

Optimization on manifolds and semidefinite relaxations

Nicolas Boumal
Princeton University

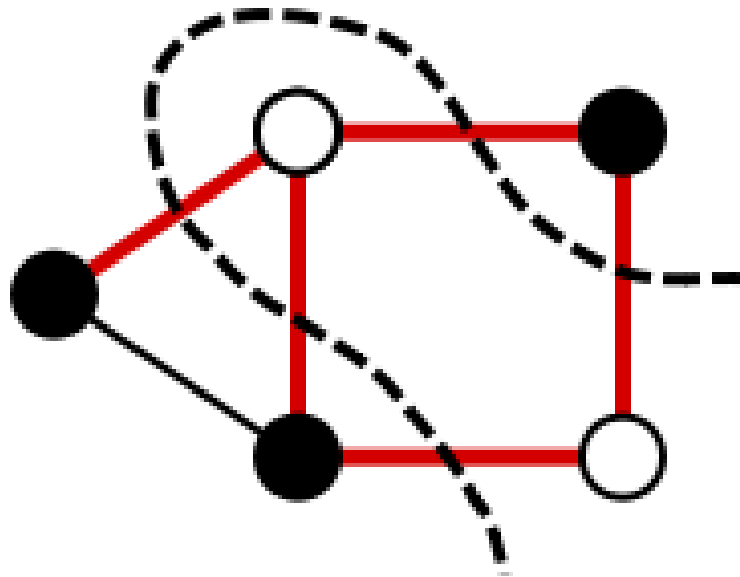
Based on work with Pierre-Antoine Absil,
Afonso Bandeira, Coralia Cartis and Vladislav Voroninski

Max-Cut relaxation

An example of global optimality on manifolds

Max-Cut

Given a graph, split its nodes in two classes, maximizing the number of in-between edges.



One of Karp's 21 NP-complete problems.

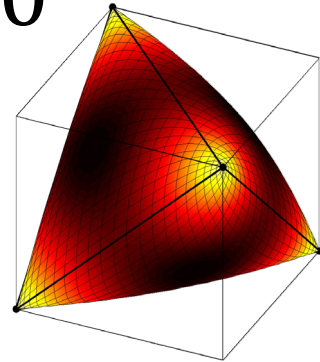
Max-Cut semidefinite relaxation

A is the adjacency matrix of the graph:

$$\min_X \text{Tr}(AX) \text{ s. t. } \text{diag}(X) = \mathbf{1}, X \succcurlyeq 0$$

Goemans & Williamson '95

Approximate the best cut within 87% by randomized projection of optimal X to $\{\pm 1\}^n$.



Convex, but IPM's run out of memory
(and time)

For a 2000 node graph (edge density 1%),
CVX runs out of memory on my former
laptop. On the new one, it returns with poor
accuracy after 3 minutes.

The methods we will discuss solve the SDP
in 6 seconds on old laptop, with certificate.

Max-Cut SDP has a low-rank solution

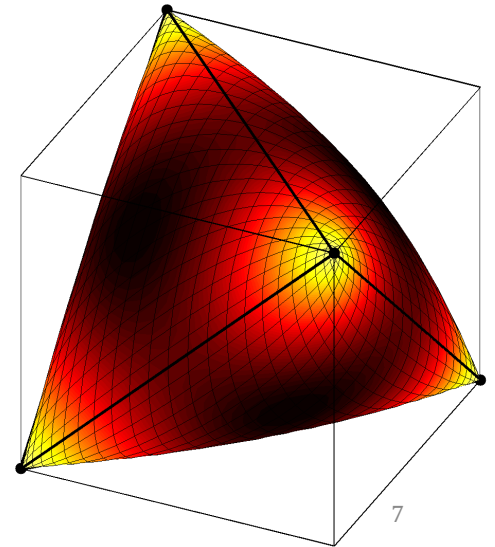
$$\min_X \text{Tr}(AX) \text{ s. t. } \text{diag}(X) = \mathbf{1}, X \succcurlyeq 0$$

Shapiro '82, Grone et al. '90, Pataki '94, Barvinok '95

There is an optimal X whose rank r satisfies

$$\frac{r(r+1)}{2} \leq n$$

A fortiori, $r \leq \sqrt{2n}$.



This justifies restricting the rank

$$\min_X \text{Tr}(AX) \text{ s. t. } \text{diag}(X) = \mathbf{1}, X \succeq 0, \text{rank}(X) \leq p$$

Parameterize as $X = YY^T$ with Y of size $n \times p$:

$$\min_{Y:n \times p} \text{Tr}(AYY^T) \text{ s. t. } \text{diag}(YY^T) = \mathbf{1}$$

Lower dimension and no conic constraint!

Burer & Monteiro '03, '05, Journée, Bach, Absil, Sepulchre '10

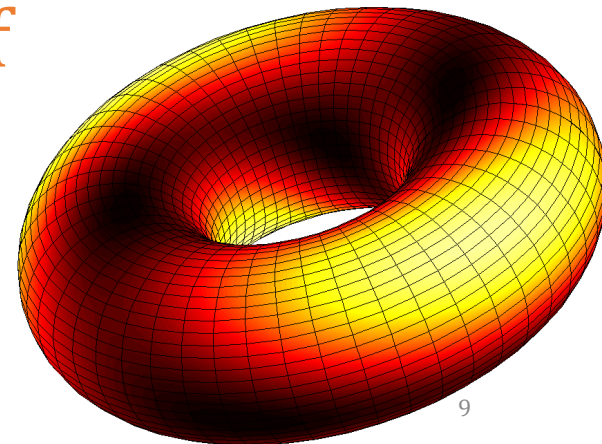
But **nonconvex**...

Key feature: search space is smooth

$$\min_{Y:n \times p} \text{Tr}(AYY^T) \text{ s. t. } \text{diag}(YY^T) = \mathbf{1}$$

Constraints \rightarrow rows of Y have unit norm.

The search space is a **product of spheres**: smooth cost function on a smooth manifold.





Our main result for Max-Cut

$$\min_{Y:n \times p} \text{Tr}(A Y Y^T) \text{ s. t. } \text{diag}(Y Y^T) = \mathbf{1}$$

If $\frac{p(p+1)}{2} > n$, for almost all A , all sop's are optimal.

If $p > n/2$, for all A , all sop's are optimal.

Main proof ingredients

1. $X = YY^T$ is optimal iff
- For all feasible \hat{X} ,
- $$\begin{aligned} 0 &\leq \text{Tr}(S\hat{X}) \\ &= \text{Tr}(A\hat{X}) - \text{Tr}(\text{ddiag}(AYY^T)\hat{X}) \\ &= \text{Tr}(A\hat{X}) - \text{Tr}(AYY^T). \end{aligned}$$

$$S = S(Y) = A - \text{ddiag}(AYY^T) \succcurlyeq 0$$

2. If Y is **sop** and **rank deficient**, $S(Y) \succcurlyeq 0$

3. For almost all A , all critical points are rank deficient (if $\frac{p(p+1)}{2} > n$).



Main result for smooth SDP's

$$\min_{X:n \times n} \text{Tr}(AX) \text{ s. t. } \text{Lin}(X) = b, X \succcurlyeq 0$$

$$\min_{Y:n \times p} \text{Tr}(AYY^T) \text{ s. t. } \text{Lin}(YY^T) = b$$

If the search space in X is **compact**
and the search space in Y is a **manifold**,
and if $\frac{p(p+1)}{2} > \# \text{constraints}$, then,
for almost all A , all sop's are optimal.

Why the manifold assumption?

What can we compute?

→ KKT points.

When are KKT conditions necessary at Y ?

→ When constraint qualifications hold at Y .

What if CQ's hold at all Y 's?

→ Set of Y 's is a smooth manifold.

Covers a range of applications

Max-Cut

\mathbf{Z}_2 -synchronization

Community detection in stochastic block model

Matrix cut norm

Phase-Cut for phase retrieval

Phase synchronization

Orthogonal-Cut (synchronization of rotations)

...

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

Optimization on manifolds

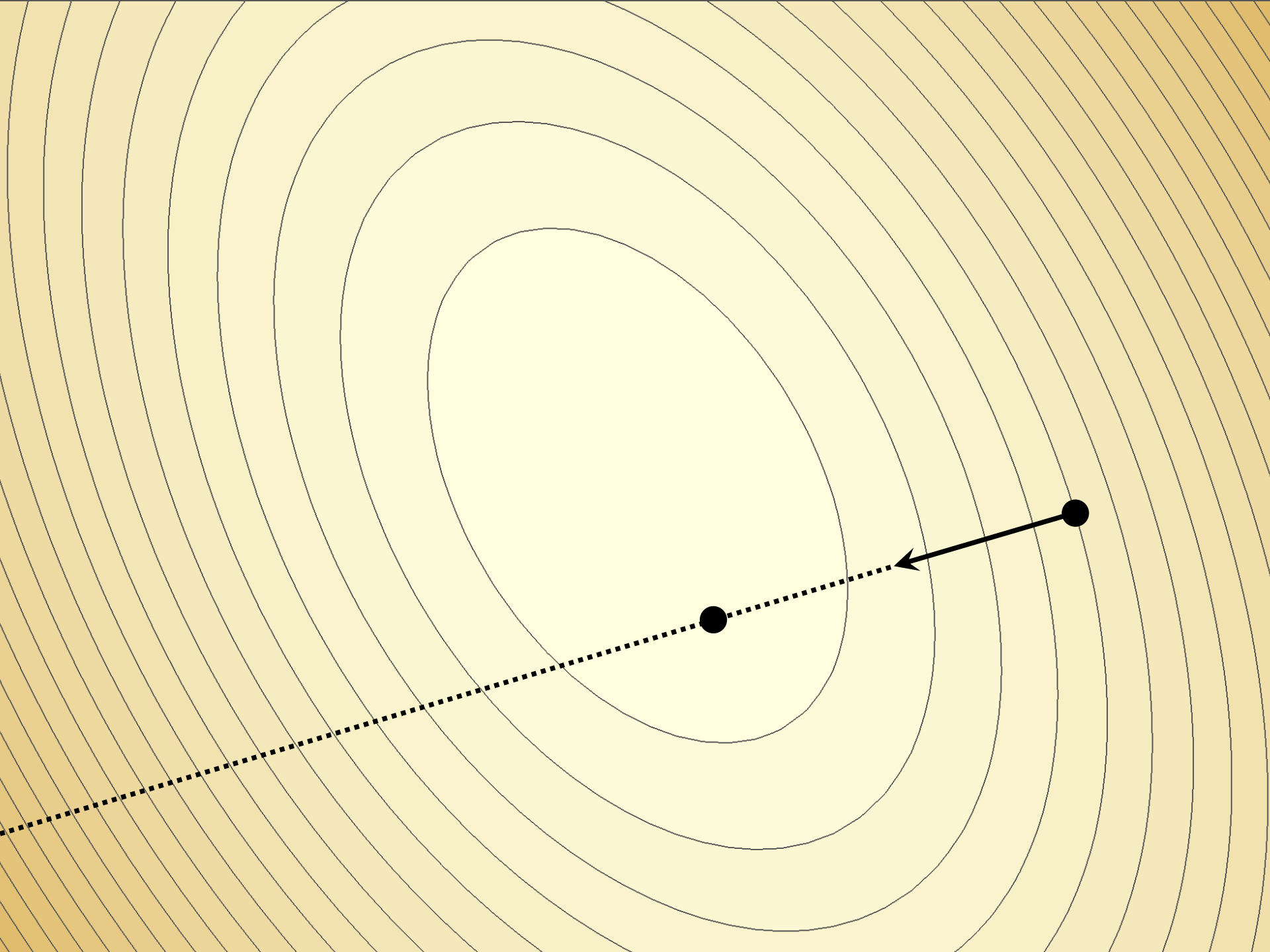
Not harder (nor easier) than unconstrained optimization

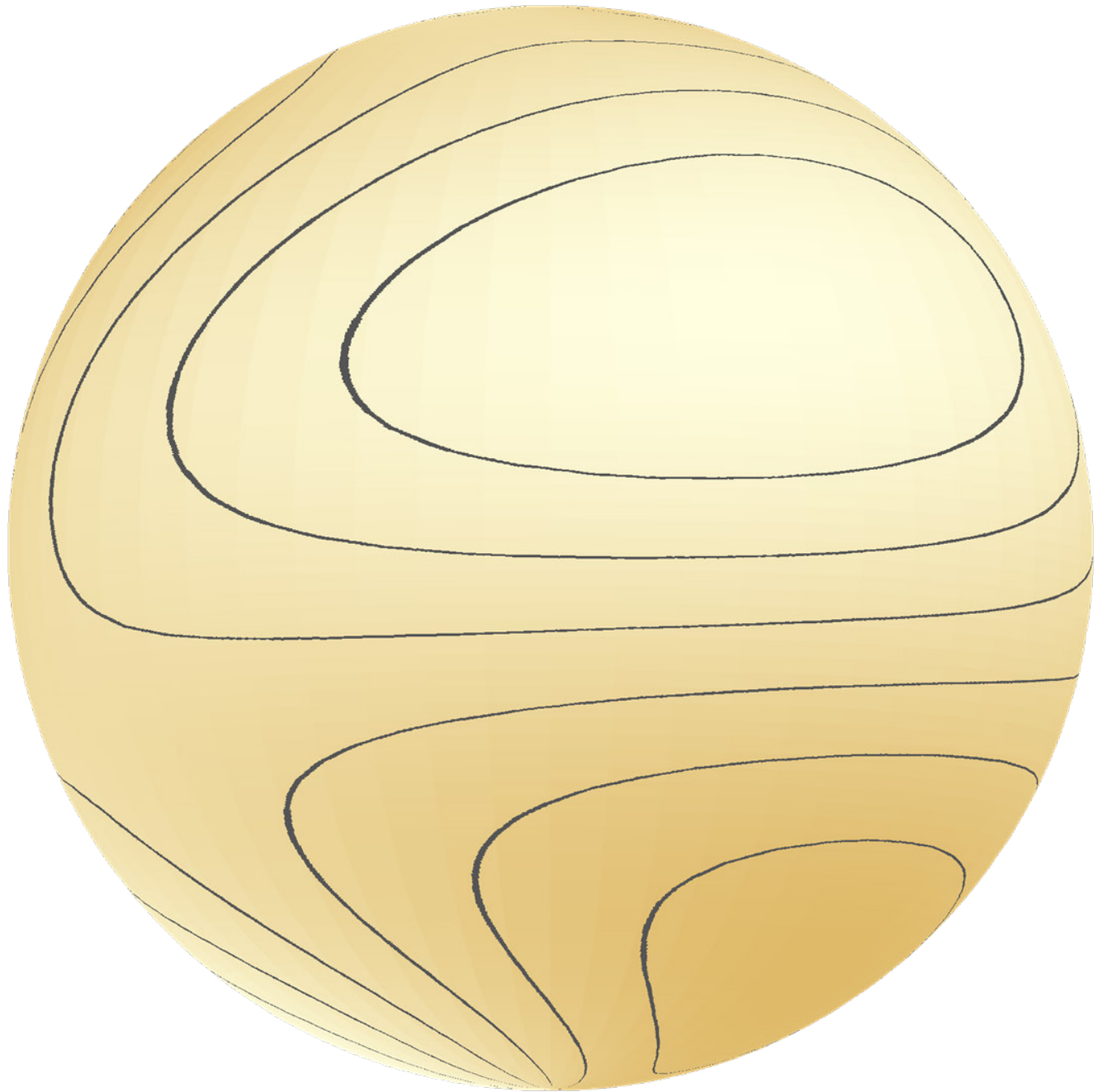
Optimization on **many** manifolds

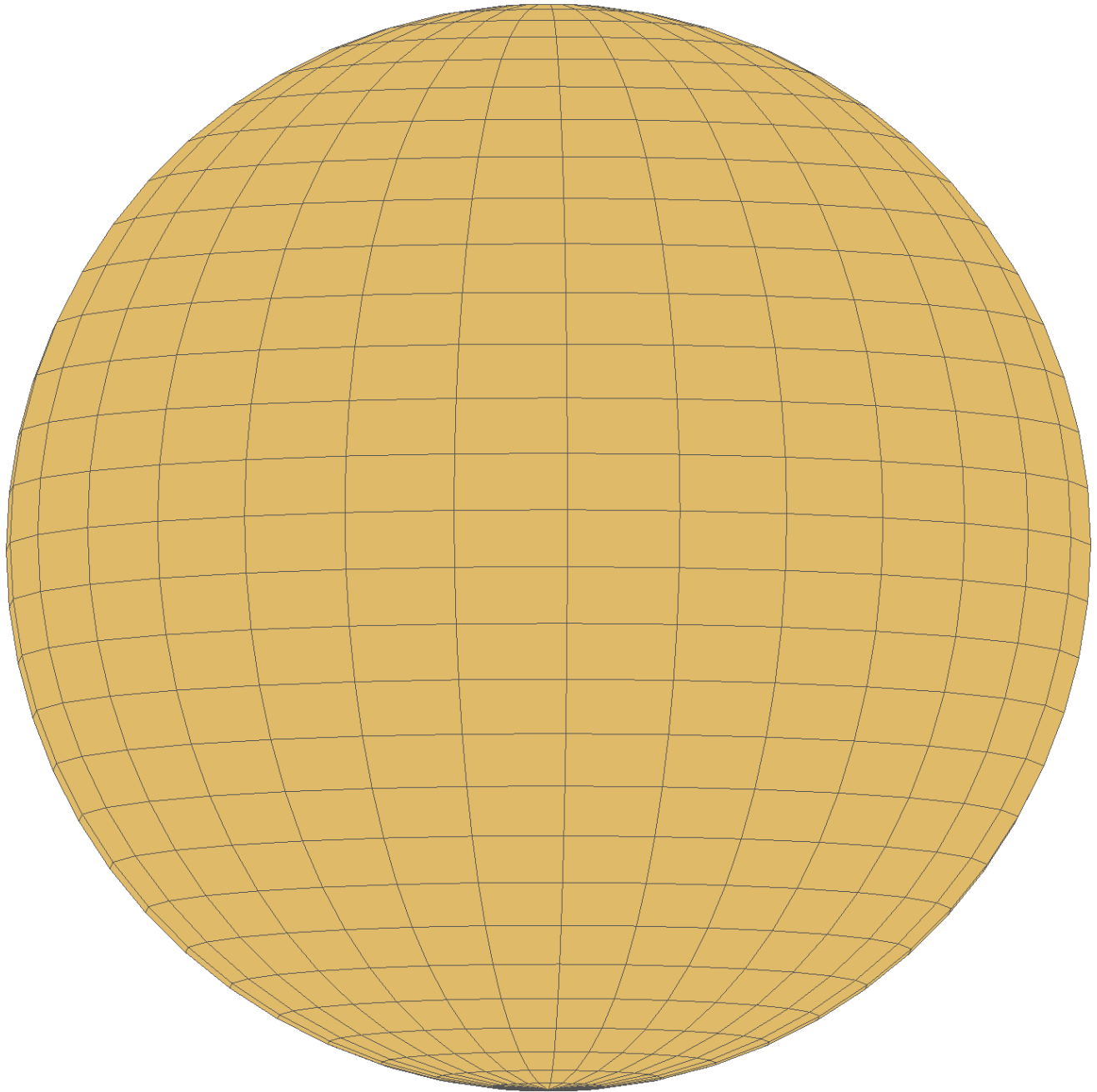
Spheres, orthonormal bases (Stiefel), rotations, positive definite matrices, fixed-rank matrices, Euclidean distance matrices, semidefinite fixed-rank matrices, shapes, linear subspaces (Grassmann), phases, essential matrices, special Euclidean group, fixed-rank tensors, Euclidean spaces...

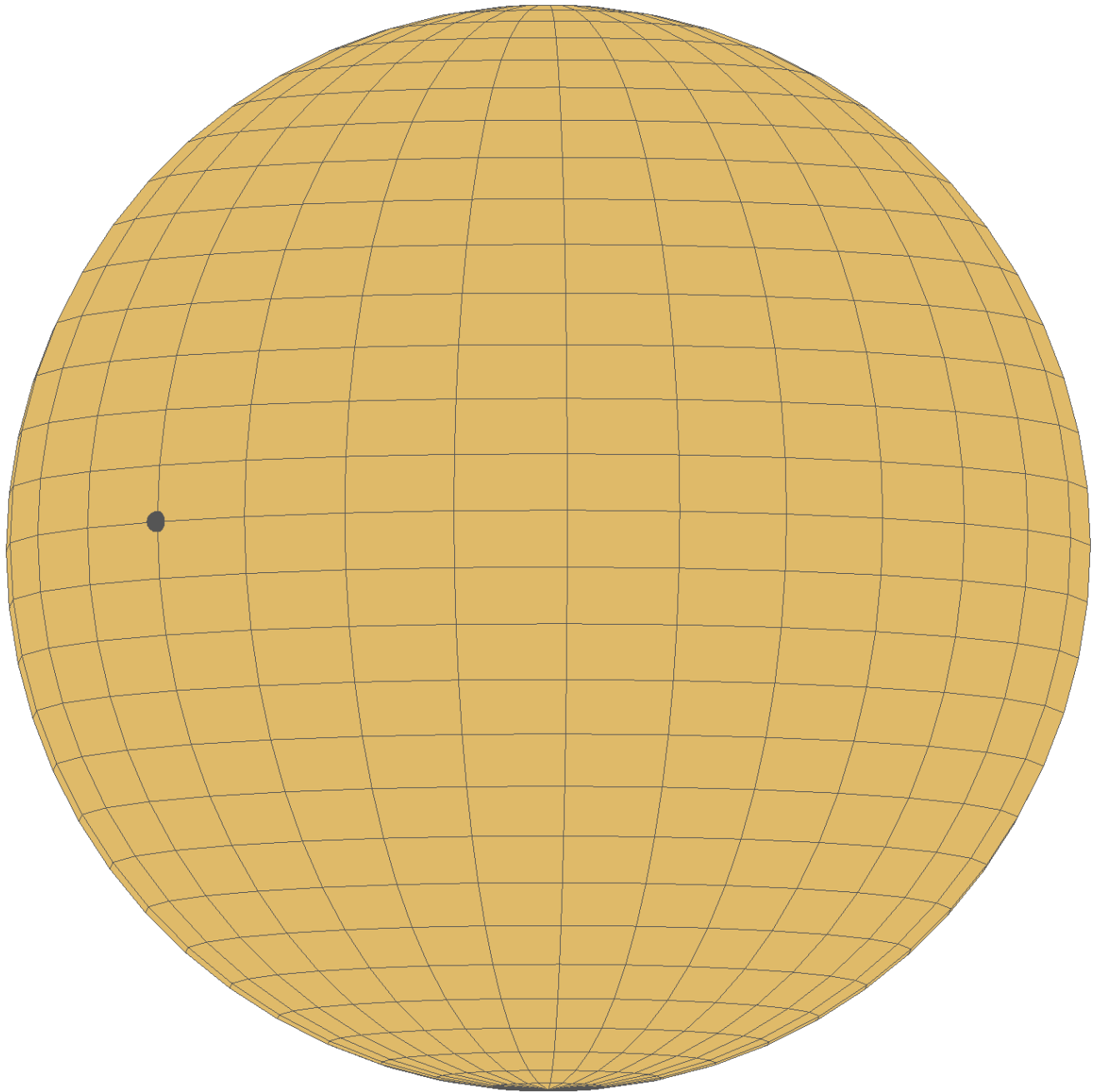
Products and **quotients** of all of these, **real** and **complex** versions...

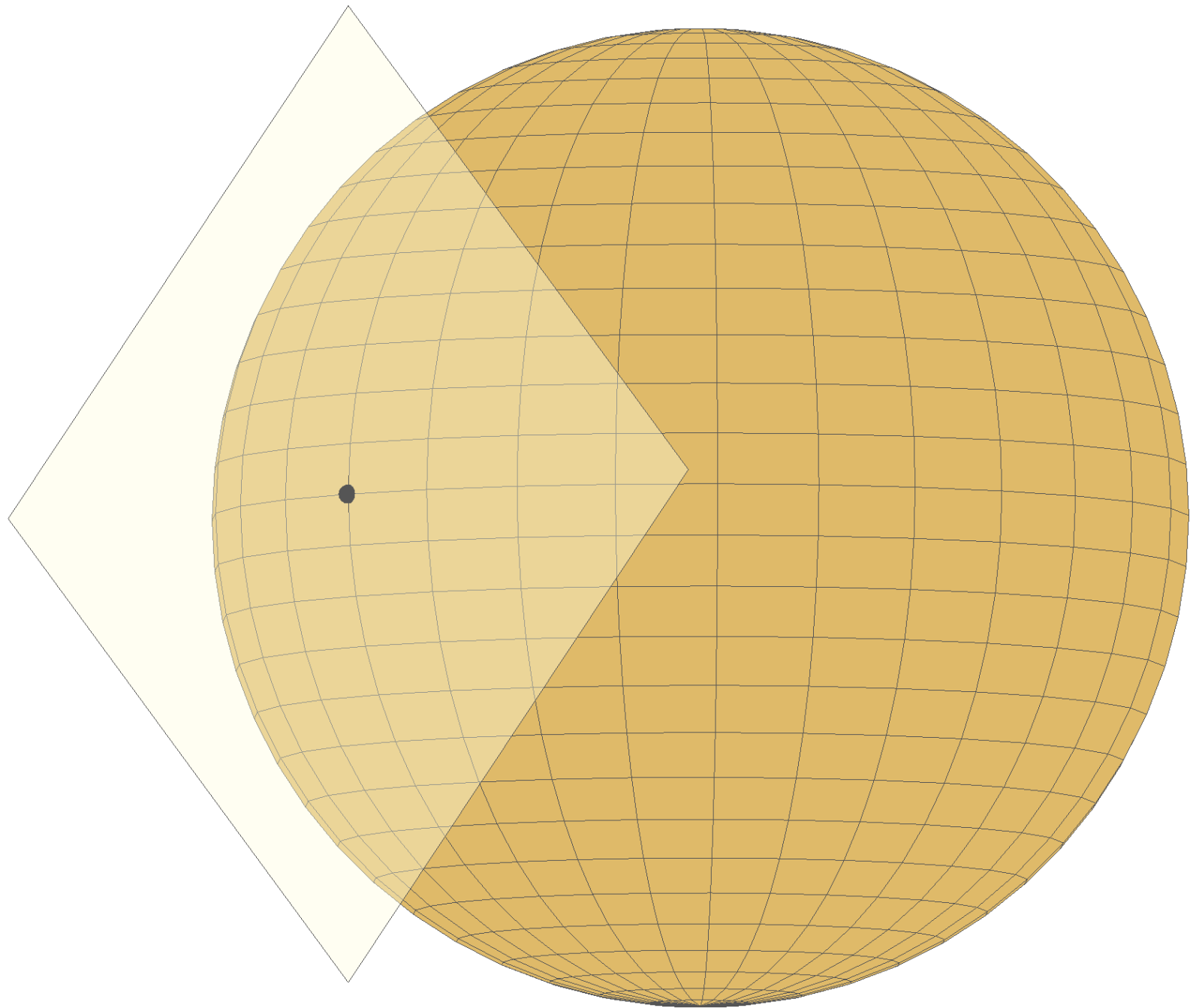
Taking a close look at gradient descent

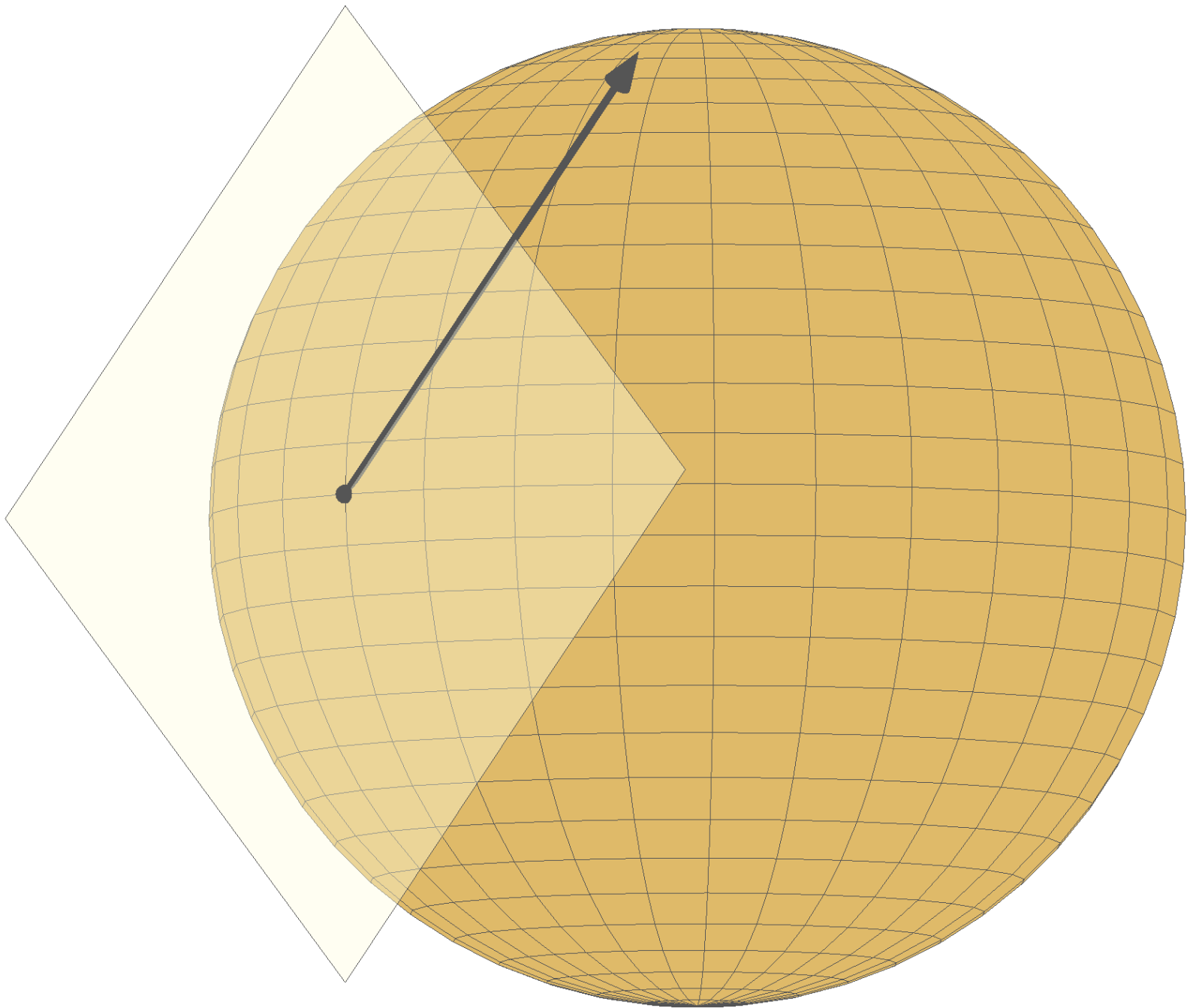


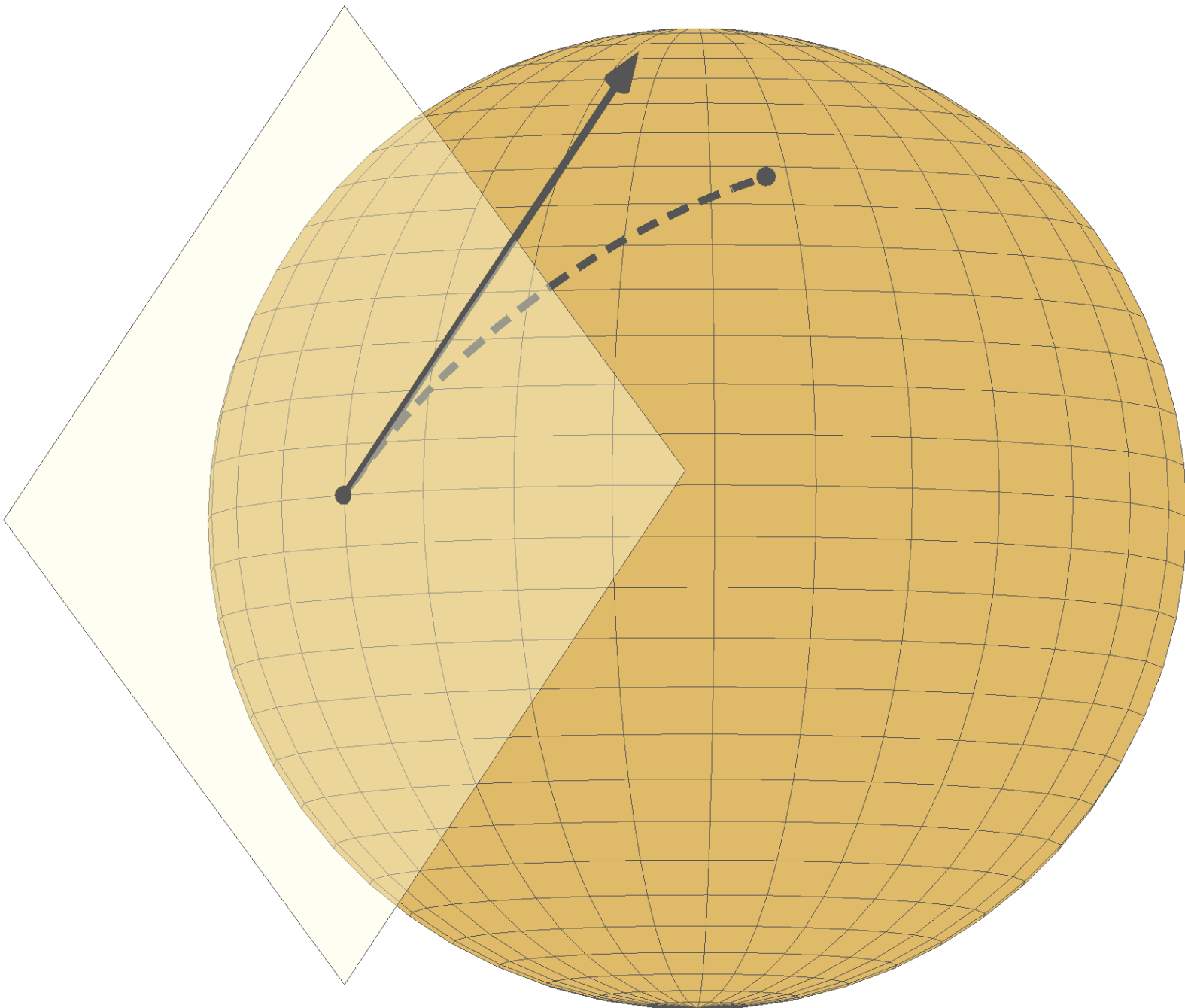












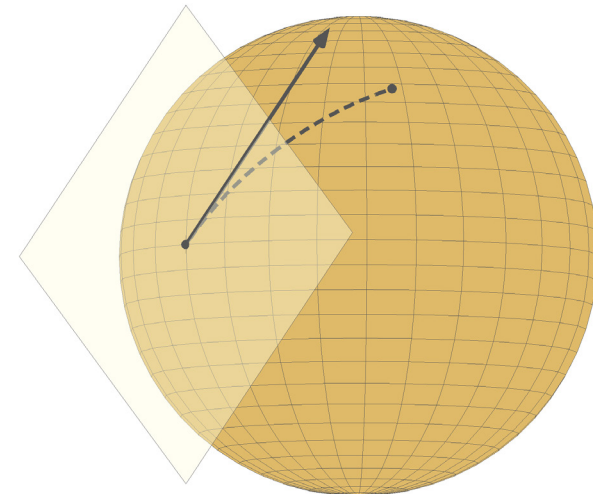
→ We need Riemannian geometry

At each point x in the search space M

We linearize M into a **tangent space** $T_x M$

And pick a **metric** on $T_x M$.

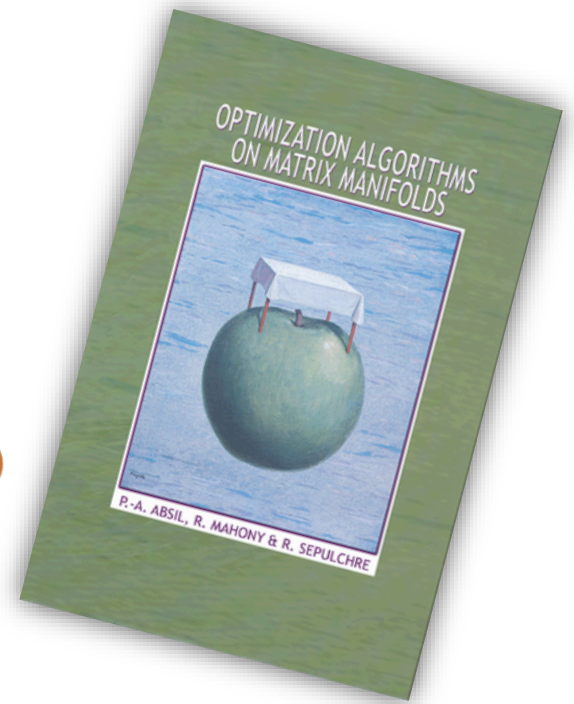
This gives intrinsic notions of **gradient** and **Hessian**.



An excellent book

Optimization algorithms on
matrix manifolds

A Matlab toolbox



www.manopt.org

Manopt [Home](#) [Tutorial](#) [Forum](#) [About](#) [Contact](#)


Welcome to Manopt!

A Matlab toolbox for optimization on manifolds

Optimization on manifolds is a powerful paradigm to address nonlinear optimization problems with various types of constraints that arise naturally in applications, such as orthonormality or low rank.

[Download](#) [Get started](#)

With Mishra, Absil & Sepulchre



Example: Max-Cut relaxation

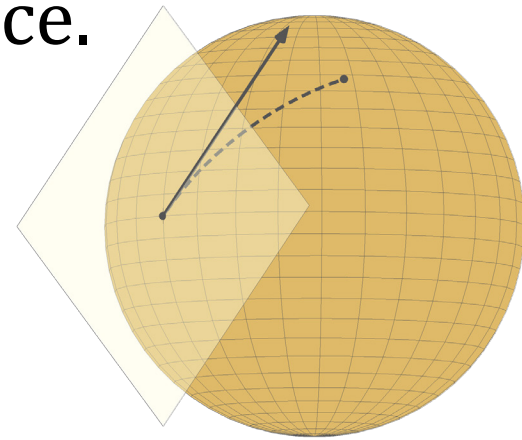
$$\min_{Y:n \times p} \text{Tr}(AYY^T) \text{ s. t. } \text{diag}(YY^T) = \mathbf{1}$$

Rows of Y have unit norm: **product of spheres.**

Tangent space: $\{\dot{Y} : \text{diag}(\dot{Y}Y^T + Y\dot{Y}^T) = \mathbf{0}\}$

Gradient: project $2AY$ to tangent space.

Retraction: normalize rows of $Y + \dot{Y}$.



```

function Y = maxcut_manopt(A)

    % Select an appropriate relaxation rank p.
    n = size(A, 1);
    p = ceil(sqrt(2*n));

    % Select the manifold to optimize over.
    problem.M = obliquefactory(p, n, true);

    % Define the cost function to be minimized.
    problem.cost = @(Y) sum(sum(Y.*(A*Y)));
    problem.egrad = @(Y) 2*(A*Y);
    problem.ehess = @(Y, Ydot) 2*(A*Ydot);

    % Call a standard solver
    % (random initialization, default parameters.)
    Y = trustregions(problem);

```

end

$$\min_{Y:n \times p} \text{Tr}(AYY^T) \text{ s.t. } \text{diag}(YY^T) = \mathbf{1}$$

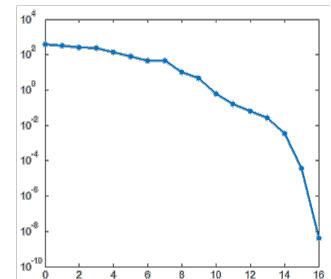
```
>> Y = maxcut_manopt(A);
```

				f: -1.189330e+01	grad : 3.969772e+02	
acc TR+	k:	1	num_inner:	1	f: -5.933834e+03	grad : 3.214287e+02
acc	k:	2	num_inner:	1	f: -1.092386e+04	grad : 2.744089e+02
acc	k:	3	num_inner:	3	f: -1.344741e+04	grad : 2.542660e+02
acc	k:	4	num_inner:	3	f: -1.541521e+04	grad : 1.351628e+02
acc	k:	5	num_inner:	5	f: -1.616969e+04	grad : 7.579978e+01
acc	k:	6	num_inner:	10	f: -1.641459e+04	grad : 4.638172e+01
REJ TR-	k:	7	num_inner:	20	f: -1.641459e+04	grad : 4.638172e+01
acc TR+	k:	8	num_inner:	6	f: -1.654937e+04	grad : 1.057115e+01
acc	k:	9	num_inner:	25	f: -1.656245e+04	grad : 3.576517e+00
acc	k:	10	num_inner:	18	f: -1.656370e+04	grad : 3.951183e-01
acc	k:	11	num_inner:	43	f: -1.656377e+04	grad : 1.330375e-01
acc	k:	12	num_inner:	48	f: -1.656378e+04	grad : 5.752944e-02
acc	k:	13	num_inner:	67	f: -1.656378e+04	grad : 2.430253e-02
acc	k:	14	num_inner:	89	f: -1.656378e+04	 grad : 2.475079e-03
acc	k:	15	num_inner:	123	f: -1.656378e+04	 grad : 1.896680e-05
acc	k:	16	num_inner:	224	f: -1.656378e+04	 grad : 1.103767e-09

Gradient norm tolerance reached; options.tolgradnorm = 1e-06.

Total time is 5.14 [s]

Optimality gap: $n \cdot \lambda_{\min}(S(Y)) = -4.2 \cdot 10^{-6}$



Convergence guarantees for Riemannian **gradient descent**

Global convergence to **critical points**.

Linear convergence rate locally.

Reach $\|\text{grad}f(x)\| \leq \varepsilon$ in $O\left(\frac{1}{\varepsilon^2}\right)$ iterations
under Lipschitz assumptions.
With Cartis & Absil (arXiv 1605.08101).



Convergence guarantees for Riemannian **trust regions**

Global convergence to
second-order critical points.

Quadratic convergence rate locally.

$\|\text{grad}f(x)\| \leq \varepsilon$ and $\text{Hess}f(x) \succcurlyeq -\varepsilon I$ in $O\left(\frac{1}{\varepsilon^3}\right)$

iterations under Lipschitz assumptions.

With Cartis & Absil (arXiv 1605.08101).



Low-rank matrix completion

Riemannian preconditioning for tensor completion

Hiroyuki Kasai*

Graduate School of Information Systems,
The university of Electro-Communications
Chofu-shi, Tokyo, 182-8585, Japan
kasai@is.uec.ac.jp

Bamdev Mishra†

Department of EECS,
University of Liège
4000 Liège, Belgium
b.mishra@ulg.ac.be

Abstract

We propose a novel Riemannian preconditioning approach for the

Journal of Machine Learning Research 11 (2010) 2057-2078

Submitted

Matrix Completion from Noisy Entries

Raghunandan H. Keshavan

Andrea Montanari*

Sewoong Oh

Department of Electrical Engineering
Stanford University
Stanford, CA 94304, USA

RAG

MON

Editor: Tommi Jaakkola

Linear Algebra and its Applications 475 (2015) 200–239



ELSEVIER

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa



Low-rank matrix completion via preconditioned optimization on the Grassmann manifold



Nicolas Boumal^{a,*}, P.-A. Absil^b

^a Inria & D.I., UMR 8548, Ecole Normale Supérieure, Paris, France

SIAM J. OPTIM.
Vol. 23, No. 2, pp. 1214–1236

© 2013 Society for Industrial and Applied Mathematics

LOW-RANK MATRIX COMPLETION BY RIEMANNIAN OPTIMIZATION*

BART VANDEREYCKEN†

Abstract. The matrix completion problem consists of finding or approximating a low-rank matrix based on a few samples of this matrix. We propose a new algorithm for matrix completion that minimizes the least-square distance on the sampling set over the Riemannian manifold of fixed-rank matrices. The algorithm is an adaptation of classical nonlinear conjugate gradients, developed within the framework of retraction-based optimization on manifolds. We describe all the necessary objects from differential geometry necessary to perform optimization over this low-rank matrix manifold, seen as a submanifold embedded in the space of matrices. In particular, we describe how metric projection can be used as retraction and how vector transport lets us obtain the conjugate search directions. Finally, we prove convergence of a regularized version of our algorithm under the assumption that the restricted isometry property holds for incoherent matrices throughout the iterations. The numerical experiments indicate that our approach scales very well for large-scale problems and compares favorably with the state-of-the-art, while outperforming most existing solvers.

Key words. matrix completion, low-rank matrices, optimization on manifolds, differential geometry, nonlinear conjugate gradients, Riemannian manifolds, Newton

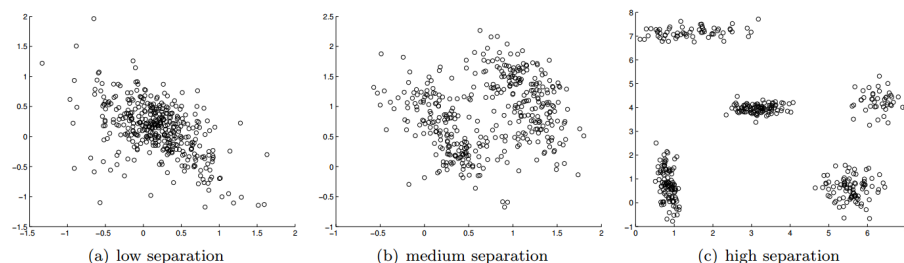
32

AMS subject classifications. 15A83, 65K05, 53B21

Gaussian mixture models

*Matrix Manifold Optimization
for Gaussian Mixture Models*

Reshad Hosseini, Suvrit Sra,
2015 (NIPS)



$$p(\mathbf{x}) := \sum_{j=1}^K \alpha_j p_{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad \mathbf{x} \in \mathbb{R}^d,$$

and where $p_{\mathcal{N}}$ is a (multivariate) Gaussian with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \succ 0$. That is,

$$p_{\mathcal{N}}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) := \det(\boldsymbol{\Sigma})^{-1/2} (2\pi)^{-d/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right).$$

Given i.i.d. samples $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, we wish to estimate $\{\hat{\boldsymbol{\mu}}_j \in \mathbb{R}^d, \hat{\boldsymbol{\Sigma}}_j \succ 0\}_{j=1}^K$ and weights $\hat{\boldsymbol{\alpha}} \in \Delta_K$, the K -dimensional probability simplex. This leads to the *GMM optimization* problem

$$\max_{\boldsymbol{\alpha} \in \Delta_K, \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \succ 0\}_{j=1}^K} \sum_{i=1}^n \log\left(\sum_{j=1}^K \alpha_j p_{\mathcal{N}}(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right). \quad (2.1)$$

Dictionary learning

Complete Dictionary Recovery over the Sphere I: Overview and the Geometric Picture

Ju Sun, *Student Member, IEEE*, Qing Qu, *Student Member, IEEE*, and John Wright, *Member, IEEE*

Abstract

We consider the problem of recovering a complete (i.e., square and invertible) matrix A_0 , from $Y \in \mathbb{R}^{n \times p}$ with $Y = A_0 X_0$, provided X_0 is sufficiently sparse. This recovery problem is central to the theoretical understanding of dictionary learning, which seeks a sparse representation for a collection of input signals, and finds numerous applications in modern

recovers A_0 when X_0 results based on efficient for any constant $\delta \in (0$

Our algorithmic problem is tractable, we shows that with high probability to design a Riemannian an arbitrary initialization shed light on other problems

This paper provides objective landscape. In are presented.

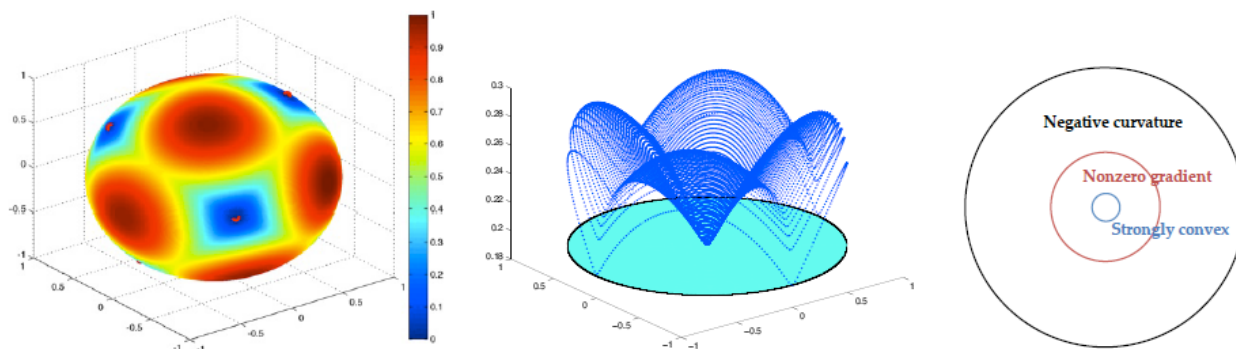


Fig. 2: Why is dictionary learning over \mathbb{S}^{n-1} tractable? Assume the target dictionary A_0 is orthogonal. **Left:** Large sample objective function $\mathbb{E}_{X_0} [f(q)]$. The only local minimizers are the columns of A_0 and their negatives. **Center:** The same function, visualized as a height above the plane a_1^\perp (a_1 is the first column of A_0 , and is also a global minimizer). **Right:** Around a_1 , the function exhibits a small region of positive curvature, a region of large gradient, and finally a region in which the direction away from a_1 is a direction of negative curvature.

Phase retrieval

A Geometric Analysis of Phase Retrieval

Ju Sun, Qing Qu, and John Wright
{js4038, qq2105, jw2966}@columbia.edu

Department of Electrical Engineering, Columbia University, New York, USA

January 31, 2016

Abstract

Can we recover a complex signal from its Fourier magnitudes? More generally, given a set of m measurements, $y_k = |a_k^* x|$ for $k = 1, \dots, m$, is it possible to recover $x \in \mathbb{C}^n$ (i.e. length- n complex vector)? This is the generalized phase retrieval problem, which is a natural non-convex task in various disciplines. Natural non-convex optimization is in practice, but lack clear theoretical explanation for this gap. We show that when the measurements are incoherent and the number of measurements is large, (1) the natural least-squares formulation for phase retrieval has no spurious local minimizers, only the global minimum and its equivalent copies; and (2) the objective function has a benign geometric structure that allows a number of algorithms to find a global minimizer without special initialization. We analyze the landscape of the least-squares formulation and propose a second-order trust-region algorithm.

Keywords. Phase retrieval, Nonconvex optimization, Ridable saddles, Trust-region method

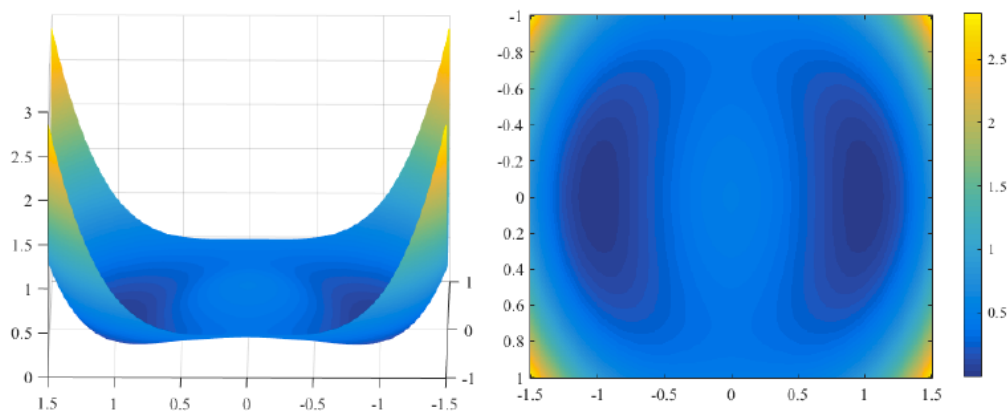


Figure 5: Function landscape of (1.1) for $x = [1; 0]$ and $m \rightarrow \infty$ for the masked Fourier transform measurements (coded diffraction model [CLS15b]). Compared to the landscape under the Gaussian model (Figure 2), the landscape here has an analogous shape qualitatively. The benign geometric structure is evident.

Phase synchronization

Nonconvex phase synchronization

Nicolas Boumal*

March 29, 2016

Abstract

We estimate n phases (angles) from noisy pairwise relative phase measurements. The task is modeled as a nonconvex least-squares optimization problem. It was recently shown that this problem can be solved in polynomial time via convex relaxation, under some conditions on the noise. In this paper, under similar but more restrictive conditions, we show that a modified version of the power method converges to the global optimizer. This is simpler and (empirically) faster than convex approaches. Empirically, they both succeed in the same regime. Further analysis shows that, in the same noise regime as previously, second-order necessary optimality conditions for this quadratically constrained quadratic program are also sufficient, despite nonconvexity.

1 Introduction

We consider the problem of estimating n phases (complex numbers with unit modulus) based on noisy measurements of the relative phases. The target parameter is

$$z \in \mathbb{C}_1^n \triangleq \{x \in \mathbb{C}^n : |x_1| = \dots = |x_n| = 1\}, \quad (1)$$

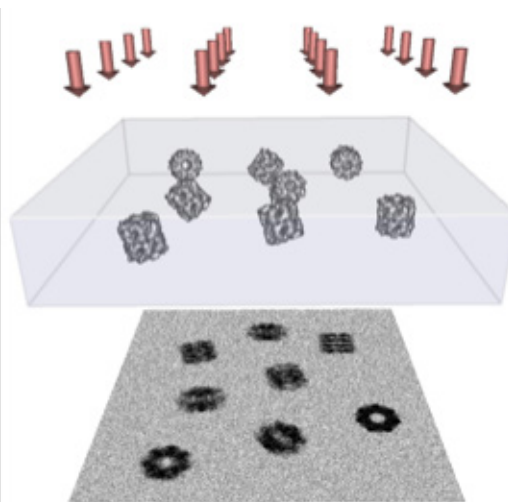
and the measurements $C_{ij} \approx z_i \bar{z}_j$ are stored in the Hermitian matrix

$$C = zz^* + \Delta, \quad (2)$$

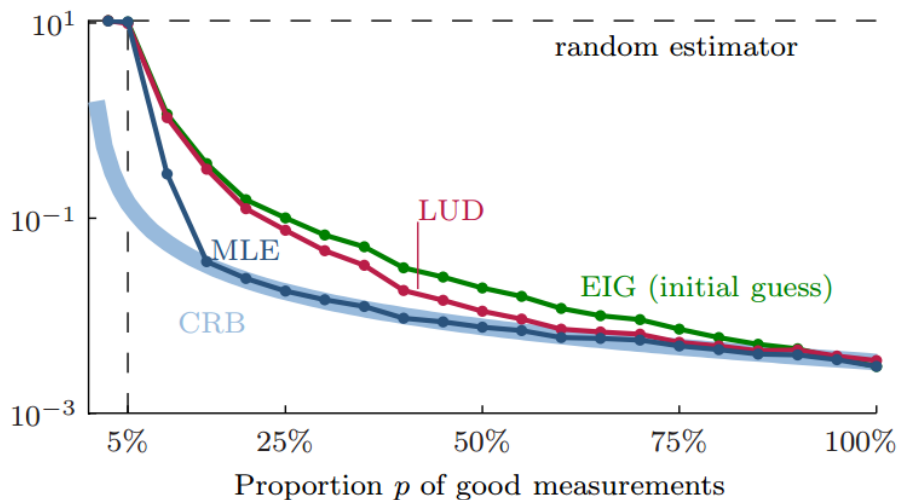
where Δ is a Hermitian perturbation. Motivated by the scenario where Δ contains white Gaussian noise, we focus on the associated maximum likelihood estimation problem (it

Synchronization of rotations

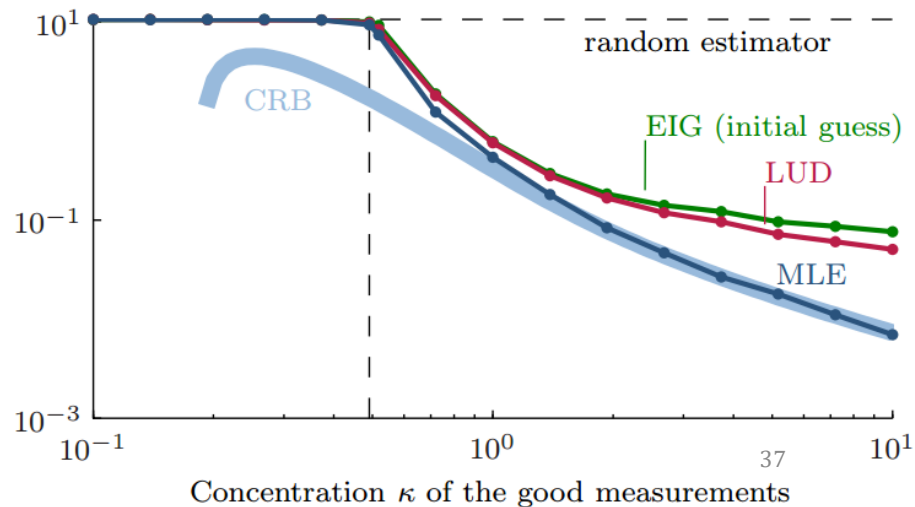
Robust estimation of rotations, 2013
B., Singer and Absil



Expected MSE, estimated over 100 realizations



Expected MSE, estimated over 60 realizations ($p = 5/\sqrt{N}$)

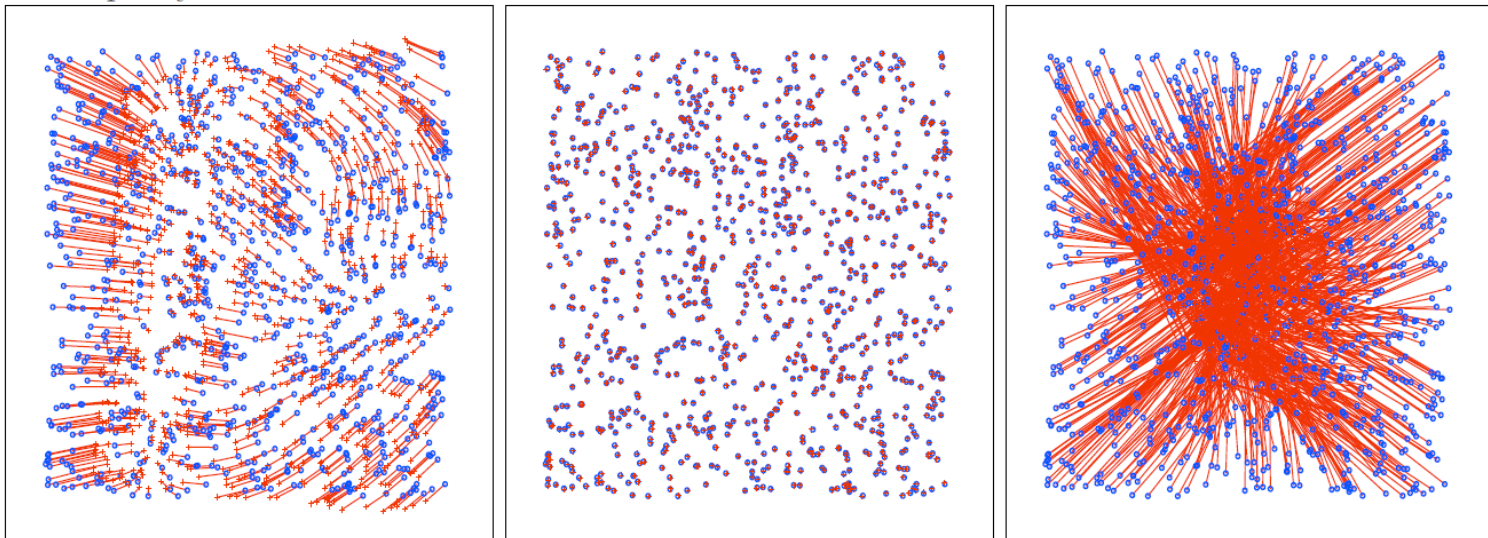


Sensor network localization

Noisy sensor network localization, robust facial reduction and the Pareto frontier

Cheung, Drusvyatskiy, Krislock and Wolkowicz 2014

Figure 2: Illustration of robust facial reduction with refinement applied on an instance with 1000 sensors (no anchors) on a $[-0.5, 0.5]^2$ box, with noise factor 0.05 and radio range 0.1. From left to right: (1) using **Algorithm 1 without refinement** (RMSD= 61.52% R); (2) using **Algorithm 1 with refinement via Manopt** (RMSD= 1.39% R); (3) using **only Manopt** (RMSD= 380.59% R). Blue: true location; red: estimated location and discrepancy.



Protein structure determination in NMR spectroscopy

Residual Dipolar Coupling, Protein Backbone Conformation and Semidefinite Programming

Yuehaw Khoo, Amit Singer and David Cowburn, 2016

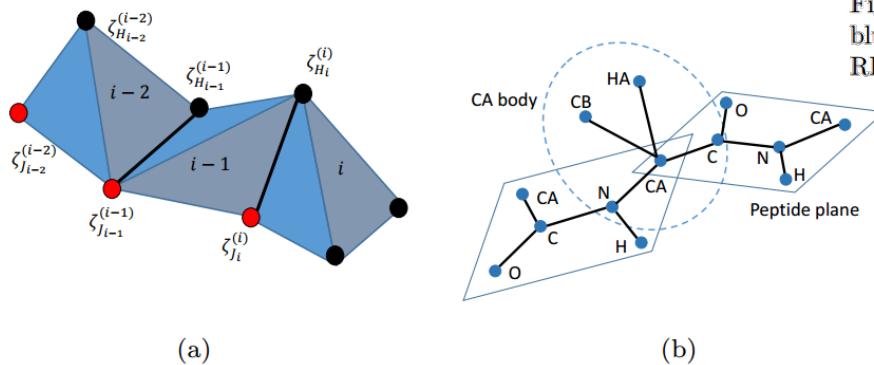


Figure 1 (a) Example of an articulated structure with joints with indices J_i 's (Red dots) and H_i 's. The hinges are represented by black bars in the figure. (b) Protein backbone consists of peptide planes and CA bodies. These rigid units are chained together at the bonds (N, CA) and (C,CA).

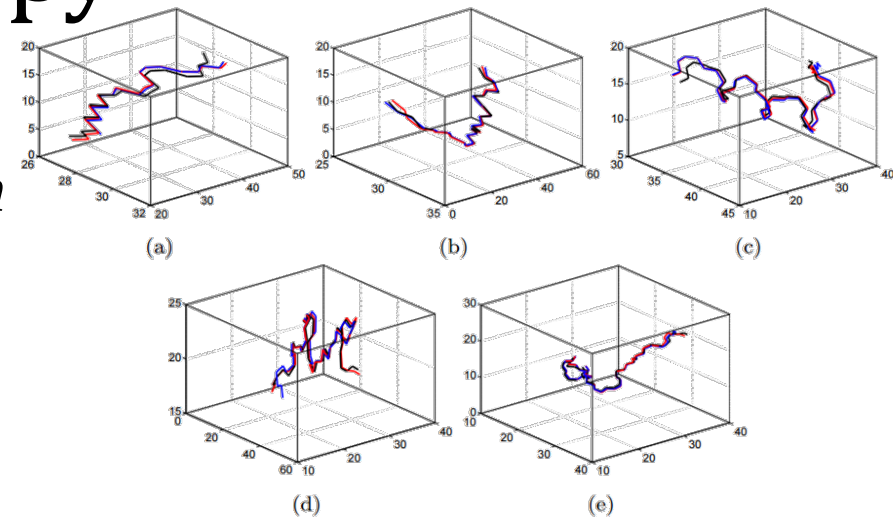
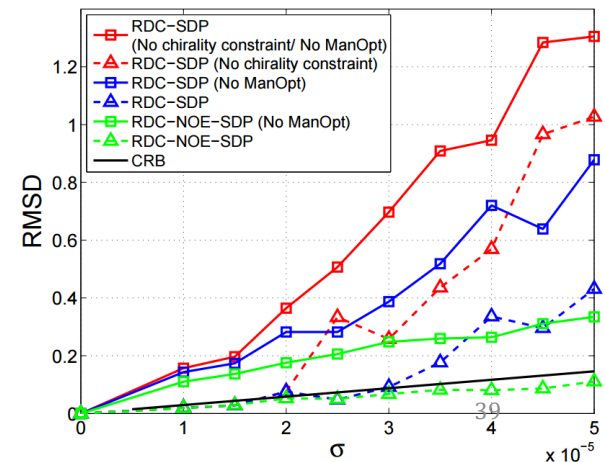


Figure 4 The trace of protein backbone drawn using N, CA and C. The black, blue and red curves come from the X-ray model 1UBQ, RDC-SDP solution and RDC-NOE-SDP respectively.



Nonsmooth with MADMM

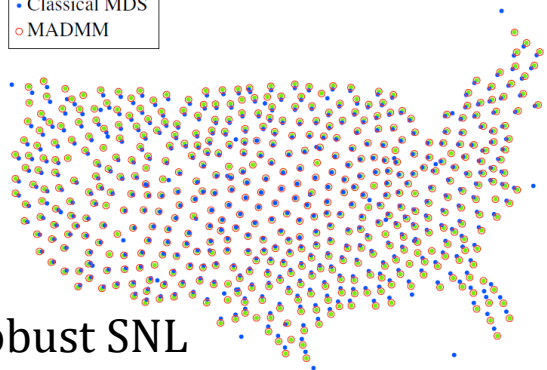
MADMM: a generic algorithm for non-smooth optimization on manifolds, Kovnatsky, Glashoff, Bronstein, 2015

Compressed modes



Figure 1: The first six compressed modes of a human mesh containing $n = 8K$ points computed using MADMM. Parameter $\mu = 10^{-3}$ and three manifold optimization iterations in X -step were used in this experiment.

- Groundtruth
- Classical MDS
- MADMM



Functional correspondence

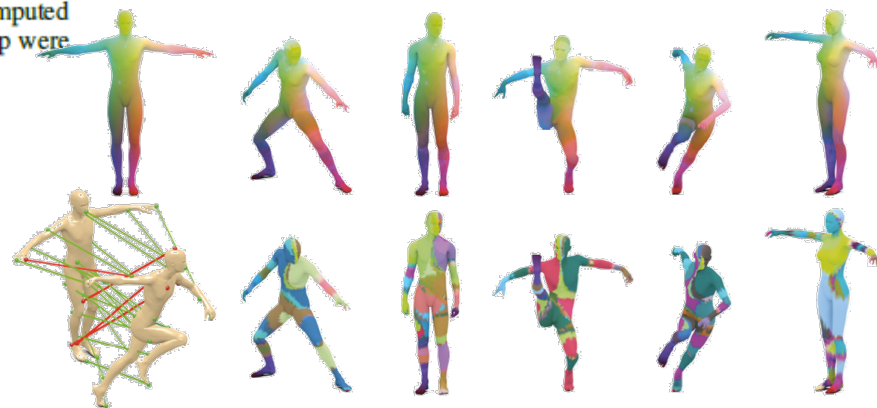


Figure 4: Examples of correspondences obtained with MADMM (top) and least-squares solution (bottom). Similar colors encode corresponding points. Bottom left: examples of correspondence between a pair of shapes (outliers are shown in red).

Take home message

Optimization on manifolds has many **applications** and is easy to try with **Manopt**.

It comes with the same **guarantees** as unconstrained nonlinear optimization.

For some problems, we get **global optimality**.

Max-Cut

A is the adjacency matrix of the graph:

$$\max_{x_1, \dots, x_n \in \{\pm 1\}} \sum_{i,j} A_{ij} \frac{1 - x_i x_j}{2}$$

$$\max_{x_1, \dots, x_n \in \{\pm 1\}} \mathbf{1}^T A \mathbf{1} - x^T A x$$

$$\min_x x^T A x \text{ s. t. } x_i^2 = 1 \forall i$$

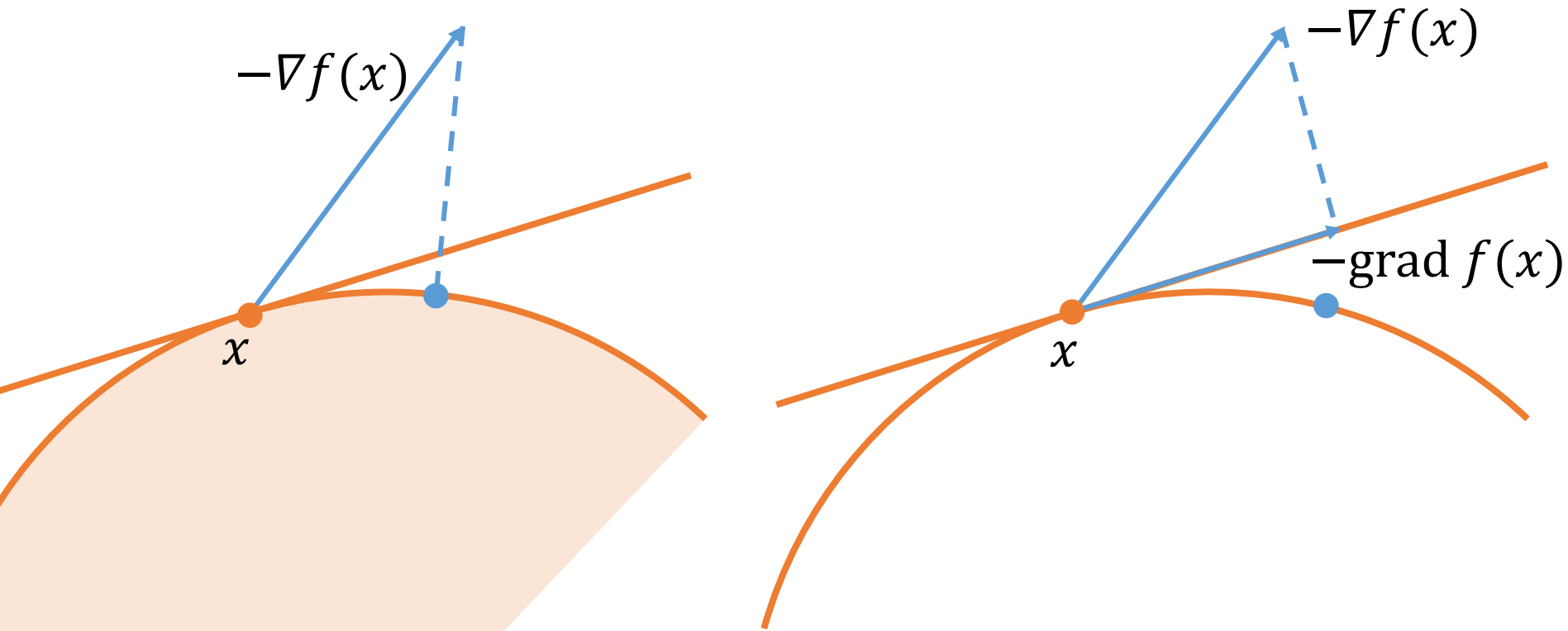
Max-Cut

$$\min_x x^T A x \quad \text{s. t.} \quad x_i^2 = 1 \quad \forall i$$

$$\min_x \text{Tr}(A x x^T) \quad \text{s. t.} \quad (x x^T)_{ii} = 1 \quad \forall i$$

$$\begin{aligned} \min_X \text{Tr}(A X) \quad \text{s. t.} \quad & \text{diag}(X) = \mathbf{1}, \\ & X \succeq 0 \\ & \text{rank}(X) = 1 \end{aligned}$$

This is *not* projected gradients



Optimization on manifolds is **intrinsic**.
There is no need for an embedding space.
Works for abstract manifolds, **quotient spaces**.