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

Qinqing Zheng and John Lafferty

NiPi 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 Monteiro (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 Nesterov's conditions:
- Burer and Monteiro (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.

Rank Minimization and SDP

Problem
Suppose $X$ is semidefinite and of rank $r$. Let $A_k = \tr(A_k^2)$ where $A_k$ is GOE symmetric matrix

\[
A_k = \begin{bmatrix} 0 & 1 \\ 0 & k \end{bmatrix}
\]

Goal is to solve
\[
\min_{X} \tr(X) \text{ subject to } \tr(A_kX) = b, \quad k = 1, \ldots, m
\]

Approach
Writing $X = Z^2$, attempt to minimize objective function

\[
\min_{Z} \tr(Z) = \min_{Z} \sum_{i,j} \tr(Z_{i,j}^2) - b_i^2
\]

Important property is

\[
\sum_{i,j} \tr(Z_{i,j}^2) = \tr(X^2) = 2 \tr(A_kX) - b_k^2 = 2 \tr(A_kX^2) - b_k^2
\]

Initialize with spectral decomposition of $A_k$, $b_k$, and then apply gradient descent.

Example: $X \in \mathbb{R}^{3 \times 3}$ is rank-1 and $Z \in \mathbb{R}^{3 \times 3}$. True vector is $Z = [1, 0, 0]^\top$. Both $Z$ and $-Z$ are minimizers.

Algorithm
Input: $(A_k, b_k, \mu, \rho, \nu)$

Initialization
Let $m$ $(m \leq n)$ and $(x_k, y_k)$ be the top $r$ eigenvectors of $\sum_{i=1}^{m} A_k$

\[
Z = [x_k^\top, y_k^\top, \ldots, x_m^\top, y_m^\top]
\]

Repeat

\[
\begin{align*}
\gamma & \leftarrow \min_{\gamma} \| Z - Z^\gamma \|_F \\
p & \leftarrow \min_{\rho} \gamma \\
Z & \leftarrow Z^\gamma
\end{align*}
\]

until convergence

Output: $X = Z^2$

Our results
Define the distance function

\[
\rho(Z, Z^\gamma) = \min_{\gamma} \| Z - Z^\gamma \|_F
\]

Let $\gamma = \rho(Z, Z^\gamma)$ denote the condition number of $X^\gamma$. There exist universal constants $c_1$ and $c_2$ such that if $\gamma \leq c_1 \rho(Z, Z^\gamma)$, with high probability the initialization $Z^\gamma$ satisfies

\[
\rho(Z, Z^\gamma) \leq c_2 \gamma
\]

Moreover, using constant step size $\gamma = \frac{1}{k}$ with $\gamma < \frac{1}{k^2}$, the $k$th iteration of the algorithm satisfies

\[
\rho(Z, Z^\gamma) \leq \frac{1}{(1 - 2\gamma)^{2k}}
\]

with high probability.

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

\[
\forall Z, \exists Z^\gamma \text{ such that } Z - Z^\gamma \leq \gamma \text{ and } \rho(Z, Z^\gamma) \leq c_1 \gamma
\]

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

\[
\forall Z, \exists Z^\gamma \text{ such that } Z - Z^\gamma \leq \gamma \text{ and } \rho(Z, Z^\gamma) \leq c_1 \gamma
\]

and a local smoothness condition

\[
\forall Z, \exists Z^\gamma \text{ such that } Z - Z^\gamma \leq \gamma \text{ and } \rho(Z, Z^\gamma) \leq c_1 \gamma
\]

where $Z = \arg\min_{\rho(Z, Z^\gamma)} \| Z - Z^\gamma \|_F$.

We exploit concentration around the mean of the Hessian $\nabla^2 f(Z)$ and matrices $Z - Z^\gamma \equiv (A_k - A_k^\gamma)A_k$.

Remark: We require $O(\log n)$ samples for the regularity conditions to hold with high probability. For the initialization to be sufficiently close, we require $O(\log \log n)$ samples. Independent work of Su et al. (2015) improves this to $O(\log n)$ 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: 400/400 random rank-2 $X$, $m = 6n$, dense A; Right: 600/600 random rank-2 $X$, $m = 7n$, sparse A.

Let $\rho$ denote the density of $A_i$. We summarize the per-iteration complexities:

\[
\begin{align*}
\text{Method} & \quad \text{Complexity} \\
\text{nuclear norm (ADMM)} & \quad O(m) \\
\text{gradient descent} & \quad O(m^2) \\
\text{SVP} & \quad O(m^3)
\end{align*}
\]

• Sample complexity:

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

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 $O(\rho)$ complexity.
• Purely first order algorithms (no SVGDs).

Acknowledgments: Supported by NSF grants 15-1116073, DAMO-151-15842, and ONR grant N00014-12-1-0762.

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