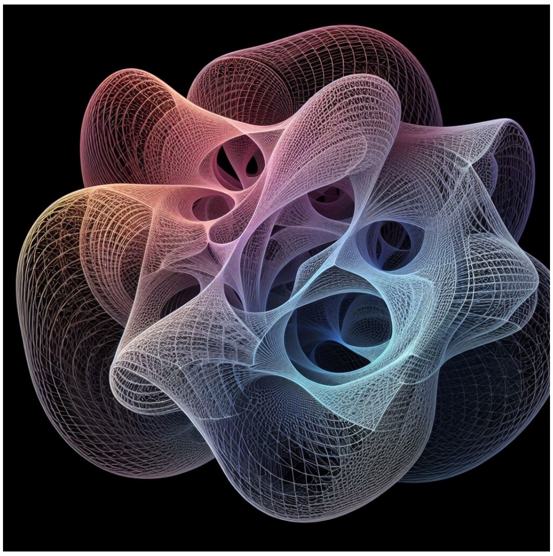

LINMA2474 - High-Dimensional Data Analysis and Optimization

Issambre L'Hermite Dumont
Simon Desmidt

Academic year 2025-2026 - Q2

UCLouvain

Contents

Introduction to 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 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.
We define a manifold as a set that can be locally approximated linearly (and therefore smooth).

1.2 Examples

1.2.1 The sphere

The sphere is a manifold: let $\mathcal{S}^{d-1} =: \{x \in \mathbb{R}^d :\ \|x\|_2^2 = 1\} = \{x \in \mathbb{R}^d :\ x^T x = 1\}$. This set is thus defined by the constraint $h(x) =: x^T x - 1 = 0$. We call this function $h : \mathbb{R}^d \to \mathbb{R}$ a defining function.
Let us use a Taylor approximation to derive the local linearization of $\mathcal{S}^{d-1}$ around any point $x \in \mathcal{S}^{d-1}$. Let $v \in \mathbb{R}^d$ be an arbitrary vector.

$$h(x + tv) = h(x) + tDh(x)[v] + \mathcal{O}(t^2) \tag{1.3}$$

Therefore, at first order, $x + tv \in \mathcal{S}^{d-1}$ iff

$$Dh(x)[v] = 2x^T v = 0 \iff v^T x = 0 \tag{1.4}$$

This means that around any point $x \in \mathcal{S}^{d-1}$, the sphere can be locally approximated by the set $\{v \in \mathbb{R}^d : \, x^T v = 0\}$, called the tangent space.
While the standard gradient descent defines the relation

$$x_{k+1} = x_k - \eta \nabla f(x_k) \tag{1.5}$$

we rather define the retraction operator for optimization on manifolds:

$$x_{k+1} = \mathcal{R}_{x_k}(-\eta \mathrm{grad} f(x_k)) \tag{1.6}$$

which will be explained later.

- A cube is not a Riemannian manifold because of the edges: they are not smooth;
- The set of matrices of rank r is a manifold, but the set of matrices of rank $\leq r$ is not, because going from rank i to $i+1$ is not smooth.

1.3 Applications

1.3.1 The Netflix problem

Let us consider that we know some ratings of users for movies, and we want to predict their rating for films based on their previous experiences. As people have hardly seen any movies in the catalogue, the rating matrix is very sparse.
Let us define $M \in \mathbb{R}^{m \times n}$ as the rating matrix, and $\Omega \subseteq \{1, \ldots, m\} \times \{1, \ldots, n\}$ as the set of indices of the known ratings. The problem is to find $X \in \mathbb{R}^{m \times n}$ that solves

$$\min_{X \in \mathbb{R}^{m \times n}} \sum_{(i,j) \in \Omega} (M_{ij} - X_{ij})^2 \tag{1.7}$$

We want to express X as the product of two low-rank matrices $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$, such that $X = UV^T$. Those matrices contain how much the users enjoy some features (long movies vs series, action vs romance, etc.) and how much those features are present in the movies.
The problem then becomes

$$\min_{X \in \mathbb{R}_k^{m \times n}} \sum_{(i,j) \in \Omega} (M_{ij} - X_{ij})^2 \tag{1.8}$$

where the set $\mathbb{R}_k^{m \times n}$ is the set of matrices of size $m \times n$ and of rank k. This set is a manifold, as will be proved later.

1.3.2 Dictionary learning

Let $x_1, \ldots, x_m$ be a collection of datapoints. The goal is to learn k atoms $b_1, \ldots, b_k$ ($k \ll m$) such that each datapoint x_i can be represented by a small number of properly chosen atoms: we want $X \approx BC$ for $B \in \mathbb{R}^{d \times k}$ and $C \in \mathbb{R}^{k \times m}$.
The problem writes

$$\min_{B,C} \|X - BC\|^2 + \lambda \|C\|_0 \qquad \text{s.t.} \qquad \|b_i\| = 1 \qquad \forall i = 1, \ldots, k \tag{1.9}$$

where the $\| \cdot \|_0$ norm is the number of non-zero entries in the matrix. The constraint is added to reduce the number of solutions and to be able to use a manifold. It defines the oblique manifold:

$$\mathcal{OB}(d,k) =: \{X \in \mathbb{R}^{d \times k} : \, \|X_{:,i}\|_2^2 = 1, \, \forall i\} \tag{1.10}$$

1.3.3 PCA

Let $x_1, \dots, x_n$ be a centered dataset in $\mathbb{R}^d$. We aim to find a collection of k orthogonal unit-norm vectors $u_1, \dots, u_k$ such that the subspace spanned by these vectors captures the most of the variance of the initial dataset. It can be expressed as an optimization problem:

$$\max_{U \in \mathbb{R}^{d \times k}} tr(U^T X X^T U) \qquad \text{s.t.} \qquad U^T U = I_k \tag{1.11}$$

This helps define the Stiefel manifold:

$$\mathcal{S}\sqcup(d,k) =: \{X \in \mathbb{R}^{d \times k} : \; X^T X = I_k\} \tag{1.12}$$

→ Note: the cost function is invariant by rotation: $f(UQ) = f(U)$.