Federated learning on Riemannian manifolds

-

March 5, 2025

Contents

Chapter 1

Introduction

Motivation

Optimization: a search space S and a cost function $f : S \to \mathbb{R}$. The goal is to find

$$x^* \in \arg\min_{x \in S} f(x).$$

Usually, S is a linear space like $\mathbb{R}^n$, for which the previous optimization problem is called *unconstrained optimization*. For f smooth enough, we have the notions of gradient, and even Hessian that give rise to gradient descent and Newton algorithms to efficiently search a solution.

When S is a smooth surface, x is not allowed to move freely in $\mathbb{R}^n$: it is *constrained optimization*. How can we generalize algorithms such as gradient descent or Newton's method in the context of smooth manifolds? How to define the gradient/Hessian?
We will take advantage of the fact that smooth manifolds can be linearized locally around every point x thanks to the tangent space at x. Endowing this tangent space with its own inner product (varying smoothly with x) leads to the construction of Riemannian manifold, on which we can properly define gradients and Hessians.

Let $x_1, \ldots, x_n \in \mathbb{R}^d$ a large collection of centered data. We wish to determine the k first principal components of such a cloud of points. Formally, if $X \in \mathbb{R}^{d \times n}$, we look for a matrix $U \in \mathbb{R}^{d \times k}$ with orthonormal columns $u_1, \ldots, u_k \in \mathbb{R}$, i.e. matrices U belonging to

$$\mathrm{St}(d, k) := \{U \in \mathbb{R}^{d \times k} : U^T U = \mathrm{I}_k\}.$$

For a smooth function $f : \mathcal{E} \to \mathbb{R}$ (where $\mathcal{E}$ is a Euclidean space), the gradient is defined w.r.t. a inner product: $\mathrm{grad} f(x)$ is the unique element of $\mathcal{E}$ s.t. for all $v \in \mathcal{E}$,

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

where $Df(x)[v] = \lim_{t \to 0} \frac{f(x+tv)-f(x)}{t}$. Consequently, the gradient of f *depends* on a choice of an inner product, while the differential does not.
To define a proper notion of gradients for $f : S^{d-1} \to \mathbb{R}$, we need to provide a meaningful concept of differential for f at x, namely $Df(x) : T_x S^{d-1} \to \mathbb{R}$ and introduce a relevant inner product $\langle \cdot, \cdot \rangle_x$ on the tangent space $T_x S^{d-1}$ so that

$$\forall v \in T_x S^{d-1},\ Df(x)[v] = \langle \mathrm{grad} f(x), v \rangle_x.$$

Here on the sphere S^{d-1} we restric the inner product of $\mathbb{R}^d$ to T_xS^{d-1}. For Riemannian submanifolds, we will define the Riemannian gradient as the orthogonal projection of the classical gradient to the tangent spaces.
Recall that the Euclidean Hessian of f at x is the linear map $\mathrm{Hess}f(x) : \mathcal{E} \to \mathcal{E}$ defined by

$$\mathrm{Hess}f(x)[v] = D(\mathrm{grad}f)(x)[v] = \lim_{t\to 0} \frac{\mathrm{grad}f(x+tv) - \mathrm{grad}f(x)}{t}.$$

Schwarz theorem implies that $\mathrm{Hess}f(x)$ is symmetric w.r.t. the inner product of $\mathcal{E}$.

Chapter 2

Riemannian geometry

2.1 Embedded submanifolds of Euclidean space

2.1.1 Embedded submanifolds and tangent spaces

Definition 2.1.1 (Embedded submanifold)**.** Let $\mathcal{E}$ be a linear space of dimension d and let $\mathcal{M} \subseteq \mathcal{E}$ (non empty). We say that $\mathcal{M}$ is a smooth *embedded submanifold* of $\mathcal{E}$ of dimension n if either of the following conditions is satisfied:

1. $n = d$ and $\mathcal{M}$ is open in $\mathcal{E}$.
2. $n = d - k$ for some $k \geq 1$ and for each $x \in \mathcal{M}$, there exists a neighborhood U of x in $\mathcal{E}$ and a smooth function $h : U \to \mathbb{R}^k$ such that

$$M \cap U = h^{-1}(0) \quad \text{and} \quad \text{rank}Dh(x) = k.$$

Remark 2.1.1. Such function h is often called a local defining function at x.

Recall the following definition of *diffeomorphism.*

Definition 2.1.2 (Diffeomorphism)**.** A diffeomorphism is a bijective map $F : U \to V$ where U, V are open sets such that F and F^{-1} are smooth.

We have the following characterisation of submanifold using local diffeomorphism.

Theorem 2.1.1. *With the same notations, $\mathcal{M}$ is a submanifold of dimension $n = d - k$ if and only if for each $x \in \mathcal{M}$, there exists a neighborhood U of x in $\mathcal{E}$, an open set $V \subset \mathbb{R}^d$ and a diffeomorphism $F : U \to V$ such that*

$$F(\mathcal{M} \cap U) = \{y \in \mathbb{R}^d : y_{n+1} = \cdots = y_d = 0\} \cap V.$$

The proof of the Theorem 2.1.1 relies on the inverse function theorem.

Definition 2.1.3 (Tangent space)**.** Let $\mathcal{M} \subseteq \mathcal{E}$. For all $x \in \mathcal{M}$, define

$$T_x\mathcal{M} := \{c'(0) \mid c : I \to \mathcal{M} \text{ smooth and } c(0) = x\}$$

where I is an open interval containing $t = 0$. The set $T_x\mathcal{M}$ the *tangent space* to $\mathcal{M}$ at x and vectors of $T_x\mathcal{M}$ are called *tangent vectors* at x.

Remark 2.1.2. A vector v is in the tangent space of x if and only if there exists a smooth curve on $\mathcal{M}$ going through x with speed v.

We have the following characterization of the tangent space on x via the differential of the local defining function. In particular, $T_x\mathcal{M}$ is a linear space.

Theorem 2.1.2. *Let $\mathcal{M}$ an embedded submanifold of $\mathcal{E}$. Let $x \in \mathcal{M}$. If $\mathcal{M}$ is an open submanifold, then $T_x\mathcal{M} = \mathcal{E}$. Otherwise,*

$$T_x\mathcal{M} = \ker Dh(x)$$

with h any local defining function at x.

Remark 2.1.3. The dimension of $T_x\mathcal{M}$ (independent of x) coincides with $\dim \mathcal{M}$.

Example 2.1.1.

- $S^{d-1} = \{x \in \mathbb{R}^d \mid x^T x = 1\} = h^{-1}(\{0\})$ where $h : \mathbb{R}^d \to \mathbb{R}, x \mapsto x^T x - 1$. Since $Dh(x)[v] = 2x^T x$, we have $\text{rank}Dh(x) = 1$ hence S^{d-1} is an embedded submanifold of $\mathbb{R}^d$ of dimension $n = d - 1$ and whose tangent spaces are given by

$$T_x S^{d-1} = \ker Dh(x) = \{x \in \mathbb{R}^d \mid x^T v = 0\}.$$

- $\text{Sym}_n^{++} := \{X \in \mathcal{M}_n(\mathbb{R}) \mid X \succ 0\}$ is an open set of $\text{Sym}(n)$ hence for all $\Sigma \in \text{Sym}_n^{++}$,

$$T_\Sigma \text{Sym}_n^{++} \simeq \text{Sym}(n).$$

- Let $O \in \mathcal{O}_n(\mathbb{R}) := \{O \in \mathcal{M}_n(\mathbb{R}) \mid OO^T = \text{I}_n\}$. Let $f : O \mapsto O^T O - \text{I}_n$ is a smooth map defining $\mathcal{O}_n(\mathbb{R})$. We have $Df(O)[v] = O^T v + v^T O$ hence

$$T_O\mathcal{O}_n(\mathbb{R}) = \left\{v \in \mathbb{R}^{p\times p} \mid O^T v + v^T O = 0_n\right\} = \left\{O\Omega \mid \Omega \in \mathbb{R}^{p\times p}, \Omega^T = -\Omega\right\}.$$

We equip embedded manifolds of $\mathcal{E}$ with the topology induced by $\mathcal{E}$.
Cartesian products of manifolds are indeed manifolds and the tangent spaces are fully known.

Proposition 2.1.1. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two embedded submanifolds of two Euclidean spaces $\mathcal{E}_1$ and $\mathcal{E}_2$ respectively. Then $\mathcal{M}_1 \times \mathcal{M}_2$ is an embedded submanifold of $\mathcal{E}_1 \times \mathcal{E}_2$ of dimension $\dim \mathcal{M}_1 + \dim \mathcal{M}_2$ with tangent spaces*

$$T_{(x_1,x_2)}(\mathcal{M}_1 \times \mathcal{M}_2) = T_{x_1}\mathcal{M}_1 \times T_{x_2}\mathcal{M}_2.$$

Example 2.1.2.

$$\text{OB}(d, k) := S^{d-1} \times \cdots \times S^{d-1} = (S^{d-1})^k$$

is an embedded submanifold of $\mathbb{R}^d$ called the *oblique manifold*.

2.1.2 Smooth maps on embedded submanifolds

In optimization, two important examples of maps between manifolds are cost functions $L : \mathcal{M} \to \mathbb{R}$ and iteration maps $\mathcal{M} \to \mathcal{M}$.

Definition 2.1.4 (Smooth maps)**.** Let $\mathcal{M}$ and $\mathcal{M}'$ be two embedded submanifolds of $\mathcal{E}$ and $\mathcal{E}'$ respectively. A map $F : \mathcal{M} \to \mathcal{M}'$ is *smooth* at $x \in \mathcal{M}$ is there exists a function $\bar{F} : U \to \mathcal{E}'$ smooth on an open neighborhood U of x in $\mathcal{E}$ and such that

$$F \equiv \bar{F} \text{ on } \mathcal{M} \cap U.$$

We call $\bar{F}$ a *smooth* extension of F around x.
The map F is *smooth* if it is smooth at all $x \in \mathcal{M}$.

Remark 2.1.4. It means that if $\bar{F}$ is a smooth map on $\mathcal{E}$, the restriction $F = \bar{F}_{|\mathcal{M}}$ is smooth on $\mathcal{M}$.

The following proposition makes it possible to consider the reverse, i.e. smooth extensions: a smooth map on $\mathcal{M}$ always admits a smooth extension to a neighborhood of $\mathcal{M}$.

Proposition 2.1.2. *With the above notations, $F : \mathcal{M} \to \mathcal{M}'$ is smooth if and only if $F = \bar{F}_{|\mathcal{M}}$ where $\bar{F}$ is some smooth map from a neighborhood of $\mathcal{M}$ in $\mathcal{E}$ to $\mathcal{E}'$.*

Definition 2.1.5 (Scalar field)**.** A *scalar field* on a manifold $\mathcal{M}$ is a function $f : \mathcal{M} \to \mathbb{R}$. If f is smooth, we call it a *smooth scalar field*. Let us denote by $\mathcal{F}(\mathcal{M})$ the set of smooth scalar fields on $\mathcal{M}$.

2.1.3 Differential of a smooth map

Recall that if $\bar{F} : U \subset \mathcal{E} \to \mathcal{E}'$ is smooth between two *vector spaces*, the differential of $\bar{F}$ at $x \in U$ is the linear map $D\bar{F}(x) : \mathcal{E} \to \mathcal{E}'$ defined by

$$D\bar{F}(x)[v] = \lim_{t \to 0} \frac{\bar{F}(x + tv) - \bar{F}(x)}{t}.$$

Let us try to generalize it to a smooth function $F : \mathcal{M} \to \mathcal{M}'$. Generally, $x + tv$ does not belong to $\mathcal{M}$. However, $c : t \mapsto x + tv$ is a curve in $\mathcal{E}$ which goes through x at velocity v. Let us use curves on $\mathcal{M}$ instead.

We know that for any tangent vector $v \in T_x\mathcal{M}$, there exists a smooth curve $c : \mathbb{R} \to \mathcal{M}$ such that $c(0) = x$ and $c'(0) = v$. We can then define the smooth curve on $\mathcal{M}'$ by considering $t \mapsto F(c(t))$, passing through $F(x)$ at velocity $\frac{d}{dt}F(c(t))_{|t=0}$ being the tangent vector of $\mathcal{M}'$ at $F(x)$.

Definition 2.1.6 (Differential)**.** The differential of $F : \mathcal{M} \to \mathcal{M}'$ at the point $x \in \mathcal{M}$ is the linear map $DF(x) : T_x\mathcal{M} \to T_{F(x)}\mathcal{M}'$ defined by

$$DF(x)[v] = \frac{d}{dt}F(c(t))_{|t=0} = (F \circ c)'(0)$$

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

Remark 2.1.5. Let us notice that

(i) this definition is independent on the choice of the curve c,

(ii) $DF(x)$ is linear,

(iii) with the same notations as above, $DF(x) = D\bar{F}(x)_{|T_x\mathcal{M}}$ and it does not depend on the choice of smooth extension $\bar{F}$.

Example 2.1.3. Let $A \in \mathrm{Sym}(d)$. Set $f : S^{d-1} \to \mathbb{R}, x \mapsto x^T Ax$. It can be smoothly extended to $\mathbb{R}^d$ by $\bar{f}(x) = x^T Ax$ hence f is smooth. Let us compute its differential. For $v \in \mathbb{R}^d$, we have

$$D\bar{f}(x)[v] = 2x^T Av$$

hence $Df(x)[v] = 2x^T Av$ for all $v \in T_xS^{d-1} = \{v \in \mathbb{R}^d \mid x^T v = 0\}$.

2.1.4 Vector fields and tangent bundle

Before defining vector fields, we need the notion of *tangent bundle*.

Definition 2.1.7 (Tangent bundle)**.** The tangent bundle of a manifold $\mathcal{M}$ is the disjoint union of the tangent spaces of $\mathcal{M}$, i.e.

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

The tangent bundle of a manifold possesses the nice property of being also a manifold.

Theorem 2.1.3. *If $\mathcal{M}$ is an embedded submanifold of $\mathcal{E}$, the tangent bundle $T\mathcal{M}$ is an embedded submanifold of $\mathcal{E} \times \mathcal{E}$ of dimension $2 \dim \mathcal{M}$.*

Now we are in a position to define vector fields.

Definition 2.1.8 (Vector field)**.** A *vector field* on a manifold $\mathcal{M}$ is a map $V : \mathcal{M} \to T\mathcal{M}$ such that

$$\forall x \in \mathcal{M},\ V(x) \in T_x\mathcal{M}.$$

If V is a smooth map (between manifolds), V is said to be a *smooth vector field*.

Let us denote by $\mathcal{X}(\mathcal{M})$ the set of smooth vector fields on $\mathcal{M}$.

Proposition 2.1.3. *Let $\mathcal{M}$ be an embedded submanifold of $\mathcal{E}$ and V a vector field on $\mathcal{M}$. The following assertions are equivalent:*

- *there exists a smooth vector field $\bar{V}$ on a neighborhood of $\mathcal{M}$ such that $V = \bar{V}_{|\mathcal{M}}$*
- *V is smooth on $\mathcal{M}$*

This means a vector field (on an embedded submanifold) is smooth if and only if it is the restriction of a smooth vector field on a neighborhood of the manifold in the embedding space.

Example 2.1.4. Let $\mathcal{M}_1 \times \mathcal{M}_2$ be a product manifold. The tangent bundle of $\mathcal{M}_1 \times \mathcal{M}_2$ is given by

$$T(\mathcal{M}_1 \times \mathcal{M}_2) = T\mathcal{M}_1 \times T\mathcal{M}_2.$$

2.1.5 Retractions

Let $x \in \mathcal{M}$ and $v \in T_x\mathcal{M}$. How to move from x along the direction v while staying on $\mathcal{M}$? There are many smooth curves c such that $c(0) = x$ and $c'(0) = v$. Retraction is a concept were we pick a particular smooth curve for each couple $(x, v) \in T\mathcal{M}$.

Definition 2.1.9 (Retraction)**.** A smooth map $R : T\mathcal{M} \to \mathcal{M} : (x, v) \mapsto R_x(v)$ is called a *retraction* if any curve $c : t \mapsto R_x(tv)$ verifies

$$c(0) = x \quad \text{and} \quad c'(0) = v.$$

Remark 2.1.6.

- For an embedded manifold $\mathcal{M} \subset \mathcal{E}$, the smoothness of R is equivalent to the existence of a smooth map $\bar{R} : \mathcal{E} \times \mathcal{E} \to \mathcal{E}$ from a neighborhood of $T\mathcal{M}$ such that $R = \bar{R}_{|T\mathcal{M}}$.
- It can be easily shown that R is a retraction if and only if for all $(x, v) \in T\mathcal{M}$, we have $R_x(0) = x$ and $DR_x(0) : T_x\mathcal{M} \to T_x\mathcal{M}$ is the identity map.

Example 2.1.5. If $x \in S^{d-1}$ and $v \in T_x S^{d-1}$, i.e. $x^T v = 0$, then

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

defines a retraction. Indeed, consider the curve $c : \mathbb{R} \to S^{d-1}$ define dby

$$c(t) = R_x(tv) = \frac{x+tv}{\sqrt{1+t^2\|v\|^2}}.$$

It is clear that $c(0) = x$ and using chain rule, $c'(0) = v$. The smoothness of R is ensured by the smoothness of its extension to $\mathbb{R}^d \times \mathbb{R}^d$.

Remark 2.1.7. One can also check that $\tilde{R}_x(v) := \cos(\|v\|)x + \frac{\sin(\|v\|)}{\|v\|}v$ is also a retraction on the sphere S^{d-1} that traces the great circle going through x at velocity v.
We will see later that those retractions R and $\tilde{R}$ are geodesics (for the right Riemannian metric) and that R is in fact the *exponential map*.

Example 2.1.6. Let $\mathcal{M}_1, \mathcal{M}_2$ be two embedded submanifolds endowed with retractions $R^{(1)}$ and $R^{(2)}$. Then, $R : T(\mathcal{M}_1 \times \mathcal{M}_2) \to \mathcal{M}_1 \times \mathcal{M}_2$ defined by

$$R_{(x_1,x_2)}(v_1, v_2) := (R^{(1)}_{x_1}(v_1), R^{(2)}_{x_2}(v_2))$$

is a licit choice of a retraction for the product manifold.

2.2 Riemannian manifolds

2.2.1 Riemannian metric

The idea behind Riemannian manifolds is to equip every tangent space with an inner product varying smoothly with each point.
We define an *inner product* on $T_x\mathcal{M}$ (bilinear, symmetric, positive definite function)

$$\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}.$$

It classically induces a norm $\|v\|_x := \sqrt{\langle v, v \rangle_x}$ for $v \in T_x\mathcal{M}$.
We refer to *metric* by the choice of inner product $\langle \cdot, \cdot \rangle_x$ for every $x \in \mathcal{M}$.

Definition 2.2.1 (Riemannian metric, Riemannian manifold)**.**

1. A metric $\langle \cdot, \cdot \rangle_x$ on $\mathcal{M}$ is called a *Riemannian metric* if for all smooth vector fields V, W on $\mathcal{M}$, the function
$$x \mapsto \langle V(x), W(x) \rangle$$
is smooth from $\mathcal{M}$ to $\mathbb{R}$.

2. $\mathcal{M}$ is a *Riemannian manifold* if it is equipped with a Riemannian metric.

Remark 2.2.1. When $\mathcal{M}$ is an embedded submanifold of a Euclidean space $\mathcal{E}$ with an inner product $\langle \cdot, \cdot \rangle$, one way of defining a Riemannian metric on each tangent space to to restrict the natural inner product of $\mathcal{E}$ to them. It is called the *induced metric*.

The previous remark is formalized with the following proposition and results from considerating smooth extensions of vector fields.

Proposition 2.2.1. *Let $\mathcal{M} \subset \mathcal{E}$ with $(\mathcal{E}, \langle \cdot, \cdot \rangle)$ a Euclidean space. The metric defined on $\mathcal{M}$ at each x by the restriction*

$$\forall u, v \in T_x\mathcal{M}, \langle u, v \rangle_x = \langle u, v \rangle$$

is a Riemannian metric.

Definition 2.2.2 (Riemannian submanifold)**.** Let $\mathcal{M} \subset \mathcal{E}$ an embedded submanifold of a Euclidean space. We call $\mathcal{M}$ a *Riemannian submanifold of* $\mathcal{E}$ when it is equipped with the Riemannian metric obtained by restricting the metric of $\mathcal{E}$ on the tangent spaces.

Example 2.2.1. S^{d-1} is a Riemannian submanifold of $\mathbb{R}^d$ when equipped with $\langle u, v \rangle_x = \langle u, v \rangle = u^T v$.

Example 2.2.2. Let $(\mathcal{M}_1, \langle \cdot, \cdot \rangle^{\mathcal{M}_1})$ and $(\mathcal{M}_2, \langle \cdot, \cdot \rangle^{\mathcal{M}_2})$ be two Riemannian manifolds. The product $\mathcal{M}_1 \times \mathcal{M}_2$ is a Riemannian manifold whose product metric is defined as

$$\forall (u_1, u_2), (v_1, v_2) \in T_{(x_1,x_2)}(\mathcal{M}_1 \times \mathcal{M}_2), \quad \langle (u_1, u_2), (v_1, v_2) \rangle_{(x_1,x_2)} = \langle u_1, v_1 \rangle_x^{\mathcal{M}_1} + \langle u_2, v_2 \rangle_{x_2}^{\mathcal{M}_2}$$

and defines a Riemannian metric.

2.2.2 Riemannian gradients

Let us consider a smooth function $f : \mathcal{M} \to \mathbb{R}$. How can we define its gradient?

Definition 2.2.3 (Riemannian gradient)**.** Let $f : \mathcal{M} \to \mathbb{R}$ a smooth map. The *Riemannian gradient* of f is the unique vector field $\mathrm{grad} f$ on $\mathcal{M}$ defined by

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

Let us detail the case where $\mathcal{M}$ is a Riemannian submanifold of a Euclidean space $\mathcal{E}$. The smoothness of f implies the existence of a smooth extension $\bar{f}$ defined on the neighborhood of $\mathcal{M}$ in $\mathcal{E}$, with Euclidean gradient $\mathrm{grad} \bar{f}$ and

$$\forall (x, v) \in T\mathcal{M}, \; \langle v, \mathrm{grad} f(x) \rangle_x = \langle v, \mathrm{grad} \bar{f}(x) \rangle.$$

Since $T_x\mathcal{M}$ is a subspace of $\mathcal{E}$ and $\mathrm{grad} \bar{f} \in \mathcal{E}$, we have the unique orthogonal decomposition:

$$\mathrm{grad} \bar{f}(x) = \underbrace{\mathrm{grad} \bar{f}(x)_{\|}}_{\in T_x\mathcal{M}} + \underbrace{\mathrm{grad} \bar{f}(x)_{\perp}}_{\in T_x\mathcal{M}^{\perp}}$$

Hence

$$\forall (x, v) \in T\mathcal{M}, \; \langle v, \mathrm{grad} f(x) \rangle_x = \langle v, \mathrm{grad} \bar{f}(x)_{\|} \rangle_x.$$

Uniqueness ensures that $\mathrm{grad} f(x) = \mathrm{grad} \bar{f}(x)_{\|}$, i.e. the orthogonal projection on the tangent space of x of the gradient of the extension.

Recall that the orthogonal projector to $T_x\mathcal{M}$ is the linear map $\mathrm{Proj}_x : \mathcal{E} \to \mathcal{E}$ such that:

1. $\mathrm{Im}(\mathrm{Proj}_x) = T_x\mathcal{M}$
2. $\mathrm{Proj}_x \circ \mathrm{Proj}_x = \mathrm{Proj}_x$
3. for all $v \in T_x\mathcal{M}$ and $u \in \mathcal{E}$, $\langle u - \mathrm{Proj}_x(u), v \rangle = 0$

Recall that orthogonal projectors are self-adjoint:

$$\forall u, v \in \mathcal{E}, \; \langle u, \mathrm{Proj}_x(v) \rangle = \langle \mathrm{Proj}_x(u), v \rangle.$$

Proposition 2.2.2. *Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function on a Riemannian submanifold of $(\mathcal{E}, \langle \cdot, \cdot \rangle)$. The Riemannian gradient of f is given by*

$$grad f(x) = Proj_x(grad \bar{f}(x)),$$

where $\bar{f}$ is any smooth extension of f to a neighborhood of $\mathcal{M}$ in $\mathcal{E}$.

Example 2.2.3. Let $f : S^{d-1} \to \mathbb{R}, x \mapsto x^T A x$ with $A = A^T \in \mathbb{R}^{d \times d}$. We have

$$D\bar{f}(x)[v] = 2x^T A x, \quad \text{therefore} \quad \text{grad}\bar{f}(x) = 2Ax.$$

Let us compute the Riemannian gradient of f. Since S^{d-1} can be seen as a Riemannian submanifold of $\mathbb{R}^d$ with the induced Riemannian metric, we need to determine the orthogonal projector on the tangent space.
Since

$$T_x S^{d-1} = \{v \in \mathbb{R}^d \mid x^T v = 0\} = \text{Vect}(x)^\perp$$

we have

$$\text{Proj}_x(u) = u - (x^T u)x = (\text{I}_x - xx^T)u.$$

Hence

$$\text{grad} f(x) = 2(Ax - (x^T Ax)x).$$

Notice that the gradient vanishes at unit-norm eigenvectors of A.

Example 2.2.4. Embedded submanifold that is not a Riemannian submanifold Let $\mathcal{E} = \mathbb{R}^d$ endowed with its Euclidean inner product $\langle u, v \rangle_{\mathcal{E}} = u^T v$.
Let $G : U \to \mathbb{R}^{d \times d}$ be a smooth map such that $G(x) \in \text{Sym}(d)^{++}$ for all $x \in U$, where U is an open subset of $\mathcal{E}$. On $\mathcal{M} = U$, we define the metric

$$\langle u, v \rangle_{\mathcal{M}} = u^T G(x) v.$$

It defines a Riemannian metric on U.
Let $f : U \to \mathbb{R}$. Let us compute the Riemannian gradient of f w.r.t. $\langle \cdot, \cdot \rangle_{\mathcal{M}}$. For $v \in \mathbb{R}^n$, we have

$$\langle v, G(x)\text{grad}_{\mathcal{M}} f(x) \rangle_{\mathcal{E}} = v^T G(x) \text{grad}_{\mathcal{M}} f(x) = \langle v, \text{grad}_{\mathcal{M}} f(x) \rangle_{\mathcal{M}} = \langle v, \text{grad}_{\mathcal{E}} f(x) \rangle_{\mathcal{E}}.$$

Unicity yields

$$\text{grad}_{\mathcal{M}} f(x) = G(x)^{-1} \text{grad}_{\mathcal{E}} f(x).$$

In the case of smooth functions on Riemannian product manifolds, the gradient is explicit.

Proposition 2.2.3. *Let $f : \mathcal{M}_1 \times \mathcal{M}_2 \to \mathbb{R}$ a smooth function defined on a Riemannian product manifold. Then*

$$grad f(x, y) = \left(grad(x \mapsto f(x, y))(x), grad(y \mapsto f(x, y))(y)\right).$$

2.2.3 Riemannian connection

Let us first define the notion of connection, which generalizes the notion of differential for vectors fields.

Definition 2.2.4 (Connection)**.** A *connection* on $\mathcal{M}$ is an operator

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

$$\nabla(u, V) \mapsto \nabla_u V$$

such that

$$u \in T_x\mathcal{M} \Rightarrow \nabla_u V \in T_x\mathcal{M}$$

and for all $U, V, W \in \mathcal{X}(\mathcal{M})$, $u, w \in T_x\mathcal{M}$ and $a, b \in \mathbb{R}$:

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

Remark 2.2.2.

- $\nabla_u V$ is a shortcut notation for $\nabla_{(x,u)} V$.
- No Riemannian metric is needed.
- There exists infinitely many connections on any manifold. Requiring additional assumptions w.r.t. the Riemannian structure narrows down to one connection, the *Levi–Civita connection.*

Example 2.2.5.

1. On a linear space $\mathcal{E}$, the following define a connection:
$$\nabla_u V = DV(x)[u].$$
2. For $\mathcal{M}$ an embedded subanifold in a Euclidean space $\mathcal{E}$, if Proj_x defines the orthogonal projector from $\mathcal{E}$ to $T_x\mathcal{M}$ and $\bar{V}$ is a smooth extension of V,
$$\nabla_u V = \mathrm{Proj}_x(D\bar{V}(x)[u])$$
defines a valid connection.

All connections coincide at critical points of a vector field.

Proposition 2.2.4. *Let $\mathcal{M}$ be a manifold with arbitrary connection ∇ and V a smooth vector field $V \in \mathcal{X}(\mathcal{M})$. Let $x \in \mathcal{M}$ such that $V(x) = 0$. Then*

$$\forall u \in T_x\mathcal{M},\ \nabla_u V = DV(x)[u].$$

Before moving on to adding two additional properties that guarantee symmetry of the Hessian, let us introduce some notations.

Definition 2.2.5. For $U, V \in \mathcal{X}(\mathcal{M})$ and $f : \mathcal{U} \subset \mathcal{M} \to \mathbb{R}$. We set:

$$(Uf) \in \mathbb{F}(\mathcal{U}) \text{ such that } (Uf)(x) = Df(x)[U(x)];$$

$$[U, V] : \mathbb{F}(\mathcal{U}) \to \mathbb{F}(\mathcal{U}) \text{ such that } [U, V]f = U(Vf) - V(Uf) :$$

$$\langle U, V\rangle \in \mathbb{F}(\mathcal{M}) \text{ such that } \langle U, V\rangle(x) = \langle U(x), V(x)\rangle_x.$$

Remark 2.2.3. The commutator $[U, V]$ is called the *Lie bracket.*
Remark that

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

The notation Uf (which should not be confused with $fU : x \mapsto f(x)U(x)$) captures the action of the vector field U on f though derivation.

Let us now state the main result of this section.

Theorem 2.2.1. *Let $\mathcal{M}$ be a Riemannian manifold. There exists a unique connection ∇ verifying that for all $U, V, W \in \mathcal{X}(\mathcal{M})$:*

1. *Symmetry: for all $f \in \mathbb{F}(\mathcal{M})$,*

$$[U, V]f = (\nabla_U V - \nabla_V U)f$$

2. *Compatibility with the metric:*

$$U\langle V, W\rangle = \langle \nabla_U V, W\rangle + \langle V, \nabla_U W\rangle.$$

It is called the Levi–Civita *or* Riemannian *connection. ∇ is characterized by the* Koszul formula*:*

$$2\langle \nabla_U V, W\rangle = U\langle V, W\rangle + V\langle W, U\rangle - W\langle U, V\rangle - \langle U, [V, W]\rangle + \langle V, [W, U]\rangle + \langle W, [U, V]\rangle.$$

Remark 2.2.4. For $\mathcal{M} = \mathcal{E}$ being a Euclidean space, we have $W\langle V, U\rangle = \langle WU, V\rangle + \langle V, WU\rangle$ so the Koszul formula comes down to

$$\langle UV, W\rangle = \langle \nabla_U V, W\rangle.$$

Example 2.2.6.

- The Riemannian connection on a Euclidean space $\mathcal{E}$ with any metric $\langle \cdot, \cdot\rangle$ is

$$\nabla_u V = DV(x)[u]$$

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

$$\nabla_u V = \text{Proj}_x(D\bar{V}(x)[u])$$

  is the Riemannian connection on $\mathcal{M}$.

2.2.4 Riemannian Hessian

Definition 2.2.6. Let $\mathcal{M}$ be a Riemannian manifold with its Riemannian connection ∇. The Riemannian Hessian of $f \in \mathbb{F}(\mathcal{M})$ at $x \in \mathcal{M}$ is the linear map

$$\mathrm{Hess} f(x) : T_x\mathcal{M} \to T_x\mathcal{M}$$

such that

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

The very properties of the Riemannian connection ensure the self-adjointness of the Riemannian Hessian.

Proposition 2.2.5. *The Riemannian Hessian is self-adjoint w.r.t. the Riemannian metric:*

$$\forall x \in \mathcal{M}, \forall u, v \in T_x\mathcal{M}, \quad \langle Hess f(x)[u], v\rangle_x = \langle u, Hess f(x)[v]\rangle_x.$$

Remark 2.2.5. Since $\mathrm{Hess} f(x)$ is symmetric, the spectral theorem ensures that all of its eigenvalues are real and its corresponding eigenvectors can be chosen to form an orthonormal basis of $T_x\mathcal{M}$ w.r.t. $\langle \cdot, \cdot \rangle_x$.

Let us see now how to compute Riemannian Hessians in the case where $\mathcal{M}$ is a Riemannian submanifold of a Euclidean space.

Proposition 2.2.6. *Let $f : \mathcal{M} \to \mathbb{R}$ and $\bar{G}$ a smooth extension of grad$f$. Then*

$$Hess f(x)[u] = Proj_x(D\bar{G}(x)[u]).$$

Example 2.2.7. Let $\bar{f}(x) = \frac{1}{2}x^T Ax$ for $A \in \mathrm{Sym}_d(\mathbb{R})$ and $f = \bar{f}_{|S^{d-1}}$ on the sphere S^{d-1}. We know that

$$\mathrm{grad} f(x) = Ax - (x^T Ax)x.$$

To get the Riemannian Hessian of f, we first differentiate a smooth extension of grad$f$ in $\mathbb{R}^d$, then project it to the tangent spaces of S^{d-1}.
Set

$$\bar{G}(x) = Ax - (x^T Ax)x.$$

We have

$$D\bar{G}(x)[u] = Au - (u^T Ax + x^T Au)x - (x^T Ax)u.$$

Hence

$$\begin{aligned} \mathrm{Hess} f(x)[u] &= \mathrm{Proj}_x(D\bar{G}(x)[u]) \\ &= \mathrm{Proj}_x(Au) - (x^T Ax)u \\ &= Au - (x^T Au)x - (x^T Ax)u. \end{aligned}$$

2.2.5 Covariant derivative

We are interested in differentiating vector fields on curves.
In order to obtain the second-order Taylor expansion of $g = f \circ c$, where $f : \mathcal{M} \to \mathbb{R}$ and $c : I \to \mathcal{M}$ (I being an open interval), we have to differentiate g', whose expression was

$$g'(t) = \langle \mathrm{grad} f(c(t)), c'(t)\rangle_{c(t)}.$$

Since $(\mathrm{grad} f) \circ c$ and c' are **not** vectors fields on $\mathcal{M}$, we cannot use the Riemannian connection directly, but instead an induced notion of differentiation of vector fields along a curve.

Definition 2.2.7 (Smooth vector field on a curve)**.** Let $c : I \to \mathcal{M}$ be a smooth curve on $\mathcal{M}$. Let $Z : I \to T\mathcal{M}$. If for all $t \in I,\ Z(t) \in T_{c(t)}\mathcal{M}$, we say that Z is a *vector field on the curve* c. It is a smooth vector field on c if it is smooth as a map $I \to T\mathcal{M}$. Let us denote by $\mathcal{X}(c)$ the set of smooth vector fields on c.

The following result states the existence of a new derivative operator on the set of smooth vector fields on a curve c.
Let $c : I \to \mathcal{M}$ be a smooth curve on a manifold endowed with a connection ∇.

Theorem 2.2.2 (Covariant derivative)**.** *With the above notations, there exists a unique operator called the* covariant derivative (induced by ∇) $\frac{D}{dt} : \mathcal{X}(c) \to \mathcal{X}(c)$ *such that for all* $Y, Z \in \mathcal{X}(c),\ U \in \mathcal{X}(\mathcal{M}),\ g \in \mathbb{F}(I),\ a, b \in \mathbb{R}$*:*

1. $\mathbb{R}$***-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:*** *for all* $t \in I$*,*
$$\left(\frac{D}{dt}(U \circ c)\right)(t) = \nabla_{c'(t)}U$$
4. *If* $\mathcal{M}$ *is a Riemannian manifold and* ∇ *is the Riemannian connection, the following* ***product rule*** *holds:*
$$\frac{d}{dt}\langle Y, Z\rangle = \langle\frac{D}{dt}Y, Z\rangle + \langle Y, \frac{D}{dt}Z\rangle.$$

Remark 2.2.6.

- Here $\langle Y, Z\rangle \in \mathbb{F}(I)$ is defined as
$$\langle Y, Z\rangle(t) = \langle Y(t), Z(t)\rangle_{c(t)}.$$
- $\frac{D}{dt}$ needs to be introduced since there exists some vector fields $Z \in \mathcal{X}(c)$ that cannot be written as $U \circ c$ with $U \in \mathcal{X}(\mathcal{M})$.

Example 2.2.8. For $f : \mathcal{M} \to \mathbb{R}$ smooth function on a Riemannian manifold with Riemannian concection ∇, we have
$$\mathrm{Hess} f(x)[u] = \nabla_u \mathrm{grad} f = \frac{D}{dt}\mathrm{grad} f(c(t))_{|t=0}$$
where $c : I \to \mathcal{M}$ is any smooth curve such that $c(0) = x$ and $c'(0) = u$.

In the case of an embedded submanifold of a Euclidean space $\mathcal{E}$, the covariant derivative is explicit.

Proposition 2.2.7. *Let* $\mathcal{M}$ *be an embedded submanifold of a Euclidean space* $\mathcal{E}$ *with connection*
$$\nabla_u V = Proj_x(D\bar{V}(x)[u]).$$
The operator $\frac{D}{dt}$ *defined as*
$$\frac{D}{dt}Z(t) = Proj_{c(t)}\left(\frac{d}{dt}Z(t)\right)$$
is the covariant derivative.
If $\mathcal{M}$ *us a Riemannian submanifold of* $\mathcal{E}$*,* $\frac{D}{dt}$ *also verifies the product rule.*

Example 2.2.9. Let $f : \mathcal{M} \to \mathbb{R}$ where $\mathcal{M}$ is a Riemannian submanifold of a Euclidean space $\mathcal{E}$. We have

$$\begin{aligned} \mathrm{Hess} f(x)[u] &= \mathrm{Proj}_x \left(\lim_{t \to 0} \frac{\mathrm{grad} f(c(t)) - \mathrm{grad} f(c(0))}{t} \right) \\ &= \lim_{t \to 0} \frac{\mathrm{Proj}_x(\mathrm{grad} f(c(t))) - \mathrm{grad} f(x)}{t} \end{aligned}$$

which leads to a finite difference approximation of the Hessian.

2.2.6 Geodesics

Having endowed $\mathcal{M}$ with the covariant derivative $\frac{D}{dt}$, we can define *acceleration* along a curve.

Definition 2.2.8 (Acceleration)**.** Let $c : I \to \mathcal{M}$ be a smooth curve. The vector field $c' \in \mathcal{X}(c)$ is called *velocity*. The smooth vector field $c'' \in \mathcal{X}(c)$ defined by

$$c'' := \frac{D}{dt} c'$$

is called the (intrinsic) *acceleration* of c.

Remark 2.2.7. If $\mathcal{M}$ is embedded in a linear space $\mathcal{E}$, we write

$$\ddot{c} = \frac{d^2}{dt^2} c$$

the classical (extrinsic) acceleration.
When $\mathcal{M}$ is a Riemannian submanifold of $\mathcal{E}$, we have

$$c''(t) = \mathrm{Proj}_{c(t)}(\ddot{c}(t)).$$

Example 2.2.10. On the sphere S^{d-1}, for $x \in S^{d-1}$ and $v \in T_x S^{d-1} \setminus \{0\}$, let us set

$$c(t) := \cos(t\|v\|)x + \frac{\sin(t\|v\|)}{\|v\|} v.$$

It can be shown that

$$\ddot{c}(t) = -\|v\|^2 c(t)$$

which, by projecting yields

$$c''(t) = \mathrm{Proj}_{c(t)} \ddot{c}(t) = (\mathrm{I}_d - c(t)c(t)^T)\ddot{c}(t) = 0.$$

Curves with acceleration equals to zero lead to the notion of *geodesic*, being the generalization of straight lines on manifolds.

Definition 2.2.9 (Geodesic)**.** Let $\mathcal{M}$ be a Riemannian manifold. A *geodesic* is a smooth curve $c : I \to \mathcal{M}$ such that for all $t \in I$,

$$c''(t) = 0.$$

Remark 2.2.8. On a Riemannian submanifold, a geodesic is a curve such that its (extrinsic) acceleration $\ddot{c}$ is normal to $\mathcal{M}$ at every point.

Example 2.2.11. Geodesics on $SO(n)$ Let $c : \mathbb{R} \to SO(n),\ t \mapsto X\exp(t\Omega)$. The curve c defines a geodesic on $SO(d)$ such that $c(0) = X$ and $c'(0) = V := X\Omega$.
Recall here that since $SO(n)$ is considered here as a Riemannian submanifold, the covariant derivative along c is $\frac{D}{dt} = \mathrm{Proj}_{c(t)}\left(\frac{d}{dt}\right)$ and here

$$\mathrm{Proj}_Z(U) = \frac{1}{2}Z(Z^TU - U^TZ), \quad (Z,U) \in TSO(n).$$

First, c is a smooth curve on $SO(n)$. Second, we need to show that $c'' = 0$.

$$\begin{aligned}\frac{D}{dt}(c'(t)) &= \mathrm{Proj}_{c(t)}\left(\frac{d}{dt}c'(t)\right)\\ &= \mathrm{Proj}_{c(t)}(c(t)\Omega^2)\\ &= \frac{c(t)}{2}(c(t)^T(c(t)\Omega^2) - (c(t)\Omega^2)^Tc(t))\\ &= \frac{c(t)}{2}(\Omega^2 - \Omega^2) = 0\end{aligned}$$

In fact, one can show that c is the only geodesic on $SO(n)$ defined on $\mathbb{R}$ with $c(0) = X$ and $c'(0) = V$.

2.2.7 Riemannian distance

Definition 2.2.10 (Length of a curve)**.** Let $\mathcal{M}$ be a Riemannian manifold and $c : [a,b] \to \mathcal{M}$ a piecewise smooth curve segment.
The length of c is given by

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

Having a notion of length can lead to defining a distance on the manifold $\mathcal{M}$ between two points x and y.

Definition 2.2.11 (Riemannian distance)**.** The function $\mathrm{dist} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ given by

$$\mathrm{dist}(x,y) := \inf_c L(c)$$

where the infimum is taken over all piecewise regular curve segments on $\mathcal{M}$ connecting x to y defines a distance if $\mathcal{M}$ is connected. Such distance is called the *Riemannian distance*.

Remark 2.2.9. Endowed with this distance, $\mathcal{M}$ is a metric space and the topology induced by it coincides with the topology given by its atlas.
If the infimum is attained for some curve segment c, such curve is called a *minimizing curve*, which coincide with geodesics.

Example 2.2.12. Let $\mathcal{M} = \mathcal{M}_1 \times \ldots \mathcal{M}_n$ a Riemannian product manifold. The Riemannian distance on $\mathcal{M}$ is given by

$$\mathrm{dist}_{\mathcal{M}}(x,y) = \left(\sum_{i=1}^n \mathrm{dist}_{\mathcal{M}_i}(x_i,y_i)^2\right).$$

Definition 2.2.12 (Completeness)**.**

- A connected Riemannian manifold is *metrically complete* if it is complete as a memtric space equipped with the Riemannian distance.
- A Riemmanian manifold is *geodesically complete* if every geodesic can be extended to a geodesic defined on the whole real line.

The following theorem makes the link between the seemingly different notions of completeness mentioned above.

Theorem 2.2.3 (Hopf–Rinow)**.** *Let $\mathcal{M}$ be a* connected *Riemannian manifold.*

(i) Metric completeness is equivalent to geodesic completeness.

(ii) $\mathcal{M}$ is complete if and only if its compact subsets are exactly its closed and bounded subsets.

One of the advantages of being a complete manifold is the possibility to connect any two points by a geodesic.

Theorem 2.2.4. *If $\mathcal{M}$ is a complete Riemannian manifold, then any two points x, y belong to the same connected component are connected by a minimizing geodesic $c : [0,1] \to \mathcal{M}$ such that $c(0) = x,\ c(1) = y$ and the distance between x and y is attained: $dist(x,y) = L(c)$.*

2.2.8 Exponential and logarithmic map

On a Riemannian manifold, it can bee shown that for every couple $(x, v) \in T\mathcal{M}$, there exists a unique (maximal) geodesic $\gamma_v : I \to \mathcal{M}$ such that

$$\gamma_v(0) = x \quad \text{and} \quad \gamma_v'(0) = v.$$

The Exponential map at a given point x and tangent vector v generalize the concept of "$x + v$" while remaining on the manifold.

Definition 2.2.13 (Exponential map)**.** Let

$$\mathcal{O} = \{(x, v) \in T\mathcal{M} : \gamma_v \text{ defined on } I \supseteq [0,1]\}.$$

The *exponential map* $\mathrm{Exp} : \mathcal{O} \to \mathcal{M}$ is defined by

$$\mathrm{Exp}(x, v) = \mathrm{Exp}_x(v) = \gamma_v(1).$$

The restriction Exp_x is defined on $\mathcal{O}_x = \{v \in T_x\mathcal{M} : (x, v) \in \mathcal{O}\}$.

The domain $\mathcal{O}$ contains all tangent space origins:

$$\{(x, 0) \in T\mathcal{M} : x \in \mathcal{M}\} \subset \mathcal{O}\}.$$

Example 2.2.13.

- In a Euclidean space, $\mathrm{Exp}_x(v) = x + v$.
- On S^{d-1}, let $\gamma : \mathbb{R} \to S^{d-1}$ such that $\gamma(0) = x$ and $\dot\gamma(0) = v \neq 0$ and

$$\gamma(t) = \cos(t\|v\|)x + \sin(t\|v\|)\frac{v}{\|v\|}.$$

  For all $t \in \mathbb{R},\ \gamma^\perp\gamma(t) = 1$ and γ has been shown to be a geodesic.
  Hence $\mathrm{Exp}_x : T_xS^{d-1} \setminus \{0\} \to S^{d-1}$ is $\mathrm{Exp}_x(\xi) = \cos(\|v\|)x + \sin(\|v\|)\frac{v}{\|v\|}$, which can be smoothly extended at $v = 0$ using $\sin(x)/x \to 1$ when $x \to 0$.

- **Exponential map on** $SO(n)$: Recall that $SO(n)$ is an embedded submanifold of $\mathbb{R}^{n\times n}$ of dimension $\frac{n(n-1)}{2}$ with tangent spaces

$$T_X SO(n) = \{V \in \mathbb{R}^{n\times n} \mid X^T V + V^T X\}\ , \quad X \in SO(n).$$

Consider the smooth matrix exponential $\exp : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ defined as

$$\exp(A) = \sum_{k=0}^{+\infty} \frac{A^k}{k!}, \quad \text{with } A^0 = \mathrm{I}_n.$$

Let $\Omega \in \mathbb{R}^{n\times n}$ be skew-symmetric, i.e. $\Omega = -\Omega^T$.

(i) $\exp(\Omega) \in SO(n)$: indeed, we have

$$\exp(\Omega)^T \exp(\Omega) = \exp(-\Omega)\exp(\Omega) = \mathrm{I}_n.$$

Moreover, $\det(\exp(\cdot))$ is constant on $\mathrm{Skew}(n)$ since $\mathrm{Skew}(n)$ is star-shaped and $\det(\exp(0) = 1$ hence $\det(\exp(\Omega)) = 1$ for any $\Omega \in \mathrm{Skew}(n)$.

(ii) $\frac{d}{dt}[\exp(t\Omega)]_{t=0} = \Omega$ since one can show that

$$\left\| \frac{\exp(t\Omega) - \exp(0)}{t} - \Omega \right\| \leq |t| \times \|\Omega\|^2 \exp(|t| \times \|\Omega\|) \to_{\to 0} 0.$$

(iii) Let $(X, V) \in TSO(n)$ and $R_X(V) = X\exp(X^T V)$. It defines a retraction. Indeed, $R_X(V) \in SO(n)$ since $X^T V$ is skew symmetric, $\exp(X^T V) \in SO(n)$ hence $X\exp(X^T V) \in SO(n)$. Besides, $R : (X, V) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \mapsto X\exp(X^T V) \in \mathbb{R}^{n\times n}$ is smooth. We have $R_X(0) = X$ and

$$\frac{d}{dt}[R_X(tV)]_{t=0} = X\frac{d}{dt}[\exp(tX^T V)]_{t=0} = XX^T V = V.$$

It can be shown that R_X is not injective for $X \in SO(n)$.

Since $\gamma_{tv}(1) = \gamma_v(t)$ for $t \in \mathbb{R}$, we have

$$\mathrm{Exp}_x(tv) = \gamma_v(t).$$

Proposition 2.2.8. *The exponential map Exp : $\mathcal{O} \to \mathcal{M}$ is smooth on $\mathcal{O}$, the latter being an open set in $T\mathcal{M}$ and which contains all tangent spaces origins.*

Remark 2.2.10. With the same notations, $\mathcal{O}$ is neighborhood of the zero-section of $T\mathcal{M}$, i.e.

$$\{(x, 0) \in T\mathcal{M} \mid x \in \mathcal{M}\} \subset \mathcal{O}.$$

Proposition 2.2.9. *The exponential map Exp : $\mathcal{O} \to \mathcal{M}$ is a second-order retraction.*

Proof. Smoothness is assumed from the previous Proposition. We know that for every $(x, v) \in T\mathcal{M}$, the following curve $c : t \mapsto \mathrm{Exp}_x(tv) = \gamma_v(t)$ satisfies $c(0) = x$ and $c'(0) = v$, which proves Exp is a retraction. Since $\forall t,\ \gamma_v''(t) = 0$, we have $c''(0) = 0$ hence Exp is second-order retraction. □

Note that any retraction at a point x is locally a diffeomorphism around the origin in the tangent space at x (indeed, it is smooth and its differential at x taken at 0 is the identity).

Example 2.2.14. Let $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$ be a Riemannian product manifold, with $\mathcal{O}_1, \mathcal{O}_2$ being the domains of the exponential maps on $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. The domain of the exponential map on $\mathcal{M}$ is $\mathcal{O}_1 \times \mathcal{O}_2$ and for all $(x_1, v_1) \in \mathcal{O}_1$ and $(x_2, v_2) \in \mathcal{O}_2$, we have

$$\mathrm{Exp}_x(v) = (\mathrm{Exp}_{x_1}(v_1), \mathrm{Exp}_{x_2}(v_2)) \quad \text{where} \quad x = (x_1, x_2) \text{ and } v = (v_1, v_2).$$

Definition 2.2.14 (Injectivity radius)**.** The *injectivity radius* of $\mathcal{M}$ at x, denoted by $\mathrm{inj}(x)$, is defined as the supremum of $r > 0$ such that Exp_x is defined and is a diffeomorphism on the open ball $B(x, r) := \{v \in T_x\mathcal{M} \mid \|v\|_x < r\}$.

Note that by the inverse function theorem, $\mathrm{inj}(x) > 0$.
Let $U := B(x, \mathrm{inj}(x)) \subseteq T_x\mathcal{M}$. Its image by Exp_x, denoted by $\mathcal{U}$, is a neighborhood of x in $\mathcal{M}$. Then, $\mathrm{Exp}_x : U \to \mathcal{U}$ is a diffeomorphism, hence posses a well-defined smooth inverse Exp_x^{-1}. In particular, $v := \mathrm{Exp}_x^{-1}(y)$ is the unique shortest (tangent) vector at x verifying $\mathrm{Exp}_x(v) = y$ and it motivates the following definition of the *logarithmic map* at x.

Definition 2.2.15 (Logarithmic map)**.** Let $x \in \mathcal{M}$.

$$\mathrm{Log}_x(y) := \arg\min_{v \in \mathcal{O}_x} \|v\|_x \quad \text{subject to } \mathrm{Exp}_x(v) = y$$

with domain such that it is uniquely defined.

Remark 2.2.11. For two points x and y, the logarithmic map generalizes the concept of "$y - x$.

Let us state some useful relations between the exponential map, the logarithmic map and the Riemannian distance.

Proposition 2.2.10. *If $\|v\|_x < inj(x)$, the geodesic $c : t \in [0, 1] \mapsto Exp_x(tv)$ is the minimizing curve connecting x to $y = Exp_x(v)$ (unique up to parametrization) and we have*

$$dist(x, y) = \|v\|_x \quad \text{and} \quad Log_x(y) = v.$$

Example 2.2.15. In a Euclidean space $\mathcal{E}$, $\mathrm{Log}_x(y) = y - x$ for all $x, y \in \mathcal{E}$.

Example 2.2.16. Let $x, y \in S^{n-1}$ such that $y \neq \pm x$.
On the sphere S^{n-1}, the exponential map is $\mathrm{Exp}_x(v) = \cos(\|v\|)x + \frac{\sin(\|v\|)}{\|v\|}v =: y$.
Let us find an expression for the inverse $\mathrm{Exp}_x^{-1}(y)$, i.e. we are looking for $v \in T_xS^{d-1}$ such that $\mathrm{Exp}_x(v) = y$. Given that $x^\perp x = 1$ and $\mathrm{x} \perp v = 0$, we have $x^\perp y = \cos(\|v\|)$, thus

$$y = (x^T y)x + \sin(\|v\|)\frac{v}{\|v\|}.$$

The orthogonal projection of y onto T_xS^{d-1} is

$$P_x(y) = (\mathrm{I}_x - xx^T)y = \sin(\|v\|)\frac{v}{\|v\|}.$$

Normalizing the projection:

$$\frac{P_x(y)}{\|P_x(y)\|} = \mathrm{sign}(\sin(\|v\|))\frac{v}{\|v\|}.$$

Now, let us restrict the domain of Exp_x to v such that $\|v\| < \pi$ leads to

$$x^T y = \cos(\|v\|) \iff \|v\| = \arccos(x^T y)$$

which implies

$$v = \arccos(x^T y)\frac{P_x(y)}{\|P_x(y)\|}.$$

Since v is unique, we have $\log_x(y) = v$ for $y \neq \pm x$ and $\log_x(y) = 0$ for $y = x$. For $y = -x$ (antipodal point of x, there is an infinite number of v such that $\mathrm{Exp}_x(v) = y$ and no logarithmic map for $y = -x$.
One can check that $\mathrm{Exp}_x : B(x, \pi) \to S^{d-1} \setminus \{-x\}$ is a diffeomorphism. Moreover,

$$d_{S^{d-1}}(x, y) = \|\log_x(y)\| = \arccos(x^T y).$$

2.2.9 Parallel transport

Let $\mathcal{M}$ equipped with a connection ∇. We wish to move a velocity $u \in T_x\mathcal{M}$ to $T_y\mathcal{M}$ along a curve c such that $c(0) = x$ and $c(1) = y$.

Let $c : I \to \mathcal{M}$ be a smooth curve such that $c(0) = x$ and $c(1) = y$. Let $Z \in \mathcal{X}(c)$ a smooth vector field on c with $Z(0) = u$.

Definition 2.2.16 (Parallel vector field)**.** With the same notation, if $\frac{D}{dt}Z = 0$, then Z is called a *parallel vector field.*

Let $\mathcal{M}$ be a manifold with a connection ∇ and covariant derivative $\frac{D}{dt}$.

Theorem 2.2.5. *For any smooth curve $c : I \to \mathcal{M}$, $t_0 \in I$ and $u \in T_{c(t_0)}\mathcal{M}$, there exists a unique parallel vector field $Z \in \mathcal{X}(\mathcal{M})$ such that $Z(t_0) = u$.*

Definition 2.2.17 (Parallel transport of tangent vectors)**.** Let c be a smooth curve on $\mathcal{M}$. The *parallel transport of tangent vectors* at $c(t_0)$ to the tangent space at $c(t_1)$ along c is the map

$$PT^c_{t_1 \leftarrow t_0} : T_{c(t_0)}\mathcal{M} \to T_{c(t_1)}\mathcal{M}$$

such that

$$PT^c_{t_1 \leftarrow t_0}(u) = Z(t_1)$$

where $Z \in \mathcal{X}(c)$ is the unique parallel vector field such that $Z(t_0) = u$.

Remark 2.2.12. Parallel transport from x to y depends on the choice of curve connecting x and y, which is chosen as the geodesic between x and y.

The parallel transport operator $PT^c_{t_1 \leftarrow t_0}$ enjoys the following properties.

Proposition 2.2.11. *Let $PT^c_{t_1 \leftarrow t_0}$ be the parallel transport operator of tangent vectors at $c(t_0)$ to the tangent space at $c(t_1)$ along the curve c.*

- *$PT^c_{t_1 \leftarrow t_0}$ is linear;*
- *$PT^c_{t_2 \leftarrow t_1} \circ PT^c_{t_1 \leftarrow t_0} = PT^c_{t_2 \leftarrow t_0}$ and $PT^c_{t \leftarrow t} = Id$;*
- *If $\mathcal{M}$ is a Riemannian manifold and ∇ is compatible with the Riemannian metric, then parallel transport is an isometry:*

$$\forall u, v \in T_{c(t_0)}\mathcal{M},\ \langle u, v\rangle_{c(t_0)} = \langle PT^c_{t_1 \leftarrow t_0}(u), PT^c_{t_1 \leftarrow t_0}(v)\rangle_{c(t_1)}.$$

Example 2.2.17. On the sphere S^{d-1}, the parallel transport of $v \in T_x\mathcal{M}$ along the geodesic $c : t \mapsto \mathrm{Exp}_x(tv)$ such that $c(0) = x,\ c(1) = 1$ and $\dot{c}(0) = v$ is

$$PT^c_{t\leftarrow 0}(u) = \left(\mathrm{I}_n + (\cos(t\|v\|) - 1)\frac{vv^T}{\|v\|^2} - \sin(t\|v\|)\frac{xv^T}{\|v\|}\right) u.$$

One application of parallel transport is Riemann conjugate gradient algorithm. Recall that the standard Riemannian gradient descent algorithm provides iterates of the form

$$x^{(k+1)} = R_{x^{(k)}}(\alpha s^{(k)})$$

where α is a step-size and $s^{(k)} = -\mathrm{grad}h(x^{(k)})$ is the Riemannian gradient of the function $h : \mathcal{M} \to \mathbb{R}$ to optimize evaluated at $x^{(k)}$. This procedure can be slow to converge and the *Riemannian conjugate gradient* adds some inertia:

$$s^{(k)} = -\mathrm{grad}h(x^{(k)}) + \beta PT_{x^{(k-1)}\leftarrow x^{(k)}}(s^{(k-1)}),\quad \beta > 0.$$

Example 2.2.18. Transporters on $SO(n)$ Recall that $SO(n) = \{X \in \mathbb{R}^{n\times n} \mid X^TX = \mathrm{I}_n\ ,\ \det(X) = 1\}$ is a Riemannian submanifold of $\mathbb{R}^{n\times n}$ and

$$T_XSO(n) = \{X\Omega \mid \Omega \in \mathbb{R}^{n\times n}, \Omega + \Omega^T = 0\}.$$

Define T as follows: for $X, Y \in SO(n)$ and $\Omega + \Omega^T = 0$,

$$T_{Y\leftarrow X}(X\Omega) = Y\Omega.$$

- T defines a transporter: let $X, Y \in SO(n)$ and $U \in T_XSO(n)$. There exists $\Omega \in \mathbb{R}^{n\times n}$ skew-symmetric such that $U = X\Omega \iff X^TU = \Omega$. Thus

  $$T_{Y\leftarrow X}(U) = T_{Y\leftarrow X}(X\Omega) = Y\Omega = YX^TU.$$

  It shows that $T_{Y\leftarrow X}$ is linear hence a transporter.

- $T_{Y\leftarrow X} : T_XSO(n) \to T_YSO(n)$ is an isometry: let $U_1 = X\Omega_1 \in T_XSO(n)$ and $U_2 = X\Omega_2 \in T_XSO(n)$. We have

  $$\begin{aligned}\langle T_{Y\leftarrow X}U_1, T_{Y\leftarrow X}U_2\rangle &= \langle T_{Y\leftarrow X}X\Omega_1, T_{Y\leftarrow X}X\Omega_2\rangle\\ &= \langle Y\Omega_1, Y\Omega_2\rangle\\ &= \mathrm{Tr}(\Omega_1^TY^TY\Omega_2) = \mathrm{Tr}(\Omega_1^T\Omega_2)\\ &= \mathrm{Tr}(\Omega_1^TX^TX\Omega_2)\\ &= \langle U_1, U_2\rangle.\end{aligned}$$

It can be shown that all geodesics of $SO(n)$ are of the form $c(t) = X\exp(t\Omega),\ c(0) = X,\ c'(0) = V = X\Omega$. Let $c : \mathbb{R} \to SO(n)$ be such a geodesic and $X = c(t_0), Y = c(t_1)$ for $t_1 \geq t_0 \geq 0$. Let us show that $T_{Y\leftarrow X}$ is not equal to parallel transport along c from t_0 to t_1.
Let $U = X\tilde{\Omega} \in T_XSO(n)$. We have $T_{c(0)\leftarrow X}(U) = U$ and $t \mapsto \frac{D}{dt}T_{c(t)\leftarrow X}(U)$ is smooth. Besides,

$T_{c(t_1)\leftarrow X}(U) = PT^c_{t_1\leftarrow t_0}(U)$ if and only if $\frac{D}{dt}T_{c(t)\leftarrow X}(U) = 0$. We have

$$\begin{aligned}
\frac{D}{dt}T_{c(t)\leftarrow X}(U) &= \mathrm{Proj}_{c(t)}\left(\frac{d}{dt}(c(t)X^TU)\right) \\
&= \mathrm{Proj}_{c(t)}(c'(t)X^TU) \\
&= \frac{c(t)}{2}(c(t)^T(c'(t)X^TU) - (c'(t)X^TU)^Tc(t)) \\
&= \frac{c(t)}{2}(\Omega X^TU - U^TX\Omega^T) \\
&= \frac{c(t)}{2}(\Omega\tilde{\Omega} - \tilde{\Omega}^T\Omega^T)
\end{aligned}$$

Hence

$$\frac{D}{dt}T_{c(t)\leftarrow X}(U) = 0 \iff (\tilde{\Omega}\Omega)^T = \tilde{\Omega}\Omega$$

which is not true in general.

Chapter 3

Optimization algorithms

In the following, let $f : \mathcal{M} \to \mathbb{R}$ be the (smooth) function of interest (objective function) defined on a Riemannian manifold $\mathcal{M}$. We want to solve

$$\arg\min_{x \in \mathcal{M}} f(x).$$

Minimizers may not exist, and if so, uniqueness is not guaranteed in general. Usually, we rather aim at a local minimizer.

3.1 First-order optimization

3.1.1 First-order Taylor expansion on curves

Let $c : I \subset \mathbb{R} \to \mathcal{M}$ a smooth curve on a Riemannian manifold $\mathcal{M}$ such that $c(0) = x$ and $c'(0) = v$ with I an open interval of $\mathbb{R}$ around 0. Let

$$g : I \to \mathbb{R} : t \mapsto g(t) := f \circ c(t)$$

The map g is a smooth map and admits a first-order Taylor expansion around 0 as follows:

$$g(t) = g(0) + tg'(0) + O(t^2) \iff f(c(t)) = f(x) + t\langle \mathrm{grad} f(x), v \rangle_x + O(t^2)$$

applying the chain rule to get $g'(t) = Df(c(t))[c'(t)] = \langle \mathrm{grad} f(c(t)), c'(t) \rangle_{c(t)}$, then evaluated at 0.
Let specify the curve c as a retraction $c(t) = R_x(tv)$, so that

$$f(R_x(tv)) = f(x) + t\langle \mathrm{grad} f(x), v \rangle_x + O(t^2).$$

Considering the variable $s = tv \in T_x\mathcal{M}$ yields alternatively

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

Remark 3.1.1. The smooth function $f \circ R_x : T\mathcal{M} \to \mathbb{R}$ is called the *pullback* of f by R_x to the tangent space $T_x\mathcal{M}$. Notice that $f \circ R$ is a map between two linear spaces.

3.1.2 Optimality conditions

How can we check if a point $x \in \mathcal{M}$ is a local minimizer for $f : \mathcal{M} \to \mathbb{R}$? Let us first state *necessary conditions*.

Definition 3.1.1 (Critical point)**.** A point $x \in \mathcal{M}$ is *critical* for $f : \mathcal{M} \to \mathbb{R}$ if for all smooth curves c on $\mathcal{M}$ satisfying $c(0) = x$, we have

$$(f \circ c)'(0) \geq 0. \tag{3.1}$$

Remark 3.1.2. Notice that condition (3.1) is equivalent to $(f \circ c)'(0) = 0$. Indeed, one can consider $t \mapsto c(t)$ and $t \mapsto c(-t)$.
Moreover, thanks to the chain rule on $f \circ c$, x is critical for $f \iff Df(x) = 0$.

Proposition 3.1.1. *If x is a local minimizer or maximizer of $f : \mathcal{M} \to \mathbb{R}$, then x is a critical point for f.*

On Riemannian manifolds, we have the following equivalence involving the nullity of the Riemannian gradient at a critical point.

Proposition 3.1.2. *A point $x \in \mathcal{M}$ is critical for f if and only if $\mathrm{grad} f(x) = 0$.*

Proof. Let $c : I \to \mathcal{M}$ a smooth curve such that $c(0) = x$ and $c'(0) = v$.
Recall that

$$(f \circ c)'(0) = \langle \mathrm{grad} f(x), v \rangle_x.$$

If $\mathrm{grad} f(x) = 0$, then x is critical. If x is critical, considering both v and $-v$ yields the wanted result. □

3.2 Riemannian gradient descent

Recall the standard procedure for gradient descent in a Euclidean space $\mathcal{E}$. Let x_0 be our initial point and $\alpha_k > 0$ some step-sizes.

$$x_0 \in \mathcal{E} \;\; ; \;\; x_{k+1} = x_k - \alpha_k \mathrm{grad} f(x_k) \quad k \geq 1, \; \alpha_k > 0.$$

Example 3.2.1. Let $\mathcal{E} = \mathbb{R}^d$ endowed with its Euclidean inner product $\langle u, v \rangle_{\mathcal{E}} = u^T v$.
Let $G : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ be a smooth map such that $G(x) \in \mathrm{Sym}(d)^{++}$ for all $x \in \mathbb{R}^d$. On $\mathcal{M} = \mathbb{R}^d$, we define the metric

$$\langle u, v \rangle_{\mathcal{M}} = u^T G(x) v.$$

It defines a Riemannian metric on $\mathcal{M}$. Recall that the Riemannian gradient of f is

$$\mathrm{grad}_{\mathcal{M}} f(x) = G(x)^{-1} \mathrm{grad}_{\mathcal{E}} f(x).$$

Let us consider a retraction $R_x(u) = x + u$ on $\mathbb{R}^d$. Riemannian gradient descent is given by

$$x_{k+1} = x_k - \eta_k G(x)^{-1} \mathrm{grad}_{\mathcal{E}} f(x).$$

Note that for $G(x) = \mathrm{Hess} f(x)$ and $\eta_k = 1$, this RGD is no other than Newton's method.

3.2.1 Algorithm and convergence

For *Riemannian gradient descent* on $\mathcal{M}$, we first choose a retraction R on $\mathcal{M}$, an initial point $x_0 \in \mathcal{M}$ and iterate for $k \geq 1$,

$$x_{k+1} = R_{x_k}(s_k) \text{ with } s_k = -\alpha_k \text{grad} f(x_k)).$$

How to choose the step-size α_k? Let

$$g(t) = f\left(R_{x_k}(-t\text{grad} f(x_k))\right).$$

The goal is to minimize g without too great a computational cost. Three stategies can be used:

Fixed step-size $\alpha_k = \alpha$ for all k

Optimal step-size α_k minimizes $g(t)$ exactly (often costly)

Backtracking start with a guess $t_0 > 0$ and at each iteration, reduce it by a factor $\tau \in (0, 1)$ to that $t_i = \tau t_{i-1}$ until "acceptable", then set $\alpha_k = t_i$.

Let us examine the convergence of this algorithms under the following assumptions on the cost function L:

A.1 There exists $f_{\text{low}} \in \mathbb{R}$ such that for all $x \in \mathcal{M}$, $f(x) \geq f_{\text{low}}$.

A.2 There exists a constant $c > 0$ such that, for all $k \geq 1$,

$$f(x_k) - f(x_{k+1}) \geq c\|\text{grad} f(x_k)\|^2.$$

Notice there are no conditions on the initialization $x_0 \in \mathcal{M}$.

Proposition 3.2.1. *With the previous notation, under **A.1** and **A.2**, we have*

$$\lim_{k \to +\infty} \|gradf(x_k)\| = 0.$$

Moreover, for all $K \geq 1$, there exists $k \in \{0, \ldots, K-1\}$ such that

$$\|gradf(x_k)\| \leq \sqrt{\frac{f(x_0) - f_{low}}{c}} \frac{1}{\sqrt{K}}$$

Proof. For $K \geq 1$, we have the telescoping

$$\begin{aligned} f(x_0) - f_{\text{low}} &\geq f(x_0) - f(x_K) \\ &= \sum_{k=0}^{K-1} f(x_k) - f(x_{k+1}) \\ &\geq Kc \min\left\{\|\text{grad} f(x_k)\|^2 \mid k = 0, \ldots, K-1\right\} \end{aligned}$$

By **A.2**, $f(x_{k+1}) \leq f(x_k)$ for all k. Taking $K \to \infty$ gives

$$f(x_0) - f_{\text{low}} \geq \sum_{k=0}^{\infty} f(x_k) - f(x_{k+1}).$$

Such a convergence implies that

$$0 \lim_{k\to\infty} f(x_k) - f(x_{k+1}) \geq c \lim_{k\to\infty} \|\mathrm{grad} f(x_k)\|^2.$$

If x is an accumulation point of $\{x_k\}_{k\geq 0}$, since the norm is a continuous function, we get

$$\|\mathrm{grad} f(x)\| = 0.$$

□

Remark 3.2.1. Assuming there exists at least one accumulation point for the cost function f, the previous property means that gradient descent converges to critical points.

Motivated by standard Taylor expansion of real functions, let us explore regularity assumptions to help guarantee sufficient decrease.

3.2.2 Regularity assumptions and complexity

Recall that, since $x_{k+1} = R_{x_k}(-\alpha_k \mathrm{grad} f(x_k))$, first-order Taylor expansion gives

$$\begin{aligned} f(x_{k+1}) &= f(R_{x_k}(-\alpha_k \mathrm{grad} f(x_k)) \\ &= f(x_k) + \langle \mathrm{grad} f(x_k), s_k \rangle + O(\|s_k^2\|). \end{aligned}$$

Assume the following Lipschitz-type assumption:

A.3 For a fixed $S \subset T\mathcal{M}$, there exists $L > 0$ such that, for all $(x, s) \in S$, we have

$$f(R_x(s)) \leq f(x) + \langle \mathrm{grad} f(x), s \rangle + \frac{L}{2}\|s\|^2.$$

Remark 3.2.2. For f a smooth function on a Euclidean space $\mathcal{E}$ equipped with the canonical retraction $R_x : s \mapsto x + s$, assuming **A.3** on $T\mathcal{E} = \mathcal{E} \times \mathcal{E}$ means that for all $x, s \in \mathcal{E}$, we have

$$f(x + s) \leq f(x) + \langle \mathrm{grad} f(x), s \rangle + \frac{L}{2}\|s\|^2.$$

It is well-known in Euclidean optimization that Lipschitz continuity of the gradient is enough to ensure such inequality.

The following result shows that under **A.3**, there exists a range of learning rates leading to sufficient decrease in the sense of assumption **A.2**.

Proposition 3.2.2. *With the previous notations, if the pairs $\{(x_k, -\alpha_k gradf(x_k))\}_{k\geq 0}$ generated by the Riemannian gradient descent algorithm are such that $\alpha_k \in [\alpha_{\min}, \alpha_{\max}] \subset (0, 2/L)$ and $(x_k, -\alpha_k gradf(x_k)) \in S$ for all $k \geq 0$, then assumption **A.2** holds with*

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

Proof. It relies on the following inequality obtained by applying the assumption **A.3**:

$$f(x_k) - f(x_{k+1}) \geq \left(\alpha_k - \frac{L}{2}\alpha_k^2\right) \|\mathrm{grad} f(x_k)\|^2$$

and the study of the quadratic expression in α_k. □

When L is known, we have the following corollary.

Corollary 3.2.1. *Let f be a smooth function verifying **A.1**. For a retraction R, assume $f \circ R$ verify **A.3** on a subset $S \subseteq T\mathcal{M}$ with constant L.*
Let $\{(x_k, s_k)\}_{k\geq 0}$ be the paris generated by the Riemannian gradient descent algorithm with constant step-size $\alpha_k = \frac{1}{L}$. *If for all $k \geq 0$, $(x_k, s_k) \in S$, then*

$$\lim_{k\to\infty} \|gradf(x_k)\| = 0.$$

Besides, for $K \geq 1$, there exists $k \in \{0, \dots, K-1\}$ such that

$$\|gradf(x_k)\| \leq \sqrt{2L(f(x_0) - f_{low}}\frac{1}{\sqrt{K}}.$$

3.2.3 Backtracking line-search

In practice, an appropriate value for L is rarely known, therefore the following *backtracking line-search* strategy is implemented to pick the step-sizes α_k in an adaptive way: given an initial step-size $\bar{\alpha}$, the procedure iteratively reduces the new candidate step-size by $\tau \in (0, 1)$ until the following condition, called the *Armijo–Goldstein* condition is ensured

$$f(x) - f(R_x(-\alpha \mathrm{grad} f(x))) \geq r\alpha \|\mathrm{grad} f(x)\|^2,$$

for some constant $r \in (0, 1)$.

Algorithm 1 Backtracking Line Search on a Manifold

Require: $\tau, r \in (0, 1)$, $x \in \mathcal{M}$, $\bar{\alpha} > 0$
1: $\alpha \leftarrow \bar{\alpha}$
2: **while** $f(x) - f(R_x(-\alpha \mathrm{grad} f(x))) < r\alpha \|\mathrm{grad} f(x)\|^2$ **do**
3: $\quad \alpha \leftarrow \tau\alpha$
4: **end while**
5: **return** α

Under **A.3**, backtracking line-search ensures **A.2**, without the need to know explicitly the regularity constant L to run the algorithm.

Lemma 3.2.1. *Let f be a smooth function on a Riemannian manifold $\mathcal{M}$. For a retraction R, a point $x \in \mathcal{M}$ and an initial step-size $\bar{\alpha} > 0$, assume **A.3** holds for $f \circ R$.*

Corollary 3.2.2.

3.2.4 RIemannian conjugate gradient

Recall the iterative process of Riemannian gradient descent:

$$x_{k+1} = R_{x_k}(\alpha_k d_k), \quad \text{where} \quad d_k = -\nabla f(x_k), \ \alpha_k > 0.$$

We want here to incorporate some inertia in the procedure. In a standard Euclidean setting, one would write

$$d_k = -\mathrm{grad} f(x_k) + \beta_{k-1} d_{k-1}.$$

Here on the manifold, $d_{k-1} \in T_{x_{k-1}}\mathcal{M}$ while $\text{grad}f(x_k) \in T_{x_k}\mathcal{M}$, hence cannot be combined directly since they do not belong to the same linear subspace. One way to counter this is to transport d_{k-1} to x_k using transport:

$$\tilde{d}_{k-1} := \mathcal{T}_{x_k \leftarrow x_{k-1}}(d_{k-1}) \in T_{x_k}\mathcal{M}$$

which gives the updates search direction

$$d_k = -\text{grad}f(x_k) + \beta_{k-1}\tilde{d}_{k-1}.$$

Algorithm 2 Riemannian Conjugate Gradient

Require: Initialization: $x_0 \in \mathcal{M}$, $d_0 = -g_0 = -\operatorname{grad} h(x_0)$
Ensure: $x_k \in \mathcal{M}$
1: **for** $k = 0$ to convergence **do**
2: **if** $\langle g_k, d_k \rangle_{x_k} \geq 0$ **then**
3: $d_k = -g_k$,
4: **end if**
5: $\alpha_k = \text{Linesearch}(x_k, d_k)$
6: $x_{k+1} = R_{x_k}(\alpha_k d_k)$
7: $g_{k+1} = \operatorname{grad} h(x^{(k+1)})$
8: $\tilde{d}_k = T_{x^{(k)}, x^{(k+1)}}(d_k)$
9: $\tilde{g}_k = T_{x^{(k)}, x^{(k+1)}}(g_k)$
10: $\beta = \max\left(0, \frac{\langle g_{k+1} - \tilde{g}_k, g_{k+1} \rangle_{x_{k+1}}}{\langle g_{k+1} - \tilde{g}_k, \tilde{d}_k \rangle_{x_{k+1}}}\right)$
11: $d_{k+1} = -g_{k+1} + \beta \tilde{d}_k$
12: **end for**

Remark 3.2.3. As parallel transport is not always available in closed form or may be expensive to compute, vector transport is a cheaper generalization that works well in practice.
The inertia parameter β is computed using the *Hestenes–Stiefel* rule but other rules could be used (See e.g. [5]).

3.3 Second-order Taylor expansion on curves and retractions

Let $f : \mathcal{M} \to \mathbb{R}$ a smooth function and $c : I \to \mathcal{M}$ a smooth curve such that $c(0) = x$ and $c'(0) = v$. Set as before the smooth function $g = f \circ c : I \subseteq \mathbb{R} \to \mathbb{R}$. Second-order Taylor expansion is

$$g(t) = g(0) + tg'(0) + tg'(0) + \frac{t^2}{2}g''(0) + O(t^3).$$

We have

$$g'(t) = Df(c(t))[c'(t)] = \langle \text{grad}f(c(t)), c'(t) \rangle_{c(t)}$$

so that

$$g'(0) = \langle \text{grad}f(x), v \rangle_x.$$

Now let us turn to $g''(t)$. Using covariant derivative property and Hessian definition, we obtain

$$\begin{aligned} g''(t) &= \frac{d}{dt}\langle \mathrm{grad} f(c(t)), c'(t)\rangle_{c(t)} \\ &= \left\langle \frac{D}{dt}(\mathrm{grad} f \circ c)(t), c'(t) \right\rangle_{c(t)} + \left\langle \mathrm{grad} f(c(t)), \frac{D}{dt}c'(t) \right\rangle_{c(t)} \\ &= \left\langle \nabla_{c'(t)} \mathrm{grad} f, c'(t)\right\rangle_{c(t)} + \langle \mathrm{grad} f(c(t)), c''(t)\rangle_{c(t)} \\ &= \langle \mathrm{Hess} f(c(t))[c'(t)], c'(t)\rangle_{c(t)} + \langle \mathrm{grad} f(c(t)), c''(t)\rangle_{c(t)} \end{aligned}$$

Having $t = 0$ gives

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

The previous computations yield

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

We can deduce the following result.

Lemma 3.3.1. *Let $c(t)$ be a geodesic connecting $x = c(0)$ to $y = c(1)$. If there exists $\mu \in \mathbb{R}$ such that for all $t \in [0,1]$, $Hess f(c(t)) \succ \mu Id$, then*

$$f(y) \geq f(x) + \langle grad f(x), v\rangle_x + \frac{\mu}{2}\|v\|_x^2.$$

Proof. Under the assumption, the mean value theorem and the fact that $\|c'(t)\|_{c(t)}$ is constant for c being a geodesic makes it possible to show that there exists $t \in (0,1)$ such that

$$f(y) = f(x) + \langle \mathrm{grad} f(x), c'(0)\rangle_x + \frac{1}{2}\langle \mathrm{Hess} f(c(t))[c'(t)], c'(t)\rangle_{c(t)} + \frac{1}{2}\langle \mathrm{grad} f(c(t)), c''(t)\rangle_{c(t)}.$$

□

Let $x \in \mathcal{M}$ and $v \in T_x\mathcal{M}$. Let us now consider the case where

$$c(t) = R_x(tv).$$

The previous second-order Taylor formula motivates the following definition.

Definition 3.3.1. Let R be a retraction. R is a *second-order retraction* on $\mathcal{M}$ if for all $x \in \mathcal{M}$, $v \in T_x\mathcal{M}$, the curve $c(t) = R_x(tv)$ verifies

$$c''(0) = 0.$$

Example 3.3.1. On the sphere S^{d-1}, the following retraction is second-order:

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

Indeed, set $c(t) = R_x(tv) = \frac{x+tv}{\sqrt{1+t^2\|v\|^2}}$. We can see that

$$\ddot{c}(t) = -\|v\|^2(x+3tv) + O(t^2)$$

so that $\ddot{c}(0) = -\|v\|^2 x$. Hence

$$c''(0) = \mathrm{Proj}_x(\ddot{c}(0)) = 0.$$

Combining second-order Taylor expansion with second-order retraction give the following important result.

Proposition 3.3.1. *Let R be a retraction on a Riemannian manifold $\mathcal{M}$ and $f : \mathcal{M} \to \mathbb{R}$ a smooth function.*

- *If x is a critical point of f:*

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

- *If R is a second-order retraction, then for all $x \in \mathcal{M}$:*

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

In both cases,

$$Hessf(x) = Hess(f \circ R_x)(0).$$

3.4 Lipschitz conditions and Taylor expansions

We present in this section Lipschitz continuity definitions on Riemannian manifolds.

Definition 3.4.1 (Lipschitz continuous function)**.** A function $f : \mathcal{M} \to \mathbb{R}$ is L-Lipschitz continuous if (and only if)

$$\forall (x, s) \in \mathcal{O},\ |f(\mathrm{Exp}_x(s)) - f(x)| \leq L\|s\|$$

where $\mathcal{O} \subset T\mathcal{M}$ is the domain of the exponential map.

Remark 3.4.1. It can be proved that this definition is equivalent to having

$$\forall x, y \in \mathcal{M},\ |f(x) - f(y)| \leq L\mathrm{dist}(x, y)$$

with dist the Riemannian distance on $\mathcal{M}$.

Should f possess a continuous gradient, Lipschitz continuity of f amounts to upper-boundedness of the gradient.

Proposition 3.4.1. *Let $f : \mathcal{M} \to \mathbb{R}$ be a function having a continuous gradient. The following assertions are equivalent:*

1. *f is L-Lipschitz continuous*
2. *$\forall x \in \mathcal{M},\ \|gradf(x)\| \leq L$.*

Proof. Let $(x, s) \in \mathcal{O}$ and $c : t \mapsto \mathrm{Exp}_x(ts)$ for $t \in [0, 1]$. We have

$$f(c(1)) - f(c(0)) = \int_0^1 \langle \mathrm{grad} f(c(t)), c'(t)\rangle dt$$

hence if assertion 2. is verified,

$$|f(\mathrm{Exp}_x(s)) - f(x)| \leq L\int_0^1 \|c'(t)\|dt = L \times L(c) = L\|s\|.$$

For the other implication, let $x \in \mathcal{M}$ and assume that the domain of Exp_x is open around the origin. Then

$$\begin{aligned}\|\mathrm{grad}f(x)\| &= \max_{s\in T_x\mathcal{M},\ \|s\|=1} \langle \mathrm{grad}f(x), s\rangle \\ &= \max_{s\in T_x\mathcal{M},\ \|s\|=1} Df(x)[s] \\ &= \max_{s\in T_x\mathcal{M},\ \|s\|=1} \lim_{t\to 0^+} \frac{f(\mathrm{Exp}_x(ts)) - f(x)}{t} \leq L \text{ using assertion 1.}\end{aligned}$$

□

Now let us define Lipschitz continuity of the *Riemannian gradient* of a function f. In order to give meaning to a Riemannian version of "$\mathrm{grad}f(x) - \mathrm{grad}f(y)$", we will resort to parallel transport. Let us give general definitions, having in mind $V = \mathrm{grad}f$.

Definition 3.4.2 (Lipschitz continuity for vector field)**.** A vector field V on $\mathcal{M}$ is L-Lipschitz continuous if and only if

$$\forall (x,s) \in \mathcal{O},\quad \|P_s^{-1}V(\mathrm{Exp}_x(s)) - V(x)\| \leq L\|s\|,$$

where $\mathcal{O} \subseteq T\mathcal{M}$ is the domain of Exp and P_s is the parallel transport along $\gamma(t) = \mathrm{Exp}_x(ts)$ from $t = 0$ to $t = 1$.

Remark 3.4.2. As before, this definition is equivalent to the more intuitive one:

$$\forall x, y \in \mathcal{M}, \mathrm{dist}(x,y) < \mathrm{inj}(x),\quad \|PT_{0\leftarrow 1}^{\gamma}V(y) - V(x)\| \leq L\mathrm{dist}(x,y)$$

with $\gamma : [0,1] \to \mathcal{M}$ being the unique minimizing geodesic connecting x et y.

In the same fashion as before, Lipschitz continuity of a vector field amounts to boundedness of its covariant derivative.

Proposition 3.4.2. *Let V be a continuously differentiable vector field on $\mathcal{M}$. The following assertions are equivalent:*

1. *V is L-Lipschitz continuous*
2. *$\forall (x,s) \in T\mathcal{M},\ \|\nabla_s V\| \leq L\|s\|$*

where ∇ is the Riemannian connection.
When either 1. or 2. holds true, for any smooth curve $c : [0,1] \to \mathcal{M}$ connecting x to y, we have

$$\|PT_{0\leftarrow 1}^{c}V(y) - V(x)\| \leq L \times L(c).$$

In particular for the Hessian, this yields the following corollary.

Corollary 3.4.1. *Let $f : \mathcal{M} \to \mathbb{R}$ a twice continuously differentiable function. The following assertions are equivalent:*

1. *$gradf$ is L-Lipschitz continuous*
2. *$\forall x \in \mathcal{M},\ \|Hessf(x)\| = \max_{s\in T_x\mathcal{M}, \|s\|=1} \|Hessf(x)[s]\| \leq L.$*

3.5 Geodesic convexity

The idea here is to adapt standard (vector space) convexity settings to Riemannian manifolds using geodesic curves instead of line segments.

Definition 3.5.1 (Subset geodesically convex)**.** Let $S \subseteq \mathcal{M}$ be a subset of a Riemannian manifold $\mathcal{M}$. The subset S is *geodesically convex* if, for every $x, y \in S$, there exists a geodesic segment $c : [0; 1] \to \mathcal{M}$ such that

$$c(0) = x,\ c(1) = y \text{ and for all} t \in [0, 1],\ c(t) \in S.$$

Remark 3.5.1. In the definition, the last line can be restated as saying that c connects x to y in S.
Beware, since S can or cannot be a manifold, c is a geodesic for $\mathcal{M}$.
The definition states that in a geodesically convex set S, any pair of points are connected in S by *at least* one geodesic segment.

Example 3.5.1.

- The empty set and singletons are geodesically-convex.
- If $\mathcal{M} = \mathcal{E}$, there is an equivalence between convexity (in the usual sense) and geodesical convexity.
- Any connected and complete Riemannian manifold is geodesically convex, in particular, it is the case of $SO(n), \mathrm{St}(n, p)$ for $p < n$ and $\mathbb{R}_+$ equipped with $\langle u, v\rangle := \frac{uv}{x^2}$.

Let us now define the concept of geodesically convex *function*.

Definition 3.5.2 (Function geodesically convex)**.** A function $f : S \to \mathbb{R}$ is *geodesically convex* if:

(i) S is geodesically convex;

(ii) $f \circ c : [0, 1] \to \mathbb{R}$ is convex for each geodesic segment $c : [0, 1] \to \mathcal{M}$ such that $c([0, 1]) \subseteq \mathcal{M}$ with $c(0) \neq c(1)$.

Remark 3.5.2.

- Let us restate the definition more explicitly: $f : S \to \mathbb{R}$ is geodesically convex if for all $x, y \in S$ and all geodesics c connecting x to y in S,

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

  We can also define strict convexity if we require that whenever $x \neq y$, the previous inequality is strict.

- In an analogous manner,

  (i) $f : S \to \mathbb{R}$ is *geodesically concave* if $-f$ is geodesically convex;

  (ii) $f : S \to \mathbb{R}$ is *geodesically linear* if its is both geodesically convex and concave;

  (iii) $f : S \to \mathbb{R}$ is *geodesically μ-strongly convex* for some $\mu > 0$ if S is geodesically convex and for each geodesic segment $c : [0, 1] \to \mathcal{M}$ such that $c([0, 1]) \subseteq \mathcal{M}$, we have

  $$f(c(t)) \leq (1 - t)f(c(0)) + tf(c(1)) - \frac{t(1 - t)}{2}\mu L(c)^2,$$

  where $L(c) = \|c'(0)\|_{c(0)}$ is the length of the geodesic segment.

As in linear space convexity, geodesic convexity ensures local minimizers are indeed global and strict convexity ensures uniqueness.

Theorem 3.5.1. *Let $f : S \to \mathbb{R}$.*

(i) If f is geodesically convex, then any local minimizer is a global minimizer.

(ii) If f is geodesically strictly convex, then f admits at most one global minimizer.

Proof.

(i) By contradiction, assume $x \in S$ is a local minimizer that is not a global minimizer. Then, there exists $y \in S$ such that $f(y) < f(x)$. There exists c a geodesic in S such that $c(0) = x$ and $c(1) = y$ and

$$\forall\ 0 < t \leq 1,\ f(c(t)) \leq (1-t)f(x) + tf(y) = f(x) + t(f(y) - f(x)) < f(x)$$

which provides the wanted contradiction.

(ii) We have to prove uniqueness. By contradiction, assume that there exists x and y such that $f(x) = f(y) =: f_*$. We also know there exists a geodesic c connecting x to y in S such that

$$\forall t \in (0,1),\ f(c(t)) < (1-t)f(x) + tf(y) = f_*.$$

The last inequality contradicts global optimality of x and y.

□

Proposition 3.5.1. *Let $\{f_i, i \in I\}$ be an arbitrary family of geodesically convex functions $f_i : S \to \mathbb{R}$. Let $\alpha_i \in R$ for $i \in I$. Set*

$$S_i := \{x \in S \mid f_i(x) \leq \alpha_i\}.$$

Then $\tilde{S} := \bigcap_{i \in I} S_i$ is geodesically convex.
Moreover, the sublevel sets of a geodesically convex function f are geodesically convex and the set of global minimizers of f is geodesically convex.

Remark 3.5.3. Let S be a geodesically convex set on $\mathcal{M}$ and $f, f_1, \ldots, f_m$ geodesically convex functions on S and $g_1, \ldots, g_p$ geodesically linear functions on S. Let $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$, $\beta_1, \ldots, \beta_p \in \mathbb{R}$. Consider

$$\min_{x \in S} f(x) \quad \text{subject to } f_i(x) \leq \alpha_i \text{ for } i = 1, \ldots, m \text{ and } g_j(x) = \beta_j \text{ for } j = 1, \ldots, p.$$

It defines a geodesically convex program.

With additional assumptions of differentiability on the function $f : \mathcal{M} \to \mathbb{R}$, we can characterize geodesically convex functions through gradient/Hessian properties. Let us first give first-order conditions.

Theorem 3.5.2. *Let S be a geodesically convex set on $\mathcal{M}$ and $f : \mathcal{M} \to \mathbb{R}$ be a differentiable function in a neighborhood of S.*

1. *$f_{|S} : S \to \mathbb{R}$ is geodesically convex if and only if for all geodesic segments $c : [0,1] \to \mathcal{M}$ contained in S (with $x = c(0)$), we have*

$$\forall t \in [0,1],\quad f(c(t)) \geq f(x) + t\langle grad f(x), c'(0)\rangle_x.$$

2. $f_{|S} : S \to \mathbb{R}$ *is geodesically μ-strongly convex for some $\mu > 0$ if and only if for all geodesic segments $c : [0,1] \to \mathcal{M}$ contained in S (with $x = c(0)$), we have*

$$\forall t \in [0,1], \quad f(c(t)) \geq f(x) + t\langle gradf(x), c'(0)\rangle_x + t^2 \frac{\mu}{2} L(c)^2.$$

3. $f_{|S} : S \to \mathbb{R}$ *is geodesically strictly convex if and only if for all geodesic segments $c : [0,1] \to \mathcal{M}$ contained in S (with $x = c(0)$), whenever $c'(0) \neq 0$, we have*

$$\forall t \in [0,1], \quad f(c(t)) > f(x) + t\langle gradf(x), c'(0)\rangle_x.$$

Proof. It relies on the fact that $f_{|S}$ is geodesically (strictly) convex if and only if for all $x, y \in S$ and all geodesics c connecting x to y in S, the function $f \circ c : [0,1] \to \mathbb{R}$ is (strictly) convex; and therefore the use of the following formula

$$\forall s, t \in [0,1], \quad f(c(t)) \geq f(c(s)) + (t-s)(f \circ c)'(s)$$

where $(f \circ c)'(s) = \langle \mathrm{grad} f(c(s)), c'(s)\rangle_{c(s)}$. □

We deduce from this theorem the following importance corollary.

Corollary 3.5.1. *Let f be a differentiable and geodesically convex function on an open geodesically convex set. The following assertions are equivalent:*

(i) x is a global minimizer of f

(ii) $gradf(x) = 0$

Sketch of the proof. If $\mathrm{grad} f(x) = 0$, from the previous Theorem, $f(x) \leq f(y)$ for all y in the domain of f.
The implication 1. ⇒ 2. follows from the fact that the domain of f is open and Proposition. □

Let us now give second order conditions of geodesic convexity.

Theorem 3.5.3. *Let $f : S \to \mathbb{R}$ be a twice differentiable function on an open geodesically convex set S.*

1. *f is geodesically convex if and only if for all $x \in S$, $Hess(fx) \succeq 0$;*
2. *f is geodesically μ-strongly convex if and only if $x \in S$, $Hess(fx) \succeq \mu Id$;*
3. *f is geodesically strictly convex if $x \in S$, $Hess(fx) \succ 0$:*

Sketch of the proof. We rely on the fact that f is geodesically convex if and only if $f \circ c : [0,1] \to \mathbb{R}$ is convex for all geodesic segments $c : [0,1] \to \mathcal{M}$ whose image is in S. This sufficient and necessary condition gives

$$\forall t \in (0,1), \quad (f \circ c)''(t) \geq 0.$$

Since $c''(t) = 0$ (geodesic property), we have $(f \circ c)''(t) = \langle \mathrm{Hess} f(c(t))[c'(t)], c'(t)\rangle_{c(t)}$. This latter inequality is crucial for the discussion to prove the theorem. □

Let S be a geodesically convex set and let $f : \mathcal{M} \to \mathbb{R}$ be a differentiable function in a neighborhood of S. Assume $f_{|S}$ is geodesically convex. Thanks to Theorem 3.5.2, for $x \in S$ and $v \in T_x\mathcal{M}$ such that $c(t) = \mathrm{Exp}_x(tv) \in S$ for all $t \in [0,1]$, we have the following

$$\forall t \in [0,1], \quad f(\mathrm{Exp}_x(tv)) \geq f(x) + t\langle \mathrm{grad} f(x), v\rangle_x.$$

- If f is geodesically strictly convex, the inequality above is strict for $f \in (0,1]$.
- If f is geodesically μ-strictly convex, we have

$$\forall t \in [0,1], \quad f(\mathrm{Exp}_x(tv)) \geq f(x) + t\langle \mathrm{grad} f(x), v\rangle_x + t^2 \frac{\mu}{2}\|v\|_x^2. \tag{3.2}$$

Recall that if the gradient of f is L-lipschitz continuous, we have

$$\forall t \in [0,1], \quad f(\mathrm{Exp}_x(tv)) \leq f(x) + t\langle \mathrm{grad} f(x), v\rangle_x + t^2 \frac{L}{2}\|v\|_x^2. \tag{3.3}$$

In the following, S is a non-empty, closed and geodesically convex set in a complete manifold $\mathcal{M}$ and $f : \mathcal{M} \to \mathbb{R}$ is differentiable in a neighborhood of S. The following results states that

(i) strong convexity ensures existence and uniqueness of a minimizer.

(i) the norm of the gradient of a geodesically strongly convex function at some point gives some information about the optimality gap at this point.

Lemma 3.5.1. *With the same notations, assume $f_{|S}$ is geodesically μ-strongly convex, for some $\mu > 0$. Then,*

1. *the sublevels sets of $f_{|S}$ are compact and $f_{|S}$ has exactly one global minimizer;*
2. *f satisfies a* Polyak–Lojasiewicz *inequality:*

$$\forall x \in S, \ f(x) - f(x_\star) \leq \frac{1}{2\mu}\|grad f(x)\|_x^2$$

with $x_\star$ denoting the minimizer of $f_{|S}$.

Sketch of the proof. 1. For some arbitrary $x_0 \in S$, we show that $S_0 := \{x \in S : f(x) \leq f(x_0)\}$ is compact since it is closed and bounded in $\mathcal{M}$, which is a complete manifold. The boundedness is proved by contradiction using (3.2). Then, using that $f_{|S_0}$ is continuous, it attains its minimum $x_\star \in S_0$, which, by examining whether any $x \in S$ is such that $x \in S_0$ or $x \notin S_0$, an be shown to minimize $f_{|S}$. The conclusion is drawn thanks to the fact that geodesic strong convexity implies geodesic strict convexity and Theorem 3.5.1.

2. By the previous point, the minimizer $x_\star \in S$ of $f_{|S}$ exists and is unique. Let $x \in S$. There exists (by geodesic convexity of S) $v_x \in T_x\mathcal{M}$ such that $x_\star = \mathrm{Exp}_x(v_x)$ and $c(t) = \mathrm{Exp}_x(tv) \in S$ for all $t \in [0,1]$. From (3.2), we have

$$\begin{aligned} f(x_\star) = f(\mathrm{Exp}_x(v_x)) &\geq f(x) + \langle \mathrm{grad} f(x), v_x\rangle_x + \frac{\mu}{2}\|v_x\|_x^2 \\ &\geq \inf_{v \in T_x\mathcal{M}} \left\{ f(x) + \langle \mathrm{grad} f(x), v\rangle_x + \frac{\mu}{2}\|v\|_x^2 \right\} \\ &= f(x) - \frac{1}{2\mu}\|\mathrm{grad} f(x)\|_x^2 \end{aligned}$$

as the infinum is attained at the critical point of the quadratic form in v, namely $v = -\frac{1}{\mu}\mathrm{grad} f(x)$.

□

The previous lemma allows to study the convergence of a simple version of Riemannian gradient descent applied to a function being geodesically strongly convex with a Lipschitz continuous gradient.

Let $f : \mathcal{M} \to \mathbb{R}$ be a differentiable and geodesically convex function on a complete manifold $\mathcal{M}$. Let $x_0 \in \mathcal{M}$ and consider $S_0 = \{x \in \mathcal{M} \mid f(x) \leq f(x_0)\}$.
Assume f has L-Lipschitz continuous gradient on a neighborhood of S_0 and $f_{|S_0}$ is geodesically μ-strongly convex with $\mu > 0$. We consider gradient descent, initialized at x_0, with exponential retraction and constant step-size $\frac{1}{L}$:

$$x_{k+1} = \mathrm{Exp}_{x_k}\left(-\frac{1}{L}\mathrm{grad} f(x_k)\right), \quad k \geq 0.$$

Theorem 3.5.4. *With the same notation, f admits a unique minimizer $x_\star$ and $\{x_k : k \geq 0\}$ converge to $x_\star$ at least linearly. Specifically, if $\kappa = L/\mu \geq 1$, for all $k \geq 0$, we have $x_k \in S_0$ and*

$$f(x_k) - f(x_\star) \leq \left(1 - \frac{1}{\kappa}\right)^k (f(x_0) - f(x_\star)) \quad \textit{and} \quad \textit{dist}(x_k, x_\star) \leq \sqrt{1 - \frac{1}{\kappa}}^k \sqrt{k}\,\textit{dist}(x_0, x_\star).$$

Proof. □

Chapter 4

Examples of embedded submanifolds

Let us describe some of the most frequently occurring examples of embedded submanifolds.

4.1 Euclidean (sub)spaces

Let $\mathcal{M}$ be a linear subspace of a Euclidean space $(\mathcal{E}, \langle \cdot, \cdot \rangle)$. It can be viewed as a Riemannian manifold whose dimension is its dimension as a linear space and whose tangent spaces are the same:

$$\forall x \in \mathcal{M},\ T_x\mathcal{M} = \mathcal{M}.$$

The orthogonal projector from $\mathcal{E}$ to $T_x\mathcal{M}$ does not depend on x (if $\mathcal{M} = \mathcal{E}$, the projector is the identity) and given a smooth function $f : \mathcal{M} \to \mathbb{R}$ with smooth extension $\bar{f} : U \to \mathbb{R}$ (with U a neighborhood of $\mathcal{M}$ in $\mathcal{E}$), we have

$$\mathrm{grad} f(x) = \mathrm{Proj}_{\mathcal{M}}(\mathrm{grad} \bar{f}(x)).$$

Example 4.1.1. Take $\mathcal{E} = \mathbb{R}^{n \times n}$ and $\mathcal{M} = \mathrm{Sym}(n)$, defining a Riemannian submanifold. We have $\mathrm{Proj}_{\mathcal{M}}(Z) = \frac{1}{2}(Z + Z^T)$.

On a Euclidean space $\mathcal{E}$, covariant derivatives $\frac{D}{dt}$ and Riemannian connection ∇ are usual vector field derivatives, the Riemannian Hessian coincides with its Euclidean Hessian and the retraction $R_x(v) = x + v$ is a second-order retraction that is also the exponential map. The Hessian of f is

$$\forall x, v \in \mathcal{M},\ \mathrm{Hess} f(x)[v] = \mathrm{Proj}_{\mathcal{M}}(\mathrm{Hess} \bar{f}(x)[v]).$$

Remark 4.1.1. A linear space can be endowed with a non-Euclidean Riemannian metric (i.e. that varies from point to point) in the following way: let $\mathcal{M} = \mathbb{R}^n$ and

$$\langle u, v \rangle_x = u^T G(x) v, \quad G(x) \in \mathrm{Sym}(n)$$

such that $G(x) \succ 0$ varies smoothly with x.

4.2 Stiefel manifold

Let $p \leq n$ and let us endow $\mathbb{R}^{n\times}$ with the standard inner product

$$\langle U, V\rangle = \mathrm{Tr}(U^T V).$$

Definition 4.2.1 (Stiefel manifold)**.** The compact *Stiefel manifold* is the set of (rectangular) matrices in $\mathbb{R}^{n\times p}$ with orthonormal columns in $\mathbb{R}^n$ w.r.t. the standard $\mathbb{R}^n$ inner product $\langle u, v\rangle = u^T v$. In other words,

$$\mathrm{St}(n,p) = \{X \in \mathbb{R}^{n\times p} \mid X^T X = \mathrm{I}_p\}.$$

Remark 4.2.1. $\mathrm{St}(n,1) = S^{n-1}$. The elements of $\mathrm{St}(n,p)$ are called orthonormal matrices.

Proposition 4.2.1.

$$\dim St(n,p) = np - \frac{p(p+1)}{2}.$$

Proof. Let $h : \mathbb{R}^{n\times p} \to \mathrm{Sym}(p) : X \mapsto X^T X - \mathrm{I}_p$. We have $\mathrm{St}(n,p) = h^{-1}(\{0\})$ and h is differentiable at every $X \in \mathrm{St}(n,p)$ with

$$Dh(X)[V] = X^T V + V^T X \in \mathrm{Sym}(p)$$

Let us show that $Dh(X)$ is surjective, so that one can conclude that $\mathrm{rank} Dh(X) = \dim \mathrm{Sym}(p) = \frac{p(p+1)}{2}$. Let $A \in \mathrm{Sym}(p)$ and consider $V = \frac{1}{2}XA$. One can check easily that $Dh(X)[V] = A$ hence the wanted surjectivity. Finally,

$$\dim \mathrm{St}(n,p) = \dim \mathbb{R}^{n\times p} - \dim \mathrm{Sym}(p) = np - \frac{p(p+1)}{2}.$$

□

Let us now explicit the tangent spaces. Before doing so, set the following

$$\mathrm{Skew}(p) := \{\Omega \in \mathbb{R}^{p\times p} \mid \Omega^T = -\Omega\}.$$

Proposition 4.2.2. *Let $X \in St(n,p)$.*

$$T_X St(n,p) = \left\{X\Omega + X_\perp B \mid \Omega \in Skew(p),\ B \in \mathbb{R}^{(n-p)\times p}\right\}.$$

Proof. Let $X \in \mathrm{St}(n,p)$. By standard characterization with defining function, we know that

$$T_X \mathrm{St}(n,p) = \ker Dh(X) = \{V \in \mathbb{R}^{n\times p} \mid X^T V + V^T X = 0\}.$$

Let us complete the orthonormal basis formed by the columns of X by a matrix $X_\perp \in \mathbb{R}^{n\times(n-p)}$ such that $[X \mid X_\perp] \in \mathbb{R}^{n\times n}$ is orthogonal, i.e.

$$X^T X = \mathrm{I}_p \ \ , \ \ X_\perp^T X_\perp = \mathrm{I}_{n-p} \ \ , \ \ X^T X_\perp = 0.$$

By invertibility of $[X \mid X_\perp]$, for any $V \in \mathbb{R}^{n\times p}$, there exists a unique $(\Omega, B) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{(n-p)\times p}$ such that

$$V = [X \mid X_\perp] \begin{bmatrix} \Omega \\ B \end{bmatrix} = X\Omega + X_\perp B.$$

Using this decomposition, one can show that

$$V \in T_X \mathrm{St}(n,p) \iff 0 = Dh(X)[V] = \Omega + \Omega^T,$$

which yields that $\Omega \in \mathrm{Skew}(p)$. □

Example 4.2.1. $SO(n) = \{X \in \mathbb{R}^{n\times n} \mid X^TX = \mathrm{I}_n,\ \det(X) = 1\}$. It defines an embedded submanifold of $\mathcal{E} = \mathbb{R}^{n\times n}$ of dimension $\frac{n(n-1)}{2}$. Indeed,

$$SO(n) = \mathrm{St}(n,n) \cap \overset{-1}{(\det}(\{-1\}))^c.$$

Since $(\det^{-1}(\{-1\}))^c$ is open, $SO(n)$ is an open subset of $\mathrm{St}(n,n)$ hence an embedded submanifold with same dimension as $\mathrm{St}(n,n)$ and with same tangent spaces:

$$T_XSO(n) = \{V \in \mathbb{R}^{n\times n} \mid X^TV + V^TX = 0\}.$$

Let us now consider two popular retractions:

- The *Q-factor retraction* is defined as

$$R_X(V) = Q, \quad X \in \mathrm{St}(n,p),\ V \in T_X\mathrm{St}(n,p),$$

where $QR = X + V$ is a QR decomposition with $Q \in \mathrm{St}(n,p)$ and $R \in \mathbb{R}^{p\times p}$ being an upper triangular matrix with nonnegative diagonal coefficients. Since $X + V$ has full rank p (given that $(X+V)^T(X+V) = I_p + V^TV \succ 0$), the QR decomposition is unique, which gives the well-definition of R_X.
Moreover, $R_X(0) = X$ and R is smooth (resulting from a composition of smooth operations in the Gram-Schmidt algorithm). Besides, $DR_X(0) = \mathrm{Id}$.

- The *polar retraction* is defined as

$$R_X(V) = (X+V)\left((X+V)^T(X+V)\right)^{-1/2} = (X+V)(\mathrm{I}_p + V^TV)^{-1/2}.$$

One has $R_X(0) = X$ and R is smooth and $DR_X(0) = \mathrm{Id}$.

Example 4.2.2. Metric polar projection retraction for Stiefel $\mathbf{St}(n,p)$ For $(X,V) \in T\mathrm{St}(n,p)$, consider the thin SVD decomposition of $X + V$, i.e.

$$X + V = U\Sigma W^T,\ \ U \in \mathrm{St}(n,p)\ ,\ W \in O(p),$$

with $\Sigma \in \mathbb{R}^{p\times p}$ a diagonal matrix with positive entries.

- UW^T is the unique (metric) projection of $X + V$ to $\mathrm{St}(n,p)$, i.e.

$$UW^T = \arg\min_{Y\in\mathrm{St}(n,p)} \|X + V - Y\|^2.$$

Indeed, thanks to the unitary invariance of the Frobenius norm and the bijectivity of $Y \mapsto YW$ on $\mathrm{St}(n,p)$, we have

$$\begin{aligned}
\inf_{Y\in\mathrm{St}(n,p)} \|X+V-Y\|^2 &= \inf_{Y\in\mathrm{St}(n,p)} \|U\Sigma W^T - Y\|^2 = \inf_{Y\in\mathrm{St}(n,p)} \|U\Sigma - YW\|^2 \\
&= \inf_{Z\in\mathrm{St}(n,p)} \|U\Sigma - Z\|^2 = \inf_{Z\in\mathrm{St}(n,p)} \left(\sum_{i=1}^p \|\sigma_iu_i - z_i\|^2\right) \\
&= \inf_{Z\in\mathrm{St}(n,p)} \left(\sum_{i=1}^p \sigma_i^2 - 2\sigma_i\langle u_i, x_i\rangle + 1\right) \\
&\geq \sum_{i=1}^p (\sigma_i^2 - 2\sigma_i + 1) \ \text{ by Cauchy–Schwarz inequality.}
\end{aligned}$$

Equality holds if and only if $u_i = z_i$. The unique minimizer is given by $YW = U$, i.e. $Y = UW^T$.

- For $(X, V) \in T\mathrm{St}(n,p)$, define $R_X(V) = UW^T$. We have
$$R_X(V) = (X+V)(I_p + V^TV)^{-1/2}.$$
Indeed, since $V \in T_X\mathrm{St}(n,p)$, we have $X^TV + V^TX = 0$ and
$$\begin{aligned}(I+V^TV)^{-1/2} &= \left[(X+V)^T(X+V)\right]^{-1/2} = (W\Sigma U^TU\Sigma W^T)^{-1/2} \\ &= (W\Sigma^2W^T)^{-1/2} = W\Sigma^{-1}W^T.\end{aligned}$$
Hence $(X+V)(I+V^TV)^{-1/2} = U\Sigma W^TW\Sigma^{-1}W^T = UW^T$.
- Let us show that $R : T\mathrm{St}(n,p) \to \mathrm{St}(n,p)$ defined as above is a retraction. First, R is smooth since $(X,V) \mapsto (X+V)(I_p+V^TV)^{-1/2}$. Second, $R_X(0) = X$ and
$$\frac{d}{dt}[R_X(tV)]_{t=0} = V + X\frac{d}{dt}[(I_p + t^2V^TV)^{-1/2}]_{t=0} = 0.$$

One can show that $R_X : T_X\mathrm{St}(n,p) \to \mathrm{St}(n,p)$ is not surjective.

Proposition 4.2.3. *Let $Proj_X$ be the orthogonal projector to $T_X St(n,p)$. For $U \in \mathbb{R}^{n\times p}$, one has*
$$Proj_X(U) = (I - XX^T)U + X\frac{X^TU - U^TX}{2}.$$

Proof. One can show that
$$N_X\mathrm{St}(n,p) := (T_X\mathrm{St}(n,p))^\perp = \{XA \mid A \in \mathrm{Sym}(p)\}$$
hence, for $U \in \mathbb{R}^{n\times p}$, the orthogonal projector verifies $U - \mathrm{Proj}_X(U) = XA$ for some $A \in \mathrm{Sym}(p)$. Moreover, we have $\mathrm{Proj}_X(U) \in T_X\mathrm{St}(n,p)$, i.e.
$$\mathrm{Proj}_X(U)^TX + X^T\mathrm{Proj}_X(U) = 0.$$
Combining both yields $U^TX + X^TU = 2A$ therefore
$$\mathrm{Proj}_X(U) = U - X\frac{X^TU + U^TX}{2} = (I - XX^T)U + X\frac{X^TU - U^TX}{2}.$$
Moreover, the projection of $U \in \mathbb{R}^{p\times p}$ to the normal space is
$$\mathrm{Proj}_X^\perp(U) = Y\frac{1}{2}(X^TU + XU^T).$$
□

Turning $\mathrm{St}(n,p)$ into a Riemannian submanifold of $\mathbb{R}^{n\times p}$ enables to use the expression of the orthogonal projector established before and get
$$\mathrm{grad}f(X) = \mathrm{grad}\bar{f}(X) - X\mathrm{sym}(X^T\mathrm{grad}\bar{f}(X))$$
where $\mathrm{sym}(M) = \frac{M+M^T}{2}$ and $\bar{f}$ is a smooth extension of f. However, other metrics can be used, such as the so-called *canonical metric*. The Stiefel manifold becomes a Riemannian manifold by introducing an inner product on its tangent spaces. There are two natural inner products for tangent spaces of Stiefel manifolds: the Euclidean inner product and the canonical inner product. See [4].

4.2.1 Euclidean inner product

Let $U, V \in T_X\mathrm{St}(n,p)$. The Euclidean inner product on $T_X\mathrm{St}(n,p)$ is defined as

$$\langle U, V\rangle := \mathrm{tr}(U^T V).$$

Let $V = XA + X_\perp B$ with $A \in \mathbb{R}^{p\times p}$ is anti-symmetric and $B \in \mathbb{R}^{(n-p)\times p}$ arbitrary. We can show that

$$\langle V, V\rangle = \mathrm{tr}(A^T A) + \mathrm{tr}(B^T B) = \sum_{i>j} 2a_{ij}^2 + \sum_{i,j} b_{ij}^2.$$

The Euclidean metric weighs the A coordinates twice as much as the B coordinates. [Absil,example 5.4.2, proof section 2.2.2 EAS]

For U a smooth vector field and $V \in T_X\mathrm{St}(n,p)$, the Levi–Civita connection is given by

$$\nabla_V U = DU[V] - X\mathrm{Sym}(X^T DU[V]).$$

Let us compute the geodesics. Let $t \mapsto Y(t)$ a curve on $\mathrm{St}(n,p)$. Derivating twice the identity $Y^T Y = \mathrm{I}_p$ yields

$$Y^Y \ddot{Y} + 2\dot{Y}^T\dot{Y} + \ddot{Y}^T Y = 0.$$

Since we look for a geodesic, $\ddot{Y}(t) \in (T_{Y(t)}\mathrm{St}(n,p))^\perp$ hence

$$\ddot{Y} + Y(\dot{Y}^T\dot{Y}) = 0$$

Solving this differential equation yields

$$Y(t) = \begin{pmatrix} Y(0) & \dot{Y}(0) \end{pmatrix} \exp\left(t \begin{pmatrix} A & -S(0) \\ I & A \end{pmatrix}\right) I_{2p,p} e^{-At}$$

where $A(t) = A(0) = Y^T\dot{Y}$ and $S(t) = \dot{Y}^T(t)\dot{Y}(t)$.

Example 4.2.3. Riemannian Gradient Descent We consider $\mathrm{St}(n,p)$ as a Riemannian submanifold of $\mathbb{R}^{n\times p}$ endowed with the usual Euclidean inner product $\langle X, Y\rangle = \mathrm{Tr}(X^T Y)$. Consider

$$f : \mathrm{St}(n,p) \to \mathbb{R},\ f(X) = \mathrm{Tr}(X^T AX),\quad A \in \mathrm{Sym}_n(\mathbb{R}).$$

Recall that the orthogonal projector $\mathrm{Proj}_X : \mathbb{R}^{n\times p} \to T_X\mathrm{St}(n,p)$ is given by

$$\mathrm{Proj}_X(U) = U - X(X^T U + U^T X)/2.$$

Remark that the computation of $\mathrm{Proj}_X(U)$ is $O(np^2)$.
Let us now compute the gradient of f. Consider the smooth extension $\bar{f} : X \in \mathbb{R}^{n\times p} \mapsto \mathrm{Tr}(X^T AX)$. We have

$$\mathrm{grad}\bar{f}(X) = 2AX$$

hence by projection on the Riemannian submanifold

$$\mathrm{grad}f(X) = 2AX - 2X(X^T AX).$$

Example 4.2.4. Riemannian Hessian Let $f : \mathrm{St}(n,p) \to \mathbb{R}$ be a smooth function and $\bar{f}$ be a smooth extension of f. One can check (using the fact that if $S \in \mathrm{Sym}(p)$, $\mathrm{Proj}_X(XS) = 0$) that

$$\mathrm{Hess}f(X) = \mathrm{Proj}_X(\mathrm{Hess}\bar{f}(X)[U]) - \mathrm{Proj}_X(U\mathrm{sym}(X^T\mathrm{grad}\bar{f}(X))).$$

Let us apply this to the cost function $f(X) = \mathrm{Tr}(X^T AX)$, $A \in \mathrm{Sym}(n)$. Define $\bar{f}(X) = \mathrm{Tr}(X^T AX)$ for $X \in \mathbb{R}^{n\times p}$. We have

$$\mathrm{grad}\bar{f}(X) = 2AX\ ,\ \mathrm{Hess}\bar{f}(X)[U] = 2AU.$$

Therefore

$$\mathrm{Hess}f(X)[U] = 2\mathrm{Proj}_X(AU) - \mathrm{Proj}_X(UX^T AX).$$

4.2.2 Canonical inner product

Contrary to the Euclidean inner product, it weighs the coordinates equally. Let us detail its construction. For $V = XA + X_\perp B$, we have $A = X^T Z$, which yields $XA = XX^T Z$. Thus

$$(I - \frac{1}{2}XX^T)V = V - \frac{1}{2}XX^T V = XA + X_\perp B - \frac{1}{2}XA = \frac{1}{2}XA + X_\perp B.$$

We have

$$\mathrm{tr}(V^T(I - \frac{1}{2}XX^T)V) = \frac{1}{2}\mathrm{tr}(A^T A) + \mathrm{tr}(B^T B) = \sum_{i>j} a_{ij}^2 + \sum_{i,j} b_{ij}^2.$$

It justifies the following definition.

Definition 4.2.2. The canonical inner product on $T_X\mathrm{St}(n,p)$ is defined for $U, V \in T_X\mathrm{St}(n,p)$ as

$$g_c(U,V) := \mathrm{tr}(U^T(I - \frac{1}{2}XX^T)V).$$

Derived from the previous computations, we have

$$g_c(U,V) = \frac{1}{2}\mathrm{Tr}(A^T A) + \mathrm{Tr}(B^T B).$$

The following lemma shows that the normal space for the canonical metric coincides with the normal space for the embedded euclidean metric.

Lemma 4.2.1. *Let $X \in St(n,p)$. The normal space N_X defined as*

$$N_X = \{N \in \mathbb{R}^{n\times p} \mid \forall U \in T_X St(n,p), g_c(N,U) = 0\}$$

is given by

$$N_X = \{XS \mid S \in Sym\}.$$

Proof. Let $N \in N_X$ and $U \in T_X\mathrm{St}(n,p)$. Remark that $g_c(N,U) = \langle \mathcal{A}(N), U\rangle_{\mathrm{Frob}}$, where $\mathcal{A} : N \mapsto (I - \frac{1}{2}XX^T)N$. We already know that

$$(T_X\mathrm{St}(n,p))^{\perp_{\mathrm{Frob}}} = \{XS \mid S = S^T\}.$$

Here

$$\mathcal{A}(N) \in (T_X\mathrm{St}(n,p))^{\perp_{\mathrm{Frob}}} \iff \mathcal{A}(N) = XS \text{ for some symmetric matrix } S.$$

Let us solve $\mathcal{A}(N) = XS$. We use that $I - \frac{1}{2}XX^T$ is invertible and

$$(I - \frac{1}{2}XX^T)^{-1} = I + XX^T.$$

Hence

$$N = (I + XX^T)(XS) = 2XS.$$

In other words, $N \in \{XS \mid S = S^T\}$. □

Let us now compute the corresponding Riemannian gradient $\mathrm{grad} f(X)$ of any smooth function $f : \mathrm{St}(n,p) \to \mathbb{R}$ at $X \in \mathrm{St}(n,p)$.

Proposition 4.2.4. *Let $f : St(n,p) \to \mathbb{R}$ be a smooth function. Let $X \in St(n,p)$. The Riemannian gradient of f associated with the canonical metric is*

$$gradf(X) = \nabla f(X) - X[\nabla f(X)]^T X.$$

Proof. We have

$$\forall V \in T_X\mathrm{St}(n,p),\ \mathrm{grad}f(X)^T(I - \frac{1}{2}XX^TV = \underbrace{\nabla f(X)^T V}_{=df(X)[V]}.$$

Taking the trace:

$$\mathrm{tr}\left(\left[(I - \frac{1}{2}XX^T)\mathrm{grad}f(X) - \nabla f(X)\right]^T V\right) = 0$$

which means that $W := (I - \frac{1}{2}XX^T)\mathrm{grad}f(X) - \nabla f(X)$ is in $(T_X\mathrm{St}(n,p))^\perp$, i.e. of the form $W = XS$ for some symmetrix matrix S. From

$$(I - \frac{1}{2}XX^T)\mathrm{grad}f(X) = \nabla f(X) + XS$$

we deduce

$$\mathrm{grad}f(X) = \nabla f(X) + XX^T\nabla f(X) + 2XS. \tag{4.1}$$

But we also know that $\mathrm{grad}f(X)$ must be in $T_X\mathrm{St}(n,p)$, hence $X^T\mathrm{grad}f(X)$ is skew-symmetric:

$$X^T\mathrm{grad}f(X) + (X^T\mathrm{grad}f(X))^T = 0.$$

Using Equation (4.1), we obtain

$$X^T\mathrm{grad}f(X) = X^T\left[\nabla f(X) + XX^T\nabla f(X) + 2XS\right] = 2X^T\nabla f(X) + 2S.$$

This last expression is skew-symmetric if and only if and only if $S = -\mathrm{sym}(X^T\nabla f(X))$. Plugging the expression of S in the previous $\mathrm{grad}f(X)$ expression gives the result. □

The Levi–Civita connection is explicit.

Proposition 4.2.5.

$$\nabla_{X'}Y = \dot{Y} - \frac{1}{2}X(\dot{Y}^TX + X^T\dot{Y}).$$

Proof.

$$\nabla_{X'}Y = \dot{Y} - \mathrm{Proj}^\perp(\dot{Y}).$$

We already know that

$$\mathrm{Proj}^\perp(\dot{Y}) = -X\mathrm{sym}(X^T\dot{Y}).$$

□

Let us compute the geodesics. Recall that a curve $X(t)$ on $\mathrm{St}(n,p)$ is a geodesic if and only if

$$\nabla_{\dot{X}(t)}\dot{X}(t) = 0.$$

Here the Levi–Civita connection is explicit and gives

$$0 = \nabla_{\dot{X}}\dot{X} = \ddot{X} - \frac{1}{2}X(\ddot{X}^TX + X^T\dot{X}) \iff \ddot{X} = \frac{1}{2}X(\ddot{X}^TX + X^T\ddot{X})$$

with the condition that $X(t)X(t)^T = I$ for all t. Let us write $X(0) = X_0$ and $\dot{X}(0) = V_0 \in T_{X_0}\mathrm{St}(n,p)$. We can uniquely write

$$V_0 = X_0\Omega_0 + N_0$$

with $\Omega_0 := X_0^T V_0 \in \mathbb{R}^{p\times p}$ skew-symmetrix, $N_0 = V_0 - X_0\Omega_0$ (normal component, i.e. $X_0^T N_0 = 0$). We can write the thin SVD decomposition of N_0 as

$$N_0 = U\Sigma W^T,\ U \in \mathrm{St}(n,r),\ W \in \mathrm{St}(p,r),\ \Sigma \in \mathbb{R}^{r\times r} \text{ diagonal.}$$

Variational calculus minimization to $L = \int_0^1 \sqrt{g_c(\dot{c(t)}, \dot{c(t)})_{c(t)}}dt$ leads to the following geodesic equation, where $c(0) = X$:

$$\ddot{c(t)} + \dot{c(t)}\dot{c(t)}^T c(t) + c(t)((c(t)^T\dot{c(t)})^2 + \dot{c(t)}^T\dot{c(t)}) = 0.$$

With the same notations as before, (i.e. $V \in T_X\mathrm{St}(n,p)$ can be written as $V = X\Omega + X_\perp B$), an explicit formula is given by

$$c(t) = [X \mid X_\perp] \exp\left(\begin{bmatrix} \Omega & -B^T \\ B & O_{n-p} \end{bmatrix} t\right) \begin{bmatrix} \mathrm{I}_p \\ O_{(n-p)\times p} \end{bmatrix}.$$

There is no explicit expression for the Riemannian logarithm on the Stiefel manifold.

4.3 Symmetric Positive Definite (SPD) Matrix Manifold

The set of symmetric positive definite matrices on size n, defined by

$$\mathrm{Sym}(n)^+ := \{X \in \mathrm{Sym}(n) \mid X \succ 0\}$$

is a convex and open set in $\mathrm{Sym}(n)$ –the latter being a Euclidean space with inner product $\langle U, V\rangle = \mathrm{Tr}(U^T V) = \mathrm{Tr}(UV)$ – hence is an open submanifold ad its tangent spaces are identified with $\mathrm{Sym}(n)$.

Let us endow $\mathrm{Sym}(n)^+$ with a Riemannian structure that makes it complete. We will provide two ways.

4.3.1 Log–euclidean metric

See [2]. Let us consider $\varphi : \mathrm{Sym}(n)^+ \to \mathrm{Sym}(n) : X \mapsto \log(X)$. Since φ is smooth and invertible ($\varphi^{-1}(Y) = \exp(Y)$), it is a diffeomorphism. Let us pull-back the Euclidean metric from $\mathrm{Sym}(n)$ to $\mathrm{Sym}(n)^+$ so that we can define the following inner product on the tangent spaces $T_X\mathrm{Sym}(n)^+ = \mathrm{Sym}(n)$:

$$\langle U, V\rangle_X^{\log} := \langle D\log(X)[U], D\log(X)[V]\rangle.$$

(i) *Geodesics*: Let $X, X' \in \mathrm{Sym}(n)^+$. The unique minimizing geodesic connecting X to X' w.r.t. the Log-Euclidean metric is

$$c(t) = \exp(\log(X) + t(\log(X') - \log(X))).$$

(ii) *Characterization of set geodesic convexity*: $S \subseteq \mathrm{Sym}(n)^+$ is geodesically convex if and only if $\log(S)$ is convex in $\mathrm{Sym}(n)$.

(iii) *Characterization of geodesic convexity for functions*: given a geodesically convex set S, a function $f : S \to \mathbb{R}$ is geodesically (strictly) convex if and only if $f \circ \exp$ is (strictly) convex on $\mathrm{Sym}(n)$

4.3.2 Affine invariant metric

See [8]This more common metric on $\mathrm{Sym}(n)^+$ is defined on $T_X\mathrm{Sym}(n)^+$ as

$$\langle U, V\rangle_X^{\mathrm{aff}} := \left\langle X^{-1/2}UX^{-1/2}, X^{-1/2}VX^{-1/2}\right\rangle = \mathrm{Tr}(X^{-1}UX^{-1}V).$$

This metric is named after the so-called affine invariant property: for all invertible matrices $M \in \mathbb{R}^{n\times n}$, we have $MXM^T \succ 0$ and

$$\langle MUM^T, MVM^T\rangle_{MXM^T}^{\mathrm{aff}} = \langle U, V\rangle_X^{\mathrm{aff}}.$$

In particular, it entails that if $c : [0, 1] \to \mathrm{Sym}(n)^+$ is a smooth curve, then its length is equal to the length of $t \mapsto Mc(t)M^T$, given that they have equal speed at each t. In a similar way, the length of c is also equal to the length of $t \mapsto c(t)^{-1}$.

The orthogonal projection from $\mathbb{R}^{p\times p}$ to $T_X\mathrm{Sym}_p^{++}$ is given by $V \mapsto \mathrm{Sym}(V) = \frac{1}{2}(V + V^T)$. Let $X, Y \in \mathcal{X}(\mathrm{Sym}_p^{++})$ be two smooth vector fields. The Levi–Civita connection on Sym_p^{++} is given by

$$(\nabla_X Y)(\Sigma) = DY(\Sigma)[X] - \frac{1}{2}\left(X\Sigma^{-1}Y(\Sigma) + Y(\Sigma)\Sigma^{-1}X\right).$$

The geodesic such that $c(0) = X$ and $c'(0) = V$ is given by

$$\mathrm{Exp}_X(tV) = c(t) = X^{1/2}\exp\left(tX^{-1/2}VX^{-1/2}\right)X^{1/2},$$

which is defined for all t, making the manifold $\mathrm{Sym}(n)^+$ complete. The exponential mapping on Sym_p^{++} at Σ is then

$$\mathrm{Exp}_\Sigma(V) = \Sigma^{1/2}\exp\left(\Sigma^{-1/2}\xi\Sigma^{-1/2}\right)\Sigma^{1/2}.$$

The geodesic γ with endpoints conditions $\gamma(0) = \Sigma_1$ and $\gamma(1) = \Sigma_2$ is given by

$$\gamma(t) = \Sigma^{1/2}\left(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}\right)^t\Sigma_1^{1/2}$$

where $A^t = \exp(t\log(A))$ (here $\log(A)$ exists and is unique since $A \in \mathrm{Sym}_p^{++}$). The *logarithmic map* of $\Sigma_2 \in \mathrm{Sym}_p^{++}$ at $\Sigma_1 \in \mathrm{Sym}_p^{++}$ is

$$\log_{\Sigma_1}(\Sigma_2) = \Sigma_1^{1/2}\log(\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2})\Sigma_1^{1/2}.$$

which gives the following Riemannian distance (also invariant to affine transformation)

$$d(\Sigma_1, \Sigma_2) = \left\|\log\left(\Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2}\right)\right\|_2.$$

The parallel transport (see [9]) between Σ_1 and Σ_2 that moves $\xi \in T_{\Sigma_1}\mathrm{Sym}_p^{++}$ to $T_{\Sigma_2}\mathrm{Sym}_p^{++}$ while preserving the Riemanning metric is given by

$$\mathcal{T}_{\Sigma_1,\Sigma_2}(\xi) = (\Sigma_2\Sigma_1^{-1})^{1/2}\xi\left((\Sigma_2\Sigma_1^{-1})^{1/2}\right)^T.$$

The retraction

$$R_\Sigma(\xi) := \Sigma + \xi + \frac{1}{2}\xi\Sigma^{-1}\xi$$

defines a second-order retraction since $r''(t) = 0$, where $r(t) = R_\Sigma(t\xi)$. Moreover, it is a second-order approximation of the exponential mapping

$$\mathrm{Exp}_\Sigma(t\xi) = R_\Sigma(t\xi) + O(t^3).$$

Moreover, we have the inverse:

$$\mathrm{Exp}_X(V) = X' \iff V = \mathrm{Log}_X(X') = X^{1/2} \log(X^{-1/2} X' X^{-1/2}) X^{1/2}.$$

In particular,

$$\mathrm{dist}(X, X') = \| \log(X^{-1/2} X' X^{-1/2} \|_{\mathrm{Frob}}.$$

Chapter 5

Fisher–Rao geometry

5.1 Background

5.1.1 Statistical model

Let $\{x_i\}_{i=1}^n \in \mathcal{X}$ a sample of observation in the sample space $\mathcal{X}$, we often want to compute an estimation of a parameter $\theta \in \mathcal{E}$ (parameter space), such parameter being a discriminant feature for a given problem of interest. Usually $\mathcal{E}$ is a finite-dimensional linear space (say dimension q), endowed with the Euclidean inner product

$$\langle \theta_1, \theta_2 \rangle = \mathrm{vec}(\theta_1)^T \mathrm{vec}(\theta_2)$$

and Euclidean distance

$$d(\theta_1, \theta_2) = \|\mathrm{vec}(\theta_1) - \mathrm{vec}(\theta_2)\|_2$$

where $\mathrm{vec} : \mathcal{E} \to \mathbb{R}^q$ stacks its coordinates into a vector.

Definition 5.1.1 (Estimator, bias, unbiased estimator)**.** An estimator $\hat{\theta}$ of θ is a mapping from $\mathcal{X}$ to $\mathcal{E}$ such that

$$\{x_i\}_{i=1,\dots,n} \mapsto \hat{\theta}(\{x_i\}_{i=1}^n.$$

The bias of $\hat{\theta} \in \mathcal{E}$ is

$$b_\theta = \mathbb{E}\left[\mathrm{vec}(\hat{\theta}) - \mathrm{vec}(\theta)\right]$$

and $\hat{\theta}$ is unbiased if $b_\theta = 0$.

Let us now assume that the sample of observation $\{x_i\}_{i=1}^n \in \mathcal{X}$ is a realization of a random variable X with PDF f parameterized by some unknown $\theta \in \mathcal{E}$, i.e.

$$X \sim f(\cdot, \theta).$$

Definition 5.1.2 (Negative log-likelihood (NLL))**.** Let $\{x_i\}_{i=1}^n \in \mathcal{X}$ be a sample. Assume that for all $\theta \in \mathcal{E},\ f(\cdot, \theta) > 0$. The negative log-likelihood function $\mathcal{L} : \mathcal{E} \to \mathbb{R}$ is defined as

$$\mathcal{L}(\theta \mid \{x_i\}_{i=1}^n) := -\log f(\{x_i\}_{i=1}^n, \theta).$$

Since maximizing $\theta \mapsto f(\{x_i\}_{i=1}^n, \theta)$ is equivalent to minimizing $\theta \mapsto -\log f(\{x_i\}_{i=1}^n, \theta)$, the *maximum likelihood estimator* (MLE) $\hat{\theta}^{\mathrm{MLE}}$ is the minimizer of the negative log-likelihood function.

Definition 5.1.3 (MLE)**.** With the same notations,

$$\hat{\theta}^{\mathrm{MLE}} := \arg\min_{\theta\in\mathcal{E}} \mathcal{L}(\theta \mid \{x_i\}_{i=1}^n).$$

For information geometry, see [1].

Definition 5.1.4 (Covariance matrix)**.** Let $\hat{\theta}$ be an unbiased estimator of θ. The covariance matrix $C_\theta \in \mathbb{R}^{q\times q}$ is the following symmetric, positive semi-definite matrix:

$$C_\theta := \mathbb{E}\left[(\mathrm{vec}(\hat{\theta}) - \mathrm{vec}(\theta))(\mathrm{vec}(\hat{\theta}) - \mathrm{vec}(\theta))^T\right].$$

Remark 5.1.1. Remark that

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Tr}(C_\theta) = \mathbb{E}\left[\|\mathrm{vec}(\hat{\theta}) - \mathrm{vec}(\theta)\|_2^2\right] =: \mathrm{MSE}(\hat{\theta}, \theta).$$

5.1.2 CES distributions and Fisher information matrix

Definition 5.1.5 (C-CES distribution)**.** A vector $x \in \mathbb{C}^p$ follows a centered C-CES distribution, denoted by $x \sim \text{C-CES}(0, \Sigma, g)$ if

$$\mathrm{law}(x) = \mathrm{law}(\sqrt{Q}\Sigma^{1/2}u),$$

where $u \sim \mathrm{Unif}(\mathbb{C}S^p), Q \in \mathbb{R}^+$ is a real random variable with probability density function f_Q and is independent of u; Σ is a covariance matrix that we assume is full-rank.

Assuming full-rank for Σ makes it possible to consider the probability density function of x, given by

$$f_x(x \mid \Sigma) \propto |\Sigma|^{-1})g\left(x^H\Sigma^{-1}x\right),$$

where $g : \mathbb{R}^+ \to \mathbb{R}^+$ is the density generator which satisfies

$$\int_0^\infty t^{p-1}g(t)dt < \infty.$$

Proposition 5.1.1 (Log-likelihood)**.** *Let $\{x_i\}_{i=1}^n$ an i.i.d. sample from $x \sim$ C-CES$(0, \Sigma, g)$. Its log-likelihood is given as*

$$\mathcal{L}_g(\{x_i\}_{i=1}^n \mid \Sigma) = \sum_{i=1}^n \log\left(g\left(x_i^H\Sigma^{-1}x_i\right)\right) - n\log(|\Sigma|).$$

Let us denote by $\Sigma(v)$ a parametrization of the covariance matrix through a real-valued vector v (typically p^2). The *score* vector is defined entry-wise as

$$s_j = \frac{\partial\mathcal{L}_g\left(\{x_i\}_{i=1}^n \mid \Sigma(v)\right)}{\partial v_j}.$$

Definition 5.1.6 (Fisher information matrix)**.** The Fisher matrix associated to $\mathcal{L}_g$ is defined as

$$F_{jk} = \mathbb{E}\left[s_j s_k\right].$$

For C-CES distribution, the Fisher information matrix is explicit.

Proposition 5.1.2. *Let $\{x_i\}_{i=1}^n$ an i.i.d. sample from $x \sim$ C-CES$(0, \Sigma, g)$, with $\Sigma = \Sigma(v)$. The entries of the Fisher information matrix are*

$$F_{jk} = n\alpha_g \, Tr(\Sigma^{-1}\xi_j\Sigma^{-1}\xi_k) + n\beta_g \, Tr(\Sigma^{-1}\xi_j) \, Tr(\Sigma^{-1}\xi_k),$$

where

$$\xi_j = \frac{\partial \Sigma(v)}{\partial v_j} \ , \quad \alpha_g = 1 - \frac{\mathbb{E}[Q^2\phi'(Q)}{p(p+1)} \quad \textit{and} \quad \beta_g = \alpha_g - 1,$$

with $\phi = g'/g$.

The Fisher information matrix in involved in the so-called Cramér–Rao inequality:

$$\mathbb{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] \succeq F^{-1} \Rightarrow \|\hat{\theta} - \theta\|^2_{\text{Frob}} \geq \text{Tr}(F^{-1})$$

where $\hat{\theta}$ is an unbiased estimator of θ built from the observation sample $\{x_i\}_{i=1}^n$.

5.2 Riemannian structure

$\mathcal{H}_p^{++}$ (set of SPD Hermitian matrices) is a smooth open submanifold of $\mathcal{H}_p$ (vector space of Hermitian matrices) of dimension p^2 and at each $\Sigma \in \mathcal{H}_p^{++}$, the tangent space is identified as follows

$$T_\Sigma \mathcal{H}_p^{++} \simeq \mathcal{H}_p.$$

5.2.1 Riemannian metric

Instead of considering the Euclidean inner product $\langle \xi, \eta \rangle_\Sigma^{\text{eucl}} = \text{Re}(\text{Tr}(\xi^H \eta))$ that do not take into account the geometry of $\mathcal{H}_p^{++}$, one can define other metrics (e.g. Affine-Invariant metric, Bures–Wasserstein, log–euclidean). Let us define the FIsher information metric (FIM), which is particularly relevant when dealing wit a statistical model, since it is a data-driven metric.

Definition 5.2.1 (FIM Riemannian metric)**.** Let $\Sigma \in \mathcal{H}_p^{++}$. For all $\xi, \eta \in T_\Sigma \mathcal{H}_p^{++}$,

$$\langle \xi, \eta \rangle_\Sigma^{\text{FIM}} := \mathbb{E}\left[D\mathcal{L}_g(x \mid \Sigma)[\xi] D\mathcal{L}_g(x \mid \Sigma)[\eta] \right].$$

In the case of C-CES distributions, we have an explicit expression for the FIM.

Theorem 5.2.1 (FIM for C-CES distributions)**.** *Let $\Sigma \in \mathcal{H}_p^{++}$ and let $\{x_i\}_{i=1}^n$ an i.i.d. sample from $x \sim$ C-CES$(0, \Sigma, g)$. For all $\xi, \eta \in T_\Sigma \mathcal{H}_p^{++}$,*

$$\langle \xi, \eta \rangle_\Sigma^{FIM} = n\alpha_g \, Tr(\Sigma^{-1}\xi\Sigma^{-1}\eta) + n\beta_g \, Tr(\Sigma^{-1}\xi) \, Tr(\Sigma^{-1}\eta).$$

Sketch of the proof. The proof relies on the stochastic representation of $x = \sqrt{Q}\Sigma^{1/2}u$ of C-CES distributed sample points, on the invariance property of the trace operator and on the two differentials

$$D \log|\Sigma|[\xi] = \text{Tr}(\Sigma^{-1}\xi) \ \text{ and } \ D(\Sigma^{-1})[\xi] = -\Sigma^{-1}\xi\Sigma^{-1}.$$

□

In the remainder of the chapter, we will rather use the generic metric (general affine-invariant metric):

$$\langle \xi, \eta \rangle_\Sigma =: g_\Sigma(\xi, \eta) = \text{Re}\left[\alpha \text{Tr}(\Sigma^{-1}\xi\Sigma^{-1}\eta) + \beta \text{Tr}(\Sigma^{-1}\xi)\text{Tr}(\Sigma^{-1}\eta) \right] \tag{5.1}$$

with $\alpha > 0$ and $\beta > -\alpha/p$ (to ensure the positive definite property of the corresponding inner product).

Remark 5.2.1. The case where $\alpha = 1$ and $\beta = 0$ coincides with the FIM of the Gaussian distribution.

5.2.2 Levi–Civita connection

The Levi–Civita connection on $\mathcal{H}_p^{++}$ associated with the general affine-invariant metric (5.1) is explicit.

Theorem 5.2.2 (Levi–Civita connection)**.** *Let $\Sigma \in \mathcal{H}_p^{++}$. For all $\xi, \eta \in \mathcal{X}(\mathcal{H}_p^{++})$,*

$$\nabla_\xi \eta = D\eta[\xi] - Herm(\eta\Sigma^{-1}\xi).$$

Sketch of the proof. Invariance properties of the trace operator are used, alongside computations of differential involved in the Koszul formula. □

Remark 5.2.2. The Levi–Civita connection does not depend on α nor β, hence remains the same for all C-CES distributions.

5.2.3 Geodesics, exponential and logarithmic maps, Riemannian distance

Chapter 6

SPD neural network

The main reference for this chapter is [6].

SPD matrices appear in various fields, such as medical imaging or visual recognition (pedestrian detection, face recognition) and arise naturally when considering covariance matrices. Since SPD matrices lie on non-Euclidean manifolds, standard Euclidean operations on SPD matrices do not take into account there geometric specificities, hence leading to several drawbacks, namely

- **Swelling effect**: determinant values can increase following Euclidean averaging.
- **Non-positive eigenvalues** can result from some operations.

Affine-invariant Riemannian metrics are expensive to compute.

6.1 Log-Euclidean geometry

In order to address the previous issues, *log-euclidean metrics* provides a computationally-friendly framework which proceeds as follows:

1. Computating the matrix logarithm of SPD matrices.
2. Performing Euclidean operations in the logarithm domain.
3. Mapping the results back to SPD space via the matrix exponential.

Recall that the matrix exponential is defined by the following (absolutely convergent) power series:

$$\exp(M) = \sum_{k=0}^{+\infty} \frac{M^k}{k!}.$$

Definition 6.1.1. Let S be an SPD matrix. The matrix logarithm is defined as

$$\log(S) = P \log(D) P^T,$$

where $S = PDP^T$ is the eigenvalue decomposition of S, $D = \mathrm{diag}(\lambda_i)$ contans the eigenvalues of S and $\log(D)$ is a diagonal matrix with elements $\log(\lambda_i)$.

Remark that $S = \exp(\log(S))$. Since standard multiplication of SPD matrices does not commute, the *log-euclidean multiplication* is defined as

$$S_1 \circ S_2 := \exp(\log S_1 + \log S_2)$$

which is makes it commutative and asociative. The inverse is defined as

$$S{-}1 := \exp(-\log S).$$

and scalar multiplication is defined as

$$\lambda \circ S := \exp(\lambda \log S).$$

In particular, such operations make SPD matrices behave as a vector space in the logarithmic domain.

The geodesic distance between two SPD matrices S_1 and S_2 induced by the log-euclidean metric is given by the Euclidean distance (for some euclidean norm $\|\cdot\|$) between the logarithms of the matrices:

$$d_{LE}(S_1, S_2) = \|\log S_1 - \log S_2\|.$$

When $\|\cdot\|$ is the Frobenius norm, one get

$$d_{LE}(S_1, S_2) = \left(\sum_{i,j=1}^{n} (\log S_1 - \log S_1)_{ij}^2 \right)^{1/2}.$$

The geodesic connecting two SPD matrices S_1 and S_2 is given by

$$\gamma(t) = \exp((1-t)\log S_1 + t \log S_2), t \in [0, 1].$$

Notice that, contrarily to affine-invariant metric, it involves no inversion and exponentiation of matrices, and only features linear interpolation in the logarithmic domain.

Definition 6.1.2 (Fréchet mean)**.** Let $S_1, \ldots, S_n$ be SPD matrices. Their *Fréchet mean* is given by

$$E_{LE}(S_1, \ldots, S_n) = \exp\left(\frac{1}{N} \sum_{i=1}^{N} \log S_i \right).$$

Remark 6.1.1. With the same notations,

$$\det(E_{LE}) = \left(\prod_{i=1}^{N} \det S_i \right)^{1/N},$$

which prevents the swelling effect.

6.2 Architecture of SPD network

The article introduces a Riemannian deep learning architecture called SPDNet that processes SPD matrices while preserving the SPD structure across the layers. The key components of this architecture are three different types of layers:

- **BiMap** layer, analogous to convolutional layers, that maps SPD matrices to SPD matrices.
- **ReEig** layer, analogous to ReLU activation function, which induces non-linearity.
- **LogEig** layer, which maps SPD matrices to a flat Euclidean space to perform computations.

6.2.1 BiMap layer

The BiMap layer performs compression of the data while preserving its SPD structure thanks to the following bilinear mapping:

$$X_k = f_b^{(k)}(X_{k-1}, W_k) := W_k X_{k-1} W_k^T,$$

where $X_{k-1} \in \mathrm{Sym}^+_{d_{k-1}}$ is the input matrix at stage k, $W_k \in \mathbb{R}_*^{d_k \times d_{k-1}}$ is a learnable weight matrix in $\mathrm{St}(d_k, d_{k-1})$ in order to ensure boundedness on matrix distances. In particular, the output X_k belongs to $\mathrm{Sym}^+_{d_k}$.

6.2.2 ReEig layer

This layer mimics and adapts the effect of standard ReLU activation function to rectify SPD matrices non-linearly thanks to some thresholding procedure. More precisely, let $\varepsilon > 0$ the rectification threshold value. At stage k, this layer is defined as

$$X_k = f_r^{(k)}(X_{k-1}) := U_{k-1} \max(\varepsilon Id, \Sigma_{k-1}) U_{k-1}^T,$$

where as before $X_{k-1} = U_{k-1}\Sigma_{k-1}U_{k-1}^T$ and

$$\max(\varepsilon Id, \Sigma_{k-1}) = \begin{cases} \Sigma_{k-1}(i,i), & \Sigma_{k-1}(i,i) > \epsilon, \\ \epsilon, & \Sigma_{k-1}(i,i) \leq \epsilon. \end{cases}$$

Remark 6.2.1. The article also breifly mentions the possibility to adapt the standard sigmoid function to this setting in an analogous way as ReLU, namely by defining

$$X_k = U_{k-1}\sigma(\Sigma_{k-1})U_{k-1}^T,$$

where $\sigma(x) = \dfrac{1}{1+e^{-\alpha(x-\beta)}}$, with $\alpha > 0, \beta > 0$ to be chosen adequately so that the SPD structure remains preserved. This approach provides smooth non-linearity, as opposed to hard-thresholding (ReEig).

6.2.3 LogEig layer

The main purpose of this layer is to flatten elements of the curved SPD manifold into some flat vector space using the matrix logarithm. Since the Log-Euclidean metric treats the matrix logarithm as an isometry between the SPD manifold and a flat Euclidean space, it makes it possible to use classical loss functions and fully connected layers, while preserving the Riemannian geometry. This layer is defined (with the same notations as before) in the following way at stage k:

$$X_k = f_l^{(k)}(X_{k-1}) = \log(X_{k-1}) = U_{k-1} \log(\Sigma_{k-1}) U_{k-1}^T.$$

Another related approach is, as for the LogEig layer, to flatten the structure via the matrix logarithm, perform operations and this time to map the results back to the SPD manifold thanks to the matrix exponential. Given $X_1, \ldots, X_N$ SPD matrices, we first compute their logarithms $\log(X_1), \ldots, \log(X_N)$. It gives rise to the following layers:

Fully connected layer for a given SPD matrix X, we first compute its logarithm $\log(X)$, then we vectorize, which can be by stacking all the columns into a single column or stacking only the elements from the upper triangular part of $\log(X)$, since it is symmetric:

$$x = \text{vec}(\log(X)).$$

Then, the FC layer applies the following transformation

$$y = Wx + b$$

where W is the weight matrix. For classification tasks, the output of the FC layer is fed into a softmax layer that outputs class probabilities.

Average Pooling layer first, we perform average pooling in the Euclidean log-space, i.e.

$$\bar{L} := \frac{1}{N}\sum_{i=1}^{N}\log(X_i).$$

Then we map it back to SPD:

$$\bar{X} = \exp(\bar{L}).$$

6.3 Backpropagation

See [7]. Recall that given a set of data points $\mathcal{D} = \{(d^{(i)}, y^{(i)}\}_{i=1}^{N}$, a loss function $L : \mathbb{R}^d \to \mathbb{R}$ and a model prediction function with parameters W for the inputs $\{d\}_{i=1}^{N}$, i.e. $f : \mathbb{R}^D \to \mathbb{R}^d$ such that $y^{(i)} = f(d^{(i)}, W)$, Empirical Risk Minimization theory states that in order to learn well-enough, it suffices to minimize

$$\arg\min_{W} \frac{1}{N}\sum_{i=1}^{N} L(f(d^{(i)}, W), y^{(i)}).$$

When L and f are continuous, (sub)-gradient descent is commonly used as a way to learn the parameters W.

Deep networks – and in particular SPDNet – models can be represented as a composition of layer functions

$$f = f^{(l)} \circ f^{(l-1)} \circ \cdots \circ f^{(1)},$$

where $f^{(k)}$ is the function involved at the k-th layer and l the number of layers. At each layer k, we have a weight matrix W_k, yielding weight parameter tuple

$$W = (W_l, W_{l-1}, \ldots, W_1).$$

The loss as a function of the k-th layer is given by

$$L^{(k)} = L \circ f^{(l)} \circ \cdots \circ f^{(k)}.$$

The central tool in the learning (update of the weights parameters) procedure is the computation of the gradient of the loss function w.r.t. the parameters. Backpropagation heavily relies on the use of the chain rule to do so. Let (d, y) be any data tuple

$$\frac{\partial L^{(k)}(X_{k-1}, y)}{\partial W_k} = \frac{\partial L^{(k+1)}(X_k, y)}{\partial X_k}\frac{\partial f^{(k)}(X_{k-1})}{\partial W_k} \tag{6.1}$$

where $x_k = f^{(k)}(x_{k-1})$ and $x_0 = d$.
Since we also need to compute the gradients in the layers below and update their parameters, we also consider

$$\frac{\partial L^{(k)}(X_{k-1}, y)}{\partial X_k} = \frac{\partial L^{(k+1)}(X_k, y)}{\partial X_k} \frac{\partial f^{(k)}(X_{k-1})}{\partial X_{k-1}}$$

where y is the output and $X_k = f^{(k)}(X_{k-1})$. Denote by $A : B = \mathrm{Tr}(A^T B)$ the matrix inner product. Recall that for a generic differentiable function $\varphi : \mathbb{R}^{m \times n} \to \mathbb{R}$, one have the following first-order Taylor expansion:

$$\varphi(X + dX) = f(X) + \frac{\partial \varphi}{\partial X} : dX + O(\|dX\|^2).$$

6.3.1 BiMap layer

Recall that the BiMap layer bilinearly maps $X_k \in \mathrm{Sym}^+_{d_{k-1}}$ to $X_k = f_b^{(k)}(X_{k-1}, W_k) = W_k X_{k-1} W_k^T$ with $W_k \in \mathrm{St}(d_{k-1}, d_k)$. This latter constraint prevents the update to be solely using Equation (6.1) since it provides no guarantee that the updated weight remains on the Stiefel manifold. This is why the authors suggest using a Riemannian SGD algorithm on the Stiefel manifold. To do so, computation of the Riemannian gradient of $L^{(k)}$ at W_k, denotd by $\tilde{\nabla} L^{(k)}_{W_k}$, is performed, which consists at projecting the Euclidean gradient on the tangent space at W_k:

$$\tilde{\nabla} L^{(k)}_{W_k} = \nabla L^{(k)}_{W_k} (\mathbf{Id} - W_k^T W_k) \in T_{W_k} \mathrm{St}(d_{k-1}, d_k).$$

Differentiating $X_k = W_k X_{k-1} W_k^T$ w.r.t. W_K yields the Euclidean gradient

$$\nabla L^{(k)}_{W_k} = 2 \frac{\partial L^{(k+1)}}{\partial X_k} W_k X_{k-1}.$$

Then, gradient descent is performed using a retraction Γ. Given the value W_k^t at iteration t, we have

$$W_k^{t+1} = \Gamma_{W_k^t}(-\lambda \tilde{\nabla} L_{W_k^t}),$$

where λ is the learning rate.

6.3.2 EIG-based layers

See [7] The EIG operation on an SPD matrix X is the eigen-decomposition

$$X = U \Sigma U^T \tag{6.2}$$

where $UU^T = \mathbf{Id}$ and Σ is diagonal with positive entries σ_i. Since it is not an element-wise operation, backpropagation is not straightforward and must account for how small variations on X affect U and Σ.

Context We have a function $\Omega \mapsto L(\Omega)$ where Ω depends on $X \in \mathrm{Sym}^+_d$ through an eigen-decomposition. Let $X = U \Sigma U^T$ where $U^T U = \mathbf{Id}$ and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$, with $\sigma_i > 0$. We assume that

$$\Omega = \Omega(X) = U f(\Sigma) U^T$$

for some scalar function f. For backpropagation, how do we compute $\frac{\partial L}{\partial X}$ given $\frac{\partial L}{\partial \Omega}$?

Idea The chain rule states that

$$dL = \frac{\partial L}{\partial U} : dU + \frac{\partial L}{\partial \Sigma} : d\Sigma.$$

The next step is to substitue the expressions for dU and $d\Sigma$ in terms of dX. To do this, define the following forward operator:

$$\mathcal{F} : dX \mapsto (dU, d\Sigma).$$

We want to determine its adjoint operator $\mathcal{F}^*$ that verifies

$$\left(\frac{\partial L}{\partial U}, \frac{\partial L}{\partial \Sigma}\right) : \mathcal{F}(dX) = \mathcal{F}^* \left(\frac{\partial L}{\partial U}, \frac{\partial L}{\partial \Sigma}\right) : dX$$

so that we obtain our quantity of interest

$$\frac{\partial L}{\partial X} = \mathcal{F}^* \left(\frac{\partial L}{\partial U}, \frac{\partial L}{\partial \Sigma}\right).$$

Method Let $\mathcal{F}$ be the functional that describes the variation of the upper layer variables w.r.t. the lower layer variables:

$$dX_k = \mathcal{F}(dX_{k-1}).$$

Here, with the same notation of in Equation (6.2), $\mathcal{F} : dX \mapsto (dU, d\Sigma)$.
The idea is to (i) project the variation dX onto an admissible space and (ii) transfer the projection onto the derivative to obtain the projected gradient, by considering the adjoint operator $\mathcal{F}^*$ of $\mathcal{F}$. Namely,

$$\frac{\partial L^{(k+1)}}{\partial X_k} : dX_k = \frac{\partial L^{(k+1)}}{\partial X_k} : \mathcal{F}(dX_{k-1}) = \mathcal{F}^* \left(\frac{\partial L^{(k+1)}}{\partial X_k}\right) : dX_{k-1}$$

Hence

$$\frac{\partial L^{(k)}(X_{k-1}, y)}{\partial X_{k-1}} = \mathcal{F}^* \left(\frac{\partial L^{(k+1)}}{\partial X_k}\right). \tag{6.3}$$

Here, for both layers ReEig and LogEig layers, the EIG operation is involved through

$$X_{k-1} = U_{k-1}\Sigma_{k-1}U_{k-1}^T.$$

Consider a virtual layer k' for the EIG operation. Using Equation (6.3) gives

$$\begin{aligned} \frac{\partial L^{(k)}}{\partial X_{k-1}} : dX_{k-1} &= \mathcal{F}^* \left(\frac{\partial L^{(k')}}{\partial U}\right) : dX_{k-1} + \mathcal{F}^* \left(\frac{\partial L^{(k')}}{\partial \Sigma}\right) : dX_{k-1} && (6.4) \\ &= \frac{\partial L^{(k')}}{\partial U} : dU + \frac{\partial L^{(k')}}{\partial \Sigma} : d\Sigma && (6.5) \end{aligned}$$

Assuming a small variation dX_{k-1} of X_{k-1}, we obtain using Equation (6.2)

$$dX_{k-1} = dU\Sigma U^T + Ud\Sigma U^T + U\Sigma dU^T.$$

The variations in the eigenvalues are given by the following result.

Proposition 6.3.1. *With the same notations,*

$$d\Sigma = (U^T dX U)_{diag}.$$

Proof. Differentiating $UU^T = \mathbf{Id}$ w.r.t. U gives

$$d(UU^T) = dU^T U + U^T dU = 0$$

hence $U^T U$ is skew-symmetric. We have

$$\begin{aligned} U^T dXU &= (U^T dU)\Sigma + d\Sigma + \Sigma(dUU) \text{ using } UU^T = \mathbf{Id} \\ &= (U^T dU)\Sigma + d\Sigma - \Sigma(U^T dU) \text{ using that } U^T U \text{ is skew-symmetric} \\ &= d\Sigma + \underbrace{[U^T dU, \Sigma]}_{\text{off-diagonal}} \end{aligned}$$

thus

$$d\Sigma = (U^T dXU)_{\text{diag}}.$$

□

The variations in the eigenvectors are given by the following result.

Proposition 6.3.2. *With the same notations,*

$$dU = 2U(P^T \circ (UdXU)_{sym}),$$

where $\circ$ denotes the Hadamard element-wise product and the matrix P is defined as

$$P_{ij} = \frac{1}{\sigma_i - \sigma_j} \mathbb{1}_{i \neq j}.$$

Combining the previous expressions , we have

$$\begin{aligned} \frac{\partial L^{(k)}}{\partial X_{k-1}} &= \frac{\partial L^{(k')}}{\partial U} : 2U(P^T \circ (UdXU)_{\text{sym}}) + \frac{\partial L^{(k')}}{\partial \Sigma} : (U^T dXU)_{\text{diag}} \\ &= \left[2U \left(P^T \circ \left(U^T \frac{\partial L^{(k')}}{\partial U} \right)_{\text{sym}} \right) U^T + U \left(\frac{\partial L^{(k')}}{\partial \Sigma} \right)_{\text{diag}} U^T \right] : dX \end{aligned}$$

Since the previous equality holds for every dX, we finally obtain (see Proposition 2 in [**?**]),

$$\frac{\partial L^{(k)}}{\partial X_{k-1}} = 2U(P^T \circ (U^T \frac{\partial L^{(k')}}{\partial U})_{\text{sym}})U^T + U \left(\frac{\partial L^{(k')}}{\partial \Sigma} \right)_{\text{diag}} U^T.$$

ReEig layer: We have $X_k = Ug(\Sigma)U^T$ with $g(\Sigma) = \max(\varepsilon \mathbf{Id}, \Sigma)$.

Since $g'(\sigma) = \mathbb{1}_{\sigma > \varepsilon}$, let us define

$$Q := \text{diag}(q_1, \ldots, q_d) \text{ where } q_i = \mathbb{1}_{\sigma_i > \varepsilon}.$$

Let us write $A := \max(\varepsilon \mathbf{Id}, \Sigma)$. Since $dA = Qd\Sigma$, we have

$$dX_k = 2(dUAU^T)_{\text{sym}} + (UQd\Sigma U^T)_{\text{sym}}.$$

Then, we have

$$\begin{aligned} \frac{\partial L^{(k+1)}}{\partial X_k} : dX_k &= \frac{\partial L^{(k+1)}}{\partial X_k} : \left[2(dUAU^T)_{\text{sym}} + (UQd\Sigma U^T)_{\text{sym}} \right] \\ &= \left[2 \left(\frac{\partial L^{(k+1)}}{\partial X_k} \right)_{\text{sym}} UA \right] : dU + \left[Q^T U^T \left(\frac{\partial L^{(k+1)}}{\partial X_k} \right)_{\text{sym}} U \right] : d\Sigma \end{aligned}$$

we obtain, by identification with Equation (6.5):

$$\frac{\partial L^{(k')}}{\partial \Sigma} = QU^T \left(\frac{\partial L^{(k+1)}}{\partial X_k}\right)_{\text{sym}} U \quad \text{and} \quad \frac{\partial L^{(k')}}{\partial U} = 2\left(\frac{\partial L^{(k+1)}}{\partial X_k}\right)_{\text{sym}} U \max(\varepsilon \mathbf{Id}, \Sigma).$$

LogEig layer: We have $X_k = U g(\Sigma) U^T$ with $g(\Sigma) = \log(\Sigma)$. Since $g'(\Sigma) = \Sigma^{-1}$, from

$$dX_k = 2(dU \log(\Sigma) U^T)_{\text{sym}} + (U\Sigma^{-1} d\Sigma U^T)_{\text{sym}}$$

we obtain in the same way

$$\frac{\partial L^{(k')}}{\partial \Sigma} = \Sigma^{-1} U^T \left(\frac{\partial L^{(k+1)}}{\partial X_k}\right)_{\text{sym}} U \quad \text{and} \quad \frac{\partial L^{(k')}}{\partial U} = 2\left(\frac{\partial L^{(k+1)}}{\partial X_k}\right)_{\text{sym}} U \log(\Sigma).$$

More generally (see section 4.2 of [3]), we considering layers of the form $P \mapsto X = f(P)$, with f being a monotonous non-linear function, the output gradient $\frac{\partial L^{(k)}}{\partial P}$ is given by the following result.

Proposition 6.3.3. *If $P = U\Sigma U^T \mapsto X = U f(\Sigma) U^T$, then*

$$\frac{\partial L^{(k)}}{\partial P} = U\left(L \odot (U^T \left(\frac{\partial L^{(k+1)}}{\partial X}\right) U)\right) U^T,$$

where L is the Loewner matrix defined by

$$L_{ij} = \frac{f(\sigma_i) - f(\sigma_j)}{\sigma_i - \sigma_j} \mathbb{1}\sigma_i \neq \sigma_j + f'(\sigma_i) \mathbb{1}_{\sigma_i = \sigma_j}.$$

Sketch of the proof. Let $P \in \text{Sym}_d^{++}$ be decomposed as $P = U\Sigma U^T$ with $U^T U = \mathbf{Id}$ and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_d)$ with $\sigma_i > 0$. We have

$$X = U f(\Sigma) U^T, \quad \text{where} f(\Sigma) = \text{diag}(f(\sigma_1), \ldots, f(\sigma_d)).$$

To prove the wanted result, it suffices to show that

$$\frac{\partial L}{\partial X} : dX = U\left(L \circ (U^T \frac{\partial L}{\partial X} U)\right) U^T$$

where L is the Loewner matrix defined above.
We have

$$d(f(\Sigma)) = f'(\Sigma) \odot d\Sigma \ ; \ d\Sigma = (U^T dP U)_{\text{diag}} \ \text{ and } \ dU = 2U\left(\Psi^T \odot (U^T dP U)_{\text{sym}}\right)$$

with $\Psi_{ij} = \frac{1}{\sigma_i - \sigma_j} \mathbb{1}_{i \neq j}$. Then, by the product rule

$$dX = (dU) f(\Sigma) U^T + U d(f(\Sigma)) U^T + U f(\Sigma) dU^T.$$

It remains to show, using the trace properties and the previous expressions, that

$$\frac{\partial L}{\partial X} : dX = \text{Tr}\left((dX)^T \frac{\partial L}{\partial X}\right) = U\left(L \odot (U^T \frac{\partial L}{\partial X} U)\right) U^T : dP.$$

□

6.4 Riemannian batch normalization

In this section, we study how we can benefit from manifold geometry in order to introduce a Batch Normalization procedure that respects the geometry of the objects involved. To do so, we resort to tools such as Fréchet mean, parallel transport and Riemannian gradient descent. Such normalization layer can be integrated into existing SPD network architectures. The main reference for this section is [3].

6.4.1 Riemannian structure

We consider the Affine-Invariant metric (AIRM). Recall that the Riemannian distance between two SPD matrices P_1 and P_2 is

$$\delta_R(P_1, P_2) = \frac{1}{2} \left\| \log(P_1^{-1/2} P_2 P_1^{-1/2} \right\|_{\text{Frob}},$$

where $\log(\cdot)$ is the matrix logarithm and $\|\cdot\|_{\text{Frob}}$ is the Frobenius norm.
Let $P_0 \in \text{Sym}_n^+$.

Exponential map: for all $S \in T_{P_0}\text{Sym}_n^+$,

$$\text{Exp}_{P_0}(S) = P_0^{1/2} \exp(P_0^{-1/2} S P_0^{-1/2}) P_0^{1/2}.$$

Logarithmic map: for all $P \in \text{Sym}_n^+$,

$$\text{Log}_{P_0}(P) = P_0^{1/2} \log(P_0^{-1/2} P P_0^{-1/2}) P_0^{1/2}.$$

Parallel transport: Let P_1, P_2 be two SPD matrices. For all $S \in T_{P_1}\text{Sym}_n^+$, we have

$$\mathcal{T}_{P_1 \to P_2}(S) = (P_2 P_1^{-1})^{1/2} S (P_2 P_1^{-1})^{1/2}.$$

Let $\{P_i\}_{i=1}^N$ be a batch of SPD matrices. Since Sym_n^+ is not a vector space, the Euclidean arithmetic mean $\frac{1}{N}\sum_{i=1}^N P_i$ is no longer SPD. To circumvent this and preserve the SPD property, we instead consider the Fréchet mean, also refered to as Riemannian barycenter.

Definition 6.4.1 (Riemannian barycenter)**.** The weighted Riemannian barycenter of a batch $\{P_i\}_{i=1}^N \in (\text{Sym}_n^+)^N$ is given by

$$\text{Bar}_w(\{P_i\}_{i=1}^N) := \arg\min_{G \in \text{Sym}_n^+} \sum_{i=1}^N w_i \delta_R^2(G, P_i),$$

where the weights are any family $\{w_i\}_{i=1}^N$ such that $w_i \geq 0$ and $\sum_{i=1}^N w_i = 1$.

Remark 6.4.1. When $N = 2$ with weights $\{w, 1-w\}$, a closed-form formula exists for the Riemannian barycenter:

$$\text{Bar}_{\{w,1-w\}}(P_1, P_2) = P_1^{1/2} (P_1^{-1/2} P_2 P_1^{-1/2})^w P_1^{1/2}$$

where $\cdot^w = \exp(w \log(\cdot))$.
When $N > 2$, no closed-form solution exists and the Riemannian barycenter is computed iteratively using *Karcher flow algorithm.*

Karcher flow algorithm See [8] Section 3.7 Equation (14). We compute iteratively

$$G = \arg\min_{G \in \mathrm{Sym}_n^+} \sum_{i=1}^{N} w_i \delta_R^2(G, P_i).$$

The key idea is to compute the weighted arithmetic average in the tangent space via the logarithmic mapping and map the result back on the manifold using the exponential mapping.

- First, set $G^{(0)} = \frac{1}{N}\sum_{i=1}^{N} P_i$ (initial value).
- At each iterative step t, the update rule is

$$G^{(t+1)} = \mathrm{Exp}_{G^{(t)}}\left(\sum_{i=1}^{N} w_i \mathrm{Log}_{G^{(t)}}(P_i)\right).$$

Convergence is guaranteed as it follows from the fact that Sym_n^+ has non-positive curvature.

6.4.2 Riemannian batch normalization algorithm

Let us now see how to center and bias a batch of SPD matrice.

- To *center* a batch $\{P_i\}_{i=1}^N$ with Riemannian barycenter G, each matrix P_i is transported from G to the identity matrix via parallel transport:

$$\bar{P}_i = \mathcal{T}_{G \to \mathbf{Id}}(P_i) = G^{-1/2} P_i G^{-1/2}.$$

  It is analogous of $x - \mu$ in a Euclidean setting.

- To *bias* a batch $\{P_i\}_{i=1}^N$ with Riemannian barycenter G, each matrix $\bar{P}_i$ is transported from the identity matrix to G via parallel transport:

$$\tilde{P}_i = \mathcal{T}_{\mathbf{Id} \to G}(P_i) = G^{1/2} \bar{P}_i G^{1/2}.$$

  It is analogous of $\gamma\hat{x} + \beta$ is Euclidean batch-normalization setting.

Let us now present the Riemannian batch normalization algorithm. Consider a batch $\{P_i\}_{i=1}^N \in (\mathrm{Sym}_n^+)^N$ of SPD matrices. The general idea is (i) to compute the Riemannian batch barycenter, (ii) center each matrix, (iii) bias each matrix towards a (learnable) SPD parameter matrix G and (iv) maintaining a running barycenter G_S for inference. The training mode adapts dynamically to inputs while the inference mode ensures consistency.

Training mode :

1. Compute the Riemannian barycenter (via Karcher flow algorithm)

$$G_B = \mathrm{Bar}(\{P_i\}_{i=1}^N) = \arg\min_{G \in \mathrm{Sym}_n^+} \sum_{i=1}^{N} \delta_R^2(G, P_i).$$

2. Keep and update a running global barycenter G_S tracking the overall training distribution. The update is performed as a weighted Riemannian average between the running barycenter and the current batch barycenter:

$$G_S \leftarrow \mathrm{Bar}_{(\eta, 1-\eta)}(G_S, G_B) \iff G_S = G_B^{1/2}(G_B^{-1/2} G_S G_B^{-1/2})^{\eta} G_B^{1/2}, \quad \eta \in (0,1).$$

3. Center each batch matrix:

$$P_i^c = \mathcal{T}_{G_B \to \mathbf{Id}}(P_i) = G_B^{-1/2} P_i G_B^{-1/2}.$$

4. Bias the previous centered matrix from $\mathbf{Id}$ to the learnable bias SPD matrix G:

$$P_i^b = \mathcal{T}_{\mathbf{Id} \to G}(P_i^c) = G^{1/2} P_i^c G^{1/2}.$$

The normalized matrices are $\{P_i^b\}_{i=1}^N$.

Inference mode : We do not recalculate the batch barycenter G_B, but solely use the precomputed running barycenter G_S and precomputed learnt bias G from the training mode. Let $\{\tilde{P}_i\}_{i=1}^N$ be a test batch.

1. Center using running barycenter:

$$\tilde{P}_i^c = G_S^{-1/2} \tilde{P}_i G_S^{-1/2}.$$

2. Apply bias G:

$$\tilde{P}_i^b = G^{1/2} \tilde{P}_i^c G^{1/2}.$$

Recall that batchnorm bias parameter G is learnable. During training mode, at each processed batch:

1. We compute batch mean G_B via Karcher flow algorithm.
2. We perform the *forward* pass by centering around G_B and biasing by G.
3. To update G, we *backpropagate* through the network by computing the gradient of the loss function w.r.t. G and applying Riemannian gradient descent to update G while staying on the manifold.

For this last backpropagation part, the Euclidean gradient ∇^{eucl} is computed using the chain rule result of Proposition 6.3.3, since the forward mapping involves non-linear monotonic SPD functions like $G \mapsto G^{1/2}$ or $G \mapsto G^{-1/2}$. Then, this Euclidean gradient is projected onto the tangent space at G via

$$\Pi_{T_G \text{Sym}_d^+}(P) = G P_{\text{sym}} G, \quad P \in \text{Sym}_d^+.$$

Finally, the descent step with learning rate $\eta > 0$ is performed and mapped back on the manifold, i.e. at update iteration t,

$$G^{(t+1)} \leftarrow \exp_{G^{(t)}} \left(-\eta \Pi_{T_{G^{(t)}} \text{Sym}_d^+} (\nabla^{\text{eucl}} L(G^{(t)}) \right).$$

Chapter 7

Federated Learning

From the definition given by Kairouz et al., *federated learning* is a machine learning setting where multiple entities (called clients) collaborate in solving a machine learning (optimization) problem, under the coordination of a server or service provider. Each client's raw data is stored locally and not exchanged or transferred .Federated analytics works by running local computations over each device's data and only make the aggregated results available to product engineers. Federated optimization relies on several important ideas:

- Communication efficiency: reduced communication costs between the server and the clients.
- Data heterogeneity: clients can have different quantities of training data and statistically heterogeneous, i.e. a non-i.i.d. data partitioning in federated settings
- Computational constraints: hardware disparities between clients can lead to computational heteogeneity.
- Privacy and security: access to information from the data is only thgough aggregates.
- System complexity: stragglers, data storage and local computation. Server orchastration.

7.1 Framework

We want to minimize the objective function

$$F(x) = \mathbb{E}_{i\sim P}[F_i(x)], \quad \text{where } F_i(x) = \mathbb{E}_{\xi\sim D_i}[f_i(x,\xi)] \tag{7.1}$$

with

- $x \in \mathbb{R}^d$ representing the parameter for the global model;
- $F_i : \mathbb{R}^d \to \mathbb{R}$ the local objective function at client i;
- P being the distribution on the population of clients $\mathcal{I}$;
- $f_i(x,\xi)$ are the local loss functions (often the same across all clients) with local data distribution D_i that can vary (data heterogeneity).

Remark that direct computation of $F(x)$ or $\nabla F(x)$ cannot be computed directly since there is only access to a random sample of the clients at each round of communication.

Empirical risk minimization (ERM) paradigm applied to (7.1) leads to considering

$$F^{ERM}(x) = \sum_{i=1}^{M} p_i F_i^{ERM}(x), \text{ where } F_i^{ERM}(x) = \frac{1}{|D_i|} \sum_{\xi \in D_i} f_i(x, \xi) \text{ and } \sum_{i=1}^{M} p_i = 1$$

with $M = |\mathcal{I}|$ denoting the total number of clients and p_i being the relative weight of client i. ERM of the objective function across the union of all the local datasets can be achieved by setting $p_i = \frac{|D_i|}{\sum_{i=1}^{M} |D_i|}$.

- Since the local datasets D_i can have different distributions and size, the local objective functions $F_i(x)$ can be different and therefore have different local minima. Besides, the D_i cannot be shared with the server or shuffled across clients.
- The total number of clients M can be either very large and not well-defined. The client distribution P, the total number of clients M or the total number of data samples $\sum_{i=1}^{M} |D_i|$ are not known *a priori* before training starts.

7.1.1 Federated averaging algorithm

Classical GD solving sets

$$x^{(t+1)} = x^{(t)} - \eta_t \nabla F(x^{(t)}), \ t \in \mathbb{N}.$$

Under regularity assumptions, we can swap differentiation and expectation;

$$\nabla F(x) = \nabla \mathbb{E}_{i \sim P}[F_i(x)] = \mathbb{E}_{i \sim P}[\nabla F_i(x)].$$

Assume that at communication round t, only a finite subset $S^{(t)}$ of clients can connect to the server. Gradient descent update rule is then

$$x^{(t+1)} = x^{(t)} - \eta_t \frac{1}{|S^{(t)}|} \sum_{i \in S^{(t)}} \nabla F_i(x^{(t)}).$$

Stochastic approximation can be used to replace the excat gradient of the local loss functions with an unbiased stochastic gradient $g_i(x^{(t)})$ such that

$$\mathbb{E}_{\xi \sim D_i}[g_i(x^{(t)})] = \nabla F_i(x^{(t)}).$$

To reduce communication costs, each active client updates its local model for τ_i steps

$$\Delta_i^{(t)} = x_i^{(t,\tau_i)} - x^{(t)}$$

before the server aggregates them.
McMahan et al. proposed FedAvg, a federated averaging algorithm dividing the training process into rounds and relying on two levels of optimization strategies: ClientOpt and ServerOpt (e.g. SGD). For every round $t \in \{0, 1, \ldots, T-1\}$:

- the server broadcasts the current global model $x^{(t)}$ to a random subset of clients $S^{(t)}$ uniformly sampled without replacement. For each sampled client $i \in S^{(t)}$, in parallel:
  - initialization of the local model $x_i^{(t,0)} = x^{(t)}$

– for $k = 0, \ldots, \tau_i - 1$: computation of local stochastic gradient $g_i(x_i^{(t,k)})$ and performation of local update

$$x_i^{(t,k+1)} = \text{ClientOpt}(x_i^{(t,k)}, g_i(x_i^{(t,k)}), \eta_{\text{client}}, t).$$

At the end, compute local model changes

$$\Delta_i^{(t)} = x_i^{(t,\tau_i)} - x_i^{(t,0)}.$$

- Aggregation of local changes

$$\Delta^{(t)} = \sum_{i \in S^{(t)}} \frac{p_i \Delta_i^{(t)}}{\sum_{i \in S^{(t)}} p_i}$$

- Update global model:

$$x^{(t+1)} = \text{ServerOpt}(x^{(t)}, -\Delta^{(t)}, \eta_{\text{server}}, t).$$

Remark 7.1.1.

1. Unlike local SGD procedures, FedAvg assume only a subset of clients participate in each training round and no assumption is made on homogeneity of local data nor on the number of local updates, which can vary from one client to another.

2. Every client is allowed to have a different ClientOpt model (*personalized model*).

3. Regularization terms can be added to the objective function F to promote sparsity or enforce various constraints (rank, monotonicity, etc.):

$$\min_x F(x) + \Omega(x)$$

with regularizer $\Omega(x)$ convex, possibly non-smooth, non-finite additive.

7.1.2 Guidelines for developing practical algorithms

- Specify the application setting
- Improve communication efficiency that can be done by combining the following methods:
  – reducing the communication frequency by allowing local updates;
  – reducing communication volume by compressing messages;
  – reducing communication traffic at server by limiting the participating clients per round.
- Design for data and computational heterogeneity
- Compatibility with system architectures and privacy-preserving protocols:
  – monitoring client inactivity and using adequate sampling strategies;
  – privacy protection and weighting scheme

7.2 Federated optimization theory

Let us introduce theoretical tools to analyse and prove federated optimization algorithm FedAvg mentioned before.

Assumptions :

A.1 At any round t, at each client i are computed τ local SGD steps with constant learning rate η:

$$x_i^{(t,k+1)} = x_i^{(t,k)} - \eta g_i(x_i^{(t,k)}), \quad k \in \{0, \dots, \tau - 1\}$$

where g_i is the stochastic gradient.

A.2 $\Delta^{(t)} = x^{(t+1)} - x^{(t)}$

A.3 Finite number of clients: $\{1, 2, \dots, M\}$ with uniform contribution of the global objective

$$F(x) = \frac{1}{M} \sum_{i=1}^{M} F_i(x).$$

A.4 Each client participates at every round, i.e. $S^{(t)} = \{1, 2, \dots, M\}$.

A.5 For each client i, F_i is convex and L-smooth.

A.6 Each client i can query an unbiased stochastic gradient with σ^2-uniformly bounded variance in $\|\cdot\|_2$:

$$\mathbb{E}\left[g_i(x_i^{(t,k)}) \mid x_i^{(t,k)}\right] = \nabla F_i(x_i^{(t,k)}) \;\; ; \;\; \mathbb{E}\left[\left\|g_i(x_i^{(t,k)}) - \nabla F_i(x_i^{(t,k)})\right\|^2 \mid x_i^{(t,k)}\right] \leq \sigma^2.$$

A.7 The difference of $\nabla F_i(x)$ (local gradient) and $\nabla F(x)$ (global gradient) is ζ-uniformly bound in $\|\cdot\|_2$:

$$\max_i \sup_x \left\|\nabla F_i(x_i^{(t,k)}) - \nabla F(x_i^{(t,k)})\right\| \leq \zeta.$$

Since in the federated learning setup, we have multiple local iterates from the clients, we introduce the following *shadow sequence*:

$$\bar{x}^{(t,k)} := \frac{1}{M} \sum_{i=1}^{M} x_i^{(t,k)}$$

so that

$$\bar{x}^{(t,k+1)} = \bar{x}^{(t,k)} - \frac{\eta}{M} \sum_{i=1}^{M} g_i(x_i^{(t,k)}).$$

In a convergence proof, we want to show that there exists a function (rate of convergence) $r(T)$ decreasing with T such that

$$\mathbb{E}\left[\frac{1}{\tau T} \sum_{t=0}^{T-1} \sum_{k=1}^{\tau} F(\bar{x}^{(t,k)} - F(x^\star)\right] \leq r(T).$$

Remark that at the end of each round, $\bar{x}^{(t,\tau)} = \bar{x}^{(t+1,0)} = x^{(t+1)}$, which means the bound on the poxy $\bar{x}$ also quantifies the convergence of the global model.

To do so,

1. we first prove that progress is made at each round, i.e. that $\mathbb{E}\left[\frac{1}{\tau}\sum_{k=1}^{\tau} F(\bar{x}^{(t,k)}) - F(x^\star)\right]$ is bounded by some function plus an additional error term;
2. second, we prove that all client iterates remain close to the global shadow sequence, i.e. in expectation, $\|x_i^{(t,k)} - \bar{x}^{(t,k)}\|^2$ is bounded;
3. we finally use telescoping over t to show that over T rounds, significant progress towards the optimal value has been made.

Bibliography

[1] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[2] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347, 2007.

[3] Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for spd neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[5] William W Hager and Hongchao Zhang. A survey of nonlinear conjugate gradient methods. *Pacific journal of Optimization*, 2(1):35–58, 2006.

[6] Zhiwu Huang and Luc Van Gool. A riemannian network for spd matrix learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[7] Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *Proceedings of the IEEE international conference on computer vision*, pages 2965–2973, 2015.

[8] Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A riemannian framework for tensor computing. *International Journal of computer vision*, 66:41–66, 2006.

[9] Suvrit Sra and Reshad Hosseini. Conic geometric optimization on the manifold of positive definite matrices. *SIAM Journal on Optimization*, 25(1):713–739, 2015.