Optimization on manifolds

Estelle Massart

UCLouvain

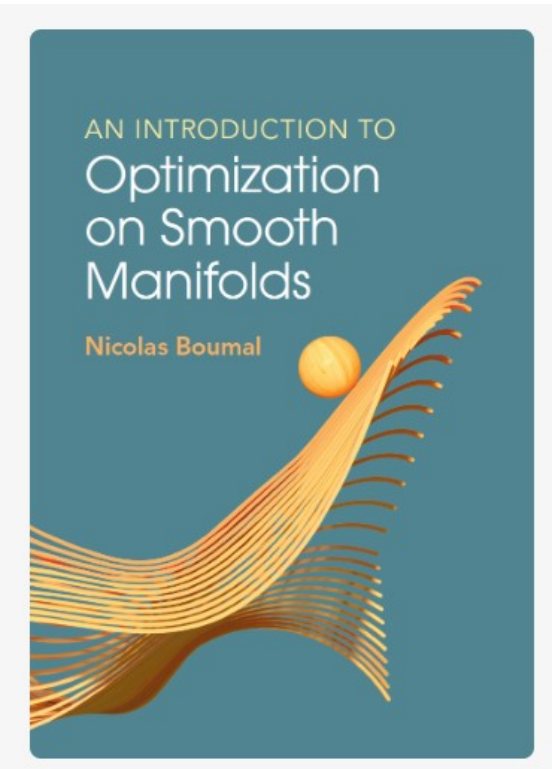
June 28, 2024

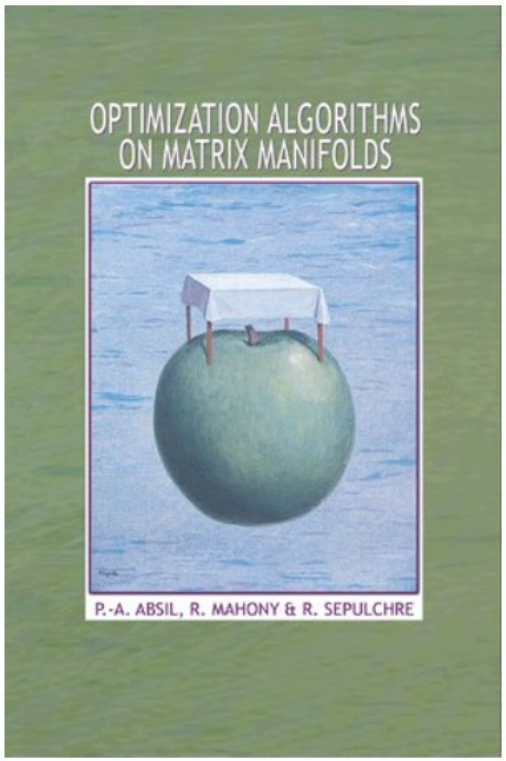
Part I: Introduction

Reference books

This course is based on the books:

- N. Boumal, *Optimization on smooth manifolds*, Cambridge University Press, 2023.
- P.-A. Absil, R. Mahony and R. Sepulchre, *Optimization algorithms on matrix manifolds*, Princeton University Press, 2008.





Optimization on manifolds

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

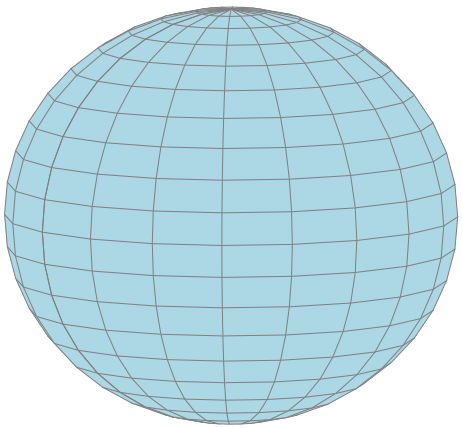
- $\mathcal{M}$ is the search space
- f is the cost function (a.k.a. objective function)

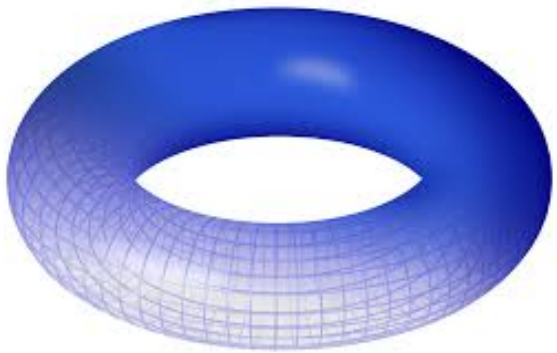
Optimization on manifolds

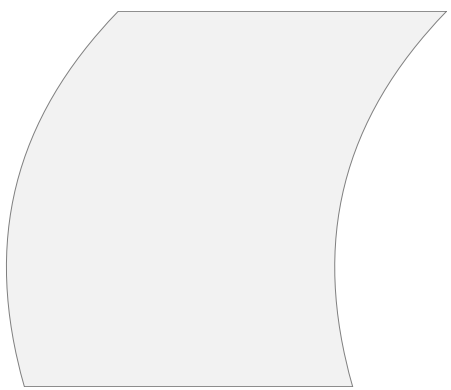
We assume that the search space and the objective function are "smooth".

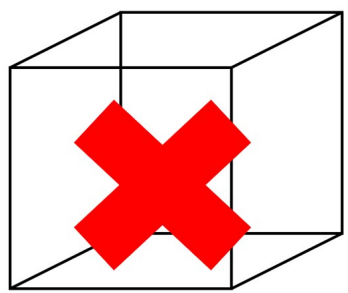
What is a manifold?

For now, just think of a manifold as a set that is well approximated locally around any point by a vector space...









Optimization on manifolds

$$\min_{x \in \mathcal{M}} f(x) \tag{1}$$

Two ways to see (P):

- A constrained optimization problem
- An unconstrained optimization problem, assuming that nothing else exists than the set $\mathcal{M}$.

Optimization on manifolds extends classical unconstrained optimization algorithms to problems whose search space is a manifold.

A 4-slide spoiler...

What do classical optimization algorithms require?

Notion of gradient, Hessian $\rightarrow$ inner product.

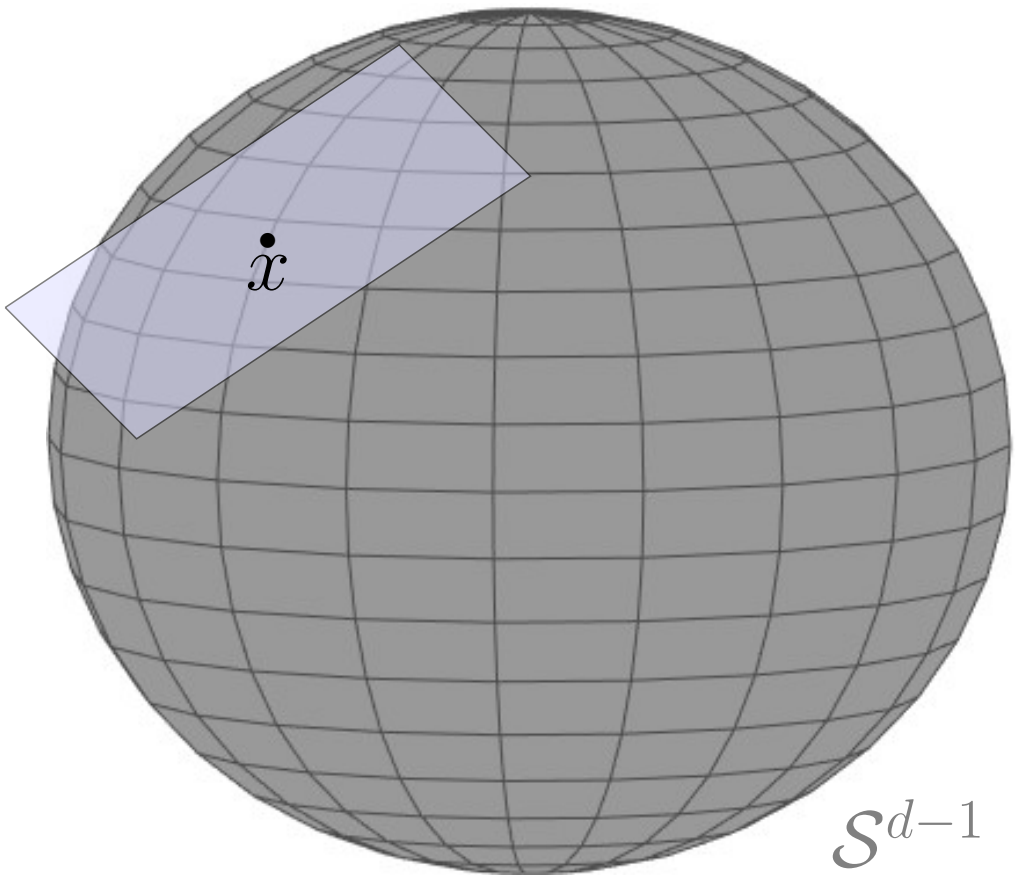
How to equip arbitrary smooth spaces with inner products?

- Manifolds can be approximated locally around any point by a vector space (tangent space)
- These vector spaces can be equipped by inner products (Riemannian metric)
- These inner products allow us to define gradients, Hessians etc. (Riemannian gradient, Riemannian Hessian, ...)

The choice of the inner product is not unique, leading to different notion of gradients, Hessians, ..., for a given set $\mathcal{M}$.

A 4-slide spoiler...

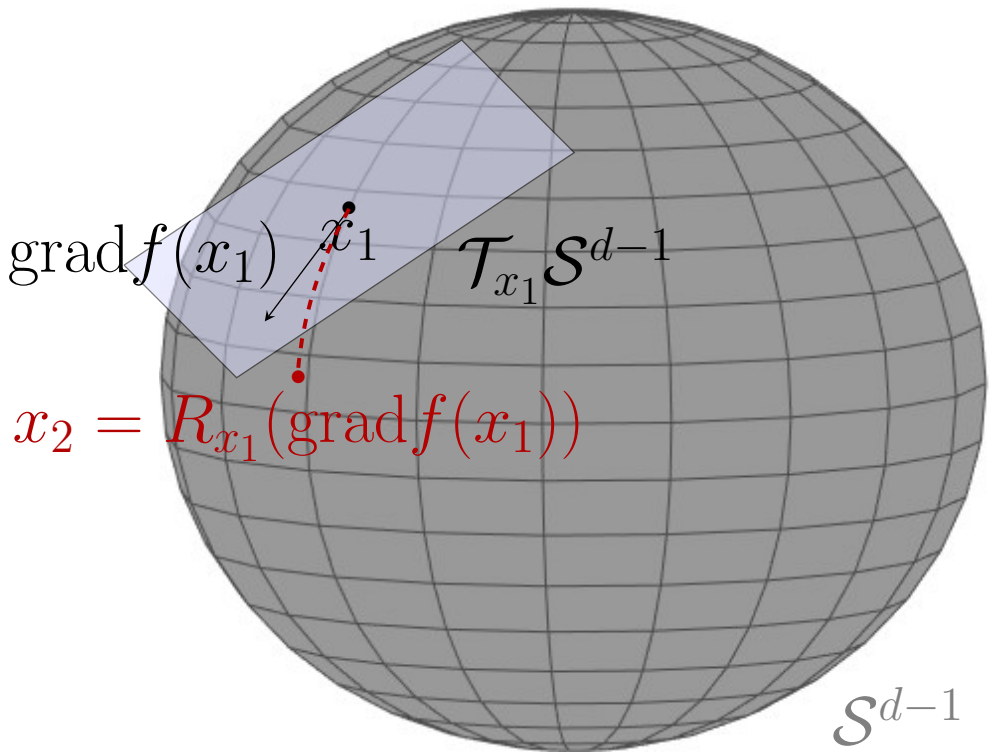
$$\mathcal{T}_x\mathcal{S}^{d-1} = \{v \in \mathbb{R}^3 : x^\top v = 0\}$$



The most often used are natural choices (e.g., simply use the usual inner product in $\mathbb{R}^2$)...

A 4-slide spoiler...

What is then an iteration of gradient descent?



A 4-slide spoiler...

What can be achieved ultimately?

- Adaptation of numerous algorithms from the Euclidean setting (Gradient descent, Newton, Conjugate gradients, BFGS, trust-region, PSO, adaptive regularization with cubics).
- General-purpose toolbox for Riemannian optimization, containing implementations for many solvers, and libraries allowing to work on many manifolds: Manopt (Matlab, Julia), Pymanopt, ROPTlib, McTorch, GeoTorch.
- Convergence and complexity results.

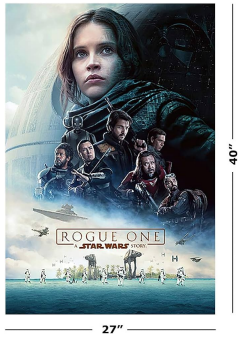
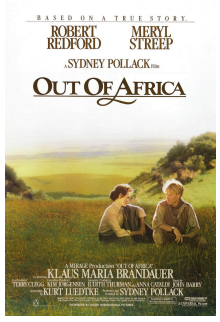
Why is this useful?

Examples of applications

Application: the Netflix problem (2006)

Famous example of recommendation system

Goal: predict user ratings for films based on previous ratings.

				
Anne	5	1	?	?
Ben	4	?	5	?
Charles	?	5	?	4
David	5	?	?	?

Dataset:

- $\simeq$ 17.000 movies
- $\simeq$ 500.000 users
- Percentage of known entries: $\simeq 1\%$.

Application: the Netflix problem (2006)

- Let $M \in \mathbb{R}^{m \times n}$ be the rating matrix
- Let $\Omega \subset \{1, \dots, m\} \times \{1, \dots, n\}$ contain the indices of known ratings.

The problem can be written as

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

Additional assumption

Assume that the matrix X has low rank (e.g., there exists k latent factors that allow making reasonable guesses).

Illustration on a simple example

Consider the following 2×2 matrix:

$$\begin{pmatrix} 1 & 5 \\ 2 & ? \end{pmatrix}$$

What is the missing entry

- if the matrix has rank two?
- if the matrix has rank one?
- if the matrix has rank zero?

Application: the Netflix problem (2006)

- Let $M \in \mathbb{R}^{m \times n}$ be the rating matrix
- Let $\Omega \subset \{1, \ldots, m\} \times \{1, \ldots, n\}$ contain the indices of known ratings.
- Let $\mathbb{R}_k^{m \times n}$ be the set of rank-k $m \times n$ matrices.

The problem can be written as

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

- The set $\mathbb{R}_k^{m \times n}$ is a "smooth set" (it is a smooth manifold)
- Its closure (the set of $m \times n$ matrices of rank at most k is not; it is a stratified space...).

Application: dictionary learning

The dictionary learning problem

Let $x_1, \ldots, x_m$ be a collection of images. The goal is to learn k atoms $b_1, \ldots, b_k$ (with $k \ll m$) such that each image x_i can be represented by a **small** number of properly chosen atoms.

We want:

$$\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix} \simeq \begin{bmatrix} & & \\ & B & \\ & & \end{bmatrix} \begin{bmatrix} & C & \end{bmatrix}$$

with

- $X := [x_1, \ldots, x_m] \in \mathbb{R}^{d \times m}$
- $B := [b_1, \ldots, b_k] \in \mathbb{R}^{d \times k}$
- $C \in \mathbb{R}^{k \times m}$ the coefficient matrix.

Application: dictionary learning

The dictionary learning problem

Let $x_1, \dots, x_m$ be a collection of images. The goal is to learn k *atoms* $b_1, \dots, b_k$ (with $k \ll m$) such that each image y_i can be represented by a **small** number of properly chosen atoms.

The problem writes:

$$\min_{\substack{B \in \mathbb{R}^{d \times k} \\ C \in \mathbb{R}^{k \times m}}} \|X - BC\|^2 + \lambda \|C\|_0$$
$$s.t.\ \|b_1\| = \|b_2\| = \cdots = \|b_k\| = 1.$$

with

- $X := [x_1, \dots, x_m] \in \mathbb{R}^{d \times m}$
- $B := [b_1, \dots, b_k] \in \mathbb{R}^{d \times k}$
- $C \in \mathbb{R}^{k \times m}$ the coefficient matrix.

Application: dictionary learning

The oblique manifold is the set

$$\begin{aligned}\mathcal{OB}(d,k) &= \{B \in \mathbb{R}^{d\times k} : \|B(:,i)\| = 1 \ \forall i\} \\ &= \{B \in \mathbb{R}^{d\times k} : \operatorname{diag}(B^\top B) = \mathbf{1}_k\}.\end{aligned}$$

The problem becomes then unconstrained:

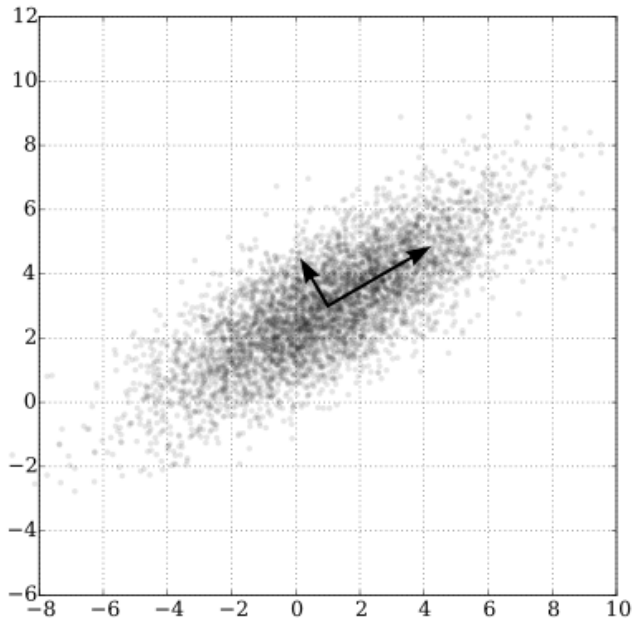
$$\min_{\substack{B\in\mathcal{OB}(d,k)\\ C\in\mathbb{R}^{k\times m}}} \|X - BC\|^2 + \lambda\|C\|_0.$$

This is an optimization problem on the product manifold $\mathcal{OB}(d,k) \times \mathbb{R}^{k\times m}$.

Principal component analysis

The principal component analysis problem

Let $x_1, \ldots, x_n$ be a centered dataset in $\mathbb{R}^d$. We aim to find a collection of k orthogonal unit-norm vectors $u_1, \ldots, u_k$ such that the (low-dimensional) subspace spanned by these vectors captures most of the variance of the initial dataset.



Principal component analysis

To find one component:

$$\begin{aligned}
\max_{\substack{u\in\mathbb{R}^d\\ \|u\|=1}} \sum_{i=1}^{n} \|uu^\top x_i\|^2 &= \max_{\substack{u\in\mathbb{R}^d\\ \|u\|=1}} \|uu^\top X\|_{\mathrm{F}}^2 \\
&= \max_{\substack{u\in\mathbb{R}^d\\ \|u\|=1}} \langle uu^\top X, uu^\top X\rangle \\
&= \max_{\substack{u\in\mathbb{R}^d\\ \|u\|=1}} \operatorname{trace}\left(X^\top uu^\top uu^\top X\right) \\
&= \max_{\substack{u\in\mathbb{R}^d\\ \|u\|=1}} \langle XX^\top u, u\rangle.
\end{aligned}$$

PCA: formulation on the Stiefel manifold

To find two components:

$$\max_{\substack{u_1,u_2\in\mathbb{R}^d\\ \|u_1\|=1\\ \|u_2\|=1\\ u_1^\top u_2=0}} \sum_{i=1}^{n} \|(u_1u_1^\top + u_2u_2^\top)x_i\|^2 = \max_{\substack{u\in\mathbb{R}^d\\ \|u_1\|=1\\ \|u_2\|=1\\ u_1^\top u_2=0}} \langle XX^\top u_1, u_1\rangle + \langle XX^\top u_2, u_2\rangle.$$

We define the Stiefel manifold

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

Then,

$$\max_{\substack{u\in\mathbb{R}^d\\ \|u_1\|=1\\ \|u_2\|=1\\ u_1^\top u_2=0}} \alpha_1\langle XX^\top u_1, u_1\rangle + \alpha_2\langle XX^\top u_2, u_2\rangle = \max_{U\in\mathrm{St}(2,d)} \langle XX^\top U, UD\rangle.$$

with α_1, α_2 weight factors and $D = \mathrm{diag}(\alpha_1, \alpha_2)$.

PCA: formulation on the Stiefel manifold

- It is well-known that this optimization problem is solved by taking the leading eigenvectors of $XX^\top$.
- The optimization formulation allows to explore variants, such as sparse PCA, robust PCA, or situations where the dataset changes or grows with time.
- In somes cases, the order of the principal components does not matter, and we only seek a basis of the k-dimensional principal subspace of $XX^\top$...

PCA: formulation on the Grassmann manifold

Note that the cost function

$$f(U) = \langle XX^\top U, U \rangle$$

is invariant under orthogonal transformations, as $\forall Q \in \mathcal{O}(k)$, with

$$\mathcal{O}(k) = \{Q \in \mathbb{R}^{k \times k} : Q^\top Q = I_k\},$$

there holds

$$f(UQ) = f(U).$$

Reminder: equivalence

This induces an equivalence relation $\sim$ on the Stiefel manifold:

$$U \sim V \Leftrightarrow V = UQ \quad \text{for some} \quad Q \in \mathcal{O}(k).$$

Reminder: an equivalence relation $\sim$ on a set $\mathcal{M}$ is a

- reflexive ($a \sim a$),
- symmetric ($a \sim b \Leftrightarrow b \sim a$), and
- transitive ($a \sim b$ and $b \sim c \Leftrightarrow a \sim c$)

binary relation.

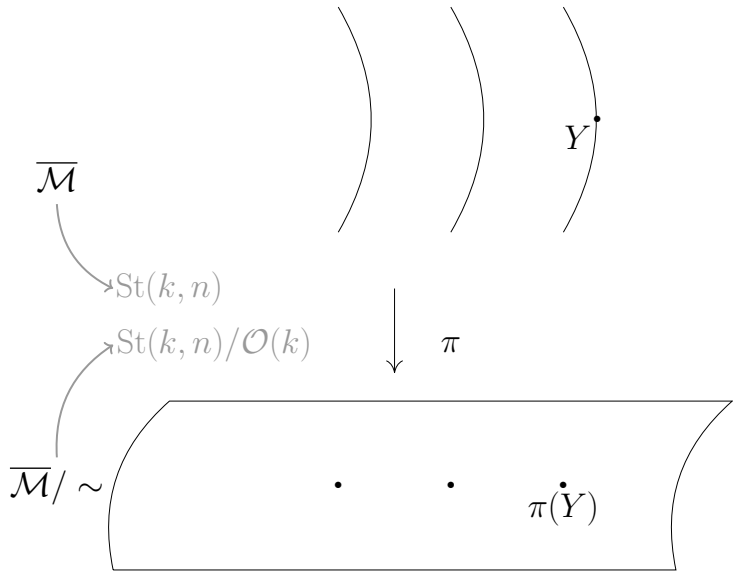
This equivalence relation partitions $\mathrm{St}(k, d)$ into equivalence classes:

$$[U] = \{V \in \mathrm{St}(k, d) : U \sim V\} = \{UQ : Q \in \mathcal{O}(k)\}.$$

PCA: formulation on the Grassmann manifold

The set of equivalence classes is called the quotient set:

$$\mathrm{St}(k,d)/\sim \;=\; \mathrm{St}(k,d)/\mathcal{O}(k) = \{[U] : U \in \mathrm{St}(k,d)\}.$$



With the right geometry, the latter is called the Grassmann manifold:

$$\mathrm{Gr}(k,d) = \{\text{subspaces of dimension } k \text{ in } \mathbb{R}^d\} \equiv \mathrm{St}(k,d)/\mathcal{O}(k).$$

Other applications involving manifolds

- The Stiefel manifold: Rotation synchronization (Boumal13), training in DNNs (Wisdom16, Lezcano-Casado19, Massart20), ...
- The fixed-rank manifold: Low-rank matrix and tensor completion (Vandereycken12), ...
- The Grassmann manifold: Matrix completion (Boumal15), system identification (Usevich14), ...
- The manifold of fixed-rank PSD matrices: Distance matrix learning (Meyer11), Role Model Extraction in graphs (Marchand16), computation of low-rank solutions to Lyapunov equations (Vandereycken10), ...

References

The main reference for this presentation is "N. Boumal, An introduction to optimization on smooth manifolds, Cambridge University Press, 2023" (Chapter 2).
Other references:

- Absil08: P.-A. Absil, R. Mahony, R. Sepulchre, *Optimization algorithms on matrix manifolds*, 2008.
- Boumal13: N. Boumal, A. Singer, P.-A. Absil, *Robust estimation of rotations from relative measurements by maximum likelihood*, CDC, 2013.
- Boumal15: N. Boumal, P.-A. Absil, *Low-rank matrix completion via preconditioned optimization on the Grassmann manifold*, Lin. Alg. and App. 475, 200–239, 2015.
- Lezcano-Casado19: *Cheap Orthogonal Constraints in Neural Networks: A Simple Parametrization of the Orthogonal and Unitary Group.*
- Marchand16: M. Marchand, et al, *A Riemannian Optimization Approach for Role Model Extraction*, Proceedings of MTNS, 2016.
- Massart20: E. Massart, V. Abrol, *Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices*, SIAM J. Matrix Anal. Appl., 41(1), 171–198.

- Meyer11: G. Meyer, S. Bonnabel, R. Sepulchre, *Regression on Fixed-Rank Positive Semidefinite Matrices: A Riemannian Approach*, Journal of Machine Learning Research 12, 593-625, 2011.
- Usevich14: K. Usevich, I. Markovski, *Optimization on a Grassmann manifold with application to system identification*, Automatica 50(6), 1656-1662, 2014.
- Vandereycken10: B. Vandereycken, S. Vandewalle, *A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations*, SIAM J. Matrix Anal. Appl., 31(5), 2553–2579, 2010.
- Vandereycken12: B. Vandereycken, *Low-rank matrix completion by Riemannian optimization—extended version*, arxiv preprint 2012.
- Vandereycken13: B. Vandereycken, P.-A. Absil, S. Vandewalle, *A Riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank*, IMA Journal of Numerical Analysis 33(2), 2013.
- Wisdom16: S. Wisdom et al, *Full-Capacity Unitary Recurrent Neural Networks*, NIPS2016.

Part II: Embedded submanifolds

Optimization on manifolds

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

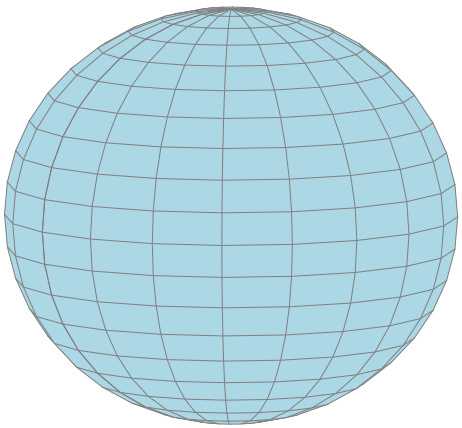
- $\mathcal{M}$ is the search space
- f is the cost function (a.k.a. objective function)

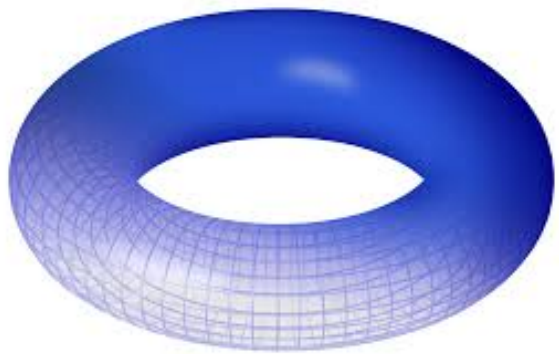
Optimization on manifolds

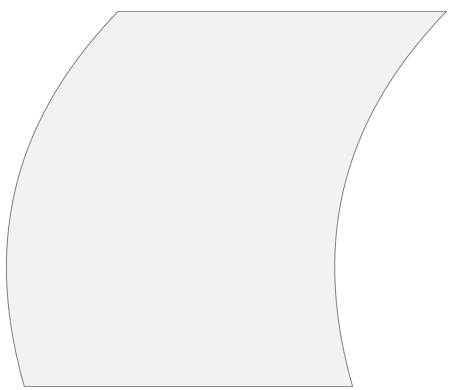
We assume that the search space and the objective function are "smooth".

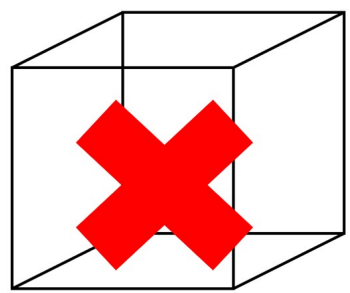
What is a manifold?

For now, just think of a manifold as a set that is well approximated locally around any point by a vector space...









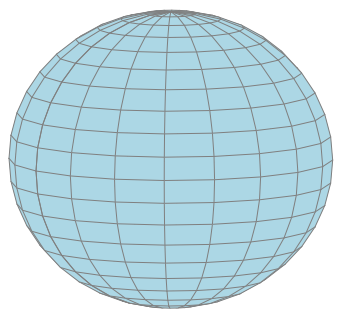
Goals of this section

Define the main geometric tools required to extend Euclidean optimization algorithms to manifolds, namely:

- Embedded submanifolds
- Tangent spaces
- Smoothness and differential of mapping between manifolds
- Retraction
- Metrics on manifolds
- Riemannian gradient
- Taylor development of functions defined on manifolds

Focus here: embedded submanifolds of $\mathbb{R}^D$.

Example:



Embedded submanifolds

Tangent spaces

Smoothness and differential of mapping between manifolds

Retraction

Metrics on manifolds

Riemannian gradient

Taylor development of functions on manifolds

The sphere is a manifold

Let

$$\mathcal{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|^2 = 1\} = \{x \in \mathbb{R}^d : x^\top x = 1\}.$$

The set $\mathcal{S}^{d-1}$ is thus defined by the constraint

$$h(x) = 0,$$

for $h(x) := x^\top x - 1$.
Differentiating the constraint gives us a linearization of $\mathcal{S}^{d-1}$. Let $x \in \mathcal{S}^{d-1}$ and $v \in \mathbb{R}^d$. Then

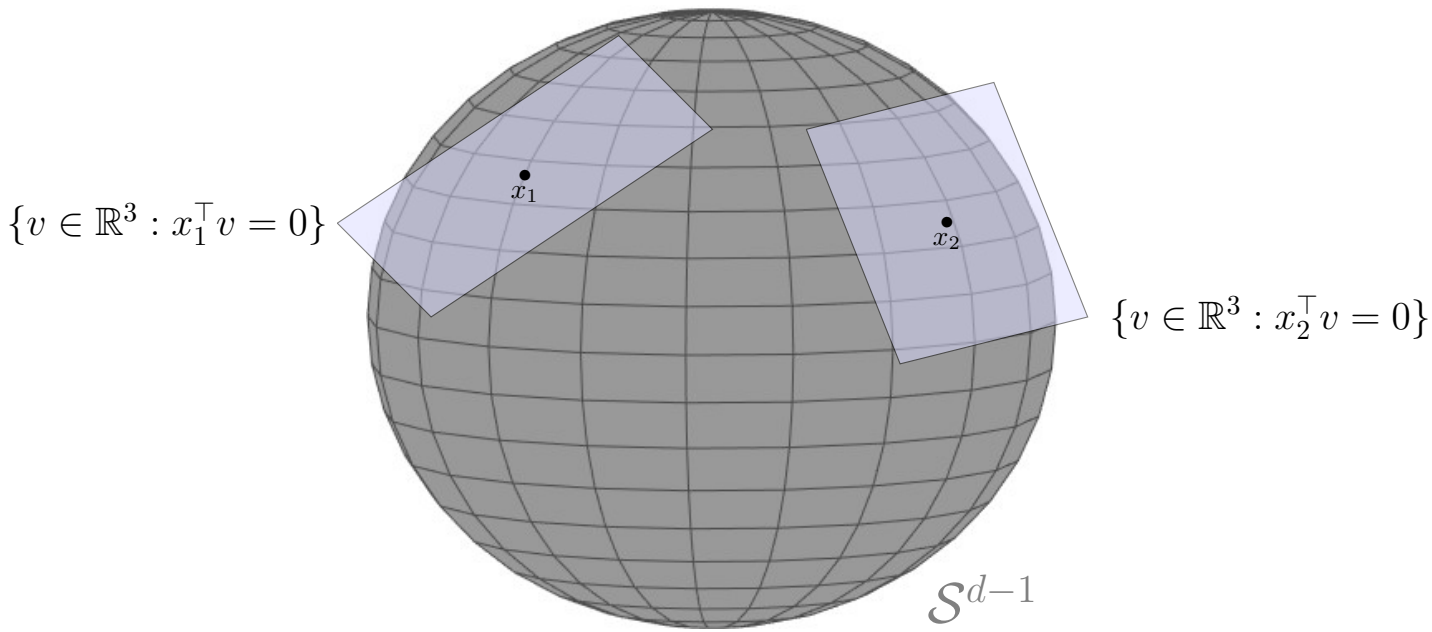
$$\begin{aligned} h(x + tv) &= h(x) + tDh(x)[v] + \mathcal{O}(t^2) \\ &= h(x) + t(v^\top x + x^\top v) + \mathcal{O}(t^2). \end{aligned}$$

Tangent space of the sphere

Around any point $x \in \mathcal{S}^{d-1}$, the sphere can be locally approximated by the set

$$\{v \in \mathbb{R}^d : x^\top v + v^\top x = 2x^\top v = 0\} = \{v \in \mathbb{R}^d : x^\top v = 0\}.$$

Graphically:



The function $h : \mathbb{R}^3 \to \mathbb{R}$ is called a local defining function of the manifold $\mathcal{S}^{d-1}$.

Is the existence of a smooth defining function sufficient for the set to be a manifold?

Counter-example:

$$\mathcal{X} = \{x \in \mathbb{R}^2 : x_1^2 - x_2^2 = 0\}.$$

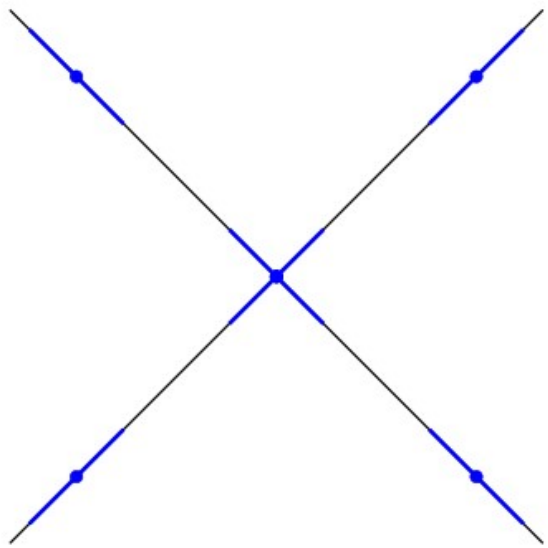


Figure: Figure courtesy of *Boumal (2023)*.

Is the existence of a smooth defining function sufficient for the set to be a manifold?

Counter-example:

$$\mathcal{X} = \{x \in \mathbb{R}^2 : x_1^2 - x_2^2 = 0\}.$$

We get:

$$\mathrm{D}h(x) = \left[\frac{\partial h}{\partial x_1}(x), \frac{\partial h}{\partial x_2}(x)\right] = \left[2x_1, -2x_2\right].$$

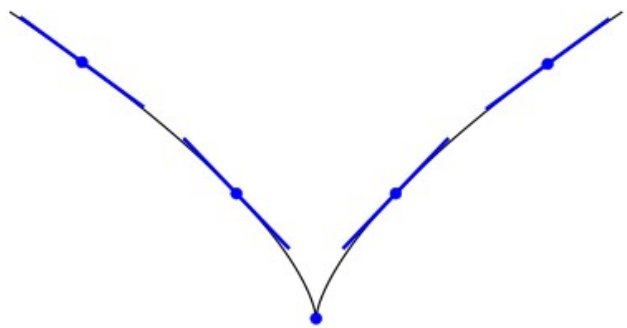
There follows that

- $\dim(\ker(\mathrm{D}h(x))) = 1 \qquad \forall x \in \mathbb{R}^2 \setminus (0,0),$
- $\dim(\ker(\mathrm{D}h(x))) = 2 \qquad \text{if } x = (0,0).$

Other counter-example

The set

$$\mathcal{X} = \{x \in \mathbb{R}^2 : x_1^2 - x_2^3 = 0\}$$



Cusp: $x_1^2 - x_2^3 = 0$

Figure: Figure courtesy of *Boumal (2023)*.

Other counter-example

The set

$$\mathcal{X} = \{x \in \mathbb{R}^2 : x_1^2 - x_2^4 = 0\}$$

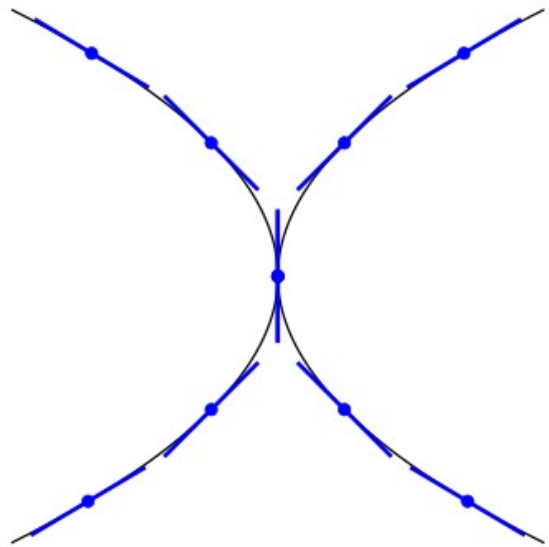


Figure: Figure courtesy of *Boumal (2023)*.

Embedded submanifold: definition

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

- $\mathcal{M} \cap \mathcal{U} = h^{-1}(0) := \{y \in \mathcal{U} : h(y) = 0\}$; and
- $\mathrm{rank}(\mathrm{D}h(x)) = k$.

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

Exercise

Can you think about a local defining function for the manifold of orthogonal matrices

$$\mathcal{O}(n) = \{X \in \mathbb{R}^{n \times n} : X^\top X = I_n\}?$$

Exercise

Can you think about a local defining function for the manifold of orthogonal matrices

$$\mathcal{O}(n) = \{X \in \mathbb{R}^{n \times n} : X^\top X = I_n\}?$$

Solution:

$$h : \mathbb{R}^{n \times n} \to \mathrm{Sym}(n) : X \mapsto X^\top X - I_n$$

Solution:

$$h : \mathbb{R}^{n \times n} \to \mathrm{Sym}(n) : X \mapsto X^\top X - I_n$$

Why?

- What are $\mathcal{U}$ and k?
- $h : \mathbb{R}^{n \times n} \to \mathbb{R}^k$ is smooth?
- $h^{-1}(0) =$?
- $\mathrm{D}h(X)[V] = V^\top X + X^\top V$ has rank k for all $X \in \mathcal{O}(n)$?

Solution:

$$h : \mathbb{R}^{n\times n} \to \mathrm{Sym}(n) : X \mapsto X^\top X - I_n$$

Why?

- What are $\mathcal{U}$ and k?
  Take $\mathcal{U} = \mathbb{R}^{n\times n}$, identify $\mathbb{R}^k$ to the vector space $\mathrm{Sym}(n)$ (i.e., take $k = n(n+1)/2$).
- $h : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ is smooth?
- $h^{-1}(0) =$?
- $\mathrm{D}h(X)[V] = V^\top X + X^\top V$ has rank k for all $X \in \mathcal{O}(n)$?

Solution:

$$h : \mathbb{R}^{n\times n} \to \mathrm{Sym}(n) : X \mapsto X^\top X - I_n$$

Why?

- What are $\mathcal{U}$ and k?
  Take $\mathcal{U} = \mathbb{R}^{n\times n}$, identify $\mathbb{R}^k$ to the vector space $\mathrm{Sym}(n)$ (i.e., take $k = n(n+1)/2$).
- $h : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ is smooth? Yes
- $h^{-1}(0) =$?
- $\mathrm{D}h(X)[V] = V^\top X + X^\top V$ has rank k for all $X \in \mathcal{O}(n)$?

Solution:

$$h : \mathbb{R}^{n\times n} \to \mathrm{Sym}(n) : X \mapsto X^\top X - I_n$$

Why?

- What are $\mathcal{U}$ and k?
  Take $\mathcal{U} = \mathbb{R}^{n\times n}$, identify $\mathbb{R}^k$ to the vector space $\mathrm{Sym}(n)$ (i.e., take $k = n(n+1)/2$).
- $h : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ is smooth? Yes
- $h^{-1}(0) =?$ $h^{-1}(0) = \mathcal{O}(n) \cap \mathcal{U} = \mathcal{O}(n)$
- $\mathrm{D}h(X)[V] = V^\top X + X^\top V$ has rank k for all $X \in \mathcal{O}(n)$?

Solution:

$$h : \mathbb{R}^{n\times n} \to \mathrm{Sym}(n) : X \mapsto X^\top X - I_n$$

Why?

- What are $\mathcal{U}$ and k?
  Take $\mathcal{U} = \mathbb{R}^{n\times n}$, identify $\mathbb{R}^k$ to the vector space $\mathrm{Sym}(n)$ (i.e., take $k = n(n+1)/2$).
- $h : \mathbb{R}^{n\times n} \to \mathbb{R}^k$ is smooth? Yes
- $h^{-1}(0) =?$ $h^{-1}(0) = \mathcal{O}(n) \cap \mathcal{U} = \mathcal{O}(n)$
- $\mathrm{D}h(X)[V] = V^\top X + X^\top V$ has rank k for all $X \in \mathcal{O}(n)$?
  The last property holds if $\mathrm{im}(\mathrm{D}h(X))$ contains a subspace of dimension k... Let $A \in \mathrm{Sym}(n)$ be arbitrary. Take $V = XA$. Then,

$$(XA)^\top X + X^\top XA = AX^\top X + X^\top XA = 2A.$$

The manifold of orthogonal matrices: conclusions

The set

$$\mathcal{O}(n) = \{X \in \mathbb{R}^{n\times n} : X^\top X = I_n\}$$

is thus an embedded submanifold of $\mathbb{R}^{n\times n}$, of dimension

$$d - k = n^2 - \frac{n(n+1)}{2} = n^2 - \frac{n^2}{2} - \frac{n}{2} = \frac{n(n-1)}{2}.$$

Embedded submanifolds

Tangent spaces

Smoothness and differential of mapping between manifolds

Retraction

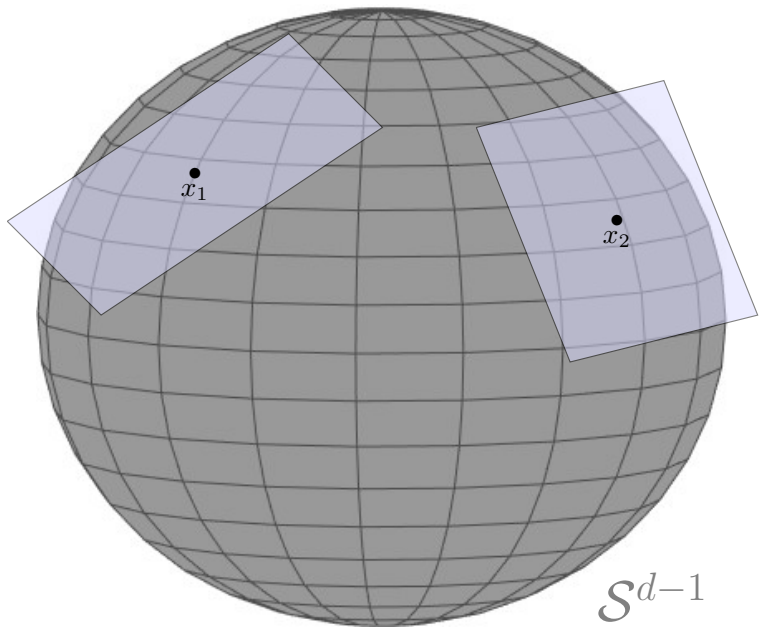
Metrics on manifolds

Riemannian gradient

Taylor development of functions on manifolds

Tangent spaces

$\mathcal{T}_{x_1}\mathcal{S}^{d-1} = \{v \in \mathbb{R}^3 : x_1^\top v = 0\}$



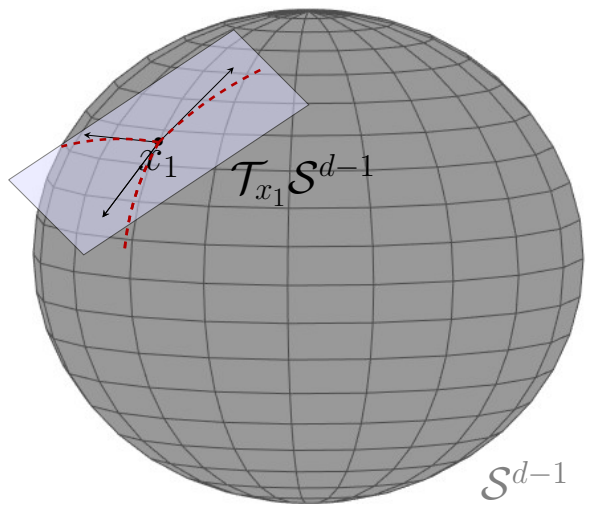
$\mathcal{T}_{x_2}\mathcal{S}^{d-1} = \{v \in \mathbb{R}^3 : x_2^\top v = 0\}$

Tangent space: definition

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

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

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



In this definition, $c'(0)$ is the derivative of the curve $c : \mathcal{I} \to \mathcal{E}$, which is defined in the usual sense...

Tangent space: how to compute it?

Let $\mathcal{M}$ be an embedded submanifold of $\mathcal{E}$. Then, for all $x \in \mathcal{M}$, there holds

$$\mathcal{T}_x\mathcal{M} = \ker(\mathrm{D}h(x)),$$

with h any local defining function at x.

Exercise

What is the tangent space at X of the orthogonal group

$$\mathcal{O}(n) = \{X \in \mathbb{R}^{n \times n} : X^\top X = I_n\}?$$

Choose among the following possibilities:

a. $\mathcal{T}_X\mathcal{O}(n) = \{\eta \in \mathbb{R}^{n \times n} : \eta^\top X = 0_{n \times n}\}$

b. $\mathcal{T}_X\mathcal{O}(n) = \{\eta \in \mathbb{R}^{n \times n} : \eta^\top X + X^\top \eta = 0_{n \times n}\}$

c. $\mathcal{T}_X\mathcal{O}(n) = \{\eta \in \mathbb{R}^{n \times n} : \eta = X\Omega, \Omega^\top = -\Omega\}.$

Exercise

What is the tangent space at X of the orthogonal group

$$\mathcal{O}(n) = \{X \in \mathbb{R}^{n\times n} : X^\top X = I_n\}?$$

Choose among the following possibilities:

a. $\mathcal{T}_X\mathcal{O}(n) = \{\eta \in \mathbb{R}^{n\times n} : \eta^\top X = 0_{n\times n}\}$

b. $\mathcal{T}_X\mathcal{O}(n) = \{\eta \in \mathbb{R}^{n\times n} : \eta^\top X + X^\top \eta = 0_{n\times n}\}$

c. $\mathcal{T}_X\mathcal{O}(n) = \{\eta \in \mathbb{R}^{n\times n} : \eta = X\Omega, \Omega^\top = -\Omega\}$.

Note that b) and c) are equivalent. Since X is invertible, we can write $\eta = XM$ for some $M \in \mathbb{R}^{n\times n}$. Then,

$$\eta^\top X + X^\top \eta = 0_{n\times n} \Leftrightarrow M^\top X^\top X + X^\top XM = M^\top + M = 0_{n\times n}$$
$$\Leftrightarrow M^\top = -M.$$

Two more definitions: Tangent bundle and vector field

The tangent bundle of a manifold $\mathcal{M}$ is the disjoint union of the tangent spaces of $\mathcal{M}$:

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

A vector field on a manifold $\mathcal{M}$ is a map $V : \mathcal{M} \to \mathcal{T}\mathcal{M}$ such that

$$V(x) \in \mathcal{T}_x\mathcal{M} \qquad \forall x \in \mathcal{M}.$$

- If V is a smooth map, we say it is a smooth vector field.
- The set of smooth vector fields on $\mathcal{M}$ is denoted by $\Xi(\mathcal{M})$.

Embedded submanifolds

Tangent spaces

Smoothness and differential of mapping between manifolds

Retraction

Metrics on manifolds

Riemannian gradient

Taylor development of functions on manifolds

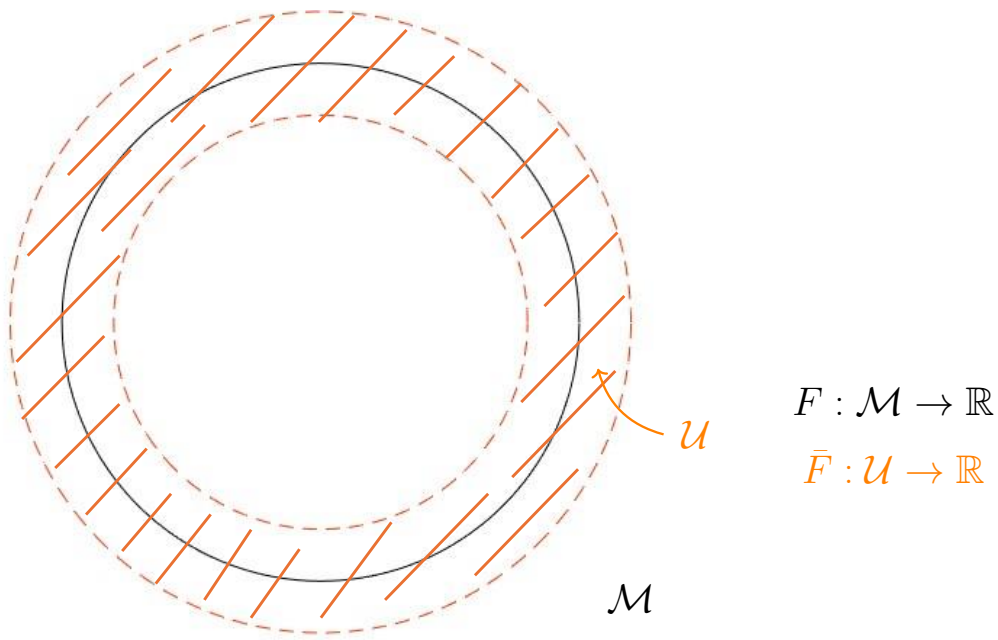
Smoothness of a map between two manifolds

Let $\mathcal{M}$ and $\mathcal{M}'$ be embedded submanifolds of $\mathcal{E}$ and $\mathcal{E}'$ respectively.

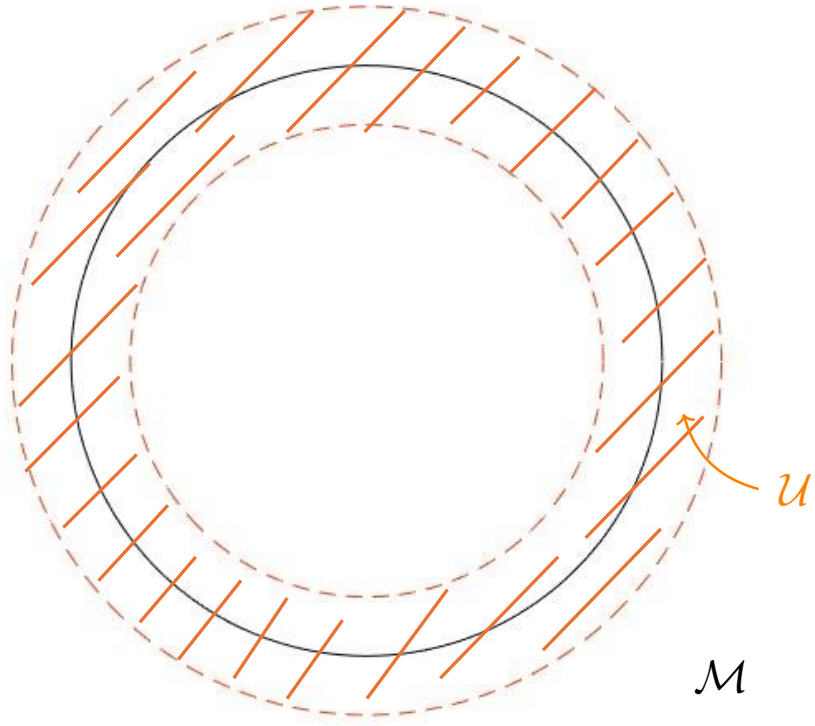
Smoothness of a map

A map $F : \mathcal{M} \to \mathcal{M}'$ is smooth at $x \in \mathcal{M}$ if there exists a function $\bar{F} : \mathcal{U} \to \mathcal{E}'$ which is smooth on a neighborhood $\mathcal{U} \subseteq \mathcal{E}$ of x and such that F and $\bar{F}$ coincide on $\mathcal{M} \cap \mathcal{U}$, that is,

$$F(y) = \bar{F}(y) \qquad \forall y \in \mathcal{M} \cap \mathcal{U}.$$



Smoothness of a map between two manifolds



$F : \mathcal{M} \to \mathbb{R}$

$\bar{F} : \mathcal{U} \to \mathbb{R}$

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

Exercise

Smoothness of a map

A map $F : \mathcal{M} \to \mathcal{M}'$ is smooth at $x \in \mathcal{M}$ if there exists a function $\bar{F} : \mathcal{U} \to \mathcal{E}'$ which is smooth on a neighborhood $\mathcal{U} \subseteq \mathcal{E}$ of x and such that F and $\bar{F}$ coincide on $\mathcal{M} \cap \mathcal{U}$, that is,

$$F(y) = \bar{F}(y) \qquad \forall y \in \mathcal{M} \cap \mathcal{U}.$$

Is the Rayleigh quotient of an arbitrary matrix $A \in \mathbb{R}^{n \times n}$,

$$F : \mathcal{S}^{d-1} \to \mathbb{R} : x \mapsto x^\top A x,$$

a smooth function on the sphere?

Exercise

Smoothness of a map

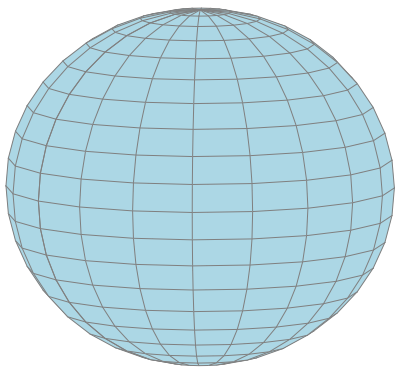
A map $F : \mathcal{M} \to \mathcal{M}'$ is smooth at $x \in \mathcal{M}$ if there exists a function $\bar{F} : \mathcal{U} \to \mathcal{E}'$ which is smooth on a neighborhood $\mathcal{U} \subseteq \mathcal{E}$ of x and such that F and $\bar{F}$ coincide on $\mathcal{M} \cap \mathcal{U}$, that is,

$$F(y) = \bar{F}(y) \qquad \forall y \in \mathcal{M} \cap \mathcal{U}.$$

Is the Rayleigh quotient of an arbitrary matrix $A \in \mathbb{R}^{n \times n}$,

$$F : \mathcal{S}^{d-1} \to \mathbb{R} : x \mapsto x^\top A x,$$

a smooth function on the sphere?
Solution:



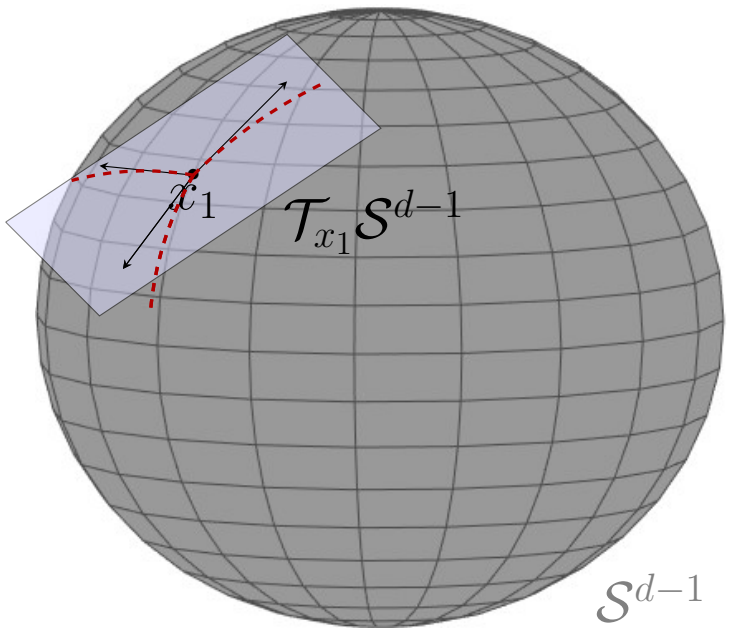
- $\mathcal{M} = \mathcal{S}^{d-1}$, $\mathcal{M}' = \mathbb{R}$.
- $\mathcal{E} = \mathbb{R}^d$, $\mathcal{E}' = \mathbb{R}$
- $\bar{F} : \mathbb{R}^d \to \mathbb{R} : x \mapsto x^\top A x$
- $\bar{F}$ smooth $\Rightarrow$ F smooth.

Differential of a smooth real-valued function on a manifold

The differential of $f : \mathcal{M} \to \mathbb{R}$ at the point $x \in \mathcal{M}$ is the linear map $Df(x) : \mathcal{T}_x\mathcal{M} \to \mathbb{R}$ defined by:

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

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



Embedded submanifolds

Tangent spaces

Smoothness and differential of mapping between manifolds

Retraction

Metrics on manifolds

Riemannian gradient

Taylor development of functions on manifolds

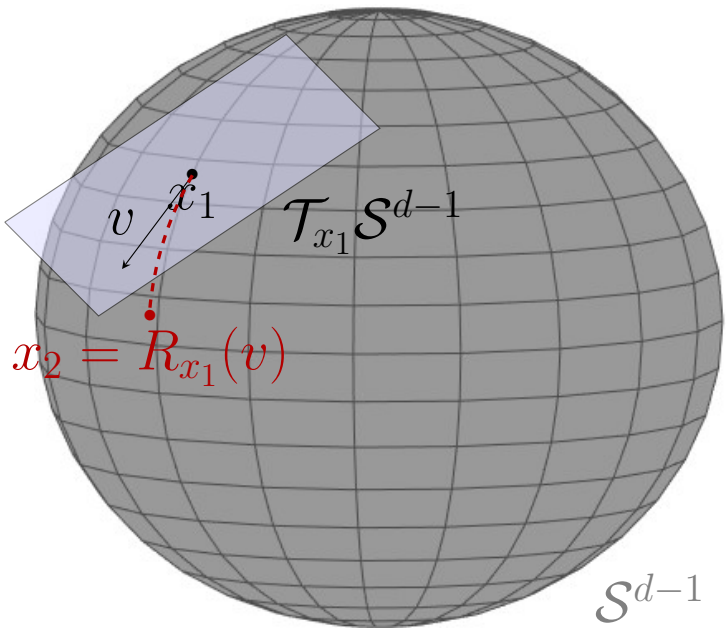
Retraction

How to move on the manifold in a given direction?

A retraction on a manifold $\mathcal{M}$ is a smooth map

$$R : \mathcal{TM} \to \mathcal{M} : (x, v) \mapsto R_x(v)$$

such that each curve $c(t) = R_x(tv)$ satisfies $c(0) = x$ and $c'(0) = v$.



Exercise

Which of the following functions are retractions on the sphere $\mathcal{S}^{d-1}$?

- $R_x(tv) := x + tv$?
- $R_x(tv) := \dfrac{x + tv}{\|x + tv\|}$?

Exercise

Which of the following functions are retractions on the sphere $\mathcal{S}^{d-1}$?

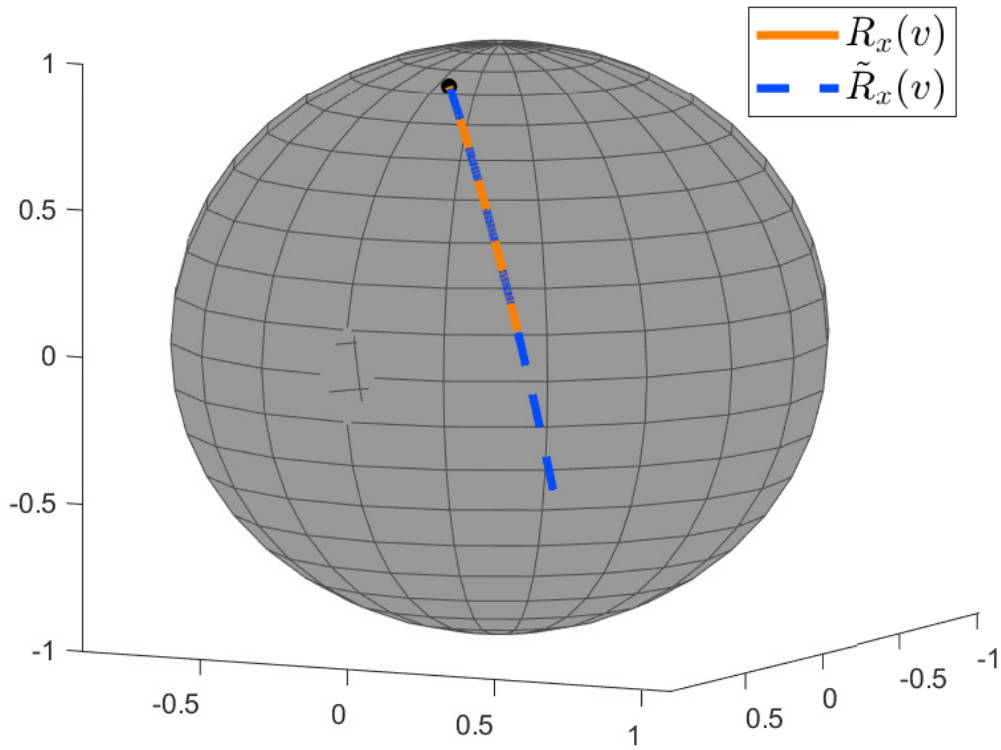
- $R_x(tv) := x + tv$?
  No: $R_x(tv) \notin \mathcal{S}^{d-1}$ if $tv \neq 0$
- $R_x(tv) := \dfrac{x + tv}{\|x + tv\|}$?

Exercise

Which of the following functions are retractions on the sphere $\mathcal{S}^{d-1}$?

- $R_x(tv) := x + tv$?
  No: $R_x(tv) \notin \mathcal{S}^{d-1}$ if $tv \neq 0$
- $R_x(tv) := \dfrac{x + tv}{\|x + tv\|}$?
  Yes:
  - $R_x(tv) \in \mathcal{S}^{d-1}\ \forall t$
  - Let us rewrite $R_x(tv) := \dfrac{x + tv}{\|x + tv\|} = \dfrac{x + tv}{\sqrt{1 + t^2\|v\|^2}}$
  - For $c(t) := R_x(tv)$, there holds $c(0) = x$
  - For $c(t) := R_x(tv)$, there holds
    $$c'(0) = \frac{v\sqrt{1 + t^2\|v\|^2} - (x + tv)t\|v\|^2}{1 + t^2\|v\|^2}\Big|_{t=0} = v.$$

Example: retraction on the sphere



Comparison between the two retractions $R_x(v) := \dfrac{x+v}{\|x+v\|}$ and $R_x(v) := \cos(\|v\|)x + \dfrac{\sin(\|v\|)}{\|v\|}v$, with the usual convention $\sin(0)/0 = 1$.

Example: retraction on the orthogonal group

Reminder:

$$\mathcal{T}_X\mathcal{O}(n) = \{X\Omega : \Omega = -\Omega^\top\}.$$

Several possibilities exist:

- Exponential map (geodesic):

$$R_X(X\Omega) = X\mathrm{expm}(\Omega)$$

- QR retraction

$$R_X(X\Omega) = Qf(X + X\Omega),$$

where $Qf(A)$ is the orthogonal factor of the QR decomposition of A.

- Cayley retraction

$$R_X(X\Omega) = X(I - \frac{1}{2}\Omega)^{-1}(I + \frac{1}{2}\Omega)$$

Retraction and directional derivative

In particular, for any tangent vector $v \in \mathcal{T}_x\mathcal{M}$, and any smooth function $f : \mathcal{M} \to \mathbb{R}$, the directional derivative $Df(x)[v]$ can be written as:

$$Df(x)[v] = \lim_{t \to 0} \frac{f(R_x(tv)) - f(x)}{t}.$$

Embedded submanifolds

Tangent spaces

Smoothness and differential of mapping between manifolds

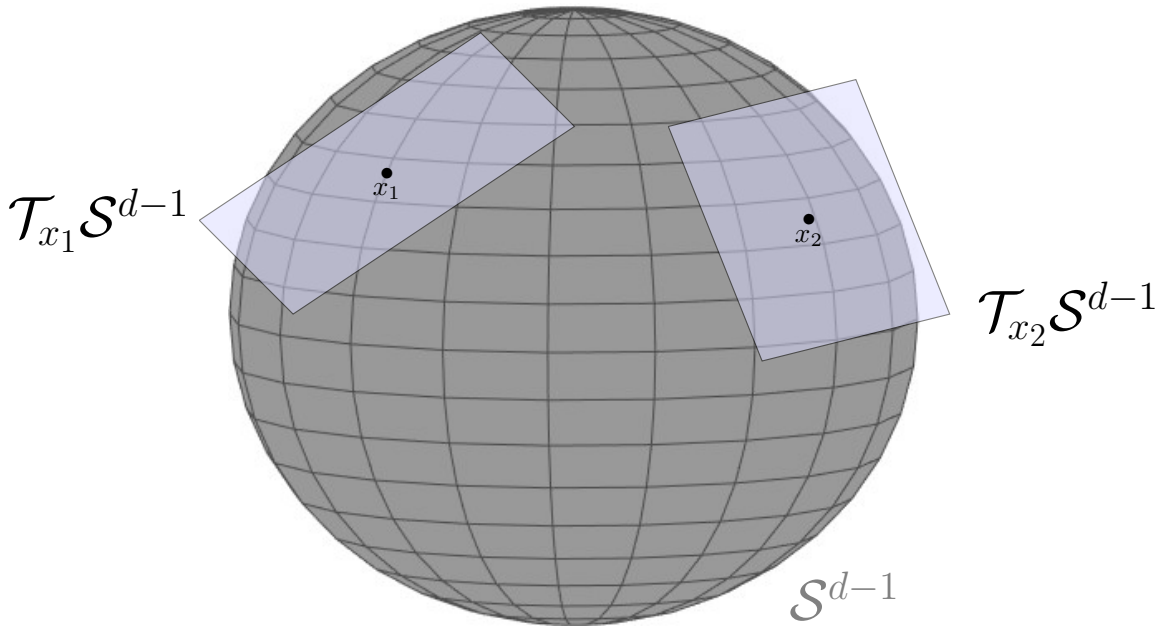
Retraction

Metrics on manifolds

Riemannian gradient

Taylor development of functions on manifolds

Metrics on manifolds



- To allow the definition of gradients and Hessians, tangent spaces must be endowed with inner products.
- One possibility: use the usual Euclidean inner product

$$\langle u, v \rangle = u^\top v \qquad \forall u, v \in \mathcal{T}_x \mathcal{S}^{d-1}.$$

Metrics on manifolds

Metric

An inner product on $\mathcal{T}_x\mathcal{M}$ is a bilinear, symmetric, positive definite function

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

It induces a norm for tangent vectors:

$$\|u\|_x^2 = \langle u, u \rangle_x.$$

A metric on $\mathcal{M}$ is a choice of inner product $\langle \cdot, \cdot \rangle_x$ for each $x \in \mathcal{M}$.

Riemannian metric

A metric $\langle \cdot, \cdot \rangle_x$ on $\mathcal{M}$ is a Riemannian metric if it varies smoothly with x, in the sense that for all smooth vector fields V, W on $\mathcal{M}$ the function $x \mapsto \langle V(x), W(x) \rangle_x$ is smooth from $\mathcal{M}$ to $\mathbb{R}$.

A Riemannian manifold is a manifold with a Riemannian metric.

Exercise

Are the following metrics on the sphere? And Riemannian metrics?

- $\langle u, v \rangle_x = 2u^\top v$?
- $\langle u, v \rangle_x = \arccos(u^\top v)$?
- $\langle u, v \rangle_x = x_1(u^\top v)$?

Exercise

Are the following metrics on the sphere? And Riemannian metrics?

- $\langle u, v \rangle_x = 2u^\top v$?
  Yes.
- $\langle u, v \rangle_x = \arccos(u^\top v)$?
- $\langle u, v \rangle_x = x_1(u^\top v)$?

Exercise

Are the following metrics on the sphere? And Riemannian metrics?

- $\langle u, v \rangle_x = 2u^\top v$?
  Yes.
- $\langle u, v \rangle_x = \arccos(u^\top v)$?
  No (not bilinear)
- $\langle u, v \rangle_x = x_1(u^\top v)$?

Exercise

Are the following metrics on the sphere? And Riemannian metrics?

- $\langle u, v \rangle_x = 2u^\top v$?
  Yes.
- $\langle u, v \rangle_x = \arccos(u^\top v)$?
  No (not bilinear)
- $\langle u, v \rangle_x = x_1(u^\top v)$?
  No (not positive definite for all $x \in \mathcal{S}^{d-1}$: take, e.g., $x = (0, 1)$)

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

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

The Riemannian distance is then defined as

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

where the infimum is taken over all (smooth) curves on $\mathcal{M}$ which connect x to y.

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 $\mathcal{E}$.

Example: Let

$$\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\},$$

which is an embedded submanifold of $\mathbb{R}^d$.
With the inherited metric $\langle u, v\rangle_x = \langle u, v\rangle = u^\top v$ on each tangent space $\mathcal{T}_x\mathcal{S}^{d-1}$, the sphere becomes a Riemannian submanifold of $\mathbb{R}^d$.

Embedded submanifolds

Tangent spaces

Smoothness and differential of mapping between manifolds

Retraction

Metrics on manifolds

Riemannian gradient

Taylor development of functions on manifolds

Riemannian gradient

In the Euclidean setting, for any smooth function $f : \mathbb{R}^d \to \mathbb{R}$, there holds:

$$Df(x)[\eta] = \langle \nabla f(x), \eta \rangle \qquad \forall x, \eta \in \mathbb{R}^n.$$

Definition: Riemannian gradient

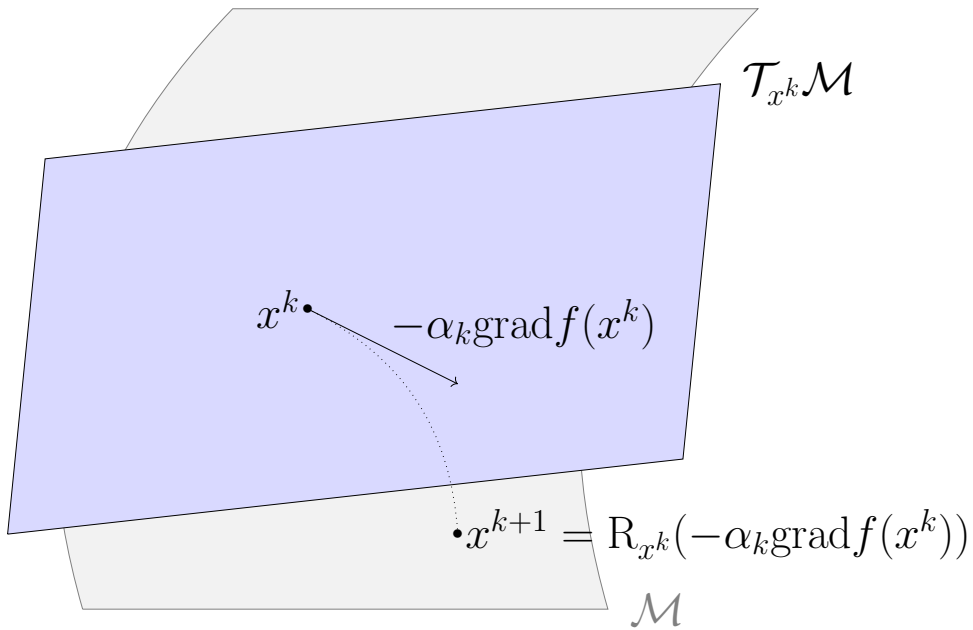
Let $f : \mathcal{M} \to \mathbb{R}$ be smooth on a Riemannian manifold $\mathcal{M}$. The Riemannian gradient of f is the vector field $\mathrm{grad} f$ on $\mathcal{M}$ uniquely defined by the following identities:

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

How to compute the Riemannian gradient?

If $\mathcal{M}$ is a Riemannian submanifold of a Euclidean space $\mathcal{E}$:

Simply take Euclidean gradient, project on the tangent space...



Orthogonal projectors

Let $\mathcal{M}$ be an embedded submanifold of a Euclidean space $\mathcal{E}$ equipped with a Euclidean metric $\langle \cdot, \cdot \rangle$.

The orthogonal projector to $\mathcal{T}_x\mathcal{M}$ is the linear map $\mathrm{Proj}_x : \mathcal{E} \to \mathcal{E}$ characterized by the following properties:

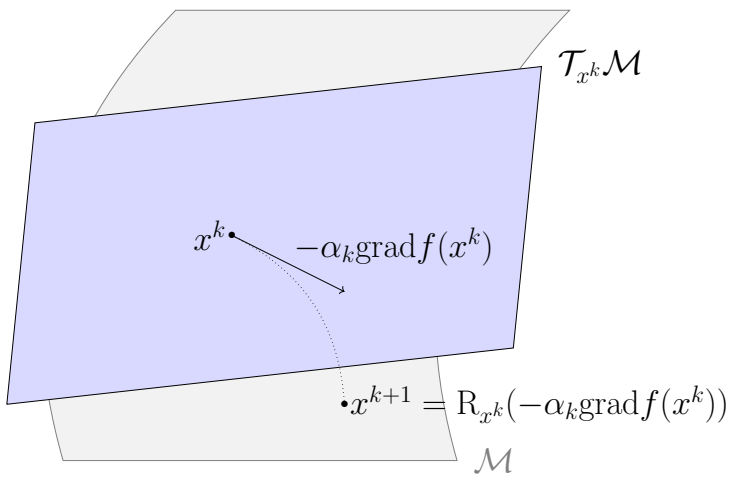
- Range: $\mathrm{im}(\mathrm{Proj}_x) = \mathcal{T}_x\mathcal{M}$;
- Projector: $\mathrm{Proj}_x \circ \mathrm{Proj}_x = \mathrm{Proj}_x$;
- Orthogonal: $\langle u - \mathrm{Proj}_x(u), v \rangle = 0$ for all $v \in \mathcal{T}_x\mathcal{M}$ and $u \in \mathcal{E}$.

How to compute the Riemannian gradient?

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

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

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



Example: the Rayleigh quotient

Let

$$f : \mathcal{S}^{d-1} \to \mathbb{R} : x \mapsto \frac{1}{2} x^\top A x$$

with

$$\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : x^\top x = 1\}.$$

We define the smooth extension

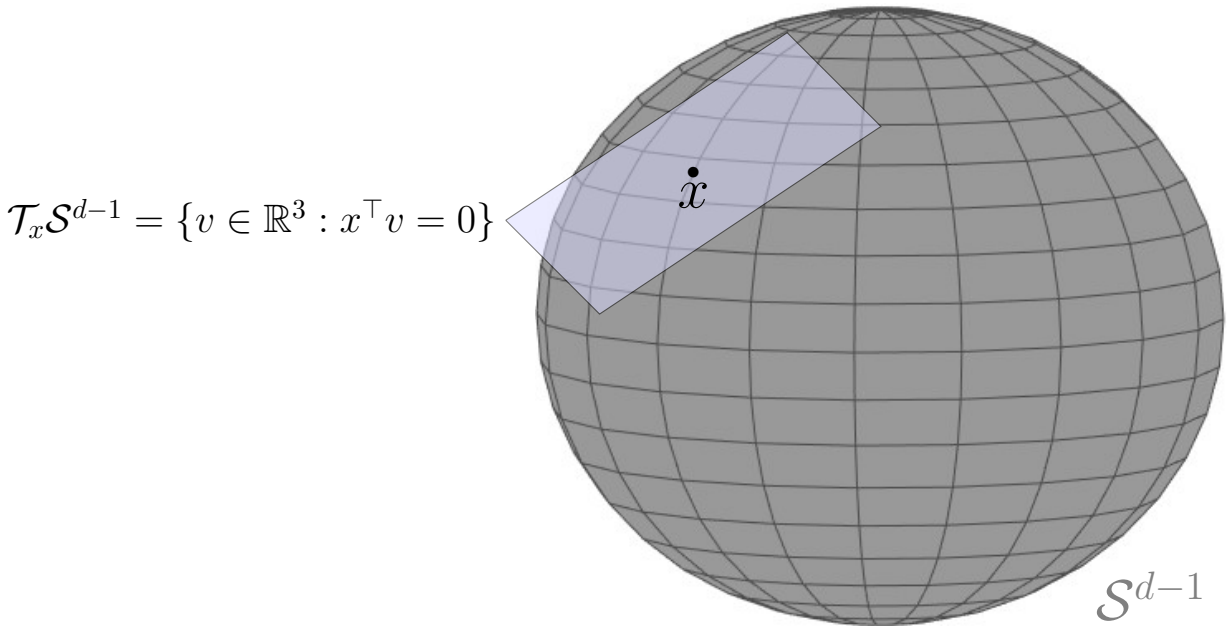
$$\bar{f} : \mathbb{R}^d \to \mathbb{R} : x \mapsto \frac{1}{2} x^\top A x.$$

whose (Euclidean) gradient is:

$$\nabla \bar{f}(x) = Ax.$$

To get the Riemannian gradient of f at $x \in \mathcal{S}^{d-1}$, we need a projector on the tangent space $\mathcal{T}_x \mathcal{S}^{d-1}$...

Projector on the tangent space to the sphere



The orthogonal projector to $\mathcal{T}_x\mathcal{S}^{d-1}$ is given by:

$$\mathrm{Proj}_x(v) = v - (x^\top v)x.$$

Example: the Rayleigh quotient (continued)

The Riemannian gradient of

$$f : \mathcal{S}^{d-1} \to \mathbb{R} : x \mapsto \frac{1}{2} x^\top A x$$

with

$$\mathcal{S}^{d-1} = \{x \in \mathbb{R}^d : x^\top x = 1\}$$

is thus simply obtained as

$$\mathrm{grad} f(x) = \mathrm{Proj}_x(\nabla \bar{f}(x)) = Ax - (x^\top A x)x.$$

Embedded submanifolds

Tangent spaces

Smoothness and differential of mapping between manifolds

Retraction

Metrics on manifolds

Riemannian gradient

Taylor development of functions on manifolds

Taylor development of functions defined on manifolds

- Let $f : \mathcal{M} \to \mathbb{R}$ be smooth, $c : \mathcal{I} \to \mathcal{M}$ 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 write $g : \mathcal{I} \to \mathbb{R} : t \mapsto g(t) = f(c(t))$.
- Since $g = f \circ c$ is smooth and maps real numbers to real numbers, it admits a Taylor expansion:

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

- By the chain rule,

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

- For $t = 0$, we get:

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

Taylor development of functions defined on manifolds

There follows that

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

If the curve c is obtained by a retraction (i.e., $c(t) = R_x(tv)\ \forall t$):

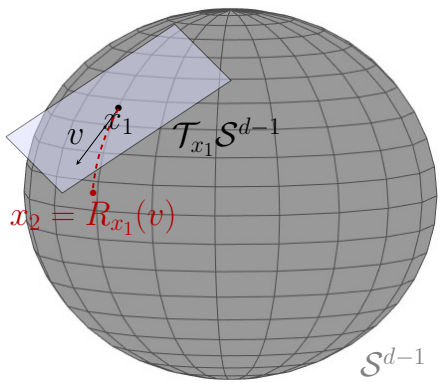
$$f(R_x(tv))) = f(x) + t\langle \operatorname{grad} f(x), v\rangle_x + \mathcal{O}(t^2).$$

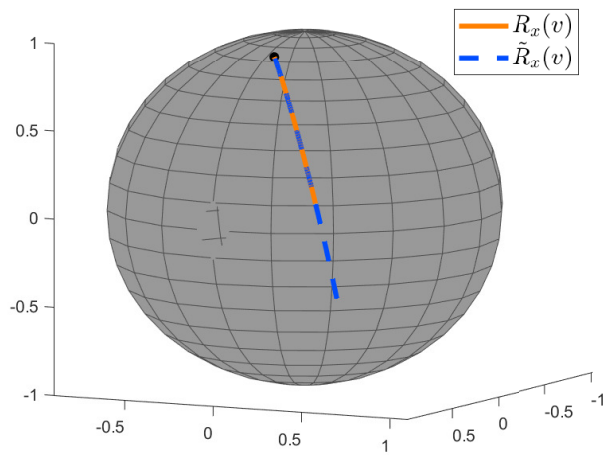
Applying the change of variable $s := tv \in \mathcal{T}_x\mathcal{M}$ gives:

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

Conclusions

- Definition of an **embedded submanifold** $\mathcal{M}$ of a vector space $\mathcal{E}$ through **local defining functions**.
- **Tangent space** $\mathcal{T}_x\mathcal{M}$**:** linear approx. to $\mathcal{M}$ at x (coincides with the kernel of the local defining function).
- **Retractions** allow to make a step on a manifold and in a given direction.





Conclusions

- A **Riemannian** metric is a **smoothly varying inner product** $\langle \eta_x, \xi_x \rangle_x$, for all $\eta_x, \xi_x \in \mathcal{T}_x\mathcal{M}$ and all $x \in \mathcal{M}$.
- When the Riemannian metric is the one of $\mathcal{E}$, $\mathcal{M}$ is a **Riemannian submanifold** of $\mathcal{E}$.
- Riemannian metrics allow to define the **Riemannian gradient** as the unique vector field satisfying

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

- For a Riemannian submanifold $\mathcal{M}$ of a Euclidean space $\mathcal{E}$,

$$\mathrm{grad}\ f(x) = \mathrm{Proj}_x(\mathrm{grad}\ \bar{f}(x)),$$

- The Taylor development of f around $x \in \mathcal{M}$ is

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

References:

The main reference for this presentation is "N. Boumal, An introduction to optimization on smooth manifolds, Cambridge University Press, 2023" (Chapter 3).

Part III: Riemannian gradient descent

Optimization on manifolds

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

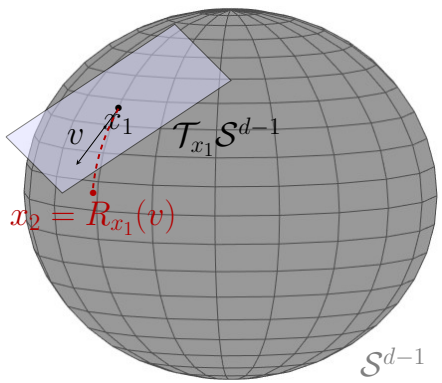
- $\mathcal{M}$ is the search space
- f is the cost function (a.k.a. objective function)

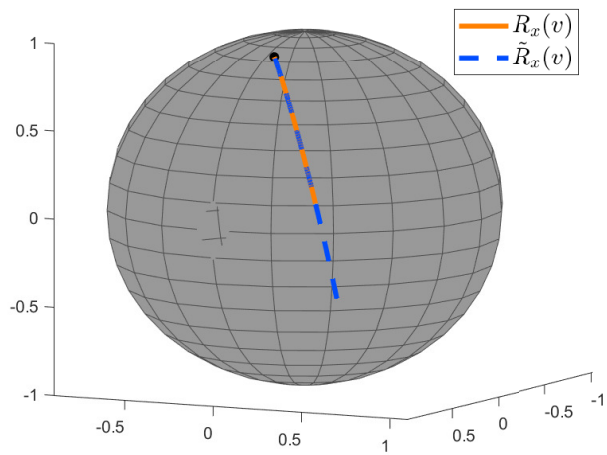
Optimization on manifolds

We assume that the search space and the objective function are "smooth".

Reminders

- Definition of an **embedded submanifold** $\mathcal{M}$ of a vector space $\mathcal{E}$ through **local defining functions**.
- **Tangent space** $\mathcal{T}_x\mathcal{M}$**:** linear approx. to $\mathcal{M}$ at x (coincides with the kernel of the local defining function).
- **Retractions** allow to make a step on a manifold and in a given direction.





Reminders

- A **Riemannian** metric is a **smoothly varying inner product** $\langle \eta_x, \xi_x \rangle_x$, for all $\eta_x, \xi_x \in \mathcal{T}_x\mathcal{M}$ and all $x \in \mathcal{M}$.
- When the Riemannian metric is the one of $\mathcal{E}$, $\mathcal{M}$ is a **Riemannian submanifold** of $\mathcal{E}$.
- Riemannian metrics allow to define the **Riemannian gradient** as the unique vector field satisfying

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

- For a Riemannian submanifold $\mathcal{M}$ of a Euclidean space $\mathcal{E}$,

$$\mathrm{grad} f(x) = \mathrm{Proj}_x(\mathrm{grad} \bar{f}(x)),$$

- The Taylor development of f around $x \in \mathcal{M}$ is

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

Optimization: basic definitions

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

- $x^* \in \mathcal{M}$ is a **global minimum** of (P) if and only if

$$f(x^*) \leq f(x) \qquad \forall x \in \mathcal{M}.$$

- $x^* \in \mathcal{M}$ is a **local minimum** of (P) if and only if

$$f(x^*) \leq f(x)$$

  for all x in a neighbourhood of x^* in $\mathcal{M}$.
- Often, finding global/local minimizers is too difficult; we target here critical points instead.

Critical points on manifolds

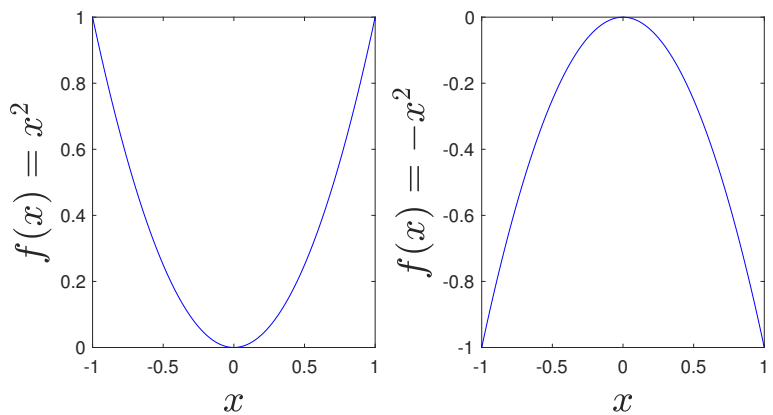
A point $x \in \mathcal{M}$ is **critical** (or **stationary**) for a smooth function $f : \mathcal{M} \to \mathbb{R}$ if

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

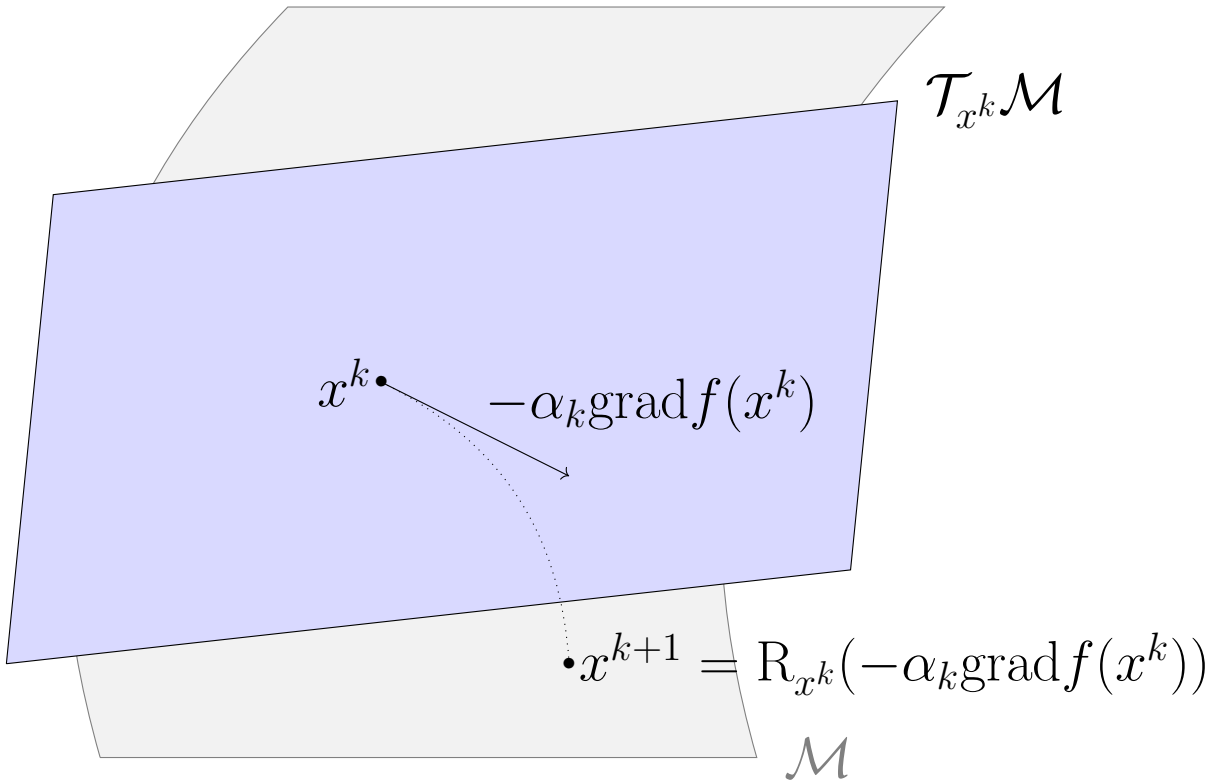
for all smooth curves c on $\mathcal{M}$ such that $c(0) = x$.
Equivalently, $x \in \mathcal{M}$ is **critical** if and only if

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

Note that being critical is a necessary condition for x to be a local minimizer of f.



Riemannian gradient descent



Riemannian gradient descent

Algorithm 4.1 RGD: the Riemannian gradient descent method

Input: $x_0 \in \mathcal{M}$

For $k = 0, 1, 2, \ldots$

Pick a step-size $\alpha_k > 0$

$x_{k+1} = \mathrm{R}_{x_k}(s_k)$, with step $s_k = -\alpha_k \mathrm{grad} f(x_k)$

Stepsize selection rules:

- Fixed stepsize: $\alpha_k = \alpha$ for all k
- Optimal stepsize: α_k minimizes exactly the function

$$g(\alpha) = f(R_{x_k}(-\alpha \mathrm{grad} f(x_k)))$$

- Backtracking: starting with a guess $\alpha_0 > 0$, iteratively reduce it by a factor $\tau \in (0, 1)$ until being deemed acceptable (see later).

Definitions: limit and accumulation points on manifolds

Consider a sequence S of points $x_0, x_1, x_2, \ldots$ on a manifold $\mathcal{M}$. Then,

- A point $x \in \mathcal{M}$ is a **limit of** S if, for every neighborhood $\mathcal{U}$ of x in $\mathcal{M}$, there exists an integer K such that $x_K, x_{K+1}, x_{K+2}, \ldots$ are in $\mathcal{U}$. The topology of a manifold is Hausdorff, hence a sequence has at most one limit. If x is the limit, we write

$$\lim_{k \to \infty} x_k = x \qquad \text{or} \qquad x_k \to x$$

  and we say 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, that is, if every neighborhood $\mathcal{U}$ of x in $\mathcal{M}$ contains an infinite number of elements of S.

Global convergence result

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

- There exists $f_{low} \in \mathbb{R}$ such that $f(x) \geq f_{low}$ for all $x \in \mathcal{M}$.
- There exists a constant $c > 0$ such that, for all k,

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

Then,

$$\lim_{k\to\infty} \|\mathrm{grad} f(x_k)\|_{x_k} = 0.$$

In particular, all accumulation points (if any) are critical points. Furthermore, for all $K \geq 1$, there exists k in $0, \ldots, K-1$ such that

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

Proof

- Based on a standard telescoping sum argument
- Note that, for all $K \geq 1$, there holds

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

- Taking $K \to \infty$ gives

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

- As each term in the summation is nonnegative, the summands must converge to zero:

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

so that $\|\mathrm{grad} f(x_k)\|_{x_k} \to 0$.

Proof (continued)

- Let us assume that there exists an accumulation point x of the sequence of iterates.
- By definition, there exists a subsequence of iterates $x_{(0)}, x_{(1)}, x_{(2)}, \ldots$ which converges to x.
- Since the norm of the gradient is continuous, it commutes with the limit and we get:

$$\begin{aligned} 0 &= \lim_{k\to\infty} \|\mathrm{grad} f(x_k)\|_{x_k} \\ &= \lim_{k\to\infty} \|\mathrm{grad} f(x_{(k)})\|_{x_k} \\ &= \|\mathrm{grad} f(x)\|_x, \end{aligned}$$

so that all accumulation points (if they exist) are critical points.

Looking further into the assumptions...

How can we ensure that at each iteration there exists a constant $c > 0$ such that, for all k,

$$f(x_k) - f(x_{k+1}) \geq c\|\mathrm{grad} f(x_k)\|_{x_k}^2 ?$$

Lipschitz-gradient type assumption

For a given subset S of the tangent bundle $\mathcal{T}\mathcal{M}$, there exists a constant $L > 0$ such that, for all $(x, s) \in S$,

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

Looking further into the assumptions...

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

Writing $x := x_k$, $s := -\alpha_k \operatorname{grad} f(x_k)$, and $x_{k+1} := R_x(s)$ gives

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \|\operatorname{grad} f(x_k)\|_{x_k}^2 + \frac{L}{2}\alpha_k^2 \|\operatorname{grad} f(x_k)\|_{x_k}^2$$

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

so that

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

This is equivalent to our initial assumption

$$f(x_k) - f(x_{k+1}) \geq c\|\operatorname{grad} f(x_k)\|_{x_k}^2, \qquad \text{with } c = \alpha_k \left(1 - \frac{L\alpha_k}{2}\right).$$

Choosing the best stepsize

We just showed that

$$f(x_k) - f(x_{k+1}) \geq c\|\mathrm{grad} f(x_k)\|_{x_k}^2, \qquad \text{with } c = \alpha_k \left(1 - \frac{L\alpha_k}{2}\right).$$

- The best objective decrease guarantee is obtained when c is maximized.
- This is achieved if $g : \mathbb{R} \to \mathbb{R} : \alpha \mapsto \alpha \left(1 - \frac{L\alpha}{2}\right)$ is maximized.

We get:

$$g'(\alpha^*) = \left(1 - \frac{L\alpha^*}{2}\right) - \alpha^* \frac{L}{2} = 1 - L\alpha^* = 0$$
$$\Leftrightarrow \alpha^* = 1/L.$$

Note that then $c = \frac{1}{2L}$.

Global convergence result (alternative)

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

- There exists $f_{low} \in \mathbb{R}$ such that $f(x) \geq f_{low}$ for all $x \in \mathcal{M}$.
- For a retraction R, there exist $S \subseteq T\mathcal{M}$ and $L > 0$ such that
  - $(x_k, s_k) \in S$ for all k, with $s_i := -(1/L)\mathrm{grad} f(x_i)$
  - $f(R_{x_k}(s_k)) \leq f(x_k) + \langle \mathrm{grad} f(x_k), s_k \rangle_{x_k} + \frac{L}{2}\|s_k\|^2_{x_k} \forall k.$

Then,

$$\lim_{k\to\infty} \|\mathrm{grad} f(x_k)\|_{x_k} = 0.$$

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

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

Conclusions so far

- This last result is equivalent to saying that, $\forall \epsilon > 0$, there exists $k \in \{0, \ldots, K-1\}$ such that

$$\|\mathrm{grad} f(x_k)\| \leq \epsilon \qquad \text{if} \qquad K \geq \frac{2L(f(x_0) - f_{low})}{\epsilon^2}.$$

- Global convergence rate in $\mathcal{O}(1/\epsilon^2)$, independent on the dimension of $\mathcal{M}$.

Connection with usual Euclidean assumptions

When $\mathcal{M} = \mathbb{R}^d$, ones classically assume that $\exists L > 0$ such that

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

which implies

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

Indeed, with $c(t) = x + ts$, we get

$$\begin{aligned} f(x+s) - f(x) &= f(c(1)) - f(c(0)) \\ &= \int_0^1 (f \circ c)'(t) \mathrm{d}t \\ &= \int_0^1 \mathrm{D}f(c(t))[c'(t)] \mathrm{d}t \quad = \int_0^1 \langle \nabla f(x+ts), s \rangle \mathrm{d}t. \end{aligned}$$

Connection with usual Euclidean assumptions

We just showed that

$$f(x+s) - f(x) = \int_0^1 \langle \nabla f(x+ts), s \rangle \mathrm{d}t.$$

Then, using Cauchy-Schwarz we get:

$$\begin{aligned} |f(x+s) - f(x) - \langle \nabla f(x), s \rangle| &= \left| \int_0^1 \langle \nabla f(x+ts) - \nabla f(x), s \rangle \mathrm{d}t \right| \\ &\leq \int_0^1 \|\nabla f(x+ts) - \nabla f(x)\| \|s\| \mathrm{d}t \\ &\leq \|s\| \int_0^1 L\|ts\| \mathrm{d}t \\ &\leq \frac{L}{2}\|s\|^2. \end{aligned}$$

Connection with usual Euclidean assumptions

Problem:

What to extend the usual Lipschitz assumption

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

to manifolds?

Natural possibility:

$$\| \underbrace{\operatorname{grad} f(R_x(s))}_{\in \mathcal{T}_{R_x(s)}\mathcal{M}} - \underbrace{\operatorname{grad} f(x)}_{\in \mathcal{T}_x\mathcal{M}} \|_{??} \leq L\|s\|_x, \qquad \forall (x, s) \in \mathcal{TM},$$

The left-hand side does not make sense on a manifold!

What if L is unknown? Backtracking line search strategy...

Algorithm 4.2 Backtracking line-search

Parameters: $\tau, r \in (0,1)$; for example, $\tau = \frac{1}{2}$ and $r = 10^{-4}$
Input: $x \in \mathcal{M}$, $\bar{\alpha} > 0$
Set $\alpha \leftarrow \bar{\alpha}$
While $f(x) - f(\mathrm{R}_x(-\alpha \mathrm{grad} f(x))) < r\alpha \|\mathrm{grad} f(x)\|^2$
    Set $\alpha \leftarrow \tau\alpha$
Output: α

- The condition

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

  is referred to as the **Armijo-Goldstein** condition.
- A convergence rate of $\mathcal{O}(1/\epsilon^2)$ can again be guaranteed, under Lipschitz-type assumption on the gradient.

Local convergence

- So far, our analyses do not depend on the initial point $x_0 \in \mathcal{M}$: these are **global convergence** results.
- The decrease of the gradient norm is slow $(1/\sqrt{k})$.
- In practice, it is common to observe an eventually exponential decrease of the gradient norm.
- This is the **local convergence**: the convergence of sequences when they are close enough to their limit.

Linear convergence rate

In a metric space with a distance dist, a sequence $a_0, a_1, a_2, \ldots$ converges at least linearly to a^* if there exist positive reals $\epsilon_0, \epsilon_1, \epsilon_2, \ldots$ converging to zero such that

$$\mathrm{dist}(a_k, a^*) \leq \epsilon_k \qquad \text{and} \qquad \lim_{k \to \infty} \frac{\epsilon_{k+1}}{\epsilon_k} = \mu$$

for some $\mu \in (0, 1)$. The infimum over such μ is the linear convergence factor. If the latter is zero, the convergence is superlinear.

Local convergence result

Let $\mathcal{M}$ be a Riemannian manifold with a retraction R and $f : \mathcal{M} \to \mathbb{R}$ be smooth. Assume that $x^* \in \mathcal{M}$ satisfies

$$\mathrm{grad} f(x^*) = 0 \qquad \text{and} \qquad \mathrm{Hess} f(x^*) \succ 0.$$

Let $0 < \lambda_{\min} \leq \lambda_{\max}$ be the extreme eigenvalues of $\mathrm{Hess} f(x^*)$, and let $\kappa = \dfrac{\lambda_{\max}}{\lambda_{\min}}$ denote its condition number. Set $L > \dfrac{1}{2}\lambda_{\max}$.
Given $x_0 \in \mathcal{M}$, define

$$x_{k+1} := R_{x_k}\left(-\frac{1}{L}\mathrm{grad} f(x_k)\right).$$

There exists a neighborhood of x^* such that, if the above sequence enters the neighborhood, then it stays in that neighborhood and it converges to x^* at least linearly. If $L = \lambda_{\max}$, the linear convergence factor is at most $1 - \dfrac{1}{\kappa}$.

Conclusions

- Riemannian gradient descent iterates

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

- Various stepsize selection rules can be used.
- Global convergence in gradient norm in $\mathcal{O}(1/\sqrt{k})$ can be proven.
- Local **linear** convergence can also be shown, provided additional assumptions.

References:

The main reference for this presentation is "N. Boumal, An introduction to optimization on smooth manifolds, Cambridge University Press, 2023" (Chapter 4).

Part IV: Beyond embedded submanifolds...

Optimization on manifolds

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

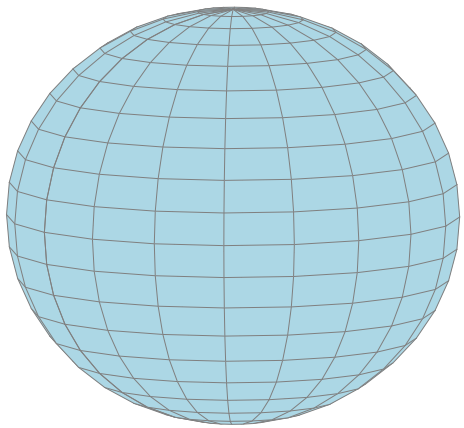
- $\mathcal{M}$ is the search space
- f is the cost function (a.k.a. objective function)

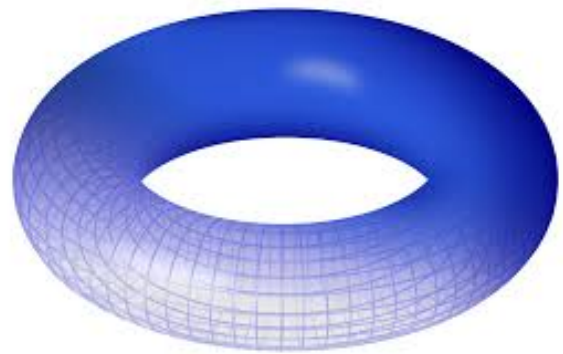
Optimization on manifolds

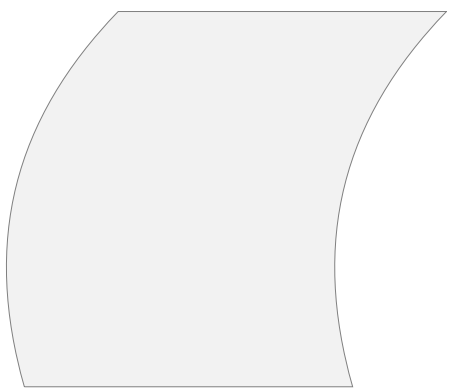
We assume that the search space and the objective function are "'smooth".

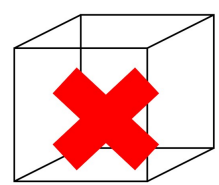
What is a manifold?

Until now, “easy” manifolds: embedded in the Euclidean space...

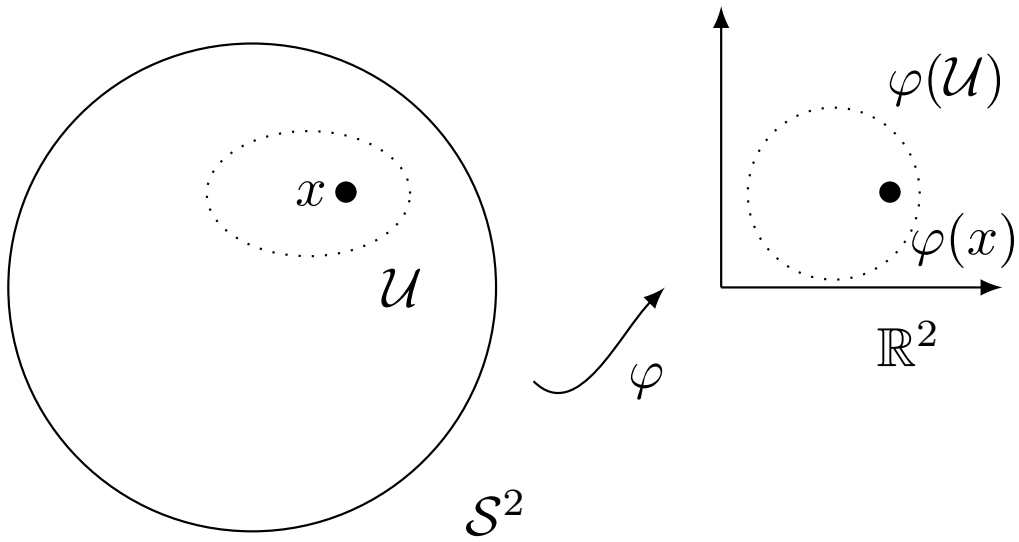








Charts


x
$\mathcal{U}$
$\mathcal{S}^2$
φ
$\varphi(\mathcal{U})$
$\varphi(x)$
$\mathbb{R}^2$

Definition of a chart

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

- $\phi(\mathcal{U})$ is open in $\mathbb{R}^d$, and
- ϕ is invertible between $\mathcal{U}$ and $\phi(\mathcal{U})$.

The map $\phi^{-1} : \phi(\mathcal{U}) \to \mathcal{U}$ is a local parameterization of $\mathcal{M}$.

Smoothness of functions on manifolds

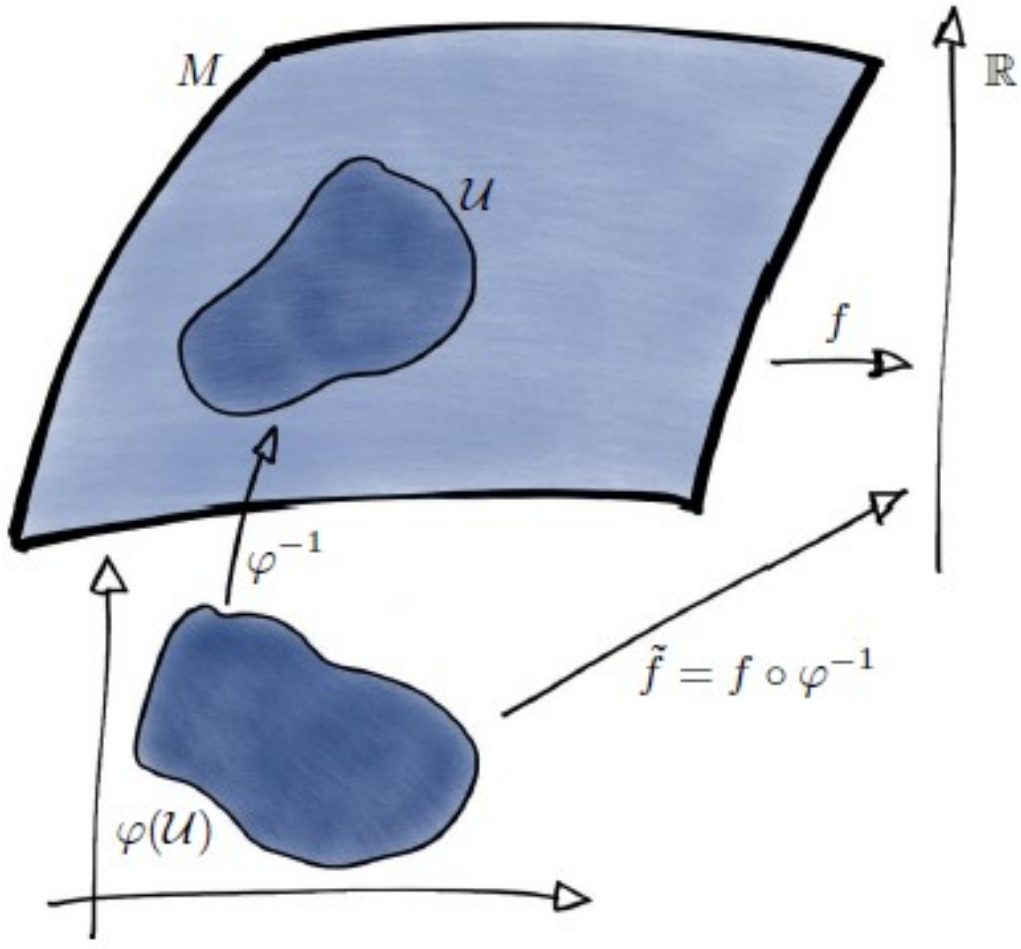


Figure: Figure courtesy of *Boumal23*.

Smoothness of functions on manifolds

Let $(\mathcal{U}, \phi)$ be a d-dimensional chart around $x \in \mathcal{M}$.

The function

$$\tilde{f} = f \circ \phi^{-1} : \phi(\mathcal{U}) \to \mathbb{R}$$

is called a coordinate representative of f in this chart.

Since $\phi(\mathcal{U})$ is open in $\mathbb{R}^d$, it makes sense to talk of differentiability of $\tilde{f}$.

We define that, with respect to this chart, f is smooth at x if $\tilde{f}$ is smooth at $\phi(x)$.

Compatibility condition between charts

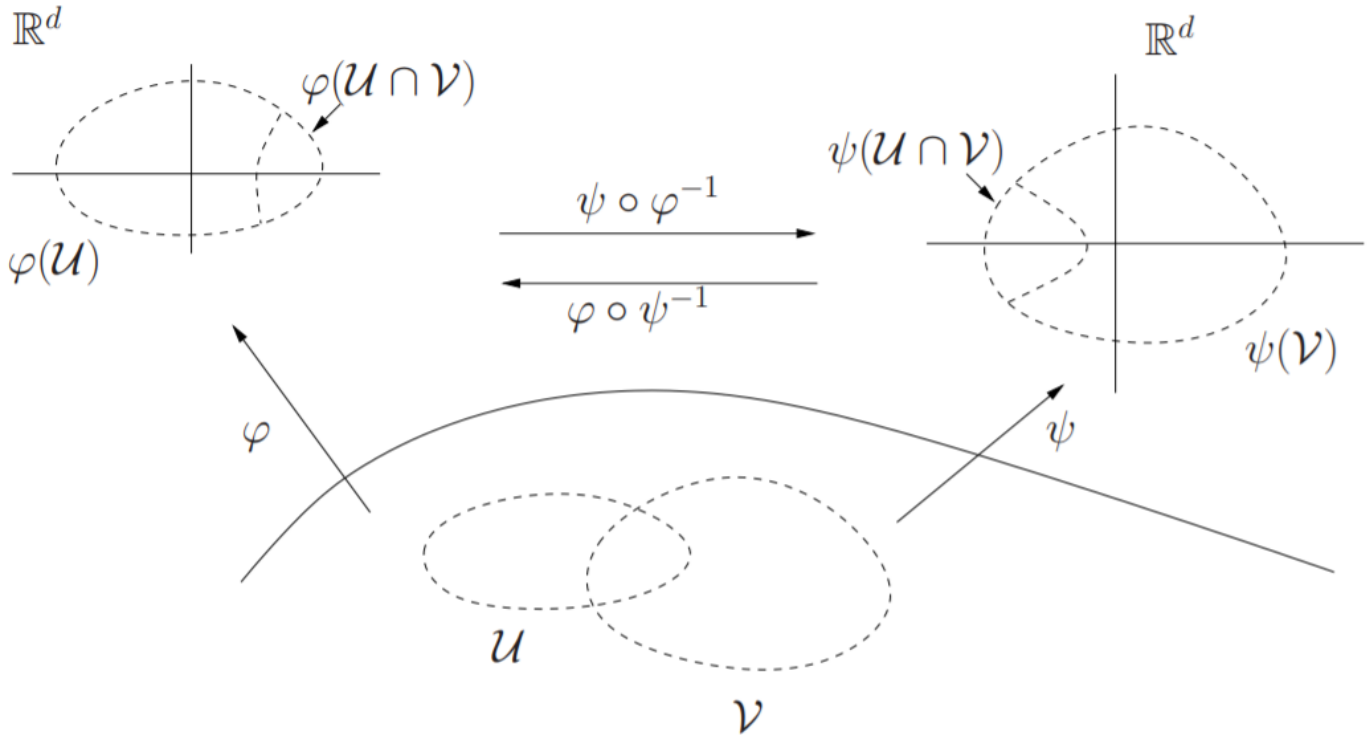


Figure: Figure courtesy of *Absil08*.

Compatibility condition between charts and atlas

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

- $\phi(\mathcal{U} \cap \mathcal{V})$ is open in $\mathbb{R}^d$; and
- $\psi(\mathcal{U} \cap \mathcal{V})$ is open in $\mathbb{R}^d$; and
- $\psi \circ \phi^{-1} : \phi(\mathcal{U} \cap \mathcal{V}) \to \psi(\mathcal{U} \cap \mathcal{V})$ is a smooth invertible function whose inverse is also smooth (i.e., it is a diffeomorphism).

(They give then the same conclusions regarding smoothness of functions at x).

Definition of an atlas

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

Exercise

Consider the unit circle,

$$\mathcal{S}^1 = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}.$$

One possible atlas is made of four charts:

- $\mathcal{U}_N = \{x \in S^1; x_2 > 0\}, \qquad \phi_N(x) = x_1,$
- $\mathcal{U}_E = \{x \in S^1 : x_1 > 0\}, \qquad \phi_E(x) = x_2,$
- $\mathcal{U}_S = \{x \in S^1 : x_2 < 0\}, \qquad \phi_S(x) = x_1,$
- $\mathcal{U}_W = \{x \in S^1 : x_1 < 0\}, \qquad \phi_W(x) = x_2.$

Questions

- What is the dimension of these charts?
- Are their images open sets?
- Are they invertible?
- Do they satisfy the compatibility condition?

Exercise: solution

- These are one-dimensional charts.
- For the 1st chart: $\phi_N(\mathcal{U}_N) = (-1, 1)$.
- For the 1st chart: $\phi_N^{-1}(z) = (z, \sqrt{1 - z^2})$
- Checking for the North and East charts, we get:
  - $\mathcal{U}_N \cap \mathcal{U}_E = \{x \in \mathcal{S}^1 : x_1 > 0 \text{ and } x_2 > 0\}$
  - $\phi_N(\mathcal{U} \cap \mathcal{N}) = (0, 1)$ is open
  - $\phi_E(\mathcal{U}_N \cap \mathcal{U}_E) = (0, 1)$ is open, and
  - $\phi_E^{-1}(z) = (\sqrt{1 - z^2}, z)$, and thus $\phi_N(\phi_E^{-1}(z)) = \sqrt{1 - z^2}$.

Manifold

Definition of a manifold

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

Tangent vectors as equivalence classes

Let C_x be the set

$$C_x = \{c | c : \mathcal{I} \to \mathcal{M} \text{ is smooth and } c(0) = x\}.$$

Let $(\mathcal{U}, \phi)$ be a chart of $\mathcal{M}$ around x and consider $c_1, c_2 \in C_x$. Then, $c_1 \sim c_2$ if and only if

$$c_1 \sim c_2 \Leftrightarrow (\phi \circ c_1)'(0) = (\phi \circ c_2)'(0).$$

The equivalence class of a curve $c \in C_x$ is the set

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

Each equivalence class is called a tangent vector to $\mathcal{M}$ at x. The tangent space to $\mathcal{M}$ at x, denoted by $\mathcal{T}_x\mathcal{M}$, is the quotient set

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

that is, the set of all equivalence classes.

Differential of a function defined on a manifold

The differential of a smooth function $f : \mathcal{M} \to \mathbb{R}$ at x is a linear map

$$\mathrm{D}f(x) : \mathcal{T}_x\mathcal{M} \to \mathbb{R}$$

defined by:

$$\mathrm{D}f(x)[v] = \frac{\mathrm{d}}{\mathrm{d}t} f(c(t))|_{t=0}$$

where c is a smooth curve on $\mathcal{M}$ passing through x at $t = 0$ such that $v = [c]$.

Retraction and other tools

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 for each $(x, v) \in \mathcal{TM}$ the curve $c(t) = R_x(tv)$ satisfies $v = [c]$, where $[c]$ is the equivalence class (tangent vector) of the curve c.

All other tools seen for embedded submanifold can be extended to manifolds

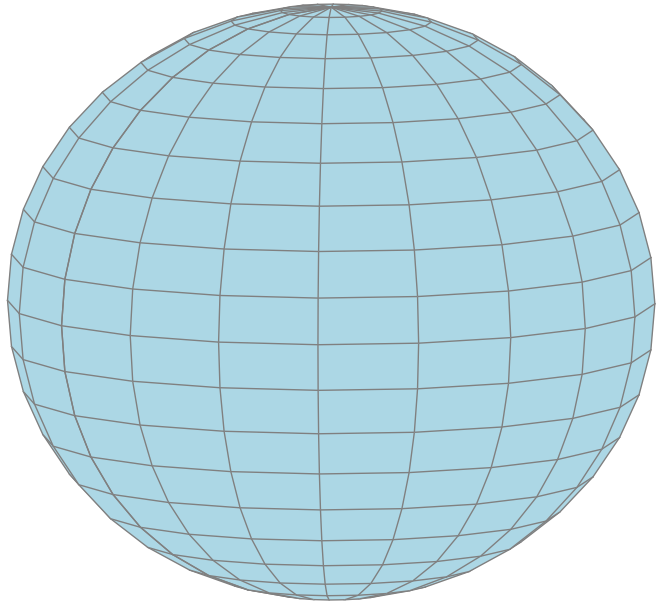
For example, one can define a Riemannian metric and a Riemannian gradient on manifolds but the relation

$$\mathrm{grad} f(x) = \mathrm{Proj}_{\mathcal{T}_x\mathcal{M}} \nabla f(x)$$

does not hold any more...

How to choose the representation of a manifold?

Other ways to define a manifold: quotient manifolds...



$\mathcal{S}^2$ embedded in $\mathbb{R}^3$

OR

$\mathcal{S}^2 \simeq \mathbb{R}^3_*/\mathbb{R}_+$

$x \sim y \Leftrightarrow x = \alpha y, \alpha \in \mathbb{R}_+$

Quotient manifolds: other examples

- Stiefel manifold (Edelman98):

$$\mathrm{St}(p,n) = \{Y \in \mathbb{R}^{n\times p} | Y^\top Y = I_p\}$$

  - embedded manifold of $\mathbb{R}^{n\times p}$
  - $\mathrm{St}(p,n) \simeq \mathcal{O}_n/\mathcal{O}_{n-p}$

$$Q = \begin{bmatrix} Y & Y_\perp \end{bmatrix}$$

- Grassmann manifold (Edelman98):

$$\mathrm{Grass}(p,n) = \text{“Set of p-dimensional subspaces in } \mathbb{R}^n\text{”}$$

  - $\mathrm{Grass}(p,n) \simeq \mathrm{St}(p,n)/\mathcal{O}_p$
  - $\mathrm{Grass}(p,n) \simeq \mathcal{O}_n/(\mathcal{O}_p \times \mathcal{O}_{n-p})$
- Manifold of rank-k $m \times n$ matrices (Vandereycken12, Meyer11LR)

Conclusions

- In general, manifolds are sets endowed with a differential structure (**atlas**).
- Different atlases results in different topologies.
- All the tools needed to run Riemannian optimization algorithms can be extended to (general manifolds).
- In practice, the atlas is often chosen as the differential structure inherited when the manifold is seen as an **embedded submanifold** or a **quotient manifold** of some known manifold (often, $\mathbb{R}^{n \times n}$).

Thanks for your attention!

References:

Most of the definitions, results and examples are extracted (often verbatim) from N. Boumal, An introduction to optimization on smooth manifolds , Cambridge University Press, 2023. (Chapter 8)

Other references:

- Absil08: P.-A. Absil, R. Mahony, R. Sepulchre, *Optimization algorithms on matrix manifolds*, 2008.
- Edelman98: A. Edelman, T. A. Arias, S. T. Smith, *The Geometry of Algorithms with Orthogonality Constraints*, SIAM J. Matrix Anal.Appl., 20(2), 303–353, 1998.
- Meyer11LR: G. Meyer, S. Bonnabel, R. Sepulchre, *Linear Regression under Fixed-Rank Constraints: A Riemannian Approach*, Proceedings of ICML 2011.
- Vandereycken12: B. Vandereycken, *Low-rank matrix completion by Riemannian optimization—extended version*, arxiv preprint 2012.