(0)}\mathcal{M}$, and there is in general no predefined correspondence between the vector spaces that allows us to compute the difference. Such a correspondence can be introduced by means of affine connections.

5.2 AFFINE CONNECTIONS

The definition of an affine connection on a manifold is one of the most fundamental concepts in differential geometry. An affine connection is an additional structure to the differentiable structure. Any manifold admits infinitely many different affine connections. Certain affine connections, however, may have particular properties that single them out as being the most appropriate for geometric analysis. In this section we introduce the concept of an affine connection from an abstract perspective and show how it generalizes the concept of a directional derivative of a vector field.

Let $\mathfrak{X}(\mathcal{M})$ denote the set of smooth vector fields on $\mathcal{M}$. An *affine connection* ∇ (pronounced "del" or "nabla") on a manifold $\mathcal{M}$ is a mapping

$$\nabla : \mathfrak{X}(\mathcal{M}) \times \mathfrak{X}(\mathcal{M}) \to \mathfrak{X}(\mathcal{M}),$$

which is denoted by $(\eta, \xi) \xrightarrow{\nabla} \nabla_\eta \xi$ and satisfies the following properties:

i)$\mathfrak{F}(\mathcal{M})$-linearity in η: $\nabla_{f\eta+g\chi}\xi = f\nabla_\eta\xi + g\nabla_\chi\xi$,
ii)$\mathbb{R}$-linearity in ξ: $\nabla_\eta(a\xi + b\zeta) = a\nabla_\eta\xi + b\nabla_\eta\zeta$,
iii)Product rule (Leibniz' law): $\nabla_\eta(f\xi) = (\eta f)\xi + f\nabla_\eta\xi$,

in which $\eta, \chi, \xi, \zeta \in \mathfrak{X}(\mathcal{M})$, $f, g \in \mathfrak{F}(\mathcal{M})$, and $a, b \in \mathbb{R}$. (Notice that ηf denotes the application of the vector field η to the function f, as defined in Section 3.5.4.) The vector field $\nabla_\eta\xi$ is called the *covariant derivative* of ξ with respect to η for the affine connection ∇.

In $\mathbb{R}^n$, the classical directional derivative defines an affine connection,

$$(\nabla_\eta\xi)_x = \lim_{t\to 0} \frac{\xi_{x+t\eta_x} - \xi_x}{t}, \tag{5.4}$$

called the *canonical (Euclidean) connection*. (This expression is well defined in view of the canonical identification $T_x\mathcal{E} \simeq \mathcal{E}$ discussed in Section 3.5.2, and it is readily checked that (5.4) satisfies all the properties of affine connections.) This fact, along with several properties discussed below, suggests that the covariant derivatives are a suitable generalization of the classical directional derivative.

Proposition 5.2.1 *Every (second-countable Hausdorff) manifold admits an affine connection.*

In fact, every manifold admits infinitely many affine connections, some of which may be computationally more tractable than others.

We first characterize all the possible affine connections on the linear manifold $\mathbb{R}^n$. Let $(e_1, \ldots, e_n)$ be the canonical basis of $\mathbb{R}^n$. If ∇ is a connection on $\mathbb{R}^n$, we have

$$\begin{aligned}\nabla_\eta\xi = \nabla_{\sum_i \eta^i e_i}\left(\sum_j \xi^j e_j\right) &= \sum_i \eta^i \nabla_{e_i}\left(\sum_j \xi^j e_j\right)\\ &= \sum_{i,j}\left(\eta^i\xi^j\nabla_{e_i}e_j + \eta^i\partial_i\xi^j e_j\right),\end{aligned}$$

where $\eta, \xi, e_i, \nabla_\eta\xi, \nabla_{e_i}e_j$ are all vector fields on $\mathbb{R}^n$. To define ∇, it suffices to specify the n^2 vector fields $\nabla_{e_i}e_j$, $i = 1, \ldots, n$, $j = 1, \ldots, n$. By convention, the kth component of $\nabla_{e_i}e_j$ in the basis $(e_1, \ldots, e_n)$ is denoted by Γ^k_{ij}. The n^3 real-valued functions Γ^k_{ij} on $\mathbb{R}^n$ are called *Christoffel symbols*. Each choice of smooth functions Γ^k_{ij} defines a different affine connection on $\mathbb{R}^n$. The Euclidean connection corresponds to the choice $\Gamma^k_{ij} \equiv 0$.

On an n-dimensional manifold $\mathcal{M}$, locally around any point x, a similar development can be based on a coordinate chart $(\mathcal{U}, \varphi)$. The following coordinate-based development shows how an affine connection can be defined on $\mathcal{U}$, at least in theory (in practice, the use of coordinates to define an affine connection can be cumbersome). The canonical vector e_i is replaced by the ith coordinate vector field E_i of $(\mathcal{U}, \varphi)$ which, at a point y of $\mathcal{U}$, is represented by the curve $t \mapsto \varphi^{-1}(\varphi(y) + te_i)$; in other words, given a real-valued function f defined on $\mathcal{U}$, $E_i f = \partial_i(f \circ \varphi^{-1})$. Thus, one has $\mathrm{D}\varphi(y)[(E_i)_y] = e_i$. We will also use the notation $\partial_i f$ for $E_i f$. A vector field ξ can be decomposed as $\xi = \sum_j \xi^j E_j$, where ξ^i, $i = 1, \ldots, d$, are real-valued functions on $\mathcal{U}$, i.e., elements of $\mathfrak{F}(\mathcal{U})$. Using the characteristic properties of affine connections, we obtain

$$\begin{aligned}\nabla_\eta \xi = \nabla_{\sum_i \eta^i E_i} \left(\sum_j \xi^j E_j \right) &= \sum_i \eta^i \nabla_{E_i} \left(\sum_j \xi^j E_j \right) \\ &= \sum_{i,j} \left(\eta^i \xi^j \nabla_{E_i} E_j + \eta^i \partial_i \xi^j E_j \right). \end{aligned} \tag{5.5}$$

It follows that the affine connection is fully specified once the n^2 vector fields $\nabla_{E_i} E_j$ are selected. We again use the Christoffel symbol Γ^k_{ij} to denote the kth component of $\nabla_{E_i} E_j$ in the basis $(E_1, \ldots, E_n)$; in other words,

$$\nabla_{E_i} E_j = \sum_k \Gamma^k_{ij} E_k.$$

The Christoffel symbols Γ^k_{ij} at a point x can be thought of as a table of n^3 real numbers that depend both on the point x in $\mathcal{M}$ and on the choice of the chart φ (for the same affine connection, different charts produce different Christoffel symbols). We thus have

$$\nabla_\eta \xi = \sum_{i,j,k} \left(\eta^i \xi^j \Gamma^k_{ij} E_k + \eta^i \partial_i \xi^j E_j \right).$$

A simple renaming of indices yields

$$\nabla_\eta \xi = \sum_{i,j,k} \eta^j \left(\xi^k \Gamma^i_{jk} + \partial_j \xi^i \right) E_i. \tag{5.6}$$

We also obtain a matrix expression as follows. Letting hat quantities denote the (column) vectors of components in the chart $(\mathcal{U}, \phi)$, we have

$$\widehat{\nabla_{\eta_x} \xi} = \hat{\Gamma}_{\hat{x}, \hat{\xi}} \hat{\eta}_{\hat{x}} + \mathrm{D}\hat{\xi}\left(\hat{x}\right)\left[\hat{\eta}_{\hat{x}}\right], \tag{5.7}$$

where $\hat{\Gamma}_{\hat{x}, \hat{\xi}}$ denotes the matrix whose (i, j) element is the real-valued function

$$\sum_k \left(\xi^k \Gamma^i_{jk} \right) \tag{5.8}$$

evaluated at x.

From the coordinate expression (5.5), one can deduce the following properties of affine connections.

1. Dependence on η_x. The vector field $\nabla_\eta \xi$ at a point x depends only on the value η_x of η at x. Thus, an affine connection at x is a mapping $T_x\mathcal{M} \times \mathfrak{X}(x) \to \mathfrak{X}(x) : (\eta_x, \xi) \mapsto \nabla_{\eta_x}\xi$, where $\mathfrak{X}(x)$ denotes the set of vector fields on $\mathcal{M}$ whose domain includes x.
2. Local dependence on ξ. In contrast, ξ_x does not provide enough information about the vector field ξ to compute $\nabla_\eta \xi$ at x. However, if the vector fields ξ and ζ agree on some neighborhood of x, then $\nabla_\eta \xi$ and $\nabla_\eta \zeta$ coincide at x. Moreover, given two affine connections ∇ and $\tilde{\nabla}$, $\nabla_\eta \xi - \tilde{\nabla}_\eta \xi$ at x depends only on the value ξ_x of ξ at x.
3. Uniqueness at zeros. Let ∇ and $\tilde{\nabla}$ be two affine connections on $\mathcal{M}$ and let ξ and η be vector fields on $\mathcal{M}$. Then, as a corollary of the previous property,

$$(\nabla_\eta \xi)_x = \left(\tilde{\nabla}_\eta \xi\right)_x \quad \text{if } \xi_x = 0.$$

This final property is particularly important in the convergence analysis of optimization algorithms around critical points of a cost function.

5.3 RIEMANNIAN CONNECTION

On an arbitrary (second-countable Hausdorff) manifold, there are infinitely many affine connections, and *a priori*, no one is better than the others. In contrast, on a vector space $\mathcal{E}$ there is a preferred affine connection, the canonical connection (5.4), which is simple to calculate and preserves the linear structure of the vector space. On an arbitrary Riemannian manifold, there is also a preferred affine connection, called the Riemannian or the Levi-Civita connection. This connection satisfies two properties (symmetry, and invariance of the Riemannian metric) that have a crucial importance, notably in relation to the notion of Riemannian Hessian. Moreover, the Riemannian connection on Riemannian submanifolds and Riemannian quotient manifolds admits a remarkable formulation in terms of the Riemannian connection in the structure space that makes it particularly suitable in the context of numerical algorithms. Furthermore, on a Euclidean space, the Riemannian connection reduces to the canonical connection—the classical directional derivative.

5.3.1 Symmetric connections

An affine connection is symmetric if its Christoffel symbols satisfy the symmetry property $\Gamma^k_{ij} = \Gamma^k_{ji}$. This definition is equivalent to a more abstract coordinate-free approach to symmetry that provides more insight into the underlying structure of the space.

To define symmetry of an affine connection in a coordinate-free manner, we will require the concept of a *Lie bracket* of two vector fields. Let ξ and ζ be vector fields on $\mathcal{M}$ whose domains meet on an open set $\mathcal{U}$. Recall that

$\mathfrak{F}(\mathcal{U})$ denotes the set of smooth real-valued functions whose domains include $\mathcal{U}$. Let $[\xi,\eta]$ denote the function from $\mathfrak{F}(\mathcal{U})$ into itself defined by

$$[\xi,\zeta]f := \xi(\zeta f) - \zeta(\xi f). \tag{5.9}$$

It is easy to show that $[\xi,\zeta]$ is $\mathbb{R}$-linear,

$$[\xi,\eta](af+bg) = a[\xi,\eta]f + b[\xi,\eta]g,$$

and satisfies the product rule (Leibniz' law),

$$[\xi,\eta](fg) = f([\xi,\eta]g) + ([\xi,\eta]f)g.$$

Therefore, $[\xi,\zeta]$ is a derivation and defines a tangent vector field, called the *Lie bracket* of ξ and ζ.

An affine connection ∇ on a manifold $\mathcal{M}$ is said to be *symmetric* when

$$\nabla_\eta \xi - \nabla_\xi \eta = [\eta,\xi] \tag{5.10}$$

for all $\eta,\xi \in \mathfrak{X}(\mathcal{M})$.

Given a chart $(\mathcal{U},\varphi)$, denoting by E_i the ith coordinate vector field, we have, for a symmetric connection ∇,

$$\nabla_{E_i}E_j - \nabla_{E_j}E_i = [E_i,E_j] = 0$$

since $[E_i,E_j]f = \partial_i\partial_j f - \partial_j\partial_i f = 0$ for all $f \in \mathfrak{F}(\mathcal{M})$. It follows that $\Gamma^k_{ij} = \Gamma^k_{ji}$ for every symmetric connection. Conversely, it is easy to show that connections satisfying $\Gamma^k_{ij} = \Gamma^k_{ji}$ are symmetric in the sense of (5.10) by expanding in local coordinates.

5.3.2 Definition of the Riemannian connection

The following result is sometimes referred to as the fundamental theorem of Riemannian geometry. Let $\langle\cdot,\cdot\rangle$ denote the Riemannian metric.

Theorem 5.3.1 (Levi-Civita) *On a Riemannian manifold $\mathcal{M}$ there exists a unique affine connection ∇ that satisfies*

(i) $\nabla_\eta\xi - \nabla_\xi\eta = [\eta,\xi]$ *(symmetry), and*
(ii) $\chi\langle\eta,\xi\rangle = \langle\nabla_\chi\eta,\xi\rangle + \langle\eta,\nabla_\chi\xi\rangle$ *(compatibility with the Riemannian metric),*

for all $\chi,\eta,\xi \in \mathfrak{X}(\mathcal{M})$. This affine connection ∇, called the Levi-Civita connection *or the* Riemannian connection *of $\mathcal{M}$, is characterized by the* Koszul formula

$$2\langle\nabla_\chi\eta,\xi\rangle = \chi\langle\eta,\xi\rangle + \eta\langle\xi,\chi\rangle - \xi\langle\chi,\eta\rangle - \langle\chi,[\eta,\xi]\rangle + \langle\eta,[\xi,\chi]\rangle + \langle\xi,[\chi,\eta]\rangle. \tag{5.11}$$

Recall that for vector fields $\eta,\xi,\chi \in \mathfrak{X}(\mathcal{M})$, $\langle\eta,\xi\rangle$ is a real-valued function on $\mathcal{M}$ and $\chi\langle\eta,\xi\rangle$ is the real-valued function given by the application of the vector field (i.e., derivation) χ to $\langle\eta,\xi\rangle$.)

Since the Riemannian connection is symmetric, it follows that the Christoffel symbols of the Riemannian connection satisfy $\Gamma^k_{ij} = \Gamma^k_{ji}$. Moreover, it follows from the Koszul formula (5.11) that the Christoffel symbols for the Riemannian connection are related to the coefficients of the metric by the formula

$$\Gamma_{ij}^{k} = \frac{1}{2} \sum_{\ell} g^{k\ell} \left(\partial_i g_{\ell j} + \partial_j g_{\ell i} - \partial_\ell g_{ij} \right), \tag{5.12}$$

where $g^{k\ell}$ denotes the matrix inverse of $g_{k\ell}$, i.e., $\sum_i g^{ki} g_{i\ell} = \delta_\ell^k$. In theory, the formula (5.12) provides a means to compute the Riemannian connection. However, working in coordinates can be cumbersome in practice, and we will use a variety of tricks to avoid using (5.12) as a computational formula.

Note that on a Euclidean space, the Riemannian connection reduces to the canonical connection (5.4). A way to see this is that, in view of (5.12), the Christoffel symbols vanish since the metric is constant.

5.3.3 Riemannian connection on Riemannian submanifolds

Let $\mathcal{M}$ be a Riemannian submanifold of a Riemannian manifold $\overline{\mathcal{M}}$. By definition, the Riemannian metric on the submanifold $\mathcal{M}$ is obtained by restricting to $\mathcal{M}$ the Riemannian metric on $\overline{\mathcal{M}}$; therefore we use the same notation $\langle\cdot,\cdot\rangle$ for both. Let ∇ denote the Riemannian connection of $\mathcal{M}$, and $\overline{\nabla}$ the Riemannian connection of $\overline{\mathcal{M}}$. Let $\mathfrak{X}(\mathcal{M})$ denote the set of vector fields on $\mathcal{M}$, and $\mathfrak{X}(\overline{\mathcal{M}})$ the set of vector fields on $\overline{\mathcal{M}}$.

Given $\eta_x \in T_x\mathcal{M}$ and $\xi \in \mathfrak{X}(\mathcal{M})$, we begin by defining the object $\overline{\nabla}_\eta \xi$. To this end, since $T_x\mathcal{M}$ is a subspace of $T_x\overline{\mathcal{M}}$, let $\overline{\eta}_x$ be η_x viewed as an element of $T_x\overline{\mathcal{M}}$; moreover, let $\overline{\xi}$ be a smooth local extension of ξ over a coordinate neighborhood $\mathcal{U}$ of x in $\overline{\mathcal{M}}$. Then define

$$\overline{\nabla}_{\eta_x} \xi := \overline{\nabla}_{\overline{\eta}_x} \overline{\xi}. \tag{5.13}$$

This expression does not depend on the local extension of ξ. However, in general, $\overline{\nabla}_{\eta_x} \xi$ does not lie in $T_x\mathcal{M}$, as illustrated in Figure 5.2. Hence the restriction of $\overline{\nabla}$ to $\mathcal{M}$, as defined in (5.13), does not qualify as a connection on $\mathcal{M}$.

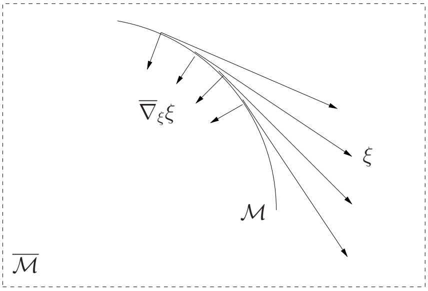

Figure 5.2 Riemannian connection $\overline{\nabla}$ in a Euclidean space $\overline{\mathcal{M}}$ applied to a tangent vector field ξ to a circle. We observe that $\overline{\nabla}_\xi \xi$ is not tangent to the circle.

Recall from Section 3.6.1 that, using the Riemannian metric on $\overline{\mathcal{M}}$, each tangent space $T_x\overline{\mathcal{M}}$ can be decomposed as the direct sum of $T_x\mathcal{M}$ and its orthogonal complement $(T_x\mathcal{M})^\perp$, called the normal space to the Riemannian submanifold $\mathcal{M}$ at x. Every vector $\xi_x \in T_x\overline{\mathcal{M}}$, $x \in \mathcal{M}$, has a unique decomposition

$$\xi_x = \mathrm{P}_x\xi_x + \mathrm{P}_x^\perp\xi_x,$$

where $\mathrm{P}_x\xi_x$ belongs to $T_x\mathcal{M}$ and $\mathrm{P}_x^\perp\xi_x$ belongs to $(T_x\mathcal{M})^\perp$. We have the following fundamental result.

Proposition 5.3.2 *Let $\mathcal{M}$ be a Riemannian submanifold of a Riemannian manifold $\overline{\mathcal{M}}$ and let ∇ and $\overline{\nabla}$ denote the Riemannian connections on $\mathcal{M}$ and $\overline{\mathcal{M}}$. Then*

$$\nabla_{\eta_x}\xi = \mathrm{P}_x\overline{\nabla}_{\eta_x}\xi \tag{5.14}$$

for all $\eta_x \in T_x\mathcal{M}$ and $\xi \in \mathfrak{X}(\mathcal{M})$.

This result is particularly useful when $\mathcal{M}$ is a Riemannian submanifold of a Euclidean space; then (5.14) reads

$$\nabla_{\eta_x}\xi = \mathrm{P}_x\left(\mathrm{D}\xi\left(x\right)\left[\eta_x\right]\right), \tag{5.15}$$

i.e., a classical directional derivative followed by an orthogonal projection.

Example 5.3.1 ***The sphere*** S^{n-1}

On the sphere S^{n-1} viewed as a Riemannian submanifold of the Euclidean space $\mathbb{R}^n$, the projection P_x is given by

$$\mathrm{P}_x\xi = (I - xx^T)\xi$$

and the Riemannian connection is given by

$$\nabla_{\eta_x}\xi = (I - xx^T)\,\mathrm{D}\xi\,(x)\,[\eta_x] \tag{5.16}$$

for all $x \in S^{n-1}$, $\eta_x \in T_xS^{n-1}$, and $\xi \in \mathfrak{X}(S^{n-1})$. A practical application of this formula is presented in Section 6.4.1.

Example 5.3.2 ***The orthogonal Stiefel manifold*** $\mathrm{St}(p, n)$

On the Stiefel manifold $\mathrm{St}(p, n)$ viewed as a Riemannian submanifold of the Euclidean space $\mathbb{R}^{n\times p}$, the projection P_X is given by

$$\mathrm{P}_X\xi = (I - XX^T)\xi + X\,\mathrm{skew}(X^T\xi)$$

and the Riemannian connection is given by

$$\nabla_{\eta_X}\xi = \mathrm{P}_X(\mathrm{D}\xi\,(x)\,[\eta_X]) \tag{5.17}$$

for all $X \in \mathrm{St}(p, n)$, $\eta_X \in T_X\,\mathrm{St}(p, n)$, and $\xi \in \mathfrak{X}(\mathrm{St}(p, n))$.

5.3.4 Riemannian connection on quotient manifolds

Let $\overline{\mathcal{M}}$ be a Riemannian manifold with a Riemannian metric $\overline{g}$ and let $\mathcal{M} = \overline{\mathcal{M}}/\sim$ be a Riemannian quotient manifold of $\overline{\mathcal{M}}$, i.e., $\mathcal{M}$ is endowed with a manifold structure and a Riemannian metric g that turn the natural projection $\pi : \overline{\mathcal{M}} \to \mathcal{M}$ into a Riemannian submersion. As in Section 3.6.2, the horizontal space $\mathcal{H}_y$ at a point $y \in \overline{\mathcal{M}}$ is defined as the orthogonal complement of the vertical space, and $\overline{\xi}$ denotes the horizontal lift of a tangent vector ξ.

Proposition 5.3.3 *Let $\mathcal{M} = \overline{\mathcal{M}}/\sim$ be a Riemannian quotient manifold and let ∇ and $\overline{\nabla}$ denote the Riemannian connections on $\mathcal{M}$ and $\overline{\mathcal{M}}$. Then*

$$\overline{\nabla_\eta \xi} = \mathrm{P}^h \left(\overline{\nabla}_{\overline{\eta}} \overline{\xi}\right) \tag{5.18}$$

for all vector fields ξ and η on $\mathcal{M}$, where P^h denotes the orthogonal projection onto the horizontal space.

This is a very useful result, as it provides a practical way to compute covariant derivatives in the quotient space. The result states that the horizontal lift of the covariant derivative of ξ with respect to η is given by the horizontal projection of the covariant derivative of the horizontal lift of ξ with respect to the horizontal lift of η.

If the structure space $\overline{\mathcal{M}}$ is (an open subset of) a Euclidean space, then formula (5.18) simply becomes

$$\overline{\nabla_\eta \xi} = \mathrm{P}^h \left(\mathrm{D}\overline{\xi}\,[\overline{\eta}]\right).$$

In some practical cases, $\overline{\mathcal{M}}$ is a vector space endowed with a Riemannian metric $\overline{g}$ that is not constant (hence $\overline{\mathcal{M}}$ is not a Euclidean space) but that is nevertheless *horizontally invariant*, namely,

$$\mathrm{D}\left(\overline{g}(\nu, \lambda)\right)(y)\,[\eta_y] = \overline{g}(\mathrm{D}\nu\,(y)\,[\eta_y]\,, \lambda_y) + \overline{g}(\nu_y, \mathrm{D}\lambda\,(y)\,[\eta_y])$$

for all $y \in \overline{\mathcal{M}}$, all $\eta_y \in \mathcal{H}_y$, and all horizontal vector fields ν, λ on $\overline{\mathcal{M}}$. In this case, the next proposition states that the Riemannian connection on the quotient is still a classical directional derivative followed by a projection.

Proposition 5.3.4 *Let $\mathcal{M}$ be a Riemannian quotient manifold of a vector space $\overline{\mathcal{M}}$ endowed with a horizontally invariant Riemannian metric and let ∇ denote the Riemannian connection on $\mathcal{M}$. Then*

$$\overline{\nabla_\eta \xi} = \mathrm{P}^h \left(\mathrm{D}\overline{\xi}\,[\overline{\eta}]\right)$$

for all vector fields ξ and η on $\mathcal{M}$.

Proof. Let $\overline{g}(\cdot,\cdot) = \langle \cdot,\cdot \rangle$ denote the Riemannian metric on $\overline{\mathcal{M}}$ and let $\overline{\nabla}$ denote the Riemannian connection of $\overline{\mathcal{M}}$. Let χ, ν, λ be horizontal vector fields on $\overline{\mathcal{M}}$. Notice that since $\overline{\mathcal{M}}$ is a vector space, one has $[\nu, \lambda] = \mathrm{D}\lambda[\nu] - \mathrm{D}\nu[\lambda]$, and likewise for permutations between χ, ν, and λ. Moreover, since it is assumed that $\overline{g}$ is horizontally invariant, it follows that

$\mathrm{D}\overline{g}(\nu,\lambda)[\chi] = \overline{g}(\mathrm{D}\nu[\chi],\lambda) + \overline{g}(\nu,\mathrm{D}\lambda[\chi])$; and likewise for permutations. Using these identities, it follows from Koszul's formula (5.11) that

$$\begin{aligned} 2\langle \overline{\nabla}_\chi \nu, \lambda\rangle &= \chi\langle \nu,\lambda\rangle + \nu\langle \lambda,\chi\rangle - \lambda\langle \chi,\nu\rangle + \langle \lambda,[\chi,\nu]\rangle + \langle \nu,[\lambda,\chi]\rangle - \langle \chi,[\nu,\lambda]\rangle \\ &= 2\overline{g}(\mathrm{D}\nu[\chi],\lambda), \end{aligned}$$

hence $\mathrm{P}^h(\overline{\nabla}_\chi \nu) = \mathrm{P}^h(\mathrm{D}\nu[\chi])$. The result follows from Proposition 5.3.3. □

Example 5.3.3 ***The Grassmann manifold***

We follow up on the example in Section 3.6.2. Recall that the Grassmann manifold $\mathrm{Grass}(p,n)$ *was viewed as a Riemannian quotient manifold of* $(\mathbb{R}_*^{n\times p}, \overline{g})$ *with*

$$\overline{g}_Y(Z_1,Z_2) = \mathrm{tr}\left((Y^TY)^{-1}Z_1^TZ_2\right). \tag{5.19}$$

The horizontal distribution is

$$\mathcal{H}_Y = \{Z \in \mathbb{R}^{n\times p} : Y^TZ = 0\} \tag{5.20}$$

and the projection onto the horizontal space is given by

$$\mathrm{P}_Y^hZ = (I - Y(Y^TY)^{-1}Y^T)Z. \tag{5.21}$$

It is readily checked that, for all horizontal vectors $Z \in \mathcal{H}_Y$*, it holds that*

$$\begin{aligned} \mathrm{D}\overline{g}(\overline{\xi},\overline{\zeta})\,(Y)\,[Z] &= \mathrm{D}_Y(\mathrm{tr}((Y^TY)^{-1}(\overline{\xi}_Y)^T\overline{\zeta}_Y))\,(Y)\,[Z] \\ &= \overline{g}(\mathrm{D}\overline{\xi}\,(Y)\,[Z]\,,\overline{\zeta}_Y) + \overline{g}(\overline{\xi}_Y,\mathrm{D}\overline{\zeta}\,(Y)\,[Z]) \end{aligned}$$

since $Y^TZ = 0$ *for all* $Z \in \mathcal{H}_Y$*. The Riemannian metric* $\overline{g}$ *is thus horizontally invariant. Consequently, we can apply the formula for the Riemannian connection on a Riemannian quotient of a manifold with a horizontally invariant metric (Proposition 5.3.4) and obtain*

$$\overline{\nabla_\eta \xi} = \mathrm{P}_Y^h\left(\mathrm{D}\overline{\xi}\,(Y)\,[\overline{\eta}_Y]\right). \tag{5.22}$$

We refer the reader to Section 6.4.2 for a practical application of this formula.

5.4 GEODESICS, EXPONENTIAL MAPPING, AND PARALLEL TRANSLATION

Geodesics on manifolds generalize the concept of straight lines in $\mathbb{R}^n$. A geometric definition of a straight line in $\mathbb{R}^n$ is that it is the image of a curve γ with zero acceleration; i.e.,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\gamma(t) = 0$$

for all t.

On manifolds, we have already introduced the notion of a tangent vector $\dot{\gamma}(t)$, which can be interpreted as the *velocity* of the curve γ at t. The mapping $t \mapsto \dot{\gamma}(t)$ defines the *velocity vector field* along γ. Next we define the acceleration vector field along γ.

Let $\mathcal{M}$ be a manifold equipped with an affine connection ∇ and let γ be a curve in $\mathcal{M}$ with domain $I \subseteq \mathbb{R}$. A *vector field on the curve* γ smoothly assigns to each $t \in I$ a tangent vector to $\mathcal{M}$ at $\gamma(t)$. For example, given any vector field ξ on $\mathcal{M}$, the mapping $t \mapsto \xi_{\gamma(t)}$ is a vector field on γ. The velocity vector field $t \mapsto \dot{\gamma}(t)$ is also a vector field on γ. The set of all (smooth) vector fields on γ is denoted by $\mathfrak{X}(\gamma)$. It can be shown that there is a unique function $\xi \mapsto \frac{\mathrm{D}}{\mathrm{d}t}\,\xi$ from $\mathfrak{X}(\gamma)$ to $\mathfrak{X}(\gamma)$ such that

1. $\frac{\mathrm{D}}{\mathrm{d}t}(a\xi + b\zeta) = a\,\frac{\mathrm{D}}{\mathrm{d}t}\,\xi + b\,\frac{\mathrm{D}}{\mathrm{d}t}\,\zeta \quad (a, b \in \mathbb{R})$,
2. $\frac{\mathrm{D}}{\mathrm{d}t}(f\xi) = f'\xi + f\,\frac{\mathrm{D}}{\mathrm{d}t}\,\xi \quad (f \in \mathfrak{F}(I))$,
3. $\frac{\mathrm{D}}{\mathrm{d}t}(\eta \circ \gamma)(t) = \nabla_{\dot{\gamma}(t)}\eta \quad (t \in I,\ \eta \in \mathfrak{X}(\mathcal{M}))$.

The *acceleration vector field* $\frac{\mathrm{D}^2}{\mathrm{d}t^2}\,\gamma$ on γ is defined by

$$\frac{\mathrm{D}^2}{\mathrm{d}t^2}\,\gamma := \frac{\mathrm{D}}{\mathrm{d}t}\,\dot{\gamma}. \tag{5.23}$$

Note that the acceleration depends on the choice of the affine connection, while the velocity $\dot{\gamma}$ does not. Specifically, in a coordinate chart $(\mathcal{U}, \varphi)$, using the notation $(x^1(t), \dots, x^n(t)) := \varphi(\gamma(t))$, the velocity $\dot{\gamma}$ simply reads $\frac{\mathrm{d}}{\mathrm{d}t}\,x^k$, which does not depend on the Christoffel symbol; on the other hand, the acceleration $\frac{\mathrm{D}^2}{\mathrm{d}t^2}\,\gamma$ reads

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\,x^k + \sum_{i,j} \Gamma^k_{ij}(\gamma)\,\frac{\mathrm{d}}{\mathrm{d}t}\,x^i\,\frac{\mathrm{d}}{\mathrm{d}t}\,x^j,$$

where $\Gamma^k_{ij}(\gamma(t))$ are the Christoffel symbols, evaluated at the point $\gamma(t)$, of the affine connection in the chart $(\mathcal{U}, \varphi)$.

A *geodesic* γ on a manifold $\mathcal{M}$ endowed with an affine connection ∇ is a curve with zero acceleration:

$$\frac{\mathrm{D}^2}{\mathrm{d}t^2}\,\gamma(t) = 0 \tag{5.24}$$

for all t in the domain of γ. Note that different affine connections produce different geodesics.

For every $\xi \in T_x\mathcal{M}$, there exists an interval I about 0 and a unique geodesic $\gamma(t; x, \xi) : I \to \mathcal{M}$ such that $\gamma(0) = x$ and $\dot{\gamma}(0) = \xi$. Moreover, we have the homogeneity property $\gamma(t; x, a\xi) = \gamma(at; x, \xi)$. The mapping

$$\mathrm{Exp}_x : T_x\mathcal{M} \to \mathcal{M} : \xi \mapsto \mathrm{Exp}_x\xi = \gamma(1; x, \xi)$$

is called the *exponential map at* x. When the domain of definition of Exp_x is the whole $T_x\mathcal{M}$ for all $x \in \mathcal{M}$, the manifold $\mathcal{M}$ (endowed with the affine connection ∇) is termed *(geodesically) complete*.

It can be shown that Exp_x defines a diffeomorphism (smooth bijection) of a neighborhood $\widehat{\mathcal{U}}$ of the origin $0_x \in T_x\mathcal{M}$ onto a neighborhood $\mathcal{U}$ of $x \in \mathcal{M}$. If, moreover, $\widehat{\mathcal{U}}$ is *star-shaped* (i.e., $\xi \in \widehat{\mathcal{U}}$ implies $t\xi \in \widehat{\mathcal{U}}$ for all $0 \le t \le 1$), then $\mathcal{U}$ is called a *normal neighborhood* of x.

We can further define

$$\mathrm{Exp} : T\mathcal{M} \to \mathcal{M} : \xi \mapsto \mathrm{Exp}_x\xi,$$

where x is the foot of ξ. The mapping Exp is differentiable, and $\mathrm{Exp}_x 0_x = x$ for all $x \in \mathcal{M}$. Further, it can be shown that $\mathrm{D}\mathrm{Exp}_x(0_x)[\xi] = \xi$ (with the canonical identification $T_{0_x}T_x\mathcal{M} \simeq T_x\mathcal{M}$). This yields the following result.

Proposition 5.4.1 *Let $\mathcal{M}$ be a manifold endowed with an affine connection ∇. The exponential map on $\mathcal{M}$ induced by ∇ is a retraction, termed the* exponential retraction.

The exponential mapping is an important object in differential geometry, and it has featured heavily in previously published geometric optimization algorithms on manifolds. It generalizes the concept of moving “straight” in the direction of a tangent vector and is a natural way to update an iterate given a search direction in the tangent space. However, computing the exponential is, in general, a computationally daunting task. Computing the exponential amounts to evaluating the $t = 1$ point on the curve defined by the second-order ordinary differential equation (5.24). In a coordinate chart $(\mathcal{U}, \varphi)$, (5.24) reads

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} x^k + \sum_{i,j} \Gamma^k_{ij}(\gamma) \frac{\mathrm{d}}{\mathrm{d}t} x^i \frac{\mathrm{d}}{\mathrm{d}t} x^j = 0, \quad k = 1, \ldots, n,$$

where $(x^1(t), \ldots, x^n(t)) := \varphi(\gamma(t))$ and Γ^k_{ij} are the Christoffel symbols of the affine connection in the chart $(\mathcal{U}, \varphi)$. In general, such a differential equation does not admit a closed-form solution, and numerically computing the geodesic involves computing an approximation to the Christoffel symbols if they are not given in closed form and then approximating the geodesic using a numerical integration scheme. The theory of general retractions is introduced to provide an alternative to the exponential in the design of numerical algorithms that retains the key properties that ensure convergence results.

Assume that a basis is given for the vector space $T_y\mathcal{M}$ and let $\mathcal{U}$ be a normal neighborhood of y. Then a chart can be defined that maps $x \in \mathcal{U}$ to the components of the vector $\xi \in T_y\mathcal{M}$ satisfying $\mathrm{Exp}_y \xi = x$. The coordinates defined by this mapping are called *normal coordinates*.

We also point out the following fundamental result of differential geometry: if $\mathcal{M}$ is a Riemannian manifold, a curve with minimal length between two points of $\mathcal{M}$ is always a monotone reparameterization of a geodesic relative to the Riemannian connection. These curves are called *minimizing geodesics*.

Example 5.4.1 ***Sphere***

Consider the unit sphere S^{n-1} endowed with the Riemannian metric (3.33) *obtained by embedding S^{n-1} in $\mathbb{R}^n$ and with the associated Riemannian connection* (5.16)*. Geodesics $t \mapsto x(t)$ are expressed as a function of $x(0) \in S^{n-1}$ and $\dot{x}(0) \in T_{x(0)}S^{n-1}$ as follows (using the canonical inclusion of $T_{x_0}S^{n-1}$ in $\mathbb{R}^n$):*

$$x(t) = x(0)\cos(\|\dot{x}(0)\|t) + \dot{x}(0)\frac{1}{\|\dot{x}(0)\|}\sin(\|\dot{x}(0)\|t). \qquad (5.25)$$

(Indeed, it is readily checked that $\frac{\mathrm{D}^2}{\mathrm{d}t^2} x(t) = (I - x(t)x(t)^T)\frac{\mathrm{d}^2}{\mathrm{d}t^2} x(t) = -(I - x(t)x(t)^T)\|\dot{x}(0)\|^2 x(t) = 0$.)

Example 5.4.2 ***Orthogonal Stiefel manifold***

Consider the orthogonal Stiefel manifold $\mathrm{St}(p,n)$ *endowed with its Riemannian metric* (3.34) *inherited from the embedding in* $\mathbb{R}^{n\times p}$ *and with the corresponding Riemannian connection* ∇. *Geodesics* $t \mapsto X(t)$ *are expressed as a function of* $X(0) \in \mathrm{St}(p,n)$ *and* $\dot{X}(0) \in T_{X(0)}\,\mathrm{St}(p,n)$ *as follows (using again the canonical inclusion of* $T_{X(0)}\,\mathrm{St}(p,n)$ *in* $\mathbb{R}^{n\times p}$*):*

$$X(t) = \begin{bmatrix} X(0) & \dot{X}(0) \end{bmatrix} \exp\left(t \begin{bmatrix} A(0) & -S(0) \\ I & A(0) \end{bmatrix} \right) \begin{bmatrix} I \\ 0 \end{bmatrix} \exp(-A(0)t), \qquad (5.26)$$

where $A(t) := X^T(t)\dot{X}(t)$ *and* $S(t) := \dot{X}^T(t)\dot{X}(t)$. *It can be shown that* A *is an invariant of the trajectory, i.e.,* $A(t) = A(0)$ *for all* t, *and that* $S(t) = e^{At}S(0)e^{-At}$.

Example 5.4.3 ***Grassmann manifold***

Consider the Grassmann manifold $\mathrm{Grass}(p,n)$ *viewed as a Riemannian quotient manifold of* $\mathbb{R}_*^{n\times p}$ *with the associated Riemannian connection* (5.22). *Then*

$$\mathcal{Y}(t) = \mathrm{span}(Y_0(Y_0^TY_0)^{-1/2}V\cos(\Sigma t) + U\sin(\Sigma t)) \qquad (5.27)$$

is the geodesic satisfying $\mathcal{Y}(0) = \mathrm{span}(Y_0)$ *and* $\overline{\dot{\mathcal{Y}}(0)}_{Y_0} = U\Sigma V^T$, *where* $U\Sigma V^T$ *is a thin singular value decomposition, i.e.,* U *is* $n\times p$ *orthonormal,* V *is* $p \times p$ *orthonormal, and* Σ *is* $p\times p$ *diagonal with nonnegative elements. Note that choosing* Y_0 *orthonormal simplifies the expression* (5.27).

Let $\mathcal{M}$ be a manifold endowed with an affine connection ∇. A vector field ξ on a curve γ satisfying $\frac{\mathrm{D}}{\mathrm{d}t}\xi = 0$ is called *parallel*. Given $a \in \mathbb{R}$ in the domain of γ and $\xi_{\gamma(a)} \in T_{\gamma(a)}\mathcal{M}$, there is a unique parallel vector field ξ on γ such that $\xi(a) = \xi_{\gamma(a)}$. The operator $P_\gamma^{b\leftarrow a}$ sending $\xi(a)$ to $\xi(b)$ is called *parallel translation along* γ. In other words, we have

$$\frac{\mathrm{D}}{\mathrm{d}t}\left(P_\gamma^{t\leftarrow a}\xi(a)\right) = 0.$$

If $\mathcal{M}$ is a Riemannian manifold and ∇ is the Riemannian connection, then the parallel translation induced by ∇ is an isometry.

Much like the exponential mapping is a particular retraction, the parallel translation is a particular instance of a more general concept termed vector transport, introduced in Section 8.1. More information on vector transport by parallel translation, including formulas for parallel translation on special manifolds, can be found in Section 8.1.1. The machinery of retraction (to replace geodesic interpolation) and vector transport (to replace parallel translation) are two of the key insights in obtaining competitive numerical algorithms based on a geometric approach.

5.5 RIEMANNIAN HESSIAN OPERATOR

We conclude this chapter with a discussion of the notion of a Hessian. The Hessian matrix of a real-valued function f on $\mathbb{R}^n$ at a point $x \in \mathbb{R}^n$ is classically defined as the matrix whose (i,j) element (ith row and jth column) is given by $\partial^2_{ij} f(x) = \frac{\partial^2}{\partial_i \partial_j} f(x)$. To formalize this concept on a manifold we need to think of the Hessian as an operator acting on geometric objects and returning geometric objects. For a real-valued function f on an abstract Euclidean space $\mathcal{E}$, the Hessian operator at x is the (linear) operator from $\mathcal{E}$ to $\mathcal{E}$ defined by

$$\operatorname{Hess} f(x)[z] := \sum_{ij} \partial^2_{ij} \hat{f}(x^1, \ldots, x^n) z^j e_i, \tag{5.28}$$

where $(e_i)_{i=1,\ldots,n}$ is an orthonormal basis of $\mathcal{E}$, $z = \sum_j z^j e_j$ and $\hat{f}$ is the function on $\mathbb{R}^n$ defined by $\hat{f}(x^1, \ldots, x^n) = f(x^1 e_1 + \cdots + x^n e_n)$. It is a standard real analysis exercise to show that the definition does not depend on the choice of the orthonormal basis. Equivalently, the Hessian operator of f at x can be defined as the operator from $\mathcal{E}$ to $\mathcal{E}$ that satisfies, for all $y, z \in \mathcal{E}$,

1. $\langle \operatorname{Hess} f(x)[y], y\rangle = \mathrm{D}^2 f(x)[y, y] := \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(x + ty)\Big|_{t=0}$,
2. $\langle \operatorname{Hess} f(x)[y], z\rangle = \langle y, \operatorname{Hess} f(x)[z]\rangle$ (symmetry).

On an arbitrary Riemannian manifold, the Hessian operator is generalized as follows.

Definition 5.5.1 *Given a real-valued function f on a Riemannian manifold $\mathcal{M}$, the* Riemannian Hessian *of f at a point x in $\mathcal{M}$ is the linear mapping* $\operatorname{Hess} f(x)$ *of $T_x\mathcal{M}$ into itself defined by*

$$\operatorname{Hess} f(x)[\xi_x] = \nabla_{\xi_x} \operatorname{grad} f$$

for all ξ_x in $T_x\mathcal{M}$, where ∇ is the Riemannian connection on $\mathcal{M}$.

If $\mathcal{M}$ is a Euclidean space, this definition reduces to (5.28). (A justification for the name "Riemannian Hessian" is that the function $m_x(y) := f(x) + \langle \operatorname{grad} f(x), \operatorname{Exp}_x^{-1}(y)\rangle_x + \frac{1}{2}\langle \operatorname{Hess} f(x)[\operatorname{Exp}_x^{-1}(y)], \operatorname{Exp}_x^{-1}(y)\rangle$ is a second-order model of f around x; see Section 7.1.)

Proposition 5.5.2 *The Riemannian Hessian satisfies the formula*

$$\langle \operatorname{Hess} f[\xi], \eta\rangle = \xi(\eta f) - (\nabla_\xi \eta) f \tag{5.29}$$

for all $\xi, \eta \in \mathfrak{X}(\mathcal{M})$.

Proof. We have $\langle \operatorname{Hess} f[\xi], \eta\rangle = \langle \nabla_\xi \operatorname{grad} f, \eta\rangle$. Since the Riemannian connection leaves the Riemannian metric invariant, this is equal to $\xi\langle \operatorname{grad} f, \eta\rangle - \langle \operatorname{grad} f, \nabla_\xi \eta\rangle$. By definition of the gradient, this yields $\xi(\eta f) - (\nabla_\xi \eta) f$. □

Proposition 5.5.3 *The Riemannian Hessian is symmetric (in the sense of the Riemannian metric). That is,*

$$\langle \operatorname{Hess} f[\xi], \eta\rangle = \langle \xi, \operatorname{Hess} f[\eta]\rangle$$

for all $\xi, \eta \in \mathfrak{X}(\mathcal{M})$.

Proof. By the previous proposition, the left-hand side is equal to $\xi(\eta f) - (\nabla_\xi \eta) f$ and the right-hand side is equal to $\langle \operatorname{Hess} f(x)[\eta], \xi\rangle = \eta(\xi f) - (\nabla_\eta \xi) f$. Using the symmetry property (5.10) of the Riemannian connection on the latter expression, we obtain $\eta(\xi f) - (\nabla_\eta \xi) f = \eta(\xi f) - [\eta, \xi] f - (\nabla_\xi \eta) f = \xi(\eta f) - (\nabla_\xi \eta) f$, and the result is proved. □

The following result shows that the Riemannian Hessian of a function f at a point x coincides with the Euclidean Hessian of the function $f \circ \operatorname{Exp}_x$ at the origin $0_x \in T_x\mathcal{M}$. Note that $f \circ \operatorname{Exp}_x$ is a real-valued function on the Euclidean space $T_x\mathcal{M}$.

Proposition 5.5.4 *Let $\mathcal{M}$ be a Riemannian manifold and let f be a real-valued function on $\mathcal{M}$. Then*

$$\operatorname{Hess} f(x) = \operatorname{Hess}(f \circ \operatorname{Exp}_x)(0_x) \tag{5.30}$$

for all $x \in \mathcal{M}$, where $\operatorname{Hess} f(x)$ *denotes the Riemannian Hessian of $f : \mathcal{M} \to \mathbb{R}$ at x and* $\operatorname{Hess}(f \circ \operatorname{Exp}_x)(0_x)$ *denotes the Euclidean Hessian of $f \circ \operatorname{Exp}_x : T_x\mathcal{M} \to \mathbb{R}$ at the origin of $T_x\mathcal{M}$ endowed with the inner product defined by the Riemannian structure on $\mathcal{M}$.*

Proof. This result can be proven by working in normal coordinates and invoking the fact that the Christoffel symbols vanish in these coordinates. We provide an alternative proof that does not make use of index notation. We have to show that

$$\langle \operatorname{Hess} f(x)[\xi], \eta\rangle = \langle \operatorname{Hess}(f \circ \operatorname{Exp}_x)(0_x)[\xi], \eta\rangle \tag{5.31}$$

for all $\xi, \eta \in T_x\mathcal{M}$. Since both sides of (5.31) are symmetric bilinear forms in ξ and η, it is sufficient to show that

$$\langle \operatorname{Hess} f(x)[\xi], \xi\rangle = \langle \operatorname{Hess}(f \circ \operatorname{Exp}_x)(0_x)[\xi], \xi\rangle \tag{5.32}$$

for all $\xi \in T_x\mathcal{M}$. Indeed, for any symmetric linear form B, we have the *polarization identity*

$$2B(\xi, \eta) = B(\xi + \eta, \xi + \eta) - B(\xi, \xi) - B(\eta, \eta),$$

which shows that the mapping $(\xi, \eta) \mapsto B(\xi, \eta)$ is fully specified by the mapping $\xi \mapsto B(\xi, \xi)$. Since the right-hand side of (5.32) involves a classical (Euclidean) Hessian, we have

$$\langle \operatorname{Hess}(f \circ \operatorname{Exp}_x)(0_x)[\xi], \xi\rangle = \frac{\mathrm{d}^2}{\mathrm{d}t^2}(f \circ \operatorname{Exp}_x)(t\xi)\bigg|_{t=0}$$
$$= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathrm{d}}{\mathrm{d}t} f(\operatorname{Exp}_x(t\xi))\right)\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\mathrm{D}f(\operatorname{Exp}_x t\xi)\left[\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{Exp}_x t\xi\right]\right)\bigg|_{t=0}.$$

It follows from the definition of the gradient that this last expression is equal to $\frac{\mathrm{d}}{\mathrm{d}t}\langle \operatorname{grad} f(\operatorname{Exp}_x t\xi), \frac{\mathrm{d}}{\mathrm{d}t}\operatorname{Exp}_x t\xi\rangle\big|_{t=0}$. By the invariance property of the metric, this is equal to $\langle \frac{\mathrm{D}}{\mathrm{d}t}\operatorname{grad} f(\operatorname{Exp}_x t\xi), \xi\rangle + \langle \operatorname{grad} f(x), \frac{\mathrm{D}^2}{\mathrm{d}t^2}\operatorname{Exp}_x t\xi\rangle$. By definition of the exponential mapping, we have $\frac{\mathrm{D}^2}{\mathrm{d}t^2}\operatorname{Exp}_x t\xi = 0$ and $\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{Exp}_x t\xi\big|_{t=0} = \xi$. Hence the right-hand side of (5.32) reduces to

$$\langle \nabla_\xi \operatorname{grad} f, \xi\rangle,$$

and the proof is complete. □

The result is in fact more general. It holds whenever the retraction and the Riemannian exponential agree to the second order along all rays. This result will not be used in the convergence analyses, but it may be useful to know that various retractions yield the same Hessian operator.

Proposition 5.5.5 *Let R be a retraction and suppose in addition that*

$$\frac{\mathrm{D}^2}{\mathrm{d}t^2} R(t\xi)\Big|_{t=0} = 0 \quad \textit{for all } \xi \in T_x\mathcal{M}, \tag{5.33}$$

where $\frac{\mathrm{D}^2}{\mathrm{d}t^2}\gamma$ denotes acceleration of the curve γ as defined in (5.23)*. Then*

$$\operatorname{Hess} f(x) = \operatorname{Hess}(f \circ R_x)(0_x). \tag{5.34}$$

Proof. The proof follows the proof of Proposition 5.5.4, replacing Exp_x by R_x throughout. The first-order ridigidity condition of the retraction implies that $\frac{\mathrm{d}}{\mathrm{d}t} R_x t\xi\big|_{t=0} = \xi$. Because of this and of (5.33), we conclude as in the proof of Proposition 5.5.4. □

Proposition 5.5.5 provides a way to compute the Riemannian Hessian as the Hessian of a real-valued function $f \circ R_x$ defined on the Euclidean space $T_x\mathcal{M}$. In particular, this yields a way to compute $\langle \operatorname{Hess} f(x)[\xi], \eta\rangle$ by taking second derivatives along curves, as follows. Let R be any retraction satisfying the acceleration condition (5.33). First, observe that, for all $\xi \in T_x\mathcal{M}$,

$$\langle \operatorname{Hess} f(x)[\xi], \xi\rangle = \langle \operatorname{Hess}(f \circ R_x)(0_x)[\xi], \xi\rangle = \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(R_x(t\xi))\Big|_{t=0}. \tag{5.35}$$

Second, in view of the symmetry of the linear operator $\operatorname{Hess} f(x)$, we have the polarization identity

$$\langle \operatorname{Hess} f(x)[\xi], \eta\rangle = \tfrac{1}{2}(\langle \operatorname{Hess} f(x)[\xi+\eta], \xi+\eta\rangle - \langle \operatorname{Hess} f(x)[\xi], \xi\rangle - \langle \operatorname{Hess} f(x)[\eta], \eta\rangle). \tag{5.36}$$

Equations (5.35) and (5.36) yield the identity

$$\langle \operatorname{Hess} f(x)[\xi], \eta\rangle = \frac{1}{2}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(f(R_x(t(\xi+\eta))) - f(R_x(t\xi)) - f(R_x(t\eta))\right)\Big|_{t=0}, \tag{5.37}$$

valid for any retraction R that satisfies the zero initial acceleration condition (5.33). This holds in particular for $R = \mathrm{Exp}$, the exponential retraction.

Retractions that satisfy the zero initial acceleration condition (5.33) will be called *second-order retractions*. For general retractions the equality of the Hessians stated in (5.34) does not hold. Nevertheless, none of our quadratic convergence results will require the retraction to be second order. The fundamental reason can be traced in the following property.

Proposition 5.5.6 *Let R be a retraction and let v be a critical point of a real-valued function f (i.e.,* $\operatorname{grad} f(v) = 0$*). Then*

$$\operatorname{Hess} f(v) = \operatorname{Hess}(f \circ R_v)(0_v).$$

Proof. We show that $\langle \operatorname{Hess} f(v)[\xi_v], \eta_v\rangle = \langle \operatorname{Hess}(f \circ R_v)(0_v)[\xi_v], \eta_v\rangle$ for all $\xi, \eta \in \mathfrak{X}(\mathcal{M})$. From Proposition 5.5.2, we have $\langle \operatorname{Hess} f(v)[\xi_v], \eta_v\rangle = \xi_v(\eta f) - (\nabla_{\xi_v}\eta) f$. The second term is an element of $T_v\mathcal{M}$ applied to f; since v is a critical point of f, this term vanishes, and we are left with $\langle \operatorname{Hess} f(v)[\xi_v], \eta_v\rangle = \xi_v(\eta f)$. Fix a basis $(e_1, \ldots, e_n)$ of $T_x\mathcal{M}$ and consider the coordinate chart φ defined by $\varphi^{-1}(y^1, \ldots, y^n) = R_v(y^1 e_1 + \cdots + y^n e_n)$. Let η^i and ξ^i denote the coordinates of η and ξ in this chart. Since v is a critical point of f, $\partial_i(f \circ \varphi^{-1})$ vanishes at 0, and we obtain $\xi_v(\eta f) = \sum_i \xi_v^i \partial_i(\sum_j \eta^j \partial_j(f \circ \varphi^{-1})) = \sum_{i,j} \xi_v^i \eta_v^j \partial_i\partial_j(f \circ \varphi^{-1})$. Since $\mathrm{D}R_v(0_v)$ is the identity, it follows that ξ_v^i and η_v^j are the components of ξ_v and η_v in the basis $(e_1, \ldots, e_n)$; thus the latter expression is equal to $\langle \operatorname{Hess}(f \circ R_v)(0_v)[\xi], \eta\rangle$. □

5.6 SECOND COVARIANT DERIVATIVE*

In the previous section, we assumed that the manifold $\mathcal{M}$ was Riemannian. This assumption made it possible to replace the differential $\mathrm{D}f(x)$ of a function f at a point x by the tangent vector $\operatorname{grad} f(x)$, satisfying

$$\langle \operatorname{grad} f(x), \xi\rangle = \mathrm{D}f(x)[\xi] \quad \text{for all } \xi \in T_x\mathcal{M}.$$

This led to the definition of $\operatorname{Hess} f(x) : \xi_x \mapsto \nabla_{\xi_x} \operatorname{grad} f$ as a linear operator of $T_x\mathcal{M}$ into itself. This formulation has several advantages: eigenvalues and eigenvectors of the Hessian are well defined and, as we will see in Chapter 6, the definition leads to a streamlined formulation (6.4) for the Newton equation. However, on an arbitrary manifold equipped with an affine connection, it is equally possible to define a Hessian as a second covariant derivative that applies bilinearly to two tangent vectors and returns a scalar. This second covariant derivative is often called "Hessian" in the literature, but we will reserve this term for the operator $\xi_x \mapsto \nabla_{\xi_x} \operatorname{grad} f$.

To develop the theory of second covariant derivative, we will require the concept of a covector. Let $T_x^*\mathcal{M}$ denote the dual space of T_xM, i.e., the set of linear functionals (linear maps) $\mu_x : T_xM \to \mathbb{R}$. The set $T_x^*\mathcal{M}$ is termed the *cotangent space* of $\mathcal{M}$ at x, and its elements are called *covectors*. The bundle of cotangent spaces

$$T^*\mathcal{M} = \cup_{x\in\mathcal{M}} T_x^*\mathcal{M}$$

is termed the *cotangent bundle*. The cotangent bundle can be given the structure of a manifold in an analogous manner to the structure of the tangent bundle. A smooth section of the cotangent bundle is a smooth assignment $x \mapsto \mu_x \in T_x^*\mathcal{M}$. A smooth section of the cotangent bundle is termed a *covector field* or a *one-form* on $\mathcal{M}$. The name comes from the fact that a one-form field μ acts on "one" vector field $\xi \in \mathfrak{X}(\mathcal{M})$ to generate a scalar field on a manifold,

$$\mu[\xi] \in \mathfrak{F}(\mathcal{M}),$$

defined by $(\mu[\xi])|_x = \mu_x[\xi_x]$. The action of a covector field μ on a vector field ξ is often written simply as a concatenation of the two objects, $\mu\xi$. A covector is a $(0, 1)$-tensor. The most common covector encountered is the differential of a smooth function $f : \mathcal{M} \to \mathbb{R}$:

$$\mu_x = \mathrm{D}f(x).$$

Note that the covector field $\mathrm{D}f$ is zero exactly at critical points of the function f. Thus, another way of solving for the critical points of f is to search for zeros of $\mathrm{D}f$.

Given a manifold $\mathcal{M}$ with an affine connection ∇, a real-valued function f on $\mathcal{M}$, a point $x \in \mathcal{M}$, and a tangent vector $\xi_x \in T_x\mathcal{M}$, the covariant derivative of the covector field $\mathrm{D}f$ along ξ_x is a covector $\nabla_{\xi_x}(\mathrm{D}f)$ defined by imposing the property

$$\mathrm{D}(\mathrm{D}f[\eta])(x)[\xi_x] = \left(\nabla_{\xi_x}(\mathrm{D}f)\right)[\eta_x] + \mathrm{D}f(x)[\nabla_{\xi_x}\eta]$$

for all $\eta \in \mathfrak{X}(\mathcal{M})$. It is readily checked, using coordinate expressions, that $\left(\nabla_{\xi_x}(\mathrm{D}f)\right)[\eta_x]$ defined in this manner depends only on η through η_x and that $\left(\nabla_{\xi_x}(\mathrm{D}f)\right)[\eta_x]$ is a linear expression of ξ_x and η_x. The *second covariant derivative* of the real-valued function f is defined by

$$\nabla^2 f(x)[\xi_x, \eta_x] = \left(\nabla_{\xi_x}(\mathrm{D}f)\right)[\eta_x].$$

(There is no risk of confusing $[\xi_x, \eta_x]$ with a Lie bracket since $\nabla^2 f(x)$ is known to apply to two vector arguments.) The notation ∇^2 rather than D^2 is used to emphasize that the second covariant derivative depends on the choice of the affine connection ∇.

With development analogous to that in the preceding section, one may show that

$$\nabla^2 f(x)[\xi_x, \eta_x] = \xi_x(\eta f) - (\nabla_{\xi_x}\eta)f.$$

The second covariant derivative is symmetric if and only if ∇ is symmetric. For any second-order retraction R, we have

$$\nabla^2 f(x) = \mathrm{D}^2\,(f \circ R_x)(0_x),$$

where $\mathrm{D}^2\,(f \circ R_x)(0_x)$ is the classical second-order derivative of $f \circ R_x$ at 0_x (see Section A.5). In particular,

$$\nabla^2 f(x)[\xi_x, \xi_x] = \mathrm{D}^2\,(f \circ \mathrm{Exp}_x)(0_x)[\xi_x, \xi_x] = \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(\mathrm{Exp}_x(t\xi))|_{t=0}.$$

When x is a critical point of f, we have

$$\nabla^2 f(x) = \mathrm{D}^2\,(f \circ R_x)(0_x)$$

for *any* retraction R.

When $\mathcal{M}$ is a Riemannian manifold and ∇ is the Riemannian connection, we have

$$\nabla^2 f(x)[\xi_x, \eta_x] = \langle \mathrm{Hess}\, f(x)[\xi_x], \eta_x \rangle.$$

When F is a function on $\mathcal{M}$ into a vector space $\mathcal{E}$, it is still possible to uniquely define

$$\nabla^2 F(x)[\xi_x, \eta_x] = \sum_{i=1}^{n} (\nabla^2 F^i(x)[\xi_x, \eta_x]) e_i,$$

where $(e_1, \ldots, e_n)$ is a basis of $\mathcal{E}$.

5.7 NOTES AND REFERENCES

Our main sources for this chapter are O'Neill [O'N83] and Brickell and Clark [BC70].

A proof of superlinear convergence for Newton's method in $\mathbb{R}^n$ can be found in [DS83, Th. 5.2.1]. A proof of Proposition 5.2.1 (the existence of affine connections) is given in [BC70, Prop. 9.1.4]. It relies on partitions of unity (see [BC70, Prop. 3.4.4] or [dC92, Th. 0.5.6] for details). For a proof of the existence and uniqueness of the covariant derivative along curves $\frac{\mathrm{D}}{\mathrm{d}t}$, we refer the reader to [O'N83, Prop. 3.18] for the Riemannian case and Helgason [Hel78, §I.5] for the general case. More details on the exponential can be found in do Carmo [dC92] for the Riemannian case, and in Helgason [Hel78] for the general case. For a proof of the minimizing property of geodesics, see [O'N83, §5.19]. The material about the Riemannian connection on Riemannian submanifolds comes from O'Neill [O'N83]. For more details on Riemannian submersions and the associated Riemannian connections, see O'Neill [O'N83, Lemma 7.45], Klingenberg [Kli82], or Cheeger and Ebin [CE75].

The equation (5.26) for the Stiefel geodesic is due to R. Lippert; see Edelman *et al.* [EAS98]. The formula (5.27) for the geodesics on the Grassmann manifold can be found in Absil *et al.* [AMS04]. Instead of considering the Stiefel manifold as a Riemannian submanifold of $\mathbb{R}^{n\times p}$, it is also possible to view the Stiefel manifold as a certain Riemannian quotient manifold of the orthogonal group. This quotient approach yields a different Riemannian metric on the Stiefel manifold, called the *canonical metric* in [EAS98]. The Riemannian connection, geodesics, and parallel translation associated with the canonical metric are different from those associated with the Riemannian metric (3.34) inherited from the embedding of $\mathrm{St}(p,n)$ in $\mathbb{R}^{n\times p}$. We refer the reader to Edelman *et al.* [EAS98] for more information on the geodesics and parallel translations on the Stiefel manifold.

The geometric Hessian is not a standard topic in differential geometry. Some results can be found in [O'N83, dC92, Sak96, Lan99]. The Hessian is often defined as a tensor of type $(0,2)$—it applies to two vectors and returns a scalar—using formula (5.29). This does not require a Riemannian metric. Such a tensor varies under changes of coordinates via a congruence transformation. In this book, as in do Carmo [dC92], we define the Hessian as a tensor of type $(1,1)$, which can thus be viewed as a linear transformation of the tangent space. It transforms via a similarity transformation, therefore its eigenvalues are well defined (they do not depend on the chart).