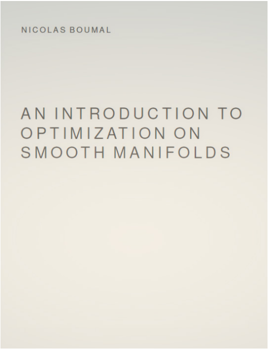


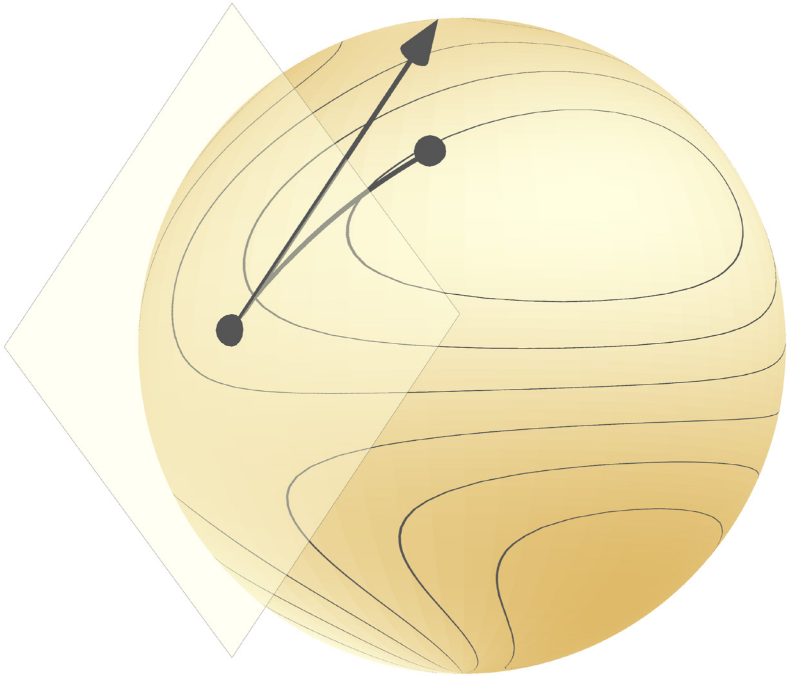

nicolasboumal.net/book

# An introduction to optimization on manifolds

Aug 30, 2021

Geometric methods in optimization and sampling,
Boot camp at Simons Institute

Nicolas Boumal – OPTIM
Institute of Mathematics, EPFL



# Step 0 in optimization

It all starts with a set $S$ and a function $f: S \to \mathbf{R}$:

$$\min_{x \in S} f(x)$$

These bare objects fully specify the problem.

Any additional structure on $S$ and $f$ may (and should) be exploited for algorithmic purposes but is not part of the problem.

# Classical unconstrained optimization

The search space *is* a linear space, e.g., $S = \mathbf{R}^n$:

$$\min_{x \in \mathbf{R}^n} f(x)$$

We can *choose* to turn $\mathbf{R}^n$ into a Euclidean space: $\langle u, v \rangle = u^\top v$.

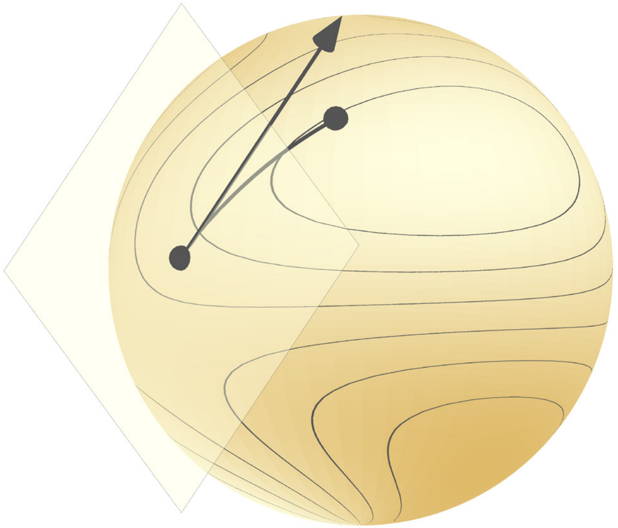
If $f$ is differentiable, this provides gradients $\nabla f$ and Hessians $\nabla^2 f$.

These objects underpin algorithms: gradient descent, Newton's method...

$$\langle \nabla f(x), v \rangle = \mathrm{D}f(x)[v] = \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}$$

$$\nabla^2 f(x)[v] = \mathrm{D}(\nabla f)(x)[v] = \lim_{t \to 0} \frac{\nabla f(x + tv) - \nabla f(x)}{t}$$

# Extend to optimization on manifolds



The search space *is* a smooth manifold, $S = \mathcal{M}$:

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

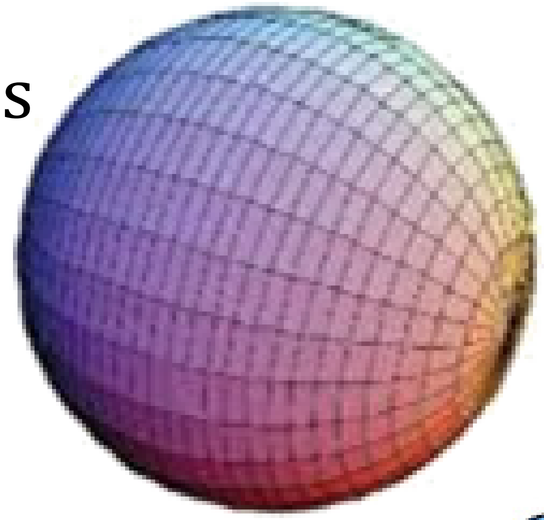
We can *choose* to turn $\mathcal{M}$ into a Riemannian manifold.

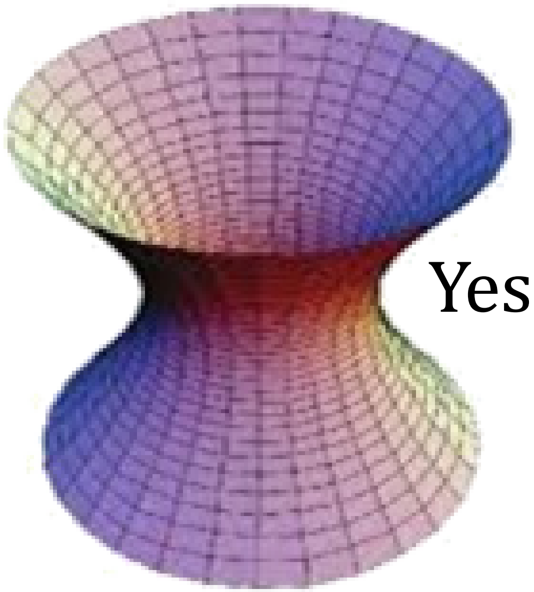
If $f$ is differentiable, this provides Riemannian gradients and Hessians.
These objects underpin algorithms: gradient descent, Newton's method...

Around since the 70s; practical since the 90s.

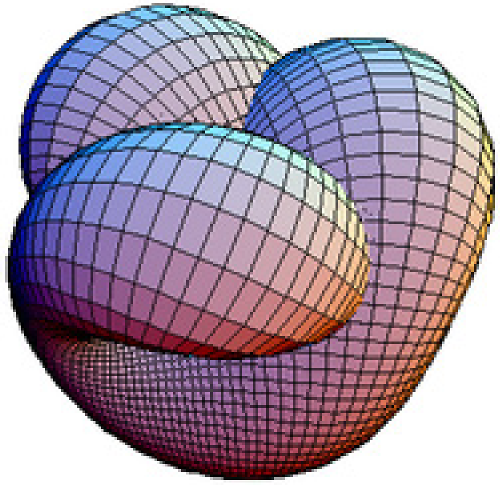
# What is a manifold? Take one:

Yes





Yes



No

# What is a manifold? Take two:

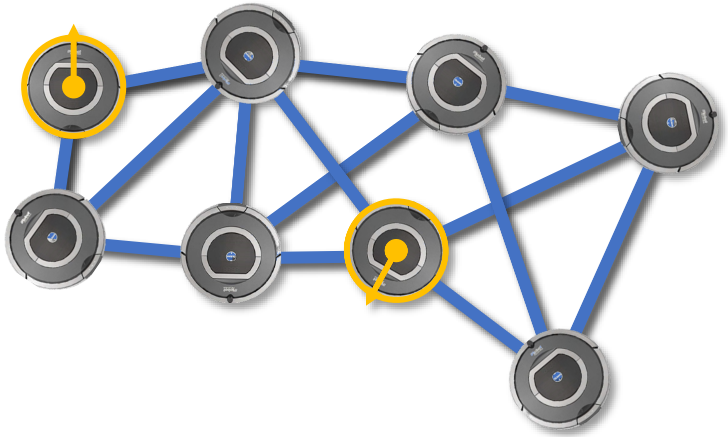
A few manifolds that come up in the wild

# Orthonormal frames and rotations

Stiefel manifold: $\mathcal{M} = \{X \in \mathbf{R}^{n\times p} : X^\top X = I_p\}$

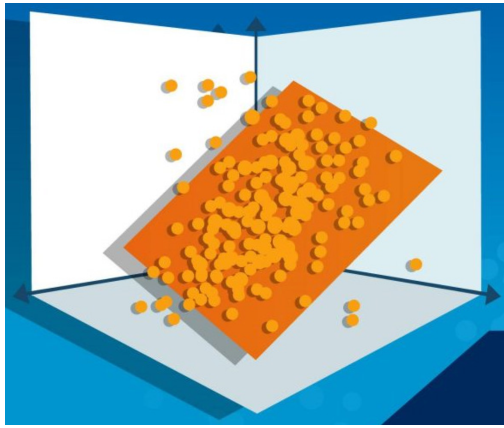
Rotation group: $\mathcal{M} = \{X \in \mathbf{R}^{3\times 3} : X^\top X = I_3 \text{ and } \det(X) = +1\}$

Applications in sparse PCA, Structure-from-Motion, SLAM (robotics)...




The singularities of Euler angles (gimbal lock) are artificial: the rotation group is smooth.

# Subspaces and fixed-rank matrices



Grassman manifold: $\mathcal{M} = \{\text{subspaces of dimension } d \text{ in } \mathbf{R}^n\}$

Fixed-rank matrices: $\mathcal{M} = \{X \in \mathbf{R}^{m\times n} : \text{rank}(X) = r\}$

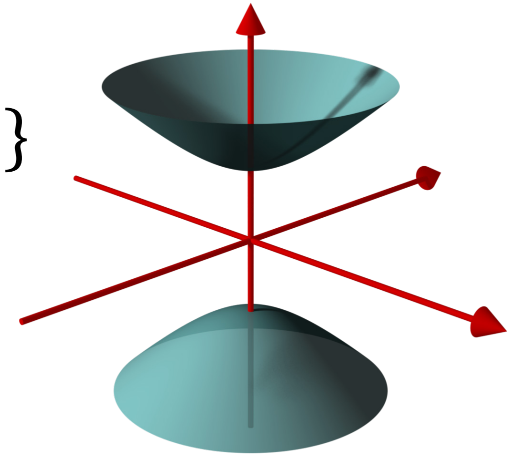
Applications to linear dimensionality reduction, data completion and denoising, large-scale matrix equations, ...

Optimization allows us to go beyond PCA (least-squares loss ≡ truncated SVD): can handle outlier-robust loss functions and missing data.



*Picture: https://365datascience.com/tutorials/python-tutorials/principal-components-analysis/*

# Positive matrices and hyperbolic space



Positive definite matrices: $\mathcal{M} = \{X \in \mathbf{R}^{n\times n}: X = X^\top \text{ and } X \succ 0\}$

Hyperbolic space: $\mathcal{M} = \{x \in \mathbf{R}^{n+1}: x_0^2 = 1 + x_1^2 + \cdots + x_n^2\}$

Used in metric learning, Gaussian mixture models, tree-like embeddings...

With appropriate metrics, these are Cartan-Hadamard manifolds:
Complete, simply connected, with non-positive (intrinsic) curvature.
Great playground for geodesic convexity.

*Picture: https://bjlkeng.github.io/posts/hyperbolic-geometry-and-poincare-embeddings*

# A tour of technical tools
## Restricted to embedded submanifolds

What is a manifold?

Tangent spaces

Smooth maps

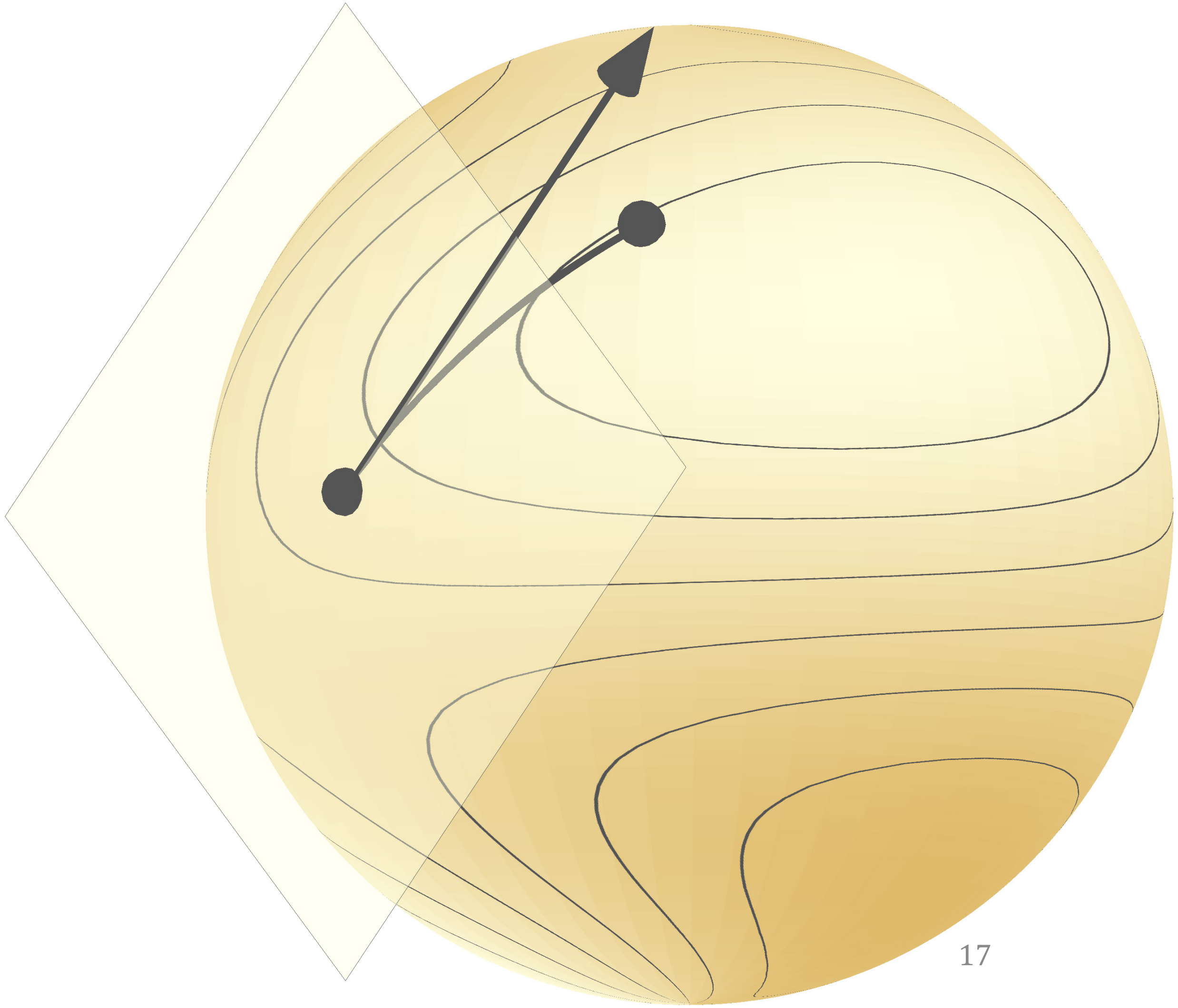
Differentials

Retractions

Riemannian manifolds

Gradients

Hessians

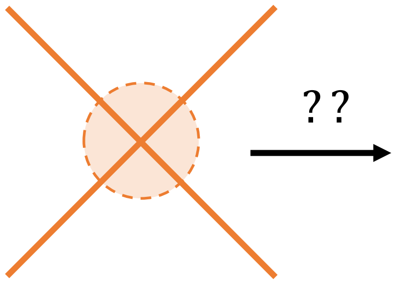


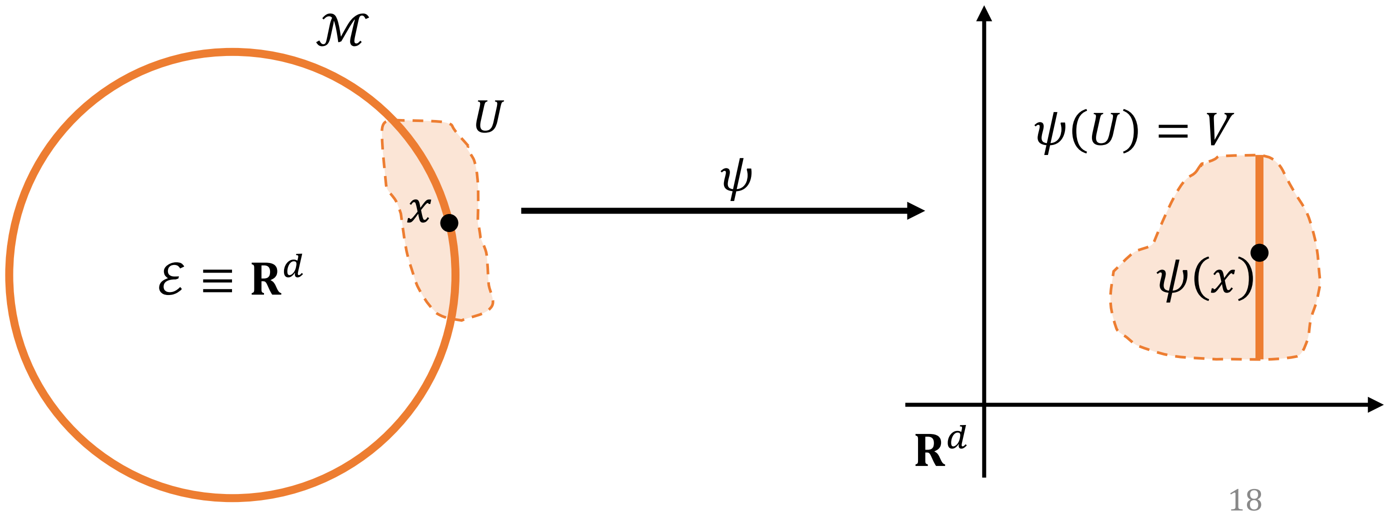
# What is a manifold? Take three:

A subset $\mathcal{M}$ of a linear space $\mathcal{E}$ of dimension $d$ is
a smooth embedded submanifold of dimension $n$ if:

For all $x \in \mathcal{M}$, there exists a neighborhood $U$ of $x$ in $\mathcal{E}$, an open set $V \subseteq \mathbf{R}^d$ and a diffeomorphism $\psi: U \to V$ such that $\psi(U \cap \mathcal{M}) = V \cap E$ where $E$ is a linear subspace of dimension $n$.

We call $\mathcal{E}$ the embedding space.



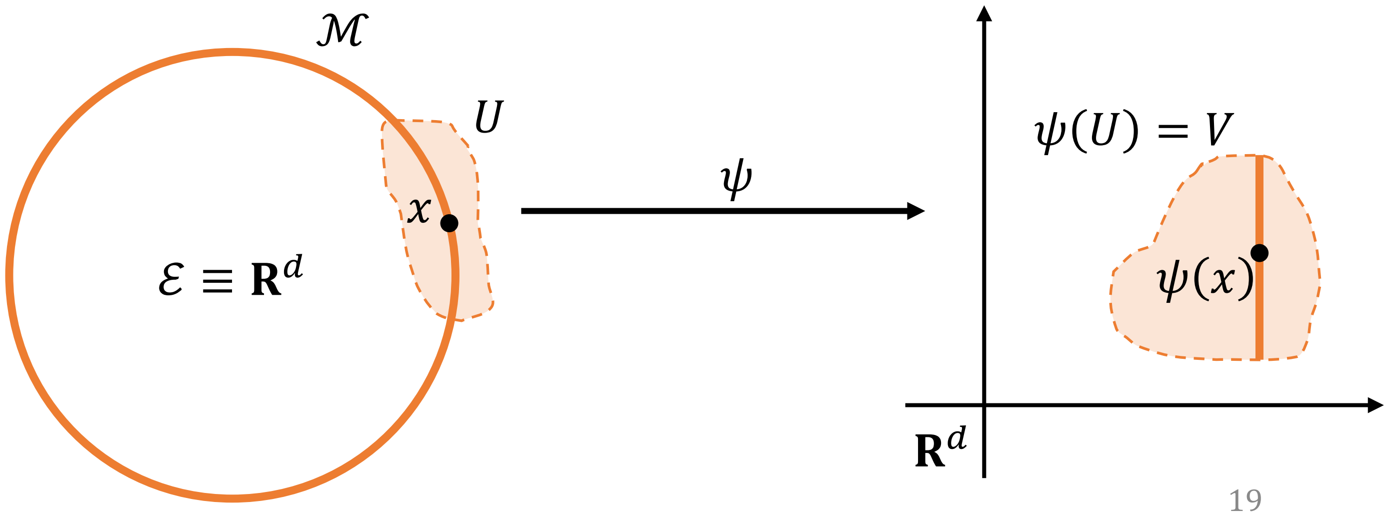
# What is a manifold? Quick facts:

Matrix sets in our list are manifolds: orthonormal, fixed-rank, positive definite...

Linear subspaces are manifolds.

Open subsets of manifolds are manifolds.

Products of manifolds are manifolds.

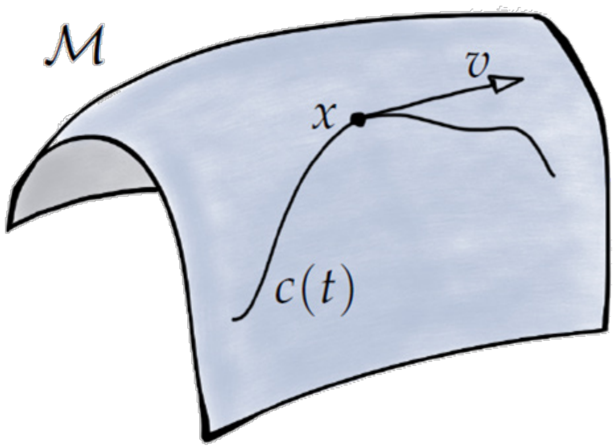


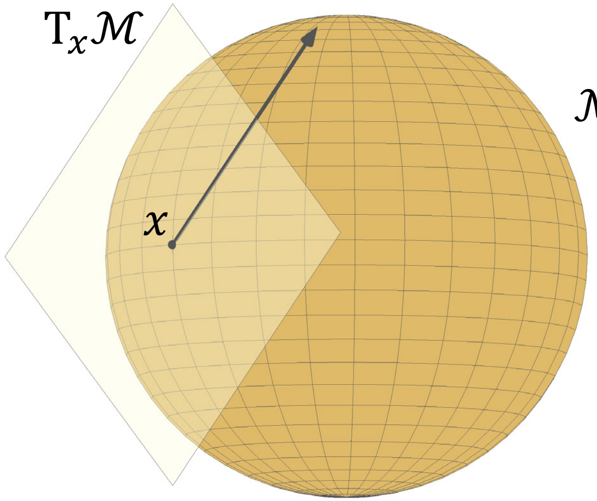
# Tangent vectors of $\mathcal{M}$ embedded in $\mathcal{E}$

A tangent vector at $x$ is the velocity $c'(0) = \lim_{t\to 0} \frac{c(t)-c(0)}{t}$ of a curve $c\colon \mathbf{R} \to \mathcal{M}$ with $c(0) = x$.

The tangent space $\mathrm{T}_x\mathcal{M}$ is the set of all tangent vectors of $\mathcal{M}$ at $x$.
It is a linear subspace of $\mathcal{E}$ of the same dimension as $\mathcal{M}$.

If $\mathcal{M} = \{x : h(x) = 0\}$ with $h\colon \mathcal{E} \to \mathbf{R}^k$ smooth and rank $\mathrm{D}h(x) = k$, then $\mathrm{T}_x\mathcal{M} = \ker \mathrm{D}h(x)$.




$h(x) = x^\top x - 1 = 0$
$\ker \mathrm{D}h(x) = \{v : x^\top v = 0\}$

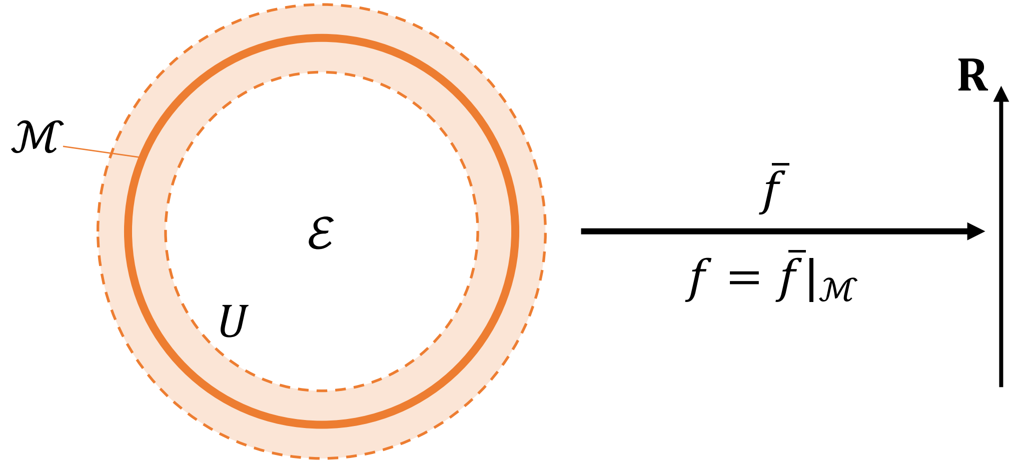
# Smooth maps on/to manifolds

Let $\mathcal{M}, \mathcal{M}'$ be (smooth, embedded) submanifolds of linear spaces $\mathcal{E}, \mathcal{E}'$.

A map $F: \mathcal{M} \to \mathcal{M}'$ is smooth if it has a smooth extension, i.e., if there exists a neighborhood $U$ of $\mathcal{M}$ in $\mathcal{E}$ and a smooth map $\bar{F}: U \to \mathcal{E}'$ such that $F = \bar{F}|_{\mathcal{M}}$.

Example: a cost function $f: \mathcal{M} \to \mathbf{R}$ is smooth if it is the restriction of a smooth $\bar{f}: U \to \mathbf{R}$.

Composition preserves smoothness.



# Differential of a smooth map $F: \mathcal{M} \to \mathcal{M}'$

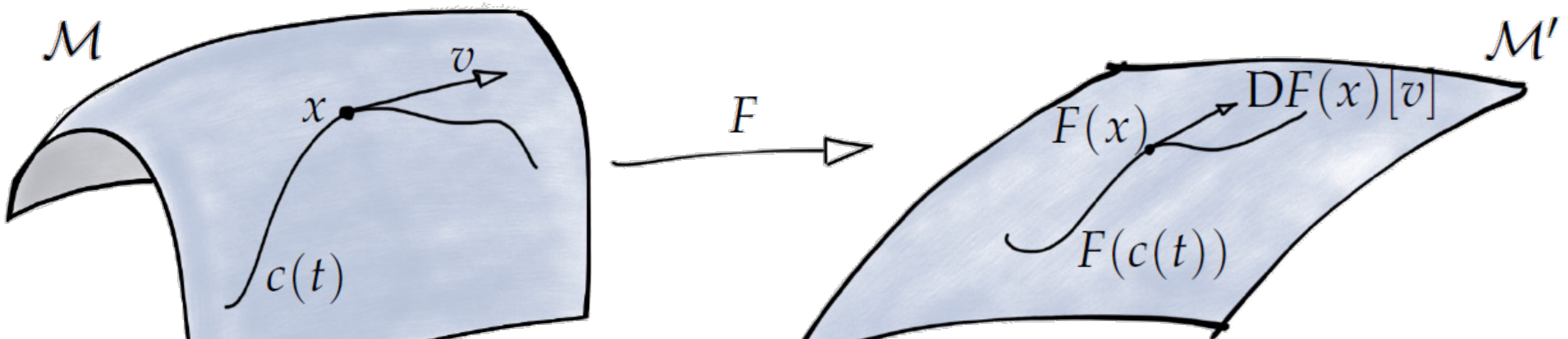
The differential of $F$ at $x$ is the map $\mathrm{D}F(x): \mathrm{T}_x\mathcal{M} \to \mathrm{T}_{F(x)}\mathcal{M}'$ defined by:

$$\mathrm{D}F(x)[v] = (F \circ c)'(0) = \lim_{t \to 0} \frac{F\big(c(t)\big) - F(x)}{t}$$

where $c: \mathbf{R} \to \mathcal{M}$ satisfies $c(0) = x$ and $c'(0) = v$.

Claim: $\mathrm{D}F(x)$ is well defined and linear, and we have a chain rule.

If $\bar{F}$ is a smooth extension of $F$, then $\mathrm{D}F(x) = \mathrm{D}\bar{F}(x)|_{\mathrm{T}_x\mathcal{M}}$.

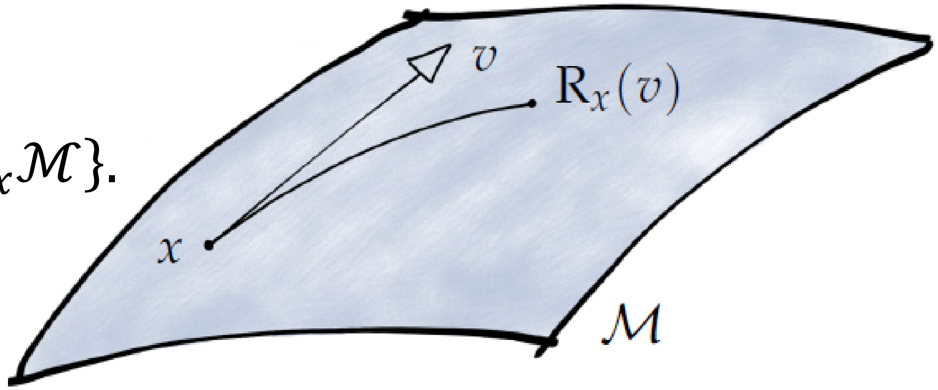


# Retractions: moving around on $\mathcal{M}$

The tangent bundle is the set

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

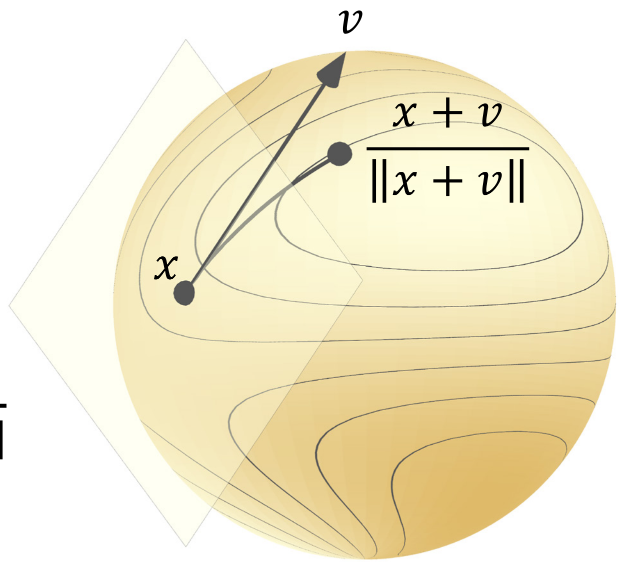
Claim: $\mathrm{T}\mathcal{M}$ is a smooth manifold embedded in $\mathcal{E} \times \mathcal{E}$.




A retraction is a smooth map $R : \mathrm{T}\mathcal{M} \to \mathcal{M} : (x, v) \mapsto R_x(v)$ such that each curve

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

satisfies $c(0) = x$ and $c'(0) = v$.




E.g., metric projection: $R_x(v)$ is the projection of $x + v$ to $\mathcal{M}$.

$\mathcal{M} = \mathbf{R}^n$: $R_x(v) = x + v$; $\quad\quad$ $\mathcal{M} = \{x : \|x\| = 1\}$: $R_x(v) = \frac{x+v}{\|x+v\|}$

$\mathcal{M} = \{X : \operatorname{rank}(X) = r\}$: $R_X(V) = \mathrm{SVD}_r(X + V)$.

# Riemannian manifolds

Each tangent space $\mathrm{T}_x\mathcal{M}$ is a linear space.
Endow each one with an inner product: $\langle u, v \rangle_x$ for $u, v \in \mathrm{T}_x\mathcal{M}$.

A vector field is a map $V: \mathcal{M} \to \mathrm{T}\mathcal{M}$ such that $V(x)$ is tangent at $x$ for all $x$.

We say the inner products $\langle \cdot, \cdot \rangle_x$ vary smoothly with $x$ if $x \mapsto \langle U(x), V(x) \rangle_x$ is smooth for all smooth vector fields $U, V$.

If the inner products vary smoothly with $x$, they form a Riemannian metric.

A Riemannian manifold is a smooth manifold with a Riemannian metric.

# Riemannian structure and optimization

A Riemannian manifold is a smooth manifold with a smoothly varying choice of inner product on each tangent space.

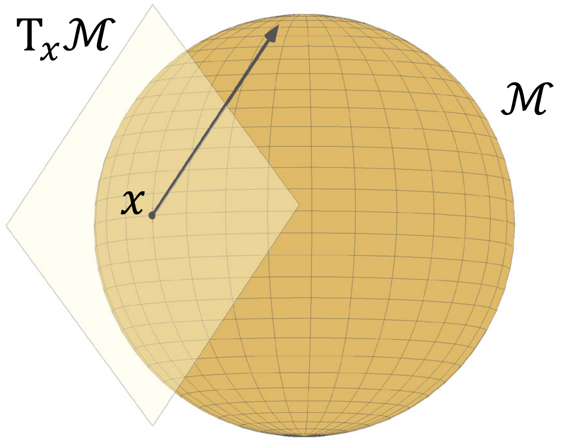
A manifold can be endowed with many different Riemannian structures.

A problem $\min_{x \in \mathcal{M}} f(x)$ is defined independently of any Riemannian structure.

We *choose* a metric for algorithmic purposes. Akin to preconditioning.

# Riemannian submanifolds

Let the embedding space of $\mathcal{M}$ be a Euclidean space $\mathcal{E}$ with metric $\langle\cdot,\cdot\rangle$.
For example: $\mathcal{E} = \mathbf{R}^n$ and $\langle u, v\rangle = u^\top v$ for all $u, v \in \mathbf{R}^n$.




A convenient choice of Riemannian structure for $\mathcal{M}$ is to let:

$$\langle u, v\rangle_x = \langle u, v\rangle.$$

This is well defined because $u, v \in \mathrm{T}_x\mathcal{M}$ are, in particular, elements of $\mathcal{E}$.

This is a Riemannian metric. With it, $\mathcal{M}$ is a Riemannian submanifold of $\mathcal{E}$.



!! A Riemannian submanifold is *not* just a submanifold that is Riemannian !!

# Riemannian gradients

The Riemannian gradient of a smooth $f : \mathcal{M} \to \mathbf{R}$ is the vector field $\mathrm{grad} f$ defined by:

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

Claim: $\mathrm{grad} f$ is a well-defined smooth vector field.

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

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

where $\mathrm{Proj}_x$ is the orthogonal projector from $\mathcal{E}$ to $\mathrm{T}_x\mathcal{M}$ and $\bar{f}$ is a smooth extension of $f$.

$$\left\langle \nabla \bar{f}(x), v \right\rangle = \mathrm{D}\bar{f}(x)[v] = \lim_{t \to 0} \frac{\bar{f}(x + tv) - \bar{f}(x)}{t}$$

$$\nabla^2 \bar{f}(x)[v] = \mathrm{D}\left(\nabla \bar{f}\right)(x)[v] = \lim_{t \to 0} \frac{\nabla \bar{f}(x + tv) - \nabla \bar{f}(x)}{t}$$

# Riemannian Hessians

The Riemannian Hessian of $f$ at $x$ should be a symmetric linear map $\mathrm{Hess}f(x)\colon \mathrm{T}_x\mathcal{M} \to \mathrm{T}_x\mathcal{M}$ describing gradient change.

Since $\mathrm{grad}f\colon \mathcal{M} \to \mathrm{T}\mathcal{M}$ is a smooth map from one manifold to another, a natural first attempt is:

$$\mathrm{Hess}f(x)[v] \stackrel{?}{=} \mathrm{D}\,\mathrm{grad}f(x)[v].$$

However, this does not produce tangent vectors in general.

To overcome this issue, we need a new derivative for vector fields: a Riemannian connection.

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

$$\begin{aligned}\mathrm{Hess}f(x)[v] &= \mathrm{Proj}_x(\mathrm{D}\,\mathrm{grad}f(x)[v]) \\ &= \mathrm{Proj}_x\big(\nabla^2 \bar{f}(x)[v]\big) + W\left(v, \mathrm{Proj}_x^\perp\left(\nabla \bar{f}(x)\right)\right)\end{aligned}$$

where $W$ is the Weingarten map of $\mathcal{M}$.

$$\langle \nabla \bar{f}(x), v \rangle = \mathrm{D}\bar{f}(x)[v] = \lim_{t\to 0} \frac{\bar{f}(x+tv) - \bar{f}(x)}{t}$$

$$\nabla^2 \bar{f}(x)[v] = \mathrm{D}\big(\nabla \bar{f}\big)(x)[v] = \lim_{t\to 0} \frac{\nabla \bar{f}(x+tv) - \nabla \bar{f}(x)}{t}$$

# Example: Rayleigh quotient optimization

Compute the smallest eigenvalue of a symmetric matrix $A \in \mathbf{R}^{n \times n}$ :

$$\min_{x \in \mathcal{M}} \tfrac{1}{2} x^\top A x \qquad \text{with} \qquad \mathcal{M} = \{x \in \mathbf{R}^n : x^\top x = 1\}$$

The cost function $f : \mathcal{M} \to \mathbf{R}$ is the restriction of the smooth function $\bar{f}(x) = \frac{1}{2} x^\top A x$ from $\mathbf{R}^n$ to $\mathcal{M}$.

Tangent spaces $\quad \mathrm{T}_x\mathcal{M} = \{v \in \mathbf{R}^n : x^\top v = 0\}$.

Make $\mathcal{M}$ into a Riemannian submanifold of $\mathbf{R}^n$ with $\langle u, v \rangle = u^\top v$.

Projection to $\mathrm{T}_x\mathcal{M}$: $\quad \mathrm{Proj}_x(z) = z - (x^\top z)x$.

Gradient of $\bar{f}$: $\quad \nabla \bar{f}(x) = Ax$.

Gradient of $f$: $\quad \mathrm{grad} f(x) = \mathrm{Proj}_x\left(\nabla \bar{f}(x)\right) = Ax - (x^\top A x)x$.

Differential of $\mathrm{grad} f$: $\quad \mathrm{Dgrad} f(x)[v] = Av - (v^\top A x + x^\top A v)x - (x^\top A x)v$.

Hessian of $f$: $\quad \mathrm{Hess} f(x)[v] = \mathrm{Proj}_x(\mathrm{Dgrad} f(x)[v]) = \mathrm{Proj}_x(Av) - (x^\top A x)v$.



The following are equivalent for $x \in \mathcal{M}$: $x$ is a global minimizer; $x$ is a unit-norm eigenvector of $A$ for the least eigenvalue; $\mathrm{grad} f(x) = 0$ and $\mathrm{Hess} f(x) \succcurlyeq 0$.

# Basic optimization algorithms

Algorithms hop around the manifold using a retraction:

$$x_{k+1} = R_{x_k}(s_k)$$

with some algorithm-specific tangent vector $s_k \in \mathrm{T}_{x_k}\mathcal{M}$.

E.g., gradient descent: $s_k = -t_k \mathrm{grad} f(x_k)$

Newton's method: $\mathrm{Hess} f(x_k)[s_k] = -\mathrm{grad} f(x_k)$

Convergence analyses rely on Taylor expansions of $f$ along retractions.

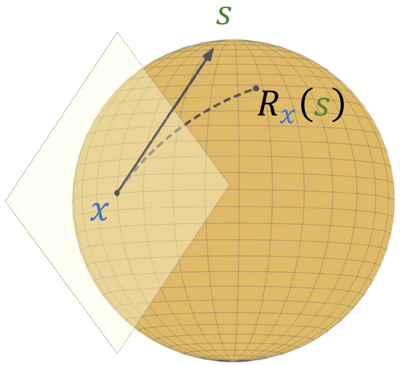
For second-order retractions (e.g., metric projection on Riemannian submanifold):

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

# Gradient descent on $\mathcal{M}$

**A1** $f(x) \geq f_{\text{low}}$ for all $x \in \mathcal{M}$

**A2** $f\big(R_x(s)\big) \leq f(x) + \langle s, \operatorname{grad} f(x) \rangle_x + \frac{L}{2}\|s\|_x^2$




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

Complexity: $\left[\min_{k<K}\|\operatorname{grad} f(x_k)\|_{x_k}\right] \leq \sqrt{\frac{2L(f(x_0)-f_{\text{low}})}{K}}$ (same as Euclidean case)

---

$$\textbf{A2} \Rightarrow f(x_{k+1}) \leq f(x_k) - \frac{1}{L}\|\operatorname{grad} f(x_k)\|_{x_k}^2 + \frac{1}{2L}\|\operatorname{grad} f(x_k)\|_{x_k}^2$$

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

$$\textbf{A1} \Rightarrow f(x_0) - f_{\text{low}} \geq f(x_0) - f(x_K) = \sum_{k=0}^{K-1} f(x_k) - f(x_{k+1}) \geq \frac{K}{2L}\min_{k<K}\|\operatorname{grad} f(x_k)\|_{x_k}^2$$

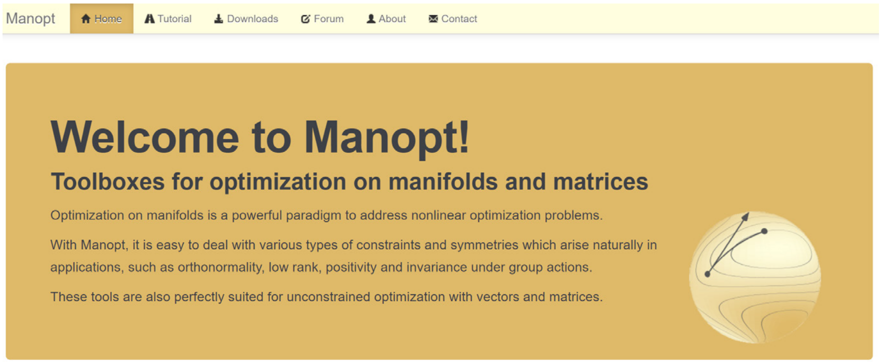
# Manopt: user-friendly software

Manopt is a family of toolboxes for Riemannian optimization.

Go to www.manopt.org for code, a tutorial, a forum, and a list of other software.

Matlab example for $\min_{\|x\|=1} x^\top A x$:

```
problem.M = spherefactory(n);
problem.cost = @(x) x'*A*x;
problem.egrad = @(x) 2*A*x;
x = trustregions(problem);
```




With Bamdev Mishra, P.-A. Absil & R. Sepulchre

Lead by J. Townsend, N. Koep & S. Weichwald

Lead by Ronny Bergmann

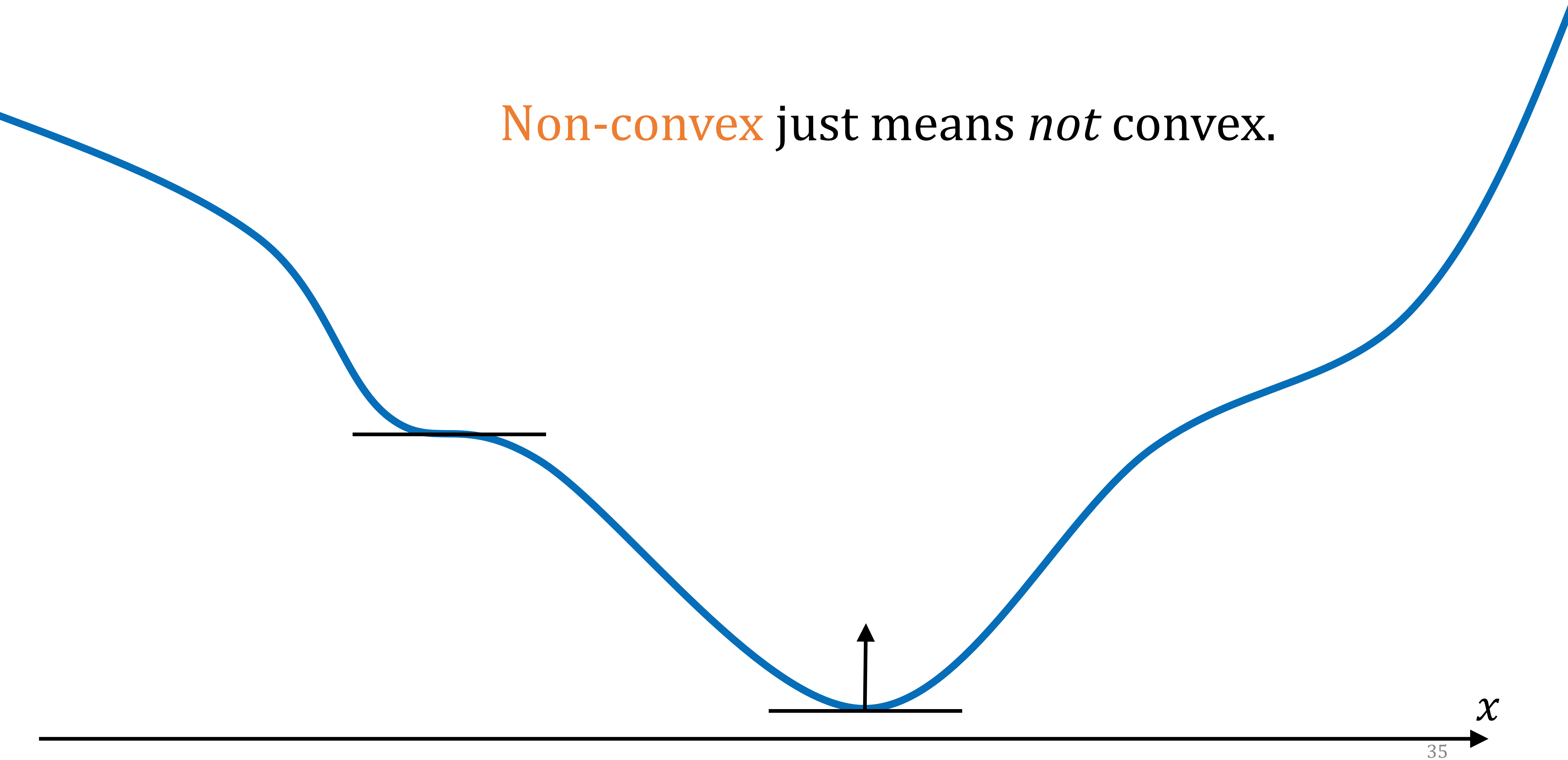
# Active research directions

- More algorithms: nonsmooth, stochastic, parallel, quasi-Newton, ...
- Constrained optimization on manifolds
- Applications, old and new (electronic structure, deep learning)
- Complexity (upper and lower bounds)
- Role of curvature
- Geodesic convexity
- Randomized algorithms
- Broader generalizations: manifolds with a boundary, algebraic varieties
- Benign non-convexity

*"... in fact, the great watershed in optimization isn't between linearity and nonlinearity, but* ***convexity*** *and* ***non-convexity****."*

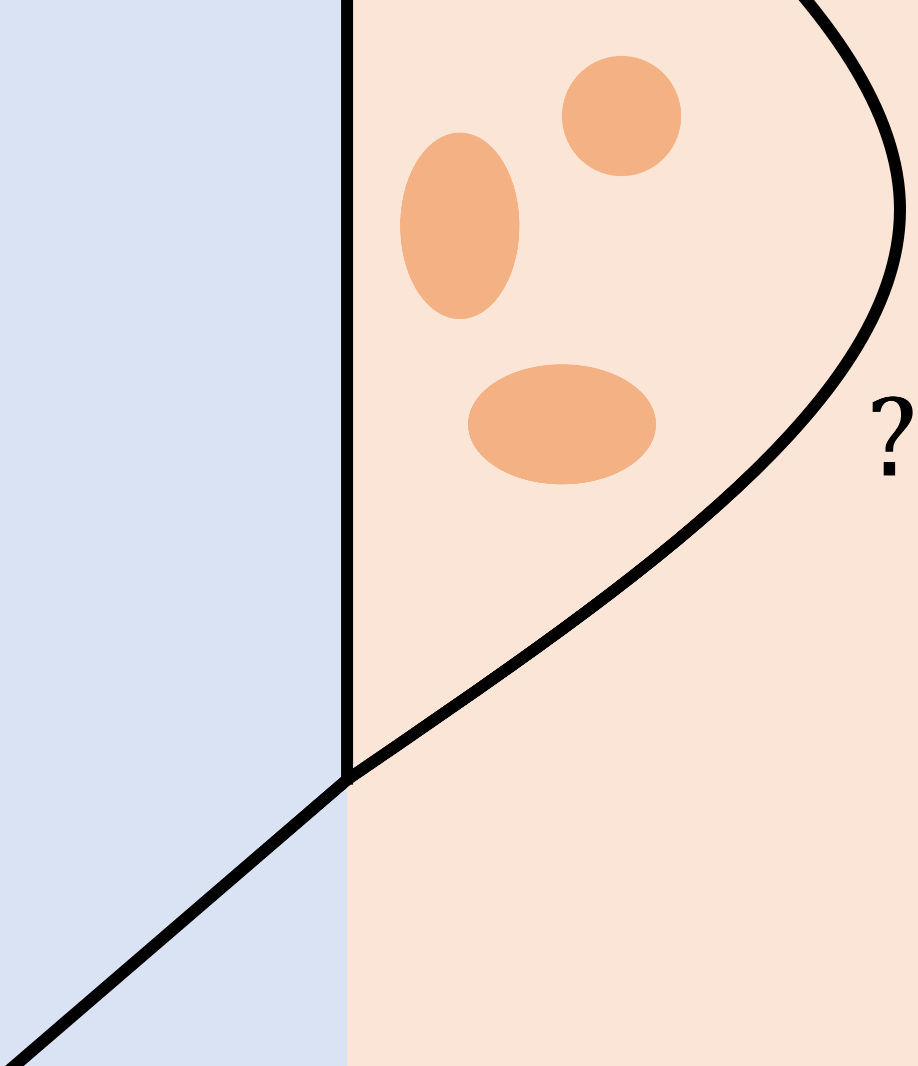
R. T. Rockafellar, in SIAM Review, 1993

Non-convex just means *not* convex.



*"... in fact, the great watershed in optimization isn't between linearity and nonlinearity, but* ***convexity*** *and* ***non-convexity****."*

R. T. Rockafellar, in SIAM Review, 1993



# Non-convexity can be benign

This can mean various things. Theorem templates are on a spectrum:

*"If {conditions}, necessary optimality conditions are sufficient."*

⋮

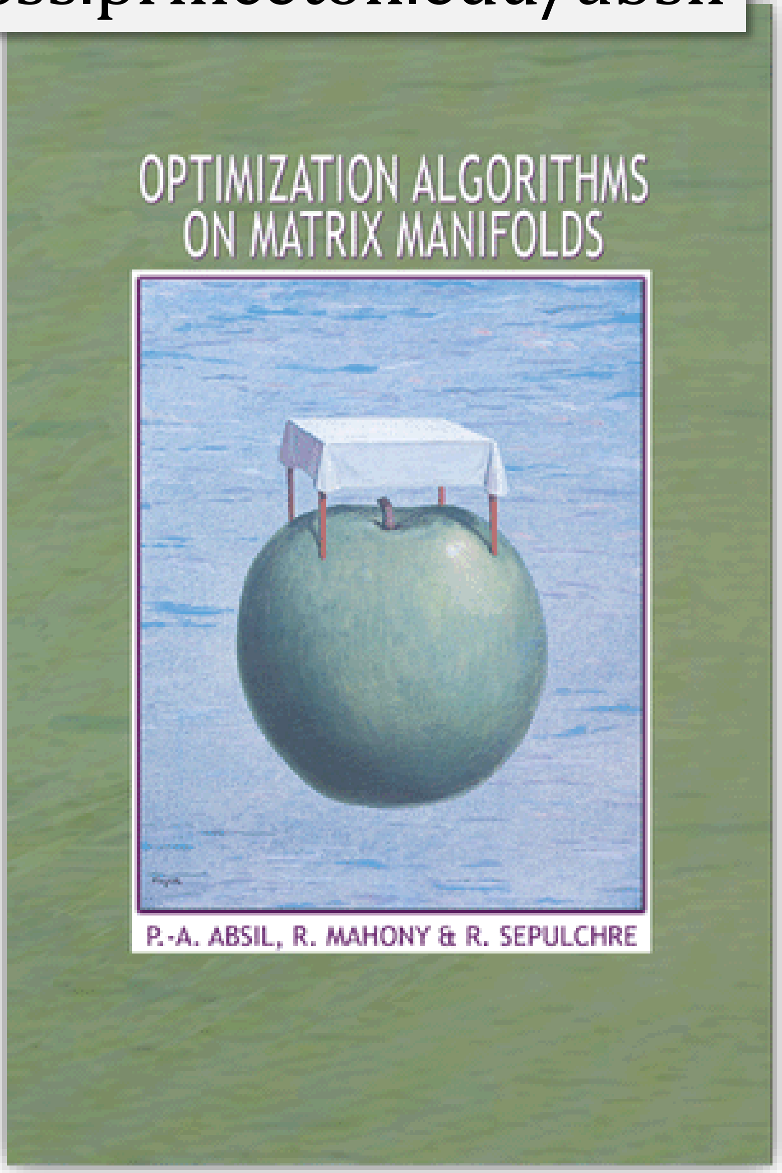
*"If {conditions}, we can initialize a specific algorithm well."*

The conditions (often on data) may be generous (e.g., genericity) or less so (e.g., high-probability event for non-adversarial distribution.)

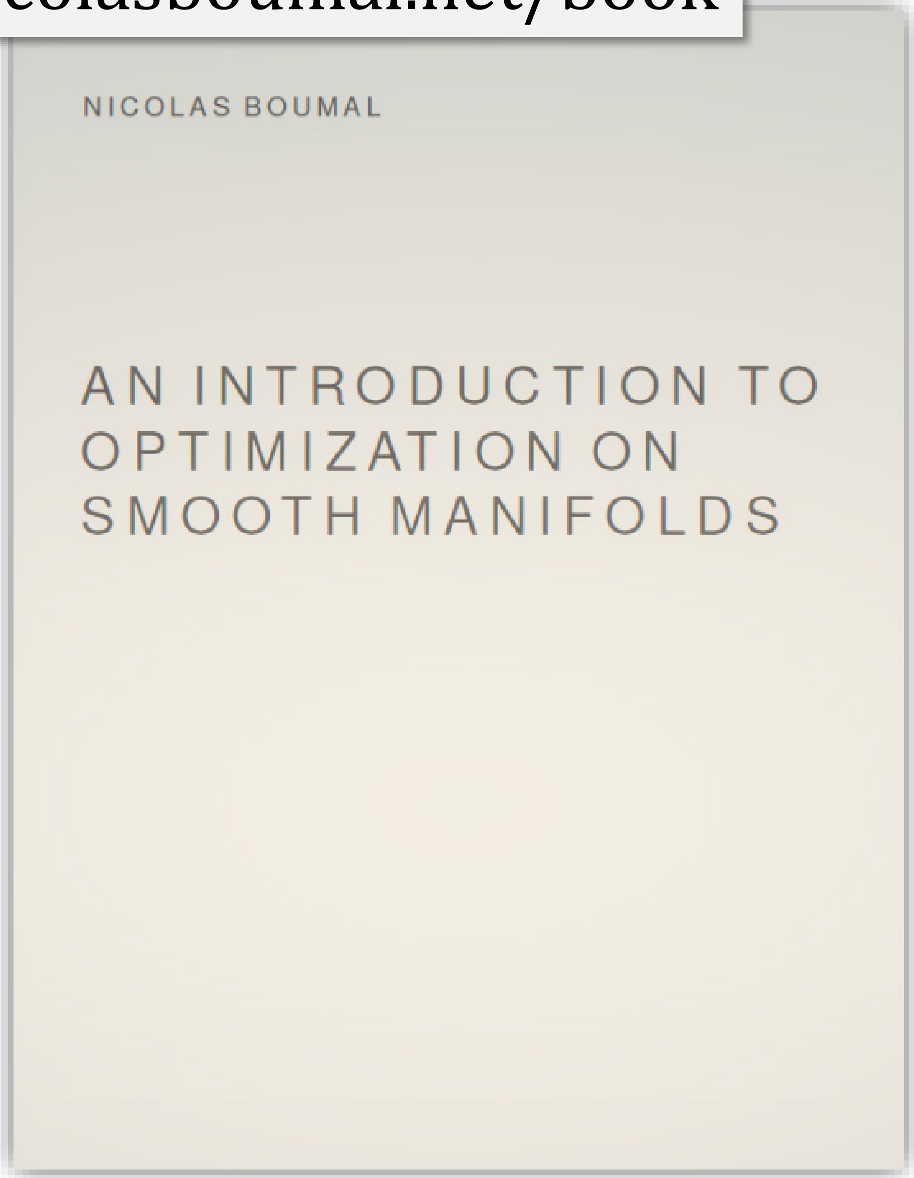
Geometry and symmetry seem to play an outsized role.

See for example Zhang, Qu & Wright, arxiv:2007.06753, for a review.

press.princeton.edu/absil




nicolasboumal.net/book




manopt.org