A Convergent Gradient Descent Algorithm for Rank Minimization and Semidefinite Programming from Random Linear Measurements

Qinqing Zheng and John Lafferty

THE UNIVERSITY OF CHICAGO

NIPS 2015, Montréal, Canada

Motivation

- Semidefinite programming is a key tool in applied mathematics, machine learning, etc. Current algorithms for SDPs do not scale to large problems. Gradient descent methods repeatedly shown to be highly effective for large scale machine learning problems. **Can first order algorithms be effective for SDPs?**
- Burer and Monterio (2003) propose general schemes for attacking SDPs with factored, nonconvex approaches, with some empirical support.
- Candès et al. (2015) develop a gradient descent procedure for *phase retrieval*, minimizing a nonconvex objective to recover complex vector from squared magnitudes of linear measurements.
- We show how similar ideas can work for rank minimization and solving certain SDPs.

Rank Minimization and SDP

$$\min_{X\in\mathbb{R}^{n\times p}} \ \operatorname{rank}(X) \quad \text{subject to } \mathcal{A}(X) = b$$

- Nonconvex and NP-hard in general
- Closely related to family of SDPs if X is semidefinite. With sufficient measurements,

$$\min \operatorname{rank}(X) \equiv \min \|X\|_* \equiv \operatorname{tr}(X).$$

Problem

Suppose $X^\star$ is semidefinite and of rank r. Let $b_i = \operatorname{tr}(A_i X^\star)$ where A_i is GOE symmetric matrix

$$A_{jk} \sim \begin{cases} \mathcal{N}(0,1) & j \neq k \\ \mathcal{N}(0,2) & j = k \end{cases}$$

Goal is to solve

$$\min_{X\succeq 0} \ \operatorname{rank}(X) \quad \text{subject to } \operatorname{tr}(A_i X) = b_i,\ i = 1,\dots,m$$

Approach

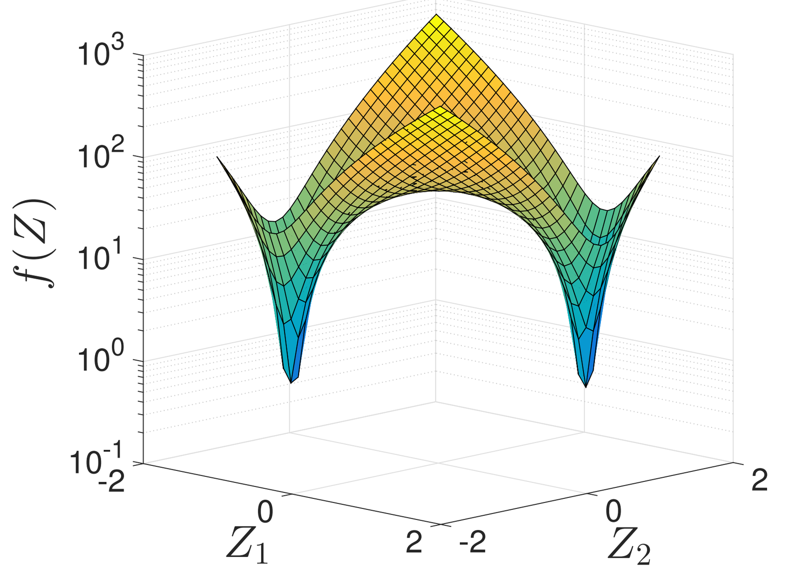
Writing $X = ZZ^\top$, attempt to minimize objective function

$$f(Z) = \frac{1}{4m}\sum_{i=1}^{m}\left(\operatorname{tr}(Z^\top A_i Z) - b_i\right)^2$$

Important property is

$$\mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m} b_i A_i\right) = 2X^\star$$

Initialize with spectral decomposition of $\frac{1}{2m}\sum_{i=1}^{m} b_i A_i$ and then apply gradient descent.



Example: $X^\star \in \mathbb{R}^{2\times 2}$ is rank-1 and $Z \in \mathbb{R}^2$. True vector is $Z^\star = [1,1]^\top$. Both $Z^\star$ and $-Z^\star$ are minimizers.

Algorithm

Input: $\{A_i, b_i\}_{i=1}^m, r, \mu$

Initialization

Let $(v_1,\lambda_1),\dots,(v_r,\lambda_r)$ to the top r eigenpairs of $\frac{1}{m}\sum_{i=1}^m b_i A_i$

$Z = [z_1,\dots,z_r]$ where $z_s = \sqrt{\frac{|\lambda_s|}{2}}\cdot v_s,\ s\in[r]$

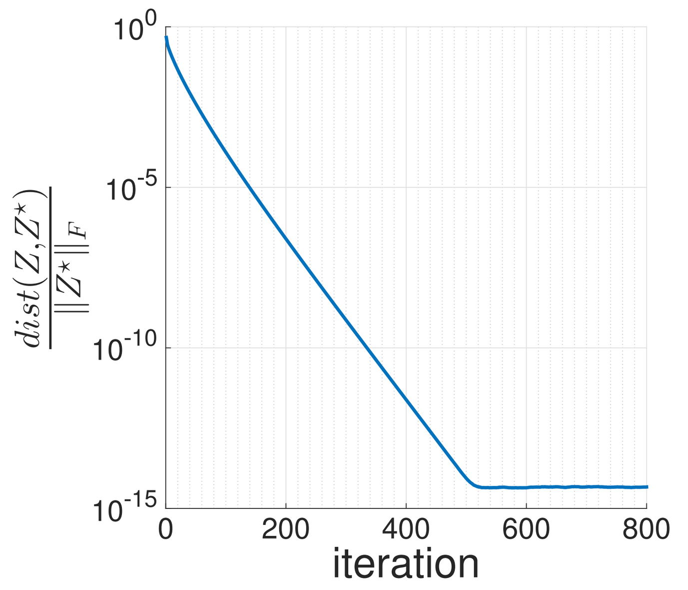
Repeat

$$\nabla f(Z) = \frac{1}{m}\sum_{i=1}^{m}\left(\operatorname{tr}(Z^\top A_i Z) - b_i\right) A_i Z$$

$$Z \leftarrow Z - \frac{\mu}{\sum_{s=1}^r |\lambda_s|/2}\nabla f(Z)$$

until convergence

Output: $\widehat{X} = ZZ^\top$



Linear convergence of the gradient scheme, where $X^\star \in \mathbb{R}^{200\times 200}$ has rank 2. The distance metric is given below.

Our results

Define the distance function

$$d(Z,Z^\star) = \min_{\text{orthogonal } U} \|Z - Z^\star U\|_F$$

Let $\kappa = \sigma_1/\sigma_r$ denote the condition number of $X^\star$. There exist universal constants c_0 and c_1 such that if $m \geq c_0\kappa^2 r^3 n\log n$, with high probability the initialization Z^0 satisfies

$$d(Z^0, Z^\star) \leq \sqrt{\frac{3}{16}}\sigma_r$$

Moreover, using constant step size $\mu/\|Z^\star\|_F^2$ with $\mu \leq \frac{c_1}{\kappa n}$, the kth iteration of the algorithm satisfies

$$d(Z^k, Z^\star) \leq \sqrt{\frac{3}{16}}\sigma_r\left(1 - \frac{\mu}{12\kappa r}\right)^{k/2}$$

with high probability.

Proof structure

We establish a *local regularity condition* similar to Nesterov's conditions:

$$\langle \nabla f(Z), Z - \widetilde{Z}\rangle \geq c_1'\left\|Z - \widetilde{Z}\right\|_F^2 + c_2'\|\nabla f(Z)\|_F^2.$$

To demonstrate this, we show that the objective f satisfies a *local curvature condition*

$$\langle \nabla f(Z), Z - \widetilde{Z}\rangle \geq C_1\left\|Z - \widetilde{Z}\right\|_F^2 + \left\|(Z - \widetilde{Z})^\top Z\right\|_F^2$$

and a *local smoothness condition*

$$\|\nabla f(Z)\|_F^2 \leq C_2\left\|Z - \widetilde{Z}\right\|_F^2 + C_3\left\|(Z - \widetilde{Z})^\top Z\right\|_F^2$$

where $\widetilde{Z} = \arg\min_{\text{solution } \widetilde{Z}}\left\|Z - \widetilde{Z}\right\|_F$.

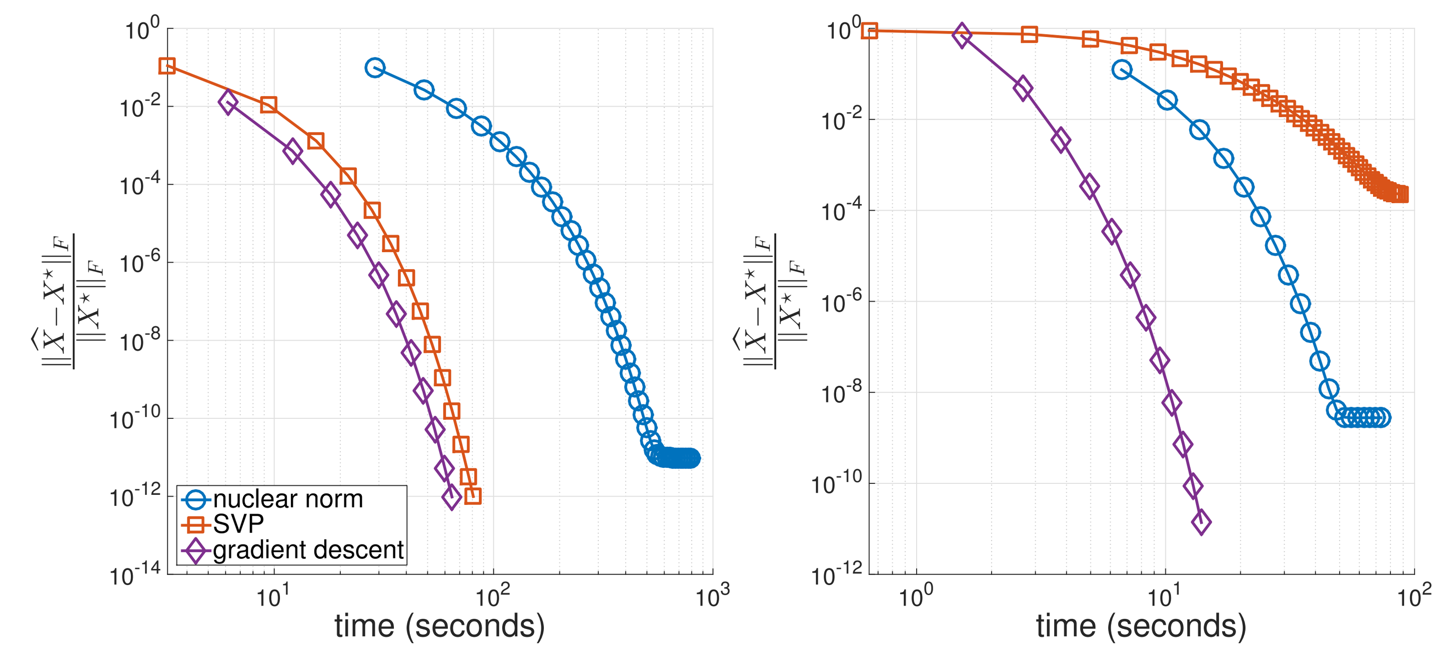
We exploit concentration around the mean of the Hessian $\nabla^2 f(Z)$ and matrices $\frac{1}{m}\sum_{i=1}^m (u^\top A_i u) A_i$.

Remark: We require $O(r^2 n\log n)$ samples for the regularity conditions to hold with high probability. For the initialization to be sufficiently close, we require $O(r^3 n\log n)$ samples. Independent work of Tu et al. (2015) improves this to $O(r^2 n)$ overall.

Simulation

We compare against the Singular Value Projection algorithm (SVP) of Jain et al. (2010) and nuclear norm relaxation of Recht et al. (2009).

- Runtime:

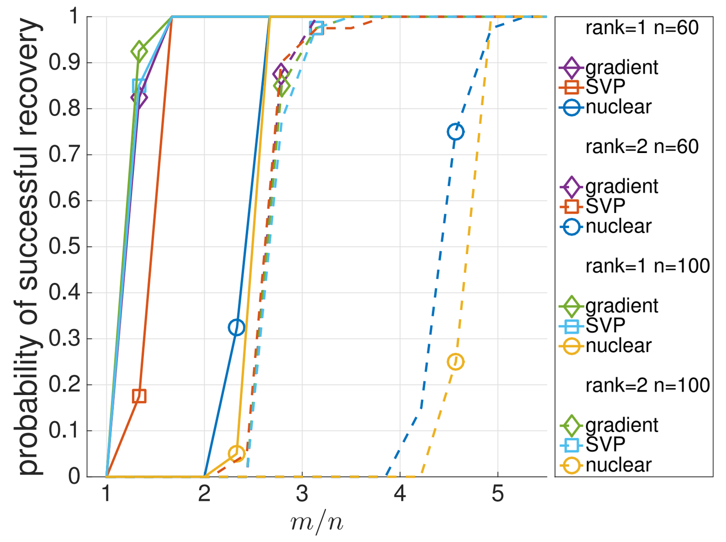


Left: 400x400 random rank-2 $X^\star$, $m = 6n$, dense A_i. Right: 600x600 random rank-2 $X^\star$, $m = 7n$, sparse A_i.

Let ρ denote the density of A_i. We summarize the per-iteration complexity:

Method	Complexity
nuclear norm (ADMM)	$O(mn^2\rho + m^2 + n^3)$
gradient descent	$O(mn^2\rho) + 2n^2 r$
SVP	$O(mn^2\rho + n^2 r)$

- Sample complexity:



We conjecture the sample complexity bound could be further improved to be $O(rn)$.

Future directions

- Many possibilities for realizing potential of factored gradient descent approaches to SDPs. Such techniques may be effective for a much wider class of SDPs.
- Explore theory for sparse or structured sensing matrices, non-random designs.
- Lower and optimal $O(nr)$ complexity.
- Purely first order algorithms (no SVDs).

Acknowledgments: Supported by NSF grants IIS-1116730, DMS-1513594, and ONR grant N00014-12-1-0762

Contacts: Qinqing Zheng (qinqing@cs.uchicago.edu) John Lafferty (lafferty@galton.uchicago.edu)