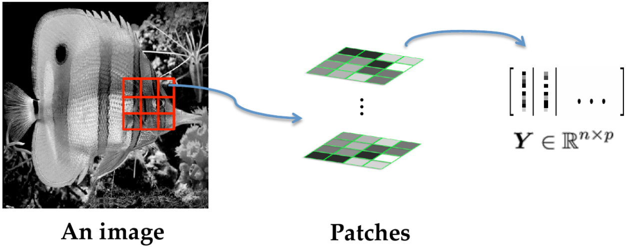
When Are Nonconvex Optimization Problems Not Scary?

Ju Sun

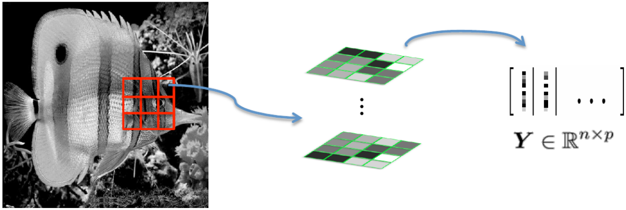
joint work with **Qing Qu**, **John Wright**
Electrical Engineering, Columbia University

Stanford University
February 1, 2016

A curious experiment



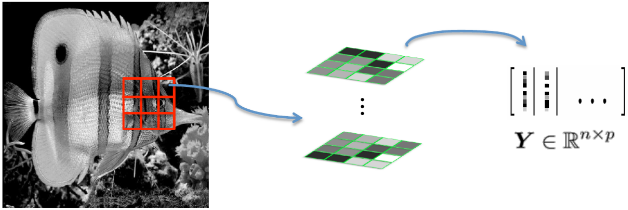
A curious experiment



An image **Patches**

Try to learn a **concise approximation**: $\boldsymbol{Y} \approx \boldsymbol{Q}\boldsymbol{X}$, with $\boldsymbol{Q} \in O_n$ and $\boldsymbol{X}$ as sparse as possible.

A curious experiment

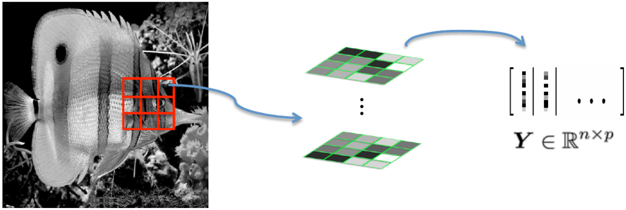


An image **Patches**

Try to learn a **concise approximation**: $\boldsymbol{Y} \approx \boldsymbol{Q}\boldsymbol{X}$, with $\boldsymbol{Q} \in O_n$ and $\boldsymbol{X}$ as sparse as possible.

... by solving $\min \quad \frac{1}{2}\left\|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\right\|_F^2 + \lambda \left\|\boldsymbol{X}\right\|_1, \quad \text{s.t. } \boldsymbol{Q} \in O_n.$

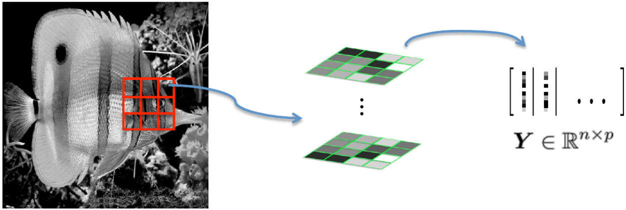
A curious experiment



An image **Patches**

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

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



An image **Patches**

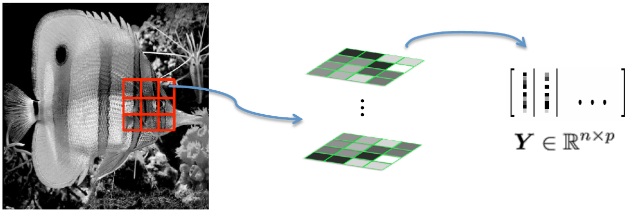
$$\min \quad f(\boldsymbol{Q}, \boldsymbol{X}) \doteq \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_1, \quad \text{s.t. } \boldsymbol{Q} \in O_n$$

Apply the naive **alternating directions**: starting from a random $\boldsymbol{Q}_0 \in O_n$

$$\boldsymbol{X}_k = \arg\min_{\boldsymbol{X}} f\left(\boldsymbol{Q}_{k-1}, \boldsymbol{X}\right)$$

$$\boldsymbol{Q}_k = \arg\min_{\boldsymbol{Q}} f\left(\boldsymbol{Q}, \boldsymbol{X}_k\right), \text{ s.t. } \boldsymbol{Q} \in O_n.$$

A curious experiment



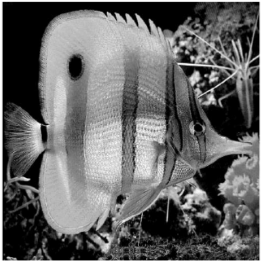
An image **Patches**

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

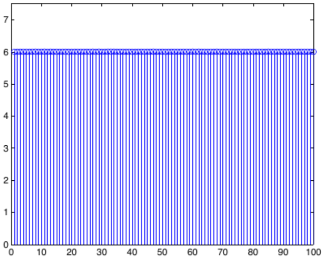
Apply the naive **alternating directions**: starting from a random $\boldsymbol{Q}_0 \in O_n$

$$\boldsymbol{X}_k = \mathcal{S}_\lambda \left[\boldsymbol{Q}_{k-1}^* \boldsymbol{Y}\right]$$
$$\boldsymbol{Q}_k = \boldsymbol{U}\boldsymbol{V}^*, \text{ where } \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \texttt{SVD}\left(\boldsymbol{Y}\boldsymbol{X}^*\right).$$

A curious experiment



An image



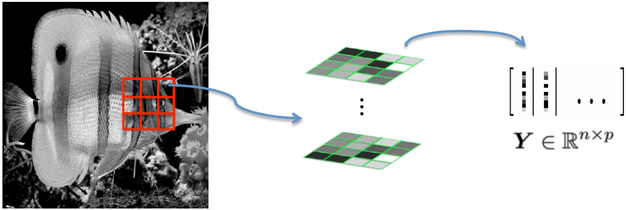
Final $f(Q_\infty, X_\infty)$, varying Q_0.

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

Apply the naive **alternating directions**: starting from a random $\boldsymbol{Q}_0 \in O_n$

$$\boldsymbol{X}_k = \mathcal{S}_\lambda \left[\boldsymbol{Q}_{k-1}^* \boldsymbol{Y}\right]$$
$$\boldsymbol{Q}_k = \boldsymbol{U}\boldsymbol{V}^*, \text{ where } \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \texttt{SVD}(\boldsymbol{Y}\boldsymbol{X}^*).$$

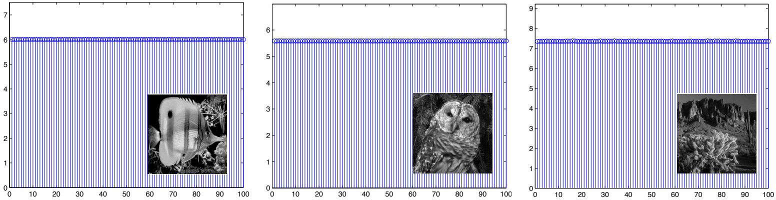
Global solutions of feature learning on real images?



An image

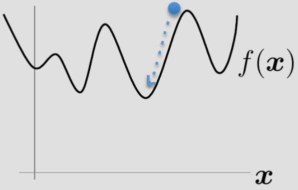
Patches

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



Nonconvex optimization in practice

- 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...



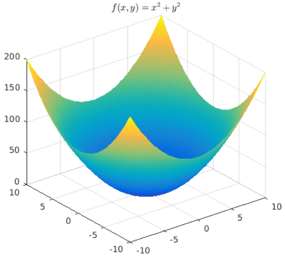
Concoct an efficient heuristic
e.g., **gradient descent**
alternating directions.

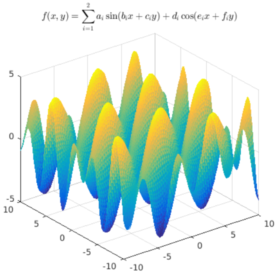
Apply it to data…
…without worrying about
convergence, recovery.

Nonconvex optimization in theory

Classical picture:

$$\min\ f(\boldsymbol{x})$$
$$\text{s. t. } \boldsymbol{x} \in \mathcal{D}.$$





"easy" **"hard"**

NCVX: Even computing a local minimizer is NP-hard! (see, e.g., [Murty and Kabadi, 1987])

This work - a step towards bridging the gap

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 "$\mathcal{X}$" (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

A classical example ...

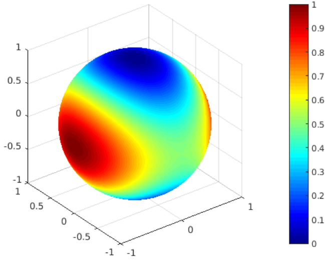
For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and

$$f(\boldsymbol{x}) \doteq \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \quad \forall\, \boldsymbol{x} : \|\boldsymbol{x}\|_2 = 1.$$

- Critical points: $\{\pm \boldsymbol{v}_i\}$

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

- The only **global minimizers** are $\pm \boldsymbol{v}_n$
- The only **global maximizers** are $\pm \boldsymbol{v}_1$
- All $\{\pm \boldsymbol{v}_i\}$ for $2 \leq i \leq n-1$ are **saddle points** with a **directional negative curvature**.



$\boldsymbol{A} = \mathrm{diag}(1, 0, -1)$

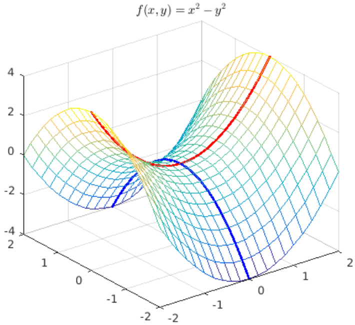
$\mathcal{X}$ (second-order convex?) functions

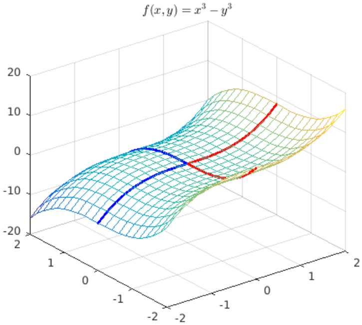
$\mathcal{X}$ **functions** (qualitative version):

- (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





$\nabla^2 f = \mathrm{diag}(2, -2)$
Ridable saddle
(**strict saddle** [Ge et al., 2015])

$\nabla^2 f = \mathrm{diag}(6x, -6y)$
local shape determined by
high-order derivatives around $\mathbf{0}$

More on (P-2): Ridable-saddle functions

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, $\operatorname{Hess} f \succ \mathbf{0}$, and
- (P-2B) For all other critical points, $\lambda_{\min}(\operatorname{Hess} f) < 0$.

- (P-2A) $\Longrightarrow$ local strong convexity around any local minimizer
- (P-2B) $\Longrightarrow$ local directional strict concavity around local maximizers and **saddle points**; particularly, **all saddles are ridable (strict)**.

Morse's examples ...

Definition

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!

Marston Morse
(1892 – 1977)

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

A typical/generic function is Morse!

More on (P-2): A quantitative definition

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

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

($T_{\boldsymbol{x}}\mathcal{M}$ is the tangent space of $\mathcal{M}$ at point $\boldsymbol{x}$)

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

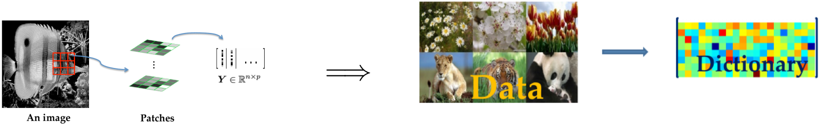
Benign structure

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

... focus on finding a local minimizer

Outline

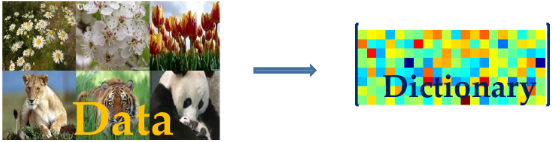
Example I: Sparse Dictionary Learning



$\boldsymbol{Y} \approx \boldsymbol{Q}\boldsymbol{X} \quad \boldsymbol{X} \in \mathbb{R}^{n \times p}$ sparse

- Algorithmic study initialized with [Olshausen and Field, 1996] in neuroscience.
- 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 - the complete case

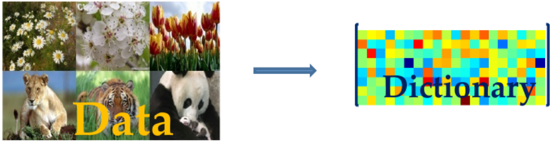


Dictionary recovery – given $\boldsymbol{Y}$ generated as $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

Our Model

$\boldsymbol{Q}_0$ complete (square and invertible),
$\boldsymbol{X}_0 = \boldsymbol{\Omega} \odot \boldsymbol{G}, \boldsymbol{\Omega} \sim_{i.i.d.} \mathrm{Ber}\left(\theta\right), \boldsymbol{G} \sim_{i.i.d.} \mathcal{N}\left(0, 1\right).$

Dictionary recovery - the complete case



Dictionary recovery – given $\boldsymbol{Y}$ generated as $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

Our Model

$\boldsymbol{Q}_0$ complete (square and invertible),
$\boldsymbol{X}_0 = \boldsymbol{\Omega} \odot \boldsymbol{G}, \boldsymbol{\Omega} \sim_{i.i.d.} \mathrm{Ber}(\theta), \boldsymbol{G} \sim_{i.i.d.} \mathcal{N}(0, 1)$.

- $\boldsymbol{Q}_0$ complete $\Longrightarrow$ $\boxed{\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)}$ $\Longrightarrow$ rows of $\boldsymbol{X}_0$ are sparse vectors in $\mathrm{row}(\boldsymbol{Y})$
- When $p \geq \Omega(n \log n)$, rows of $\boldsymbol{X}_0$ are the sparsest vectors in $\mathrm{row}(\boldsymbol{Y})$ [Spielman et al., 2012]

Dictionary recovery - the complete case

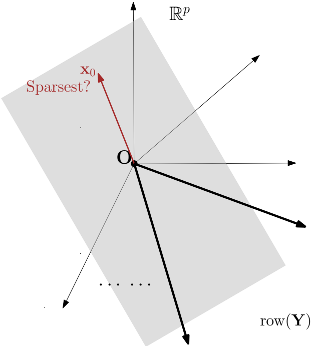
Dictionary recovery – given $\boldsymbol{Y}$ generated as $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

Our Model

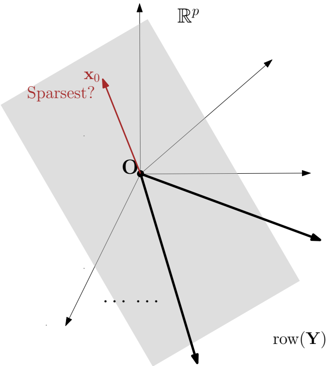
$\boldsymbol{Q}_0$ complete (square and invertible),
$\boldsymbol{X}_0 = \boldsymbol{\Omega} \odot \boldsymbol{G}, \boldsymbol{\Omega} \sim_{i.i.d.} \mathrm{Ber}(\theta), \boldsymbol{G} \sim_{i.i.d.} \mathcal{N}(0,1)$.

$$\boxed{\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)}$$

Find the sparse vectors in $\mathrm{row}(\boldsymbol{Y})$**!**



Dictionary learning: the complete case



$$\min \quad \|\boldsymbol{q}^*\boldsymbol{Y}\|_0 \quad \text{s.t. } \boldsymbol{q} \neq \mathbf{0}.$$

- Nonconvex "relaxation":

Model problem

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

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

Towards geometric understanding

Model problem

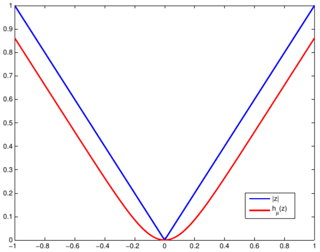
$$\min \quad \frac{1}{p}\|\boldsymbol{q}^*\boldsymbol{Y}\|_1 = \frac{1}{p}\sum_{i=1}^{p}|\boldsymbol{q}^*\boldsymbol{y}_i| \quad \text{s.t. } \|\boldsymbol{q}\|_2^2 = 1. \quad \boldsymbol{Y} \in \mathbb{R}^{n\times p}$$

Slightly modified model problem

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

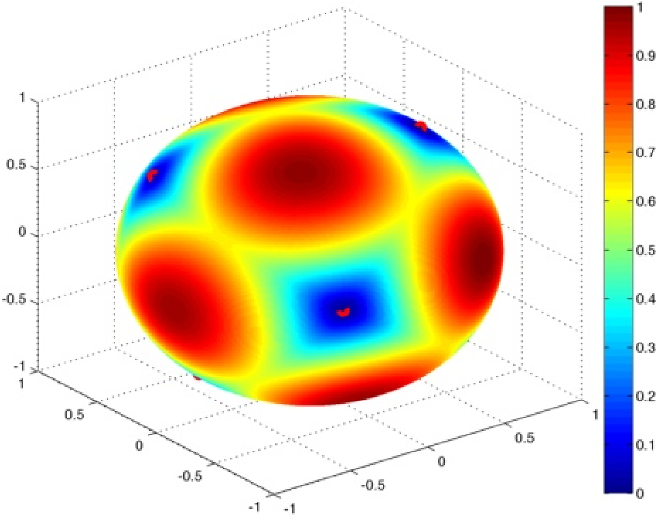
- Work with a *smooth surrogate* for $|z|$:

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



An $\mathcal{X}$ function!

A low-dimensional example ($n = 3$) of the landscape when the target dictionary $\boldsymbol{A}_0$ is orthogonal



The results

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

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

When p is reasonably large, and θ constant, with high probability,

- *All local minimizers produce close approximations to rows of* $\boldsymbol{X}_0$
- 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] $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$
  [Agarwal, Anandkumar, Netrapali,'13] $\boldsymbol{Q} \in \mathbb{R}^{m \times n}$ $(m \leq n)$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$
  [Arora, Ge, Moitra,'13] $\boldsymbol{Q} \in \mathbb{R}^{m \times n}$ $(m \leq n)$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$
  [Arora, Ge, Ma, Moitra,'15] $\boldsymbol{Q} \in \mathbb{R}^{m \times n}$ $(m \leq n)$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$

- **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, $\theta = \tilde{O}\left(1\right)$

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

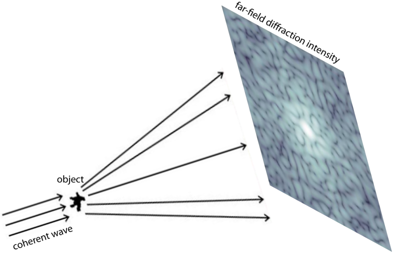
This work: the first **polynomial-time** algorithm for complete $\boldsymbol{A}$ with $\theta = \Omega(1)$.

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



Coherent Diffraction Imaging[1]

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

For a complex signal $\boldsymbol{x} \in \mathbb{C}^n$, given $|\mathcal{F}\boldsymbol{x}|$, recover $\boldsymbol{x}$.

[1]Image courtesy of [Shechtman et al., 2015]

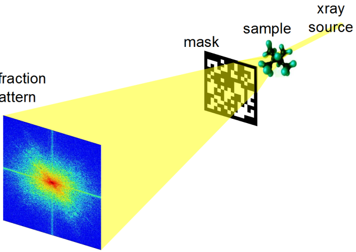
Generalized phase retrieval

For a complex signal $\boldsymbol{x} \in \mathbb{C}^n$, given $|\mathcal{F}\boldsymbol{x}|$, recover $\boldsymbol{x}$.

Generalized phase retrieval:

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

... in practice, generalized measurements by design such as masking, grating, structured illumination, etc [2]



[2]Image courtesy of [Candès et al., 2015a]

A nonconvex formulation

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

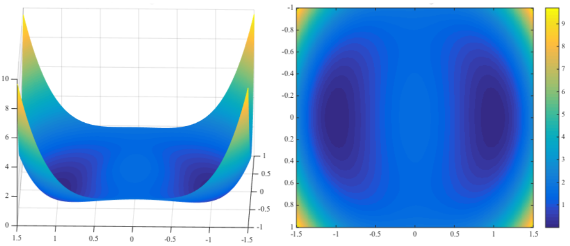
$$\boxed{\min_{\boldsymbol{z} \in \mathbb{C}^n} f(\boldsymbol{z}) \doteq \frac{1}{2m} \sum_{k=1}^{m} (y_k - |\boldsymbol{a}_k^* \boldsymbol{z}|^2)^2.}$$

A nonconvex formulation

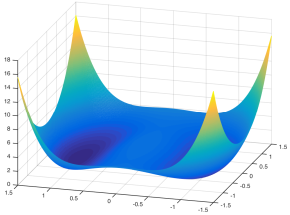
- Given $y_k = |\boldsymbol{a}_k^* \boldsymbol{x}|^2$ for $k = 1, \ldots, m$, recover $\boldsymbol{x}$ (**up to a global phase**).
- A natural **nonconvex** formulation (see also [Candès et al., 2015a])

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} f(\boldsymbol{z}) \doteq \frac{1}{2m} \sum_{k=1}^{m} \left(y_k - |\boldsymbol{a}_k^* \boldsymbol{z}|^2\right)^2.$$

When $\boldsymbol{a}_k$'s are iid standard complex Gaussian vectors and m large



The results



$$\min_{\boldsymbol{z}\in\mathbb{C}^n} f(\boldsymbol{z}) \doteq \frac{1}{2m}\sum_{k=1}^{m}\left(y_k - |\boldsymbol{a}_k^*\boldsymbol{z}|^2\right)^2.$$

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

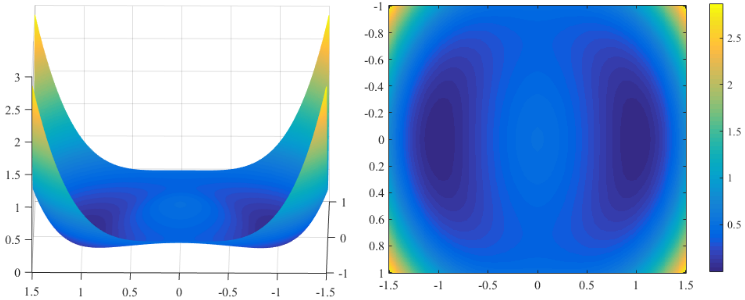
When $m \geq \Omega(n\mathrm{polylog}(n))$, with high probability,

- *All local (and global) minimizers are $\boldsymbol{x}$ 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]



- Convolutional model (with Prof. Yonina Eldar)

$$\boldsymbol{y} = |\boldsymbol{a} \circledast \boldsymbol{x}|^{.2}$$

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 III: Orthogonal tensor decomposition

... generalizes eigen-decomposition of matrices

Orthogonally decomposable (OD) d-th order tensors

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

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

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

Applications: independent component analysis (ICA), blind source separation, latent variable model learning, etc (see, e.g., [Anandkumar et al., 2014])

One component each time

Focus on OD tensors of the form

$$\mathcal{T} = \sum_{i=1}^{r} \boldsymbol{a}_i^{\otimes 4}, \quad \boldsymbol{a}_i^\top \boldsymbol{a}_j = \delta_{ij} \; \forall \; i, j, (\boldsymbol{a}_i \in \mathbb{R}^n \; \forall \; i)$$

Consider

$$\boxed{\min f(\boldsymbol{u}) \doteq -\mathcal{T}(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) = -\sum_{i=1}^{r} (\boldsymbol{a}_i^\top \boldsymbol{u})^4 \quad \text{s. t.} \quad \|\boldsymbol{u}\|_2 = 1}$$

[Ge et al., 2015] proved that

- f is $(7/r, 1/\text{poly}(r), 3, 1/\text{poly}(r))$-ridable over $\mathbb{S}^{n-1}$
- $\pm \boldsymbol{a}_i$'s are the only minimizers

All components in one shot

Focus on OD tensors of the form

$$\mathcal{T} = \sum_{i=1}^{r} \boldsymbol{a}_i^{\otimes 4}, \quad \boldsymbol{a}_i^\top \boldsymbol{a}_j = \delta_{ij} \ \forall\, i, j, (\boldsymbol{a}_i \in \mathbb{R}^n \ \forall\, i)$$

Consider

$$\begin{aligned} \min \ g(\boldsymbol{u}_1, \dots, \boldsymbol{u}_r) &\doteq \sum_{i \neq j} \mathcal{T}(\boldsymbol{u}_i, \boldsymbol{u}_i, \boldsymbol{u}_j, \boldsymbol{u}_j) \\ &= \sum_{i \neq j} \sum_{k=1}^{r} (\boldsymbol{a}_k^\top \boldsymbol{u}_i)^2 (\boldsymbol{a}_k^\top \boldsymbol{u}_j)^2, \\ \text{s. t. } \ \|\boldsymbol{u}_i\| &= 1 \ \forall i \in [r]. \end{aligned}$$

[Ge et al., 2015] proved that

- g is $(1/\mathrm{poly}(r), 1/\mathrm{poly}(r), 1, 1/\mathrm{poly}(r))$-ridable
- All local minimizers of g are equivalent (i.e., signed permuted) copies of $[\boldsymbol{a}_1, \dots, \boldsymbol{a}_r]$

Outline

Benign structure

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

... 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]
- ...

Second-order method can escape ridable saddles

Taylor expansion at a saddle point $\boldsymbol{x}$:

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

Choosing $\boldsymbol{\delta} = \boldsymbol{v}_{\mathrm{neg}}$, then

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}) - f(\boldsymbol{x}) \le -\frac{1}{2}|\lambda_{\mathrm{neg}}| \|\boldsymbol{v}_{\mathrm{neg}}\|^2.$$

Function value decreasing is guaranteed when **movement is small** such that the **approximation is reasonably good**.

Similarly for the maximizers we consider.

Trust region method - Euclidean Space

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \; f(\boldsymbol{x})$$

Consider an iterate sequence $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$

- At the current iterate $\boldsymbol{x}_k$, form a second-order approximation:

$$\widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}_k) = f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{\delta} \rangle + \frac{1}{2} \boldsymbol{\delta}^* \boldsymbol{B}_k \boldsymbol{\delta}.$$

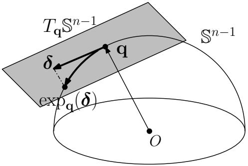
  and minimize the approximation within a small radius - the trust region

$$\boldsymbol{\delta}_\star = \arg\min_{\|\boldsymbol{\delta}\| \le \Delta} \widehat{f}(\boldsymbol{\delta}; \boldsymbol{x}_k) \qquad (\textbf{Trust-region subproblem})$$

- Next iterate is $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{\delta}_\star$
- $\boldsymbol{B}_k$ can be chosen to be the Hessian, or approximations.

We focus on $\boldsymbol{B}_k = \nabla^2 f(\boldsymbol{x}^{(k)})$.

Trust region method - Riemannian Manifold

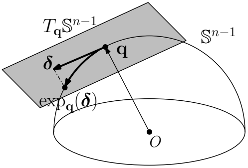


Take an example: $f : \mathbb{S}^{n-1} \mapsto \mathbb{R}$.

$$\exp_{\boldsymbol{q}}(\boldsymbol{\delta}) \doteq \boldsymbol{q} \cos \|\boldsymbol{\delta}\| + \boldsymbol{\delta}/\|\boldsymbol{\delta}\| \cdot \sin \|\boldsymbol{\delta}\|$$

For $\boldsymbol{q} \in \mathbb{S}^{n-1}$ and $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$, define $f_{\boldsymbol{q}} : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \mapsto \mathbb{R}$ as $f_{\boldsymbol{q}} \doteq f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta}))$

Trust region method - Riemannian Manifold



Take an example: $f : \mathbb{S}^{n-1} \mapsto \mathbb{R}$.

$$\exp_{\boldsymbol{q}}(\boldsymbol{\delta}) \doteq \boldsymbol{q} \cos \|\boldsymbol{\delta}\| + \boldsymbol{\delta}/\|\boldsymbol{\delta}\| \cdot \sin \|\boldsymbol{\delta}\|$$

For $\boldsymbol{q} \in \mathbb{S}^{n-1}$ and $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$, define $f_{\boldsymbol{q}} : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \mapsto \mathbb{R}$ as $f_{\boldsymbol{q}} \doteq f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta}))$

Taylor's theorem implies

$$\begin{aligned} f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) &= f(\boldsymbol{q}) + \boldsymbol{\delta}^* \nabla f(\boldsymbol{q}) + \frac{1}{2}\boldsymbol{\delta}^*(\nabla^2 f(\boldsymbol{q}) - \boldsymbol{q}^* \nabla f(\boldsymbol{q}) \boldsymbol{I})\boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3) \\ &= \underbrace{f(\boldsymbol{q}) + \boldsymbol{\delta}^* \operatorname{grad} f(\boldsymbol{q}) + \frac{1}{2}\boldsymbol{\delta}^* \operatorname{Hess} f(\boldsymbol{q})\boldsymbol{\delta}}_{\doteq \widehat{f}_{\boldsymbol{q}_k}(\boldsymbol{\delta};\boldsymbol{q})} + O(\|\boldsymbol{\delta}\|^3). \end{aligned}$$

Basic **Riemannian trust-region method**:

$$\boldsymbol{\delta}_\star \in \operatorname*{arg\,min}_{\boldsymbol{\delta} \in T_{\boldsymbol{q}_k}\mathbb{S}^{n-1}, \|\boldsymbol{\delta}\| \le \Delta} \widehat{f}_{\boldsymbol{q}_k}(\boldsymbol{\delta}; \boldsymbol{q}_k)$$

$$\boldsymbol{q}_{k+1} = \exp_{\boldsymbol{q}_k}(\boldsymbol{\delta}_\star).$$

More details on Riemannian TRM in [Absil et al., 2007] and [Absil et al., 2009].

The trust-region subproblem

$$\boldsymbol{\delta}_\star \in \mathop{\arg\min}_{\boldsymbol{\delta} \in T_{\boldsymbol{q}_k}\mathbb{S}^{n-1}, \|\boldsymbol{\delta}\| \le \Delta} \widehat{f}_{\boldsymbol{q}_k}(\boldsymbol{\delta}; \boldsymbol{q}_k) \qquad \text{(\textbf{Trust-region subproblem})}$$

- If the norm is ℓ^2, quadratic constrained quadratic program (QCQP - hard in general)
- This case can be **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.

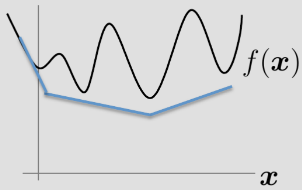
Theorem (Very informal)

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].

Outline

Convexification



Find a **tractable convex surrogate** for f.

Minimize the surrogate.

Prove that for well-structured instances, the solution is accurate.

- 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

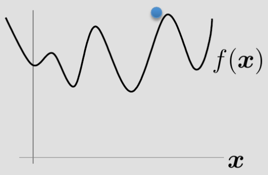
But... sometimes the recipe doesn't work

- **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



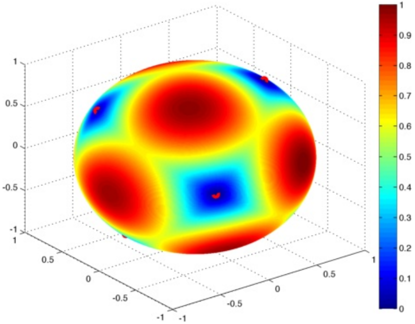
1. Use problem structure to find a **clever initial guess.**

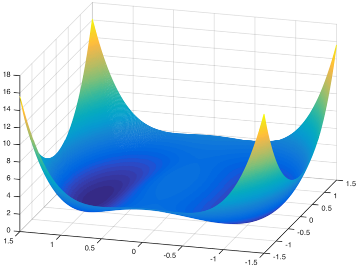
2. Analyze iteration-by-iteration **in the vicinity of the optimum**.

- **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**

References I

[Absil et al., 2007] Absil, P.-A., Baker, C. G., and Gallivan, K. A. (2007). Trust-region methods on Riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330.

[Absil et al., 2009] Absil, P.-A., Mahoney, R., and Sepulchre, R. (2009). *Optimization Algorithms on Matrix Manifolds*. Princeton University Press.

[Aharon et al., 2006] Aharon, M., Elad, M., and Bruckstein, A. (2006). K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Trans. Sig. Proc.*, 54(11):4311–4322.

[Amelunxen et al., 2014] Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, page iau005.

[Anandkumar et al., 2014] Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832.

[Bach, 2010] Bach, F. R. (2010). Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, pages 118–126.

References II

[Berthet and Rigollet, 2013] Berthet, Q. and Rigollet, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*.

[Candès et al., 2015a] Candès, E., Li, X., and Soltanolkotabi, M. (2015a). Phase retrieval via wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007.

[Candès et al., 2015b] Candès, E. J., Li, X., and Soltanolkotabi, M. (2015b). Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299.

[Chandrasekaran et al., 2012] Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849.

[Conn et al., 2000] Conn, A. R., Gould, N. I., and Toint, P. L. (2000). *Trust region methods*, volume 1. Siam.

[Engan et al., 1999] Engan, K., Aase, S. O., and Hakon Husoy, J. (1999). Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE.

References III

[Ge et al., 2015] Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842.

[Goldfarb, 1980] Goldfarb, D. (1980). Curvilinear path steplength algorithms for minimization which use directions of negative curvature. *Mathematical programming*, 18(1):31–40.

[Hillar and Lim, 2013] Hillar, C. J. and Lim, L.-H. (2013). Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6):45.

[Lewicki and Sejnowski, 2000] Lewicki, M. S. and Sejnowski, T. J. (2000). Learning overcomplete representations. *Neural computation*, 12(2):337–365.

[Mairal et al., 2014] Mairal, J., Bach, F., and Ponce, J. (2014). Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283.

[Moré and Sorensen, 1983] Moré, J. J. and Sorensen, D. C. (1983). Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing*, 4(3):553–572.

[Mu et al., 2013] Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2013). Square deal: Lower bounds and improved relaxations for tensor recovery. *arXiv preprint arXiv:1307.5870*.

References IV

[Murty and Kabadi, 1987] Murty, K. G. and Kabadi, S. N. (1987). Some np-complete problems in quadratic and nonlinear programming. *Mathematical programming*, 39(2):117–129.

[Negahban et al., 2009] Negahban, S., Yu, B., Wainwright, M. J., and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356.

[Olshausen and Field, 1996] Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

[Oymak et al., 2012] Oymak, S., Jalali, A., Fazel, M., Eldar, Y. C., and Hassibi, B. (2012). Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753*.

[Rendl and Wolkowicz, 1997] Rendl, F. and Wolkowicz, H. (1997). A semidefinite framework for trust region subproblems with applications to large scale minimization. *Mathematical Programming*, 77(1):273–299.

[Shechtman et al., 2015] Shechtman, Y., Eldar, Y. C., Cohen, O., Chapman, H. N., Miao, J., and Segev, M. (2015). Phase retrieval with application to optical imaging: a contemporary overview. *Signal Processing Magazine, IEEE*, 32(3):87–109.

References V

[Spielman et al., 2012] Spielman, D. A., Wang, H., and Wright, J. (2012). Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory*.

[Sun et al., 2015a] Sun, J., Qu, Q., and Wright, J. (2015a). Complete dictionary recovery over the sphere. *arXiv preprint arXiv:1504.06785*.

[Sun et al., 2015b] Sun, J., Qu, Q., and Wright, J. (2015b). A geometric analysis of phase retreival. *In preparation*.

[Vavasis, 2009] Vavasis, S. A. (2009). On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377.

[Zibulevsky and Pearlmutter, 2001] Zibulevsky, M. and Pearlmutter, B. (2001). Blind source separation by sparse decomposition in a signal dictionary. *Neural computation*, 13(4):863–882.