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 22, 2020
PDE and inverse problem methods in machine learning
Workshop at IPAM, UCLA

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 $\mathbb{R}^m$.

Low-rank semidefinite programming

We assume that there is an optimal solution with low rank $r \ll n$.

Motivations

Various inverse problems naturally have an approximate low-rank structure.

This notably appears when "lifting" to the matricial space inverse problems with quadratic measurements (e.g. phase retrieval).

Numerical solvers

General SDPs can be solved at arbitrary precision in polynomial time.
But the order of the polynomial is 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 : assume that there exists a low-rank solution and exploit this fact.

- [Pataki, 1998] : There is always a solution with rank

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

- In many situations, there is actually a solution with rank

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

Exploiting the low rank

Two main strategies :

- Frank-Wolfe methods ;
  [Frank and Wolfe, 1956]
- Burer-Monteiro factorization.
  [Burer and Monteiro, 2003]

Burer-Monteiro factorization

- Assume that there is a solution with rank r_{opt}.
- Choose some integer $p \geq r_{opt}$.
- Write X under the form

$$X = VV^T,$$

  with V an $n \times p$ matrix.
- Minimize $\mathrm{Trace}(CVV^T)$ over V.

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

$$\Updownarrow$$

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

Remark : p is the *factorization rank*. It must be chosen, and can be equal to or larger than r_{opt}.

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

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 $p \ll n$.
$\rightarrow$ Riemannian algorithms can be much faster than SDP solvers.

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

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 unless $p \gtrsim \sqrt{2m}$.
   - But $r_{opt} \ll \sqrt{2m}$. 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 unless $p \gtrsim \sqrt{2m}$.
   - But $r_{opt} \ll \sqrt{2m}$. Why this gap ?
2. Optimal rank for the Burer-Monteiro formulation
   - A minor improvement is possible over previous general guarantees.
   - The improved result is optimal.
     $\to$ If $p \lesssim \sqrt{2m}$, Riemannian algorithms cannot be certified correct without assumptions on C.
   - Idea of proof.

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 unless $p \gtrsim \sqrt{2m}$.
   - But $r_{opt} \ll \sqrt{2m}$. Why this gap ?
2. Optimal rank for the Burer-Monteiro formulation
   - A minor improvement is possible over previous general guarantees.
   - The improved result is optimal.
     - $\rightarrow$ If $p \lesssim \sqrt{2m}$, Riemannian algorithms cannot be certified correct without assumptions on C.
   - Idea of proof.
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 $p \approx \sqrt{2m}$; Riemannian 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.
$\rightarrow$ With high probability, $r_{opt} = 1$.
If $p = 2$, Riemannian algorithms find the global minimizer.

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

Strong guarantees, but in very specific situations only.

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

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

The only assumption is (approximately) that

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

is a manifold.

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

$$\boxed{\begin{array}{c}\text{minimize } \operatorname{Trace}(CVV^T),\\ \text{for } V \in \mathcal{M}_p.\end{array}}$$

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

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

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

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

Theorem

For almost all matrices C, if

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

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

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

Theorem

For almost all matrices C, if

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

all second-order critical points 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.

→ 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

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

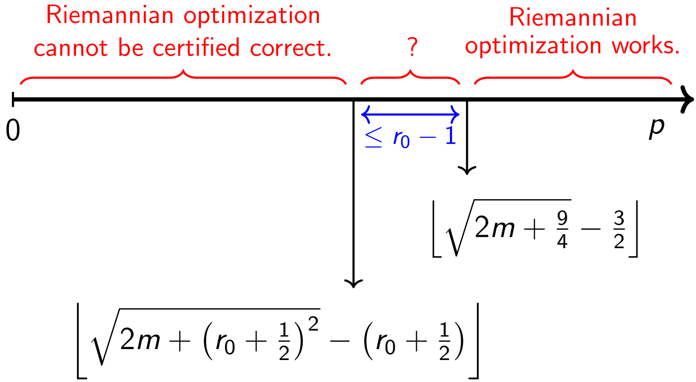
$$p \leq \left\lfloor \sqrt{2m + \left(r_0 + \frac{1}{2}\right)^2} - \left(r_0 + \frac{1}{2}\right) \right\rfloor ,$$

there is a set of matrices C with non-zero Lebesgue measure for which :

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

Comments

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



Example : MaxCut relaxations

Relaxes the *Maximum Cut* problem from graph theory.
[Delorme and Poljak, 1993]

Also appears as a convex relaxation of phase retrieval or $\mathbb{Z}_2$ synchronization problems.

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

Example : MaxCut relaxations

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

(Original SDP)

$$\Downarrow$$

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

(Burer-Monteiro factorization)

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

Example : MaxCut relaxations

- For almost all C, if

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

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

- If

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

  even when assuming a rank 1 solution, there are matrices C for which Riemannian algorithms can fail.

Idea of proof

We consider

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

We want to construct a set of matrices C with non-zero Lebesgue measure for which :

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

Idea of proof

Step 1

Construct one such matrix C.

Step 2

Show that, in a ball around C, all matrices satisfy these properties.
$\rightarrow$ Classical geometrical arguments
(implicit function theorem).

Idea of proof

Step 1

Construct one such matrix C.

Step 2

Show that, in a ball around C, all matrices satisfy these properties.
$\rightarrow$ Classical geometrical arguments
(implicit function theorem).

Idea of proof : construct a "bad" C

- Fix a feasible X_0 with rank r_0.
- Fix a feasible $V \in \mathcal{M}_p$.

Idea of proof : construct a "bad" C

- Fix a feasible X_0 with rank r_0.
- Fix a feasible $V \in \mathcal{M}_p$.
- Construct C such that
  - The SDP problem has X_0 as a unique global minimizer.
  - The factorized problem has V as a non-optimal second-order critical point.

It turns out that constructing such a C is possible for almost any X_0, V.

Idea of proof : construct a bad C

We want C such that

- X_0 is the unique global minimizer of the SDP ;
- V is a second-order critical point.

Using the analytical expressions of the gradient and Hessian, we rewrite these properties under more explicit forms.

Idea of proof : construct a bad C

We want C such that

- X_0 is the unique global minimizer of the SDP ;
- V is a second-order critical point.

Using the analytical expressions of the gradient and Hessian, we rewrite these properties under more explicit forms.

After simplification, we see that it is possible to construct such a C as soon as there exists $\mu \in \mathbb{R}^m$ such that

$$\boxed{V^T \mathcal{A}^*(\mu) V \succeq 0 \qquad \text{and} \qquad X_0^T \mathcal{A}^*(\mu) V = 0.}$$

Idea of proof : construct a bad C

Does there exist μ such that

$$\boxed{V^T \mathcal{A}^*(\mu) V \succeq 0 \qquad \text{and} \qquad X_0^T \mathcal{A}^*(\mu) V = 0\,?}$$

Idea of proof : construct a bad C

Does there exist μ such that

$$\boxed{V^T \mathcal{A}^*(\mu) V \succeq 0 \qquad \text{and} \qquad X_0^T \mathcal{A}^*(\mu) V = 0\,?}$$

Consider the map

$$\begin{array}{ccccc} \overbrace{\mathbb{R}^m}^{\text{dimension } m} & \to & \overbrace{\mathrm{Sym}^{p\times p} \quad \times \quad \mathbb{R}^{r_0 \times p}}^{\text{dimension } \frac{p(p+1)}{2} + pr_0} \\ \mu & \to & (V^T \mathcal{A}^*(\mu) V, \quad X_0^T \mathcal{A}^*(\mu) V) \end{array}$$

Idea of proof : construct a bad C

Does there exist μ such that

$$\boxed{V^T \mathcal{A}^*(\mu) V \succeq 0 \qquad \text{and} \qquad X_0^T \mathcal{A}^*(\mu) V = 0\,?}$$

Consider the map

$$\begin{array}{ccccc} \overbrace{\mathbb{R}^m}^{\text{dimension } m} & \to & \overbrace{\mathrm{Sym}^{p\times p} \quad \times \quad \mathbb{R}^{r_0\times p}}^{\text{dimension } \frac{p(p+1)}{2} + pr_0} \\ \mu & \to & (V^T \mathcal{A}^*(\mu) V, \quad X_0^T \mathcal{A}^*(\mu) V) \end{array}$$

If $m \geq \frac{p(p+1)}{2} + pr_0$, it is generically surjective and μ exists.

Idea of proof : construct a bad C

Does there exist μ such that

$$\boxed{V^T\mathcal{A}^*(\mu)V \succeq 0 \qquad \text{and} \qquad X_0^T\mathcal{A}^*(\mu)V = 0\,?}$$

Consider the map

$$\begin{array}{ccc} \overbrace{\mathbb{R}^m}^{\text{dimension } m} & \to & \overbrace{\mathrm{Sym}^{p\times p} \quad \times \quad \mathbb{R}^{r_0\times p}}^{\text{dimension } \frac{p(p+1)}{2} + pr_0} \\ \mu & \to & (V^T\mathcal{A}^*(\mu)V, \quad X_0^T\mathcal{A}^*(\mu)V) \end{array}$$

If $m \geq \frac{p(p+1)}{2} + pr_0$, it is generically surjective and μ exists.

$$\iff \quad p \leq \sqrt{2m + \left(r_0 + \tfrac{1}{2}\right)^2} - \left(r_0 + \tfrac{1}{2}\right)$$

Burer-Monteiro factorization : summary

- [Boumal, Voroninski, and Bandeira, 2018]

  When $p \gtrsim \sqrt{2m}$, for almost any cost matrix, all second-order critical points are minimizers.

  Numerical experiments suggest it could be true for

$$p = O(r_{opt}) \ll \sqrt{2m}.$$

- [Our result]

  When $p \lesssim \sqrt{2m}$, it is not true.

Open questions

1. Refined understanding of the regime $p < \sqrt{2m}$
2. Application to phase retrieval problems

Understanding the regime $p < \sqrt{2m}$

Two types of theoretical guarantees exist for the Burer-Monteiro factorization :

- Specific problems and strong assumptions on C.
  $\rightarrow$ Works for $p = r_{opt}$ or $p = r_{opt} + 1$.

  "When C is very nice, it works for $p \approx r_{opt}$."

- No assumption on C.
  $\rightarrow$ Works for $p \gtrsim \sqrt{2m}$ and not below.
  [Our result]

  "When C is very bad, $p \gtrsim \sqrt{2m}$ is necessary."

Understanding the regime $p < \sqrt{2m}$

Can we have something in between, closer to realistic settings ?

"Under moderate assumptions on C, it works for $p = O(r_{opt})$" ?

or

"For most C, it works for $p = O(r_{opt})$" ?

Application to phase retrieval problems

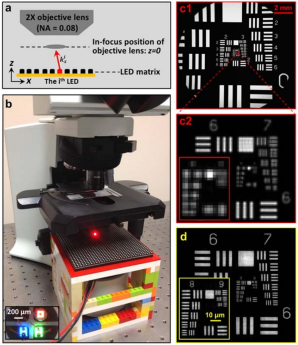
Reconstruct $x \in \mathbb{C}^d$ from $|\langle a_k, x\rangle|, 1 \leq k \leq m$.

Here,

- $a_1, \dots, a_m \in \mathbb{C}^d$ are known ;
- $|.|$ is the complex modulus.

Important applications in optics.

Phase retrieval algorithms based on convex relaxations usually offer good reconstruction quality, but are too slow.



Application to phase retrieval problems

Can we speed up the convex relaxations with Burer-Monteiro ?

- Which factorization rank ?
- Which solver ?

Thank you !

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