MATH-512 Optimization on Manifolds
Prof. Nicolas Boumal

Notes by Markus Renoldner — spring 2025, EPFL

Disclaimer: These notes were written by Markus Renoldner. They may contain typos or mistakes; any such errors are the author's responsibility.

Contents

1 Introduction

The following text is a summary of the lecture *Optimization on Manifolds* by Prof. Nicolas Boumal at EPFL, Spring 2025. The lecture is based on the book *Optimization on Manifolds* by Nicolas Boumal [Bou23]. Being a summary, its aim is to be both concise and reasonable exhaustive. On the other hand, it fails to be a good textbook for self-studying. The book [Bou23] is much better suited for that purpose, as it contains motivating paragraphs that convey the bigger picture.
After following the course, this summary can be used as a quick reference for the most important definitions, theorems and examples as well as a text that links the concepts from the lecture to the exact position and chapter in the book.
In some cases we refer and compare to the lecture notes of the course *Differential Geometry II* by Prof. Tsakanikas, which were given in winter semester 2024 at EPFL. The lecture notes are based on the book *Introduction to Smooth Manifolds* by John M. Lee [Lee03], which we also refer to from time to time.
Remark: Each subsection corresponds to one lecture slide PDF. PDF Numbers starting with 1 refer to geometry topics, numbers with 2 refer to optimization topics.
The order of topics follows the lecture and not the book. Most results are linked to the corresponding theorems in the book, however, to make it further reading easier.

2 Notation

For the sake of being concise, we sometimes do not repeat definitions of variables we use. This includes the following objects:

- $\mathcal{M}$ ... smooth manifold, either Riemannian or not, embedded or not.
- x a point on $\mathcal{M}$
- $T_x\mathcal{M}$ the tangent space at x
- v a tangent vector
- $\mathcal{E}$ a linear vector space or a Euclidean space

3 Embedded Geometry: first order

3.1 PDF 101, smooth (sub)manifolds

Definition 1: Submanifold in linear space (book Def3.10)

Let $\mathcal{E}$ be a d-dim vec-space. $\mathcal{M} \subset \mathcal{E}$ is an n-dim smooth, embedded, submanifold if either

1. $n = d$ and $\mathcal{M}$ is open in $\mathcal{E}$ (*open submfd.*); or
2. $n = d - k$ for some $k \geq 1$ and, for each $\forall x \in \mathcal{M}$ $\exists$ nbhd $U \ni x$ in $\mathcal{E}$ and $h : U \to \mathbb{R}^k$,

smooth, such that both

- $h|_U = 0 \iff y \in \mathcal{M}$
- $\operatorname{rank} \mathrm{D}h(x) = k$.

h is called *local defining function* of $\mathcal{M}$.

Remark: the 2. part of this is Prop5.11 in DGII lecture notes [TR24], or Prop5.16 in the book from Lee [Lee03].

Example 2. *Let* $\mathcal{E} = \mathbb{R}^3$ *and* $\mathcal{M} = \{(x, y, z) \in \mathbb{R}^3 \mid z = 0\} \cong \mathbb{R}^2$. *Then* $\mathcal{M} = h^{-1}(\{0\})$ *for*

$$\begin{aligned} h : U = \mathcal{E} &\to \mathbb{R} \\ (x, y, z) &\mapsto z \end{aligned}$$

Here $k = 1$. *Hence,* $\mathcal{M}$ *is a submanifold of* $\mathcal{E}$ *of dimension* 2.

Theorem 3: Charaterization of submanifolds (book Thm3.12)

$\mathcal{M} \subset \mathcal{E}$ is an n-dim embedded submfd of $\mathcal{E}$ iff. $\forall x \in \mathcal{M}$ there exists

- a nbhd $U \subset \mathcal{E}$ of x
- $V \in \mathbb{R}^d$, open, and
- diffeomorphism $F : U \to V$

such that

$$F(\mathcal{M} \cap U) = E \cap V,$$

with $E = \left\{y \in \mathbb{R}^d \ : \ y_{n+1} = \cdots = y_d = 0\right\}$ being a linear subspace of $\mathbb{R}^d$.

Remark: this is the *slice chart Lemma*, see Thm5.6 in DGII lecture notes [TR24], or Thm5.8 in the book from Lee [Lee03].

Proof. See book for details. The idea is:
" $\Longrightarrow$ ": Given h on U, build a diffeomorphism, that linearizes $\mathcal{M}$ around $x \in U$. This is possible, as we require Dh to have full rank, so we can use the inverse function theorem.
" $\Longleftarrow$ ": Play with F to build a candidate for h and check its properties. □

3.2 PDF 102, tangent spaces

Definition 4: Tangent space (book, Def3.14)

Let $\mathcal{M}$ be a subset of $\mathcal{E}$. For all $x \in \mathcal{M}$, define:

$$\mathrm{T}_x\mathcal{M} = \left\{c'(0) \mid c : I \to \mathcal{M} \text{ is smooth and } c(0) = x\right\}$$

where $I \subset \mathcal{R}$, open interval around 0.
So: $v \in \mathrm{T}_x\mathcal{M}$ iff. exists smooth curve on $\mathcal{M}$ with $c'(x) = v$.

Remark: this is Proposition 3.16-3.18 in the DGII lecture notes [TR24].

Theorem 5: Charaterization of tangent spaces (book Thm3.15)

Let $\mathcal{M}$ be an embedded submfd of $\mathcal{E}$.

- If $\mathcal{M}$ is an open submfd, then $\mathrm{T}_x\mathcal{M} = \mathcal{E}$
- Otherwise, $\mathrm{T}_x\mathcal{M} = \ker \mathrm{D}h(x)$ with h ... any loc. defining function at x.

Proof. See book/slides for details. Idea:

1. $T_x\mathcal{M} \subset \ker Dh(x)$:
   $v \in T_x\mathcal{M} \implies \exists$ curve c, smooth, s.t. $c(0) = x$ and $c'(0) = v$.
   $c(t) \in \mathcal{M} \implies h(c(t)) = 0 \implies Dh(c(t))c'(t) = 0$. Check at $t = 0$ and find that $v \in \ker Dh(x)$.
2. Show that $T_x\mathcal{M}$ contains an $n = d - k$ dimensional subspace of $\mathcal{E}$. Use the diffeomorphism F from Theorem 3 to construct a curve with velocity in $\mathcal{R}^n$.

□

Example 6. *The sphere*

$$\mathbb{S}^{d-1} := \left\{ x \in \mathbb{R}^d : x^\intercal x = 1 \right\} = h^{-1}(\{0\})$$

of $h(x) = x^\intercal x - 1$, *smooth on* $\mathbb{R}^d$. *Since* $\mathrm{D}h(x)[v] = 2x^\intercal v$, *it is clear that* $\operatorname{rank} \mathrm{D}h(x) = 1$ *for all* $x \in \mathbb{S}^{d-1}$.
As a result, $\mathbb{S}^{d-1}$ *is an embedded submanifold of* $\mathbb{R}^d$ *of dimension* $n = d - 1$. *Furthermore, its tangent spaces are given by* $\mathrm{T}_x\mathbb{S}^{d-1} = \ker \mathrm{D}h(x) = \left\{ v \in \mathbb{R}^d : x^\intercal v = 0 \right\}$.
In other words: tangent vectors of the sphere are really tangent/orthogonal to elements of the sphere.

3.3 PDF 103, Smooth maps and differentials

Definition 7: Smooth map (book, def3.30)

Let $\mathcal{M}, \mathcal{M}'$ be embed. submfds.
$F : \mathcal{M} \to \mathcal{M}'$ is smooth at x if there exists a smooth extension $\bar{F} : U \to \mathcal{E}'$ on a neighborhood $U \subset \mathcal{E}$ of x, s.t. $\bar{F}|_{\mathcal{M} \cap U} = F$.
The map F is smooth if it is smooth at all $x \in \mathcal{M}$.

Its easy to show the chain rule: $F : \mathcal{M} \to \mathcal{M}'$, and $G : \mathcal{M}' \to \mathcal{M}''$ smooth $\implies G \circ F$ smooth.

Theorem 8: Unique extension (book, prop3.31)

F is smooth iff. $\exists \bar{F}$, smooth on nbhd $V \subset \mathcal{E}$ of $\mathcal{M}$ s.t.

$$F = \bar{F}|_{\mathcal{M}}$$

In other words: the smooth extension of F on all $x \in \mathcal{M}$ is unique (but not necessarily smooth globally on $\mathcal{E}$!).

To define something analogous to the total derivative (in the book differetial) of a smooth map between vector spaces, we can either:

- Use $T_x\mathcal{M}$ (here: space of curve velocities), and define the differential as the derivative along a curve, or
- Use the smooth extension $\bar{F}$ and take the total derivative of that.

Definition 9: Differential of smooth map (book, def3.34)

The differential of $F : \mathcal{M} \to \mathcal{M}'$ at x is the map $DF(x) : T_x\mathcal{M} \to T_{F(x)}\mathcal{M}'$ defined by:

$$v \mapsto \frac{\mathrm{d}}{\mathrm{d}t} F(c(t))\Big|_{t=0}$$

where c is any smooth curve on $\mathcal{M}$ with $c(0) = x$ and $c'(0) = v$.

This definition corresponds to Prop. 3.16-3.18 in the DGII lecture notes [TR24]. See also Def. 4

Does this def. make sense?

Theorem 10: Differential is linear and indep. of curve c (book, prop3.35)

This definition coincides with $D\bar{F}|_{T_x\mathcal{M}}$, i.e. the total derivative (in the usual sense) of the smooth extension of F, restricted to the domain of DF.
Hence: $DF(x)$ is indep of choice of c, and v-linear, as $D\bar{F}(x)$ is v-linear.

Proof. As c maps into $\mathcal{M}$, we have $F(c) = \bar{F}(c)$. So

$$\begin{aligned} DF(x)(v) &= \frac{\mathrm{d}}{\mathrm{d}t} F(c(t))\Big|_{t=0} \\ &= \frac{\mathrm{d}}{\mathrm{d}t} \bar{F}(c(t))\Big|_{t=0} \\ &= D\bar{F}(c(0)) \cdot c'(0) \qquad \text{(chain rule)} \\ &= D\bar{F}(x)(v). \end{aligned}$$

□

3.4 PDF 104, Tangent bundle, vector fields, retractions

Definition 11: Tangent bundle (book, def3.42)

The tangent bundle of $\mathcal{M}$ is the set

$$T\mathcal{M} = \bigsqcup_x T_x\mathcal{M} = \{(x, v) : x \in \mathcal{M}, v \in T_x\mathcal{M}\}.$$

Fact: $T\mathcal{M}$ is an ebedded submfd in $\mathcal{E} \times \mathcal{E}$ with $\dim(T\mathcal{M}) = 2 \dim \mathcal{M}$.

Definition 12: Vector field (book def3.44)

A vector field on a mfd is a map $V : \mathcal{M} \to T\mathcal{M}$ s.t. $V(x) \in T_x\mathcal{M}$ for all x. We denote $\mathfrak{X}(\mathcal{M})$ the set of *smooth* vector fields.

Fact: If $\mathcal{M}$ is embedded in $\mathcal{E}$, then a vector field V is smooth iff there exists a smooth vector field $\overline{V}$ on a nbhd of $\mathcal{M}$ s.t. $V = \overline{V}|_{\mathcal{M}}$.

Definition 13: Retraction (book def 3.47)

A retraction is a smooth map $R : T\mathcal{M} \to \mathcal{M}$, such that each curve $c : t \mapsto R(x, tv)$ satisfies $c(0) = x, c'(0) = v$.

Example 14. *On the sphere $\mathbb{S}^{d-1}$, the map*

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

is a retraction, as one can check easily.

In the exercise *Metric projection retraction for Stiefel* from week 3 and in section 5.12 in the book, a special type of retraction is discussed (compare to def 13).

3.5 PDF 105, Riemannian metrics and gradients

Definition 15: Metric, Riemannian metric (book def3.51, 3.52, 3.53)

A metric is an inner product $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ (bilin., symm., pos-def) for each $x \in \mathcal{M}$.
We call it a Riemannian metric, if it is smooth (what does that mean using def of smooth maps??).
This is equivalent to requiring that for all smooth vector fields V, W on the mfd, the function $x \mapsto \langle V(x), W(x) \rangle_x$ is smooth.
A Riemannian manifold is a manifold with a Riemannian metric.
The metric $\langle \cdot, \cdot \rangle_{\mathcal{E}}$ of a Euclidean space $\mathcal{E}$ is called Euclidean metric.

The equivalence of the smoothness definition above to the one from Def. 7 is shown in chapter 3 in the book (in particular, Prop3.70). The proof is based on local frames, see Def. 23 below.

Remark: a metric in the geometric sense is not a metric in the topological sense, i.e. in the sense of a *distance function* on a *metric space*. Of course it induces a distance function on the tangent space (which are isomorphic to $\mathbb{R}^d$), in the usual way. But more importantly, the norm induced by the metric will give a way to compute distances on the manifold itself! (infimum over all curves between points, take norm of curve velocity and integrate etc.)

Theorem 16: Euclidean metric induces Riemannian mfd (book prop3.54)

he metric on $\mathcal{M}$ (embed. in $\mathcal{E}$) obtained by restricting the Euclidean metric $\langle \cdot, \cdot \rangle_{\mathcal{E}}$ to $T_x\mathcal{M}$ is a Riemannian metric.
With this (and only with this) metric we call $\mathcal{M}$ a Riemannian submanifold of $\mathcal{E}$.

Example 17 (Sphere as Riemannian mfd)**.** *The sphere $\mathbb{S}^{d-1}$ is a Riemannian submanifold of $\mathbb{R}^d$ with the metric induced by the Euclidean metric $\langle u, v \rangle_x := u \cdot v$ for $u, v \in T_x\mathbb{S}^{d-1}$.*

Example 18 (A non-smooth inner product)**.**

$$\langle u, v \rangle_x := u^\intercal G(x) v,$$

where $G(x) \in \mathbb{R}^{d \times d}$ is pos-definite in x. But if G is not smooth in x, the metric is not smooth.

Definition 19: Riemannian gradient (book def3.58)

Let $f : \mathcal{M} \to \mathbb{R}$, smooth. The Riemannian gradient of f is the vector field grad f on $\mathcal{M}$ uniquely defined by

$$\forall (x, v) \in T\mathcal{M}, \quad Df(x)[v] = \langle v, \operatorname{grad} f(x) \rangle_x .$$

Df is the differential, see Def 9 and Thm 10.
We skip the proof of the fact that grad f is well-defined (i.e. uniquely defined by the relation above).

Actually, in many situations there are three ways to compute gradients:

1. by definition, Def 19
2. using retractions, Theorem 20
3. using orthogonal projection of smooth extensions, Theorem 21

Theorem 20: Riemannian gradient using retractions (book prop3.59)

Let R be a retraction on a Riem. mfd $\mathcal{M}$. Then

$$\operatorname{grad} f(x) = \operatorname{grad}(f \circ R_x)(0).$$

The gradient on the right is the usual gradient on $T\mathcal{M}$ (or rather $T_x\mathcal{M}$). The precomposition $f \circ R_x$ is also called pullback of f by R_x.

Remarks:

- The above result is independent of the choice of R! This is true, as all retractions have the same behavior at zero: $d/dt R(x, tv)|_{t=0} = v$, indep of R.
- Of course, theorem 20 does not tell us how to find a retraction, it just tells us how to compute the gradient once we have a retraction.

Proof. The proof follows by def of R_x and grad, and chain rule, see book. □

Reminder on orthogonal projections (using an inner product/metric):

- Range: $\operatorname{im}(\operatorname{Proj}_x) = T_x\mathcal{M}$.
- Projector: $\operatorname{Proj}_x \circ \operatorname{Proj}_x = \operatorname{Proj}_x$.
- Orthogonal: $\langle u - \operatorname{Proj}_x(u), v \rangle = 0$ for all $v \in T_x\mathcal{M}$ and $u \in \mathcal{E}$.

Theorem 21: Riemannian gradient using projections (book, prop3.61)

Let $\operatorname{Proj}_x : \mathcal{E} \to T_x\mathcal{M}$ be the orthogonal projection. Then

$$\operatorname{grad} f(x) = \operatorname{Proj}_x[\operatorname{grad} \bar{f}(x)],$$

where $\bar{f}$ is the extension from 8.

Proof. The idea is, that for $v \in T_x\mathcal{M}$ we have

$$\begin{aligned} \langle v, \operatorname{grad} f(x) \rangle &= \langle v, \operatorname{grad} \bar{f}(x) \rangle \\ &= \langle v, \operatorname{grad} \bar{f}(x)_{\|} \rangle + \underbrace{\langle v, \operatorname{grad} \bar{f}(x)_{\perp} \rangle}_{=0}, \end{aligned}$$

where $\|, \perp$ are understood wrt. $\langle \cdot, \cdot \rangle$ and $T_x\mathcal{M}$. The parallel part is the image of the ortho. projection. □

Example 22 (Sphere). *$\mathcal{M} = \mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$, and let $\langle \cdot, \cdot \rangle$ be the standard inner product. In the setting of $\mathbb{R}^d$, the projection is just*

$$\mathrm{Proj}_x : z \mapsto z - x \langle z, x \rangle = (\mathbb{I} - xx^\intercal)z.$$

For smooth f, we have

$$\mathrm{grad}\, f(x) = \mathrm{Proj}_x[\mathrm{grad}\, \bar{f}(x)] = \mathrm{grad}\, \bar{f}(x) - \left(x^\intercal \mathrm{grad}\, \bar{f}(x)\right) x.$$

3.6 PDF 106, Local frames

Definition 23: Local frame (book def3.68)

Let $\dim(\mathcal{M}) = n$. A local frame around $x \in \mathcal{M}$ is a set of smooth vector fields $W_1, \ldots, W_n$ defined on a nbhd of x in $\mathcal{M}$, such that $W_1(y), \ldots, W_n(y)$ form a basis for the tangent space $T_y\mathcal{M}$ for all y.

Theorem 24: Local frame on embedded submanifolds (book prop3.69)

Let $\mathcal{M}$ be an embedded submanifold of a linear space $\mathcal{E}$. There exists a local frame around any $x \in \mathcal{M}$.
If $\mathcal{M}$ is Riemannian, there exists an orthonormal (local) frame.

Remark: the hairy ball theorem prevents global frames for general manifolds. Mfds with global frames are called parallelizable.

4 First-order optimization algorithms

Optimization algorithms move from point to point on a manifold by following smooth curves. In order to analyze these algorithms, we need to understand how the cost function varies along those curves. For this we will use Taylore expansions, see Section 4.1, [Bou23].

4.1 PDF 201, Taylor expansions of first order

Let $c : I \to \mathcal{M}$ be a smooth curve with $c(0) = x, c'(0) = v$ where I is an open interval around 0. Then evaluating f along this curve yields a real function $g : I \to \mathbb{R},\ t \mapsto (f \circ c)(t)$.

Since g is smooth on $\mathbb{R}$, it admits a Taylor expansion, given as

$$g(t) = g(0) + tg'(0) + \mathcal{O}(t^2). \tag{1}$$

Firstly, we know that $g(0) = f(x)$, and secondly, we get (chain rule) that

$$g'(t) = Df(c(t))[c'(t)] = \left\langle \mathrm{grad}\, f(c(t)), c'(t) \right\rangle_{c'(t)}, \tag{2}$$

so equation (1) becomes

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

Example 25. *Consider the special case where* $c : t \mapsto R(x, tv)$*, with* R *being a retraction. Then*

$$f(R(x,tv)) = f(x) + t\,\langle \operatorname{grad} f(x), v\rangle_x + \mathcal{O}\left(t^2\right).$$

One cann set $s := tv$*, and write this as*

$$f(R(x,s)) = f(x) + \langle \operatorname{grad} f(x), vs\rangle_x + \mathcal{O}\left(\|s\|_x^2\right).$$

4.2 PDF 202, Optimality conditions, first order

Definition 26: Critical/stationary point (book def4.4)

x is critical/stationary for f if $\forall$ smooth curves c on $\mathcal{M}$ with $c(0) = x$ we have

$$(f \circ c)'(0) \geq 0$$

Theorem 27: First order necessary condition (book prop4.5)

If x is a local (i.e. on a nbhd of x) minimizer of f, then x is critical for f.

Theorem 28: First order necessary condition for Riemannian mfds (book prop4.6)

Let $f : \mathcal{M} \to \mathbb{R}$ be smooth on a Riemannian manifold $\mathcal{M}$. Then x is a critical point of f if and only if $\operatorname{grad} f(x) = 0$.

4.3 PDF 203, Gradient descent

Goal: minimize $f : \mathcal{M} \to \mathbb{R}$, smooth. For this, we choose a retraction R, a Riemannian metric $\langle \cdot, \cdot \rangle_x$ and a starting point x_0.

Algorithm:

Definition 29: Riemannian gradient descent (RGD) (book Algo4.1)

Input: $x_0 \in \mathcal{M}$
For: $k = 1, ..., n$

- Pick step-size $\alpha > 0$ (e.g. backtracking line search)
- $x_{k+1} = R_{x_k}(s_k)$, with $T_x\mathcal{M} \ni s_k := -\alpha \operatorname{grad} f(x_k)$

In the lecture, the following two theorems are presented in reverse order. We stick to the book ordering.

Theorem 30: RGD converges for nice functions (book prop4.7)

Assuming

1. $\exists f_{\text{low}}$ s.t. $f(x) > f_{\text{low}}\ \forall x$
2. the algorithm achieves *sufficient decrease*: $\exists c > 0$, s.t. $f(x_k) - f(x_{k+1}) \geq c\,\|\operatorname{gradf}(x_k)\|^2$,

then:

$$\lim_{k\to\infty} \|\operatorname{grad} f(x_k)\| = 0,$$

i.e. all accumulation (=limit) points are critical points.

Furthermore $\forall K \geq 1$ there exists $k \in \{0, ..., K-1\}$ s.t.

$$\|\operatorname{grad} f(x_k)\| \leq \sqrt{\frac{f(x_0) - f_{\text{low}}}{cK}}.$$

Remark: this does **not** imply convergence of the sequence x_k to a critical point! It only states that, if limit points exists, they are critical.
In particular: if there exists at least one limit point, then RGD converges to a critical points.

Proof. The proof is based on a simple telescoping sum argument, see book. □

In the book and in the lecture notes, there is another assumption, under which the algorithm achieves *sufficient decrease* (as in thm 30). Using thm 20, as well as $\operatorname{grad} f(x) = \operatorname{grad}(f \circ R_x)(0)$, the Taylor expansion yields:

$$f(x_{k+1}) = f(R_{x_k}(s_k)) = f(x_k) + \langle \operatorname{grad} f(x_k), s_k \rangle + \mathcal{O}(\|s_k\|^2).$$

If the quadratic remainder term stays under control during all iterations, we may deduce a guarantee on the progress $f(x_k) - f(x_{k+1})$.

Theorem 31: RGD convergence 2 (book prop4.8)

Assuming $\exists L > 0$ s.t. forall $(x, s) \in S \subset T\mathcal{M}$

$$f(R_x(s)) \leq f(x) + \langle \operatorname{grad} f(x), s \rangle_x + L \|s\|^2,$$

then, if $(x_0, s_0), (x_1, s_1), ... \subset S$, and $\alpha_k \in \left(0, \frac{2}{L}\right)$, then the algo produces *sufficient decrease* as in thm 30 with

$$c = \min\left(\alpha_{\min} - \frac{L}{2}\alpha_{\min}^2, \alpha_{\max} - \frac{L}{2}\alpha_{\max}^2\right) > 0.$$

In particular: $\alpha_k = \frac{1}{L}\ \forall k \implies c = \frac{1}{2L}$.

Remark: The assumption from above is called a Lipschitz-type condition. It is a generalization of the Lipschitz continuity of the gradient. In the Euclidean setting $\mathcal{M} = \mathbb{R}^d$, this condition is implied by Lipschitz continuity of the gradient: $\|\operatorname{grad} f(y) - \operatorname{grad} f(x)\| \leq L \|y - x\|$, which is a common assumption.
However, the version of the theorem can be generalized to manifolds very easily, while the version with $\|\operatorname{grad} f(y) - \operatorname{grad} f(x)\|$ needs more theory of manifolds, as the gradients are not in the same space.

4.4 PDF 204, Step size

Definition 32: Backtracking line-search algorithm (book algo 4.2)

Input: x_0, $\bar{\alpha} > 0$, τ (e.g. 0.5), r (e.g. 10^{-4})
Set $\alpha = \bar{\alpha}$
While: $f(x) - f(R_x(-\alpha \operatorname{grad} f(x))) < r\alpha \|\operatorname{grad} f(x)\|^2$ (Armijo rule)
Set $\alpha = \tau\alpha$
Output: α

Why does this make sense?

Theorem 33: Backtracking line-search guarantees (book Lemma 4.12)

Assuming that the Lipschitz-condition from thm 31 holds for $f \circ R$ with constant L on the set

$$\{(x, -\alpha \operatorname{grad} f(x)) : \alpha \in [0, \overline{\alpha}]\}.$$

Then, the Algorithm from def 32 with parameters $\tau, r \in (0,1)$ outputs a step-size α such that

$$f(x) - f(R_x(-\alpha \operatorname{grad} f(x))) \geq r \min\left(\overline{\alpha}, \frac{2\tau(1-r)}{L} \ldots\right) \|\operatorname{grad} f(x)\|^2,$$

after a bounded nr of steps (see book). This is sufficient decrease, as in thm 30.

4.5 PDF 212, Momentum methods and nonlinear conjugate gradients

The slides mention some remarks about *Momentum methods and nonlinear conjugate gradients.*

5 Embedded Geometry: second order

In Section 4.1, we used Taylor expansions of functions along curves, to understand their *first order behavior*. We want to extend this to higher order derivatives, i.e. we need a notion of Hessian. For this we will need to differentiate gradient fields, see chapter 5 in [Bou23].
In this section we introduce the necessary tools for this. The second order Taylor expansion is then discussed in chapter 6.1

5.1 PDF 107, Differentiating vector fields

Consider again example 25: $f : \mathcal{M} \to \mathbb{R}$, and a curve c with $c(0) = x, c'(0) = v$, and define $g(t) := f(c(t))$. Then

$$g(t) = g(0) + tg'(0) + \frac{t^2}{2} g''(0) + \mathcal{O}\left(t^3\right) \tag{3}$$

$$= f(x) + t \langle \operatorname{grad} f(x), v \rangle_x + \frac{t^2}{2} \frac{\mathrm{d}}{\mathrm{d}t} \left\langle \operatorname{grad} f(c(t)), c'(t) \right\rangle_{c(t)} \Big|_{t=0} + \mathcal{O}\left(t^3\right) \tag{4}$$

In the Euclidean case $\mathcal{E}$ (with std. inner product), and with the choice of a straight curve $c(t) = x + tv$ we have

$$g(t) = f(x) + t \langle \operatorname{grad} f(x), v \rangle_x + \frac{t^2}{2} \langle \operatorname{Hess} f(x)[v], v \rangle_x + \mathcal{O}\left(t^3\right).$$

The Hessians describes the rate of change of the gradient. Consider the following example:

Example 34 (Sphere, see Section 5.1, [Bou23].)**.** *Consider $\mathbb{S}^{d-1}$ with the std. inner product. Define on $\mathbb{S}^{d-1}$*

$$f : x \mapsto \frac{1}{2} x^\intercal A x,$$

where $A \in \mathbb{R}^{d \times d}$, symm. The Riemannian grad. of f is the sm. vec-field:

$$x \mapsto \operatorname{grad} f(x) = Ax - (x^\intercal A x)x =: V(x),$$

(see book example 3.62). So $V : \mathbb{S}^{d-1} \to T\mathbb{S}^{d-1}$, smooth. Hence, using def 9, its differential, using the smooth extension $\overline{V}(x) = Ax - (x^\intercal Ax)x$ defined on $\mathbb{R}^d$ (and for any $u \in T_x\mathbb{S}^{d-1}$) is

$$\begin{aligned}
\mathrm{D}V(x)[u] &= \mathrm{D}\overline{V}(x)[u] \\
&= Au - (x^\intercal Ax)u - (u^\intercal Ax + x^\intercal Au)x \\
&= \underbrace{\mathrm{Proj}_x(Au)}_{\in T_x\mathbb{S}^{d-1}} - \underbrace{(x^\intercal Ax)u}_{\in T_x\mathbb{S}^{d-1}} - \underbrace{(u^\intercal Ax)x}_{\notin T_x\mathbb{S}^{d-1}\ !!} \\
&\notin T_x\mathbb{S}^{d-1}.
\end{aligned}$$

So the differential of the Riemannian gradient is not a tangent vector at x!

Conclusion: with this notion of derivative, $\mathrm{Hess}\, f(x)$ would not be a linear map from $T_x\mathbb{S}^{d-1}$ to itself, and terms such as $\langle \mathrm{Hess}\, f(x)[u], u\rangle_x$ would not make sense. We aim to define a better derivative for vector fields.

5.2 PDF 108, Connections

In linear space $\mathcal{E}$ the differential of a smooth vector field V is:

$$\mathrm{D}V(x)[u] = \lim_{t\to 0} \frac{V(x+tu) - V(x)}{t}.$$

Given smooth vector fields U, V, W, vectors $u, w \in T_x\mathcal{E} \cong \mathcal{E}$, $a, b \in \mathbb{R}$ and f smooth on $\mathcal{E}$, we have

1. $x \mapsto \mathrm{D}V(x)[U(x)]$ is a smooth vector field
2. $\mathrm{D}V(x)[au + bw] = a\mathrm{D}V(x)[u] + b\mathrm{D}V(x)[w]$;
3. $\mathrm{D}(aV + bW)(x)[u] = a\mathrm{D}V(x)[u] + b\mathrm{D}W(x)[u]$; and
4. $\mathrm{D}(fV)(x)[u] = \mathrm{D}f(x)[u] \cdot V(x) + f(x)\mathrm{D}V(x)[u]$.

We will now introduce a similar concept on manifolds:

Definition 35: Connection, covariant derivative (book def5.1)

A connection on mfd $\mathcal{M}$ is an operator

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

such that $\nabla_u V := \nabla(x, u, V) \in T_x\mathcal{M}$ if $u \in T_x\mathcal{M}$ and

1. $(\nabla_U V)(x) := \nabla_{U(x)} V$ is a smooth vec field $\nabla_U V$
2. Lin. in u: $\nabla_{au+bw} V = a\nabla_u V + b\nabla_w V$
3. Lin. in V: $\nabla_u(aV + bW) = a\nabla_u V + b\nabla_u W$
4. Leibniz rule: $\nabla_u(fV) = Df(x)[u] \cdot V(x) + f(x)\nabla_u V$

The vector field $\nabla_U V$ is also called the covariant derivative of V along U.

There is an alternative definition (as a pointwise derivative), see book section 5.6.

Example 36 (Linear space)**.** *By construction, $DV(x)[u]$ is a connection on $\mathcal{E}$.*

We consider the embedded case. It seems reasonable to believe, that the projection of the differential of a vector field onto $T_x\mathcal{M}$ (see e.g. in Example 34) is a connection.

Theorem 37: Connection on embedded mfd (book Thm5.2)

Let $\mathcal{M}$ be an embedded submanifold of a Euclidean space $\mathcal{E}$. The operator ∇ defined by

$$\nabla(x,u,\cdot) = \nabla_u : V \mapsto \text{Proj}_x \, \text{D}\overline{V}(x)[u],$$

with Proj_x being a projection onto $T_x\mathcal{M}$, is a connection on $\mathcal{M}$.

The example illustrates, that (at least on emebed.) manifolds many (infinite) connections exist. E.g. one can construct different ones by using a different projection. Which one is the right one?

5.3 PDF 109, Riemannian Connections

We will require 2 more properties of connections, which will make them unique. In particular, they will make sure that Riemannian Hessians (defined later) will be symmetric/self-adjoint.

In the lecture the material was presented in reverse order: we first required symmetry of the Hessian, and then derived the two extra conditions.

Definition 38: Riemannian connection (book def5.5)

For $U, V \in \mathfrak{X}(\mathcal{M})$ and f smooth on U, with U open in $\mathcal{M}$, define:

- action of field on function: $Uf : U \to \mathbb{R}$, smooth with $Uf : x \mapsto \text{D}f(x)[U(x)]$
- Lie bracket: $[U,V] : f \mapsto U(Vf) - V(Uf)$
- $\langle U, V\rangle : \mathcal{M} \to \mathbb{R}$ such that $\langle U,V\rangle(x) = \langle U(x), V(x)\rangle_x$.

Def3.4 and section 7.1.1 in DGII lecture notes [TR24] give more background on the meaning of Uf. Notice that def 19 gives

$$Uf = \langle \text{grad}\, f, U\rangle.$$

The following theorem defines the unique Riemannian connection based on the 2 extra requirements:

Theorem 39: Fundamental theorem of Riemannian geometry (book thm5.6)

On a Riemannian manifold $\mathcal{M}$, there exists a unique connection ∇ which satisfies two additional properties for all $U, V, W \in \mathfrak{X}(\mathcal{M})$:

1. *Symmetry:* $[U,V]f = (\nabla_U V - \nabla_V U)f$ for all smooth functions f on $\mathcal{M}$,
2. *Compatibility with the metric:* $U\langle V,W\rangle = \langle \nabla_U V, W\rangle + \langle V, \nabla_U W\rangle$.

This connection is called the *Levi-Civita* or *Riemannian connection.*

We summarize some useful results:

Theorem 40: Important Riemannian connections (book thm5.8-5.10)

The connection $\nabla_u V = \text{D}V(x)[u]$ is

1. symmetric on a linear space $\mathcal{E}$,
2. compatible with the metric on a Euclidean space.

The connection $\nabla_u V = \text{Proj}_x \, \text{D}\overline{V}(x)[u]$ is

1. symmetric on an embedded submanifold of a Euclidean space,
2. compatible with the metric on a Riemannian submanifold of a Euclidean space.

5.4 PDF 110, Riemannian Hessians

In the following, we denote by ∇ the unique Riemannian connection on $\mathcal{M}$.

Definition 41: Riemannian Hessian (book Def5.14)

Let $\mathcal{M}$ be a Riemannian manifold with its Riemannian connection ∇. The Riemannian Hessian of $f \in \mathcal{F}(\mathcal{M})$ at $x \in \mathcal{M}$ is the linear map

$$\operatorname{Hess} f(x) : T_x\mathcal{M} \to T_x\mathcal{M}, u \mapsto \nabla_u \operatorname{grad} f.$$

Equivalently, $\operatorname{Hess} f$ maps $\mathfrak{X}(\mathcal{M})$ to $\mathfrak{X}(\mathcal{M})$ as

$$\operatorname{Hess} f : U \mapsto \nabla_U \operatorname{grad} f.$$

The two special properties from Theorem 39 lead to symmetry of the Hessian. By the spectral theorem, this implies that the $\dim \mathcal{M}$ eigenvalues of $\operatorname{Hess} f(x)$ are real, and that corresponding eigenvectors may be chosen to form a basis of $T_x\mathcal{M}$ orthonormal wrt. $\langle \cdot, \cdot \rangle_x$.

Theorem 42: Self-adjointness of the Riemannian Hessian (book Prop5.15)

The Riemannian Hessian is self-adjoint wrt the Riem. metric: for all $x \in \mathcal{M}$ and $u, v \in T_x\mathcal{M}$,

$$\langle \operatorname{Hess} f(x)[u], v \rangle_x = \langle u, \operatorname{Hess} f(x)[v] \rangle_x.$$

On embed. mfds, one can compute Hessians of a function using a smooth extensions: we use the second connection mentioned in theorem 40.

Example 43 (book Ex. 5.17)**.** *Pick $\overline{f}(x) = \frac{1}{2} x^\intercal A x$ on $\mathbb{R}^d$, where $A \in \mathbb{R}^{d \times d}$ symmetric and its restriction $f = \overline{f}|_{\mathbb{S}^{d-1}}$ (sphere as Riemannian submanifold of $\mathbb{R}^d$).*
Consider the Euclidean and Riemannian grad (see exercises from week 1 and Example 22):

$$\begin{aligned} \operatorname{grad} \overline{f}(x) &= Ax \\ \operatorname{grad} f(x) &= \operatorname{Proj}_x(\operatorname{grad} \overline{f}(x)) = (\mathbb{I} - xx^\intercal)Ax = Ax - (x^\intercal Ax)x. \end{aligned}$$

To apply the second part of theorem 40, we choose an extension of the gradient, e.g. the obvious one:

$$\overline{G}(x) = Ax - (x^\intercal Ax)x, \quad \forall x \in \mathbb{R}^d.$$

The differential of $\overline{G}$ follows from the product rule (see the discussion after Definition 19 as well as book Section 4.7):

$$\mathrm{D}\overline{G}(x)[u] = Au - (u^\intercal Ax + x^\intercal Au)x - (x^\intercal Ax)u.$$

We use now first Definition 41 and then the second part of theorem 40: We orthogonally project to tangent space at x to get Hessian (defined on $T_x\mathbb{S}^{d-1}$):

$$\begin{aligned} \operatorname{Hess} f(x)[u] &= \operatorname{Proj}_x \mathrm{D}\overline{G}(x)[u] \\ &= \operatorname{Proj}_x [Au] - \operatorname{Proj}_x \underbrace{[(u^\intercal Ax + x^\intercal Au)x]}_{\in \ker \operatorname{Proj}_x} - \operatorname{Proj}_x [(x^\intercal Ax)u] \\ &= \operatorname{Proj}_x(Au) - (x^\intercal Ax)u \\ &= Au - (x^\intercal Au)x - (x^\intercal Ax)u. \end{aligned}$$

Example 44 (from the lecture)**.** *More abstractly speaking, if we take a general f on $\mathbb{S}^{d-1}$, then (by def of the projection), we have the following procedure:*

1. *Project the gradient:*
$$\operatorname{grad} f(x) = \operatorname{Proj}_x(\operatorname{grad} \bar{f}(x)) = \operatorname{grad} \bar{f}(x) - (x^\intercal \operatorname{grad} \bar{f}(x))_x.$$
2. *Extend it:*
$$\bar{G}(x) := \operatorname{grad} \bar{f}(x) - (x^\intercal \operatorname{grad} \bar{f}(x))_x \quad \forall x \in \mathbb{R}^d$$
3. *Compute the usual Hessian of $\bar{G}$:*
$$D\bar{G}(x)(u) = \operatorname{Hess} \bar{f}(x)(u) - (x^\intercal \operatorname{grad} \bar{f}(x))u - \left(u^\intercal \operatorname{grad} \bar{f}(x) + x^\intercal \operatorname{Hess} \bar{f}(x)(u)\right) x$$
4. *Use Def. 41 and the second part of Thm. 40 (i.e. project this Hessiand to get the Riemannian Hessian):*
$$\begin{aligned}\operatorname{Hess} f(x)(u) &= \operatorname{Proj}\left[D\bar{G}(x)(u)\right] \\ &= \operatorname{Proj}_x\left[\operatorname{Hess} \bar{f}(x)(u) - (x^\intercal \operatorname{grad} \bar{f}(x))u - \left(u^\intercal \operatorname{grad} \bar{f}(x) + x^\intercal \operatorname{Hess} \bar{f}(x)(u)\right) x\right] \\ &= \operatorname{Proj}_x\left[\operatorname{Hess} \bar{f}(x)(u)\right] - (x^\intercal \operatorname{grad} \bar{f}(x))u\end{aligned}$$

This should emphasize that the Riemannian Hessian is not just the projection of the Hessian of the smooth extension $\bar{f}$ of f! There is a correction term, called the Weingarten map*, that depends on the $T_x\mathbb{S}^{d-1}$-orthogonal component of the usual gradient of the extension of f!*

5.5 PDF 111, Differentiating vector fields along curves

As mentioned in the introduction to Section 5, we want to Taylor expand smooth functions f along curves up to the second order term. Again, let $g := f \circ c$. We know from equation (2) that $g'(t) = \langle \operatorname{grad} f(c(t)), c'(t)\rangle_{c'(t)}$. To compute g'', we have to differentiate the gradient not on the manifold, but only on the image of c.

Definition 45: Vector field along a curve (book Def5.28)

Let $c : I \to \mathcal{M}$ be a smooth curve. A map $Z : I \to T\mathcal{M}$ is a *vector field on/along c* if $Z(t) \in T_{c(t)}\mathcal{M}$ for all $t \in I$. The set of *smooth* vector fields on c is denoted by $\mathfrak{X}(c)$.

Example 46.

- *the curve velocity c' of a curve c is a vector field along the curve c itself.*
- *given $U \in \mathfrak{X}(\mathcal{M})$, the curve $U \circ c$ is a vector field along c*
- *in particular: for f smooth on $\mathcal{M}$, the curve $\operatorname{grad} f \circ c$ is a vector field along c.*

Theorem 47: Induced covariant derivative (book Thm5.29)

Let $c : I \to \mathcal{M}$ be smooth on a mfd equipped with connection ∇. There exists a unique operator
$$\frac{D}{dt} : \mathfrak{X}(c) \to \mathfrak{X}(c)$$
which satisfies the following properties for all $Y, Z \in \mathfrak{X}(c)$, $U \in \mathfrak{X}(\mathcal{M})$, $g \in \mathcal{F}(I)$, and $a, b \in \mathbb{R}$:

1. R-linearity: $\frac{D}{dt}(aY + bZ) = a\frac{D}{dt}Y + b\frac{D}{dt}Z$;
2. Leibniz rule: $\frac{D}{dt}(gZ) = g'Z + g\frac{D}{dt}Z$;
3. Chain rule: $\frac{D}{dt}(U \circ c)(t) = \nabla_{c'(t)}U$ for all $t \in I$.

We call $\frac{D}{dt}$ the *induced covariant derivative* (induced by ∇). If moreover $\mathcal{M}$ is a Riemannian

manifold and ∇ is compatible with its metric $\langle \cdot, \cdot \rangle$ (e.g., if ∇ is the Riemannian connection), then the induced covariant derivative also satisfies:

4. Product rule: $\frac{d}{dt}\langle Y, Z\rangle = \langle \frac{D}{dt}Y, Z\rangle + \langle Y, \frac{D}{dt}Z\rangle$,

where $\langle Y, Z\rangle \in \mathcal{F}(I)$ is defined by $\langle Y, Z\rangle(t) = \langle Y(t), Z(t)\rangle_{c(t)}$.

Remark: Considering the chain rule above, it looks like, the induced cov. derivative can always be computed through an application of ∇. This would make the introduction of $\frac{D}{dt}$ obsolete. However, this assumption is not true:
Not all vector fields $Z \in \mathfrak{X}(c)$ are of the form $U \circ c$ for some $U \in \mathfrak{X}(\mathcal{M})$. Indeed, consider a smooth curve c such that $c(t_1) = c(t_2) = x$ (it crosses itself), then one does not necessarily have that $Z(t_1) = Z(t_2)$. And therefore its not clear how to define $U(x)$ at that point. It could be either $Z(t_1)$ or $Z(t_2)$.
Remark:

- For a Euclidean space and $Z \in \mathfrak{X}(c)$,
$$\frac{DZ(t)}{dt} = \frac{dZ(t)}{dt} = \lim_{\delta \to 0} \frac{Z(t+\delta) - Z(t)}{\delta}.$$
- For a Riemannian submanifold of a Euclidean space,
$$\frac{DZ(t)}{dt} = \mathrm{Proj}_{c(t)}\left(\frac{dZ(t)}{dt}\right). \tag{5}$$

As the operator $\frac{D}{dt}$ is unique, the above two facts can be proven by checking if the properties of the induced covariant derivative are satisfied.

In the following we will discuss how to approximate the Riemannian Hessian by finite difference formulas (see also Section 10.6 in the book [Bou23]). Let $c : I \to \mathcal{M}$ be any smooth curve such that $c(0) = x$ and $c'(0) = u$. Then, using Def41, Thm47, and (assuming additionally, that we are on a Riemannian submfd of a Eukl. space) equation (5), we have

$$\begin{aligned}
\operatorname{Hess} f(x)[u] &\overset{41}{=} \nabla_u \operatorname{grad} f \\
&\overset{47}{=} \frac{D}{dt}(\operatorname{grad} f \circ c)(t)\big|_{t=0} \\
&\overset{(5)}{=} \mathrm{Proj}_x\left(\frac{\mathrm{d}}{\mathrm{d}t}(\operatorname{grad} f \circ c)(0)\right) \\
&= \mathrm{Proj}_x\left(\lim_{t\to 0} \frac{\operatorname{grad} f(c(t)) - \operatorname{grad} f(c(0))}{t}\right) \\
&= \lim_{t\to 0} \frac{\mathrm{Proj}_x[\operatorname{grad} f(c(t))] - \operatorname{grad} f(c(0))}{t}.
\end{aligned}$$

In the last steps we have just used the definition of the usual derivative and continuity of the projection. Also, $\operatorname{grad} f(c(0)) = \operatorname{grad} f(x) \in T_x\mathcal{M}$, so applying the projection has no effect. We can therefore approximate the Hessian by taking a finite (but small) t in the above formula:

$$\operatorname{Hess} f(x)[u] \approx \frac{\mathrm{Proj}_x[\operatorname{grad} f(c(t))] - \operatorname{grad} f(c(0))}{t}. \tag{6}$$

We have now all the tools to define the notion of curve acceleration and geodesics.

Definition 48: Acceleration along a curve (book Def5.36)

Let $c : I \to \mathcal{M}$ be a smooth curve. Its velocity is the vector field $c' \in \mathfrak{X}(c)$ along c, see Def 45. The (*intrinsic*) *acceleration* of c is the smooth vector field $c'' \in \mathfrak{X}(c)$ defined by:

$$c'' = \frac{Dc'}{dt}.$$

We denote by $\ddot{c} = \frac{\mathrm{d}^2}{\mathrm{d}t^2} c$ the usual (extrinsic) acceleration of the curve c in the ambient space $\mathcal{E}$. The first derivatives coincide, so $\dot{c} = c'$.

Definition 49: Geodesic (book Def5.38)

On a Riemannian mfd, a geodesic is a smooth curve $c : I \to \mathcal{M}$ such that $c''(t) = 0$ for all $t \in I$.

Example 50 (Geodesics on the 2D sphere $\mathbb{S}^2$)**.** *Geodesics on $\mathbb{S}^2$ are great circles, i.e., intersections of $\mathbb{S}^2$ with planes through the origin. These curves have no intrinsic acceleration ($c''(t) = 0$) and locally minimize distance, analogous to straight lines in Euclidean space.*

6 Second-order optimization algorithms

As mentioned in the introduction to Section 6, we want to understand the second order behavior of functions on manifolds. We will now use the framework developed there (Hessians, second order retractions etc.) to derive second order optimization algorithms.

6.1 PDF 205, Taylor expansions and retractions

Let $\mathcal{M}$ be a Riemannian mfd with ∇ and $\frac{D}{dt}$ and $f \in \mathcal{F}(\mathcal{M})$. Let c be a smooth curve with $c(0) = x$ and $c'(0) = v$. We have the following Taylor expansion. Similar as in equation (3), we pull back the function to $\mathcal{E}$: $g(t) := f(c(t))$. We have

$$
\begin{aligned}
g(t) &= g(0) + tg'(0) + \frac{t^2}{2} g''(0) + \mathcal{O}\left(t^3\right) \\
g'(t) &= DF(c(t))[c'(t)] = \langle \operatorname{grad} f(c(t)), c'(t) \rangle_{c(t)} \\
g''(t) &= \frac{D}{dt} \langle \operatorname{grad} f(c(t)), c'(t) \rangle_{c(t)} \\
&\overset{*3}{=} \left\langle \frac{D}{\mathrm{d}t} (\operatorname{grad} f \circ c)(t), c'(t) \right\rangle_{c(t)} + \left\langle \operatorname{grad} f(c(t)), \frac{D}{\mathrm{d}t} c'(t) \right\rangle_{c(t)} \\
&\overset{*4}{=} \left\langle \nabla_{c'(t)} \operatorname{grad} f, c'(t) \right\rangle_{c(t)} + \left\langle \operatorname{grad} f(c(t)), c''(t) \right\rangle_{c(t)} \\
&\overset{**}{=} \left\langle \operatorname{Hess} f(c(t))[c'(t)], c'(t) \right\rangle_{c(t)} + \left\langle \operatorname{grad} f(c(t)), c''(t) \right\rangle_{c(t)}.
\end{aligned}
$$

Where in step *3 and *4, we have used the 3rd and 4th bullet from Thm 47, and in step ** we have used Def 41, respectively.
Evaluating $g''(t)$ at $t = 0$ yields:

$$
(f \circ c)''(0) = g''(0) = \langle \operatorname{Hess} f(x)[v], v \rangle_x + \langle \operatorname{grad} f(x), c''(0) \rangle_x.
$$

Combining all the above steps, we obtain the Taylor expansion of f along the curve c:

$$
f(c(t)) = f(x) + t\langle \operatorname{grad} f(x), v \rangle_x + \frac{t^2}{2} \langle \operatorname{Hess} f(x)[v], v \rangle_x + \frac{t^2}{2} \langle \operatorname{grad} f(x), c''(0) \rangle_x + \mathcal{O}(t^3). \tag{7}
$$

We are now interested in Retraction curves $c(t) = R_x(tv)$, similar as in Example 25.

Definition 51: Second-order retraction (book Def5.42)

A second-order retraction R on a Riemannian manifold $\mathcal{M}$ is a retraction such that, for all $x \in \mathcal{M}$ and all $v \in T_x\mathcal{M}$, the curve $c(t) = R_x(tv)$ has zero acceleration at $t = 0$, that is, $c''(0) = 0$.

Applying equation (7) to the retraction curve $c(t) = R_x(tv)$, we obtain:

Theorem 52: Taylor expansion with retraction curve (book Prop5.44)

If R is a second-order retraction, then for all $x \in \mathcal{M}$ and $s \in T_x\mathcal{M}$, we have:

$$f(R_x(s)) = f(x) + \langle \operatorname{grad} f(x), s \rangle_x + \frac{1}{2} \langle \operatorname{Hess} f(x)[s], s \rangle_x + \mathcal{O}(\|s\|_x^3).$$

Example 53 (Second order retraction on the sphere)**.** *Consider the following retraction on the sphere* $\mathbb{S}^{d-1}$*:*

$$R_x(v) = \frac{x + v}{\|x + v\|}.$$

That retraction is second order, see book Example 5.43, [Bou23].

For the pullback $f \circ R_x$, we already had $\operatorname{grad} f(x) = \operatorname{grad}(f \circ R_x)(0)$, see Thm 20. That holds for all retractions and for all $x \in \mathcal{M}$. The expansion above yields a corollary that completes the picture:

Theorem 54: Second-order retraction and Hessian equivalence (book Prop5.45)

If R is a second-order retraction, or if $\operatorname{grad} f(x) = 0$, then

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

The latter is the classical Hessian.

6.2 PDF 206, Optimality conditions

Before we move on to discuss second-order optimization algorithms, we discuss second-order necessary optimality conditions.

Definition 55: Second-order critical point (book Def6.1)

A point $x \in \mathcal{M}$ is *second-order critical/stationary* for a smooth function $f : \mathcal{M} \to \mathbb{R}$ if

$$(f \circ c)'(0) = 0 \quad \text{and} \quad (f \circ c)''(0) \geq 0$$

for all smooth curves c on $\mathcal{M}$ such that $c(0) = x$.

Theorem 56: Second-order necessary condition (book Prop6.2)

Any local minimizer of a smooth function $f : \mathcal{M} \to \mathbb{R}$ is a second-order critical point of f.

We denote by $\succ 0, \succeq 0$ positive definiteness/semidefiniteness.

Theorem 57: Second-order critical point characterization (book Prop6.3)

Let $f : \mathcal{M} \to \mathbb{R}$ be smooth on a Riemannian manifold $\mathcal{M}$. Then, x is a second-order critical point of f if and only if

$$\operatorname{grad} f(x) = 0 \quad \text{and} \quad \operatorname{Hess} f(x) \succeq 0.$$

Theorem 57 does not characterize minimizers, though.

Example 58 (Counterexample)**.** *Consider the smooth function* $f : x \mapsto x^3$ *on* $\mathbb{R}$*. At* $x = 0$*, both* $\operatorname{grad} f(x) = 0$*, and* $\operatorname{Hess} f(x) \succeq 0$*, meaning the point i second order critical. But the point is not a minimizer.*

Theorem 59: Second order sufficient condition (book Prop6.5-6.6)

On a Riemannian mfd, if $\operatorname{grad} f(x) = 0$ and $\operatorname{Hess} f(x) \succ 0$, then x is a local minimum of f.

The result from Theorem 59 is actually discussed in a bit more detail in the book [Bou23].

6.3 PDF 207, Newton's method

We want to exploit *second order information* to speed up RGD (see Definition 29). Algorithms can be built using

$$x_{k+1} = R_{x_k}(s_k).$$

But instead of the gradient, we want to choose a better s_k.
Recall the second-order Taylor expansion along a Retraction curve from equation (7), as well as Theorems 56 and 59.
Close to critical points, we approximate the function along the retraction curve by $m_x : T_x\mathcal{M} \to \mathbb{R}$ with

$$f(R_x(s)) \approx m_x(s) := f(x) + \langle \operatorname{grad} f(x), s \rangle_x + \frac{1}{2} \langle s, \operatorname{Hess} f(x)[s] \rangle_x .$$

We want to find s such that $m_x(s)$ is small. Notice that m_x is a quadratic on $T_x\mathcal{M}$.
A minimizer of m_x, if one exists, must be a critical point. To determine its gradient, we use self-adjointness of $\operatorname{Hess} f(x)$:

$$\langle \operatorname{grad} m_x(s), u \rangle_x = Dm_x(s)[u] = \langle \operatorname{grad} f(x), u \rangle_x + \langle \operatorname{Hess} f(x)[s], u \rangle_x.$$

The above holds for all $u \in T_x\mathcal{M}$, so:

$$\operatorname{grad} m_x(s) = \operatorname{grad} f(x) + \operatorname{Hess} f(x)[s].$$

Thus, a tangent vector $s \in T_x\mathcal{M}$ is a critical point of m_x if and only if

$$\operatorname{Hess} f(x)[s] = -\operatorname{grad} f(x),$$

which is called the Newton equation/Newton step. If $\operatorname{Hess} f(x) \succ 0$, the Newton step finds the minimizer of m_x.

Definition 60: Riemannian Newton's method (book Algo6.1)

Input: $x_0 \in \mathcal{M}$
For: $k = 0, 1, 2, \ldots$

- Solve $\operatorname{Hess} f(x_k)[s_k] = -\operatorname{grad} f(x_k)$ for $s_k \in T_{x_k}\mathcal{M}$
- Update $x_{k+1} = R_{x_k}(s_k)$

Theorem 61: Convergence of Newton's method (book Thm6.7)

Let $f : \mathcal{M} \to \mathbb{R}$ be smooth on a Riemannian mfd. If x^* is such that $\operatorname{grad} f(x^*) = 0$ and $\operatorname{Hess} f(x^*)$ invertible, then there exists a nbhd U of x^* on $\mathcal{M}$ such that, for all $x_0 \in U$, Newton's method converges at least quadratically to x^*.

Quadratically means something like $\operatorname{dist}(x_{k+1}, x^*) \leq C \operatorname{dist}(x_k, x^*)^2$ for some constant $C > 0$.
The theorem is only local. Just like Newton's method on Euklidean space, its global behavior can be very bad/unpredictable.

6.4 PDF 208, Computing the Newton step

In the Newton step, we have to find $s \in T_x\mathcal{M}$ such that

$$\operatorname{Hess} f(x)[s] = -\operatorname{grad} f(x).$$

This is a linear system, but we dont always have access to $\operatorname{Hess} f(x)$ as a matrix.
Assume H is positive definite.
Instead of trying to construct the matrix (by first constructing a basis of the tangent space etc, which is costly), we will try to find a minimizer of

$$m_x(v) = \frac{1}{2} \langle v, \operatorname{Hess} f(x)[v] \rangle_x + \langle \operatorname{grad} f(x), v \rangle_x + f(x).$$

We can use gradient descent or also Conjugate Gradients for that. Denote $b := -\operatorname{grad} f(x)$ and $H = \operatorname{Hess} f(x)$. For this we compute

$$g(v) := \operatorname{grad} m_x = Hv - b.$$

We would then compute $v_0 = 0$ and $v_n = v_{n-1} - \alpha_n \operatorname{grad} m_x$. One can use line-search for α_n, but here in this simple case (m_x is quadratic), we can compute it explicitly.
More generally, taking a direction r, we have

$$\begin{aligned} g(v + \alpha r) &= \frac{1}{2} \langle v + \alpha r, H(v + \alpha r) \rangle_x - \langle b, v + \alpha r \rangle_x \\ &= \dots \\ &= g(v) + \alpha \langle r, Hv - b \rangle_x + \frac{\alpha^2}{2} \langle r, Hr \rangle_x \end{aligned}$$

Using the solution formulas for quadratic equations:

$$\alpha^* = -\frac{\langle r, Hv - b \rangle_x}{\langle r, Hr \rangle_x}.$$

In the special case of gradient descent, $r = -\operatorname{grad} m_x = -(Hv - b)$, we have

$$\alpha^* = \frac{\|r\|^2}{\langle r, Hr \rangle_x}.$$

Gradient descent becomes:
Initialize: $v_0 = 0$, and $r_0 = -\operatorname{grad} m_x(0) = -b$.

For $n = 1, 2, 3 \dots$

- $\alpha_n = \frac{\|r_{n-1}\|^2}{\langle r_{n-1}, Hr_{n-1} \rangle_x}$

- $v_n = v_{n-1} + \alpha_n r_{n-1}$
- $r_n = -\operatorname{grad} m_x(v_n) = b - Hv_n$
- **If** $\|r_n\|_x < \text{tol} \cdot \|b\|_x$, **output** v_n.

Conjugate gradients (Algo6.2, [Bou23])
Initialize: $v_0 = 0$, $r_0 = b$, $p_0 = r_0$
For $n = 1, 2, 3, \ldots$

- $\alpha_n = \frac{\|r_{n-1}\|_x^2}{\langle p_{n-1}, Hp_{n-1}\rangle_x}$
- $v_n = v_{n-1} + \alpha_n p_{n-1}$
- $r_n = r_{n-1} - \alpha_n H p_{n-1}$
- **If** $\|r_n\|_x \leq \text{tol} \cdot \|b\|_x$, **output** v_n
- $\beta_n = \frac{\|r_n\|_x^2}{\|r_{n-1}\|_x^2}$
- $p_n = r_n + \beta_n p_{n-1}$

6.5 PDF 209, Conjugate gradients

CG minimizes in at most $\dim \mathcal{M}$ iterations if using exact arithmetic, see end of Section 6.4 of [Bou23]. This fails numerically though, and it is irrelevant in high dimension.
However, CG is much faster than GD, for about same cost per iteration, also in practise.
For the CG error bound, we need the H-norm

$$\|u\|_H := \sqrt{\langle u, u\rangle_H} = \sqrt{\langle u, Hu\rangle_x}.$$

Theorem 62: CG error estimate (book end of Sec6.3)

If $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues of H, then $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$ is the condition number of H, and it can be shown that

$$\|v_n - s\|_H \leq \|s\|_H \cdot 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \leq \|s\|_H \cdot 2e^{-n/\sqrt{\kappa}},$$

so that the error decreases exponentially fast as CG iterates.

Proof. Let $s \in T_x\mathcal{M}$ minimize $Hs = b$. For $v \in T_x\mathcal{M}$,

$$\begin{aligned}\|v - s\|_H^2 &= \langle v, Hv\rangle_x - 2\langle v, b\rangle_x + \langle s, Hs\rangle_x, \\ &= 2g(v) + \langle s, Hs\rangle\end{aligned}$$

where H is self-adjoint and $Hs = b$. The last term is indep of v, so minimizing over v the term $\|v - s\|_H$ is equivalent to minimizing $g(v)$. We denote by

$$K_n := \operatorname{span}\{b, Hb, \ldots, H^{n-1}b\}$$

the nth Krylov subspace. As $v_n \in K_n$ (to see this, look at the CG algo),

$$v_n = \arg\min_{v \in K_n} \|v - s\|_H.$$

Expanding v_n in a basis of K_n gives,

$$\begin{aligned}v_n - s &= (a_0 I + a_1 H + \cdots + a_{n-1}H^{n-1})Hs - s \\ &= (a_0 H + a_1 H^2 + \cdots + a_{n-1}H^n - I)s \\ &= q_n(H)s,\end{aligned}$$

where $q_n \in Q_n := \{\text{polynomials of deg. up to n: } q(0) = -1\}$. So

$$\|v_n - s\|_H = \min_{q_n \in H_n} \|q_n(H)s\|_H \,. \tag{8}$$

Using orthog. eigenvectors $u_1, \ldots, u_d$ of H with eigenvalues $\lambda_1, \ldots, \lambda_d$, and writing $s = \sum_i a_i u_i$ we get

$$\|g(H)s\|_H^2 = \langle g(H)s, Hg(H)s\rangle_x = \sum_i a_i^2 \lambda_i q(\lambda_i)^2.$$

We conclude that, for any polynomial q,

$$\frac{\|q(H)s\|_H^2}{\|s\|_H^2} = \frac{\sum_{i=1}^d q(\lambda_i)^2 \lambda_i \langle u_i, s\rangle_x^2}{\sum_{i=1}^d \lambda_i \langle u_i, s\rangle_x^2} \leq \max_{1\leq i\leq d} q(\lambda_i)^2,$$

Using equation (8), we get

$$\|v_n - s\|_H \leq \|s\|_H \cdot \min_{q\in Q_n} \max_{1\leq i\leq d} |q(\lambda_i)|.$$

If H has k distinct eigenvalues, CG terminates in k iterations. For condition number, we have $\kappa = \lambda_{\max}/\lambda_{\min}$,

$$\|v_n - s\|_H \leq \|s\|_H \cdot 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \leq \|s\|_H \cdot 2e^{-n/\sqrt{\kappa}}.$$

□

This is useful, as for each degree n, one can find a polynomial in Q_n with maximal absolute value less than

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n$$

over the interval $[1, \kappa]$. See Fig. 6.1 in the book [Bou23] for details.

6.6 PDF 210, Trust-region methods

Can we transition from the nice global GD-like to the fast local Newton-like behavior adaptively to get the best of both worlds? The problem with Newton is that we approximate the pullback $f \circ R_x$ by a quadratic function

$$m_k(s) = f(x) + \langle \operatorname{grad} f(x), s\rangle_x + \frac{1}{2}\langle s, \operatorname{Hess} f(x)[s]\rangle_x \,, \tag{9}$$

see section 6.3. This works well only if $s \in T_x\mathcal{M}$ is small (inside a *trust region*), and $\operatorname{Hess} f(x)$ is positive definite. So instead of minimizing m_x on the whole tangent space, we solve the following constrained problem, the *trust-region subproblem* (TRS):

$$\min_{s\in T_{x_k}\mathcal{M}} m_k(s) \quad \text{subject to} \quad \|s\|_{x_k} \leq \Delta_k, \tag{10}$$

where Δ_k is the radius of the trust region at iteration k. Consider the following algorithm:

Definition 63: Riemannian trust-region method, (book Algo 6.3)

Input: $\rho' \in \left(0, \frac{1}{4}\right)$; $\overline{\Delta} > 0$
Initialize: $x_0 \in \mathcal{M}$; $\Delta_0 \in (0, \overline{\Delta}]$
For: $k = 1, 2, \ldots$

- Find s_k as the solution to TRS in equ. (10)
- Tentative next iterate $x_k^+ = R_{x_k}(s_k)$
- Assess its quality $\rho_k = \frac{f(x_k) - f(x_k^+)}{m_k(0) - m_k(s_k)}$

- Accept or reject: $x_{k+1} = \begin{cases} x_k^+ & \text{if } \rho_k > \rho' \text{ (accept)} \\ x_k & \text{otherwise} \end{cases}$
- Update the trust region radius: $\Delta_{k+1} = \begin{cases} \frac{1}{4}\Delta_k & \text{if } \rho_k < \frac{1}{4} \\ \min(2\Delta_k, \overline{\Delta}) & \text{if } \rho_k > \frac{3}{4} \text{ and } \|s_k\|_{x_k} = \Delta_k \\ \Delta_k & \text{otherwise.} \end{cases}$

The following material deviates from the book, but follows the lecture slides.

We would like to invest minimial effort into the TRS, equ. (10): minimizing m_k is only a means to an end, the actual endgoal is to minimize f. We can perform a *Cauchy step*: let

$$s^C := -t \cdot \operatorname{grad} f(x) \in T_x\mathcal{M} \tag{11}$$

with optimal t. The optimal t can be found, as m_k is a quadratic function: it will either correspond to the minimum of the parabola, or to a boundary point of the trust region. If we choose t appropriately, we obtain the following decrease estimate

Theorem 64: Minimal decrease with Cauchy step (PDF210)

The Cauchy step $s^C \in T_x\mathcal{M}$ satisfies the following minimal progress garuantee

$$m_x(0) - m_x\left(s^C\right) \geq \frac{1}{2} \min\left(\Delta, \frac{\|\operatorname{grad} f(x)\|_x}{\|\operatorname{Hess} f(x)\|_x}\right) \|\operatorname{grad} f(x)\|_x .$$

So: Cauchy step is a cheap way to get a decent decrease. We can make a global statement now. For this assume:

1. $f(x) > f_{\text{low}}$ for all $x \in \mathcal{M}$
2. R is a second order retraction (not necessary, but simplifies argument)
3. $|f(R_x(v)) - f(x) - \langle \operatorname{grad} f(x), v\rangle_x| \leq \frac{L}{2}\|v\|_x^2$ for all $(x, v) \in T\mathcal{M}$
4. Subproblem solver ensures $m_x(0) - m_x\left(s^C\right) \geq$ Cauchy step decrease (Thm 64)

Under these assumptions, we get that the trust region method is not worse than Gradient Descent (see Def 29). We have the following worst case bound on the iteration number:

Theorem 65: Iterion bound (PDF210)

The algorithm from Definition 63 finds x_k with $\|\operatorname{grad} f(x)\|_{x_k} \leq \epsilon$ for some

$$k \leq \frac{48L(f(x_0) - f_{\text{low}})}{\rho'} \frac{1}{\epsilon^2} + \frac{1}{2} \log_2\left(\frac{16L\Delta_0}{\epsilon}\right),$$

given any $\epsilon \leq 16L\Delta_0$.

Proof idea. The main ideas are:

1. The trust region radius cannot become arbitrarily small.
   Indeed the only way Δ_k can become smaller, is if $\rho_k < \frac{1}{4}$. ρ_k measures how the model (m_k) decrease predicts the actual decrease of f. If s is small, m_k is a good approximation. And if Δ_k is small, so is s, which pushes ρ_k close to 1. This idea can be made rigorous giving a lower bound on Δ_k.

2. Successful (i.e. "accepted") steps yield good decrease when the gradient is large.

$$\begin{aligned} f(x_k) - f(x_{k+1}) &= f(x_k) - f(x_k^+) && \text{(accept } x_k^+\text{)} \\ &= \rho_k \left(m_k(0) - m_k(s_k)\right) && \text{(by def. of } \rho_k\text{)} \\ &\geq \rho' \cdot (\text{Cauchy step decrease}) && \text{(Assump. 4)} \end{aligned}$$

   The Cauchy step decrease bound depended on $\min\left(\Delta, \frac{\|\operatorname{grad} f(x)\|_x}{\|\operatorname{Hess} f(x)\|_x}\right)$. But we know that Δ_k is bounded from below. If Δ is indeed larger, then the decrease is proportional to $\|\operatorname{grad} f(x_k)\|_{x_k}^2$, similar as in RGD, see Theorem 30. Otherwise we still have good decrease.

3. Most steps are successful.
   Each time we reject a step x_k, we had $\rho_k > \rho'$, but $\rho' < \frac{1}{4}$, so the update of Δ_k will choose $\Delta_{k+1} = \frac{1}{4}\Delta_k$, until we make progess again. But we already "proved" in step 1, that Δ_k is bounded from below. So asympotically, there must be at least 2 updates of type $\Delta_{k+1} = 2\min(2\Delta_k, \overline{\Delta})$, for each decrease by $\frac{1}{4}$, to compensate. So roughly 2/3 of all steps are successful.

A rigorous version of these ideas gives the bound. □

Under some stronger assumptions (in particular, if one does better than the Cauchy step), one can show that the trust region method is much better than GD.

More practical (See section 6.4.6 in [Bou23] for details):

- Riemannian trust-regions (RTR) is available in Manopt.
- Default parameters: $\rho' = 0.1$ and $\Delta_{\max} = \operatorname{diam}(\mathcal{M})$ or $\overline{\Delta} = \dim(\mathcal{M})$.
- Default initialization: $\Delta_0 = \overline{\Delta}/8$ and x_0 random on $\mathcal{M}$.
- Default subproblem solver: truncated conjugate gradients (tCG).
- Don't check $\|s_k\|_{x_k} = \Delta_k$ in floating point arithmetic: have the TRS solver return s_k and a boolean `limitedbyTR`.
- Computing ρ_k in floating point arithmetic is tricky: regularize.

Relaxed assumotions and finer guarantiess:

1. $\operatorname{Hess} f(x)$ in the definition of m_k in equation (9) can be replaced by other functions, e.g. a finite difference approximation of the Hessian, see equation (6).
2. We argued that $\|\operatorname{grad} f(x_k)\|_{x_k}$ gets small. Section 6.4.5 of [Bou23] shows some mild assumptions to prove that $\|\operatorname{grad} f(x_k)\|_{x_k} \to 0$.
3. RTR can find *approximately second-order cricital points*, see Theorem 57.

6.7 PDF 211, Truncated conjugate gradients for TRS

The trust region method assumes a choice: how much work do we invest into the TRS? The Cauchy step is cheap, and we get RGD behavior. We discuss how to get behavior *up to* Newton-like speed at the cost of finding a better s_k.
Recall the definition of m_x in equation (9), and that we aim to solve

$$\min_{s \in T_x\mathcal{M}} m_x(s) \quad \text{subject to} \quad \|s\|_x \leq \Delta,$$

Strategy: run conjugate gradients, cautiously (maybe $\operatorname{Hess} f(x) \not\succ 0$, and satisfy the constraint) and opportunistically (stop early). More details in section 6.5, [Bou23]

To do this, recall what CG does (see end of section 6.4): applying CG on $m(v) = \frac{1}{2}\langle v, Hv\rangle_x - \langle b, v\rangle_x$ generates $v_0, v_1, v_2, \ldots$ such that

$$v_n = \operatorname{argmin}_{v \in \mathcal{K}_n} g(v), \quad \text{with } \mathcal{K}_n = \operatorname{span}(p_0, \ldots, p_{n-1}) = \operatorname{span}(b, Hb, \ldots, H^{n-1}b).$$

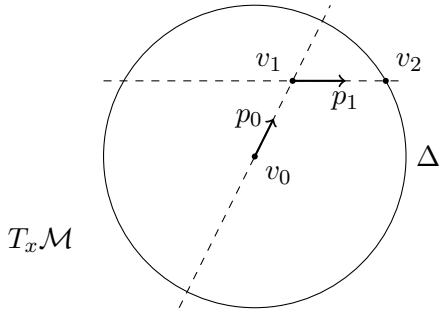
As

$$v_n = v_{n-1} + \alpha_n p_{n-1},$$

we have that v_n is optimal along $t \mapsto v_{n-1} + tp_{n-1}$. Thus, we can reinterpret CG like this:

- It generates a special set of vectors $p_0, p_1, p_2, \ldots$
- At each step, v_n minimizes g on the line $t \mapsto v_{n-1} + tp_{n-1}$.

This approach still works when constrained to a ball, like Δ, or if we lose convexity! If we hit the boundary (e.g., if m_x is concave along the line), abort. Also, if v_n is good enough, abort.
We introduce truncated conjugate gradients (compare to CG in section 6.4).



Definition 66: Truncated Conjugate Gradients (tCG)(book Algo6.4)

Initialize: $v_0 = 0$, $r_0 = b$, $p_0 = r_0$
For $n = 1, 2, 3, \ldots$

- $\alpha_n = \frac{\|r_{n-1}\|_x^2}{\langle p_{n-1}, Hp_{n-1}\rangle_x}$
- $v_n = v_{n-1} + \alpha_n p_{n-1}$
- **If** $\langle p_{n-1}, Hp_{n-1}\rangle_x \leq 0$ **or** $\|v_n\|_x \geq \Delta$, **then** return min of g restricted to $t \mapsto v_{n-1} + tp_{n-1}$
- **If** $\|r_n\|_x \leq tol \cdot \|b\|_x$, **output** v_n, with $tol = \min(0.1, \|b\|_x)$
- $\beta_n = \frac{\|r_n\|_x^2}{\|r_{n-1}\|_x^2}$
- $p_n = r_n + \beta_n p_{n-1}$

Remarks:

- The first iterate of tCG is the Cauchy step, equation (11), then it only gets better.
- Close to a strict minimizer, tCG makes essentially Newton steps.
- Hess f can be something else than Hess $f(x)$, e.g., finite differences, see equation (6).
- Often need fewer iterations than GD because it works harder at each x_k (fewer retractions, can reuse computations at x_k)
- It's useful to ensure x_{k-1} is numerically in $T_x\mathcal{M}$ periodically.

7 From embedded to general mfds

7.1 PDF 501,502, Smooth sets and functions

We generalize embedded mfds to general (non-embedded) mfds. See also chapter 8 in [Bou23] or [Lee03] and [TR24] for a more general treatment.
Here, we will build everything on top of three conpcets:

- smooth sets
- smooth functions
- tangent vectors

Definition 67: Chart (book Def8.1)

A d-dimensional chart on a set $\mathcal{M}$ is a pair (U, φ) consisting of a subset U of $\mathcal{M}$ (called the domain) and a map $\varphi : U \to \mathbb{R}^d$ such that:

1. $\varphi(U)$ is open in $\mathbb{R}^d$, and
2. φ is invertible between U and $\varphi(U)$.

The numbers $(\varphi(x)_1, \ldots, \varphi(x)_d)$ are the coordinates of the point $x \in U$ in the chart φ. The map $\varphi^{-1} : \varphi(U) \to U$ is a local parameterization of $\mathcal{M}$.

Definition 68: Compatible charts (book Def8.2)

Two charts (U, φ) and (V, ψ) of $\mathcal{M}$ are compatible if they have the same dimension d and either $U \cap V = \emptyset$, or $U \cap V \neq \emptyset$ and:

1. $\varphi(U \cap V)$ is open in $\mathbb{R}^d$;
2. $\psi(U \cap V)$ is open in $\mathbb{R}^d$; and
3. $\psi \circ \varphi^{-1} : \varphi(U \cap V) \to \psi(U \cap V)$ is a diffeomorphism.

Definition 69: Atlas (book Def8.3)

An atlas $\mathcal{A}$ on a set $\mathcal{M}$ is a compatible collection of charts on $\mathcal{M}$ whose domains cover $\mathcal{M}$. In particular, for every $x \in \mathcal{M}$, there is a chart $(U, \varphi) \in \mathcal{A}$ such that $x \in U$.

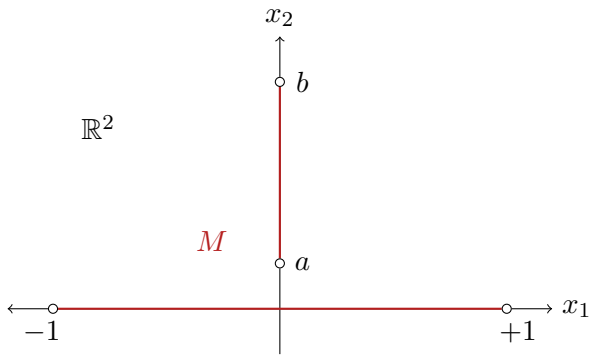
Given an atlas A, one can show that the collection A^+ of all charts of M which are compatible with A is itself an atlas of M, called a *maximal atlas*.

Definition 70: Atlas topology (book Def8.17)

Given a maximal atlas $\mathcal{A}^+$ on a set $\mathcal{M}$, the atlas topology on $\mathcal{M}$ states that a subset of $\mathcal{M}$ is open if and only if it is the union of a collection of chart domains.

Example 71 (The atlas topology can be uncomfortable)**.** *Consider the manifold*

$$\mathcal{M} := \{(0, y) : y \in (a, b)\} \cup \{(x, 0) : x \in (-1, 1)\}.$$



Let $\mathcal{U} = \{(t, 0) : -1 < t < 1\}$ be the lower *part of $\mathcal{M}$. It's possible to pick an atlas for $\mathcal{M}$ such that, in the atlas topology, the sequence*

$$\left(\frac{1}{2}, 0\right), \left(\frac{1}{3}, 0\right), \left(\frac{1}{4}, 0\right), \left(\frac{1}{5}, 0\right), \ldots$$

converges to $(0, \alpha)$ for all $\alpha \in \{0\} \cup \{(a, b)\}$: the limit is not unique. Indeed, we define:

$$\mathcal{U}_\alpha = (\mathcal{U} \setminus \{(0, 0)\}) \cup \{(0, \alpha)\},$$

and

$$\varphi_\alpha : \mathcal{U}_\alpha \to (-1, 1), \quad (x_1, x_2) \mapsto x_1,$$

and the Atlas: $\{(\mathcal{U}_\alpha, \varphi_\alpha) : \alpha = 0$ or $\alpha \in \{(a, b)\}\}$. The nonuniqueness implies the topology is not Hausdorff. (And btw one can show it is also not second-countable.)

Definition 72: Manifold (book Def8.21)

A manifold is a pair $\mathcal{M} = (\mathcal{M}, \mathcal{A}^+)$ consisting of a set $\mathcal{M}$ and a maximal atlas $\mathcal{A}^+$ on $\mathcal{M}$ such that the atlas topology is Hausdorff and second-countable.

A subset $\mathcal{M}$ of a linear space $\mathcal{E}$ may admit many genuinely different maximal atlases, yielding different manifolds (for the same set!). For example: $\mathbb{R}$ with the chart $x \mapsto x$ and $\mathbb{R}$ with the chart $x \mapsto x^3$ yield different maximal atlases on $\mathbb{R}$. However, if $\mathcal{M}$ is what we called an embedded submanifold of $\mathcal{E}$, then there exists a unique maximal atlas that turns $\mathcal{M}$ into a manifold such that:

1. The atlas topology is equivalent to the subspace topology, and
2. A function on $\mathcal{M}$ is smooth (as seen through charts) if and only if it admits a smooth extension.

Definition 73: Smooth map (book Def8.5)

A map $F : \mathcal{M} \to \mathcal{M}'$ is smooth at $x \in \mathcal{M}$ if

$$\widetilde{F} = \psi \circ F \circ \varphi^{-1} : \varphi(U) \to \psi(V)$$

is smooth at $\varphi(x)$, where (U, φ) is a chart of $\mathcal{M}$ around x and (V, ψ) is a chart of $\mathcal{M}'$ around $F(x)$. The map F is smooth if it is smooth at every point $x \in \mathcal{M}$. We call $\widetilde{F}$ a coordinate representative of F.

Fact: This definition does not depend on the choice of charts.

7.2 PDF 503, Tangent vectors

Definition 74: Tangent vectors and tangent space (book Def8.33)

We define an equivalence relation on C_x denoted by $\sim$. Let (U, φ) be a chart of $\mathcal{M}$ around x and consider $c_1, c_2 \in C_x$. Then, $c_1 \sim c_2$ if and only if $\varphi \circ c_1$ and $\varphi \circ c_2$ have the same derivative at $t = 0$, that is,

$$c_1 \sim c_2 \iff (\varphi \circ c_1)'(0) = (\varphi \circ c_2)'(0).$$

The equivalence class of a curve $c \in C_x$ is the set of curves that are equivalent to c as per the above relation:

$$[c] = \{\hat{c} \in C_x : c \sim \hat{c}\}.$$

Each equivalence class is called a *tangent vector* to $\mathcal{M}$ at x. The *tangent space* to $\mathcal{M}$ at x,

denoted by $T_x\mathcal{M}$, is the quotient set

$$T_x\mathcal{M} = C_x/\sim = \{[c] : c \in C_x\},$$

that is, the set of all equivalence classes.

Fact: the equivalence relation is independent of the choice of chart, see lecture PDF 503, [Bou25]. Define the map

$$\theta_x^\varphi : T_x\mathcal{M} \to \mathbb{R}^d : [c] \mapsto (\varphi \circ c)'(0).$$

One can easily check that it is bijective. This bijection naturally induces a linear space structure over $T_x\mathcal{M}$, by copying the linear structure of $\mathbb{R}^d$:

$$a \cdot [c_1] + b \cdot [c_2] := (\theta_x^\varphi)^{-1} \left(a \cdot \theta_x^\varphi([c_1]) + b \cdot \theta_x^\varphi([c_2])\right).$$

This structure, again, is independent of the choice of chart.
We now have two notions of tangent space: Definition 4, Theorem 5 and Definition 74. They are equivalent:

Theorem 75: Equivalence of tangent spaces (book Thm8.35)

If $\mathcal{M}$ is an embedded submanifold of a linear space $\mathcal{E}$, then the map $c \mapsto c'(0)$ is a linear bijection from $C_x/\sim$ to $\ker \mathrm{D}h(x)$, where h is a local defining function for $\mathcal{M}$ around x. Thus, both formalisms yield the same conclusions, always.

8 PDF 801, Optimization on nonsmooth sets through lifts

This is from [LKB24; LKB25], (but not covered in the book). Say we want to minimize a smooth $f : \mathbb{R}^n \to \mathbb{R}$ on the simplex:

$$\min_{x \in \Delta^{n-1}} f(x)$$

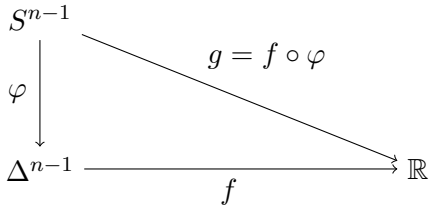
where

$$\Delta^{n-1} = \{x \in \mathbb{R}^n : x_1, \ldots, x_n \geq 0 \text{ and } x_1 + \cdots + x_n = 1\}.$$

The simplex is not a smooth manifold. Can we still use Riemannian optimization? Yes, if we smoothly parameterize the simplex. Let

$$\phi : \mathcal{S}^{n-1} \to \mathbb{R}^n, y \mapsto (y_1^2, \ldots, y_n^2).$$

We have that $\phi(\mathcal{S}^{n-1}) = \Delta^{n-1}$, and the following commuting diagram:



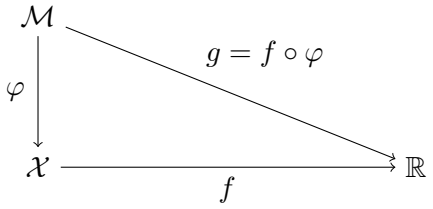
Smooth lifts parameterize nonsmooth sets. Let $\mathcal{S}$ be a (possibly nonsmooth) subset of a Euclidean space $\mathcal{E}$.

Definition 76: Smooth lift (not in book)

A smooth lift of $\mathcal{S}$ is a smooth map $\phi : \mathcal{M} \to \mathcal{E}$ such that $\phi(\mathcal{M}) = \mathcal{X}$, where $\mathcal{M}$ is a smooth manifold, and $\mathcal{X}$ is the set we want to optimize over.

We now have two problems:

"upstairs problem"	(Q)	$\min_{y \in \mathcal{M}} g(y)$
"downstairs problem"	(P)	$\min_{x \in \mathcal{X}} f(x)$



Lets summarize our goal and try to find out what could go wrong: We want to solve the problem (P) downstairs, but to do so, we run a Riemannian optimization algorithm upstairs on (Q). If y is "good" for (Q), is $x = \phi(y)$ "good" for (P)? Do local mins map to local mins? No. See lecture PDF for a visual counterexample.

However, for some special lifts, we do have that local mins map to local mins.

Definition 77: Local $\implies$ local lift

A lift is called "*local* $\implies$ *local*" if for all y we have:

$$y \text{ is a local min of (Q)} \implies x = \phi(y) \text{ is a local min of (P).}$$

Theorem 78: Open map implies local $\implies$ local

If $\phi : \mathcal{M} \to \mathcal{E}$ is an open map (maps open sets to open sets), then it is local $\implies$ local.

Proof. Suppose y is a local minimum for (Q). This implies there exists a neighborhood U of y on $\mathcal{M}$ such that $g(y) \leq g(y')$ for all $y' \in U$.
Since ϕ is open, $\phi(U)$ is a neighborhood of $\phi(y)$ in $\mathcal{X}$.
Now, for all $x' \in \phi(U)$, there exists $y' \in U$ such that $\phi(y') = x'$. Thus:

$$f(\phi(y)) = g(y) \leq g(y') = f(\phi(y')) = f(x').$$

Hence, $\phi(y)$ is a local minimum for (P). □

We will skip the rest of the lecture, which deals with

- First-order optimality conditions, i.e. conditions on ϕ, such that criticality of points transfers from $\mathcal{M}$ to $\mathcal{X}$
- Second-order optimality requirements
- Benign nonconvexity

9 Geodesic convexity

See book [Bou23], chapter 11.

9.1 PDF 701, 114, 702, Geodesic convexity: motivation and basics

Definition 79: Convex set and convex function

A set $S \subset E$ is convex if for all $x, y \in S$, the segment $(1-t)x + ty$ for $t \in [0,1]$ lies in S. A function $f : S \to \mathbb{R}$ is convex if S is convex and for all $x, y \in S$,

$$\forall t \in [0,1], \quad f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

Likewise, f is strictly convex if for $x \neq y$ the $\leq$ is a strict $<$.

This is a fruitful notion because:

1. Local minima are global minima.
2. This comes up in applications and it's easy to spot.
3. We have good algorithms.

Definition 80: Geodesically convex set (book Def11.2)

A subset S of a Riemannian manifold $\mathcal{M}$ is *geodesically convex* if, for every $x, y \in S$, there exists a geodesic segment $c : [0,1] \to \mathcal{M}$ such that $c(0) = x$, $c(1) = y$, and $c(t) \in S$ for all $t \in [0,1]$.

Remark: the intersection of geodesically convex sets is not necessarily geodesically convex! Also, on a manifold, there may exist zero, one or many geodesic segments connecting two given points.

Recall now the definition of a Geodesic (Definition 49), where the intrinsic acceleration of c is defined as $c'' = \frac{Dc'}{dt}$, see also Def 47.

Definition 81: Length of a curve (book eq10.1)

Let $\mathcal{M}$ be a Riemannian manifold. Given a piecewise smooth curve segment $c : [a,b] \to \mathcal{M}$, the length of c is defined as

$$L(c) = \int_a^b \|c'(t)\|_{c(t)}\, dt.$$

The notion of length of a curve leads to a natural notion of distance on $\mathcal{M}$, called the *Riemannian distance*:

$$\operatorname{dist}(x,y) = \inf_c L(c), \tag{12}$$

where the infimum is taken over all piecewise smooth curves $c : [a,b] \to \mathcal{M}$ with $c(a) = x$ and $c(b) = y$.

Theorem 82: Distance induced by geodesics (book Thm10.3)

If $\mathcal{M}$ is connected (meaning each pair of points is connected by a curve segment), then equation (12) defines a metric distance-function (i.e. homogeneous, symmetric, positive, triangle-inequality). Equipped with this distance, $\mathcal{M}$ is a metric space whose metric topology coincides with its atlas topology.

Theorem 83: Hopf-Rinow (book Thm10.8)

A connected Riemannian manifold $\mathcal{M}$ is metrically complete if and only if it is geodesically complete. Additionally, $\mathcal{M}$ is complete (in either sense) if and only if its compact subsets are exactly its closed and bounded subsets.

Remark: one can show that embedded, closed mfds $\mathcal{M}$ in $\mathcal{E}$ are complete.

If the infimum in (12) is attained for some curve segment c, we call c a *minimizing curve*. Remarkably, up to parameterization, these are geodesics as in Definition 49! In other words, two competing generalizations of the notion of straight line from linear spaces to manifolds turn out to be equivalent:

1. one based on shortest paths,
2. one based on zero acceleration.

Theorem 84: Minimizing geodesics (book Thm10.4)

Every minimizing curve admits a constant-speed parameterization such that it is a geodesic, called a *minimizing geodesic.*

One can show that on complete and connected mfds, each pair x, y is connected by a minimizing geodesic.

Definition 85: Geodesically convex function (book Def11.3)

A function $f : S \to \mathbb{R}$ is geodesically (strictly) convex if S is geodesically convex and $f \circ c : [0, 1] \to \mathbb{R}$ is (strictly) convex for each geodesic segment $c : [0, 1] \to \mathcal{M}$ whose image is in S (with $c(0) \neq c(1)$).

Properties:

1. Sublevel sets of g-convex functions are g-convex sets.
2. Intersections of such sublevel sets are g-convex sets.
3. Sums of nonnegatively scaled g-convex functions are g-convex.
4. The pointwise maximum of g-convex functions is g-convex.

Theorem 86: Geodesic convexity: local minimizers are global (book Thm11.6)

If $f : S \to \mathbb{R}$ is geodesically convex, then any local minimizer is a global minimizer.

Proof. Assume for contradiction that $x \in S$ is a local minimizer but not a global minimizer. Then, there exists $y \in S$ such that $f(y) < f(x)$. There also exists a geodesic c connecting $c(0) = x$ to $c(1) = y$ in S such that, for all $t \in (0, 1]$,

$$\begin{aligned} f(c(t)) &\leq (1 - t)f(x) + tf(y) \\ &= f(x) + t(f(y) - f(x)) \\ &< f(x), \end{aligned}$$

which contradicts the claim that x is a local minimizer. □

Theorem 87: Continuity of geodesically convex functions (book Prop11.9)

If $f : S \to \mathbb{R}$ is geodesically convex, then f is continuous on the interior of S.

Compact manifolds are complete, and continuous functions on compact sets attain their maximum. Consider geodesics through the maximizer to conclude:

Theorem 88: Geodesic convexity on compact manifolds (book Corollary 11.10)

If $\mathcal{M}$ is a connected, compact Riemannian manifold and $f : \mathcal{M} \to \mathbb{R}$ is geodesically convex, then f is constant.

The take-away is that on compact manifolds, geodesic convexity is only interesting on subsets of a connected component.

The slides now discuss several more definitions of convex functions.

9.2 PDF 213, Linear convergence with Polyak-Lojasiewicz

The Polyak–Lojasiewicz (PL) condition on a Riemannian manifold $\mathcal{M}$ ensures that the squared gradient norm bounds the optimality gap of a differentiable function f, and implies linear convergence when paired with a sufficient decrease condition in iterative optimization.

In other words, if for all iterations the function value decreases by a fixed amount proportional to the squared gradient norm (sufficient decrease), and the PL condition holds, then the function values converge linearly to the minimum.

f satisfies the PL condition if there exists $\mu > 0$ on a set $S \subseteq \mathcal{M}$ such that

$$f(x) - f^* \leq \frac{1}{2\mu} \| \operatorname{grad} f(x) \|_x^2 \qquad \text{for all } x \in S,$$

where $f^* = \inf_{x \in S} f(x)$. In words: within S, the squared gradient norm bounds the optimality gap.

The PL condition is satisfied for geodesically strongly convex functions, where the Riemannian Hessian is bounded below by a multiple of the identity. An application includes computing Fréchet means on manifolds by minimizing the average squared distance to given points, where the behavior of the Hessian of the objective function plays a critical role.

See slides for details.

References

[Bou23] Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge: Cambridge University Press, 2023 (cit. on pp. 2, 8, 11, 16, 18, 19, 21, 22, 24, 25, 29).

[Bou25] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. 2025. URL: https://www.nicolasboumal.net/book/#lectures (visited on 02/04/2025) (cit. on p. 28).

[Lee03] John M. Lee. *Introduction to Smooth Manifolds*. 1st ed. Vol. 218. Graduate Texts in Mathematics. Springer Book Archive, 63 b/w illustrations. New York, NY: Springer, 2003, pp. XVII + 631. ISBN: 978-0-387-21752-9. DOI: 10.1007/978-0-387-21752-9. URL: https://doi.org/10.1007/978-0-387-21752-9 (cit. on pp. 2, 3, 25).

[LKB24] Eitan Levin, Joe Kileel and Nicolas Boumal. "The effect of smooth parametrizations on nonconvex optimization landscapes". In: *Mathematical Programming* 209.1–2 (Mar. 2024), pp. 63–111. ISSN: 1436-4646. DOI: 10.1007/s10107-024-02058-3. URL: http://dx.doi.org/10.1007/s10107-024-02058-3 (cit. on p. 28).

[LKB25] Eleny Levin, Joey Kileel and Nicolas Boumal. "The effect of smooth parametrizations on nonconvex optimization landscapes". In: *Mathematical Programming* 209 (2025), pp. 63–111. DOI: 10.1007/s10107-024-02058-3. URL: https://doi.org/10.1007/s10107-024-02058-3 (cit. on p. 28).

[TR24] N. Tsakanikas and L. E. Rösler. *Differential Geometry II - Smooth Manifolds*. Lecture notes, Differential Geometry, Winter Term 2024/2025, EPFL. Dec. 2024 (cit. on pp. 3, 5, 13, 25).