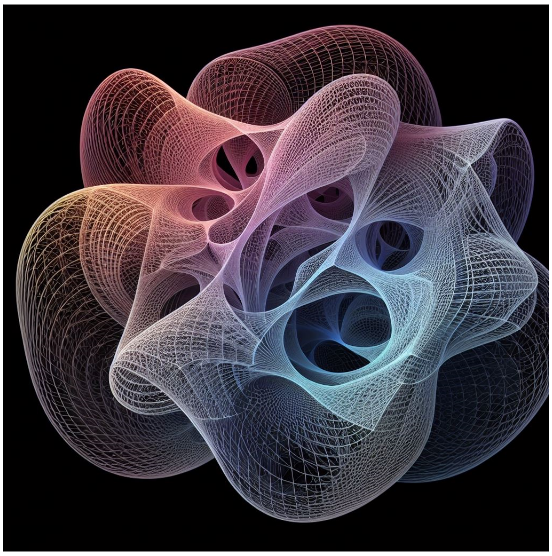

# LINMA2474 - High-Dimensional Data Analysis and Optimization

Issambre L'Hermite Dumont
Simon Desmidt

This summary may not be up-to-date, the newer version is available at this address:
https://github.com/SimonDesmidt/Syntheses

Academic year 2025-2026 - Q2

UCLouvain

# Contents

# Optimization on manifolds

## 1.1 Introduction

Classical optimization methods like the gradient descent solve problems of the form

$$\min_{x \in \mathcal{M}} f(x) \tag{1.1}$$

for a set $M$. The methods rely on two key properties:

- Linearity: $x_k$ and $\nabla f(x_k)$ belong to some vector space, in which they can be combined with linear operations;
- Inner product: $\nabla f(x_k)$ is the unique element of $\mathbb{R}^D$ such that

  $$\forall v \in \mathbb{R}^D,\ Df(x)[v] = \langle v, \nabla f(x) \rangle \tag{1.2}$$

  where $Df(x)[v] = \lim_{t \to 0} \frac{f(x+tv)-f(x)}{t}$ is the directional derivative of $f$ at $x$ in the direction $v$.

There are two ways to see the problem 1.1: as a constrained optimization problem, or as an unconstrained optimization problem assuming that nothing else exists outside the set $\mathcal{M}$. Optimization on manifolds extends the classical unconstrained optimization algorithms to problems whose search space is a manifold (will be defined later).

## 1.2 Definitions

### 1.2.1 Definition and properties of a manifold

**Definition 1.1 (Optimisation on manifolds).** To minimize a function $f$ on a manifold $\mathcal{M}$, we need several objects:

- A local defining function $h : \mathbb{R}^d \to \mathbb{R}$ such that $h^{-1}(0) = \mathcal{M}$;
- The tangent space of $\mathcal{M}$ at some point $x$ is the local linearization of $\mathcal{M}$ at $x \in \mathcal{M}$: $u \in$tangent space of $\mathcal{M}$ at $x$ iff $u \in Ker(Dh(x))$;
- An inner product on the tangent spaces to define a new notion of gradient:

  $$Df(x)[v] = \langle \nabla f(x), v \rangle$$

- A retraction function, i.e. a tool that allows to make a step on a manifold in a given tangent direction.

**Definition 1.2 (Smoothness).** $F : \mathcal{U} \subseteq \mathcal{E} \to \mathcal{V} \subseteq \mathcal{E}'$, with $\mathcal{U}, \mathcal{V}$ open, is said to be smooth if it is $\mathcal{C}^\infty$ on its domain.

Let us explain the concepts needed for our optimization:

**Definition 1.3 (Embedded submanifold and local defining function).** Let $\mathcal{E}$ be a linear space of dimension $d$. A non-empty subset $\mathcal{M}$ of $\mathcal{E}$ is a smooth embedded submanifold of $\mathcal{E}$ of dimension $n$ if either:

- $n = d$ and $\mathcal{M}$ is open in $\mathcal{E}$ (open submanifold);
- $n = d - k$ for some $k \geq 1$ and, for each $x \in \mathcal{M}$, there exists a neighbourhood $\mathcal{U}$ of $x$ in $\mathcal{E}$ and a smooth function $h : \mathcal{U} \to \mathbb{R}^k$. In that case,
  - $\mathcal{M} \cap \mathcal{U} = h^{-1}(0) = \{y \in \mathcal{U} : h(y) = 0\}$ and
  - $rank(Dh(x)) = k$.

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

**Definition 1.4 (Tangent space).** Let $\mathcal{M}$ be a subset of $\mathcal{E}$. For all $x \in \mathcal{M}$, define

$$\mathcal{T}_x\mathcal{M} = \{c'(0) \mid c : \mathcal{I} \to \mathcal{M} \text{ is smooth and } c(0) = x\} \tag{1.3}$$

where $\mathcal{I}$ is any open interval containing $t = 0$. That means that $v$ is in $\mathcal{T}_x\mathcal{M}$ iff there exists a smooth curve on $\mathcal{M}$ passing through $x$ with velocity $v$.

Consider $\mathcal{M}$ an embedded submanifold of $\mathcal{E}$, $x \in \mathcal{M}$ and the set $\mathcal{T}_x\mathcal{M}$.

- If $\mathcal{M}$ is an open submanifold of $\mathcal{E}$, then $\mathcal{T}_x\mathcal{M} = \mathcal{E}$;
- Otherwise, $\mathcal{T}_x\mathcal{M} = Ker(Dh(x))$ with $h$ any local defining function at $x$.

**Definition 1.5 (Tangent bundle).** The tangent bundle is the set of all tangent spaces: $\mathcal{T}\mathcal{M} = \{(x, v) : v \in \mathcal{T}_x\mathcal{M}\}$.

**Definition 1.6 (Map between manifolds).** Let $\mathcal{M}$ and $\mathcal{M}'$ be embedded submanifolds of $\mathcal{E}$ and $\mathcal{E}'$. A map $F : \mathcal{M} \to \mathcal{M}'$ is smooth iff $F = \bar{F}|_{\mathcal{M}}$ where $\bar{F}$ is some smooth map from a neighbourhood of $\mathcal{M}$ in $\mathcal{E}$ to $\mathcal{E}'$.

**Definition 1.7 (Differential of a map between manifolds).** The differential of $F : \mathcal{M} \to \mathcal{M}'$ at the point $x \in \mathcal{M}$ is the linear map $DF(x) : \mathcal{T}_x\mathcal{M} \to \mathcal{T}_{F(x)}\mathcal{M}'$ defined by

$$DF(x)[v] = \frac{d}{dt}F(c(t))|_{t=0} = (F \circ c)'(0) \tag{1.4}$$

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

$\to$ Note: the definition does not depend on the choice of the curve $c$: $DF(x) = D\bar{F}(x)|_{\mathcal{T}_x\mathcal{M}}$.

**Definition 1.8** (**Retraction function**). A retraction on a manifold $\mathcal{M}$ is a smooth map $R : \mathcal{TM} \to \mathcal{M} : (x, v) \to R_x(v)$ such that each curve $c(t) = R_x(tv)$ satisfies $c(0) = x$ and $c'(0) = v$.

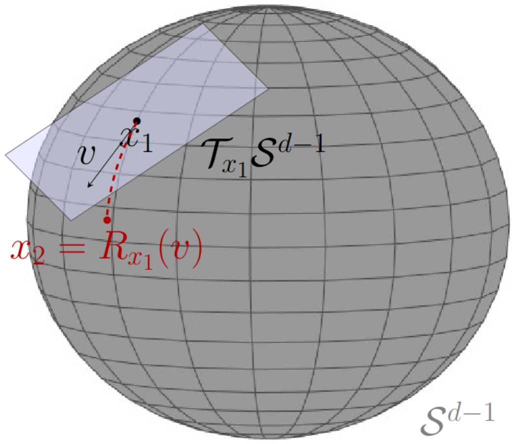


Figure 1.1: Retraction on a sphere.

### 1.2.2 Riemannian manifolds and metrics

**Definition 1.9** (**Inner product**). As seen in previous courses, an inner product on $\mathcal{T}_x\mathcal{M}$ is a bilinear, symmetric, positive definite function $\langle \cdot, \cdot \rangle_x : \mathcal{T}_x\mathcal{M} \times \mathcal{T}_x\mathcal{M} \to \mathbb{R}$. Note that the inner product depends on the point of linearization. It induces some norm for tangent vectors: $\|u\|_x = \sqrt{\langle u, u \rangle_x}$. A metric on $\mathcal{M}$ is a choice of inner product $\langle \cdot, \cdot \rangle_x$ for each $\mathcal{M}$.

**Definition 1.10** (**Metric**). A metric $x \to \langle \cdot, \cdot \rangle_x$ on $\mathcal{M}$ is a Riemannian metric if it varies smoothly with $x$, i.e. for all smooth vector fields $V, W$ on $\mathcal{M}$, the function $x \to \langle V(x), W(x) \rangle_x$ is smooth from $\mathcal{M}$ to $\mathbb{R}$.

**Definition 1.11** (**Riemannian manifold**). A Riemannian manifold is a manifold with a Riemannian metric.

**Definition 1.12** (**Riemannian distance**). Let $\mathcal{M}$ be a Riemannian manifold. Given a smooth curve $c : [a, b] \to \mathcal{M}$, we define the length of $c$ as

$$L(c) = \int_a^b \|c'(t)\|_{c(t)} dt \tag{1.5}$$

The Riemannian distance is then defined as $dist(x, y) = \inf_c L(c)$.

**Definition 1.13** (**Riemannian submanifolds**). Let $\mathcal{M}$ be an embedded submanifold of a Euclidean space $\mathcal{E}$. Equipped with the Riemannian metric obtained by restriction of the metric of $\mathcal{E}$, we call $\mathcal{M}$ a Riemannian submanifold of $\mathcal{E}$.

### 1.2.3 Gradient on manifolds

**Definition 1.14 (Riemannian gradient).** Let $f : \mathcal{M} \to \mathbb{R}$ be smooth on a Riemannian manifold $\mathcal{M}$. The Riemannian gradient of $f$ is the vector field $\mathrm{grad} f$ on $\mathcal{M}$ uniquely defined by the following identities:

$$\forall (x, v) \in \mathcal{TM}, \qquad Df(x)[v] = \langle v, \mathrm{grad} f(x) \rangle_x \tag{1.6}$$

where $Df(x)$ is the differential of $f$ and $\langle \cdot, \cdot \rangle_x$ is the Riemannian metric.

**Theorem 1.15.** Let $\mathcal{M}$ be a Riemannian submanifold of $\mathcal{E}$ endowed with the metric $\langle \cdot, \cdot \rangle$ and let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function. The Riemannian gradient of $f$ is given by

$$\mathrm{grad} f(x) = \mathrm{Proj}_x(\nabla \bar{f}(x)) \tag{1.7}$$

where $\bar{f}$ is any smooth extension of $f$ to a neighborhood of $\mathcal{M}$ in $\mathcal{E}$, and $\nabla \bar{f}(x)$ is the Euclidean gradient of $\bar{f}$ at $x$.

→ Note: for $\mathcal{E} = \mathbb{R}^d$ and using the usual metric $\langle u, v \rangle = u^T v$, the projection operator is

$$\mathrm{Proj}_x(v) = v - (x^T v)x \tag{1.8}$$

**Proposition 1.16.** Let $f : \mathcal{M}_1 \to \mathcal{M}_2$ and $G : \mathcal{M}_2 \to \mathcal{M}_3$ be smooth, where $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ are embedded submanifolds of $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ respectively. Then

$$G \circ F : \mathcal{M}_1 \to \mathcal{M}_3 : x \to G(F(x)) \tag{1.9}$$

is smooth and the chain rule applies:

$$D(G \circ F)(x)[v] = DG(F(x))[DF(x)[v]] \tag{1.10}$$

### 1.2.4 Taylor development of functions defined on manifolds

Let $f : \mathcal{M} \to \mathbb{R}$ be smooth and $c : \mathcal{I} \to \mathcal{M}$ be a smooth curve with $c(0) = x$ and $c'(0) = v$, with $\mathcal{I} \subseteq \mathbb{R}$ an open interval around $t = 0$ and $\|v\|_x = 1$. Let us define $g : \mathcal{I} \to \mathbb{R} : t \to g(t) = f(c(t))$. Since $g = f \circ c$ is smooth and maps real numbers to real numbers, it admits a Taylor expansion:

$$g(t) = g(0) + tg'(0) + \mathcal{O}(t^2) \tag{1.11}$$

By the chain rule,

$$g'(t) = Df(c(t))[c'(t)] = \langle \mathrm{grad} f(c(t)), c'(t) \rangle_{c(t)} \tag{1.12}$$

and for $t = 0$,

$$g(0) = f(x) \qquad g'(0) = \langle \mathrm{grad} f(x), v \rangle_x \tag{1.13}$$

Therefore,

$$\begin{aligned} f(c(t)) &= f(x) + t\langle \mathrm{grad} f(x), v \rangle_x + \mathcal{O}(t^2) \\ f(R_x(tv)) &= f(x) + \langle \mathrm{grad} f(x), tv \rangle_x + \mathcal{O}(t^2) \end{aligned} \tag{1.14}$$

And defining $s := tv \in \mathcal{T}_x\mathcal{M}$,

$$f(R_x(s)) = f(x) + \langle \mathrm{grad} f(x), s \rangle_x + \mathcal{O}(\|s\|_x^2) \tag{1.15}$$

This allows to define the Riemannian gradient descent in the next chapter.

# Riemannian gradient descent

## 2.1 Topology tools

### 2.1.1 General topology

**Definition 2.1 (Topology).** A topology on a set $X$ is a collection $T$ of subsets of $X$, called open sets, such that

- $X$ and $\emptyset$ belong to $X$;
- the union of the elements of any subcollection of $T$ is in $T$;
- the intersection of the elements of any finite subcollection of $T$ is in $T$.

**Definition 2.2 (Hausdorff topological space.).** A topological space is a couple $(X, T)$ where $X$ is a set and $T$ is a topology on $X$. The topological space $(X, T)$ is Hausdorff if any two distinct points of $X$ have disjoint neighborhoods. If $X$ is Hausdorff, then eveery sequence of points of $X$ converges to at most one point in $X$.

### 2.1.2 Topology for embedded submanifolds

**Definition 2.3 (Open set.).** A subset $\mathcal{U}$ of $\mathcal{M}$ is open (respectively closed) if it is the intersection of $\mathcal{M}$ with an open (respectively closed) set of $\mathcal{E}$.

**Definition 2.4 (Neighborhood.).** A neighborhood of $x$ in $\mathcal{M}$ is an open subset of $\mathcal{M}$ that contains $x$.
A neighborhood of a subset of $\mathcal{M}$ is an open subset of $\mathcal{M}$ that contains that subset.

**Definition 2.5 (Limit.).** Consider a sequence $s$ of points $x_0, x_1, x_2, \ldots$ on a manifold $\mathcal{M}$.

- We way that a point $x$ in $\mathcal{M}$ is a limit of $s$ if, for all neighborhood $\mathcal{U}$ of $x$, there exists some $K \in \mathbb{Z}$ such that $x_k$ is in $\mathcal{U}$ for all $k \geq K$.
  The topology of a manifold is Hausdorff, hence a sequence has at most one limit. If $x$ is the limit, we write
  $$\lim_{k\to\infty} x_k = x \qquad \text{or} \qquad x_k \to x \tag{2.1}$$
  and we say that the sequence converges to $x$.
- A point $x \in \mathcal{M}$ is an accumulation point of $s$ if it is the limit of a subsequence of $s$.

## 2.2 Riemannian gradient descent algorithm

The basic algorithm is the following:

**Algorithm 1** Riemannian gradient descent

1: **Input:** $x_0 \in \mathcal{M}$;
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Pick a step size $\alpha_k > 0$;
4:
$$x_{k+1} = R_{x_k}\left(-\alpha_k \mathrm{grad} f(x_k)\right) \tag{2.2}$$
5: **end for**

There are several ways to choose the step size $\alpha_k$:

- a fixed step size: $\alpha_k = \alpha$ for all $k$;
- optimal step size: compute $\alpha_k$ that minimizes exactly the function
$$g(\alpha) = f\left(R_{x_k}(-\alpha \mathrm{grad} f(x_k))\right) \tag{2.3}$$
- Backtracking: starting with a guess $\alpha_0 > 0$, iteratively reduce it by a factor $\tau \in (0, 1)$ until it is deemed acceptable.

### 2.2.1 Global convergence

Let $f : \mathcal{M} \to \mathbb{R}$ be a smooth function on a Riemannian manifold $\mathcal{M}$, and assume that

- there exists a lower bound $f_{low} \in \mathbb{R}$ such that $f(x) \geq f_{low}$ for all $x \in \mathcal{M}$;
- there exists a constant $c > 0$ such that, for all $k \in \mathbb{Z}$,
$$f(x_k) - f(x_{k+1}) \geq c\|\mathrm{grad} f(x_k)\|_{x_k}^2 \tag{2.4}$$

Then,
$$\lim_{k\to\infty} \|\mathrm{grad} f(x_k)\|_{x_k}^2 = 0 \tag{2.5}$$

Furthermore, for all $K \geq 1$, there exists $k \in \{0, \ldots, K-1\}$ such that
$$\|\mathrm{grad} f(x_k)\|_{x_k} \leq \sqrt{\frac{f(x_0) - f_{low}}{cK}} \tag{2.6}$$

This means that if a precision $\epsilon$ is required on the solution, the number of iterations that we will need has a complexity $\mathcal{O}(\epsilon^{-2})$.

The proof of these results is the same as done in the courses LINMA2471 or LINMA2460.

### 2.2.2 Ensuring the assumptions

In classical optimisation, the Lipschitz-smoothness assumption is that there exists a constant $L > 0$ such that

$$\|\nabla f(x+s) - \nabla f(x)\| \leq L\|s\| \qquad \forall x, s \in \mathbb{R}^d \tag{2.7}$$

implied by the expression

$$f(x+s) \leq f(x) + \langle \nabla f(x), s\rangle + \frac{L}{2}\|s\|^2 \qquad \forall x, s \in \mathbb{R}^d \tag{2.8}$$

However, on manifolds, the gradient is defined differently and with a different support. We need to modify this condition.
On a manifold $\mathcal{M}$, for a given subset $S$ of the tangent bundle $\mathcal{TM}$, there exists a constant $L > 0$ such that, forall $(x, s) \in S$,

$$f(R_x(s)) \leq f(x) + \langle \mathrm{grad} f(x), s\rangle + \frac{L}{2}\|s\|_x^2 \tag{2.9}$$

From this, we can ensure the assumption of equation (2.4): define $x =: x_k$, $s =: -\alpha_k \mathrm{grad} f(x_k)$ and $R_x(s) =: R_x(s)$. Then,

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \left(1 - \frac{L\alpha_k}{2}\right) \|\mathrm{grad} f(x_k)\|_{x_k}^2 \tag{2.10}$$

and this gives

$$f(x_k) - f(x_{k+1}) \geq \alpha_k \left(1 - \frac{L\alpha_k}{2}\right) \|\mathrm{grad} f(x_k)\|_{x_k}^2 \tag{2.11}$$

Therefore, the assumption is true taking $c = \alpha_k \left(1 - \frac{L\alpha_k}{2}\right)$.

### 2.2.3 Optimal step size

From the previous proof, we know that the decrease of the iterate is best when the constant $c$ is maximized. We can thus define the function

$$g : \mathbb{R} \to \mathbb{R} : \alpha \to \alpha \left(1 - \frac{L\alpha}{2}\right) \tag{2.12}$$

which is maximized for $\alpha^* = 1/L$, the same value seen for optimization in $\mathbb{R}^d$. For that value, the constant is $c = 1/2L$.

### 2.2.4 Backtracking

If the Lipschitz constant $L$ is unknown, a possible algorithm for the optimal step size if backtracking line-search using the Armijo-Goldstein condition. This condition is

$$f(x) - f(R_x(-\alpha \mathrm{grad} f(x))) \geq r\alpha \|\mathrm{grad} f(x)\|_x^2 \tag{2.13}$$

and the algorithm is

**Algorithm 2** Backtracking line-search

**Parameters:** $\tau, r \in (0,1)$, e.g. $\tau = 1/2$ and $r = 10^{-4}$;
**Input:** $x \in \mathcal{M}$ and $\bar{\alpha} > 0$;
Set $\alpha \leftarrow \bar{\alpha}$;
**while** $f(x) - f(R_x(-\alpha \mathrm{grad} f(x))) < r\alpha \|\mathrm{grad} f(x)\|^2$ **do**
    $\alpha \leftarrow \tau\alpha$
**end while**
**return** $\alpha$

Under an assumption similar to the Lipschitz condition on the gradient, a convergence rate of $\mathcal{O}(1/\epsilon^2)$ can still be guaranteed.