The non-convex Burer-Monteiro approach works on smooth semidefinite programs*

ORIE 7191

Mateo Díaz

February 7, 2018

***This talk is based on Boumal, Voroninski, and Bandeira (2016).**

Quiz

1. Which one of the following assumptions of the solution set is required for the main result?

    a) Smoothness b) Prox-regularity
    c) Outer-semicontinuity d) None of these.

2. The main technique used to reduce the dimension was?

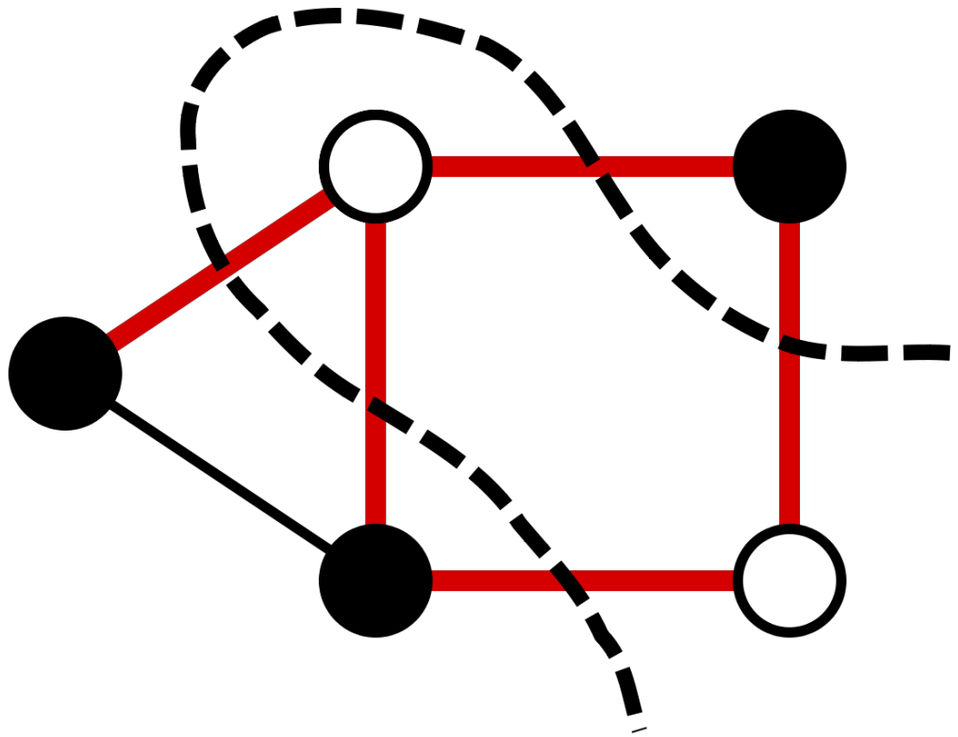
    a) Random projection b) Burer-Monteiro formulation
    c) Fourier transform d) JPEG compression.

3. The last name of the third author is

    a) Hernandez b) Chen
    c) Bandeira d) Smith.

A motivating example

Given a graph the **Max-Cut** problem asks one to cluster the nodes in two classes, $\{-1, +1\}$, such that as many edges as possible join nodes of different signs.



As you probably imagine, Max-Cut is **NP-hard**.

A motivating example

But **fear not!** In a remarkably beautiful paper Goemans and Williamson (1995) proved that there is an 0.878-approximation algorithm, i.e.

$$\text{Number of cuts of the algorithm} \geq 0.878 \times \text{Optimal number of cuts}$$

The algorithm is based on a relaxation of

$$\max_{x \in \mathbb{R}^n} \frac{1}{4} \sum_{i,j=1}^{n} C_{ij}(1 - x_i x_j) \qquad \text{subject to} \qquad x_1^2 = \cdots = x_n^2 = 1.$$

to the SDP

$$\max_{x \in \mathbb{R}^{n \times n}} \langle C, X \rangle \qquad \text{subject to} \qquad \operatorname{diag}(X) = \mathbf{1},\ X \succeq 0.$$

A motivating example

$$\max_{x \in \mathbb{R}^{n \times n}} \langle C, X \rangle \qquad \text{subject to} \qquad \text{diag}(X) = \mathbf{1},\ X \succeq 0.$$

This is great, there are **polynomial time methods** to solve SDPs. But in **practice** the exponents are too high and we can only solve **small problems.**

Also, we **lifted the problem** from $\mathbb{R}^n$ to $\mathbb{R}^{n^2}$:(

To keep in mind: if we assume the complexity was linear in the dimension and the computer executed 1 operation per second.

$$\text{If with } d = n, \text{ it takes one day} \implies d = n^2, \text{ it takes } 236 \text{ years.}$$

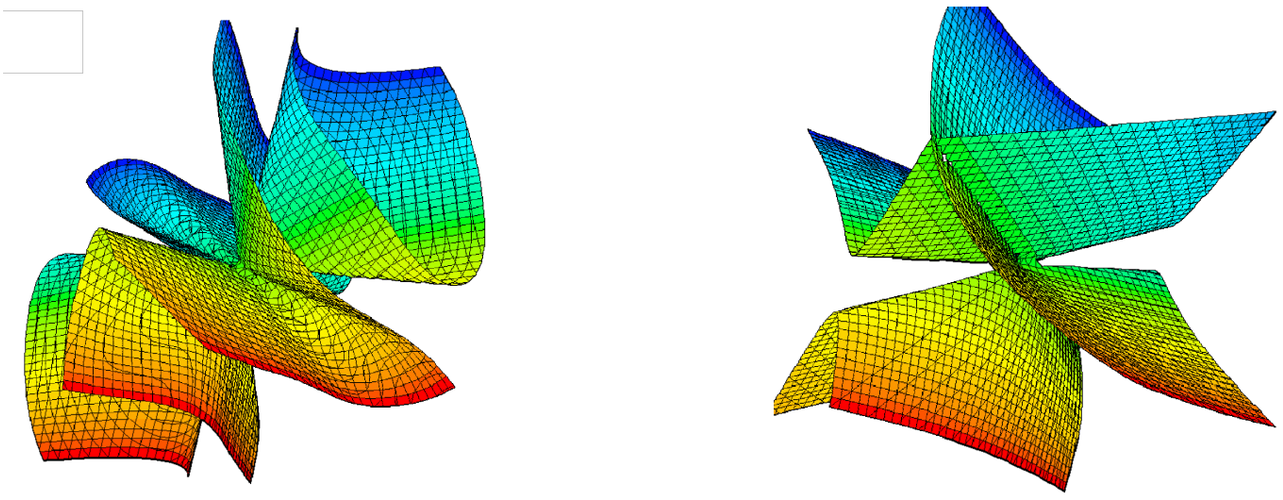
Burer-Monteiro factorization

Possible solution. Burer and Monteiro (2005) proposed to consider a factored formulation

$$\min_{x \in \mathbb{R}^{n \times p}} \langle C, YY^\top \rangle \qquad \text{subject to} \qquad \operatorname{diag}(YY^\top) = \mathbf{1}.$$

The good parts. Smaller dimension, no semidefinite constraint, we can run gradient descent.

The scary parts. Nonconvexity :O



Step back: a general setting

Goal. We wish to solve the following semidefinite program **(SDP)**

$$f^* = \min_{X \in \mathbb{S}^{n \times n}} \langle C, X \rangle \quad \text{subject to} \quad \mathcal{A}(X) = b,\ X \succeq 0,$$

where $\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ and $b \in \mathbb{R}^m$ are fixed and known.

Approach. Apply Burer-Monteiro factorization **(BM)**

$$q^* = \min_{Y \in \mathbb{R}^{n \times p}} \langle CY, Y \rangle \quad \text{subject to} \quad \mathcal{A}(YY^\top) = b.$$

Questions.

1. How to choose p?

2. Guarantees on the set of critical points of BM?

First question

How to choose p? This answer was answered by Burer and Monteiro. In fact, due to Pataki (1995) the know that if the **SDP search space is compact** then there exists a solution with rank r with

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

Then, setting

$$\frac{p(p+1)}{2} \geq m$$

suffices to have

$$f^* = q^*.$$

Second question (Main takeaway)

Theorem (Informal).

Assume $p(p+1)/2 \geq m$, further the search space of (SDP) is **compact**, and the search space of (BM) is a **regularly defined smooth manifold**. Then for **almost all** C we have that

$$Y \text{ is a } \textbf{second-order critical point} \text{ of (BM)}$$
$$\implies Y \text{ is a } \textbf{global solution} \text{ of (SDP)}$$

Remarks.

1. This is a geometrical statement, not a computational one.
2. We are not talking about local minima.

Roadmap

- **Preliminaries**
- **Main results**
- **Examples**
- **Sketch of the proof**

Some remarks on the linear map

The operator is linear

$$\mathcal{A}(X) = \left[\langle A_1, X\rangle, \ldots, \langle A_m, X\rangle\right]^\top$$

Furthermore, we only care about symmetric matrices, then

$$\mathcal{A}(X) = \mathcal{A}\left(\frac{X + X^\top}{2}\right)$$

Recall that $\mathbb{R}^{n\times n} = \mathcal{S}^{n\times n} \oplus \left(\mathcal{S}^{n\times n}\right)^\circ$, hence

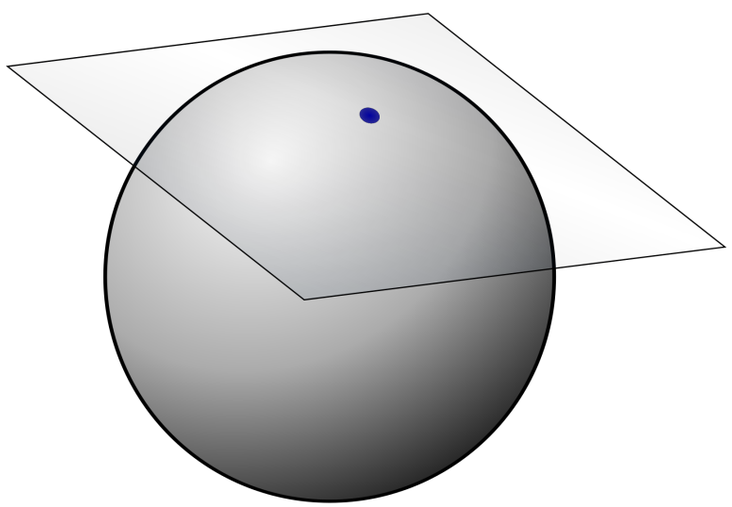
$$\ker\mathcal{A} = \left(\mathcal{S}^{n\times n}\right)^\circ.$$

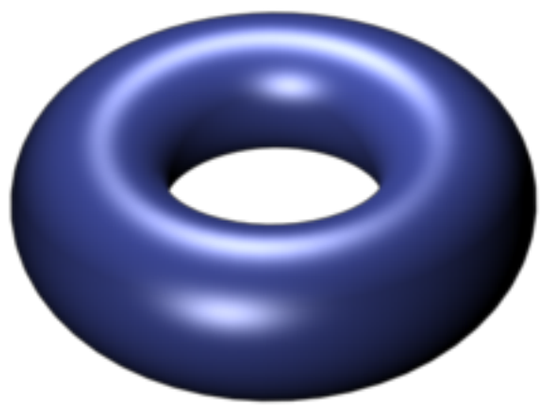
Manifolds

Heuristically think of a manifold $\mathcal{M}$ as **higher dimensional surface**. Locally it looks like a plane, which gives a notion of dimension.

At each point $Y \in \mathcal{M}$, one can construct a **tangent plane** (exactly what you are thinking) denoted by $T_Y \mathcal{M}$.

In $\mathbb{R}^n$ we can describe them locally as zeros of smooth functions or images of smooth functions.





Manifolds

For our problem, we assume that the search space to be a smooth manifold

$$\mathcal{M} = \mathcal{M}_p = \{Y \in n \times p \mid \mathcal{A}(YY^\top) = b\}$$

Additionally, we assume the manifold is “regular”

$$A_1 Y, \ldots, A_m Y \text{ are linearly independent for all } Y \in \mathcal{M}.$$

Then, the tangent space at any point can be written as

$$T_Y \mathcal{M} = \left\{ \dot{Y} \mid \mathcal{A}\left(\dot{Y}Y^\top + Y\dot{Y}^\top\right) = 0 \right\}$$

Some remarks on the objective

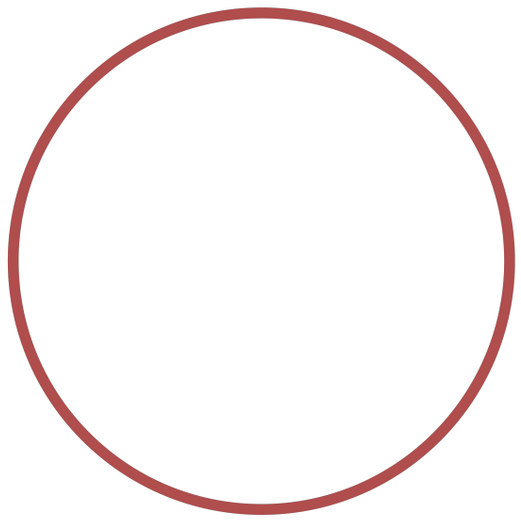
Notice that in the case of BM, the objective is quadratic

$$f(Y) = \langle C, YY^\top \rangle$$

As a simple example

$$C = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \qquad A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad b = 1.$$

If $p = 1$, what is the problem finding?

$\mathcal{M}_1 =$ 

Minimum singular value and vector of C!

Gradient and Hessian for manifolds

Just as in Euclidean spaces, inside manifolds we can also define gradients and Hessians of smooth functions.

- The **gradient** at $Y \in \mathcal{M}$ is an element of the tangent space $\operatorname{grad} f(Y) \in T_Y\mathcal{M}$, given by

$$\operatorname{grad} f(Y) = \operatorname{Proj}_Y \nabla f(Y).$$

- The **Hessian** at $Y \in \mathcal{M}$ is a linear mapping acting on the tangent $\operatorname{Hess} f(Y) : T_Y\mathcal{M} \to T_Y\mathcal{M}$, given by

$$\operatorname{Hess} f(Y)[\dot{Y}] = \operatorname{Proj}_Y D\left(\tilde{Y} \mapsto \operatorname{grad} f(\tilde{Y})\right)(Y)[\dot{Y}]$$

SAY WHAT NOW?

Think of them as restriction of the classical gradient and Hessian to the manifold.

Critical points

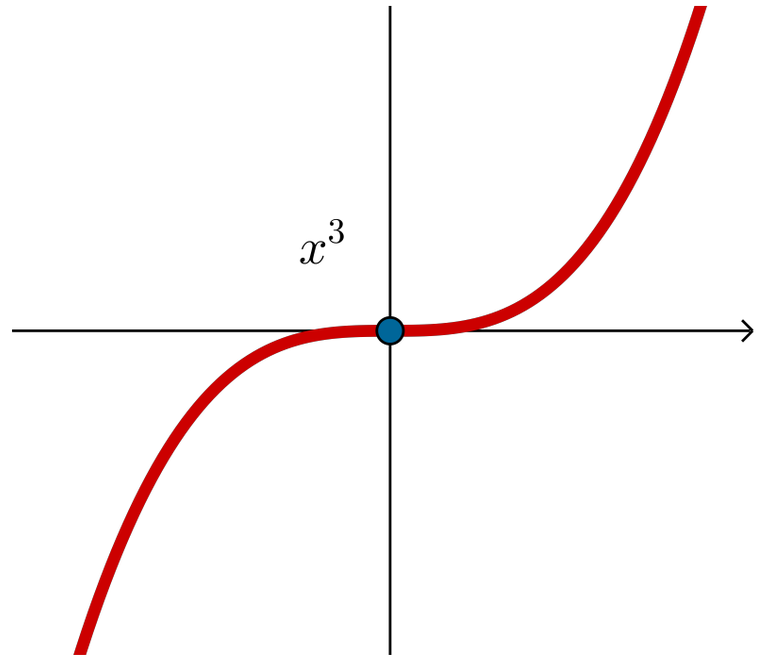
We say that a point $Y \in \mathcal{M}$ is a **first-order critical point** for a function $f : \mathcal{M} \subseteq \mathbb{R}^n \to \mathbb{R}$ restricted to $\mathcal{M}$ if

$$\operatorname{grad} f(Y) = 0$$

We say that a point $Y \in \mathcal{M}$ is a **second-order critical point** for a function $f : \mathcal{M} \subseteq \mathbb{R}^n \to \mathbb{R}$ restricted to $\mathcal{M}$ if it is first-order critical and

$$\operatorname{Hess} f(Y) \succeq 0$$

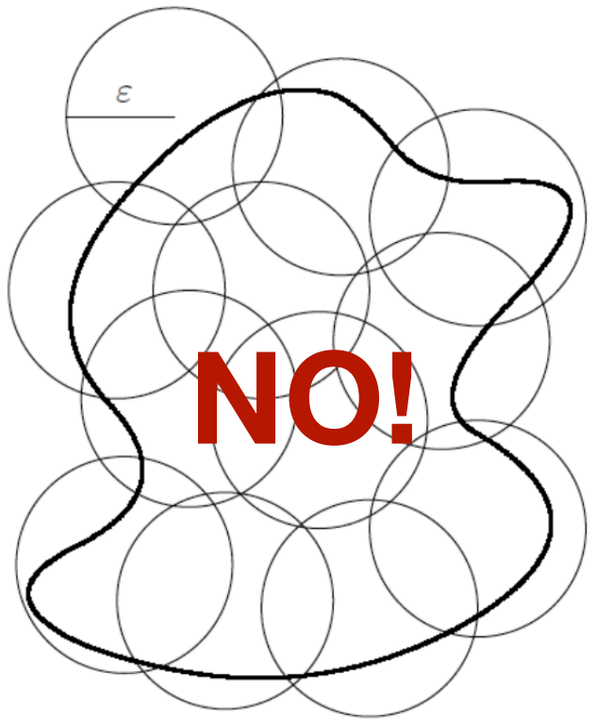
Remark. Second-order critical points are **different** from local minima!



Sets of measure zero

We say that a set $S \subseteq \mathbb{R}^n$ has **measure zero** if for any $\delta > 0$ it can be cover by balls (or cubes) $B_1, B_2, \ldots$ satisfying

$$\sum_{i=1}^{\infty} \mathrm{vol}(B_i) \leq \delta.$$



Remark.

1. Countable unions of negligible sets are negligible (why?).
2. If you add a random perturbation, you get out of these set.

Examples. Low dimensional subspaces and manifolds, rank deficient matrices, objective functions with multiple solutions in a linear program.

Main result

Theorem.

Assume $p(p+1)/2 \geq m$, further the search space of (SDP) is **compact**, and the search space of (BM) is a **regularly defined smooth manifold**. Then for **almost all** C we have that

$$Y \text{ is a } \textbf{second-order critical point} \text{ of (BM)}$$
$$\implies Y \text{ is a } \textbf{global solution} \text{ of (SDP)}$$

How about algorithms?

There exists algorithms to do optimization on Manifolds.

MANOPT!

Proposition.

Under the aforementioned assumptions, the Riemannian-trust region method (RTR) initialized with any point $Y_0 \in \mathcal{M}$, returns in $O(1/\varepsilon_g^2 \varepsilon_H + 1/\varepsilon_H^3)$ iterations a point $Y \in \mathcal{M}$ satisfying

$$f(Y) \leq f(Y_0), \qquad \|\text{grad } f(Y)\| \leq \varepsilon_g$$

$$\text{Hess } f(Y) \succeq -\varepsilon_H I.$$

Quantitative result

Theorem.

Let $R < \infty$ be the maximal trace of any feasible X for (SDP). For any p such that $\mathcal{M}_p$ and $\mathcal{M}_{p+1}$ are smooth manifolds (no more assumptions are needed) and for any $Y \in \mathcal{M}_p$, form

$$\tilde{Y} = [Y | 0_{n \times 1}] \in \mathcal{M}_{p+1}$$

The optimality gap is bounded as

$$0 \leq 2\,(f(Y) - f^*) \leq \sqrt{R}\|\mathrm{grad}\ f(Y)\| - R\lambda_{\min}(\mathrm{Hess}\ f(Y)).$$

Example 1: Fixed trace/Sphere

Consider the **SDP** problem

$$\min_X \langle C, X \rangle \qquad \text{subject to} \qquad \text{trace}(X) = 1, X \succeq 0;$$

Then **Burer-Monteiro** gives

$$\min_{Y \in \mathbb{R}^{n \times p}} \langle CY, Y \rangle \qquad \text{subject to} \qquad \|Y\| = 1.$$

Then the solution is just the eigenvector associated to the smaller singular value!

When do we have higher rank solutions?

Example 2:
Fixed diagonal/Product of spheres

Consider the **SDP** problem

$$\min_X \langle C, X \rangle \qquad \text{subject to} \qquad \operatorname{diag}(X) = \mathbf{1},\ X \succeq 0.$$

Then **Burer-Monteiro** gives

$$\min_{Y \in \mathbb{R}^{n \times p}} \langle CY, Y \rangle \qquad \text{subject to} \qquad Y^\top = [y_1, \dots, y_n], \|y_i\| = 1.$$

Applications. Max-Cut, $\mathbb{Z}_2$-synchronization, community detection, and phase retrieval, among others.

Example 3:
Fixed blocks/Product of Stiefel

Consider the **SDP** problem

$$\min_X \langle C, X \rangle \qquad \text{subject to} \qquad X_{ii} = I_d,\ X \succeq 0.$$

Then **Burer-Monteiro** gives

$$\min_{Y \in \mathbb{R}^{qd \times p}} \langle CY, Y \rangle \qquad \text{subject to} \qquad Y^\top = [Y_1, \ldots, Y_q],\ Y_i{}^\top Y_i = I_d.$$

Applications. Orthogonal-cut, synchronization problems with the rotation and permutation groups.

Sketch of the proof

The proof consists of two big steps:

Step 1.

Proposition. If Y is a column rank-deficient second-order critical point for (BM), then it is optimal for (BM) and $YY^\top$ is optimal for (SDP)

Step 2.

Lemma. Under the assumptions of the main theorem. For almost all C, all the critical points of (BM) are rank deficient.

Conclusions and questions

- Nonconvexity is not terrible in some scenarios.

- We saw that under the right assumptions on the dimension and search spaces, it doesn't hurt to consider a nonconvex formulation.

- BM gives you a access to faster, storage-efficient algorithm.

- **Open question:** Is RTR the best we can do? $O(1/\varepsilon_g^2 \varepsilon_H + 1/\varepsilon_H^3)$ sounds bad.

- **Open question:** is there a bigger class of SDPs for which BM works?

Thank you!

Questions?