CS 468 – Lecture V

Non-Euclidean Methods in Machine Learning

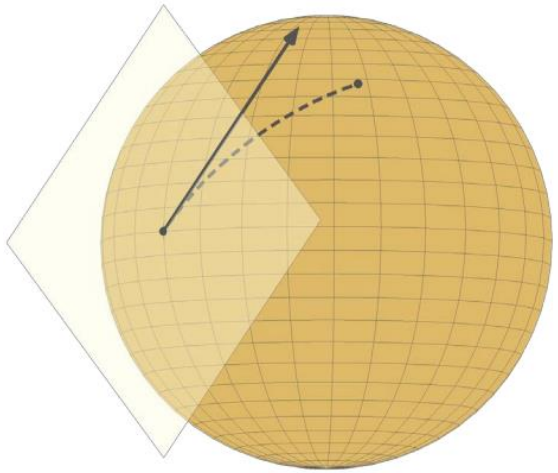
Optimization on Riemannian Manifolds

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

Tolga Birdal

Ines Chami

Leonidas Guibas



cs468.stanford.edu

Optimization on Riemannian Manifolds

Partial Slide Credits: Nicolas Boumal

An introduction to optimization on smooth manifolds

https://web.math.princeton.edu/~nboumal/book/index.html

Constrained vs. Riemannian Optimization

$$\min_x f(x) \quad \text{subject to} \quad \underbrace{x \in \mathcal{M}}$$

Constraints on Parameters

Linear spaces

Unconstrained; linear equality constraints

Low rank (matrices, tensors)

Recommender systems, large-scale Lyapunov equations, …

Orthonormality (Grassmann, Stiefel, rotations)

Dictionary learning, SfM, SLAM, PCA, ICA, SBM, Electr. Struct. Comp.…

Positivity (positive definiteness, positive orthant)

Metric learning, Gaussian mixtures, diffusion tensor imaging, …

Symmetry (quotient manifolds)

Invariance under group actions

Constrained vs. Riemannian Optimization

$$\min_x f(x) \quad \text{subject to} \quad \underbrace{x \in \mathcal{M}}$$

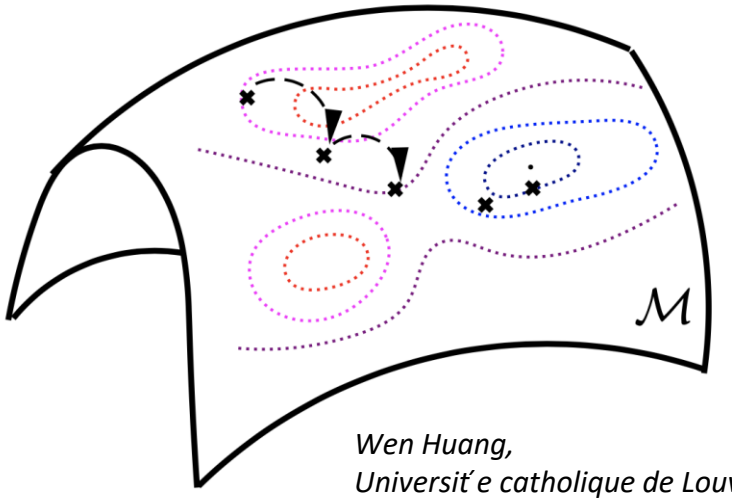
Constraints on Parameters

Unconstrained Optimization

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

Constrained vs. Riemannian Optimization

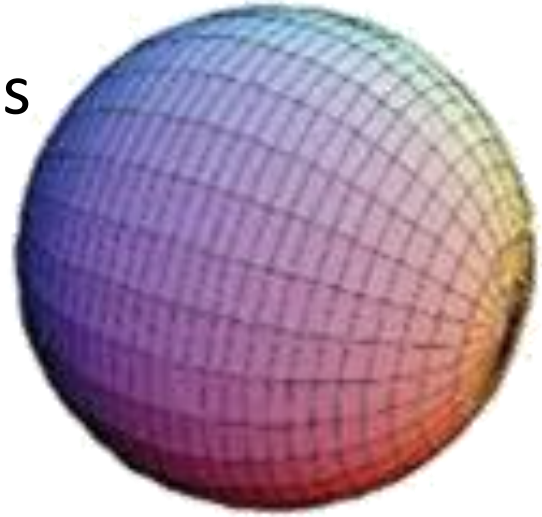
1. We can use unconstrained optimization tools (gradient descent, newton etc.).
2. No need to consider Lagrange multipliers or penalty functions.
3. Theoretical guarantees usually transfer from Euclidean space to Riemannian manifolds.
4. Can be cheaper in terms of resource use depending upon the application.

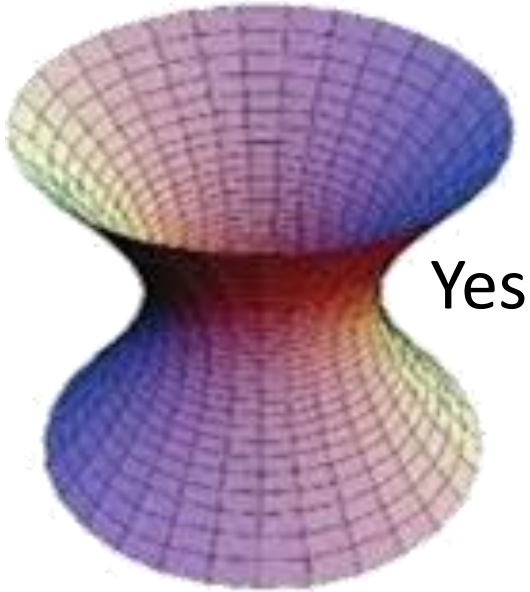


Wen Huang,
Universit´e catholique de Louvain

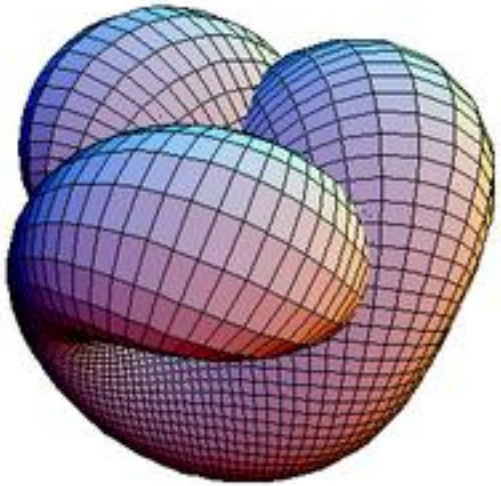
Which Manifolds?

Yes





Yes



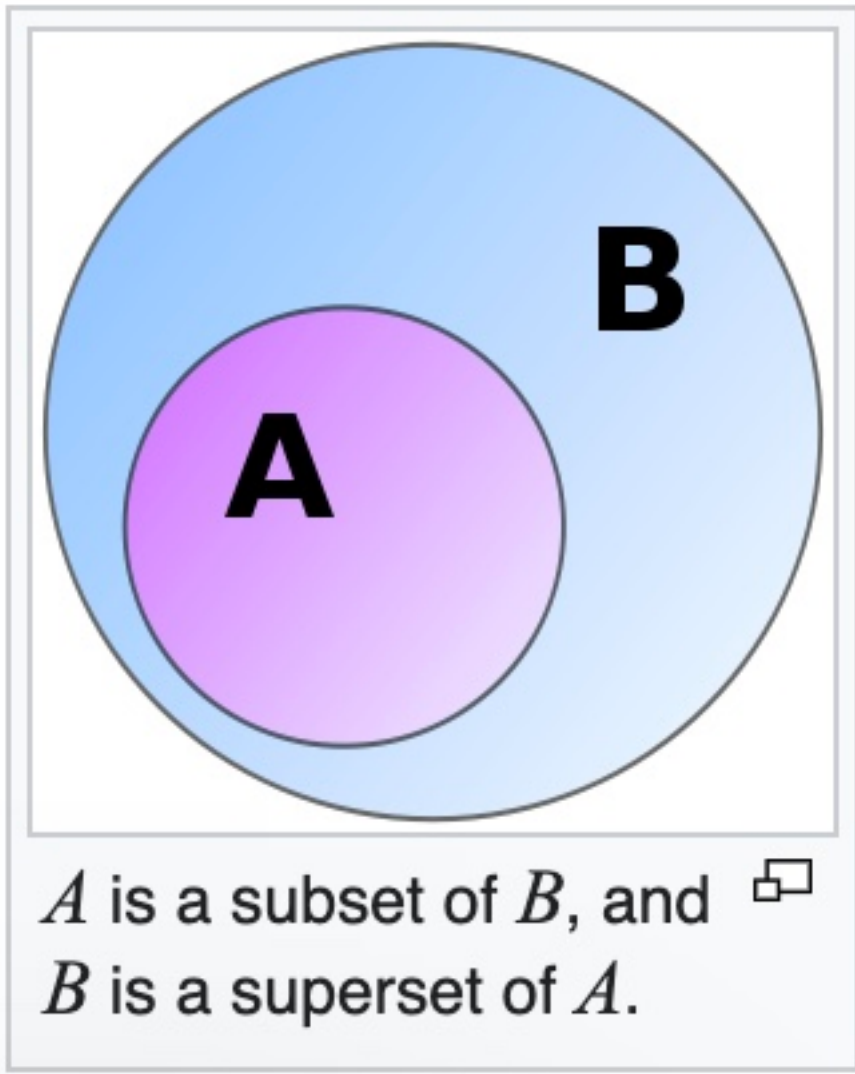
No

We restrict ourselves to *embedded submanifolds* of **linear spaces**.

Let N be a smooth manifold. One possible definition (I believe) for an embedded submanifold of N is some $M \subset N$ that is a (smooth) manifold such that the inclusion $i : M \hookrightarrow N$ is an embedding.

In mathematics, if A is a subset of B, then the **inclusion map** (also **inclusion function**, **insertion**,[1] or **canonical injection**) is the function ι that sends each element x of A to x, treated as an element of B:

$$\iota : A \to B, \qquad \iota(x) = x.$$



A is a subset of B, and
B is a superset of A.

For smoothness of $\mathcal{M}$, our model space is the unit sphere in $\mathbb{R}^n$:

$$\mathrm{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}, \tag{3.2}$$

where $\|x\| = \sqrt{x^\top x}$ is the Euclidean norm on $\mathbb{R}^n$. Intuitively, we think of S^{n-1} as a smooth nonlinear space in $\mathbb{R}^n$. Our definitions below are compatible with this intuition, and we call S^{n-1} an *embedded submanifold* of $\mathbb{R}^n$.

An important element in these definitions is to capture the idea that S^{n-1} can be locally approximated by a linear space around any point x: we call these *tangent spaces*, denoted by $T_x S^{n-1}$. This is as opposed to a cube for which no good linearization exists at the edges. More specifically for our example, S^{n-1} is defined by the constraint $x^\top x = 1$, and we expect that differentiating this constraint should yield a suitable linearization:

$$T_x S^{n-1} = \{v \in \mathbb{R}^n : v^\top x + x^\top v = 0\} = \{v \in \mathbb{R}^n : x^\top v = 0\}. \quad (3.3)$$

In the same spirit, it stands to reason that linear spaces, or open subsets of linear spaces, should also be considered smooth.

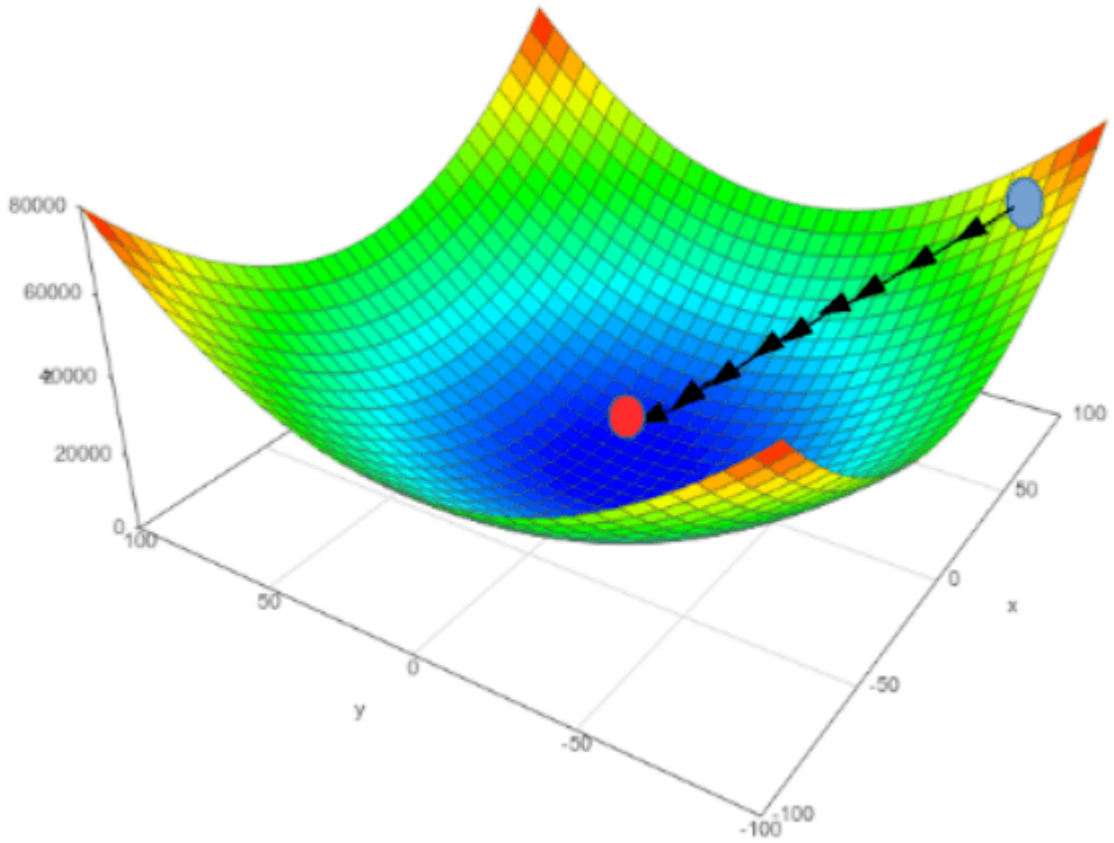
Optimization on a Euclidean Space

$$x_{k+1} = x_k + \Delta x_k = x_k + \tau_k d_k$$

movement

step

previous iterate



Optimization on a Euclidean Space

$$x_{k+1} = x_k + \Delta x_k = x_k + \tau_k d_k$$

movement

previous iterate

step

- Steepest (Gradient) Descent:

update step size

$$x_{k+1} = x_k - \tau_k \nabla f(x_k)$$

- Newton's Method:

Hessian Gradient

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

- Other methods to compute the direction and the step size.

A Naïve Manifold Optimizer: Projected Gradient Descent

Loop:

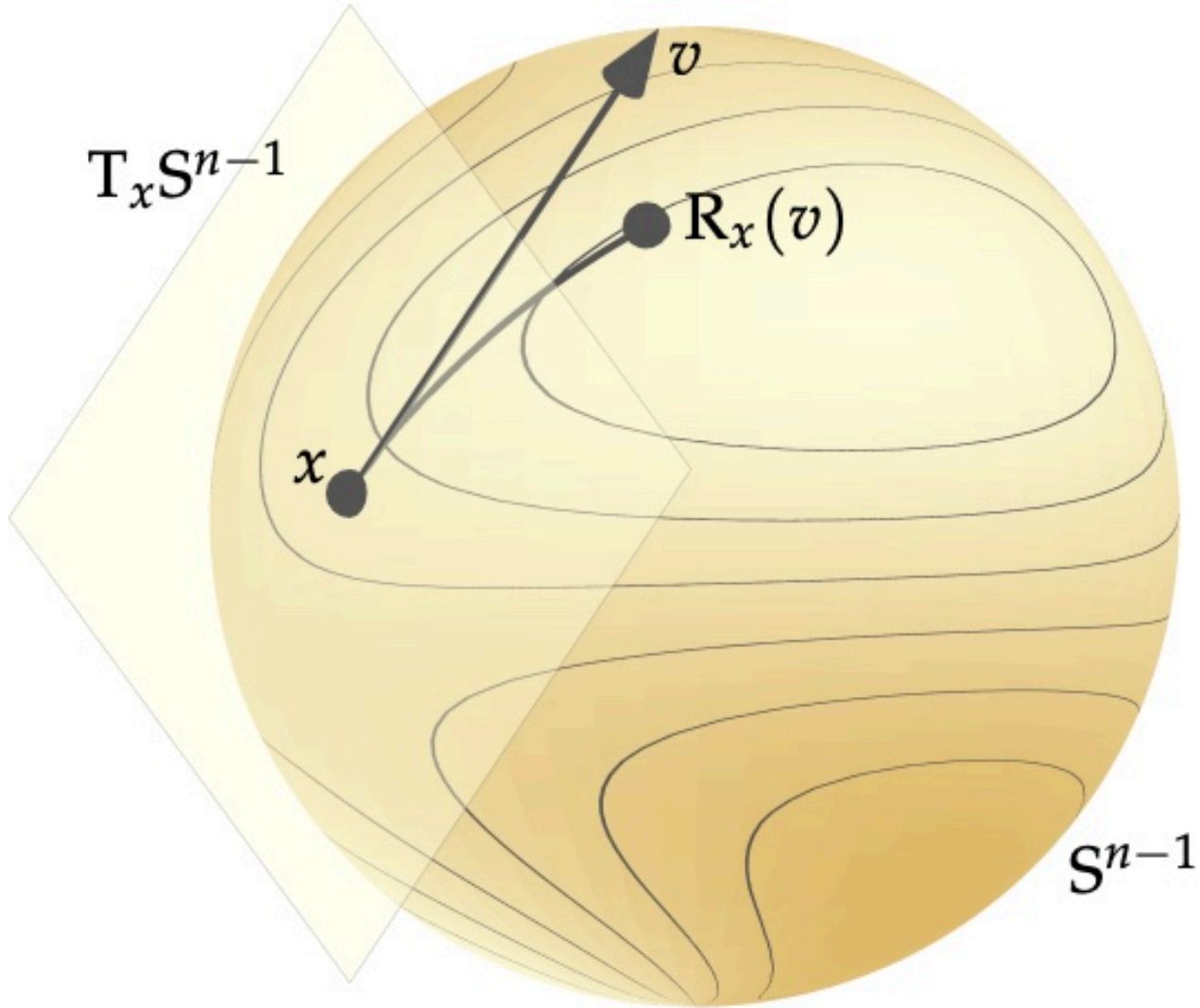
1. Perform the gradient update:

$$\hat{x} = x_k - \tau_k \nabla f(x_k)$$

2. Project on the manifold:

$$x_{k+1} = \Pi_{\mathcal{M}}(\hat{x})$$

Inaccurate & Slow Convergence



Retraction $\mathrm{R}_x(v) = \frac{x+v}{\|x+v\|}$ on the sphere.

Smooth Map

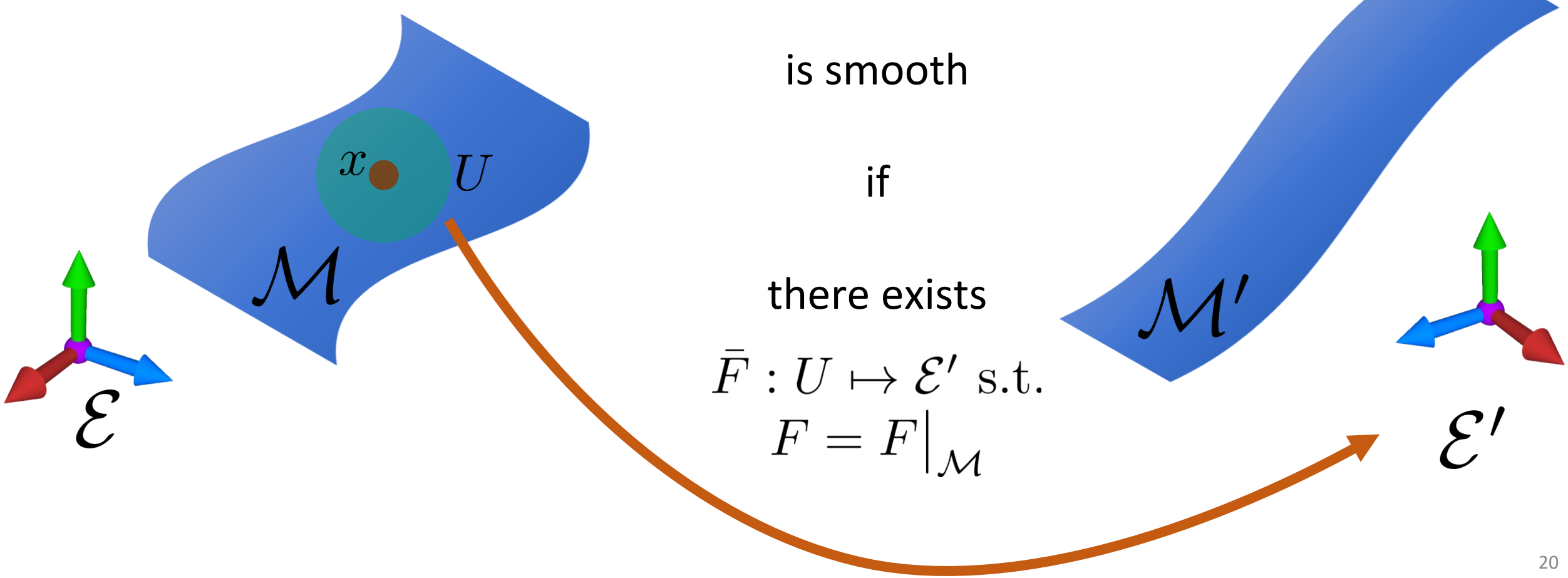
$$F : \mathcal{M} \mapsto \mathcal{M}'$$

is smooth

if

there exists

$$\bar{F} : U \mapsto \mathcal{E}' \text{ s.t. } F = F\big|_{\mathcal{M}}$$



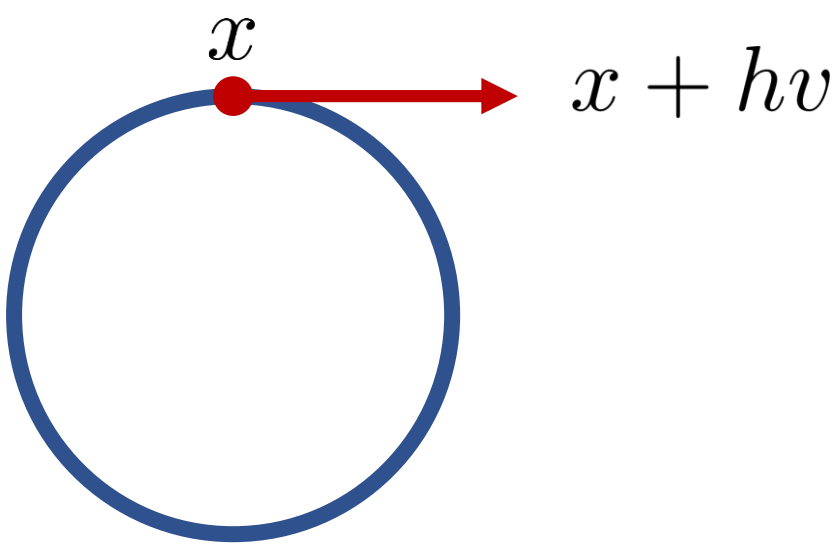
The Differential of a Smooth Map

Directional derivative of a smooth function $f : \mathbb{R}^n \mapsto \mathbb{R}^n$ along a vector v:

$$D_v f = Df(x) \cdot v = \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h}$$

Cannot apply to a smooth manifold $\mathcal{M}$ as $(x + hv)$ might not be belong to $\mathcal{M}$.

For $\bar{F}$, this is always defined, but not for F.



The Differential of a Smooth Map

For $\bar{F} : \mathcal{E} \mapsto \mathcal{E}'$ we have:

$$D_v \bar{F} = D\bar{F}(x) \cdot v = \lim_{h \to 0} \frac{\bar{F}(x + hv) - \bar{F}(x)}{h}$$

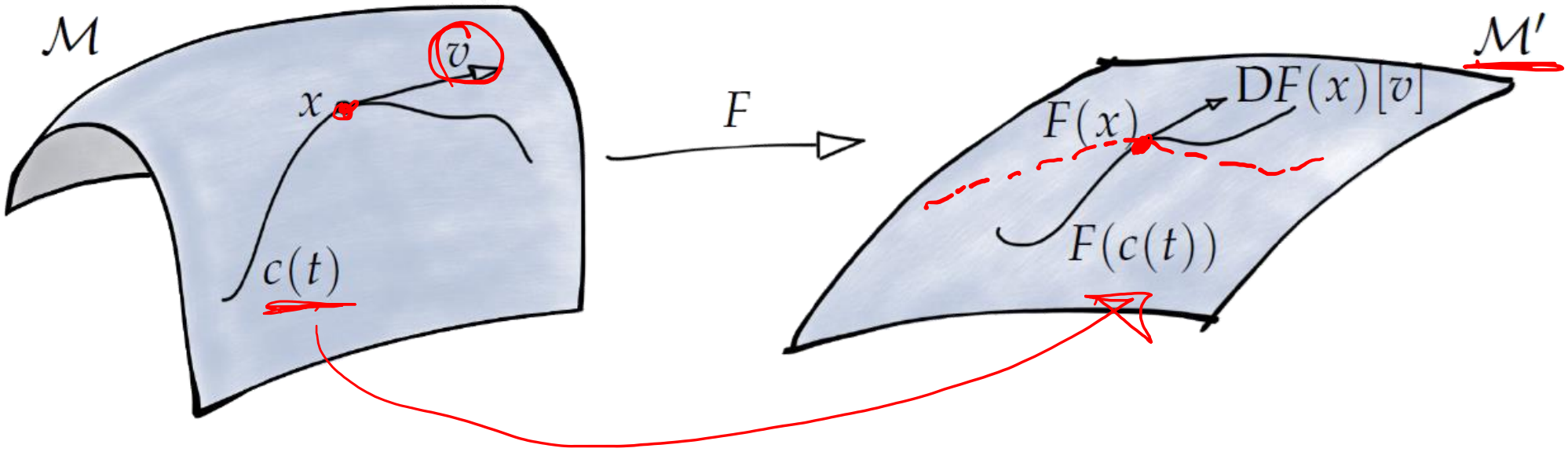
For $F : \mathcal{M} \mapsto \mathcal{M}'$ and $v \in \mathcal{T}_x \mathcal{M}$, we write:

$$D_v F = D_v \bar{F}$$

Does not depend on the choice of the smooth extension!

The Differential of a Smooth Map

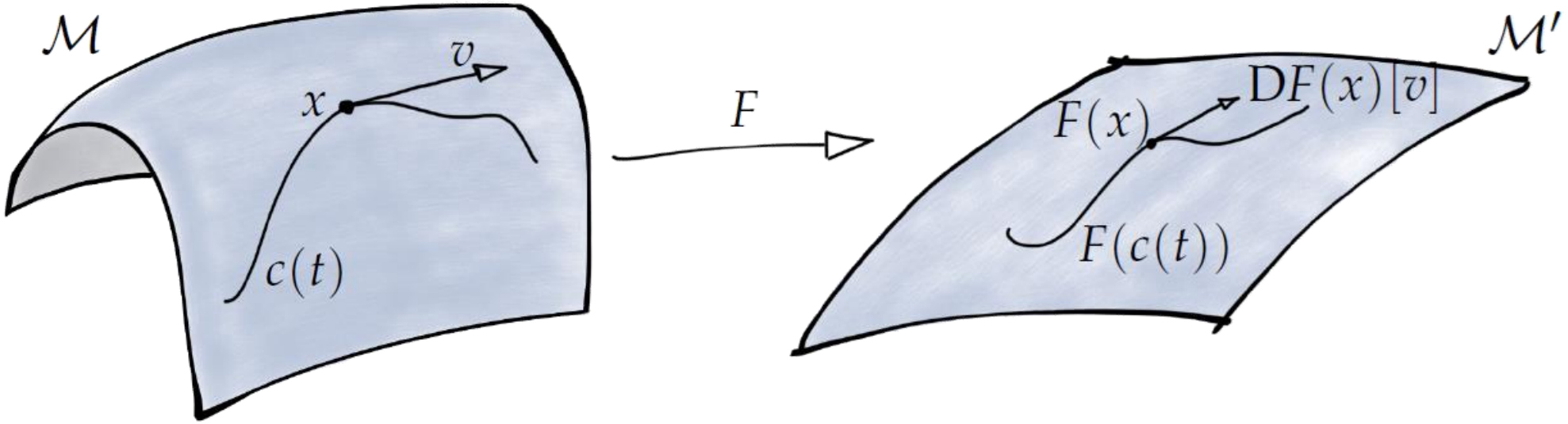
Equivalently, consider a smooth curve $c(t) \in \mathcal{M}$ with $F(c(t)) \in \mathcal{M}'$ being a curve in $\mathcal{M}'$ passing through $F(x)$ with velocity $\mathrm{D}F(x) \cdot v$.



$$DF(x)[v] : \mathcal{T}_x\mathcal{M} \mapsto \mathcal{T}_{F(x)}\mathcal{M}' = \frac{d}{dt} F(c(t))\Big|_{t=0}$$

Does not depend on the choice of the curve!

The Differential of a Smooth Map



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

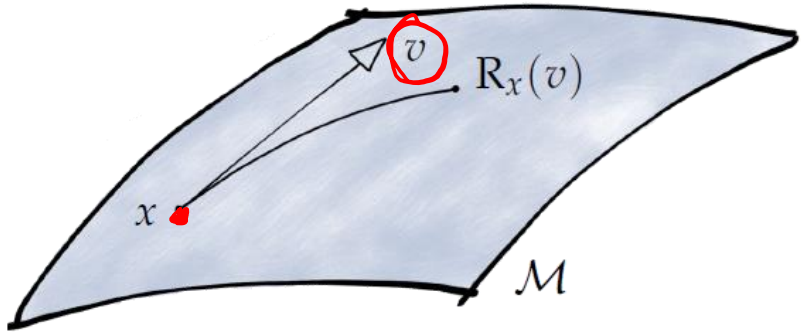
Does not depend on the choice of the curve!

$$DF(x) \cdot v = DF(x)[v] = \lim_{h \to 0} \frac{F(c(t+h)) - F(c(t))}{h}|_{t=0} = (F \circ c)'(0)$$

$$DF(x) \cdot v = DF(x)[v] = \lim_{h \to 0} \frac{F(c(t+h)) - F(c(t))}{h}|_{t=0} = (F \circ c)'(0)$$

How to choose these curves == Retractions

- Smooth choice of curves over the tangent bundle.
- Maps tangent vectors back to the manifold.
- Defines curves in a given direction.



$$c(t) = R(x, tv) = R_x(tv)$$
$$c(0) = x \text{ and } c'(0) = v$$

A **Retraction** map $R : \mathcal{T}_x\mathcal{M} \mapsto \mathcal{M}$ satisfies:

1. R is continuously differentiable.
2. $R_x(0) = x$ *(centering)*
3. $\mathrm{D}R_x(0)[v] = v$ *(local rigidity)*

Identity

In differential geometry, the **tangent bundle** of a differentiable manifold M is a manifold TM which assembles all the tangent vectors in M. As a set, it is given by the disjoint union[note 1] of the tangent spaces of M. That is,

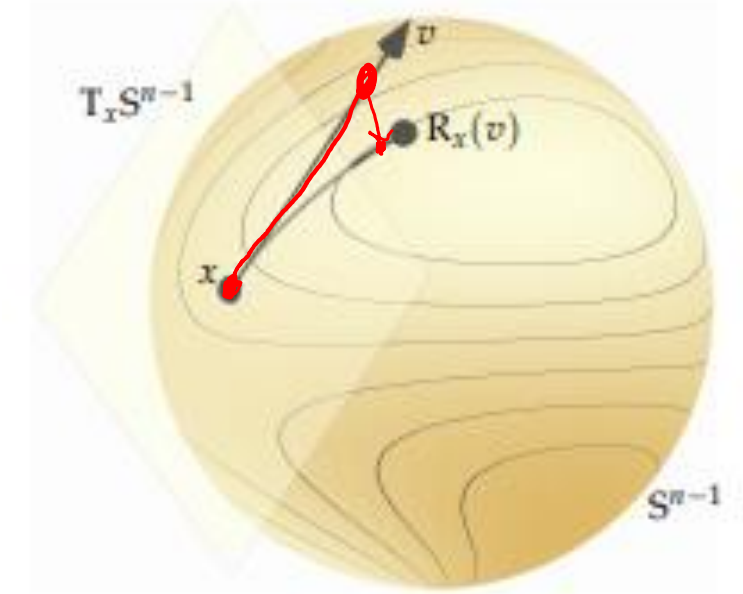
$$\begin{aligned} TM &= \bigsqcup_{x \in M} T_x M \\ &= \bigcup_{x \in M} \{x\} \times T_x M \\ &= \bigcup_{x \in M} \{(x, y) \mid y \in T_x M\} \\ &= \{(x, y) \mid x \in M,\, y \in T_x M\} \end{aligned}$$

Projection as a Retraction

One way to define a retraction on the sphere $\mathbb{S}^{d-1}$:

$$R_x(v) = \frac{x+v}{\|x+v\|}$$

Exercise: Verify that this is indeed a valid retraction.



Retraction $R_x(v) = \frac{x+v}{\|x+v\|}$ on the sphere.

Example 3.40. *Let x be a point on the sphere S^{n-1} and let v be tangent at x, that is, $x^\top v = 0$. To move away from x along v while remaining on the sphere, one way is to take the step in $\mathbb{R}^n$ then to project back to the sphere:*

$$\mathrm{R}_x(v) \triangleq \frac{x+v}{\|x+v\|} = \frac{x+v}{\sqrt{1+\|v\|^2}}. \tag{3.31}$$

Consider the curve $c\colon \mathbb{R} \to \mathrm{S}^{n-1}$ *defined by:*

$$c(t) = \mathrm{R}_x(tv) = \frac{x + tv}{\sqrt{1 + t^2\|v\|^2}}.$$

Evidently, $c(0) = x$. *This holds because* $\mathrm{R}_x(0) = x$. *Furthermore, one can compute* $c'(0) = v$, *that is: locally around* x, *up to first order, the retraction curve moves along* v. *Another way to state this is via the chain rule:*

$$v = c'(0) = \mathrm{D}\mathrm{R}_x(0)[v],$$

where $\mathrm{D}\mathrm{R}_x(0)$ *is understood as per Definition 3.27 for* $\mathrm{R}_x\colon \mathrm{T}_x\mathrm{S}^{n-1} \to \mathrm{S}^{n-1}$. *In other words:* $\mathrm{D}\mathrm{R}_x(0)\colon \mathrm{T}_x\mathrm{S}^{n-1} \to \mathrm{T}_x\mathrm{S}^{n-1}$ *is the identity map.*

Needed for Optimization

1. A representation for points $x \in \mathcal{M}$, for tangent spaces $\mathcal{T}_x\mathcal{M}$ and Riemannian metrics $g_x(\cdot, \cdot)$,
2. A map from the tangent space to the manifold: $R_x : \mathcal{T}_x\mathcal{M} \mapsto \mathcal{M}$
3. Expressions for $f(x)$, $\mathrm{grad} f(x)$ and $\mathrm{Hess} f(x)$
4. Notion of *vector transport* for second order methods

Do we need the true geodesics to move?

NO!

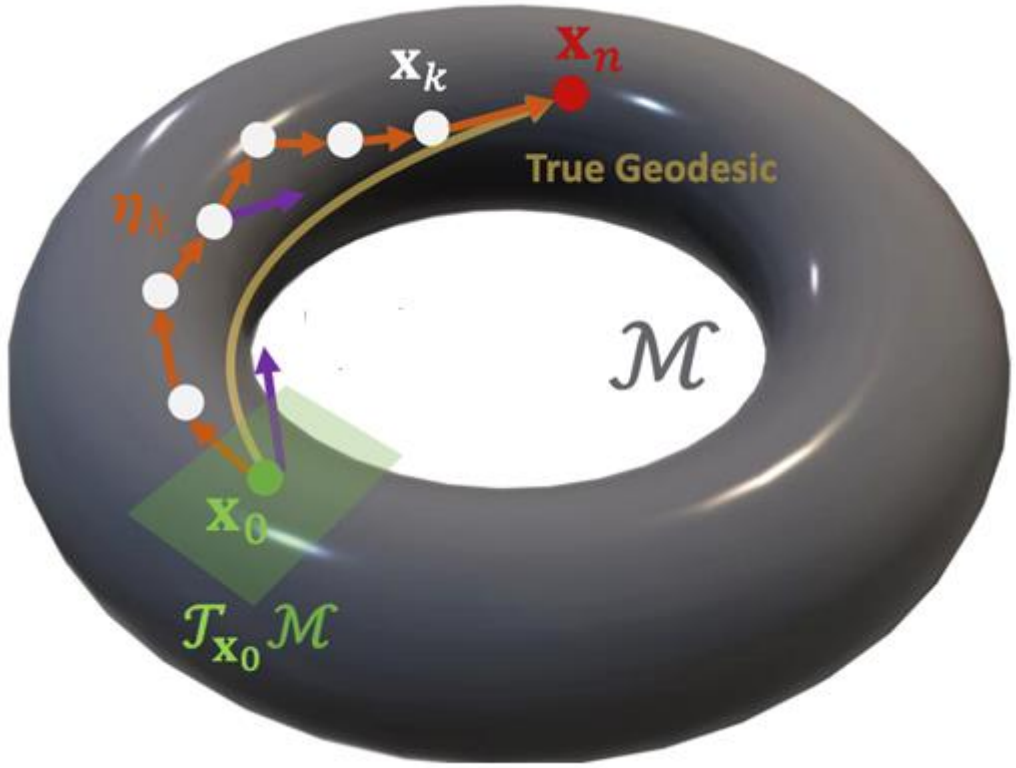
A Less Naïve Manifold Optimizer

Euclidean

$$x_{k+1} = x_k + \tau_k d_k$$

Riemannian

$$x_{k+1} = R_{x_k}(\tau_k \eta_k), \;\; \eta_k \in \mathcal{T}_x\mathcal{M}$$



Birdal, Tolga, and Umut Simsekli. "Probabilistic permutation synchronization using the riemannian structure of the birkhoff polytope." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11105-11116. 2019.

Riemannian Gradient & Hessian

Riemannian gradient of $f(x)$ at x is the unique tangent vector in $\mathcal{T}_x\mathcal{M}$ satisfying $\forall\, v \in \mathcal{T}_x\mathcal{M}$:

$$\mathrm{D}F(x)[v] = \langle \mathrm{grad} f(x), v \rangle_x$$

Direction of steepest ascent

$$\text{If } x \text{ is a local optimum of } f, \text{ then } \mathrm{grad} f(x) = 0.$$

Riemannian Hessian of $f(x)$ at x is a linear operator from $\mathcal{T}_x\mathcal{M} \mapsto \mathcal{T}_x\mathcal{M}$:

$$\mathrm{Hess} f(x) : \mathcal{T}_x\mathcal{M} \mapsto \mathcal{T}_x\mathcal{M} : v \mapsto \underbrace{\nabla_v \mathrm{grad} f}$$

Affine Connection

More on that later.

Riemannian Gradient

1. $\mathrm{D}(f \circ R_x)(0)[v] = \mathrm{D}f(\underbrace{R_x(0)}_{x})[\underbrace{\mathrm{D}R_x(0)}_{\text{Identity}}[v]] = \mathrm{D}f(x)[v]$

2. Hence, $\langle \mathrm{grad}(f \circ R_x)(0), v \rangle_x = \langle \mathrm{grad} f(x), v \rangle_x$

3. $\mathrm{D}f(x)[v] = \mathrm{D}\bar{f}(x)[v] = \langle v, \mathrm{grad}\bar{f}(x) \rangle$ (Euclidean)

Riemannian Gradient

Observe that $\mathcal{T}_x\mathcal{M} \subset \mathcal{E}$ and $\operatorname{grad}\bar{f}(x) \in \mathcal{E}$. Hence:

$$\operatorname{grad}\bar{f}(x) = \operatorname{grad}\bar{f}(x)_{||} + \operatorname{grad}\bar{f}(x)_{\perp}$$

Tangential Orthogonal

Since $\forall v \in \mathcal{T}'_x \quad \langle v, \operatorname{grad}\bar{f}(x)_{\perp}\rangle = 0$:

$$\begin{aligned}\langle v, \operatorname{grad} f(x)\rangle_x &= \langle v, \operatorname{grad}\bar{f}(x)\rangle_x \\ &= \langle v, \operatorname{grad}\bar{f}(x)_{||} + \operatorname{grad}\bar{f}(x)_{\perp}\rangle \\ &= \langle v, \operatorname{grad}\bar{f}(x)_{||}\rangle\end{aligned}$$

Riemannian Gradient

$$\mathrm{grad} f(x) = grad\bar{f}(x)_{||}$$

To compute the **Riemannian gradient**:

1. Obtain an expression for the classical gradient.

2. Orthogonally project to the tangent space (cancel the normal component)

Riemannian Gradient

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

where $\Pi_x \triangleq \mathrm{Proj}_x$ is a **projector**:

$$\langle u - \Pi_x(u), v \rangle = 0 \ \forall v \in \mathcal{T}_x\mathcal{M} \text{ and } u \in \mathcal{E}$$

$$\Pi_x \circ \Pi_x = \Pi_x$$

Consider the Sphere

$$T_x\mathcal{S}^{d-1} = \{v \in \mathbb{R}^n \,:\, x^\top v = 0\} = \{v \in \mathbb{R}^n \,:\, \langle x, v\rangle = 0\}$$

$$\Pi_x(u) = u - (x^\top u)x = (I - xx^\top)u$$

A Less Naïve Manifold Optimizer

Loop:

1. Perform the gradient update:

$$\hat{x} = x_k - \tau_k \mathrm{grad} f(x_k)$$

2. Project on the manifold:

$$x_{k+1} = \Pi_{\mathcal{M}}(\hat{x})$$

Inaccurate & Slow Convergence

Riemannian Gradient Descent

Loop:

$$x_{k+1} = R_{x_k}(-\tau_k \operatorname{grad} f(x_k))$$

Provable convergence to local minimum.

Riemannian Gradient Descent – Taylor Perspective

The composition $f \circ R_x : \mathrm{T}_x\mathcal{M} \to \mathbf{R}$ is on a linear space, hence we may Taylor expand it:

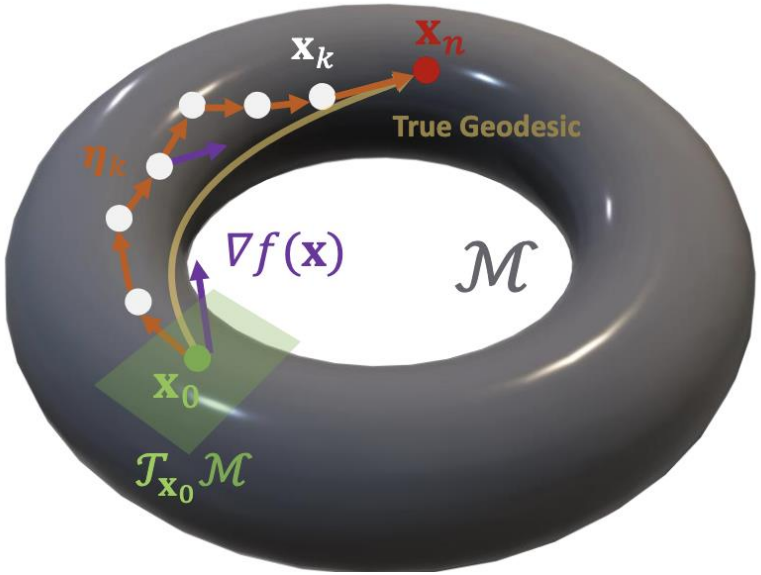
$$
\begin{aligned}
& f\big(R_x(v)\big) \\
&= f\big(R_x(0)\big) + \langle \mathrm{grad}(f \circ R_x)(0), v \rangle_x + O(\|v\|_x^2) \\
&= f(x) + \langle \mathrm{grad} f(x), v \rangle_x + O(\|v\|_x^2)
\end{aligned}
$$

This will take us to 2nd order methods.

A Generic Riemannian Optimizer

Sufficient to describe a large family of optimization methods.

input: A Riemannian manifold $\mathcal{M}$, a retraction operator R
while $\mathbf{x}_k$ *does not sufficiently minimize* f **do**
  Pick a gradient related descent direction $\boldsymbol{\eta}_k \in \mathcal{T}_{\mathbf{x}_k}\mathcal{M}$.
  Choose a retraction $R_{\mathbf{x}_k} : \mathcal{T}_{\mathbf{x}_k}\mathcal{M} \to \mathcal{M}$.
  Choose a step length $\tau_k \in \mathbb{R}$.
  Set $\mathbf{x}_{k+1} \leftarrow R_{\mathbf{x}_k}(\tau_k \boldsymbol{\eta}_k)$.
  $k \leftarrow k + 1$.



Birdal, Tolga, and Umut Simsekli. "Probabilistic permutation synchronization using the riemannian structure of the birkhoff polytope." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11105-11116. 2019.

Riemannian Steepest Descent with Armijo Line Search

- Accelerated scheme with adaptive step size
- Determines the step size via Armijo Condition

input: A Riemannian manifold $\mathcal{M}$, a retraction operator R, the projection operator onto the tangent space $\Pi_{\mathbf{x}_k} : \mathbb{R}^n \to \mathcal{T}_{\mathbf{x}_k}\mathcal{M}$, a real-valued, differentiable potential energy f, initial iterate $\mathbf{x}_0 \in \mathcal{M}$ and the Armijo line search scalars including c.

while $\mathbf{x}_k$ *does not sufficiently minimize* f **do**

// Euclidean gradient to Riemannian direction

$\boldsymbol{\eta}_k \leftarrow -\text{grad} f(\mathbf{x}_k) \triangleq \Pi_{\mathbf{x}_k}(-\nabla f(\mathbf{x}_k))$.

Select $\mathbf{x}_{k+1}$ such that:

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq c\big(f(\mathbf{x}_k) - f(R_{\mathbf{x}}(\tau_k \boldsymbol{\eta}_k))\big), \tag{22}$$

where τ_k is the Armijo step size.

$k \leftarrow k + 1$.

$x_{k+1} = x_k + \tau_k \eta_k$

Armijo condition

Birdal, Tolga, and Umut Simsekli. "Probabilistic permutation synchronization using the riemannian structure of the birkhoff polytope." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11105-11116. 2019.

Determining the Step Size

Options:

1. Fix $\tau_k = \tau \, \forall k$
2. Optimal τ_k that minimizes $f(R_{x_k}(-t\text{grad}f(x_k))$ can be found cheaply in rare cases.
3. **Backtracking**: Start with an initial guess and iteratively reduce until the step size is deemed acceptable. (conditions vary)