Rank optimality for the Burer-Monteiro factorization

Irène Waldspurger

CNRS and CEREMADE (Université Paris Dauphine)
Équipe MOKAPLAN (INRIA)

Joint work with Alden Waters (Bernoulli Institute, Rijksuniversiteit Groningen)

April 3, 2019

Imaging and machine learning
The mathematics of imaging semester
Institut Henri Poincaré

Semidefinite programming

$$\begin{aligned} \text{minimize } & \mathrm{Trace}(CX) \\ \text{such that } & \mathcal{A}(X) = b, \\ & X \succeq 0. \end{aligned}$$

Here,

- X, the unknown, is an $n \times n$ matrix ;
- C is a fixed $n \times n$ matrix (cost matrix) ;
- $\mathcal{A} : \mathrm{Sym}_n \to \mathbb{R}^m$ is linear ;
- b is a fixed vector in $\mathbb{R}^m$.

Motivations

Various difficult problems can be "lifted" to SDPs, and solving these lifted SDPs may solve the original problems.

Particularly important example : relaxation of *MaxCut*.

$$\begin{aligned} \text{minimize } & \mathrm{Trace}(CX) \\ \text{such that } & \mathrm{diag}(X) = 1, \\ & X \succeq 0. \end{aligned}$$

Relaxes the *Maximum Cut* problem from graph theory.
[Delorme and Poljak, 1993]
Appears also in phase retrieval, $\mathbb{Z}_2$ synchronization ...

Numerical solvers

SDPs can be solved at a given precision in polynomial time.
But the order of the polynomial may be large.

Interior point solvers, for instance, have a per iteration complexity of $O(n^4)$ in full generality (when m and n are of the same order).

First-order ones, applied to a smoothed problem, have a $O(n^3)$ complexity, but require more iterations.

$\rightarrow$ Numerically, high dimensional SDPs are difficult to solve.

Exploiting the low rank

To speed up these algorithms : exploit the structure of the problem.

Here, the “structure” we consider is the fact that there exists a low-rank solution.

- There is always a solution with rank r_{opt} at most

$$\left\lfloor \sqrt{2m + 1/4} - 1/2 \right\rfloor.$$

  [Pataki, 1998]
- In many situations, there is actually a solution with rank $r_{opt} = O(1)$.

Burer-Monteiro factorization

We focus on one heuristic that takes advantage of the low rank : the Burer-Monteiro factorization.
[Burer and Monteiro, 2003]

If there is a solution with rank r_{opt}, we can write X under the form

$$X = VV^T,$$

with V an $n \times p$ matrix, and $p \geq r_{opt}$.

$\rightarrow$ We optimize over V instead of optimizing over X.

$$\text{minimize } \mathrm{Trace}(CX)$$
$$\text{for } X \in \mathbb{R}^{n \times n} \text{ such that } \mathcal{A}(X) = b,$$
$$X \succeq 0.$$

$$\Updownarrow$$

$$\text{minimize } \mathrm{Trace}(CVV^T)$$
$$\text{for } V \in \mathbb{R}^{n \times p} \text{ such that } \mathcal{A}(VV^T) = b.$$

Remark : The factorization rank p must be chosen. It can be different from r_{opt}, the rank of the solution.

$$\text{minimize } \mathrm{Trace}(CVV^T)$$
$$\text{for } V \in \mathbb{R}^{n \times p} \text{ such that } \mathcal{A}(VV^T) = b.$$

We assume that $\{V \in \mathbb{R}^{n \times p}, \mathcal{A}(VV^T) = b\}$ is a "nice" manifold.
$\rightarrow$ Riemannian optimization algorithms.

Main advantage of the factorized formulation

The number of variables is not $O(n^2)$ anymore, but $O(np)$, with possibly $p \ll n$.
$\rightarrow$ Less computationally-demanding algorithms can be used.

$$\text{minimize } \mathrm{Trace}(CVV^T)$$
$$\text{for } V \in \mathbb{R}^{n \times p} \text{ such that } \mathcal{A}(VV^T) = b.$$

Main drawback of the factorized formulation

Contrarily to the SDP, this problem is non-convex.
$\rightarrow$ Riemannian optimization algorithms may get stuck at a critical point instead of finding a global minimizer.

This issue can arise or not, depending on the factorization rank p.
$\Rightarrow$ How to choose p ?

Outline

1. Literature review
   - In practice, algorithms work when $p = O(r_{opt})$.
   - In particular situations, this phenomenon is understood.
   - In a general setting, no guarantees for $p \lesssim \sqrt{2m}$.
   - Why this gap ?

Outline

1. Literature review
   - In practice, algorithms work when $p = O(r_{opt})$.
   - In particular situations, this phenomenon is understood.
   - In a general setting, no guarantees for $p \lesssim \sqrt{2m}$.
   - Why this gap ?
2. Optimal rank for the Burer-Monteiro formulation
   - Up to a minor improvement, $p \approx \sqrt{2m}$ is the optimal rank for which general guarantees can be derived.
   - Consequently, when $p \lesssim \sqrt{2m}$, Riemannian optimization algorithms cannot be certified correct without assumptions on C.

Outline

1. Literature review
   - In practice, algorithms work when $p = O(r_{opt})$.
   - In particular situations, this phenomenon is understood.
   - In a general setting, no guarantees for $p \lesssim \sqrt{2m}$.
   - Why this gap ?
2. Optimal rank for the Burer-Monteiro formulation
   - Up to a minor improvement, $p \approx \sqrt{2m}$ is the optimal rank for which general guarantees can be derived.
   - Consequently, when $p \lesssim \sqrt{2m}$, Riemannian optimization algorithms cannot be certified correct without assumptions on C.
3. Open questions

Empirical observations

1. [Burer and Monteiro, 2003]
   Numerical experiments on various problems, notably MaxCut and minimum bisection relaxations.
   The factorization rank is $p \approx \sqrt{2m}$, and algorithms always find a global minimizer.
   (The authors do not test smaller values of p.)

2. [Journée, Bach, Absil, and Sepulchre, 2010]
   Numerical experiments on MaxCut relaxations (with a particular initialization scheme).
   The algorithm proposed by the authors always finds a global minimizer when $p = r_{opt}$.

Empirical observations (continued)

3. [Boumal, 2015]
   Numerical experiments on problems coming from orthogonal synchronization.
   Here, $r_{opt} = 3$ and the algorithm finds the global minimizer as soon as $p \geq 5$.

4. Similar results on "SDP-like" problems.
   See for example [Mishra, Meyer, Bonnabel, and Sepulchre, 2014].

Theoretical explanations in particular cases

[Bandeira, Boumal, and Voroninski, 2016]
SDP instances coming from $\mathbb{Z}_2$ synchronization and community detection problems, under specific statistical assumptions.
→ With high probability, $r_{opt} = 1$.
If $p = 2$, Riemannian algorithms find the global minimizer.

Other particular SDP-like problems have been studied.
→ Under strong assumptions, $p \geq r_{opt}$ is enough so that a global minimizer is found.
[Ge, Lee, and Ma, 2016] ...

General case : one main result [Boumal, Voroninski, and Bandeira, 2018]

$$\text{minimize } \mathrm{Trace}(CVV^T)$$
$$\text{for } V \in \mathbb{R}^{n\times p} \text{ such that } \mathcal{A}(VV^T) = b.$$

Main hypothesis (approximately)

$\mathcal{M}_p \overset{déf}{=} \{V \in \mathbb{R}^{n\times p}, \mathcal{A}(VV^T) = b\}$ is a manifold.

General case : one main result [Boumal, Voroninski, and Bandeira, 2018]

$$\text{minimize } \mathrm{Trace}(CVV^T)$$
$$\text{for } V \in \mathbb{R}^{n\times p} \text{ such that } \mathcal{A}(VV^T) = b.$$

Main hypothesis (approximately)

$\mathcal{M}_p \overset{déf}{=} \{V \in \mathbb{R}^{n\times p}, \mathcal{A}(VV^T) = b\}$ is a manifold.

[More precisely : for all $V \in \mathcal{M}_p$,

$$\phi_V : \dot{V} \in \mathbb{R}^{n\times p} \to \mathcal{A}(V\dot{V}^T + \dot{V}V^T) \in \mathbb{R}^m$$

is surjective.]

General case : one main result [Boumal, Voroninski, and Bandeira, 2018]

$$\text{minimize } \mathrm{Trace}(CVV^T), \quad \text{for } V \in \mathcal{M}_p.$$

Riemannian optimization algorithms typically converge to second-order critical points :

A matrix $V_0 \in \mathcal{M}_p$ is a second-order critical point if

- $\nabla f_C(V_0) = 0_{n,p}$;
- $\mathrm{Hess}\, f_C(V_0) \succeq 0$,

where $f_C \overset{déf}{=} \left(V \in \mathcal{M}_p \to \mathrm{Trace}(CVV^T)\right)$.

General case : one main result [Boumal, Voroninski, and Bandeira, 2018]

Theorem

Under suitable hypotheses, for almost all matrices C, if

$$p > \left\lfloor \sqrt{2m + \frac{1}{4}} - \frac{1}{2} \right\rfloor ,$$

all second-order critical points of the factorized problem are global minimizers.
Consequently, Riemannian optimization algorithms always find a global minimizer.

General case : one main result [Boumal, Voroninski, and Bandeira, 2018]

Theorem

Under suitable hypotheses, for almost all matrices C, if

$$p > \left\lfloor \sqrt{2m + \frac{1}{4}} - \frac{1}{2} \right\rfloor ,$$

all second-order critical points of the factorized problem are global minimizers.
Consequently, Riemannian optimization algorithms always find a global minimizer.

Remark : The value of p does not depend on r_{opt}.

Summary

- In empirical experiments, as well as in the few particular cases that have been studied, algorithms seem to always work when

$$p = O(r_{opt}).$$

- The only available general result guarantees that algorithms work when

$$p \gtrsim \sqrt{2m}.$$

Summary

- In empirical experiments, as well as in the few particular cases that have been studied, algorithms seem to always work when
$$p = O(r_{opt}).$$
- The only available general result guarantees that algorithms work when
$$p \gtrsim \sqrt{2m}.$$

As r_{opt} is often much smaller than $\sqrt{2m}$, this leaves a big gap.

$\rightarrow$ Is it possible to obtain general guarantees for $p \ll \sqrt{2m}$?

Overview of our results

- A minor improvement is possible over the result by [Boumal, Voroninski, and Bandeira, 2018], but it does not change the leading order term

$$p \gtrsim \sqrt{2m}.$$

Overview of our results

- A minor improvement is possible over the result by [Boumal, Voroninski, and Bandeira, 2018], but it does not change the leading order term

$$p \gtrsim \sqrt{2m}.$$

- With this improvement, the result is essentially optimal, even if $r_{opt} \ll \sqrt{2m}$.

Improving [Boumal, Voroninski, and Bandeira, 2018]

Theorem

Under suitable hypotheses, for almost all matrices C, if

$$p > \left\lfloor \sqrt{2m + \frac{9}{4}} - \frac{3}{2} \right\rfloor ,$$

all second-order critical points of the factorized problem are global minimizers.

In [Boumal, Voroninski, and Bandeira, 2018], we had $\left\lfloor \sqrt{2m + \frac{1}{4}} - \frac{1}{2} \right\rfloor$. Our result is better by one unit for most values of m.

Theorem (Quasi-optimality of the previous result)

Let $r_0 = \min\{\mathrm{rank}(X), \mathcal{A}(X) = b, X \succeq 0\}$.
Under suitable hypotheses, if

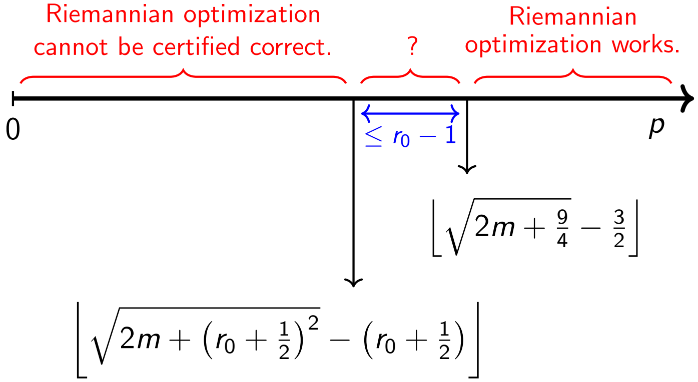
$$p \leq \left\lfloor \sqrt{2m + \left(r_0 + \frac{1}{2}\right)^2} - \left(r_0 + \frac{1}{2}\right) \right\rfloor ,$$

then there exists a set of matrices C with non-zero Lebesgue measure such that :

1. The global minimizer has rank r_0.
2. There is a second order critical point that is not a global minimizer.

Comments

- In most applications, r_0 is small, possibly $r_0 = 1$.
- We have the following picture :



Technical comment : "under suitable hypotheses"

There must exist $U_0 \in \mathbb{R}^{n \times r_0}$, $V \in \mathbb{R}^{n \times p}$ such that

$$\mathcal{A}(U_0 U_0^T) = \mathcal{A}(VV^T) = b,$$

and

$$\psi_V : (T, R) \in \mathrm{Sym}_p \times \mathbb{R}^{r_0 \times p} \to \mathcal{A}\left(\begin{pmatrix} V & U_0 \end{pmatrix} \begin{pmatrix} T \\ R \end{pmatrix} V^T + V \begin{pmatrix} T \\ R \end{pmatrix}^T \begin{pmatrix} V & U_0 \end{pmatrix}^T \right) \in \mathbb{R}^m$$

is injective.

Technical comment : "under suitable hypotheses"

There must exist $U_0 \in \mathbb{R}^{n \times r_0}$, $V \in \mathbb{R}^{n \times p}$ such that

$$\mathcal{A}(U_0 U_0^T) = \mathcal{A}(VV^T) = b,$$

and

$$\psi_V : (T, R) \in \mathrm{Sym}_p \times \mathbb{R}^{r_0 \times p} \rightarrow \mathcal{A}\left(\begin{pmatrix} V & U_0 \end{pmatrix} \begin{pmatrix} T \\ R \end{pmatrix} V^T + V \begin{pmatrix} T \\ R \end{pmatrix}^T \begin{pmatrix} V & U_0 \end{pmatrix}^T \right) \in \mathbb{R}^m$$

is injective.

Because $\dim\left(\mathrm{Sym}_p \times \mathbb{R}^{r_0 \times p}\right) \leq \dim(\mathbb{R}^m)$, this condition is a priori generically satisfied.

Example : MaxCut relaxations

$$\begin{aligned} \text{minimize } & \mathrm{Trace}(CX), \\ \text{such that } & \mathrm{diag}(X) = 1, \\ & X \succeq 0. \end{aligned}$$

(Original SDP)

$$\Downarrow$$

$$\begin{aligned} \text{minimize } & \mathrm{Trace}(CVV^T), \\ \text{such that } & \mathrm{diag}(VV^T) = 1, V \in \mathbb{R}^{n \times p}. \end{aligned}$$

(Burer-Monteiro factorization)

- In this case, $r_0 = 1$.
- The “suitable hypotheses” are satisfied.

Example : MaxCut relaxations

- For almost all C, if

$$p > \left\lfloor \sqrt{2m + \frac{9}{4}} - \frac{3}{2} \right\rfloor,$$

  no bad second-order critical point exists : Riemannian optimization algorithms work.

- If

$$p \leq \left\lfloor \sqrt{2m + \frac{9}{4}} - \frac{3}{2} \right\rfloor,$$

  bad second-order critical points may exist, even when there is a rank 1 solution : Riemannian algorithms cannot be certified correct without additional assumptions on C.

Burer-Monteiro factorization : summary

- [Literature]
  In particular cases, with strong statistical assumptions on C, the Burer-Monteiro factorization works as soon as

$$p = r_{opt} \text{ or } p = r_{opt} + 1.$$

- [Our result]
  There are matrices C for which it can fail, unless

$$p \gtrsim \sqrt{2m},$$

even if $r_{opt} = O(1)$.

Burer-Monteiro factorization : summary

- [Literature]
  In particular cases, with strong statistical assumptions on C, the Burer-Monteiro factorization works as soon as
  $$p = r_{opt} \text{ or } p = r_{opt} + 1.$$
- [Our result]
  There are matrices C for which it can fail, unless
  $$p \gtrsim \sqrt{2m},$$
  even if $r_{opt} = O(1)$.
- [Empirically]
  The Burer-Monteiro factorization usually works for
  $$p = O(r_{opt}).$$

→ Apparently, the matrices we have constructed for which the Burer-Monteiro factorization admits bad second-order critical points are somewhat pathological, and not encountered in practice.

Questions

Questions

- Compute the volume, in the space of cost matrices, of matrices for which bad second-order critical points exist, as a function of n and p ?

Questions

- Compute the volume, in the space of cost matrices, of matrices for which bad second-order critical points exist, as a function of n and p?
- Develop guarantees for the Burer-Monteiro factorization with assumptions on C, but only mild ones?

  [Intermediate between very specific settings, for which we have strong guarantees, and the general case, where guarantees are only for $p \gtrsim \sqrt{2m}$.]

Thank you !

I. Waldspurger and A. Waters (2018). Rank optimality for the Burer-Monteiro factorization. arXiv preprint arXiv :1812.03046.