When Are Nonconvex Optimization Problems Not Scary?

Ju Sun

joint work with Qing Qu, John Wright
Electrical Engineering, Columbia University

Stanford University
February 1, 2016
A curious experiment

Try to learn a concise approximation: $Y \approx QX$, with $Q \in O$ and $X$ as sparse as possible.

... by solving $\min_{Q \in O} \frac{1}{2} \|Y - QX\|_F^2 + \lambda \|X\|_1$, s.t. $Q \in O$.

An image

Patches

$Y \in \mathbb{R}^{n \times p}$
A curious experiment

Try to learn a concise approximation: $Y \approx QX$, with $Q \in O_n$ and $X$ as sparse as possible.
A curious experiment

Try to learn a concise approximation: \( Y \approx QX \), with \( Q \in O_n \) and \( X \) as sparse as possible.

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

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

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

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

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

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

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

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

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

\[
X_k = S_\lambda [Q_{k-1}^* Y]
\]
\[
Q_k = UV^*, \quad \text{where } U\Sigma V^* = \text{SVD} (Y X^*) .
\]
A curious experiment

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

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

\[ X_k = S_\lambda \left[ Q_{k-1}^* Y \right] \]
\[ Q_k = U V^*, \text{ where } U \Sigma V^* = \text{SVD} (YX^*) \]
Global solutions of feature learning on real images?

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

An image

Patches

\[ Y \in \mathbb{R}^{n \times p} \]
Many problems in modern **signal processing, data analysis, statistical estimation, ...**, are most naturally formulated as **nonconvex** (possibly also nonsmooth) optimization problems.

Heuristic algorithms are often surprisingly successful...

Concoct an efficient heuristic, e.g., gradient descent alternating directions.
Apply it to data... …without worrying about convergence, recovery.
Nonconvex optimization in theory

Classical picture:

\[ \min f(\mathbf{x}) \quad \text{s.t. } \mathbf{x} \in \mathcal{D}. \]

“easy”

“hard”

NCVX: Even computing a local minimizer is NP-hard! (see, e.g., [Murty and Kabadi, 1987])
In practice: Heuristic algorithms are often surprisingly successful...

In theory: Even computing a local minimizer is NP-hard!

*Which nonconvex optimization problems are easy?*

**Working hypothesis**

- Certain nonconvex optimization problems have a **benign structure** when the input data are **large** and/or **random/generic**.
- This benign structure allows *"initialization-free"* iterative methods to **efficiently** find a *"global"* minimizer.
1. The “$\mathcal{X}$” (second-order convex?) functions

2. Examples from practical problems
   - Sparse (complete) dictionary learning [Sun et al., 2015a]
   - Generalized phase retrieval [Sun et al., 2015b]
   - Orthogonal tensor decomposition [Ge et al., 2015]

3. Algorithms: Riemannian trust-region method

4. Comparison with alternatives
Outline

1. The "X" (second-order convex?) functions

2. Examples from practical problems
   - Sparse (complete) dictionary learning [Sun et al., 2015a]
   - Generalized phase retrieval [Sun et al., 2015b]
   - Orthogonal tensor decomposition [Ge et al., 2015]

3. Algorithms: Riemannian trust-region method

4. Comparison with alternatives
A classical example ...

For a symmetric matrix $A \in \mathbb{R}^{n \times n}$ and

$$f(x) = x^\top A x \quad \forall x : \|x\|_2 = 1.$$  

- Critical points: $\{\pm v_i\}$

Suppose $\lambda_1 > \lambda_2 \geq \ldots \lambda_{n-1} > \lambda_n$.

- The only global minimizers are $\pm v_n$
- The only global maximizers are $\pm v_1$
- All $\{\pm v_i\}$ for $2 \leq i \leq n - 1$ are saddle points with a directional negative curvature.

$$A = \text{diag}(1, 0, -1)$$
\( \mathcal{X} \) (second-order convex?) functions

\( \mathcal{X} \) functions (qualitative version):

- (P-1) All local minimizers are also global
- (P-2) All saddle points have directional negative curvature

Thanks to (P-1), focus on finding a local minimizer!
More on (P-2): Saddle points

\[ \nabla^2 f = \text{diag}(2, -2) \]

**Ridable saddle**

*(strict saddle [Ge et al., 2015])*

\[ f(x, y) = x^2 - y^2 \]

---

\[ \nabla^2 f = \text{diag}(6x, -6y) \]

Local shape determined by high-order derivatives around 0

\[ f(x, y) = x^3 - y^3 \]
Consider twice continuously differentiable function \( f: \mathcal{M} \mapsto \mathbb{R} \), where \( \mathcal{M} \) is a Riemannian manifold.

### (P-2)+

- (P-2A) For all local minimizers, \( \text{Hess} f > 0 \), and
- (P-2B) For all other critical points, \( \lambda_{\text{min}}(\text{Hess} f) < 0 \).

- (P-2A) \( \implies \) local strong convexity around any local minimizer
- (P-2B) \( \implies \) local directional strict concavity around local maximizers and **saddle points**; particularly, **all saddles are ridable (strict)**.
A smooth function $f : \mathcal{M} \mapsto \mathbb{R}$ is called Morse if all critical points are nondegenerate.

All Morse functions are ridable (strict)-saddle functions!

The Morse functions form an open, dense subset of all smooth functions $\mathcal{M} \mapsto \mathbb{R}$.

A typical/generic function is Morse!
Ridable-saddle (strict-saddle) functions A function $f : \mathcal{M} \mapsto \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-ridable $(\alpha, \beta, \gamma, \delta > 0)$ if any point $x \in \mathcal{M}$ obeys at least one of the following:

1) [Strong gradient] $\| \nabla f(x) \| \geq \beta$;
2) [Negative curvature] There exists $v \in T_x \mathcal{M}$ with $\| v \| = 1$ such that $\langle \text{Hess } f(x)[v], v \rangle \leq -\alpha$;
3) [Strong convexity around minimizers] There exists a local minimizer $x_\star$ such that $\| x - x_\star \| \leq \delta$, and for all $y \in \mathcal{M}$ that is in $2\delta$ neighborhood of $x_\star$, $\langle \text{Hess } f(y)[v], v \rangle \geq \gamma$ for any $v \in T_y \mathcal{M}$ with $\| v \| = 1$.

($T_x \mathcal{M}$ is the tangent space of $\mathcal{M}$ at point $x$)
1. The “$\mathcal{X}$” (second-order convex?) functions

2. Examples from practical problems
   - Sparse (complete) dictionary learning [Sun et al., 2015a]
   - Generalized phase retrieval [Sun et al., 2015b]
   - Orthogonal tensor decomposition [Ge et al., 2015]

3. Algorithms: Riemannian trust-region method

4. Comparison with alternatives
(P-1) All local minimizers are also global,
(P-2A) For all local minimizers, $\text{Hess } f \succ 0$, and
(P-2B) For all other critical points, $\lambda_{\min}(\text{Hess } f) < 0$.

... focus on finding a local minimizer
Outline

1. The “$\mathcal{X}$” (second-order convex?) functions

2. Examples from practical problems
   - Sparse (complete) dictionary learning [Sun et al., 2015a]
   - Generalized phase retrieval [Sun et al., 2015b]
   - Orthogonal tensor decomposition [Ge et al., 2015]

3. Algorithms: Riemannian trust-region method

4. Comparison with alternatives

Important algorithmic contributions from many researchers: e.g., [Lewicki and Sejnowski, 2000, Engan et al., 1999, Aharon et al., 2006], many others

Widely used in image processing, recently used in visual recognition, compressive signal acquisition, deep architecture for signal classification (see, e.g., [Mairal et al., 2014])
Dictionary recovery - given $Y$ generated as $Y = Q_0 X_0$, recover $Q_0$ and $X_0$.

Our Model

$Q_0$ complete (square and invertible),
$X_0 = \Omega \odot G$, $\Omega \sim \text{i.i.d. Ber}(\theta)$, $G \sim \text{i.i.d. N}(0, 1)$.
Dictionary recovery – given \( Y \) generated as \( Y = Q_0 X_0 \), recover \( Q_0 \) and \( X_0 \).

Our Model

\( Q_0 \) complete (square and invertible),

\[ X_0 = \Omega \odot G, \Omega \sim_{i.i.d.} \text{Ber} (\theta), G \sim_{i.i.d.} \mathcal{N} (0, 1). \]

- \( Q_0 \) complete \( \implies \) \( \text{row} (Y) = \text{row} (X_0) \) \( \implies \) rows of \( X_0 \) are sparse vectors in \( \text{row} (Y) \)

- When \( p \geq \Omega (n \log n) \), rows of \( X_0 \) are the sparsest vectors in \( \text{row} (Y) \) [Spielman et al., 2012]
Dictionary recovery – given $Y$ generated as $Y = Q_0 X_0$, recover $Q_0$ and $X_0$.

**Our Model**

$Q_0$ complete (square and invertible),

$X_0 = \Omega \odot G$, $\Omega \sim \text{i.i.d.} \text{Ber}(\theta)$, $G \sim \text{i.i.d.} \mathcal{N}(0, 1)$.

\[
\text{row}(Y) = \text{row}(X_0)
\]

Find the sparse vectors in $\text{row}(Y)$!
Dictionary learning: the complete case

Nonconvex “relaxation”:

Model problem

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

many precedents, e.g., [Zibulevsky and Pearlmutter, 2001] in blind source separation.
Towards geometric understanding

**Model problem**

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

**Slightly modified model problem**

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

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

  \[
h_\mu(z) = \mu \log \cosh \frac{z}{\mu}
\]

Sun, Qu, and Wright

When Are Nonconvex Optimization Problems Not Scary?
An \( \mathcal{X} \) function!

A low-dimensional example \((n = 3)\) of the landscape when the target dictionary \( A_0 \) is orthogonal.
The results

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

Theorem (Informal, [Sun et al., 2015a])

When \( p \) is reasonably large, and \( \theta \) constant, with high probability,

- All local minimizers produce close approximations to rows of \( X_0 \)
- \( f \) is \((c\theta, c\theta, c\theta/\mu, \sqrt{2\mu/7})\)-ridable over \( \mathbb{S}^{n-1} \) for some \( c > 0 \)

Algorithms later ...
Comparison with the DL Literature

- **Efficient algorithms** with performance guarantees
  - [Spielman, Wang, Wright,’12] $Q \in \mathbb{R}^{n \times n}, \theta = \tilde{O}(1/\sqrt{n})$
  - [Agarwal, Anandkumar, Netrapali,’13] $Q \in \mathbb{R}^{m \times n} (m \leq n), \theta = \tilde{O}(1/\sqrt{n})$
  - [Arora, Ge, Moitra,’13] $Q \in \mathbb{R}^{m \times n} (m \leq n), \theta = \tilde{O}(1/\sqrt{n})$
  - [Arora, Ge, Ma, Moitra,’15] $Q \in \mathbb{R}^{m \times n} (m \leq n), \theta = \tilde{O}(1/\sqrt{n})$

- **Quasipolynomial algorithms** with better guarantees
  - [Arora, Bhaskara, Ge, Ma,’14] different prob. model, $\theta = O\left(1/\text{polylog}\left(n\right)\right)$
  - [Barak, Kelner, Steurer,’14] sum-of-squares, $\theta = \tilde{O}(1)$

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

This work: the first **polynomial-time** algorithm for complete $A$ with $\theta = \Omega(1)$. 
Outline

1. The “$\mathcal{X}$” (second-order convex?) functions

2. Examples from practical problems
   - Sparse (complete) dictionary learning [Sun et al., 2015a]
   - Generalized phase retrieval [Sun et al., 2015b]
   - Orthogonal tensor decomposition [Ge et al., 2015]

3. Algorithms: Riemannian trust-region method

4. Comparison with alternatives
Example II: Generalized phase retrieval

Phase retrieval: Given phaseless information of a complex signal, recover the signal

Applications: X-ray crystallography, diffraction imaging (left), optics, astronomical imaging, and microscopy

Coherent Diffraction Imaging\(^1\)

For a complex signal \(x \in \mathbb{C}^n\), given \(|Fx|\), recover \(x\).

\(^1\)Image courtesy of [Shechtman et al., 2015]
For a complex signal $x \in \mathbb{C}^n$, given $|Fx|$, recover $x$.

**Generalized phase retrieval:**

For a complex signal $x \in \mathbb{C}^n$, given measurements of the form $|a_k^*x|$ for $k = 1, \ldots, m$, recover $x$.

... in practice, generalized measurements by design such as masking, grating, structured illumination, etc.

---

$^2$Image courtesy of [Candès et al., 2015a]
A nonconvex formulation

- Given $y_k = |a_k^* x|^2$ for $k = 1, \ldots, m$, recover $x$ (up to a global phase).
- A natural nonconvex formulation (see also [Candès et al., 2015a])

$$\min_{z \in \mathbb{C}^n} f(z) = \frac{1}{2m} \sum_{k=1}^{m} (y_k - |a_k^* z|^2)^2.$$
A nonconvex formulation

- Given $y_k = |a_k^* x|^2$ for $k = 1, \ldots, m$, recover $x$ (up to a global phase).
- A natural nonconvex formulation (see also [Candès et al., 2015a])

$$ \min_{z \in \mathbb{C}^n} f(z) = \frac{1}{2m} \sum_{k=1}^{m} (y_k - |a_k^* z|^2)^2. $$

When $a_k$'s are iid standard complex Gaussian vectors and $m$ large
The results

\[ \min_{z \in \mathbb{C}^n} f(z) = \frac{1}{2m} \sum_{k=1}^{m} (y_k - |a_k^* z|^2)^2. \]

**Theorem (Informal, [Sun et al., 2015b])**

*When \( m \geq \Omega(n \text{polylog}(n)) \), with high probability,*

- All local (and global) minimizers are \( x \) with a global phase shift
- \( f \) is \((c, c/(n \log m), c, c/(n \log m))\)-ridable over \( \mathbb{C}^n \) for some \( c > 0 \)
Other measurement models for GPR

Other measurements

- Coded diffraction model [Candès et al., 2015b]

\[ y = |a \otimes x|^2 \]
1. The “\( \mathcal{X} \) (second-order convex?) functions

2. Examples from practical problems
   - Sparse (complete) dictionary learning [Sun et al., 2015a]
   - Generalized phase retrieval [Sun et al., 2015b]
   - Orthogonal tensor decomposition [Ge et al., 2015]

3. Algorithms: Riemannian trust-region method

4. Comparison with alternatives
Example III: Orthogonal tensor decomposition

... generalizes eigen-decomposition of matrices

Orthogonally decomposable (OD) $d$-th order tensors

$$\mathcal{T} = \sum_{i=1}^{r} \lambda_i a_i^{\otimes d}, \quad a_i^\top a_j = \delta_{ij} \ \forall \ i, j, (a_i \in \mathbb{R}^n \ \forall \ i)$$

where $\otimes$ generalizes the usual outer product of vectors.

Orthogonal tensor decomposition: given OD tensor $\mathcal{T}$, find the components $a_i$'s (up to sign).

Applications: independent component analysis (ICA), blind source separation, latent variable model learning, etc (see, e.g., [Anandkumar et al., 2014])
Focus on OD tensors of the form

\[ T = \sum_{i=1}^{r} a_i \otimes^4, \quad a_i^\top a_j = \delta_{ij} \ \forall \ i, j, (a_i \in \mathbb{R}^n \ \forall \ i) \]

Consider

\[
\min f(u) = -T(u, u, u, u) = -\sum_{i=1}^{r} (a_i^\top u)^4 \quad \text{s. t.} \quad \|u\|_2 = 1
\]

[Ge et al., 2015] proved that

- \( f \) is \((7/r, 1/poly(r), 3, 1/poly(r))\)-ridable over \( \mathbb{S}^{n-1} \)
- \( \pm a_i \)'s are the only minimizers
All components in one shot

Focus on OD tensors of the form

\[
\mathcal{T} = \sum_{i=1}^{r} a_i \otimes^4, \quad a_i^\top a_j = \delta_{ij} \forall i, j, (a_i \in \mathbb{R}^n \forall i)
\]

Consider

\[
\min g(u_1, \ldots, u_r) = \sum_{i \neq j} \mathcal{T}(u_i, u_i, u_j, u_j)
\]

\[
= \sum_{i \neq j} \sum_{k=1}^{r} (a_k^\top u_i)^2 (a_k^\top u_j)^2
\]

s. t. \( \|u_i\| = 1 \forall i \in [r] \).

[Ge et al., 2015] proved that

- \( g \) is \((1/\text{poly}(r), 1/\text{poly}(r), 1, 1/\text{poly}(r))\)-ridable
- All local minimizers of \( g \) are equivalent (i.e., signed permuted) copies of \([a_1, \ldots, a_r]\)
1 The “$\mathcal{X}$” (second-order convex?) functions

2 Examples from practical problems
   - Sparse (complete) dictionary learning [Sun et al., 2015a]
   - Generalized phase retrieval [Sun et al., 2015b]
   - Orthogonal tensor decomposition [Ge et al., 2015]

3 Algorithms: Riemannian trust-region method

4 Comparison with alternatives
(P-1) All local minimizers are also global,
(P-2A) For all local minimizers, $\text{Hess } f \succ 0$, and
(P-2B) For all other critical points, $\lambda_{\text{min}}(\text{Hess } f) < 0$.

... focus on **escaping saddle points** and maximizers and finding a **local minimizer**.
Algorithmic possibilities

- Second-order trust-region method (described here, [Conn et al., 2000])
- Curvilinear search [Goldfarb, 1980]
- Noisy/stochastic gradient descent [Ge et al., 2015]
- ...

Sun, Qu, and Wright
When Are Nonconvex Optimization Problems Not Scary?
Taylor expansion at a saddle point $x$:

$$\hat{f}(\delta; x) = f(x) + \frac{1}{2} \delta^* \nabla^2 f(x) \delta.$$ 

Choosing $\delta = v_{neg}$, then

$$\hat{f}(\delta; x) - f(x) \leq -\frac{1}{2} |\lambda_{neg}| \|v_{neg}\|^2.$$  

Function value decreasing is guaranteed when movement is small such that the approximation is reasonably good.
Similarly for the maximizers we consider.
Consider an iterate sequence \( x_0, x_1, x_2, \ldots \)

- At the current iterate \( x_k \), form a second-order approximation:

\[
\hat{f}(\delta; x_k) = f(x_k) + \langle \nabla f(x_k), \delta \rangle + \frac{1}{2} \delta^* B_k \delta.
\]

and minimize the approximation within a small radius - the trust region

\[
\delta_* = \arg \min_{\|\delta\| \leq \Delta} \hat{f}(\delta; x_k) \quad \text{(Trust-region subproblem)}
\]

- Next iterate is \( x_{k+1} = x_k + \delta_* \)

- \( B_k \) can be chosen to be the Hessian, or approximations.

We focus on \( B_k = \nabla^2 f(x^{(k)}) \).
Take an example: $f : \mathbb{S}^{n-1} \mapsto \mathbb{R}$.

$$\exp_q(\delta) = q \cos \|\delta\| + \frac{\delta}{\|\delta\|} \cdot \sin \|\delta\|$$

For $q \in \mathbb{S}^{n-1}$ and $\delta \in T_q\mathbb{S}^{n-1}$, define $f_q : T_q\mathbb{S}^{n-1} \mapsto \mathbb{R}$ as $f_q = f(\exp_q(\delta))$.
Take an example: \( f : \mathbb{S}^{n-1} \mapsto \mathbb{R} \).

\[
\exp_q(\delta) = q \cos \|\delta\| + \frac{\delta}{\|\delta\|} \cdot \sin \|\delta\|
\]

For \( q \in \mathbb{S}^{n-1} \) and \( \delta \in T_q \mathbb{S}^{n-1} \), define \( f_q : T_q \mathbb{S}^{n-1} \mapsto \mathbb{R} \) as \( f_q = f(\exp_q(\delta)) \).

Taylor’s theorem implies

\[
f(\exp_q(\delta)) = f(q) + \delta^* \nabla f(q) + \frac{1}{2} \delta^* (\nabla^2 f(q) - q^* \nabla f(q) I) \delta + O(\|\delta\|^3)
\]

\[
= f(q) + \delta^* \text{grad} f(q) + \frac{1}{2} \delta^* \text{Hess} f(q) \delta + O(\|\delta\|^3).
\]

\[
\hat{f}_{q_k}(\delta;q) = \hat{\delta}^* f_{q_k} (\delta; q_k)
\]

Basic Riemannian trust-region method:

\[
\delta_* \in \arg \min_{\delta \in T_{q_k} \mathbb{S}^{n-1}, \|\delta\| \leq \Delta} \hat{f}_{q_k}(\delta; q_k)
\]

\[
q_{k+1} = \exp_{q_k}(\delta_*).
\]

More details on Riemannian TRM in [Absil et al., 2007] and [Absil et al., 2009].
The trust-region subproblem

\[ \delta^*_x \in \arg \min_{\delta \in T_{q_k}S^{n-1}, \|\delta\| \leq \Delta} \widehat{f}_{q_k}(\delta; q_k) \quad \text{(Trust-region subproblem)} \]

- If the norm is \( \ell^2 \), quadratic constrained quadratic program (QCQP - hard in general)
- This case can be \textbf{exactly} solved by root finding [Moré and Sorensen, 1983] or SDP relaxation [Rendl and Wolkowicz, 1997].
- In practice, only approximate solution (with controllable quality) needed to ensure convergence.
Proof of convergence

- When the gradient is strong or the curvature is negative, function value decrease by at least a fixed amount;
- Under mild conditions, the sequence will ultimately move into the strongly convex region around a local minimizer;
- The algorithm acts like a typical second-order method on convex function and local quadratic convergence in sequence is observed.

Theorem (Very informal)

For ridable-saddle functions, starting from an arbitrary initialization, the iteration sequence with sufficiently small step size (trust-region size) converges to a local minimizer in polynomial number of steps.

worked out examples in [Sun et al., 2015a, Sun et al., 2015b]; see also promise of 1-st order method [Ge et al., 2015].
1. The “$\mathcal{X}$” (second-order convex?) functions

2. Examples from practical problems
   - Sparse (complete) dictionary learning [Sun et al., 2015a]
   - Generalized phase retrieval [Sun et al., 2015b]
   - Orthogonal tensor decomposition [Ge et al., 2015]

3. Algorithms: Riemannian trust-region method

4. Comparison with alternatives
Convexification

Convexity allows separation of formulations/analysis from algorithms.

Vast array of beautiful mathematical results, substantial applied impact:
- Important examples: sparse recovery, low-rank matrix recovery/completion
- General frameworks: atomic norms [Chandrasekaran et al., 2012], submodular sparsity inducers [Bach, 2010], restricted strong convexity [Negahban et al., 2009], conic statistical dimensions [Amelunxen et al., 2014], etc

Find a **tractable convex surrogate** for \( f \).

**Minimize** the surrogate.

**Prove** that for well-structured instances, the solution is accurate.
But... sometimes the recipe doesn’t work

- The natural convex surrogates may be intractable:
  - Tensor recovery [Hillar and Lim, 2013]
  - Nonnegative low-rank approximation [Vavasis, 2009]

- Or the natural relaxations subject to fundamental limitations:
  - Simultaneous structure estimation [Oymak et al., 2012]
  - Tensor recovery [Mu et al., 2013]
  - Sparse PCA [Berthet and Rigollet, 2013]
  - Dictionary learning [Spielman et al., 2012]

In all these cases, there are substantial and provable gaps between the performance of known convex relaxations and the information theoretic optimum.

In addition, computations are often expensive and impractical (e.g., SDP lifting) even for medium-scale problems.
Prior work: proving nonconvex recovery

1. Use problem structure to find a clever initial guess.
2. Analyze iteration-by-iteration in the vicinity of the optimum.

- **Matrix completion/recovery**: [Keshevan, Oh, Montanari.’09], [Jain, Netrapali, Sanghavi. ’13], [Hardt’13], [Hardt, Wooters. ’14], [Netrapalli et al. ’14], [Jain + Netrapalli,’14], [Zheng, Lafferty.’15], [Tu et al’15]. Also [Meta, Jain, Dhillon.’09]

- **Dictionary learning**: [Agarwal, Anandkumar, Netrapali. ’13 ], [Arora, Ge, Moitra. ’13], [Agarwal, Anandkumar, Jain, Netrapali.’13], [Arora, Ge, Ma, Moitra. ’15]

- **Tensor recovery**: [Jain, Oh. ’13], [Anandkumar, Ge, Janzamin. ’14]

- **Phase retrieval**: [Netrapali, Jain, Sanghavi.’13], [Candes, Li, Soltanokoltabi. ’14], [Chen, Candes.’15]

Also recovery in statistical sense, ..., e.g., [Loh + Wainwright’12]
Our approach

- We characterize the **geometry**, which is critical to algorithm design whether initialization is used or not.
- The geometry effectively allows **arbitrary initialization**.


