

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



Contents

1	Introduction to optimization on manifolds	2
1.1	Introduction	2
1.2	Examples	2
1.3	Applications	3
1.4	Definitions	4

Introduction to optimization on manifolds

1.1 Introduction

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

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

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

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

where $Df(x)[v] = \lim_{t \rightarrow 0} \frac{f(x+tv)-f(x)}{t}$ 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 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 $S^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2^2 = 1\} = \{x \in \mathbb{R}^d : x^T x = 1\}$. This set is thus defined by the constraint $h(x) := x^T x - 1 = 0$. We call this function $h : \mathbb{R}^d \rightarrow \mathbb{R}$ a defining function.

Let us use a Taylor approximation to derive the local linearization of S^{d-1} around any point $x \in S^{d-1}$. Let $v \in \mathbb{R}^d$ be an arbitrary vector.

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

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

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

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

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

we rather define the retraction operator for optimization on manifolds:

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

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 $\Omega \subseteq \{1, \dots, m\} \times \{1, \dots, n\}$ as the set of indices of the known ratings. The problem is to find $X \in \mathbb{R}^{m \times n}$ that solves

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

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

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

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, \dots, x_m be a collection of datapoints. The goal is to learn k atoms b_1, \dots, 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

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

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:

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

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:

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

This helps define the Stiefel manifold:

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

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

1.4 Definitions

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

- A local defining function $h : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $u \in \text{Ker}(Dh(x))$;
- 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). $F : \mathcal{U} \subseteq \mathcal{E} \rightarrow \mathcal{V} \subseteq \mathcal{E}'$, with \mathcal{U}, \mathcal{V} open, is said to be smooth if it is \mathcal{C}^∞ on its domain.

Definition 1.3 (Embedded submanifold). 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 $h : \mathcal{U} \rightarrow \mathbb{R}^k$. In that case,
 - $\mathcal{M} \cap \mathcal{U} = h^{-1}(0) = \{y \in \mathcal{U} : h(y) = 0\}$ and
 - $\text{rank}(Dh(x)) = k$.

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

$$\mathcal{T}_x \mathcal{M} = \{c'(0) \mid c : \mathcal{I} \rightarrow \mathcal{M} \text{ is smooth and } c(0) = x\} \quad (1.13)$$

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, $\mathcal{T}_x\mathcal{M} = \text{Ker}(Dh(x))$ 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 $F : \mathcal{M} \rightarrow \mathcal{M}'$ 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 $F : \mathcal{M} \rightarrow \mathcal{M}'$ at the point $x \in \mathcal{M}$ is the linear map $DF(x) : \mathcal{T}_x\mathcal{M} \rightarrow \mathcal{T}_{F(x)}\mathcal{M}'$ defined by

$$DF(x)[v] = \frac{d}{dt} F(c(t))|_{t=0} = (F \circ c)'(0) \quad (1.14)$$

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

→ 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 $R : \mathcal{T}\mathcal{M} \rightarrow \mathcal{M} : (x, v) \rightarrow R_x(v)$ such that each curve $c(t) = T_x(tv)$ satisfies $c(0) = x$ and $c'(0) = v$.

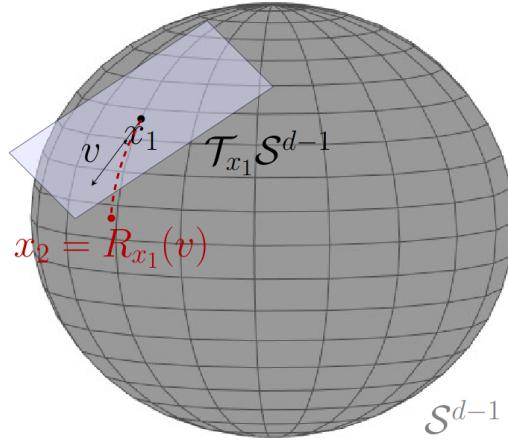


Figure 1.1: Retraction on a sphere.