Liam O'Carroll January 2022

Optimization on Smooth Manifolds Crash Course

Contents

1 Introduction

This is a crash course on optimization on smooth manifolds. Basically all of the definitions and theorems are taken from [Bou20], and I've done my best to provide the exact places in the book where you can find them. Unfortunately [Bou20] is a preprint, and newer versions have been released since I made these notes. The version I used to create these notes was compiled on November 9, 2020.[1] Thus, be warned that theorem/definition numbering in the most recent version of [Bou20] may not always exactly match what is given here (but sections should stay the same).

An important note about [Bou20]: They structure their textbook by first defining and proving everything for *embedded submanifolds* (manifolds that are subsets of a Euclidean space). It isn't until Chapter 8 that they go back and define things more generally. As embedded submanifolds are much simpler to work with and make up most applications, this write-up focuses on them nearly exclusively. When it makes sense to do so though I will try to provide intuition about how definitions generalize beyond the embedded case, since all of the generalizations are very natural and not something to be afraid of.

1.1 Running examples

We will use the following optimization problem as a running example:

$$\min_{y \in \mathcal{M}_0} f(y)$$

where

$$\mathcal{M}_0 = \{y \in \mathbb{R}^d \mid h(y) = 0\}.$$

Here $f : \mathcal{M}_0 \to \mathbb{R}$ and $h : \mathbb{R}^d \to \mathbb{R}^m$. We let $h_i : \mathbb{R}^d \to \mathbb{R}$ denote the ith component function of h for $i \in [m]$. The properties of this kind of manifold (i.e., a manifold defined by the zero set of a function h) are covered in "Section 7.7: Manifolds defined by $h(x) = 0$" in [Bou20], so we will borrow heavily from that section in examples.

To illustrate the particularization of a manifold defined by the zero set of a function to an application, we will use the Burer-Monteiro optimization problem:

$$\min_{Y \in \mathcal{M}_p} \left\langle C, YY^T \right\rangle$$

where

$$\mathcal{M}_p = \{Y \in \mathbb{R}^{n \times p} \mid \mathscr{A}(YY^T) = b\}$$

Here $\mathscr{A} : \mathbb{S}^{n \times n} \to \mathbb{R}^m$ is linear, $C \in \mathbb{S}^{n \times n}$, and $b \in \mathbb{R}^m$. We can break $\mathscr{A}$ into coordinate functions via $A_1, \ldots, A_m \in \mathbb{S}^{n \times n}$ by letting $\mathscr{A}(YY^T)_i = \left\langle A_i, YY^T \right\rangle$. The Burer-Monteiro problem can be viewed as a particularization of the more general problem above by setting $f(Y) = \left\langle C, YY^T \right\rangle$ and $h(Y) = \mathscr{A}(YY^T) - b$, or equivalently, $h_i(Y) = \left\langle A_i, YY^T \right\rangle - b_i$. We define $f(Y)$ and $h(Y)$ in this way, where the case of Y makes it clear whether we are referring to the more general case or the Burer-Monteiro problem.

[1] You can see the compilation date on the first page of the PDF of the textbook.

1.2 Notation

$\mathbb{S}^{n\times n}$ denotes the space of $n \times n$ symmetric matrices. We use $\nabla f, \nabla^2 f$ to denote the Euclidean gradient and Euclidean Hessian respectively,[2] whereas $\operatorname{grad} f, \operatorname{Hess} f$ denote the Riemannian gradient and Riemannian Hessian. We typically use $\langle \cdot, \cdot \rangle$ to denote the inner product, but $A \bullet B$ may be used as well on occasion to denote the trace inner product between matrices (A and B in this case). I_p denotes the $p \times p$ identity matrix. $\mathcal{M}$ is used to denote a general or unspecified manifold.

1.3 Terminology

[Bou20] is careful to use the term "embedded submanifold" when they specifically mean a submanifold of a linear space, as opposed to a submanifold of a more general manifold (see Section 8.14 in that reference for the latter). Unfortunately these notes may not always adhere to that convention, although generally when we say "submanifold" we mean an embedded submanifold in particular. When in doubt, we advise referring to [Bou20].

2 First-order geometry

In this section we build up to defining the Riemannian gradient. First, it is necessary to define the *differential* of a function between linear spaces. Later we will extend this definition to a function between manifolds.

Definition 2.1 (Differential of a function between linear spaces) *Let $F : \mathbb{R}^d \to \mathbb{R}^{d'}$ be smooth (infinitely differentiable). The differential[3] of F at x is the linear map $DF(x) : \mathbb{R}^d \to \mathbb{R}^{d'}$ defined as:*

$$DF(x)[v] = \lim_{t\to 0} \frac{F(x+tv) - F(x)}{t} = \frac{d}{dt} F(x+tv)\Big|_{t=0}.$$

Intuitively, just think of $d' \leftarrow 1$, in which case the differential eats a vector v and spits out the directional derivative of F in the direction v. Thus, it is just the linear map defined by $\nabla F(x)$. More generally, the differential tells you "how F changes" as you start at x and move along v. (Specifically, it is the total derivative or equivalently the Jacobian viewed as a linear map.) As we will see, this intuition generalizes to the definition of the differential of a map between manifolds.

We can already define what a smooth embedded submanifold (of a linear space) is:

Definition 2.2 (Smooth embedded submanifold of a linear space) *Let $\mathcal{M}$ be a subset of $\mathbb{R}^d$. $\mathcal{M}$ is a smooth embedded submanifold[4] of $\mathbb{R}^d$ if either:*

[2] Technically this isn't necessary since the Euclidean derivatives coincide with the Riemannian ones in linear spaces, but we will do it for illustrative purposes. [Bou20] uses $\operatorname{grad} f, \operatorname{Hess} f$ to denote the Euclidean gradient and Euclidean Hessian because ∇ is used as the symbol for a *connection* on a manifold. However, these notes only briefly cover connections, and I think the clarity obtained by using the usual ∇ for the Euclidean derivatives outweighs the potential confusion arising from it also being the symbol for a connection.

[3] (3.12) in [Bou20].

[4] Definition 3.6 in [Bou20]. I should mention that Boumal also includes a rather nice illustration to accompany this definition on the same page.

1. *$\mathcal{M}$ is an open subset of $\mathbb{R}^d$, in which case it is called an open submanifold.*
2. *For a fixed integer $k \geq 1$ and for each $x \in \mathcal{M}$ there exists a neighborhood U of x in $\mathbb{R}^d$ and a smooth function $h : U \to \mathbb{R}^k$ such that*
   (a) *If $y \in U$, then $h(y) = 0$ if and only if $y \in \mathcal{M}$; and*
   (b) $\text{rank}\mathrm{D}h(x) = k$. *(In other words, $\mathrm{D}h(x)$ is full rank.)*

   h is called a local defining function for $\mathcal{M}$ at x.

For now on we may drop the word "smooth," but know that when we refer to an "embedded submanifold," we really always mean a smooth embedded submanifold. (The same omission is made in [Bou20].)

Example 2.3 ($\mathcal{M}_0$ is a smooth embedded submanifold if LICQ holds) *In this case the whole manifold is defined by a single defining function: h. Condition (a) under Definition 2.2 is trivially satisfied. As for Condition (b), we derive*

$$\mathrm{D}h(y)[v] = \left(\langle \nabla h_1(y), v\rangle, \ldots, \langle \nabla h_m(y), v\rangle\right)$$

Then $\mathrm{D}h(Y)$ is full rank if and only if $\nabla h_1(y), \ldots \nabla h_m(y)$ are linearly independent. In other words, if $\nabla h_1(y), \ldots \nabla h_m(y)$ are linearly independent at all $y \in \mathcal{M}_p$, then $\mathcal{M}_0$ is indeed a smooth embedded submanifold. Note that this is equivalent to the constraint qualification LICQ holding at every point. Or using the nonlinear programming language of [LY16], it is equivalent to saying that every $y \in \mathcal{M}_0$ is a regular point. (See page 325 in that reference.)

Example 2.4 ($\mathcal{M}_p$ is a smooth embedded submanifold if LICQ holds) *We have*

$$\begin{aligned}\mathrm{D}h(Y)[V] &= \left(\left\langle \nabla_Y (A_1 \bullet YY^T - b_1), V\right\rangle, \ldots, \left\langle \nabla_Y (A_m \bullet YY^T - b_m), V\right\rangle\right) \\ &= 2\left(\langle A_1 Y, V\rangle, \ldots, \langle A_m Y, V\rangle\right)\end{aligned}$$

Then $\mathrm{D}h(Y)$ is full rank if and only if $A_1 Y, \ldots A_m Y$ are linearly independent. In other words, if $A_1 Y, \ldots A_m Y$ are linearly independent at all $Y \in \mathcal{M}_p$, then $\mathcal{M}_p$ is indeed a smooth embedded submanifold.

Regarding Definition 2.2: It turns out that it is necessary for $\mathrm{D}h(x)$ to be full rank on x—it is *not* sufficient for $\mathrm{D}h(x)$ to just have constant rank at all $x \in \mathcal{M}_p$. However, it *is* sufficient for $\mathrm{D}h(x)$ to have constant rank if it has constant rank not just on the zero set of h but on a neighborhood of the zero set as well.[5] This is exactly Case (b) in Assumption 1.1 in [BVB18]! (Case (a) in Assumption 1.1 is of course just Condition (b) in Definition 2.2.) This is all to say that the manifold approach can give you a way to deal with "redundant constraints" which the usual nonlinear programming approach cannot (as far as we know).

Finally, it is also worth noting that Condition (b) in Definition 2.2 can be shown to be equivalent to Definition 1 in [WW20]. This is to say that everything here is standard even if it might appear in slightly different forms in the literature.

[5] See Sections 3.10 and 8.14 in [Bou20].

Definition 2.5 (Velocity of a curve) *Let $c : I \subseteq \mathbb{R} \to \mathbb{R}^d$ be a curve. Its velocity*[6] *at time t is $c'(t) = \dfrac{\mathrm{d}}{\mathrm{d}t}c(t)$. (This is just the ordinary Euclidean derivative.)*

Definition 2.6 (Tangent space, tangent vector) *Let $\mathcal{M}$ be a smooth embedded submanifold of $\mathbb{R}^d$. For all $x \in \mathcal{M}$, define*

$$\mathrm{T}_x\mathcal{M} = \{c'(0) \mid c : I \subseteq \mathbb{R} \to \mathcal{M} \text{ is smooth around } 0 \text{ and } c(0) = x\}.$$

In other words, $v \in \mathrm{T}_x\mathcal{M}$ if and only if there exists a smooth curve on $\mathcal{M}$ passing through x with velocity v.[7] *$\mathrm{T}_x\mathcal{M}$ is called the tangent space to $\mathcal{M}$ at x. $v \in \mathrm{T}_x\mathcal{M}$ is called a tangent vector to $\mathcal{M}$ at x.*[8]

It can be shown that the tangent space is a linear space (aka a vector space).

The fact that the tangent space is a linear space is key because it allows us to easily extend many definitions from Euclidean spaces to manifolds.

Theorem 2.7 (Expression for the tangent space of a smooth embedded submanifold) *Let $\mathcal{M} \subseteq \mathbb{R}^d$ be an embedded submanifold. Then for $x \in \mathcal{M}$, if M is an open submanifold (first case in Definition 2.2), then $\mathrm{T}_x\mathcal{M} = \mathbb{R}^d$. Otherwise, $\mathrm{T}_x\mathcal{M} = \ker \mathrm{D}h(x)$ where h is any local defining function at x.*[9]

Note that this is exactly the definition of the "tangent plane" given in the nonlinear programming textbook [LY16]—see page 325! Thus, we see that notions in smooth manifold optimization (smooth embedded submanifold, tangent space) reduce to notions in nonlinear programming (LICQ, tangent plane), although the manifold approach is more general and gives us better geometric intuition.

Example 2.8 (Tangent space to $\mathcal{M}_0$) *Recall from Example 2.3 that we have*

$$\mathrm{D}h(y)[v] = \left(\langle \nabla h_1(y), v\rangle, \ldots, \langle \nabla h_m(y), v\rangle\right)$$

Theorem 2.7 instantly gives

$$\mathrm{T}_y\mathcal{M}_0 = \left\{\dot{v} \in \mathbb{R}^d \mid \langle \nabla h_i(y), \dot{v}\rangle = 0 \text{ for } i \in [m]\right\}.$$

Example 2.9 (Tangent space to $\mathcal{M}_p$) *Recall from Example 2.4 that we have*

$$\mathrm{D}h(Y)[V] = 2\left(\langle A_1 Y, V\rangle, \ldots, \langle A_m Y, V\rangle\right).$$

Theorem 2.7 instantly gives

$$\mathrm{T}_Y\mathcal{M}_p = \left\{\dot{V} \in \mathbb{R}^{n\times p} \mid \left\langle A_i Y, \dot{V}\right\rangle = 0 \text{ for } i \in [m]\right\}.$$

As a check, note that this is exactly how they define the tangent space in [BVB18] (Lemma 2.1). It is also exactly how the tangent space is defined in [WW20] (page 5), although they express it in a more complicated way.

[6] (3.12) in [Bou20].
[7] Definition 3.7 in [Bou20].
[8] Definition 3.10 in [Bou20].
[9] Theorem 3.8 in [Bou20].

Definition 2.10 (Normal space) *Let $\mathcal{M} \subseteq \mathbb{R}^d$ be an embedded submanifold. The normal space*[10] *at x, denoted $\mathrm{N}_x\mathcal{M}$, is simply the orthogonal complement of $\mathrm{T}_x\mathcal{M}$ in $\mathbb{R}^d$.*

Example 2.11 (Normal space to $\mathcal{M}_0$) *Referencing Example 2.8, it is easy to see that*

$$\mathrm{N}_y\mathcal{M}_0 = \mathrm{span}\{\nabla h_1(y), \ldots, \nabla h_m(y)\}.$$

Example 2.12 (Normal space to $\mathcal{M}_p$) *Referencing Example 2.9, it is easy to see that*

$$\mathrm{N}_Y\mathcal{M}_p = \mathrm{span}\{A_1 Y, \ldots, A_m Y\}.$$

Definition 2.13 (Dimension of a manifold) *The dimension of a manifold $\mathcal{M} \subseteq \mathbb{R}^d$, denoted $\dim \mathcal{M}$, is just the dimension of its tangent space.*[11] *Note that the tangent space is linear (aka a vector space) and the dimension of $\mathrm{T}_x\mathcal{M}$ does not depend on x, so this is well-defined. In particular, this means that $\dim \mathcal{M} = d$ if $\mathcal{M}$ is an open embedded submanifold, and $\dim \mathcal{M} = \dim \ker \mathrm{D}h(x) = d - \mathrm{rank}\mathrm{D}h(x)$ otherwise (due to Theorem 2.7).*

Example 2.14 (Dimension of $\mathcal{M}_0$) *Remember that we need $\mathrm{D}h(y)$ to be full rank for $\mathcal{M}_0$ to be a smooth embedded submanifold in the first place (see Example 2.3). Since $\mathrm{D}h(y) : \mathbb{R}^d \to \mathbb{R}^m$, we have that $\mathrm{rank}\mathrm{D}h(y) = m$. So $\dim \mathcal{M}_0 = d - m$.*

Example 2.15 (Dimension of $\mathcal{M}_p$) *It immediately follows from Example 2.14 that $\dim \mathcal{M}_p = np - m$.*

As a check, this is confirmed by [BVB18] (Proposition 1.2) and [WW20] (page 5). (Although [BVB18] makes things slightly more complicated because they allow $\mathrm{D}h(Y)$ to not be full rank as long as it has constant rank on a neighborhood of $\mathcal{M}_p$ [see the discussion of this below Example 2.4 above]. This affects $\dim \mathcal{M}_p$, although it still just ends up being $d - \mathrm{rank}\mathrm{D}h(x)$.)

We now define what it means for a map between two manifolds to be smooth, which we will do by smoothly extending the function in question from the manifold to the linear embedding space (also called the ambient space). In fact, the notion of a smooth extension will end up being the more important part of the definition below for our purposes!

Definition 2.16 (Smooth map between manifolds, smooth extension) *Let $\mathcal{M} \subseteq \mathbb{R}^d$ and $\mathcal{M}' \subseteq \mathbb{R}^{d'}$ be embedded submanifolds. A map $F : M \to \mathcal{M}'$ is smooth if there exists a function $\overline{F} : U \to \mathbb{R}^{d'}$ which is smooth (in the usual sense, aka infinitely differentiable) on a neighborhood U of $\mathcal{M}$ in $\mathbb{R}^d$. Furthermore, this map $\overline{F}$ must be such that F and $\overline{F}$ coincide on $\mathcal{M} \cap U$, i.e., $F(y) = \overline{F}(y)$ for all $y \in \mathcal{M} \cap U$. (Equivalently, F is the restriction of $\overline{F}$ to $\mathcal{M}$: $F = \overline{F}\big|_{\mathcal{M}}$.) We call $\overline{F}$ a smooth extension of F.*[12]

To gain intuition for the above definition, it is best to think of $\mathcal{M}'$ as just being $\mathbb{R}$. (Note that $\mathbb{R}$ is an open embedded submanifold per Definition 2.2.)

[10] Defined in the paragraph above (5.15) in [Bou20].

[11] Theorem 3.10 in [Bou20].

[12] Proposition 3.24 in [Bou20]. Note that Boumal first defines what it means for F to be smooth at a point in Definition 3.23, but I have omitted this for simplicity.

Example 2.17 (Smooth extension of our objective function on $\mathcal{M}_p$) *Clearly a smooth extension of $f : \mathcal{M}_p \to \mathbb{R}$ defined as $f(Y) = \langle C, YY^T \rangle$ is $\overline{f} : \mathbb{R}^{np} \to \mathbb{R}$ defined as $\overline{f}(Y) = \langle C, YY^T \rangle$. In other words, the only thing that changes when we take a smooth extension is the domain! Thus, we can really view ourselves as starting with a smooth extension in this application, and then obtaining the objective function on the manifold by restricting the original function to the manifold.*

We are now ready to extend Definition 2.1 to manifolds.

Definition 2.18 (Differential of a map between manifolds) *The differential of $F : \mathcal{M} \subseteq \mathbb{R}^d \to \mathcal{M}' \subseteq \mathbb{R}^{d'}$ at x is a linear operator $\mathrm{D}F(x) : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_{F(x)}\mathcal{M}'$ defined by*

$$\mathrm{D}F(x)[v] = \frac{\mathrm{d}}{\mathrm{d}t} F(c(t))\Big|_{t=0},$$

where c is a smooth curve on $\mathcal{M}$ passing through x at $t = 0$ with velocity v.[13]

Intuitively, $\mathrm{D}F(x)$ is just a map which tells you how F changes as you start at x and move along the manifold $\mathcal{M}$ in the direction $v \in T_x\mathcal{M}$. Although we can't actually move along v—we have to move along a curve with velocity v called $c(t)$. Then, $F(c(t))$ is just a curve on $\mathcal{M}$, and we would like to return how F changes as we move along this curve. Well, let's just return the velocity of this curve at $t = 0$—this is exactly $\mathrm{D}F(x)[v]$.

It is worth mentioning that the differential of Theorem 2.18 satisfies all of the properties we would expect it to satisfy (the chain rule, the product rule, etc.). See Section 4.7 in [Bou20] for details.

The next theorem provides even more intuition:

Theorem 2.19 (Differential of a map between manifolds via a smooth extension) *With the same notation as in 2.18, let $\overline{F} : U \subseteq \mathbb{R}^d \to \mathbb{R}^{d'}$ be a smooth extension of F. (We know such an extension exists since F is smooth.) Then $\mathrm{D}F(x)$ is the restriction of $\mathrm{D}\overline{F}(x)$ to $\mathrm{T}_x\mathcal{M}$, i.e.,*

$$\mathrm{D}F(x) = D\overline{F}(x)\big|_{\mathrm{T}_x\mathcal{M}}.$$

Note that $\mathrm{D}\overline{F}(x)$ is a differential between linear spaces, so it is defined via Definition 2.1.[14]

In a sense, this says that we could equivalently define the differential between manifolds via smooth extensions. This also helps us gain much better intuition as to what $\mathrm{D}F(x)$ is if F is a map between manifolds. Let $d' \leftarrow 1$ so that $F : \mathcal{M} \subseteq \mathbb{R}^d \to \mathcal{M}' \subseteq \mathbb{R}$. Let $\overline{F}$ be a smooth extension of F. Recall from the discussion below Definition 2.1 that $\mathrm{D}\overline{F}(x)$ eats a vector v and spits out the directional derivative of $\overline{F}$ in the direction v. Then, Theorem 2.19 basically says that $DF(x)$ does the exact same thing, except you are only allowed to feed in vectors in $T_x\mathcal{M}$. This restriction makes sense because intuitively, we are only allowed to move in directions that are in $T_x\mathcal{M}$, so $DF(x)$ is essentially a fancy way of forcing us to only pay attention to how F changes in directions which we actually care about!

[13] Definition 3.27 in [Bou20]. Once again, there is an accompanying illustration which is excellent!
[14] Proposition 3.28 in [Bou20].

Example 2.20 (Differential of the objective function over $\mathcal{M}_p$) *Define $\overline{f}$ as in Example 2.17. Using Definition 2.1 (and particularly the discussion below it), we have that for $Y \in \mathbb{R}^{np}$, $\mathrm{D}\overline{f}(Y)[V] : \mathbb{R}^{np} \to \mathbb{R}$ is defined via*

$$\begin{aligned}\mathrm{D}\overline{f}(Y)[V] &= \left\langle \nabla_Y \overline{f}(Y), V \right\rangle \\ &= 2\left\langle CY, V \right\rangle\end{aligned}$$

Then using Theorem 2.19, we have that $\mathrm{D}f(Y) : \mathrm{T}_Y\mathcal{M}_p \to \mathbb{R}$ is

$$\mathrm{D}f(Y) = 2\left\langle CY, V \right\rangle|_{\mathrm{T}_Y\mathcal{M}_p}$$

where $\mathrm{T}_Y\mathcal{M}_p$ is defined in Example 2.9. We will obtain a more explicit expression for $\mathrm{D}f(Y)$ after we define the Riemannian gradient!

We need only a few more definitions before we can define the Riemannian gradient!

Definition 2.21 (Tangent bundle) *The tangent bundle*[15] *$\mathrm{T}\mathcal{M}$ of a manifold $\mathcal{M}$ is the disjoint union of the tangent spaces of $\mathcal{M}$:*

$$\mathrm{T}\mathcal{M} = \{(x, v) \mid x \in \mathcal{M} \text{ and } v \in \mathrm{T}_x\mathcal{M}\}.$$

Following the conventions of [Bou20], we may abuse notation at times and conflate v and (x, v) for a tangent vector $v \in \mathrm{T}_x\mathcal{M}$. We may write $(x, v) \in \mathrm{T}_x\mathcal{M}$, or even $v \in \mathrm{T}\mathcal{M}$ if it is clear from context that the foot or base of v is x.

Theorem 2.22 *If $\mathcal{M}$ is an embedded submanifold of $\mathbb{R}^d$, then the tangent bundle $\mathrm{T}\mathcal{M}$ is an embedded submanifold of $\mathbb{R}^d \times \mathbb{R}^d$.*[16] *(See Appendix B for more information about product manifolds.)*

We don't really need Theorem 2.22 in these notes beyond one use case: In multiple places (e.g., Definition 2.23, Definition 4.1), we will need to introduce a map from or to $\mathrm{T}\mathcal{M}$, and we will want such a map to be smooth per Definition 2.16. Since Definition 2.16 only applies to maps between embedded submanifolds, we need to recognize that $\mathrm{T}\mathcal{M}$ is itself an embedded submanifold.

Definition 2.23 (Vector field on a manifold) *A vector field*[17] *on a manifold $\mathcal{M}$ is a map $V : \mathcal{M} \to \mathrm{T}\mathcal{M}$ such that $V(x) \in \mathrm{T}_x\mathcal{M}$ for all $x \in \mathcal{M}$. If V is a smooth map, we say it is a smooth vector field.*

In other words, a vector field is just a formal way of assigning to each $x \in \mathcal{M}$ a tangent vector at that point.

[15] Definition 3.35 in [Bou20].
[16] Theorem 3.36 in [Bou20].
[17] Definition 3.37 in [Bou20].

Definition 2.24 (Riemannian gradient) *Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function on a Riemannian manifold*[18] *$\mathcal{M}$. The Riemannian gradient*[19] *of f is the vector field $\operatorname{grad} f$ on $\mathcal{M}$ uniquely defined by the following: For all $(x, v) \in \mathrm{T}\mathcal{M}$,*

$$\mathrm{D}f(x)[v] = \langle \operatorname{grad} f(x), v \rangle ,$$

where $\mathrm{D}f(x)$ is as defined in 2.18 (or Theorem 2.19).

Note that this definition critically implies that $\operatorname{grad} f(x) \in \mathrm{T}_x\mathcal{M}$, i.e., it is a tangent vector.

Definition 2.25 (Orthogonal projector to the tangent space) *Let $\mathcal{M} \subseteq \mathbb{R}^d$ be an embedded submanifold. We denote the orthogonal projector*[20] *from $\mathbb{R}^d$ to $\mathrm{T}_x\mathcal{M}$ as*

$$\operatorname{Proj}_x : \mathbb{R}^d \to \mathrm{T}_x\mathcal{M} \subseteq \mathbb{R}^d.$$

Being a projector, Proj_x is a linear operator such that $\operatorname{Proj}_x \circ \operatorname{Proj}_x = \operatorname{Proj}_x$. Being orthogonal, the matrix corresponding to Proj_x is symmetric.

Example 2.26 (Deriving Proj_y for $\mathcal{M}_0$) *Recall from Theorem 2.7 that*

$$\mathrm{T}_y\mathcal{M}_0 = \ker \mathrm{D}h(y).$$

Also, recall that the orthogonal projector onto the kernel of a matrix A can be represented as $I - A^+A$, where A^+ denotes the Moore-Penrose inverse of A (aka the pseudoinverse). Thus, one expression for Proj_y is

$$\operatorname{Proj}_y(V) = V - Dh(y)^+[Dh(y)[v]]. \tag{1}$$

We will also derive a second expression for Proj_y. Note that equivalently, $\mathrm{T}_y\mathcal{M}_p = \ker((Dh(y)^T)^T)$. Recall that the orthogonal projector onto the kernel of the transpose of a matrix, A^T, can be expressed as $I - AA^+$. Setting $A \leftarrow Dh(y)^T$, this gives

$$\operatorname{Proj}_y(v) = v - Dh(y)^T \left[(Dh(y)^T)^+[v]\right]. \tag{2}$$

(I am giving both of these expressions because although (1) is simpler, Boumal actually only gives (2) in Section 7.7 of [Bou20]. It might be because typically $Dh(V)^T$ is typically simpler to work with, although I'm not sure.)

Example 2.27 (Deriving Proj_Y for $\mathcal{M}_p$) *Particularizing (1) or (2) to the Burer-Monteiro problem just involves plugging in what $Dh(Y)$ is in that context—see Example 2.4.*

It is worth mentioning that at first glance, neither (1) nor (2) look equivalent to the expression for Proj_Y given in Lemma 2.2 in [BVB18]. Indeed, [BVB18] uses a more unorthodox method to derive their expression for Proj_Y. However, their expression for Proj_Y can be shown to be equivalent to (2) using these properties of the Moore-Penrose inverse.

[18] See Section 2.1 for what a Riemannian manifold is.
[19] Definition 3.50 in [Bou20].
[20] Definition 3.52 in [Bou20].

And finally:

Theorem 2.28 (Connecting the Riemannian gradient and Euclidean gradient) *Let $\mathcal{M} \subseteq \mathbb{R}^d$ be a Riemannian submanifold[21] of $\mathbb{R}^d$. The Riemannian gradient of f is given by*

$$\mathrm{grad} f(x) = \mathrm{Proj}_x(\nabla \overline{f}(x)),$$

where $\overline{f}$ is any smooth extension of f to a neighborhood of $\mathcal{M}$ in $\mathbb{R}^d$.[22]

It is not hard to see that Theorem 2.28 follows quickly from Definition 2.24 and Theorem 2.19.

Example 2.29 (Riemannian gradient in the optimization problem over M_0) *Let $\overline{f} : \mathbb{R}^d \to \mathbb{R}$ denote a smooth extension of $f : \mathcal{M}_0 \to \mathbb{R}$. Theorem 2.28 yields*

$$\mathrm{grad} f(y) = \nabla \overline{f}(y) - Dh(y)^T \left[(Dh(y)^T)^+ [\nabla \overline{f}(y)] \right]$$

To really see what is going on here, it is helpful to reexpress this as follows:

$$\mathrm{grad} f(y) = \nabla \overline{f}(y) - \sum_{i=1}^{m} \lambda_i(y) \nabla h_i(y) \tag{3}$$

where

$$\lambda(y) := (Dh(y)^T)^+ [\nabla \overline{f}(y)].$$

(Recall from Example 2.3 that the ith column of $Dh(y)^T$ is simply $\nabla h_i(y)$.)

How should we think of the "multipliers" that make up $\lambda(Y)$? Well, remember how we derived (3) in the first place via Theorem 2.28—we took $\nabla \overline{f}(Y)$ and deleted its normal component. Thus, you should just think of the $\lambda_i(y)$'s as being the multipliers which delete the component of $\nabla \overline{f}(y)$ that lies in

$$\mathrm{N}_y \mathcal{M}_0 = \mathrm{span}\{\nabla h_1(y), \ldots, \nabla h_m(y)\}.$$

Indeed, this is exactly what the pseudoinverse does. And this makes complete sense because $\mathrm{grad} f(Y)$ must be a tangent vector by definition, so the $\lambda_i(y)$'s couldn't be anything else.

It is also basically immediately apparent at this point from (3) that the first-order criticality condition from nonlinear programming is equivalent to $\mathrm{grad} f(y) = 0$. More on this later.

Example 2.30 (Riemannian gradient in the Burer-Monteiro problem) *Define $\overline{f}$ as in Example 2.17. Theorem 2.28 yields*

$$\begin{aligned} \mathrm{grad} f(Y) &= \nabla \overline{f}(Y) - Dh(Y)^T \left[(Dh(Y)^T)^+ [\nabla \overline{f}(Y)] \right] \\ &= 2CY - Dh(Y)^T \left[(Dh(Y)^T)^+ [2CY] \right]. \end{aligned}$$

[21] See Section 2.1 for what a Riemannian manifold is.

[22] Proposition 3.53 in [Bou20].

This is not particularly informative, but following the same path as in Example 2.29, we can reexpress this as

$$\begin{aligned}\operatorname{grad} f(Y) &= 2\left[CY - \sum_{i=1}^{m} \lambda_i(Y) A_i Y\right] \\ &= 2\left[\left(C - \sum_{i=1}^{m} \lambda_i(Y) A_i\right) Y\right]\end{aligned}$$

where

$$\lambda(Y) := (Dh(Y)^T)^+[2CY].$$

As a sanity check, this is exactly the expression derived in [BVB18]—see (2.6) and (2.5) in that paper. Although note that due to their unorthodox expression for Proj_Y *(see the discussion of this in Example 2.27), their expression for* $\lambda(Y)$ *(which they call* μ*) is different (though equivalent).*

2.1 First-order geometry: material not covered

This section covered a large portion of Chapter 3 from [Bou20] as well as parts of Section 7.7 from [Bou20] through the examples. That said, we still made some omissions. Probably the most important omission has to do with the notation of a "Riemannian submanifold," which was mentioned in several places (e.g., Definition 2.24, Theorem 2.28). Basically, [Bou20] stresses that you can equip manifolds with different inner products. More accurately, you equip each tangent space with an inner product (since tangent spaces are vector spaces), and the inner product on $T_x\mathcal{M}$ can even depend on x.[23] If the inner product on $T_x\mathcal{M}$ varies smoothly with x, then we call it a *Riemannian metric.*[24] A manifold equipped with a Riemannian metric is a *Riemannian submanifold.*[25] It turns out that if you take an embedded submanifold of $\mathbb{R}^d$ and define the inner products on the tangent spaces of the embedded submanifold to just be the restrictions of an inner product on $\mathbb{R}^d$ to those spaces, you obtain a Riemannian submanifold. We refer to this particular type of Riemannian submanifold, where the inner product on $T_x\mathcal{M}$ doesn't vary with x and is inherited from the inner product on $\mathbb{R}^d$, a *Riemannian submanifold of* $\mathbb{R}^d$.[26] Because, as noted in [Bou20], this is the most common type of Riemannian submanifold in practice and it is very natural to work with, we are only covering it briefly here.

It is important to make clear that in these notes, if a theorem/definition in [Bou20] involves a Riemannian submanifold, our "version" of the theorem/definition will basically always additionally assume that it is a Riemannian submanifold of $\mathbb{R}^d$. This way we don't need to add the point where you are taking the inner product as a subscript to each inner product. For example, compare Definition 2.24 to Definition 3.50 in [Bou20]: The latter writes $\langle v, \operatorname{grad} f(x)\rangle_x$ whereas we omit the subscript x. Once again, we do this because Riemannian submanifolds of linear spaces are the most common use case, and it also helps to think about theorems in terms of the inner products we are all used to (i.e., inner products inherited from linear spaces).

[23] Definition 3.43 in [Bou20].
[24] Definition 3.44 in [Bou20].
[25] Definition 3.45 in [Bou20].
[26] Definition 3.47 in [Bou20].

We also didn't cover retractions (Section 3.6 in [Bou20])—this will wait until a later section of these notes.

2.2 First-order geometry: extending beyond the embedded case

To be done.

3 Second-order geometry

In this section we build up to defining the Riemannian Hessian, although we will not end up defining it in full generality. This is due to the fact that the general definition takes a lot of work—you need to build up certain mathematical objects which satisfy nice properties and then prove uniqueness, etc. As expected however, working with the Riemannian Hessian on embedded submanifolds is much easier than in the general case, so we rarely need this machinery in practice. Still, we have sought to give a flavor of how it is actually defined.

Toward doing this, it will be helpful to remember how we should think about the regular Euclidean Hessian, $\nabla^2 f(x)$ for $f : \mathbb{R}^d \to \mathbb{R}$. It can be thought of as an $n \times n$ matrix of second-order partial derivatives, but just as it is typically better to think of $\nabla f(x)$ in terms of the linear map it induces, it is better to think of $\nabla^2 f(x)$ as a linear map. Indeed, we can express $\nabla^2 f(x)$ as a linear map, $\nabla^2 f(x) : \mathbb{R}^d \to \mathbb{R}^d$, where $\nabla^2 f(x)[v]$ is precisely the directional derivative of $\nabla f(x)$ along v. (Of course, behind the scenes $\nabla^2 f(x)[v]$ just corresponds to multiplying the $n \times n$ matrix "version" of $\nabla^2 f(x)$ by $v \in \mathbb{R}^d$. For a more in-depth discussion of all of this, see the beginning of Chapter 5 in [Bou20].) Likewise, the Riemannian Hessian will end up being a linear map, $\mathrm{Hess} f(x) : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_x\mathcal{M}$, where $\mathrm{Hess} f(x)[v]$ should somehow be the directional derivative (whatever that means for manifolds) of the Riemannian gradient $\mathrm{grad} f(x)$ along v.

A good first try for defining the Riemannian Hessian is to take the differential, per Definition 2.18, of the vector field $\mathrm{grad} f(x)$. After all, remember, per Definitions 2.24 and 2.23, that $\mathrm{grad} f(x)$ is really a map: $\mathrm{grad} f(x) : \mathcal{M} \to \mathrm{T}\mathcal{M}$. In other words, it is a map between two embedded submanifolds,[27] and Definition 2.18 tells us exactly how to take the differential of a map between manifolds. However, there is a problem. Note that $D(\mathrm{grad} f)(x) : \mathrm{T}_x\mathcal{M} \to T_{\mathrm{grad} f(x)}\mathrm{T}\mathcal{M}$ per Definition 2.18. So if we define $\mathrm{Hess} f(x)$ as $D(\mathrm{grad} f)(x)$, the domain will indeed be $\mathrm{T}_x\mathcal{M}$ as desired, but it is not clear that the codomain, $T_{\mathrm{grad} f(x)}\mathrm{T}\mathcal{M}$, is equivalent to $\mathrm{T}_x\mathcal{M}$, the latter being what we want the codomain to be. In fact they aren't equivalent, as proven through a simple example in Section 5.1 of [Bou20].

So to recap, we have a vector field $\mathrm{grad} f(x)$, and we want to take the derivative of it in a way such that the codomain is $\mathrm{T}_x\mathcal{M}$. The differential does not do this, so we need a new kind of derivative, which we will call a *connection*.

Definition 3.1 (Connection, informal) *A connection on a manifold $\mathcal{M}$ is an operator*

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

[27] $\mathrm{T}\mathcal{M}$ is indeed an embedded submanifold—see Theorem 3.36 in [Bou20].

where $\mathcal{X}(\mathcal{M})$ is the set of all smooth vector fields on $\mathcal{M}$.[28] *So that we can talk about specific inputs and outputs, we will say that it maps $(u, V) \in \mathrm{T}\mathcal{M} \times \mathcal{X}(\mathcal{M})$ to $\nabla_u V \in \mathrm{T}\mathcal{M}$. It must satisfy the property that if $u \in \mathrm{T}_x\mathcal{M}$, then $\nabla_u V \in \mathrm{T}_x\mathcal{M}$—in other words, the tangent vector that it spits out must lie in the same tangent space as the tangent vector given in the input. Furthermore, a connection needs to satisfy a list of other properties (smoothness, linearity, etc.) so that it really behaves like a derivative. See Definition 5.1 in [Bou20] for the full definition of a connection, which includes a list of these properties.*

So we have a notion of the derivative of a vector field which enforces our rule that it should map vectors in $\mathrm{T}_x\mathcal{M}$ to vectors in $\mathrm{T}_x\mathcal{M}$. Indeed, for a vector field V over a manifold $\mathcal{M}$, this is obtained by fixing the second argument in the connection to be V, so that for a given $x \in \mathcal{M}$ we obtain a map from $\mathrm{T}_x\mathcal{M}$ to $\mathrm{T}_x\mathcal{M}$. We are only a few steps away from being able to define the Riemannian Hessian. It turns out that the list of properties a connection must satisfy (see Definition 5.1 in [Bou20]) is not expansive enough for our purposes. In particular, we can only define the Riemannian Hessian (and the Riemannian gradient for that matter) after introducing an inner product which makes our manifold into a Riemannian manifold.[29] We would like the connection we use to define the Riemannian Hessian to play nicely with this inner product/metric. This is the subject of the next theorem:

Theorem 3.2 (Fundamental theorem of Riemannian geometry, informal) *On a Riemannian manifold $\mathcal{M}$, there exists a unique connection which "plays nicely" with the metric on your manifold. This connection is called the Levi-Civita or Riemannian connection. See Theorem 5.5 in [Bou20] for a formal statement.*

Now we can define the Riemannian Hessian:

Definition 3.3 (Riemannian Hessian) *Let $\mathcal{M}$ be a Riemannian manifold with its Riemannian connection ∇. Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function. The Riemannian Hessian*[30] *of f at $x \in \mathcal{M}$ is a linear operator $\mathrm{Hess} f(x) : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_x\mathcal{M}$ defined as follows*

$$\mathrm{Hess} f(x)[u] = \nabla_u \mathrm{grad} f.$$

Recall from Definition 3.1 that $\nabla_u \mathrm{grad} f$ denotes the result you get from feeding $u \in \mathrm{T}_x\mathcal{M}$ and $\mathrm{grad} f \in \mathcal{X}(\mathcal{M})$ into the connection ∇ as its input.

The following intuitively had to be true, but let's make it explicit:

Theorem 3.4 (The Riemannian Hessian is self-adjoint) *The Riemannian Hessian is self-adjoint.*[31] *In particular, when we are working with real numbers, the matrix corresponding to the Riemannian Hessian is symmetric.*

[28] [Bou20] uses similar notation for the set of all vector fields on $\mathcal{M}$—see Definition 3.37.
[29] See Section 2.1.
[30] Definition 5.13 in [Bou20].
[31] Proposition 5.16 in [Bou20].

Now we begin the process of simplifying all we have done above for the case where $\mathcal{M}$ is a Riemannian submanifold of a Euclidean space. First, we obtain an explicit expression for the Riemannian connection:

Theorem 3.5 (Riemannian connection of a Riemannian submanifold of $\mathbb{R}^d$) *Let $\mathcal{M}$ be a Riemannian submanifold of a Euclidean space (i.e., $\mathbb{R}^d$). Then the Riemannian connection is precisely*[32]

$$\nabla_u V = \mathrm{Proj}_x(D\overline{V}(x)[u]),$$

where $\overline{V}$ is any smooth extension of V. (Here V is a smooth vector field on $\mathcal{M}$ and $u \in \mathrm{T}\mathcal{M}$.)

This is incredibly intuitive based on the discussion we had toward the beginning of this chapter. Remember that the whole problem with using the differential to obtain the derivative of a vector field was that the result might not be in $\mathrm{T}_x\mathcal{M}$. Well, Theorem 3.5 is basically saying that the "correct" way to fix this for Riemannian submanifolds of Euclidean spaces is to just force the output to lie in $\mathrm{T}_x\mathcal{M}$ through orthogonal projection!

Definition 3.3 and Theorem 3.5 combine to instantly yield the following corollary:

Corollary 3.6 (Riemannian Hessian on a Riemannian submanifold of $\mathbb{R}^d$) *Let $\mathcal{M}$ be a Riemannian submanifold of the Euclidean space $\mathbb{R}^d$. Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function. Let $\overline{\mathrm{grad} f} : \mathbb{R}^d \to \mathbb{R}^d$ denote a smooth extension of $\mathrm{grad} f$. Then*[33]

$$\mathrm{Hess} f(x)[u] = \mathrm{Proj}_x(D(\overline{\mathrm{grad} f})(x)[u]).$$

In other words, for a Riemannian submanifold of $\mathbb{R}^d$, you can compute the Riemannian Hessian by differentiating *in the classical sense* (i.e., using Definition 2.1) a smooth extension of the Riemannian gradient vector field, and then orthogonally projecting the result to $\mathrm{T}_x\mathcal{M}$.

Note that $D(\overline{\mathrm{grad} f})(x)[u]$ is *not* equivalent to $\nabla^2 f(x)[u]$, meaning Corollary 3.6 is *not* saying that $\mathrm{Hess} f(x)[u]$ is just the projection of the Euclidean Hessian to $T_x\mathcal{M}$, aka $\mathrm{Proj}_x(\nabla^2 f(x)[u])$. In particular, $D(\overline{\mathrm{grad} f})(x)[u]$ and $\nabla^2 f(x)[u]$ are not equivalent because $\overline{\mathrm{grad} f}$ and ∇f are not equivalent. The latter are not equivalent because they need not coincide on the manifold $\mathcal{M}$ itself. Indeed, by definition $\overline{\mathrm{grad} f}$ and $\mathrm{grad} f$ must coincide on $\mathcal{M}$, and we know from Theorem 3.5 that $\mathrm{grad} f$ and ∇f need not coincide on $\mathcal{M}$. (After all, what then would even be the point of defining the Riemannian gradient, as it would always be equivalent to the Euclidean gradient?)

However, it is still possible to directly relate the Riemannian Hessian to the Euclidean Hessian. We only provide an informal statement here:

Theorem 3.7 (Relating the Riemannian Hessian to the Euclidean Hessian, informal) *Let $\mathcal{M}$ be a Riemannian submanifold of $\mathbb{R}^d$. Let $f : \mathcal{M} \to \mathbb{R}$ be smooth with smooth extension $\overline{f} : \mathbb{R}^d \to \mathbb{R}$. The Riemannian Hessian of f is given by:*

$$\mathrm{Hess} f(x)[u] = \mathrm{Proj}_x(\nabla^2 \overline{f}(x)[u]) + \mathrm{correction}(\mathrm{Proj}_x^{\perp}(\nabla \overline{f}(x))),$$

[32] Theorem 5.8 in [Bou20].
[33] Corollary 5.14 in [Bou20].

where $\mathrm{Proj}_x^{\perp} = \mathrm{Id} - \mathrm{Proj}_x$ *denotes the orthogonal projector onto the normal space* $\mathrm{N}_x\mathcal{M}$*, and* correction $: \mathbb{R}^d \to \mathrm{T}_x\mathcal{M}$ *is some "correctional function" which we do not specify here. (*Id *is the identity map.)*

See Corollary 5.17 in [Bou20] for the formal statement.

In other words, this says that for Riemannian submanifolds of Euclidean spaces, the Riemannian Hessian is the projected Euclidean Hessian plus a correctional term which depends only on the normal part of the Euclidean gradient.

Example 3.8 (Riemannian Hessian in the optimization problem over $\mathcal{M}_0$) *Using the notation of Example 2.29, recall that*

$$\mathrm{grad} f(y) = \nabla \overline{f}(y) - \sum_{i=1}^{m} \lambda_i(y) \nabla h_i(y) \tag{4}$$

where

$$\lambda(y) := (Dh(y)^T)^+ [\nabla \overline{f}(y)]. \tag{5}$$

We now wish to apply Corollary 3.6 to obtain an expression for $\mathrm{Hess} f(x)$*. To do so we need to obtain a smooth extension of* $\mathrm{grad} f$*, which we will denote* $\overline{\mathrm{grad} f}$*. Well, the smooth extensions of* $\nabla \overline{f}(y)$ *and* $\nabla h_i(y)$ *can just be* $\nabla \overline{f}(y)$ *and* $\nabla h_i(y)$ *respectively.*[34] *The only question remaining is how to smoothly extend* $\lambda(y)$*. Well, from (5) we can view* $\lambda(y)$ *as a smooth function on the open subset of* $\mathbb{R}^d$ *consisting of all points* y *where* $Dh(y)$ *has full rank. This is an open neighborhood of* $\mathcal{M}_0$*, so we have our smooth extension. In other words,*

$$\overline{\mathrm{grad} f(x)} = \nabla \overline{f}(y) - \sum_{i=1}^{m} \lambda_i(y) \nabla h_i(y), \tag{6}$$

where we have replaced $\nabla \overline{f}(y), \lambda_i(y), \nabla h_i(y)$ *with smooth extensions of themselves.*

Then we can differentiate (6) normally (i.e., using Definition 2.1) to get

$$D(\overline{\mathrm{grad} f(x)})(x)[u] = \nabla^2 \overline{f}(y)[u] - \sum_{i=1}^{m} D\lambda_i(x)[u] \cdot \nabla h_i(x) - \sum_{i=1}^{m} \lambda_i(y) \nabla^2 h_i(y)[u].$$

Applying Corollary 3.6 yields

$$\begin{aligned} \mathrm{Hess} f(y)[u] &= \mathrm{Proj}_x \left(\nabla^2 \overline{f}(y)[u] - \sum_{i=1}^{m} D\lambda_i(x)[u] \cdot \nabla h_i(x) - \sum_{i=1}^{m} \lambda_i(y) \nabla^2 h_i(y)[u] \right) \\ &= \mathrm{Proj}_x \left(\nabla^2 \overline{f}(y)[u] - \sum_{i=1}^{m} \lambda_i(y) \nabla^2 h_i(y)[u] \right), \end{aligned} \tag{7}$$

[34] Formally, we are extending the domain of $\nabla \overline{f}(y)$ and $\nabla h_i(y)$ as they appear in (4) from $\mathcal{M}$ to $\mathbb{R}^d$. Although in other sense, the $\nabla \overline{f}(y)$ and $\nabla h_i(y)$ which appear in (4) are really restrictions of the "original" $\nabla \overline{f}(y)$ and $\nabla h_i(y)$ from $\mathbb{R}^d$ to $\mathcal{M}$, so it is more like we are lifting this restriction.

where the second equality follows because

$$\mathrm{Proj}_x\left(\sum_{i=1}^{m} D\lambda_i(x)[u]\cdot\nabla h_i(x)\right) = 0$$

since each $\nabla h_i(x) \in \mathrm{N}_y\mathcal{M}$ *(see Example 2.11).*

Note that (7) looks super similar to the second-order criticality condition in nonlinear programming (in other words, the Hessian of the Lagrangian)—more on this later.

Example 3.9 (Riemannian Hessian in the Burer-Monteiro problem) *Applying (7), we obtain*

$$\begin{aligned}\mathrm{Hess} f(Y)[U] &= 2\cdot\mathrm{Proj}_Y\left(CU - \sum_{i=1}^{m}\lambda_i(Y)A_iU\right)\\ &= 2\cdot\mathrm{Proj}_Y\left(\left[C - \sum_{i=1}^{m}\lambda_i(Y)A_i\right]U\right).\end{aligned}$$

where $\lambda(Y)$ *is as in Example 2.30 and* Proj_Y *is as in Example 2.27.*

As a sanity check, this is precisely what the authors of [BVB18] derive—see (2.7) in that paper.

3.1 Second-order geometry: material not covered

We went over much of the first half of Chapter 5 of [Bou20] in this section. Obviously some formality regarding connections was left out, but these details aren't really necessary for working with Riemannian submanifolds of Euclidean spaces due to Corollary 3.6 and Theorem 3.7. It is worth mentioning that Definition 3.1 (and the corresponding Definition 5.1 in [Bou20]) is not actually the standard definition of the connection. However, it is equivalent to the standard definition (Definition 5.19 in [Bou20]), as proven in Section 5.6 of [Bou20]. Boumal chooses to start with a nonstandard definition because it makes it easier to define the Riemannian Hessian.

The second half of Chapter 5 in [Bou20] contains other important material: differentiating vector fields along curves, acceleration, geodesics, second-order Taylor expansions on curves, second-order retractions, etc. We will see some of this later in these notes.

Finally, we followed Section 7.7 of [Bou20] in Example 3.8. It is worth noting that in Section 7.8 of [Bou20], an alternative derivation of the Riemannian Hessian for this particular problem is given. The latter approach mirrors the derivation of the second-order necessary condition in nonlinear programming (see page 333 in [LY16]), and as noted in [Bou20], can be extended more easily to higher-order derivatives.

3.2 Second-order geometry: extending beyond the embedded case

To be done.

4 Retractions, Taylor expansions, and optimality conditions

We now discuss the tool used to move along manifolds in a particular direction. Formally, given a point $x \in \mathcal{M}$ and a tangent vector $v \in \mathrm{T}_x\mathcal{M}$, starting at x and moving in the direction v corresponds to picking a curve $c : I \subseteq \mathbb{R} \to \mathcal{M}$ such that $c(0) = x$ and $c'(0) = v$, and then moving along c. There may be many such curves. A *retraction* picks a particular curve for each $(x, v) \in \mathrm{T}\mathcal{M}$ in a smooth way.

Definition 4.1 (Retraction) *A retraction*[35] *on $\mathcal{M}$ is a smooth map $\mathrm{R} : \mathrm{T}\mathcal{M} \to \mathcal{M}$ with the following properties. For each $x \in \mathcal{M}$, let $\mathrm{R}_x : \mathrm{T}_x\mathcal{M} \to \mathcal{M}$ denote the restriction of R at x, so that $\mathrm{R}_x(v) = \mathrm{R}(x, v)$. Then*

1. *$\mathrm{R}_x(0) = x$, and*
2. *$\mathrm{DR}_x(0) : \mathrm{T}_x\mathcal{M} \to \mathrm{T}_x\mathcal{M}$ is the identity map: $\mathrm{DR}_x(0)[v] = v$.*

Equivalently, each curve $c(t) = \mathrm{R}_x(tv)$ satisfies $c(0) = x$ and $c'(0) = v$.

First of all, note that $\mathrm{DR}_x(0)$ is indeed a map from $\mathrm{T}_x\mathcal{M}$ to $\mathrm{T}_x\mathcal{M}$ per Definition 2.18. In particular, the domain is $\mathrm{T}_x\mathcal{M}$ because the tangent space to $\mathrm{T}_x\mathcal{M}$ is $\mathrm{T}_x\mathcal{M}$, due to the fact that $\mathrm{T}_x\mathcal{M}$ is a linear space.

Next, let us interpret the conditions of Definition 4.1. The first condition is easy to interpret, but what of $\mathrm{DR}_x(0)[v] = v$? Well, it is helpful to go back to Definition 2.18 and the text below it. Remember that $\mathrm{DR}_x(0)[v]$ intuitively tells us how R_x changes when we start with $0 \in \mathrm{T}_x\mathcal{M}$ as the input and push the input in the direction v. So the fact that $\mathrm{DR}_x(0)[v] = v$ says that if we push the input along v, the output (which starts at $\mathrm{R}_x(0)$) had better also be pushed along v. (Or more precisely, e.g., the curve $\mathrm{R}_x(tv) : \mathbb{R} \to \mathcal{M}$ better have velocity v when $t = 0$, which is exactly the text at the bottom of Definition 4.1.) In other words, the $\mathrm{DR}_x(0)[v] = v$ condition intuitively forces the retraction to actually "send us in the direction we specify."

Example 4.2 (Retractions on a sphere) *The following are natural retractions on the sphere S^{n-1}:*

$$\mathrm{R}_x(v) = \frac{x + v}{||x + v||_2} = \frac{x + v}{\sqrt{1 + ||v||^2}}$$

and

$$\mathrm{R}_x(v) = \cos(||v||)x + \frac{\sin(||v||)}{||v||}v. \tag{8}$$

The former is simply projection. The latter has you move along the great circle traced out by v. See Example 3.40 in [Bou20] for more details.

[35] Definition 3.41 in [Bou20]

Example 4.3 (Retraction on $\mathcal{M}_0$) *A natural retraction on $\mathcal{M}_0$ is the metric projection:*

$$\mathrm{R}_x(v) = \operatorname*{argmin}_{y \in \mathcal{M}_0} ||x + v - y||_2. \tag{9}$$

In other words, we return the point on the manifold closest to $x + v$. This is well-defined for small v, but $\mathrm{R}_x(v)$ may not be uniquely defined for larger v. It may also be difficult to compute this retraction.

Example 4.4 (Retractions on $\mathcal{M}_p$) *As far as I know, the particularization of (9) to $\mathcal{M}_p$ does not yield a nice expression in general.*[36] *However, for particular instantiations of $\mathcal{M}_p$, (9) might yield a nice expression (which I'm guessing is why manifold-based algorithms for optimization on $\mathcal{M}_p$ seem to be limited to these special cases) or other nice retractions might exist. See [Bou16] and [JBAS10] for examples.*

The following object is key to the general method of designing manifold optimization algorithms:

Definition 4.5 (Pullback) *Let $f : \mathcal{M} \to \mathbb{R}$ be smooth and let $\mathrm{R} : \mathrm{T}\mathcal{M} \to \mathcal{M}$ be a retraction on $\mathcal{M}$. The composition $f \circ \mathrm{R} : \mathrm{T}\mathcal{M} \to \mathbb{R}$ is called the pullback*[37] *of f to the tangent spaces. In particular, $f \circ \mathrm{R}_x : \mathrm{T}_x\mathcal{M} \to \mathbb{R}$ is the pullback of f to the tangent space at x. Importantly, this is a smooth function on a linear space.*

Since typical optimization algorithms on manifolds are retraction based, meaning they iterate

$$x_{k+1} = \mathrm{R}_{x_k}(s_k)$$

for some step s_k, the change in the cost function value from one iterate (x_k) to the next (x_{k+1}) can be understood through the pullback $f \circ \mathrm{R}_{x_k}$. In other words, we want to pick $s_k \in \mathrm{T}_{x_k}\mathcal{M}$ so as to minimize $f \circ \mathrm{R}_{x_k}$. However, $f \circ \mathrm{R}_{x_k}$ may in general be very complicated, so we seek to approximate it using Taylor expansions. We then minimize the approximation instead of $f \circ \mathrm{R}_{x_k}$ itself.[38] This is the basis for both Riemannian gradient descent and Riemannian Newton's method—more on this later. First we need to write out the relevant Taylor expansions, which will also help us derive optimality conditions.

Theorem 4.6 (First-order Taylor expansion on curves) *Let $\mathcal{M}$ be a Riemannian submanifold of $\mathbb{R}^d$,*[39] *and let $f : \mathcal{M} \to \mathbb{R}$ be smooth. Let $c : I \to \mathcal{M}$ be a smooth curve on $\mathcal{M}$ with $c(0) = x$ and $c'(0) = v$, where I is an open interval of $\mathbb{R}$ around $t = 0$. Then*[40]

$$f(c(t)) = f(x) + t \left\langle \operatorname{grad} f(x), v \right\rangle + O(t^2).$$

The big O notation here is with respect to t going to 0, meaning that the residual goes to zero at least as fast as t^2 as $t \to 0$.

[36] I haven't actually tried to compute this though. But I'm guessing this is the case since I have not seen such a result.

[37] Section 4.1 in [Bou20].

[38] For reference, I took this discussion from Section 6.2 of [Bou20].

[39] See Section 2.1 for a reminder of what this means.

[40] Section 4.1 in [Bou20].

Theorem 4.7 (First-order Taylor expansion of the pullback) *With the same setup as in Theorem 4.6, suppose we additionally have a retraction* $\mathrm{R} : \mathrm{T}\mathcal{M} \to \mathcal{M}$. *Then*[41]

$$f(\mathrm{R}_x(s)) = f(x) + t\,\langle \mathrm{grad} f(x), s\rangle + O(||s||^2).$$

The proofs of Theorems 4.6 and 4.7 are very short and simple. Theorem 4.6 is proven by a simple application of the chain rule. Theorem 4.7 is proven by applying Theorem 4.6 to the curve $c(t) = \mathrm{R}_x(tv)$ and then introducing $s = tv$. See Section 4.1 in [Bou20] for the details.

Theorem 4.8 (First-order necessary condition) *Let* $f : \mathcal{M} \to \mathbb{R}$ *be a smooth function on a Riemannian manifold. If* x *is a local minimizer of* f, *then* $\mathrm{grad} f(x) = 0$.[42]

Once again, the proof is simple. If $\mathrm{grad} f(x)$ were not zero, then the first-order Taylor expansion reveals that f will decrease along any curve with velocity $-\mathrm{grad} f(x)$.

Definition 4.9 (Critical/stationary point) *Given a smooth function* f *on a Riemannian manifold* $\mathcal{M}$, *we call* $x \in \mathcal{M}$ *a critical point or stationary point*[43] *of* f *if* $\mathrm{grad} f(x) = 0$.

Now we move into the second-order realm where things inevitably get more complicated:

Theorem 4.10 (Second-order Taylor expansion on curves) *With the same setup as in Theorem 4.6, we have*[44]

$$f(c(t)) = f(x) + t\,\langle \mathrm{grad} f(x), v\rangle + \frac{t^2}{2}\,\langle \mathrm{Hess} f(x)[v], v\rangle + \frac{t^2}{2}\,\big\langle \mathrm{grad} f(x), c''(0)\big\rangle + O(t^3). \tag{10}$$

Clearly this is not as clean of an expression as the first-order Taylor expansion due to the annoying $\frac{t^2}{2}\,\langle \mathrm{grad} f(x), c''(0)\rangle$ term (which in particular depends on c beyond just the velocity of c at $t = 0$). In fact, we have not yet even defined what $c''(0)$ means exactly. $c'(t)$ can be thought of as a function which maps each t in some open interval of $\mathbb{R}$ containing 0 to a tangent vector at $c(t)$. Indeed, $c'(t) \in \mathrm{T}_{c(t)}\mathcal{M}$ by the literal definition of the tangent space (Definition 2.6). More generally, $c'(t)$ is an example of a *smooth vector field on a curve* which is essentially produced by taking a curve c and composing it with a smooth vector field. Formally:

Definition 4.11 (Smooth vector field on a curve) *Given a smooth curve* $c : I \to \mathcal{M}$ *(I is open and connected in $\mathbb{R}$), the map* $Z : I \to \mathrm{T}\mathcal{M}$ *is a smooth vector field on* c *if* $Z(t)$ *is in* $\mathrm{T}_{c(t)}\mathcal{M}$ *for all* $t \in I$, *and if it is smooth as a map from* I *(an open submanifold of* $\mathbb{R}$*) to* $\mathrm{T}\mathcal{M}$.[45]

Just as we needed to define a new kind of derivative, the connection, in Section 3 so that the Hessian spits out tangent vectors, we need to define a new kind of derivative to "properly" take

[41] Section 4.1 in [Bou20].
[42] Proposition 4.4 in [Bou20].
[43] Definition 4.5 in [Bou20].
[44] Section 5.9 in [Bou20].
[45] Definition 5.27 in [Bou20].

the derivative of a smooth vector field along a curve. It is called the *induced covariant derivative*, and it is defined in a similar way to how the Riemannian connection (Theorem 3.2) is defined: You basically show that there exists a unique operator which satisfies the "nice properties" which the derivative of a vector field along a curve "should satisfy." See Theorem 5.28 in [Bou20] for the formal statement.

Just like for the connection, one of the most important properties of the induced covariant derivative is that the result of applying it to a smooth vector field along a curve should produce a smooth vector field along a curve, meaning in particular that the result is still assigning tangent vectors to each $t \in I$ and not arbitrary vectors in the linear embedding space. Furthermore, just like for the connection (Theorem 3.5), in the cases we care about in these notes, the induced covariant derivative just corresponds to composing the regular Euclidean derivative with an orthogonal projection onto the tangent space. (See Proposition 5.29 in [Bou20] for a formal statement.) For this reason, *the $c''(0)$ in Theorem 4.10 is just the regular Euclidean derivative of $c'(0)$ composed with an orthogonal projection onto the tangent space.* Formally,

$$c''(t) := \mathrm{Proj}_{c(t)}\left(\frac{\mathrm{d}}{\mathrm{d}t}c'(t)\right), \tag{11}$$

where $\frac{\mathrm{d}}{\mathrm{d}t}c'(t)$ is the Euclidean derivative of $c'(t)$. In fact, $c'(t)$ is formally called the *acceleration*[46] of the curve c. The classical or *extrinsic* acceleration of c is denoted in [Bou20] (see Section 5.8) as

$$\ddot{c}(t) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}c(t) = \frac{\mathrm{d}}{\mathrm{d}t}c'(t)$$

Then (11) is equivalent to

$$c''(t) := \mathrm{Proj}_{c(t)}\left(\ddot{c}(t)\right), \tag{12}$$

which says that for a Riemannian submanifold of $\mathbb{R}^d$, the acceleration of a curve c is the tangential part of its extrinsic acceleration in the embedding space.

This is all great, but we are still left with this ugly $\frac{t^2}{2}\langle \mathrm{grad} f(x), c''(0)\rangle$ term in Theorem 4.10. Well, fortunately it turns out that in most of the cases we care about, this term will get zeroed out. The first case where this will happen is when we are at a stationary point, since then $\mathrm{grad} f(x) = 0$. This yields the following:

Theorem 4.12 (Second-order necessary condition) *Consider a smooth function $f : \mathcal{M} \to \mathbb{R}$ defined on a Riemannian manifold $\mathcal{M}$. If x is a local minimizer of f, then* $\mathrm{grad} f(x) = 0$ *and* $\mathrm{Hess} f(x) \succeq 0$.[47]

The proof is simple. We already know from Theorem 4.8 that x being a local minimum implies $\mathrm{grad} f(x) = 0$, so (10) becomes

$$f(c(t)) = f(x) + \frac{t^2}{2}\langle \mathrm{Hess} f(x)[v], v\rangle + O(t^3). \tag{13}$$

Then if Hess weren't positive semidefinite, (13) implies that f decreases along a curve whose velocity is an eigenvector of Hess with a negative corresponding eigenvalue.

[46] Definition 5.35 in [Bou20].
[47] Proposition 6.1 in [Bou20].

Definition 4.13 (Second-order critical/stationary point) *Given a smooth function f on a Riemannian manifold $\mathcal{M}$, we call $x \in \mathcal{M}$ a second-order critical point or a second-order stationary point*[48] *of f is $\operatorname{grad} f(x) = 0$ and $\operatorname{Hess} f(x) \succeq 0$.*

Example 4.14 (Necessary conditions for optimization over $\mathcal{M}_0$) *We derive the criticality conditions for optimization over $\mathcal{M}_0$. Combining (3) and Theorem 4.8/Definition 4.9, we have that $y \in \mathcal{M}_0$ is a first-order critical point if and only if*

$$\operatorname{grad} f(y) = \nabla \overline{f}(y) - \sum_{i=1}^{m} \lambda_i(y) \nabla h_i(y) = 0 \tag{14}$$

where

$$\lambda(y) := (Dh(y)^T)^+ [\nabla \overline{f}(y)].$$

Note that this is equivalent to the first-order nonlinear programming criticality condition with respect to the optimization of $\overline{f}$ over $\mathcal{M}_0$. Indeed, we say that y is a first-order critical point with respect to the latter problem if $Dh(y)$ has full rank (i.e., LICQ holds) and there exists unique multipliers $\nu \in \mathbb{R}^m$ such that

$$\nabla \overline{f}(y) - \sum_{i=1}^{m} \nu_i \nabla h_i(y) = 0$$

The fact that $\nabla h_1(y), \ldots, \nabla h_m(y)$ are linearly independent (see Example 2.3) implies that $\lambda(y) = \nu$ at a first-order critical point.

From (7) and Theorem 4.12/Definition 4.13, we have that $y \in \mathcal{M}_0$ is a second-order critical point if and only if (14) holds and

$$\operatorname{Proj}_x \circ \left(\nabla^2 \overline{f}(y) - \sum_{i=1}^{m} \lambda_i(y) \nabla^2 h_i(y) \right) \circ \operatorname{Proj}_x \succeq 0. \tag{15}$$

(Note that we can precompose (7) with Proj_x to get the equivalent expression (15) since the input to the Riemannian Hessian lies in $\mathrm{T}_x\mathcal{M}$ anyway by the definition of the Riemannian Hessian. We follow [Bou20]—see (7.79) in that reference—and make this change since (15) is "pleasantly symmetric.")

It is immediately apparent that this is equivalent to the second-order criticality condition in nonlinear programming. See Exercise 7.8 in [Bou20] for the details.

Example 4.15 (Necessary conditions for the Burer-Monteiro problem) *Combining Example 2.30 with Theorem 4.8/Definition 4.9, we have that $Y \in \mathcal{M}_p$ is a first-order critical point if and only if*

$$\operatorname{grad} f(Y) = 2 \left[\left(C - \sum_{i=1}^{m} \lambda_i(Y) A_i \right) Y \right] = 0 \tag{16}$$

[48] Definition 6.2 in [Bou20].

where

$$\lambda(Y) := (Dh(Y)^T)^+[2CY].$$

From applying (15) to the Burer-Monteiro problem, we get that $Y \in \mathcal{M}_p$ is a second-order critical point if and only if (16) holds and

$$2\left[C - \sum_{i=1}^{m} \lambda_i(Y) A_i\right] \bullet UU^T \geq 0$$

for all $U \in \mathrm{T}_Y\mathcal{M}_p$. (See Example 2.9 for an expression for $\mathrm{T}_Y\mathcal{M}_p$.) Clearly these are equivalent to the nonlinear programming criticality conditions for the Burer-Monteiro problem. (See also Example 4.14 above.)

As a sanity check, these are indeed equivalent to the conditions given in [BVB18]—see Definition 2.3 in that reference.

It turns out that even if we aren't at a critical point, there is a way to remove the ugly $\frac{t^2}{2}\langle \mathrm{grad} f(x), c''(0)\rangle$ term. If we can't make $\mathrm{grad} f(x)$ equal to zero, we just need to make the acceleration $c''(0)$ equal to zero! This leads to a special kind of curve:

Definition 4.16 (Geodesic) *On a Riemannian manifold $\mathcal{M}$, a geodesic[49] is a smooth curve $c : I \to \mathcal{M}$ such that $c''(t) = 0$ for all $t \in I$, where I is an open interval of $\mathbb{R}$.*

Due to (12), a curve c on a Riemannian submanifold $\mathcal{M}$ is a geodesic if and only if its classical (or extrinsic) acceleration $\ddot{c}$ is everywhere normal to $\mathcal{M}$.

We can also use this idea to define a special kind of retraction:

Definition 4.17 (Second-order retraction) *A second-order retraction[50] R on a Riemannian manifold $\mathcal{M}$ is a retraction such that, for all $x \in \mathcal{M}$ and all $v \in \mathrm{T}_x\mathcal{M}$, the curve $c(t) = \mathrm{R}_x(tv)$ has zero acceleration at $t = 0$, that is, $c''(0)$.*

What can we say about the existence of geodesics and second-order retractions? Well, it turns out that geodesics exist between any two sufficiently close points on an arbitrary Riemannian manifold.[51] If the Riemannian manifold is additionally complete[52] (which is satisfied in particular by compact manifolds[53] and finite-dimensional Euclidean spaces[54]), then any two points x, y in the same connected component can be connected by a geodesic.[55] It turns out that every Riemannian manifold admits a special second-order retraction called the exponential map.[56] In general, the

[49] Definition 5.37 in [Bou20].
[50] Definition 5.41 in [Bou20]
[51] See the note to the right of the proof of Proposition 6.3 in [Bou20].
[52] See Definitions 10.5 and 10.6 as well as Theorem 10.7 in [Bou20].
[53] Example 10.9 in [Bou20].
[54] Example 10.10 in [Bou20].
[55] Theorem 10.8 in [Bou20].
[56] Definition 10.13 in [Bou20].

exponential map may only be defined on an open subset of $\mathrm{T}\mathcal{M}$, but if the Riemannian manifold is additionally complete, then it is defined on all of $\mathrm{T}\mathcal{M}$.[57] As an example, (8) is the exponential map on S^{n-1}.[58] The curve $c(t) = \mathrm{R}_x(tv)$ defined by (8) is a geodesic.[59]

Using a second-order retraction, we can obtain a second-order version of Theorem 4.7:

Theorem 4.18 (Second-order Taylor expansion of the pullback) *Consider a Riemannian manifold $\mathcal{M}$ of $\mathbb{R}^d$ equipped with a second-order retraction R. Then for all points $x \in \mathcal{M}$ we have*[60]

$$f(\mathrm{R}_x(s)) = f(x) + \langle \mathrm{grad} f(x), s\rangle + \frac{1}{2}\langle \mathrm{Hess} f(x)[s], s\rangle + O(||s||^3).$$

Finally, we give a second-order sufficient condition:

Theorem 4.19 (Second-order sufficient condition) *Consider a smooth function $f : \mathcal{M} \to \mathbb{R}$ defined on a Riemannian manifold $\mathcal{M}$. If $\mathrm{grad} f(x) = 0$ and $\mathrm{Hess} f(x) \succ 0$, then x is a strict local minimizer of f.*[61]

The fact that there exists a geodesic between any two sufficiently close points on a Riemannian manifold is useful for proving this, since it lets us remove dependence on the acceleration c'' from our Taylor expansions (in a neighborhood at least).

5 Quotient manifolds

Quotient manifolds are a powerful tool to gain further geometric insight into manifolds which can be decomposed into equivalence classes. Very informally speaking, a quotient manifold is the manifold formed by the equivalence classes of another manifold. The latter is known as the *total space*. As each element of a quotient manifold is an equivalence class, quotient manifolds are rather abstract objects. As a result, we seek to understand properties of and operators on quotient manifolds in terms of the corresponding properties of and operators on the total space. Indeed, this is much of what Chapter 9 of [Bou20] focuses on.

One might worry that the abstractness of quotient manifolds would hinder the development of effective algorithms on the quotient manifold. Fortunately this isn't the case. Running Riemannian gradient descent on the quotient manifold is actually equivalent to running Riemannian gradient descent on the total space. So in practice, we never actually run Riemannian gradient descent on the quotient manifold itself but instead run it on the simpler total space. Still, identifying Riemannian gradient descent as being equivalent on both can lead to better theoretical guarantees,[62] since while the Riemannian Hessian on the total space will never be positive definite, it may be positive definite

[57] See the text below Definition 10.13 in [Bou20].
[58] Example 10.18 in [Bou20].
[59] Example 5.36 in [Bou20].
[60] Proposition 5.43 in [Bou20].
[61] Proposition 6.3 in [Bou20].
[62] See Section 9.9 in [Bou20] for a discussion of this.

on the quotient manifold. Running Riemannian Newton's method on the total space *isn't* equivalent to running it on the quotient manifold, although the two return numerically close solutions.[63]

To get started, let $\sim$ be an equivalence relation on a manifold $\overline{\mathcal{M}}$ with equivalence classes

$$[x] = \{y \in \overline{\mathcal{M}} \mid x \sim y\},$$

and let

$$\mathcal{M} = \overline{\mathcal{M}}/\sim = \{[x] : x \in \overline{\mathcal{M}}\}$$

denote the resulting quotient set. $\overline{\mathcal{M}}$ is known as the *total space*,[64] and the following is the key operator which links the total space $\overline{\mathcal{M}}$ to the quotient set $\mathcal{M}$:

Definition 5.1 (Canonical/natural projection) *The canonical projection or natural projection $\pi : \overline{\mathcal{M}} \to \mathcal{M}$ links the total space $\overline{\mathcal{M}}$ to the quotient $\mathcal{M}$ in the following way:*

$$\pi : x \mapsto \pi(x) = [x].$$

In other words, π sends $x \in \overline{\mathcal{M}}$ to its equivalence class.

So far we have been careful to refer to $\mathcal{M}$ as a quotient set, but we would like to know when it is also a quotient manifold. Unfortunately, fully understanding the definition of a quotient manifold requires one to also understand the definition of a general manifold (as opposed to just the special case of an embedded submanifold of a linear space—Definition 2.2—which we have covered in these notes). This is because even if the total space $\overline{\mathcal{M}}$ is an embedded submanifold of $\mathbb{R}^d$, the quotient set $\mathcal{M}$ may not be identifiable as an embedded submanifold of a linear space. (At least, I think this is the case due to the fact that it is really a set of equivalence classes.) Indeed, this is why the chapter on quotient manifolds (Chapter 9) comes after the chapter on general manifolds (Chapter 8) in [Bou20]. Still, as we will see, we can get very far in our practical understanding of quotient manifolds if we just accept the definition of a general manifold to be a black box.

Definition 5.2 (Quotient manifold) *Suppose we endow the quotient set $\mathcal{M} = \overline{\mathcal{M}}/\sim$ with "smooth structure" so that it is a manifold*[65] *in its own right. If additionally the canonical projection π is smooth and its differential $\mathrm{D}\pi(x) : \mathrm{T}_x\overline{\mathcal{M}} \to \mathrm{T}_{[x]}\mathcal{M}$ has rank $\dim \mathcal{M}$ (i.e., it is full rank) for all $x \in \overline{\mathcal{M}}$, then we say that $\mathcal{M}$ is a quotient manifold.*[66]

$\mathrm{D}\pi(x)$ is a very important operator in its own right: Just as π provides a connection between points on $\overline{\mathcal{M}}$ and points on $\mathcal{M}$, $D\pi(x)$ provides a connection between tangent vectors on $\overline{\mathcal{M}}$ and tangent vectors on $\mathcal{M}$ (or more accurately, between tangent vectors in $\mathrm{T}_x\overline{\mathcal{M}}$ and tangent vectors in $\mathrm{T}_{[x]}\mathcal{M}$). Indeed, we will not actually formally cover in these notes what a tangent vector on

[63] See Section 9.12 in [Bou20].

[64] Section 9.1 in [Bou20]. We follow the notational convention of [Bou20] in that the total space is always denoted $\overline{\mathcal{M}}$ and the quotient set is always denoted $\mathcal{M}$.

[65] Roughly, you can think of a general manifold as being a space which locally resembles Euclidean space. See Definition 8.20 in [Bou20] for a formal definition.

[66] Definition 9.1 in [Bou20].

$\mathcal{M}$ is, since as discussed earlier, $\mathcal{M}$ may not be an embedded submanifold of a linear space even if $\overline{\mathcal{M}}$ is. (Definition 2.6 is only for embedded submanifolds of a linear space, although the general definition of a tangent vector, Definition 8.29 in [Bou20], is a natural extension.) That said, we can still understand tangent vectors on $\mathcal{M}$ using $\mathrm{D}\pi(x)$. In fact, even with the definition of a general tangent vector, it is still typically more helpful to understand tangent vectors on $\mathcal{M}$ through $\mathrm{D}\pi(x)$!

Typically we do not actually use Definition 5.2 to check that a quotient set is a quotient manifold; instead, we use properties of the equivalence relation $\sim$. In fact, there exists a characterization of which equivalence relations give rise to quotient manifolds,[67] but it is also unwieldy. Fortunately, there exists a class of equivalence relations defined through *group actions* on manifolds that are simple to identify and ubiquitous in practice. We will not formally go over this sufficient condition for $\mathcal{M}$ to be a quotient manifold,[68] but we will provide intuition for it through an example.

First, we will informally go over a few definitions. A *Lie group* is a set that is both a group (in the abstract algebra sense) and a manifold. Examples include $\mathrm{O}(n)$ (the orthogonal group or the set of $n \times n$ orthogonal matrices), $\mathrm{SO}(n)$ (the rotation group or the set of $n \times n$ orthogonal matrices with determinant 1), and $\mathrm{GL}(n)$ (the general linear group or the set of invertible $n \times n$ matrices.) Elements of a group can be used to transform points on a manifold—formally, these are called left group actions and right group actions. For example, under the right conditions, right-multiplying by an orthogonal matrix is a right-group action and left-multiplying by an orthogonal matrix is a left-group action. These actions naturally induce equivalence relations on the set that is getting acted on.[69] (See Section 9.2 in [Bou20] for a formal treatment of the contents of this paragraph.)

If a Lie group $\mathcal{G}$ acts on a smooth manifold $\overline{\mathcal{M}}$ in a way that satisfies certain properties, then $\overline{\mathcal{M}}/\mathcal{G}$ is a quotient manifold. We will give the formal statement next, although we will not formally define what all of the words in the theorem mean. That said, we will get some intuition for them in Example 5.5.

Theorem 5.3 (Quotient manifold through a group action) *If a Lie group $\mathcal{G}$ acts smoothly, freely, and properly on a smooth manifold $\overline{\mathcal{M}}$, then the quotient space $\overline{\mathcal{M}}/\mathcal{G}$ is a quotient manifold of dimension $\dim \overline{\mathcal{M}} - \dim \mathcal{G}$.*[70] *For manifolds that arise this way, the quotient space $\overline{\mathcal{M}}/\mathcal{G}$ is also known as the orbit space.*[71]

We will gain some intuition behind this theorem in Example 5.5, but first, a theorem that will be useful in the upcoming example:

Theorem 5.4 (Open subsets of embedded submanifolds are embedded submanifolds) *Let $\mathcal{M}$ be an embedded submanifold of $\mathbb{R}^d$. Any open subset of $\mathcal{M}$ is also an embedded submanifold of $\mathbb{R}^d$, with the same dimension and tangent spaces as $\mathcal{M}$.*[72]

Example 5.5 ($\mathcal{M}_p^{full}/\mathrm{O}(p)$ is a quotient manifold) *Let $\mathcal{M}_p^{full}$ denote the open subset of $\mathcal{M}_p$ that contains its rank p elements. By Theorem 5.4, we have that $\mathcal{M}_p^{full}$ is an embedded submanifold of $\mathbb{R}^{np}$ with the same dimension and tangent spaces as $\mathcal{M}_p$.*

[67] Proposition 3.4.2 in [AMS07].
[68] See Section 9.2 in [Bou20] for a formal treatment.
[69] Definition 9.12 in [Bou20].
[70] Theorem 8.17 in [Bou20].
[71] Definition 9.12 in [Bou20].
[72] Proposition 3.17 in [Bou20].

We claim that $\mathcal{M}_p^{full}/\mathrm{O}(p)$ *is a quotient manifold, where the equivalence classes are formed by* $\mathrm{O}(p)$ *acting on* $\mathcal{M}_p^{full}$ *through right multiplication. In other words,*

$$Y \sim V \iff Y = VQ \text{ for some } Q \in \mathrm{O}(p). \tag{17}$$

To prove that $\mathcal{M}_p^{full}/\mathrm{O}(p)$ *is indeed a quotient manifold, it is sufficient to check that* $\mathrm{O}(p)$ *acts smoothly, freely, and properly on* $\mathcal{M}_p^{full}$*. The group action is smooth since the map from* $\mathcal{M}_p^{full} \times \mathrm{O}(p)$ *to* $\mathcal{M}_p^{full}$ *defined by* $(Y, Q) \mapsto YQ$ *is smooth. The group action is free because only the identity element of the group fixes any given* Y*. In other words,*

$$YQ = Y \text{ for some } Y \in \mathcal{M}_p^{full} \Rightarrow Q = I_p. \tag{18}$$

The group action is proper because $\mathrm{O}(p)$ *is compact, and every smooth action by a compact Lie group is proper.*[73] *We conclude that* $\mathcal{M}_p^{full}/\mathrm{O}(p)$ *is indeed a quotient manifold. As a sanity check,* $\mathcal{M}_p^{full}$ *is stated to be a quotient manifold in [JBAS10] (see Section 4) and [WW20] (see page 18).*

We can also obtain the dimension of $\mathcal{M}_p^{full}/\mathrm{O}(p)$ *using Theorem 5.3:*

$$\begin{aligned}\dim(\mathcal{M}_p^{full}/\mathrm{O}(p)) &= \dim \mathcal{M}_p^{full} - \dim \mathrm{O}(p) \\ &= np - m - \frac{p(p-1)}{2},\end{aligned}$$

where we used the fact $\dim \mathcal{M}_p^{full} = \dim \mathcal{M}_p$ *due to Theorem 5.4, and the result from Example 2.15. The expression for* $\dim \mathrm{O}(p)$ *can be found in Section 7.4 in [Bou20]. To confirm our expression, see page 18 of [WW20].*

A natural question is whether $\mathcal{M}_p/\mathrm{O}(p)$ *(with equivalence classes defined as in (17)) is also a quotient manifold. It turns out that if* $\mathcal{M}_p$ *contains rank-deficient matrices, then* $\mathcal{M}_p/\mathrm{O}(p)$ *is not a quotient manifold, as noted in the footnote at the bottom of page 18 in [WW20].*[74]

For the next definition, it will be useful to be able to talk about the set containing all points in the total space $\overline{\mathcal{M}}$ which map into the same equivalence class on the quotient manifold $\mathcal{M}$. The notation we will use for this is $\pi^{-1}(\pi(x))$. (π is defined in Definition 5.1.) In other words,

$$\pi^{-1}(\pi(x)) = \{y \in \overline{\mathcal{M}} \mid x \sim y\}.$$

Definition 5.6 (Fiber, orbit) *Let* $\mathcal{M} = \overline{\mathcal{M}}/\sim$ *be a quotient manifold. For any* $x \in \overline{\mathcal{M}}$*, the equivalence class* $\mathcal{F} = \pi^{-1}(\pi(x))$*, also called a fiber, is closed in* $\overline{\mathcal{M}}$ *and it is an embedded submanifold of* $\overline{\mathcal{M}}$*. Its tangent spaces are given by*

$$\mathrm{T}_y\mathcal{F} = \ker \mathrm{D}\pi(y) \subseteq \mathrm{T}_y\overline{\mathcal{M}}.$$

[73] Proposition 9.16 in [Bou20].

[74] They prove that $\mathcal{M}_p/\mathrm{O}(p)$ is not a quotient manifold by showing that it violates Condition (i) of Proposition 3.4.2 in [AMS07]. (Proposition 3.4.2 is a complete characterization of which equivalence relations lead to quotient manifolds.) I think another way to prove it would be to show that the fibers (Definition 5.6) that would arise were $\mathcal{M}_p/\mathrm{O}(p)$ a quotient manifold could have different dimensions, which is a contradiction. This is discussed on page 203 of [Bou20]; see also Exercise 9.7 in that reference.

In particular, $\dim \mathcal{F} = \dim \overline{\mathcal{M}} - \dim \mathcal{M}$.[75]

If $\mathcal{M}$ *arises as the action of a Lie group* $\mathcal{G}$ *on* $\overline{\mathcal{M}}$ *as in Theorem 5.3, then the fiber* $\pi^{-1}(\pi(x))$ *is also known as the orbit of* $x \in \overline{\mathcal{M}}$ *under the action associated with* $\mathcal{G}$.[76]

This definition (which is really part theorem) says that the equivalence classes on the total space partition the total space into many embedded submanifolds. Furthermore, we begin to see through Definition 5.6 the importance of the operator $\mathrm{D}\pi(y)$, which was discussed earlier.

Example 5.7 (Fibers of $\mathcal{M}_p^{full}/\mathrm{O}(p)$) *The fiber containing* $Y \in \mathcal{M}_p^{full}$ *is precisely*

$$\mathcal{F} = \{Z \in \mathcal{M}_p^{full} \mid Y \sim Z\} = \{YQ \mid Q \in \mathrm{O}(p)\}.$$

In this case, it is not hard to calculate $\mathrm{T}_V\mathcal{F}$ *directly for some* $V \in \mathcal{F}$ *using Definition 2.6. All tangent vectors in* $\mathrm{T}_V\mathcal{F}$ *take the form* $c'(0)$ *for some smooth curve* $c : I \subseteq \mathbb{R} \to \mathcal{F}$ *with* $c(0) = V$. *Any such curve is necessarily of the form* $c(t) = VQ(t)$ *where* $Q : I \to \mathrm{O}(p)$ *is a smooth curve on the manifold* $\mathrm{O}(p)$ *with* $Q(0) = I_p$. *Then, all tangent in* $\mathrm{T}_V\mathcal{F}$ *are of the form* $VQ'(0)$. *The tangent space to* $\mathrm{O}(p)$ *at* $Q(0) = I_p$ *turns out to be the set of* $p \times p$ *skew-symmetric matrices,*[77] *so we have that*

$$\mathrm{T}_V\mathcal{F} = \{VS \mid S \in \mathbb{R}^{p\times p}, S + S^T = 0\}.$$

It is not hard to check that $\mathrm{T}_V\mathcal{F}$ *is indeed a subset of* $\mathrm{T}_V\mathcal{M}_p^{full}$, *as expected. (See Example 2.9, and recall that* $\mathrm{T}_V\mathcal{M}_p^{full} = \mathrm{T}_V\mathcal{M}_p$ *due to Theorem 5.4.)*

For a derivation of $\mathrm{T}_V\mathcal{F}$ *involving* $\ker \mathrm{D}\pi(V)$, *see Example 9.4 in [Bou20]. (The above also basically follows Example 9.4, although we are working with a slightly different, albeit functionally the same, manifold.)*

As a sanity check, this is actually the expression obtained in [WW20]—see the bottom of page 6 in that paper. (Note that they use the term "orbit" instead of "fiber.")

Finally, we calculate $\dim \mathcal{F}$ *using our expression for* $\mathrm{T}_V\mathcal{F}$ *and the fact that by definition, the dimension of a manifold is just the dimension of its tangent space (Definition 2.13). This immediately implies that* $\dim \mathcal{F} = \binom{p}{2} = \frac{p(p-1)}{2}$. *As a quick check, this is indeed equal to*

$$\dim \mathcal{M}_p^{full} - \dim(\mathcal{M}_p^{full}/\mathrm{O}(p)) = (np - m) - \left(np - m - \frac{p(p-1)}{2}\right).$$

We would now like to obtain a correspondence between tangent vectors of the total space $\overline{\mathcal{M}}$ and tangent vectors of the quotient manifold $\mathcal{M} = \overline{\mathcal{M}}/\sim$. The appropriate tool to do this is $\mathrm{D}\pi(x) : \mathrm{T}_x\overline{\mathcal{M}} \to \mathrm{T}_{[x]}\mathcal{M}$, which is surjective due to Definition 5.2. However, it is not one-to-one, so to fix this we need to restrict its domain. Definition 5.6 provides a very natural way to do this—we can partition the domain of $\mathrm{D}\pi(x)$, aka $\mathrm{T}_x\overline{\mathcal{M}}$, into those vectors in $\mathrm{T}_x\mathcal{F} = \ker \mathrm{D}\pi(x)$ and those vectors in its orthogonal complement, $(\mathrm{T}_x\mathcal{F})^\perp = (\ker \mathrm{D}\pi(x))^\perp$. Then, the restriction $\mathrm{D}\pi(x)$ to $(\mathrm{T}_x\mathcal{F})^\perp = (\ker \mathrm{D}\pi(x))^\perp$ forms a bijection between $(\mathrm{T}_x\mathcal{F})^\perp = (\ker \mathrm{D}\pi(x))^\perp$ and $\mathrm{T}_{[x]}\mathcal{M}$. We formalize this below with proper terminology:

[75] Proposition 9.3 in [Bou20].
[76] Definition 9.12 in [Bou20].
[77] Section 7.4 in [Bou20].

Definition 5.8 (Vertical space, horizontal space) *For a quotient manifold $\mathcal{M} = \overline{\mathcal{M}}/\sim$, the vertical space*[78] *at $x \in \overline{\mathcal{M}}$ is the subspace*

$$\mathrm{V}_x = \mathrm{T}_x\mathcal{F} = \ker \mathrm{D}\pi(x)$$

where $\mathcal{F} = \{y \in \overline{\mathcal{M}} \mid y \sim x\}$ is the fiber of x. If $\overline{\mathcal{M}}$ is Riemannian submanifold of $\mathbb{R}^d$ (so that we have an inner product), we call the orthogonal complement of V_x the horizontal space at x:

$$\mathrm{H}_x = (\mathrm{V}_x)^\perp = \{u \in \mathrm{T}_x\overline{\mathcal{M}} \mid \langle u, v\rangle = 0 \text{ for all } v \in \mathrm{V}_x\}.$$

Intuitively, the vertical directions (aka V_x) are all of the "uninteresting" directions of $\mathrm{T}_x\overline{\mathcal{M}}$.

Recall from the text above Definition 5.8 that the restriction of $\mathrm{D}\pi(x) : \mathrm{T}_x\overline{\mathcal{M}} \to \mathrm{T}_{[x]}\mathcal{M}$ to $\mathrm{H}_x = (\mathrm{T}_x\mathcal{F})^\perp = (\ker \mathrm{D}\pi(x))^\perp$, aka $\mathrm{D}\pi(x)|_{\mathrm{H}_x}$, is a bijection. This bijection allows us to use "concrete" horizontal vectors in $\mathrm{T}_x\overline{\mathcal{M}}$ to represent "abstract" vectors in $\mathrm{T}_{[x]}\mathcal{M}$.

Definition 5.9 (Horizontal lift) *Consider a point $x \in \overline{\mathcal{M}}$ and a tangent vector $\xi \in \mathrm{T}_{[}x]\mathcal{M}$. The horizontal lift*[79] *of ξ at x is the unique horizontal vector $u \in \mathrm{H}_x$ such that $\mathrm{D}\pi(x)[u] = \xi$. We write*

$$u = \left(\mathrm{D}\pi(x)|_{\mathrm{H}_x}\right)^{-1}[\xi] = \mathrm{lift}_x(\xi).$$

Horizontal lifts are critical for connecting objects on the quotient manifold to objects on the total space (e.g., retractions, vector fields, connections, etc.) We only include two examples in these notes:

Theorem 5.10 (Riemannian gradient result for quotient manifolds) *The Riemannian gradient of f on a Riemannian quotient manifold is related to the Riemannian gradient of the lifted function $\overline{f} = f \circ \pi$ on the total space via*[80]

$$\mathrm{lift}_x(\mathrm{grad} f([x])) = \mathrm{grad}\overline{f}(x)$$

for all $x \in \overline{\mathcal{M}}$.

We haven't actually talked about objective functions at all in this section, so in particular we haven't talked about what it means to lift a function. It is super intuitive though—precomposing a function f defined on the quotient manifold with the canonical projection π instantly gives you an "equivalent" function on the total space. And in fact, the only objective functions on the total space that are "interesting" in the first place (with respect to the quotient geometry at least) are those that are the lift of some function on the quotient manifold. If this isn't the case, then your objective function on the total space has different values among elements in the same equivalence class, at which point the quotient geometry doesn't help much.

Note that Theorem 5.10 implies that the Riemannian gradient on the total space, aka $\mathrm{grad}\overline{f}(x)$, is always a horizontal vector (because the output of a lift is always a horizontal vector). This makes

[78] Definition 9.23 in [Bou20].
[79] Definition 9.24 [Bou20].
[80] Proposition 9.38 in [Bou20].

sense intuitively—the Riemannian gradient should be orthogonal to the "flat" directions, aka the vertical space V_x. (After all, the Euclidean gradient is always orthogonal to the level curves—we should expect the same of the Riemannian gradient.)

We now give a similar result for the Hessian without much commentary:

Theorem 5.11 (Riemannian Hessian result for quotient manifolds) *The Riemannian Hessian of f on a Riemannian quotient manifold is related to the Riemannian Hessian of the lifted function $\overline{f} = f \circ \pi$ on the total space as*

$$\mathrm{lift}_x(\mathrm{Hess} f([x])[\xi]) = \mathrm{Proj}_x^{\mathrm{H}}(\mathrm{Hess}\overline{f}(x)[u])$$

for all $x \in \overline{\mathcal{M}}$ and $\xi \in \mathrm{T}_{[x]}\mathcal{M}$, with $u = \mathrm{lift}_x(\xi)$. Here $\mathrm{Proj}_x^{\mathrm{H}}$ is the orthogonal projector onto H_x.

The following theorems are useful for analyzing critical points in the total space:

Theorem 5.12 (Critical points result for quotient manifolds) *Let $f : \mathcal{M} \to \mathbb{R}$ be smooth on a Riemannian quotient manifold $\mathcal{M} = \overline{\mathcal{M}}/\sim$ with canonical projection π, and denote the lifted function $\overline{f} = f \circ \pi$. From Theorem 5.10, it is clear that $x \in \overline{\mathcal{M}}$ is a first-order critical point for $\overline{f}$ if and only if $[x]$ is a first-order critical point for f. Furthermore, it can be shown that x is a second-order critical point for $\overline{f}$ if and only if $[x]$ is a second-order critical point for f.*[81]

We can say more about the Riemannian Hessian at a first-order critical point:

Theorem 5.13 (Riemannian Hessian at a critical point result for quotient manifolds) *With the same setup as in Theorem 5.12, let $x \in \overline{\mathcal{M}}$ be a first-order critical point. Then the eigenvalues of $\mathrm{Hess}\overline{f}(x)$ are exactly the eigenvalues of $\mathrm{Hess} f([x])$ together with a set of $\dim \mathrm{V}_x = \dim \overline{\mathcal{M}} - \dim \mathcal{M}$ eigenvalues equal to zero. In particular, the vertical space V_x is included in the kernel of $\mathrm{Hess}\overline{f}(x)$. For every eigenvector of $\mathrm{Hess} f([x])$, there is a corresponding eigenvector of $\mathrm{Hess}\overline{f}(x)$ which lies in H_x and has the same eigenvalue.*[82] *(Indeed, the latter is just the horizontal lift of the former.)*

The above implies that if $\dim \mathcal{M} < \dim \overline{\mathcal{M}}$, the cost function $\overline{f}$ on the total space $\overline{\mathcal{M}}$ cannot admit second-order critical points where the Hessian is positive definite. Fortunately this isn't an issue because as explained in Section 9.9 in [Bou20], running Riemannian gradient descent on the total space is equivalent to running it on the Hessian. Thus, if the Hessian is positive definite on the quotient manifold, we will get the same benefits as if it were positive definite on the total space.

It is also possible to extend Theorem 4.19 to this scenario:

Theorem 5.14 (Second-order sufficient condition for quotient manifolds) *With the same setup as in Theorem 5.12, let $x \in \overline{\mathcal{M}}$ be such that $\mathrm{grad}\overline{f}(x) = 0$ (which is true if and only if $\mathrm{grad} f([x]) = 0$ due to Theorem 5.10) and $\mathrm{Hess} f([x]) \succ 0$. Then $x \in \overline{\mathcal{M}}$ is a local minimum.*[83]

[81] Exercise 9.45 in [Bou20].

[82] Exercise 9.45 and Lemma 9.40 in [Bou20]. Although part of this theorem I derived myself when solving Exercise 9.45.

[83] This result is not given in [Bou20], but I was able to prove it myself. Note that even if x is a local minimum, it will never be a strict local minimum due to the "flat" directions corresponding to the vertical space V_x.

Note that due to Theorem 5.13, we do not need to check that $\mathrm{Hess} f([x]) \succ 0$ *directly. It is instead sufficient to show that the only eigenvectors of* $\mathrm{Hess} \overline{f}(x)$ *with 0 as their eigenvalue are those* $\dim \mathrm{V}_x = \dim \overline{\mathcal{M}} - \dim \mathcal{M}$ *of them which lie in* V_x*.*

For example, Theorem 5.14 is used heavily in [WW20] to show that certain second-order critical points are additionally local minima. See Definition 3 and Remark 1 in that paper.

5.1 Quotient manifolds: material not covered

The material in this section is heavily based off of Chapter 9 of [Bou20]. That said, Chapter 9 is very long and we skipped a ton of material. Nearly all of the material skipped has to do with connecting objects on the total space to the corresponding objects on the quotient manifold (e.g., smooth maps, vector fields, retractions, connections, etc.) Section 9.13 is worth reading in particular since they talk about the specific scenario where the total space $\overline{\mathcal{M}}$ is an embedded submanifold of $\mathbb{R}^d$.

6 Riemannian gradient descent

To be done. For now, refer to Chapter 4 in [Bou20].

7 Riemannian second-order methods

To be done. For now, refer to Chapter 6 in [Bou20].

A Examples of manifolds

To be done. For now, refer to Chapter 7 in [Bou20], which is full of examples.

B Regarding product manifolds

Definition B.1 (Product manifold) *Let* $\mathcal{M}, \mathcal{M}'$ *be embedded submanifolds of* $\mathbb{R}^d, \mathbb{R}^{d'}$ *respectively. Then* $\mathcal{M} \times \mathcal{M}'$ *is an embedded submanifold of* $\mathbb{R}^d \times \mathbb{R}^{d'}$ *of dimension* $\dim \mathcal{M} + \dim \mathcal{M}'$ *such that*

$$\mathrm{T}_{(x,x')}(\mathcal{M} \times \mathcal{M}') = \mathrm{T}_x \mathcal{M} \times \mathrm{T}_{x'} \mathcal{M}'.$$

Basically, product manifolds behave exactly as you would expect them to: "X" for a product manifold is typically just the product or concatenation of "X" over each individual manifold which makes up the product manifold.

More details/examples to be added later.

References

[AMS07] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, USA, 2007.

[Bou16] Nicolas Boumal. A riemannian low-rank method for optimization over semidefinite matrices with block-diagonal constraints, 2016.

[Bou20] Nicolas Boumal. An introduction to optimization on smooth manifolds. Available online, Nov 2020.

[BVB18] Nicolas Boumal, Vladislav Voroninski, and Afonso Bandeira. Deterministic guarantees for burer-monteiro factorizations of smooth semidefinite programs. *Communications on Pure and Applied Mathematics*, 73, 04 2018.

[JBAS10] M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

[LY16] David G. Luenberger and Yinyu Ye. *Linear and nonlinear programming, fourth edition*, volume 228. New York, NY: Springer, 2016.

[WW20] Irène Waldspurger and Alden Waters. Rank optimality for the Burer-Monteiro factorization. *SIAM J. Optim.*, 30(3):2577–2602, 2020.