Sesión 1:
Optimización con restricciones en una variedad

Daniel López-Castaño
Johns Hopkins University

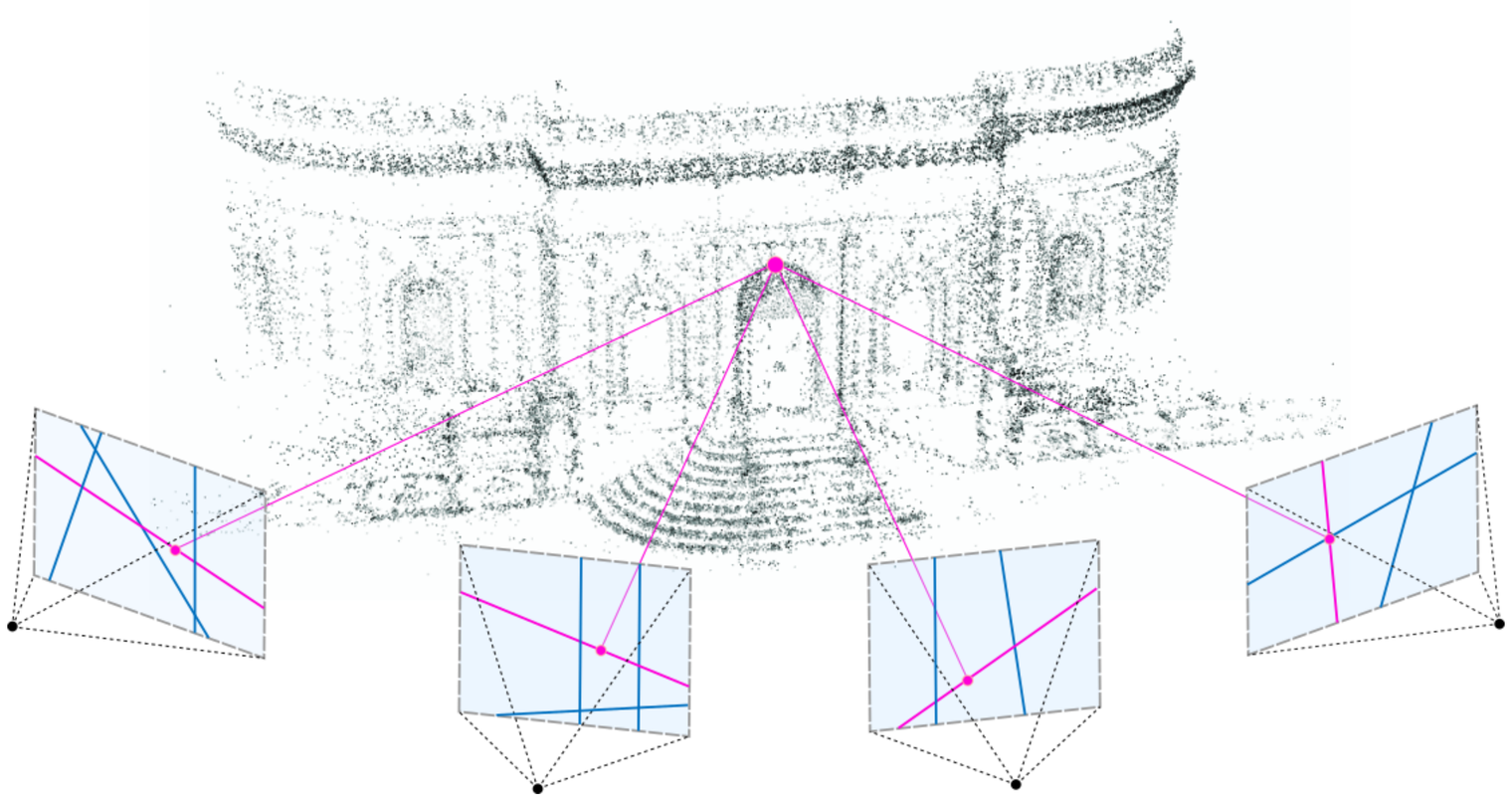
Geometría en Ciencias de Datos
Universidad Nacional de Colombia
Agosto 18-21

Restricciones usuales en ML/Computer Vision

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

- Espacio euclideano $\mathbb{R}^d$
- Esfera unitaria $\mathbb{S}^{d-1}$
- Variedad de Stiefel $\mathrm{St}(d, k)$ (marcos ortonormales)
- Variedad Grasmaniana $\mathrm{Gr}(k, V)$ (subespacios lineales de V de dimensión k)
- Grupo de Rotaciones $\mathrm{SO}(n)$
- Matrices semidefinidas positivas con rango fijo $S_+^k(d)$

Ejemplo: Structure-from-motion



Espacio de búsqueda: grupo de rotaciones

Ejemplo: localizacion en redes de sensores

Espacio de búsqueda: nube de puntos modulo movimientos rígidos

Lo típico

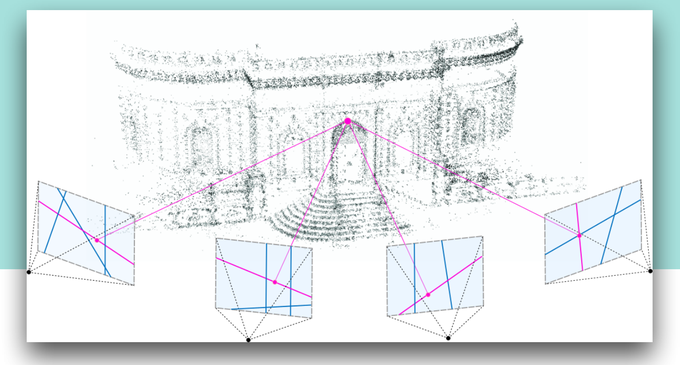
$$\min_{x \in \mathcal{M}} f(x)$$

restricción

Ejemplo:

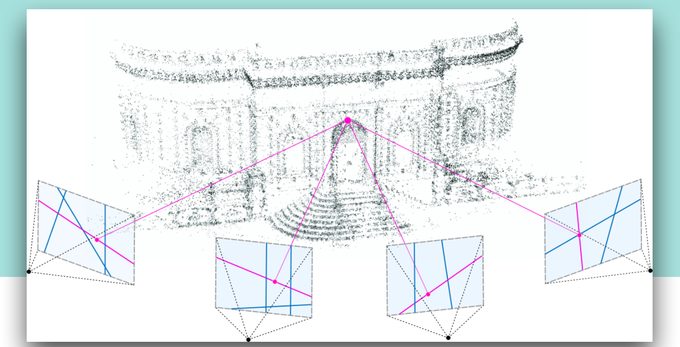
$$\min_{x \in \mathrm{SO}(3)} f(R) \longmapsto \min_{R^\top R = I_3} f(R)$$

$$\det(R) = 1$$



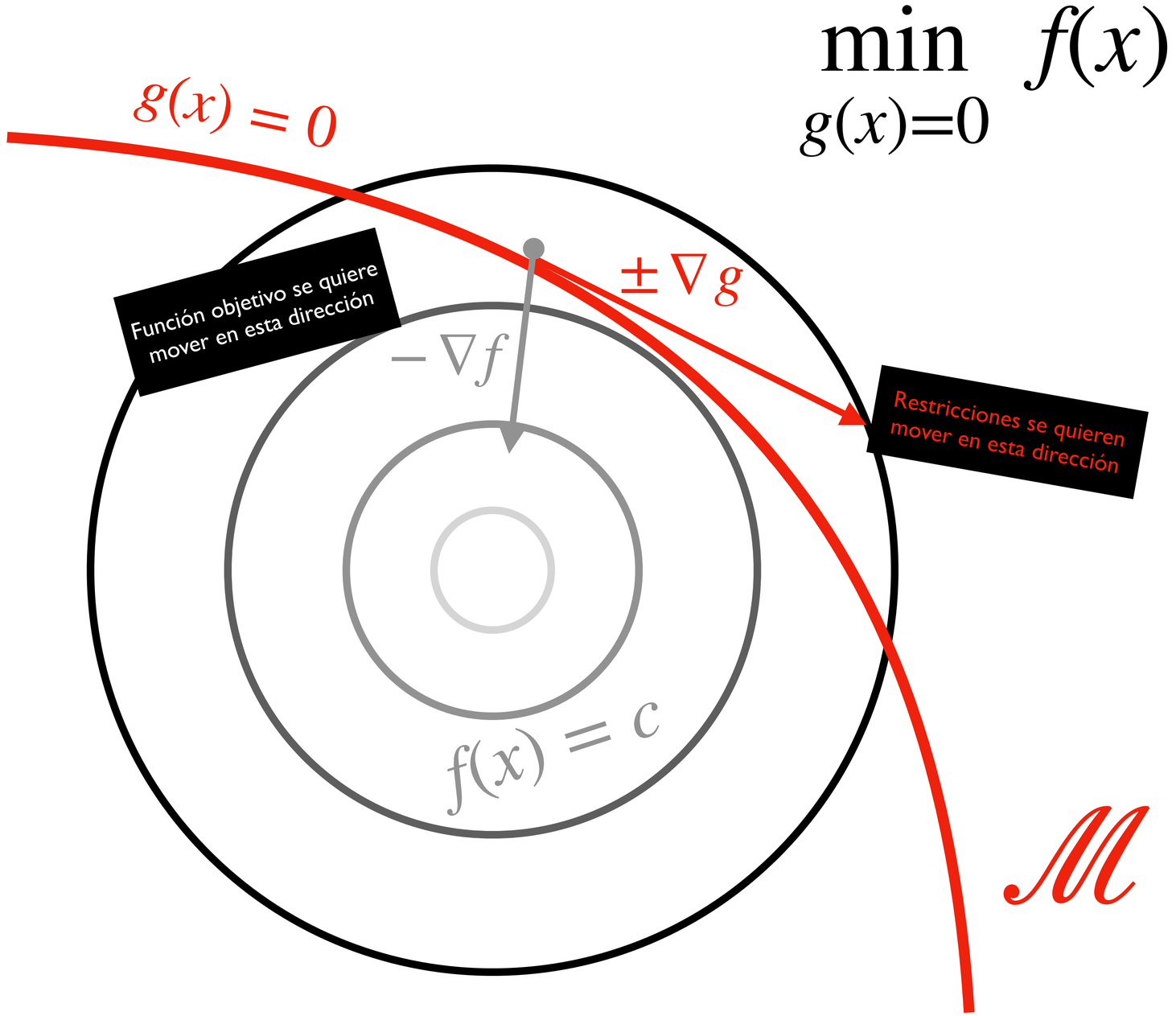
$$\min_{x \in \mathrm{SO}(3)} f(R) \quad \longmapsto \quad \min_{R^\top R = I_3} f(R)$$

$$\det(R) = 1$$

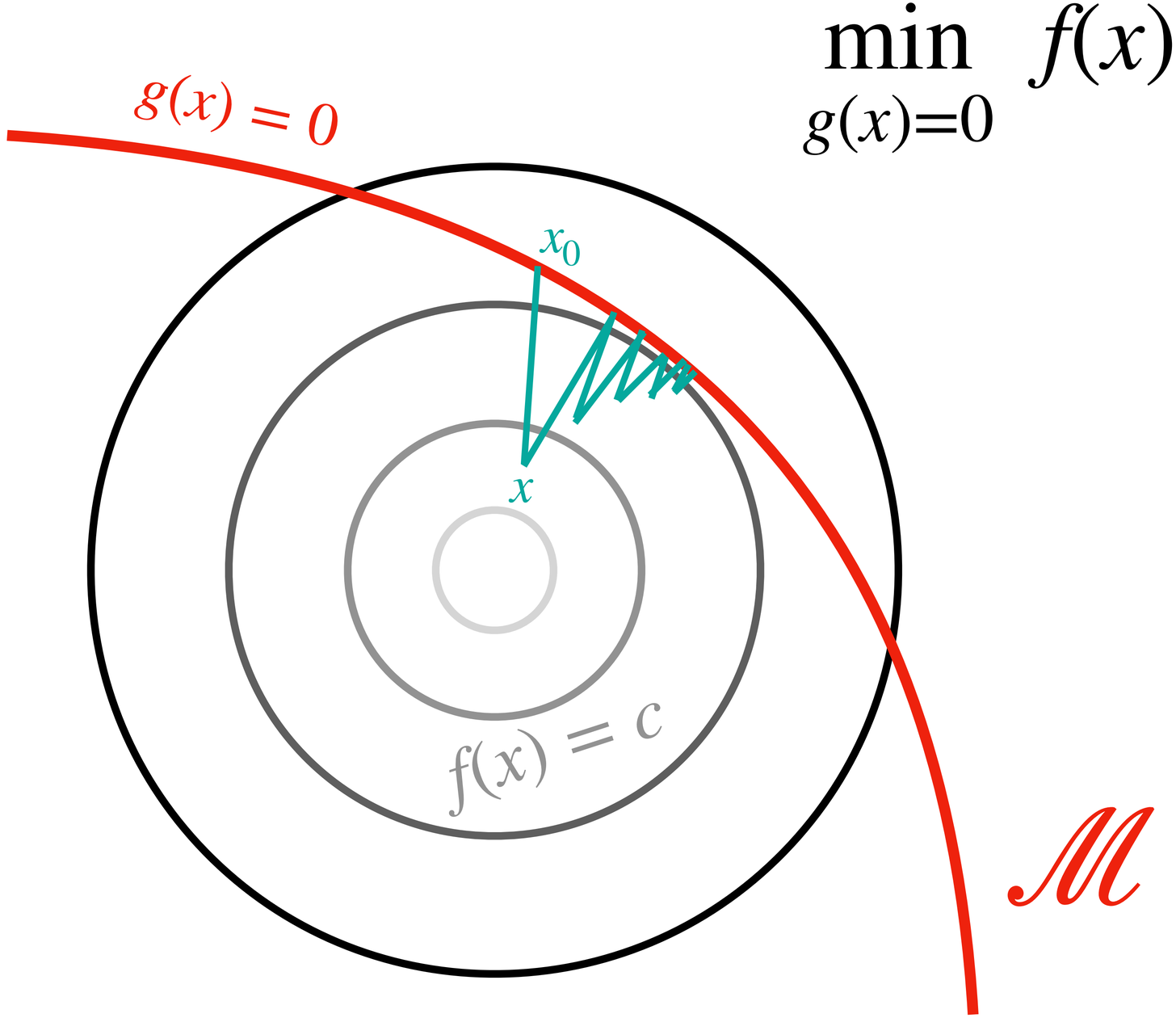


Estas restricciones “recortan” un conjunto curvo y no convexo en $\mathbb{R}^{3\times3}$

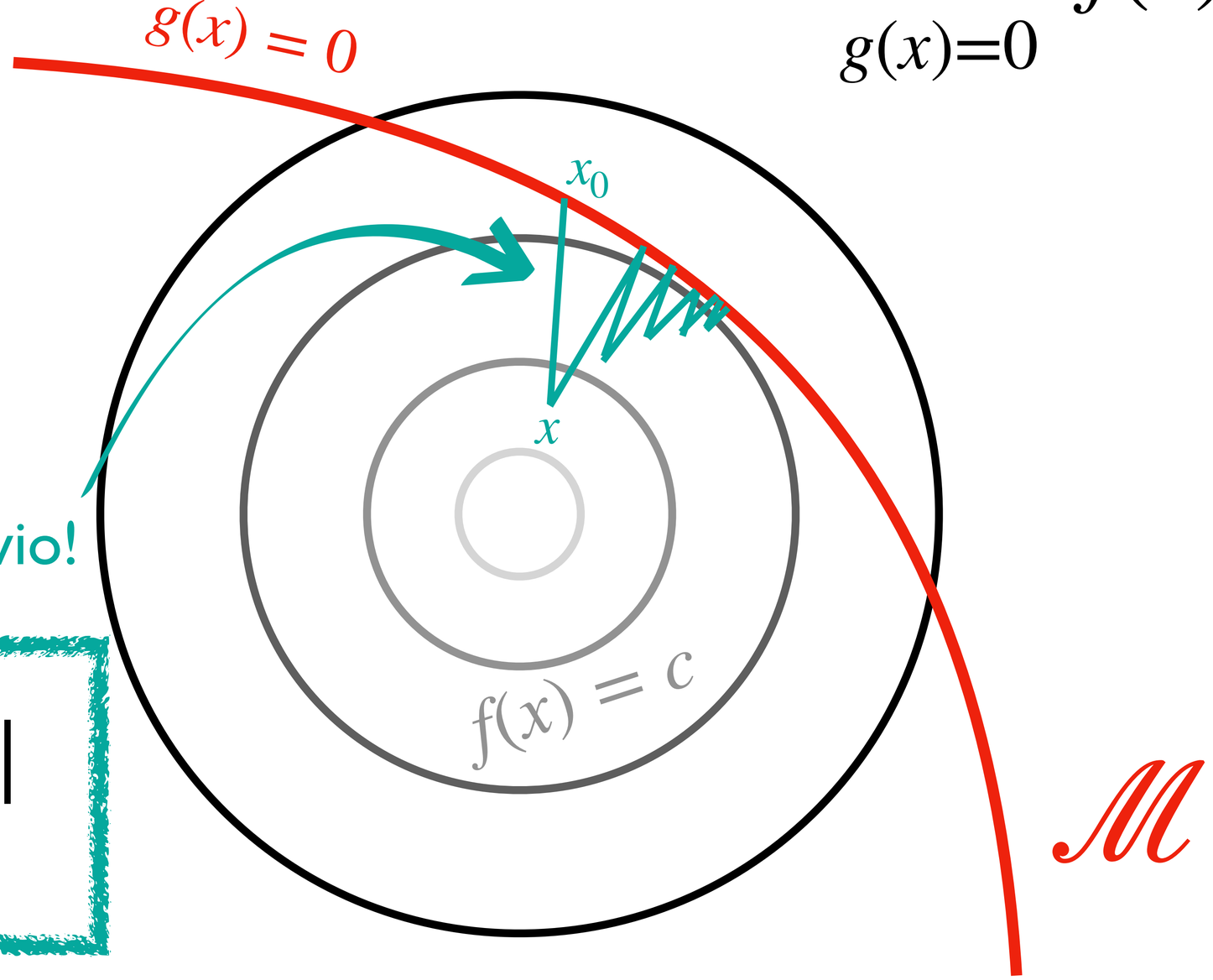
$$\min_{g(x)=0} f(x)$$


$g(x) = 0$
Función objetivo se quiere mover en esta dirección
$\pm \nabla g$
$-\nabla f$
Restricciones se quieren mover en esta dirección
$f(x) = c$
$\mathcal{M}$

$$\min_{g(x)=0} f(x)$$

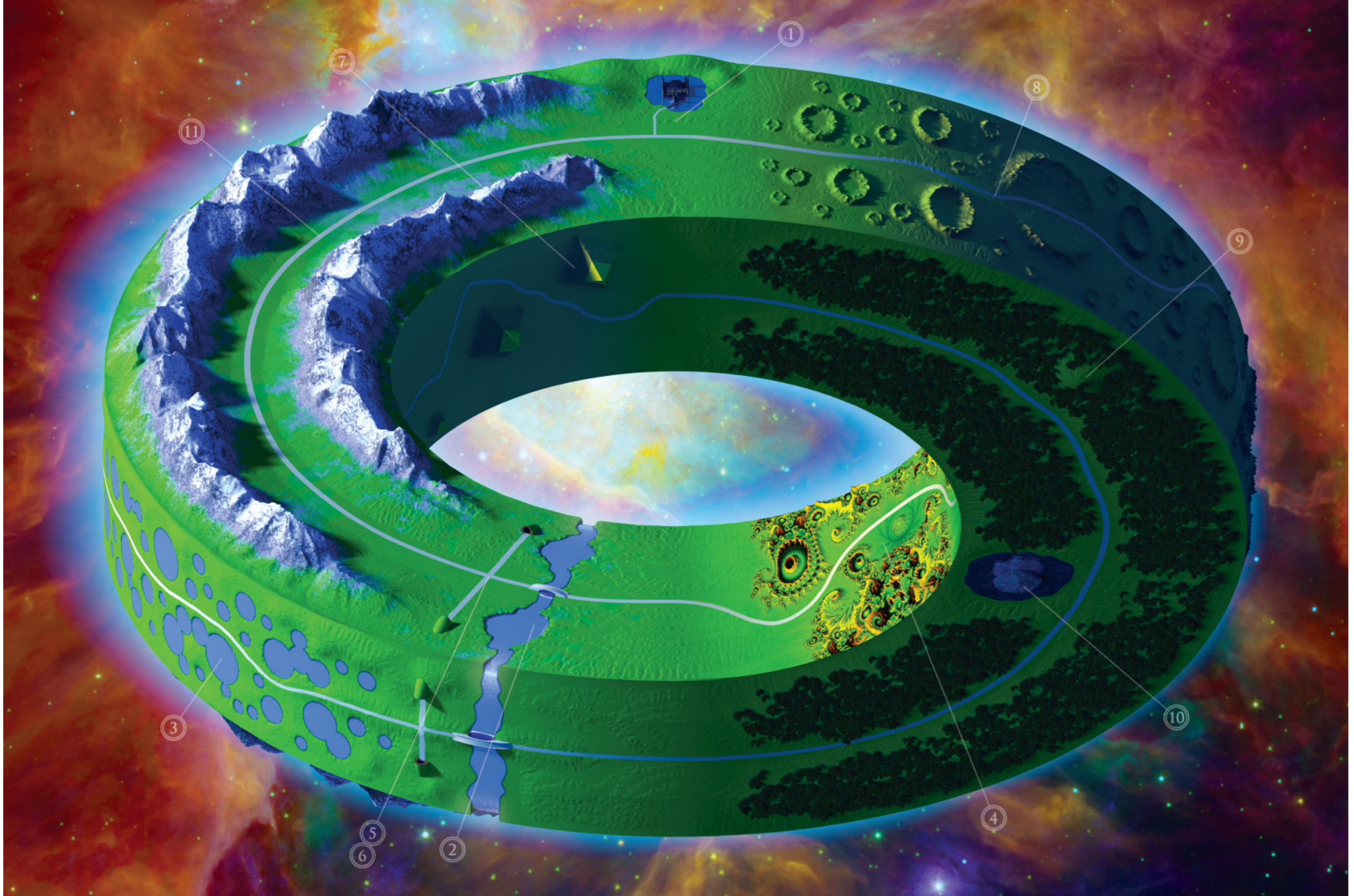


$$\min_{g(x)=0} f(x)$$

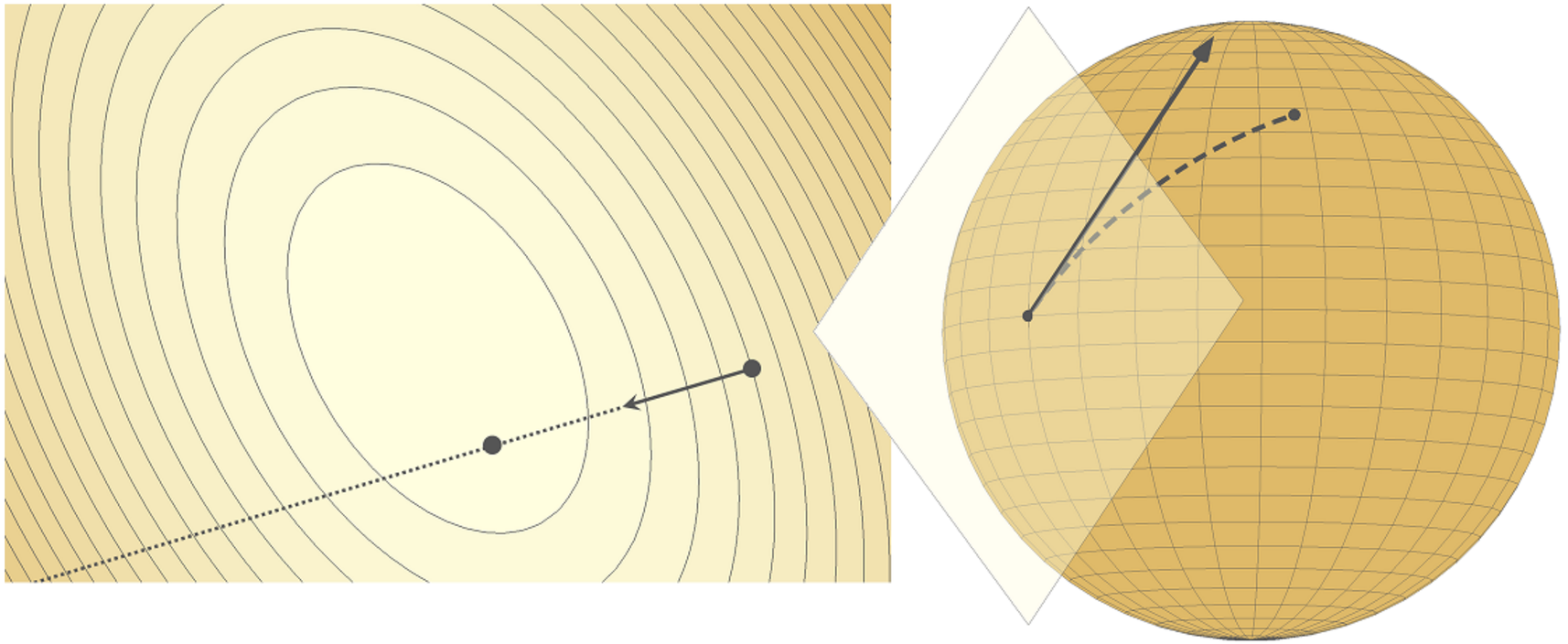


Projectar no es obvio!

$$\min_{x \text{ s.t. } g(x)=0} \|x - x_0\|$$


1
7
8
11
9
3
10
5
4
6
2

Intrinsic approach



https://afonsobandeira.wordpress.com/wp-content/uploads/2015/03/steepestdescent_compare_euclidean_sphere.png

Optimizacion sin restricciones:

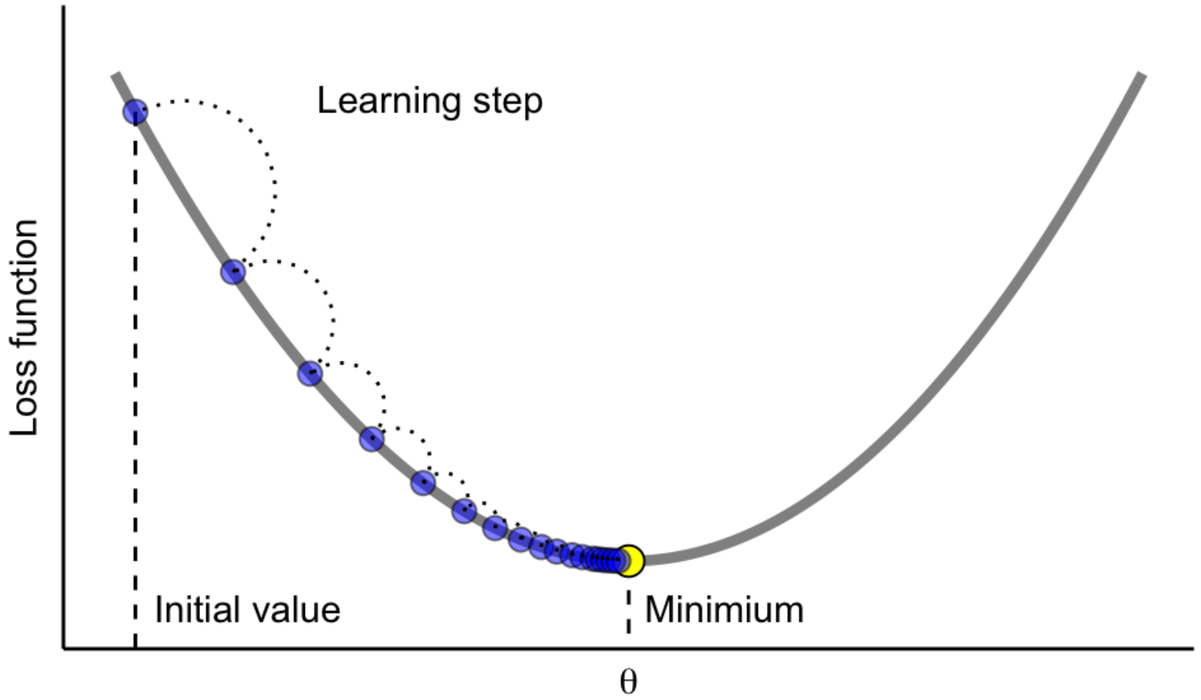
sin caminar "por fuera de la variedad"

Génesis

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

"gradiente descendente euclideo"

caminar dirección



https://www.linkedin.com/pulse/cracking-code-machine-learning-beginners-guide-alves-estrella/

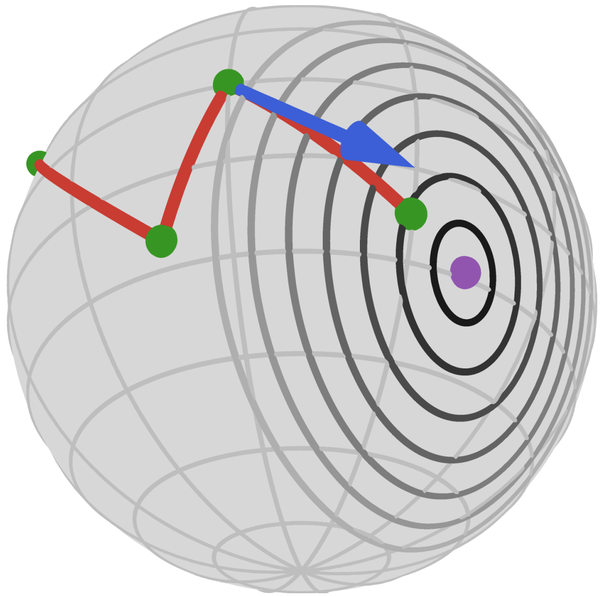
Optimización de primer orden en variedades

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$ "gradiente descendente euclideo"

caminar dirección

"gradiente descendente en variedades"

$$x_{k+1} = \exp_{x_k}(-\alpha_k \nabla_{\mathscr{M}} f(x_k))$$



Por qué optimización en variedades?

"gradiente descendente en variedades"

$$x_{k+1} = \exp_{x_k}(-\alpha_k \nabla_{\mathcal{M}} f(x_k))$$

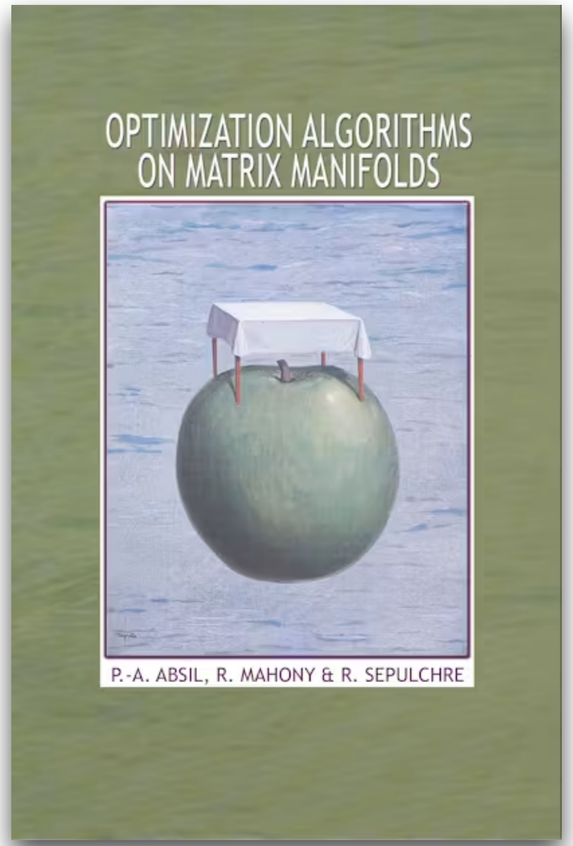
- Desde la práctica:

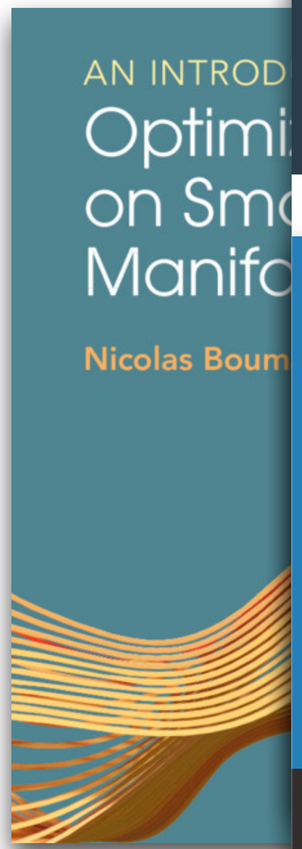
Mejores algoritmos
satisfacción de restricciones automática, especializados en el espacio

- Desde la teoría:

Descripcion matematica elegante
generaliza convexidad, gradiente descendente

Literatura - paquetes

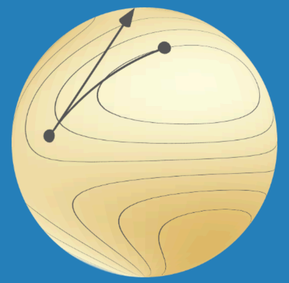




Manopt

Toolboxes for optimization on manifolds and linear spaces

Pymanopt



stable (2.2.1)

Search docs

GETTING STARTED

Quickstart

Examples

API Reference

Contributing

NOTEBOOKS

Riemannian Optimization for Inference in MoG models

» Pymanopt

View page source

Pymanopt

Pymanopt is a Python toolbox for optimization on Riemannian manifolds. The project aims to lower the barrier for users wishing to use state-of-the-art techniques for optimization on manifolds by leveraging automatic differentiation for computing gradients and Hessians, thus saving users time and reducing the potential of calculation and implementation errors.

We encourage users and developers to report problems, request features, ask for help, or leave general comments either on github or gitter. Please refer to our contributing guide if you wish to extend Pymanopt's functionality and/or contribute to the project.

Pymanopt is distributed under the open source 3-clause BSD license. Please cite this JMLR paper if you publish work using Pymanopt:

```
@article{JMLR:v17:16-177,
    author = {James Townsend and Niklas Koep and Sebastian Weichwald},
    journal = {Journal of Machine Learning Research},
    number = {137},
    pages = {1-5},
    title = {Pymanopt: A Python Toolbox for Optimization on Manifolds using Automatic Differen
    url = {http://jmlr.org/papers/v17/16-177.html},
    volume = {17},
    year = {2016}
}
```

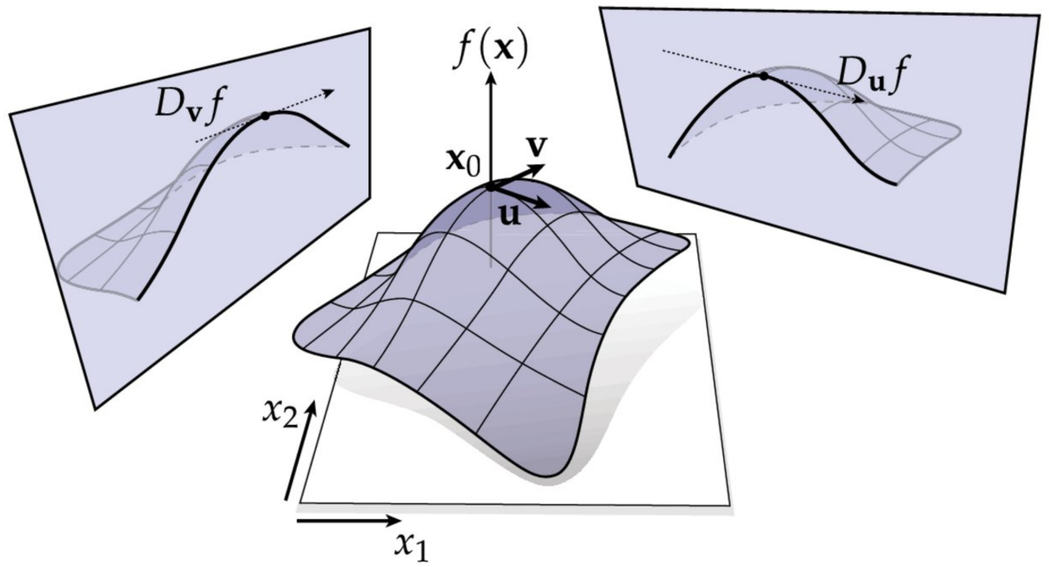
Repaso: cálculo multivariado

$$df_{x_0}(v) := \lim_{h \to 0} \frac{f(x_0 + hv) - f(x_0)}{h}$$

Proposición:

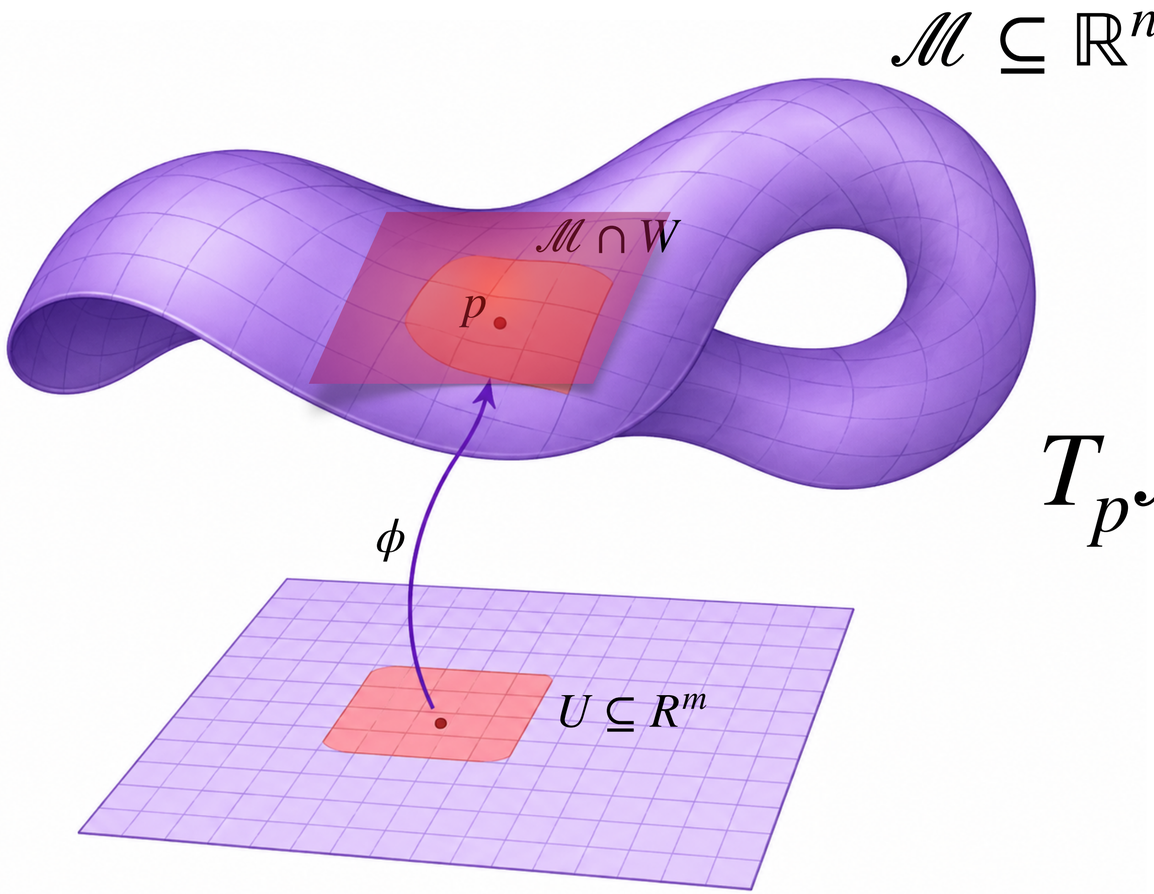
$df_{x_0} : \mathbb{R}^n \longrightarrow \mathbb{R}$ es un operador lineal

$$df_{x_0}(v) = \nabla f(x_0) \cdot v$$



https://15462.courses.cs.cmu.edu/fall2018content/lectures/03_vectorcalc/images/slide_024.jpg

Repaso: espacio tangente



$$T_p\mathscr{M} = \{\gamma'(0) \text{ where } \gamma(0) = p\}$$
$$= \text{image}(d\phi_{\phi^{-1}(p)})$$

Cortesía de ChatGPT

De vuelta a la optimización

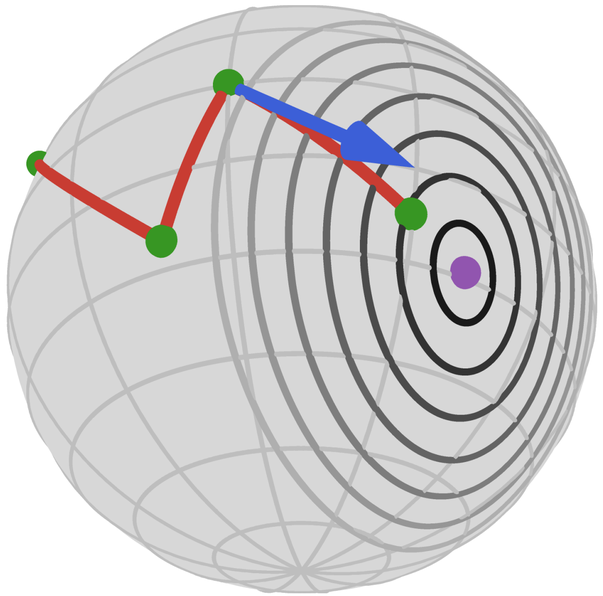
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

"gradiente descendente euclideo"

caminar dirección

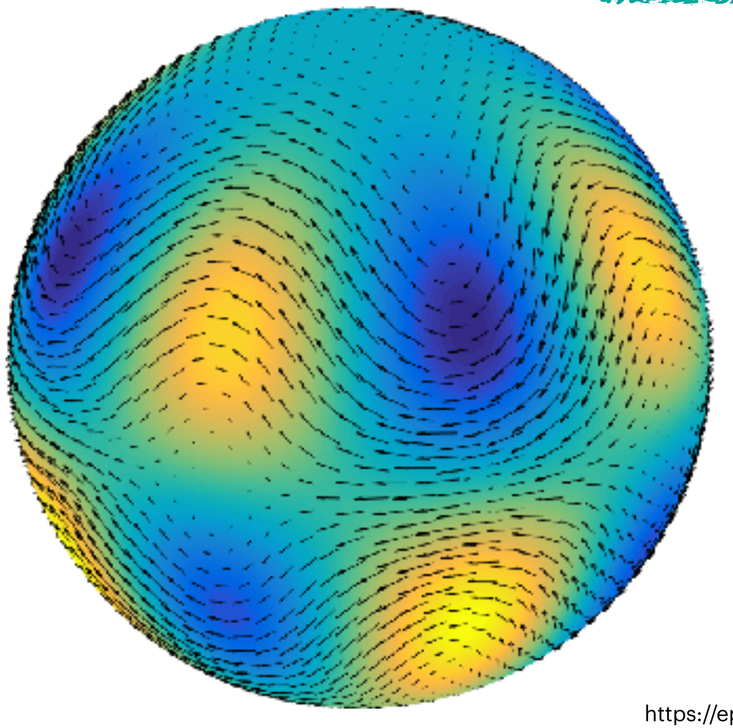
"gradiente descendente en variedades"

$$x_{k+1} = \exp_{x_k}(-\alpha_k \nabla_{\mathscr{M}} f(x_k))$$



Repaso: gradiente

$$x_{k+1} = \exp_{x_k}(-\alpha_k \nabla_{\mathscr{M}} f(x_k))$$



https://epubs.siam.org/doi/10.1137/15M1045855

Proposición:

Para cada $p \in \mathscr{M}$, existe un único vector $\nabla f(p) \in T_p\mathscr{M}$ tal que $df_p(v) = v \cdot \nabla f(p)$ para todo $v \in T_p\mathscr{M}$.

Moverse sobre la variedad: aplicación exponencial

$$x_{k+1} = \exp_{x_k}(-\alpha_k \nabla_{\mathcal{M}} f(x_k))$$

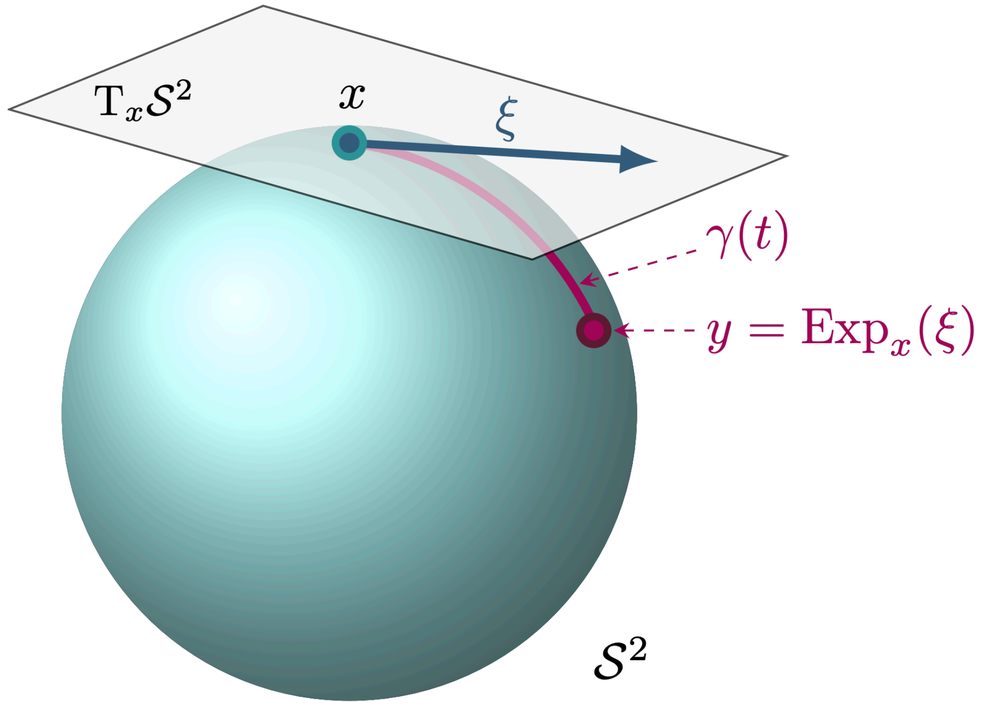


FIG. 2.1. *The Riemannian exponential map on the sphere.*

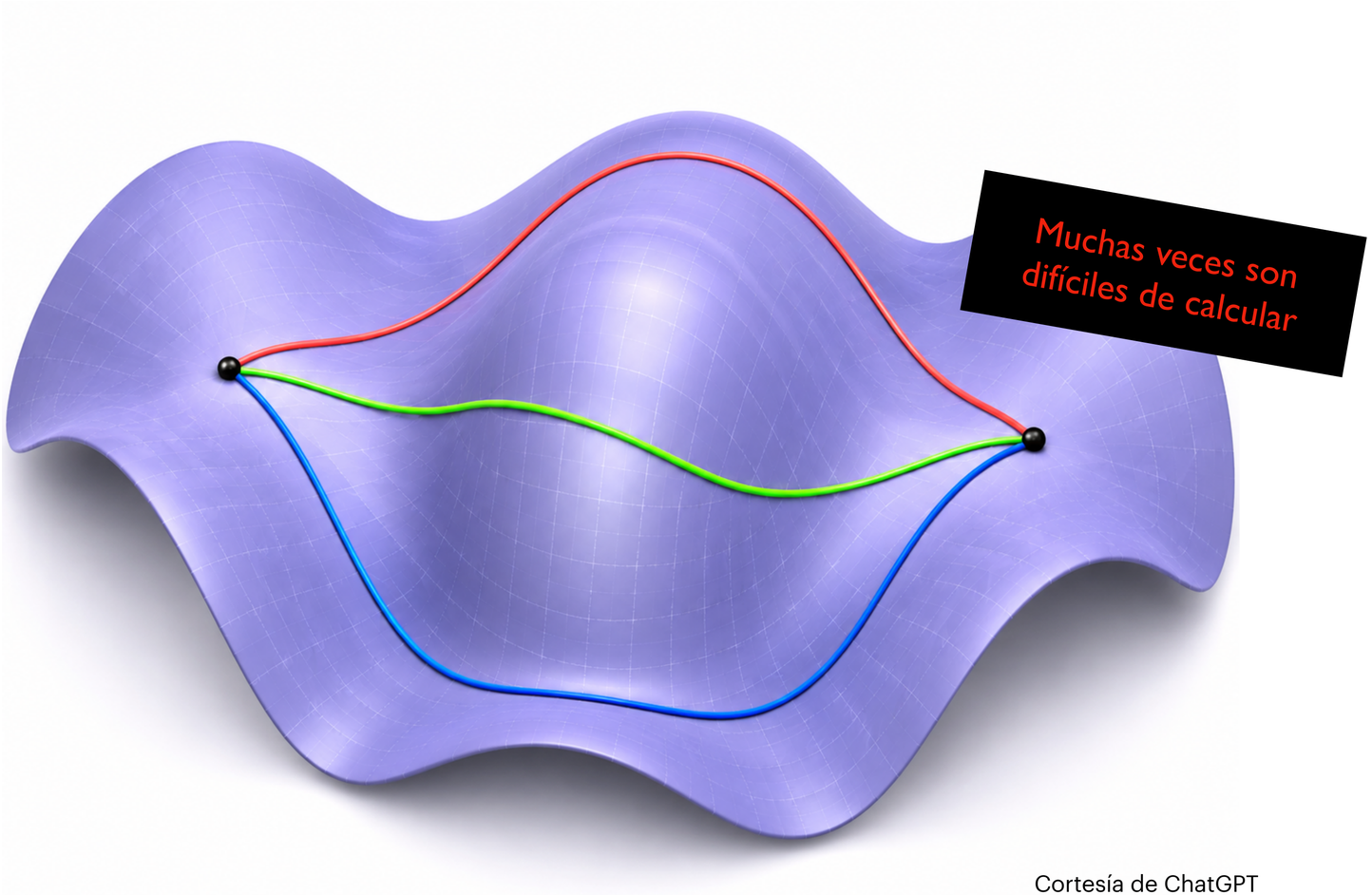
https://epub.oeaw.ac.at/?arp=0x003f5dd2

$$\exp_p(v) := \gamma_v(1)$$

donde γ_v es la (única) geodésica definida en p con velocidad v

Problema: las geodesicas son complicadas :c

$$x_{k+1} = \exp_{x_k}(-\alpha_k \nabla_{\mathcal{M}} f(x_k))$$



Cortesía de ChatGPT

Noción más débil: Retracción

$$x_{k+1} = \text{retraction}_{x_k}(-\alpha_k \nabla_{\mathcal{M}} f(x_k))$$

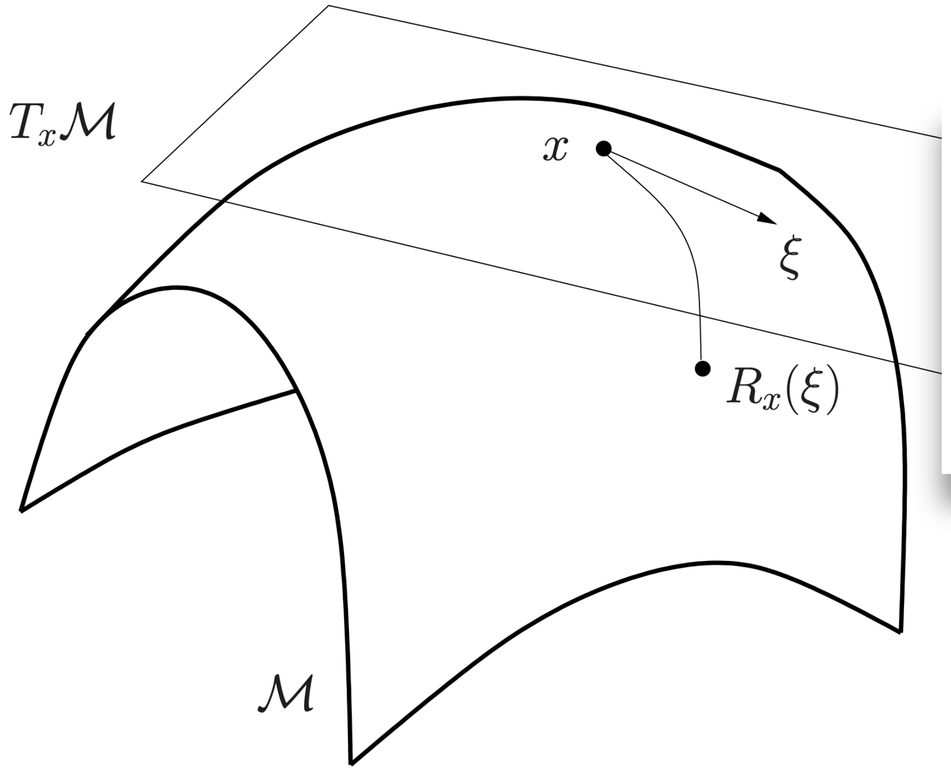


Figure 4.1 Retraction.

Definition 4.1.1 (retraction) *A* retraction *on a manifold $\mathcal{M}$ is a smooth mapping R from the tangent bundle $T\mathcal{M}$ onto $\mathcal{M}$ with the following properties. Let R_x denote the restriction of R to $T_x\mathcal{M}$.*

(i) $R_x(0_x) = x$, where 0_x denotes the zero element of $T_x\mathcal{M}$.
(ii) With the canonical identification $T_{0_x}T_x\mathcal{M} \simeq T_x\mathcal{M}$, R_x satisfies

$$\mathrm{D}R_x(0_x) = \mathrm{id}_{T_x\mathcal{M}}, \tag{4.2}$$

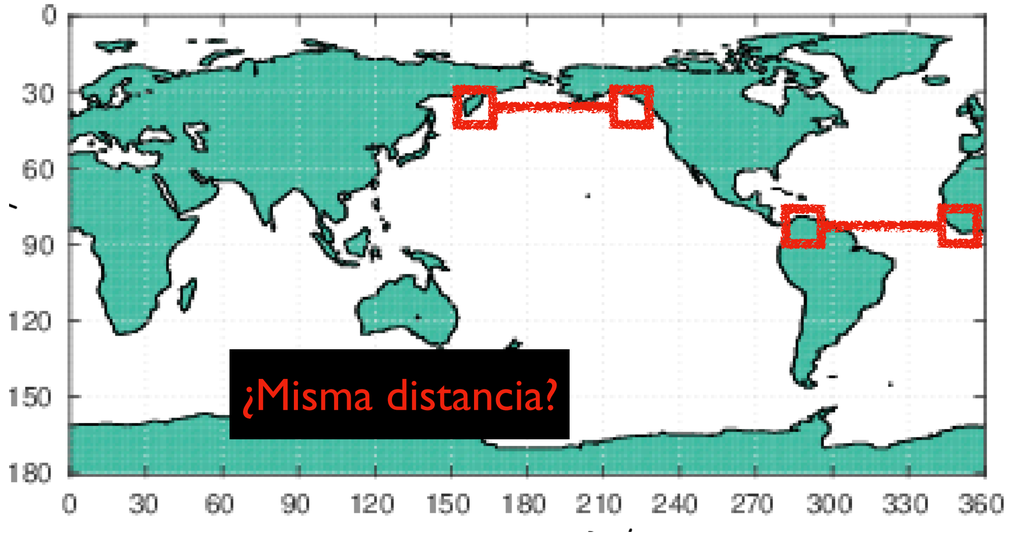
where $\mathrm{id}_{T_x\mathcal{M}}$ denotes the identity mapping on $T_x\mathcal{M}$.

Definition 3.47. *A* retraction *on a manifold $\mathcal{M}$ is a smooth map*

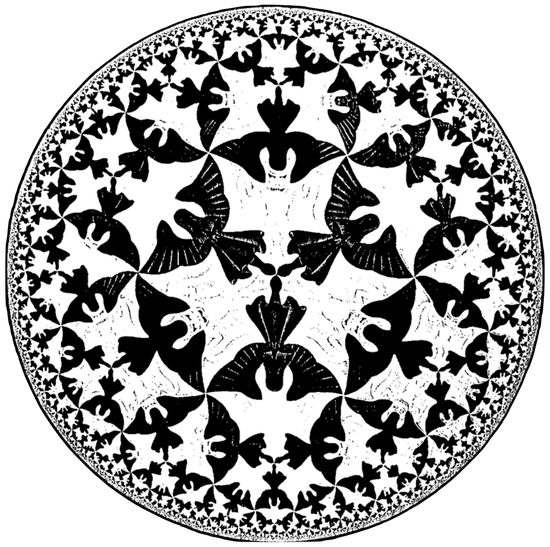
$$\mathrm{R}\colon \mathrm{T}\mathcal{M} \to \mathcal{M}\colon (x, v) \mapsto \mathrm{R}_x(v)$$

such that each curve $c(t) = \mathrm{R}_x(tv)$ satisfies $c(0) = x$ and $c'(0) = v$.

Escenario más general: Variedades Riemannianas

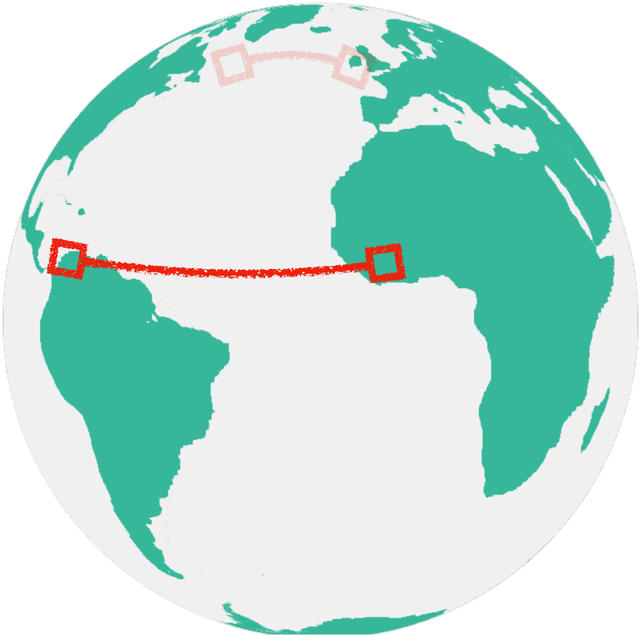


Par $(\mathscr{M}, g)$ conformado por una variedad diferenciable $\mathscr{M}$ y un producto interno definido puntualmente
$g_p(\cdot, \cdot) : T_p\mathscr{M} \times T_p\mathscr{M} \to \mathbb{R}$.



$\cong$

$$ds^2 = \frac{dx^2 + dy^2}{(1 - x^2 - y^2)^2}$$



https://epubs.siam.org/doi/10.1137/15M1045855

Gradiente Riemanniano

- Tensor métrico $g \in \mathbb{R}^{n \times n}$
- Gradiente en coordenadas $\nabla f \in \mathbb{R}^n$

$$\nabla_g f = g^{-1} \nabla f$$

De vuelta a la optimización

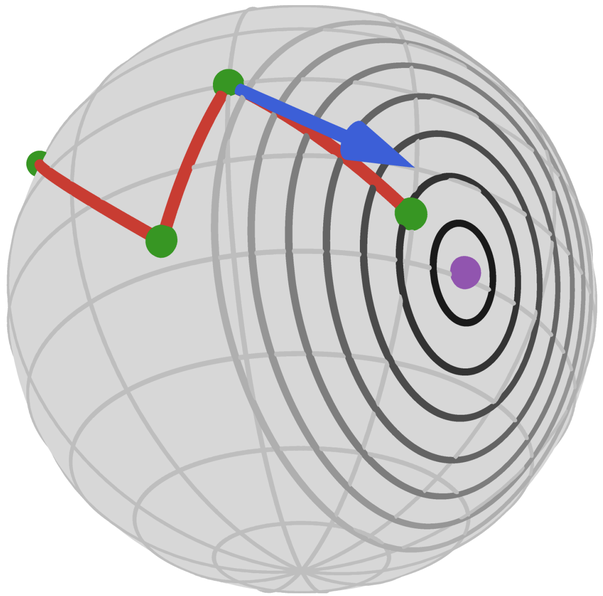
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

"gradiente descendente euclideo"

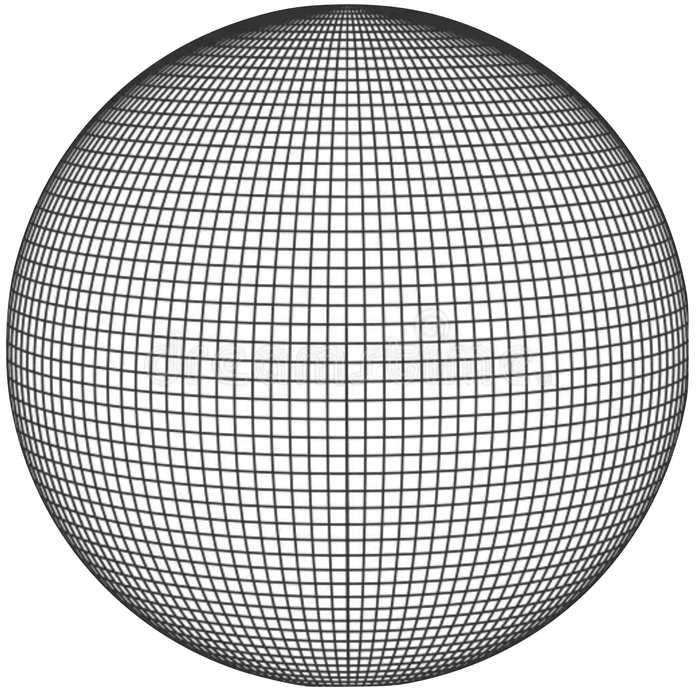
caminar dirección

"gradiente descendente en variedades"

$$x_{k+1} = \exp_{x_k}(-\alpha_k \nabla_{\mathscr{M}} f(x_k))$$



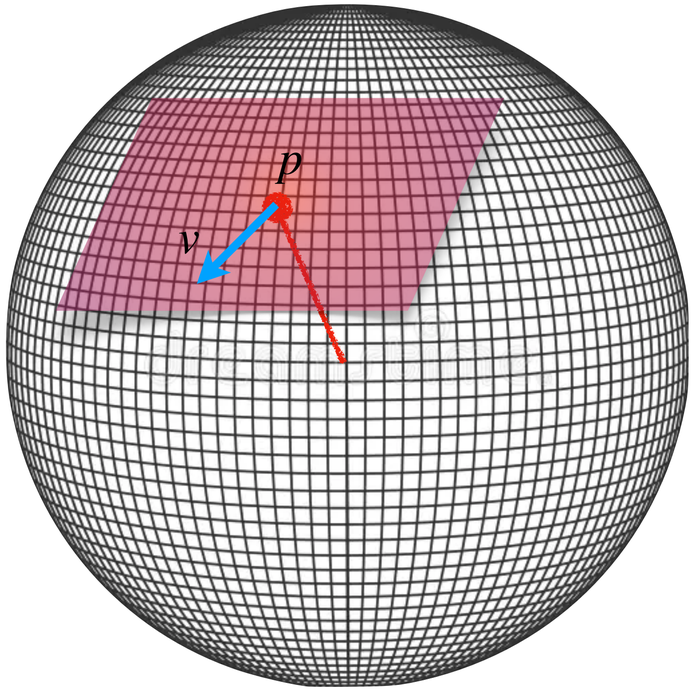
Ejemplo: Esfera



$$\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$$

Ejemplo: Esfera

$$\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$$



$$T_p\mathbb{S}^{n-1} := \{v \in \mathbb{R}^n : v \cdot p = 0\}$$

Ejemplo: Esfera

Variedad $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$

Tangente $T_p\mathbb{S}^{n-1} := \{v \in \mathbb{R}^n : v \cdot p = 0\}$

Restricción de $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla_{\mathbb{S}^{n-1}} f(p) = (I_{n\times n} - pp^\top)\, \nabla_{\mathbb{R}^n} f(p)$$

Proyectar el gradiente del espacio ambiente al espacio tangente

Ejemplo: Esfera

Variedad $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$

Tangente $T_p\mathbb{S}^{n-1} := \{v \in \mathbb{R}^n : v \cdot p = 0\}$

Gradiente Intrinseco $\nabla_{\mathbb{S}^{n-1}} f(p) = (I_{n\times n} - pp^\top)\nabla_{\mathbb{R}^n} f(p)$

$$\exp_p(v) = p\cos\|v\|_2 + \frac{v\sin\|v\|_2}{\|v\|_2}$$

$$R_p(v) = \frac{p+v}{\|p+v\|_2}$$

Ejemplo: Esfera

Variedad $$\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$$

Tangente $$T_p\mathbb{S}^{n-1} := \{v \in \mathbb{R}^n : v \cdot p = 0\}$$

Gradiente Intrinseco $$\nabla_{\mathbb{S}^{n-1}} f(p) = (I_{n\times n} - pp^\top)\nabla_{\mathbb{R}^n} f(p)$$

Retracción $$R_p(v) = \frac{p+v}{\|p+v\|_2}$$

Ejemplo: variedad de Stiefel

$$\mathrm{St}(n,k) := \{U \in \mathbb{R}^{n\times k} : U^\top U = I_{k\times k}\}$$

$$\mathrm{St}(n,k) := \{U \in \mathbb{R}^{n\times k} : U^\top U = I_{k\times k}\}$$

$$T_U\mathrm{St}(n,k) := \{V \in \mathbb{R}^{n\times k} : U^\top V + V^\top U = 0_{k\times k}\}$$

$$R_U(V) := (U+V)(I_{k\times k} + V^\top V)^{-1/2}$$

Ejemplo: Esfera

Variedad $$\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$$

Tangente $$T_p\mathbb{S}^{n-1} := \{v \in \mathbb{R}^n : v \cdot p = 0\}$$

Gradiente Intrinseco $$\nabla_{\mathbb{S}^{n-1}} f(p) = (I_{n\times n} - pp^\top)\nabla_{\mathbb{R}^n} f(p)$$

Retracción $$R_p(v) = \frac{p+v}{\|p+v\|_2}$$

Optimizacion del cociente de Rayleigh: $$\min_{v\in\mathbb{S}^{n-1}} \frac{1}{2} x^\top A x$$

$$\min_{v \in \mathbb{S}^{n-1}} \frac{1}{2} x^\top A x$$

$$\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$$

$$T_p \mathbb{S}^{n-1} := \{v \in \mathbb{R}^n : v \cdot p = 0\}$$

$$\nabla_{\mathbb{S}^{n-1}} f(p) = (I_{n \times n} - pp^\top) \nabla_{\mathbb{R}^n} f(p)$$

$$R_p(v) = \frac{p + v}{\|p + v\|_2}$$

$$\mathrm{St}(n,k) := \{U \in \mathbb{R}^{n\times k} : U^\top U = I_{k\times k}\}$$

$$T_U\mathrm{St}(n,k) := \{V \in \mathbb{R}^{n\times k} : U^\top V + V^\top U = 0_{k\times k}\}$$

$$R_U(V) := (U+V)(I_{k\times k} + V^\top V)^{-1/2}$$

Análisis de Componentes Principales (PCA):

$$\min_{U\in\mathrm{St}(n,k)} \|U^\top A\|_{\mathrm{Fr}}^2$$

$$\min_{U \in \mathrm{St}(n,k)} \|U^\top A\|_{\mathrm{Fr}}^2$$

$$\mathrm{St}(n,k) := \{U \in \mathbb{R}^{n \times k} : U^\top U = I_{k \times k}\}$$

$$T_U \mathrm{St}(n,k) := \{V \in \mathbb{R}^{n \times k} : U^\top V + V^\top U = 0_{k \times k}\}$$

$$R_U(V) := (U + V)(I_{k \times k} + V^\top V)^{-1/2}$$

Sparse PCA:

$$\max_{U \in \mathrm{St}(n,k)} \|U^\top A\|_{\mathrm{Fr}}^2 - \alpha \sum_{ij} |x_{ij}|$$

Robust PCA:

$$\min_{U \in \mathrm{St}(n,k)} \sum_i \|a_i - UU^\top a_i\|_2$$

Journal of Machine Learning Research 16 (2015) 2859-2900 Submitted 5/14; Revised 3/15; Published 12/15

Linear Dimensionality Reduction: Survey, Insights, and Generalizations

John P. Cunningham JPC2181@COLUMBIA.EDU
Department of Statistics
Columbia University

Method	Objective $f_X(M)$	Manifold $\mathcal{M}$
PCA (§3.1.1)	$\|\|X - MM^\top X\|\|_F^2$	$\mathcal{O}^{d\times r}$
MDS (§3.1.2)	$\sum_{i,j}\left(d_X(x_i,x_j) - d_Y(M^\top x_i, M^\top x_j)\right)^2$	$\mathcal{O}^{d\times r}$
LDA (§3.1.3)	$\frac{\mathrm{tr}(M^\top \Sigma_B M)}{\mathrm{tr}(M^\top \Sigma_W M)}$	$\mathcal{O}^{d\times r}$
Traditional CCA (§3.1.4)	$\mathrm{tr}\left(M_a^\top (X_a X_a^\top)^{-1/2} X_a X_b^\top (X_b X_b^\top)^{-1/2} M_b\right)$	$\mathcal{O}^{d_a\times r}\times\mathcal{O}^{d_b\times r}$
Orthogonal CCA (§3.1.4)	$\frac{\mathrm{tr}\left(M_a^\top X_a X_b^\top M_b\right)}{\sqrt{\mathrm{tr}(M_a^\top X_a X_a^\top M_a)\mathrm{tr}(M_b^\top X_b X_b^\top M_b)}}$	$\mathcal{O}^{d_a\times r}\times\mathcal{O}^{d_b\times r}$
MAF (§3.1.5)	$\frac{\mathrm{tr}(M^\top \Sigma_\delta M)}{\mathrm{tr}(M^\top \Sigma M)}$	$\mathcal{O}^{d\times r}$
SFA (§3.1.6)	$\mathrm{tr}(M^\top \dot{X}\dot{X}^\top M)$	$\mathcal{O}^{d\times r}$
SDR (§3.1.7)	$\mathrm{tr}\left(\bar{K}_Z\left(\bar{K}_{M^\top X} + n\epsilon I\right)^{-1}\right)$	$\mathcal{O}^{d\times r}$
LPP (§3.1.8)	$\mathrm{tr}\left(M^\top (XDX^\top)^{-\top/2} XLX^\top (XDX^\top)^{-1/2} M\right)$	$\mathcal{O}^{d\times r}$
UICA (§3.2.1)	$\frac{1}{2}\log\|M^\top M\| + \frac{1}{n}\sum_{i=1}^n\sum_{k=1}^r \log f_\theta\left(m_k^\top x_n\right)$	$\mathbb{R}^{d\times r}$
PPCA (§3.2.2)	$\log\|MM^\top + \sigma^2 I\| + \mathrm{tr}\left(XX^\top (MM^\top + \sigma^2 I)^{-1}\right)$	$\mathbb{R}^{d\times r}$
FA (§3.2.3)	$\log\|MM^\top + D\| + \mathrm{tr}\left(XX^\top (MM^\top + D)^{-1}\right)$	$\mathbb{R}^{d\times r}$
LR (§3.2.4)	$\|\|X_{out} - MX_{in}\|\|_F^2 + \lambda\|\|M\|\|_p$	$\mathbb{R}^{d\times r}$
DML (§3.2.5)	$\sum_{i,j\in\eta(i)}\left\{d_M(x_i,x_j)^2 + \lambda\sum_\ell \mathbb{1}(z_i\neq z_\ell)\left[1 + d_M(x_i,x_j)^2 - d_M(x_i,x_\ell)^2\right]_+\right\}$	$\mathbb{R}^{d\times r}$

2011 50th IEEE Conference on Decision a
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Low-rank opt

***Abstract*— This paper addresses the**
distance matrix completion. This proble
the missing entries of a distance matrix
the data embedding space is possibly un
pared to the number of considered data
high-dimensional problems. We recast th
into an optimization problem over the s
semidefinite matrices and propose two e
low-rank distance matrix completion. In
a strategy to determine the dimension of
The resulting algorithms scale to high-
and monotonically converge to a global s

Sub

Abstract

Making sense of Wasserstein d
discrete measures in high-dimen
mains a challenge. Recent wo
a two-step approach to improv
facilitate the computation of opti
ing for instance projections on
or a preliminary quantization o
reduce the size of their suppo
in this work a "max-min" robu
Wasserstein distance by conside
possible distance that can be
two measures, assuming they ca
thogonally on a lower k-dime
Alternatively, we show that the corresponding

Wasserstein approximation of measures via orthogonal frames: Statistical and computational complexities

Mateo Díaz*, Nicolás García Trillos†, Daniel López-Castaño‡, and Congwei Yang†

Abstract.
Wasserstein approximation of measures with orthogonal frames connects optimal transport geometry to the problem of learning interpretable factors from data. We use this connection to recover an optimal frame from random samples, both through empirical risk minimization and an online Riemannian stochastic gradient method with provable finite-sample convergence. We obtain near-parametric convergence guarantees despite the intrinsic non-smoothness of the empirical objective, which arises from discontinuities in the nearest-axis projection. The key contribution is replacing the classical Lipschitz gradient analysis with a more general quasi-Lipschitz analysis based on Clarke subdifferentials and a universal geometric covering of switching boundaries on the unit sphere. This structure yields polynomial rather than exponential dependence on dimension, avoiding the curse of dimensionality endemic to Wasserstein empirical approximation without structural assumptions. We complement our theoretical findings with an open source implementation and numerical examples.

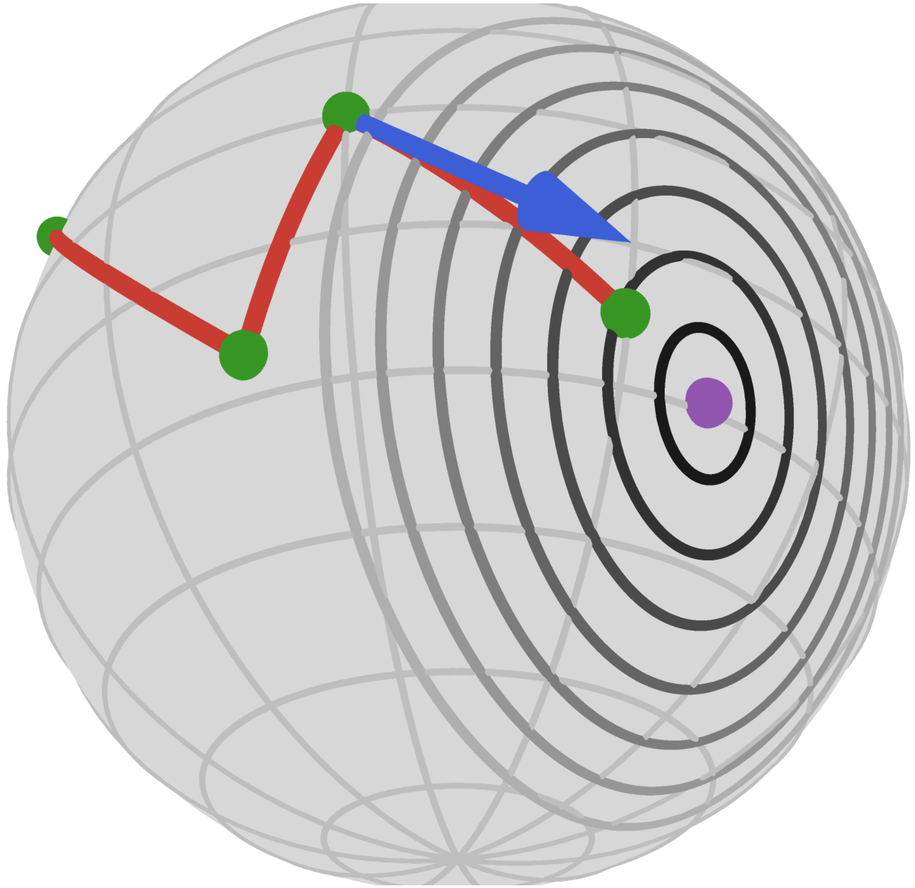
1. Introduction. Approximating probability measures with simpler, more structured distributions is a fundamental problem in machine learning and statistics. Readily, this amounts to solving an infinite-dimensional optimization problem of the form

$$\inf_{\nu\in\mathcal{F}} D(\mu,\nu),$$

where μ denotes the target probability measure, $\mathcal{F}$ represents the prespecified family of approximating distributions, and D is a metric or divergence on the space of probability measures assessing the quality of approximation. Transport-based distances, like the 2-Wasserstein distance W_2, are common choices for D, as they provide rigorous and effective frameworks for quantifying the quality of approximation in settings where the approximating measures and the target are anticipated to fail to be absolutely continuous with respect to each other. Such scenarios often arise when the family $\mathcal{F}$ is specifically chosen to extract certain geometric features from the target μ; we will discuss some examples below as well as in Section 1.1.

In this paper, we will focus on the approximation problem with the 2-Wasserstein distance W_2 and a specific family of distributions $\mathcal{F}$ that we will discuss shortly. The general approximation problem with W_2 reads

$$\inf_{\nu\in\mathcal{F}} W_2^2(\mu,\nu), \tag{1.1}$$



Gracias!

(: