When Nonconvex Optimization Meets Big Data

Ju Sun
[Advisor: Prof. John Wright]

November 21, 2014
A curious experiment

Try to learn a concise approximation:

\[ Y \approx QX, \text{ with } Q \in O(n)^{n \times p} \text{ and } X \text{ sparse}. \]

... by solving

\[
\min_{Q \in O(n)} \left\{ \frac{1}{2} \|Y - QX\|_F^2 + \lambda \|X\|_1, \text{ s.t. } Q \in O(n) \right\}.
\]
A curious experiment

Try to learn a concise approximation: $Y \approx QX$, with $Q \in O(n)$ and $X$ sparse.
A curious experiment

Try to learn a **concise approximation**: \( Y \approx QX \), with \( Q \in O(n) \) and \( X \) sparse.

... by solving \( \min \frac{1}{2} \| Y - QX \|_F^2 + \lambda \| X \|_1 \), s.t. \( Q \in O(n) \).
A curious experiment

\[
\min f(Q, X) = \frac{1}{2} \|Y - QX\|^2_F + \lambda \|X\|_1, \quad \text{s.t. } Q \in O(n).
\]

- Objective is **nonconvex**: \((Q, X) \mapsto QX\) is bilinear
- Orthogonal group \(O(n)\) is a **nonconvex** set
- Combinatorially many isolated global minima: \((Q, X)\) or \((Q\Pi, \Pi^*X)\) (\(2^n n!\) many signed permutations \(\Pi\))
A curious experiment

An image

Patches

\[
\begin{aligned}
\min f (Q, X) &= \frac{1}{2} \| Y - QX \|_F^2 + \lambda \| X \|_1, \quad \text{s.t. } Q \in O(n)
\end{aligned}
\]

Apply the naive **alternating directions**: starting from a random \(Q_0 \in O(n)\)

\[
X_k = \arg \min_X f (Q_{k-1}, X)
\]

\[
Q_k = \arg \min_Q f (Q, X_k), \quad \text{s.t. } Q \in O(n).
\]
A curious experiment

\[ f(Q, X) = \frac{1}{2} \| Y - QX \|_F^2 + \lambda \| X \|_1, \quad \text{s.t. } Q \in O(n) \]

Apply the naive alternating directions: starting from a random \( Q_0 \in O(n) \)

\[ X_k = \arg \min_X f(Q_{k-1}, X) \]

\[ Q_k = \arg \min_Q f(Q, X_k), \quad \text{s.t. } Q \in O(n). \]
What is going on here?

You can find me and see thousands more!
Theme of this thesis

- Certain nonconvex optimization problems become tractable when the input data are large and random (generic).
- Geometry of the function landscape provides important clues for algorithm design and analysis.

... starting with \textit{sparse dictionary learning}!
Sparse dictionary learning


Important algorithmic contributions from many researchers: [Lewicki, Sejnowski.’99], [Engan et al. ’99], [Aharon, Elad, Bruckstein. ’06], many others

Widely used in image processing, vision, audio, and machine learning

\[ Y \approx QX \quad X \in \mathbb{R}^{n \times p} \quad \text{sparse} \]
Assume $Q$ is square and nonsingular, then $\text{row}(Y) = \text{row}(X)$.

When $p \geq \Omega(n \log n)$, rows of $X$ are the sparsest vectors in $\text{row}(Y)$ [Spielman, Wang, Wright. '12]

$$\min \|q^* Y\|_0 s.t. q \neq 0.$$ If we recover one row (up to scaling) of $X$, then we use deflation to find the others.

$Y \approx QX \quad X = \Omega \odot V, \quad \Omega \sim \text{Ber}(\theta), \quad V \sim \mathcal{N}(0, 1)$.
Assume $Q$ is square and nonsingular, then $\text{row}(Y) = \text{row}(X)$.

When $p \geq \Omega(n \log n)$, rows of $X$ are the sparsest vectors in $\text{row}(Y)$ [Spielman, Wang, Wright. '12]
Dictionary learning - the complete case

\[ Y \approx QX \quad X = \Omega \odot V, \quad \Omega \sim \text{Ber}(\theta), \quad V \sim \mathcal{N}(0, 1). \]

- Assume \( Q \) is square and nonsingular, then \( \text{row} (Y) = \text{row} (X) \).
- When \( p \geq \Omega(n \log n) \), rows of \( X \) are the sparsest vectors in \( \text{row} (Y) \) [Spielman, Wang, Wright. '12]

\[
\min \| q^* Y \|_0 \quad \text{s.t. } q \neq 0.
\]

if we recover one row (up to scaling) of \( X \), then we use deflation to find the others.
Dictionary learning: the complete case

\[
\min \| q^* Y \|_0 \quad \text{s.t. } q \neq 0.
\]

- Convex relaxation:

\[
\min \| q^* Y \|_1 \quad \text{s.t. } r^* q = 1
\]

Provably succeeds when \( \theta n = O(\sqrt{n}) \), provably fails if \( \theta n = \Omega(\sqrt{n \log n}) \) [Spielman, Wang, Wright.’12].

- Nonconvex relaxation:

Model problem

\[
\min \| q^* Y \|_1 \quad \text{s.t. } \| q \|_2^2 = 1.
\]

many precedents, e.g., [Zibulevsky-Perlmutter, ’01] in source separation.
The model problem

**Model problem**

\[
\min \quad \frac{1}{p} \|q^*Y\|_1 = \frac{1}{p} \sum_{i=1}^{p} |q^*y_i| \quad \text{s.t.} \quad \|q\|_2 = 1. \quad Y \in \mathbb{R}^{n \times p}
\]

- Convex objective function, but nonconvex constraint \( q \in S^{n-1}. \)
- If \( p \geq \Omega(n \log n) \), w.h.p. every global optimizer \( q^* \) produces \( q^*Y \) that recovers one row of \( X \) (up to scaling)
Towards geometric understanding

Model problem

\[
\min \quad \frac{1}{p} \|q^*Y\|_1 = \frac{1}{p} \sum_{i=1}^{p} |q^*y_i| \quad \text{s.t.} \quad \|q\|_2^2 = 1. \quad Y \in \mathbb{R}^{n \times p}
\]

Slightly modified model problem

\[
\min \quad \frac{1}{p} \sum_{i=1}^{p} h_{\mu}(q^*y_i) \quad \text{s.t.} \quad \|q\|_2^2 = 1. \quad Y \in \mathbb{R}^{n \times p}
\]

- Work with a smooth surrogate for \(|z|\):

\[
h_{\mu}(z) = \mu \log \left( \frac{e^{z/\mu} + e^{-z/\mu}}{2} \right)
\]

- Recognize the objective as a normalized sum of independent random variables → expectation, asymptotically
Why might this work?

W.l.o.g., assume for analysis that $Y = X$ (i.e., $A = I$); we correctly recover a row iff an algorithm produces $q_\diamond = \pm e_i$, $i = 1, \ldots, n$

$$f(q) = \frac{1}{p} \sum_{i=1}^{p} h_\mu (q^* x_i)$$

$$q(w) = \left( w_1, \ldots, w_{n-1}, \sqrt{1 - \|w\|_2^2} \right)$$

$$g(w) = f(g(w))$$

Target points are $w_\diamond = 0, \pm e_1, \ldots, \pm e_{n-1}$
W.l.o.g., assume for analysis that $Y = X$ (i.e., $A = I$); we correctly recover a row iff an algorithm produces $q_\diamond = \pm e_i$, $i = 1, \ldots, n$

Target points are $w_\diamond = 0, \pm e_1, \ldots, \pm e_{n-1}$
**Lemma:** Suppose $\theta \in \left[ \frac{1}{n}, \frac{1}{2} \right)$, and $\mu < cn^{-5/4}$. Then

1. $\nabla^2 \mathbb{E}[g(w)] \preceq \frac{c'\theta}{\mu} I$, for $\|w\| \leq \frac{\mu}{4\sqrt{2}}$
2. $\frac{w^* \nabla \mathbb{E}[g(w)]}{\|w\|} \geq c'\theta$, for $\frac{\mu}{4\sqrt{2}} \leq \|w\| \leq R$
3. $\frac{w^* \nabla \mathbb{E}[g(w)]w}{\|w\|^2} \leq -c'\theta$, for $\|w\| \geq R$

($R = \Theta(1)$).

and so, **every local optimizer of** $\mathbb{E}[g(w)]$ **is a target point.**
When does the finite-sample objective converge to the asymptotic one, in optimization sense? ...informally, is the function geometry “nice” for some large but finite $p$?
Objective $g(w) = \frac{1}{p} \sum_{i=1}^{p} h_{\mu}(q(w)^* x_i)$ is a sum of independent RVs.

The proof follows a typical concentration-expectation path

Lemma

Suppose $\theta \in \left[\frac{1}{n}, \frac{1}{2}\right]$, if $\mu < cn^{-5/4}$, and $p \geq \frac{Cn^3}{\mu^2 \theta^2} \log n$, it holds uniformly w.h.p. that

\[ \nabla^2 \mathbb{E}[g(w)] \leq \frac{c'' \theta}{\mu} I, \text{ for } \|w\| \leq \frac{\mu}{4\sqrt{2}} \]

\[ \frac{w^* \nabla \mathbb{E}[g(w)]}{\|w\|} \geq c'' \theta, \text{ for } \frac{\mu}{4\sqrt{2}} \leq \|w\| \leq R \]

\[ \frac{w^* \nabla \mathbb{E}[g(w)] w}{\|w\|^2} \leq -c'' \theta, \text{ for } \|w\| \geq R (R = \Theta(1)). \]
... following intuition we build from the geometry:

- Don’t know the chart $q(w)$ ahead of time
  $\implies$ work directly on $q \in \mathbb{S}^{n-1}$.
- Pull the “niceness” back to the sphere:
  descent direction in $w \iff$ descent direction in $q$
  along some curve
- Need to escape saddle points $\implies$ Use
  second-order information. Here, via the
  trust region method.

Trust-region on manifolds [Absil, Baker, Gallivan.
’07], also [Absil, Mahoney, Sepulchre. ’08]
Consider $q \in S^{n-1}$; for $\delta \perp q$, calculus gives

$$f(\exp_q(\delta)) = f(q) + \langle \delta, \nabla f(q) \rangle + \frac{1}{2} \delta^* (\nabla^2 f(q) - \langle q, \nabla f(q) \rangle) \delta + O(\|\delta\|_2^3)$$

$$= \tilde{f}(q; \delta) + O(\|\delta\|_2^3)$$

where $\exp_q(\delta) = q \cos(\|\delta\|_2) + \frac{\delta}{\|\delta\|_2} \sin(\|\delta\|_2)$.

Basic Riemannian trust region method:

$$\delta^* \in \arg\min_{\delta \in T_{q_k}S^{n-1}, \|\delta\|_2 \leq \Delta} \tilde{f}(q_k; \delta)$$

$$q_{k+1} = \exp_{q_k}(\delta^*).$$

The trust region subproblem involves a (possibly nonconvex) quadratic and one quadratic constraint. Solvable in polynomial time by root finding [More+Sorensen ’82] or SDP relaxation.
Theorem (informal)

Suppose that \( \theta \in \left[ \frac{1}{n}, \frac{1}{2} \right), \mu < cn^{-5/4}, Y = QX \) with \( Q \in O(n) \). If we observe \( p \geq \text{poly}(n) \) samples, then applying the trust region method with fixed radius \( \Delta = \frac{1}{\text{poly}(n)} \) for \( T = \text{poly}(n) \) iterations. W.h.p, the algorithm produces a \( \hat{q} \) such that

\[
\| \hat{q} - q_\diamond \| \leq C \frac{\mu}{\theta} \sqrt{\frac{n \log p}{p}}
\]

for some target solution \( q_\diamond \) satisfying \( q_\diamond Y = \pm e_i^* X \).

- Using linear programming rounding + deflation, one can recover all of \( X \), and subsequently \( Q \).
- If \( Q \) is not an orthobasis, apply preconditioning, but need \( p \geq \text{poly}(n, \sigma_{\min}(Q)^{-1}) \).
Comparison with the Literature

- **Efficient algorithms** with performance guarantees
  - [Spielman, Wang, Wright,’12] \( Q \in \mathbb{R}^{n \times n}, \theta = \tilde{O} \left( \frac{1}{\sqrt{n}} \right) \)
  - [Agarwal, Anandkumar, Netrapali,’13] \( Q \in \mathbb{R}^{m \times n} (m \leq n), \theta = \tilde{O} \left( \frac{1}{\sqrt{n}} \right) \)
  - [Arora, Ge, Moitra,’13] \( Q \in \mathbb{R}^{m \times n} (m \leq n), \theta = \tilde{O} \left( \frac{1}{\sqrt{n}} \right) \)

- **Quasipolynomial algorithms** with better guarantees
  - [Spielman, Wang, Wright,’12] \( Q \in \mathbb{R}^{n \times n}, \theta = \tilde{O} \left( \frac{1}{\sqrt{n}} \right) \)
  - [Arora, Bhaskara, Ge, Ma,’14] different prob. model, \( \theta = O \left( \frac{1}{\text{polylog}(n)} \right) \)
  - [Barak, Kelner, Steurer,’14] sum-of-squares, \( \theta = \tilde{O} \left( 1 \right) \)

- Other theoretic work on **local geometry**: [Gribonval, Schnass’11], [Geng, Wright, ’11], [Schnass’14]

This work: a polynomial algorithm for squared \( Q, \theta = O(1) \).
What we have done so far...

\[
\min \frac{1}{p} \sum_{i=1}^{p} h_\mu (q^* y_i) \quad \text{s.t. } \|q\|_2^2 = 1. \quad Y \in \mathbb{R}^{n \times p}
\]

- Prove as \( p \) becomes large, the nonconvex program becomes tractable under our probabilistic setting.
- Geometry has guided our analysis and algorithm design.

Related publications:
- Sun, Qu, Wright. Complete dictionary recovery over the sphere. In preparation.
For sparse dictionary learning:

- Streamline the proof and work directly in manifold language
- With directly with the $\| \cdot \|_1$ function (partial progress)
- Does similar thing happens if we look at structured dictionary directly? (orthogonal group - very likely; tight frame - likely)
- Algorithm side: how to understand the surprisingly successful alternating direction method in this setting?
- ...

Ju Sun [Advisor: Prof. John Wright]  When Nonconvex Optimization Meets Big Data
Proving nonconvex recovery

Theme of this thesis

- Certain nonconvex optimization problems become tractable when the input data are large and random (generic).
- Geometry of the function landscape provides important clues for algorithm design and analysis.

Matrix completion: [Keshavan, Oh, Montanari.'09], [Jain, Netrapali, Sanghavi. '13], [Hardt'13], [Hardt, Wooters. '14]. Also [Meta, Jain, Dhillon.'09]

Dictionary learning: [Agarwal, Anandkumar, Netrapali. '13 ], [Arora, Ge, Moitra. '13], [Agarwal, Anandkumar, Jain, Netrapali.'13]

Tensor recovery: [Jain, Oh. '13], [Anandkumar, Ge, Janzamin. '14]

Phase retrieval: [Netrapali, Jain, Sanghavi.'13], [Candes, Li, Soltanolkotabi. '14]
For the analytic strategy:

Generalized Model Problem

\[
\min \frac{1}{p} \sum_{i=1}^{p} f_i(q) \quad \text{s.t. } q \in \mathcal{M}.
\]

\(f_i\)'s are independent, and \(\mathcal{M}\) is some Riemannian manifold
\((S^{n-1}, O(n), \{X : \text{rank}(X) = r\}, \text{etc})\)

- Other problems: phase retrieval, matrix/tensor recovery, recovery of signal with simultaneous structures, **blind deconvolution**, etc
THANKS to Prof. John Wright and Mr. Qing Qu.

Questions?