When Are Nonconvex Problems Not Scary?

A few friendly nonconvex optimization problems

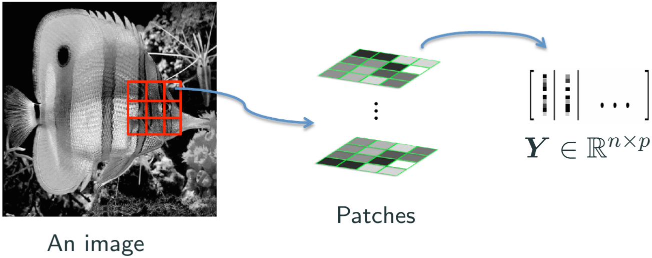
Ju Sun

Electrical Engineering

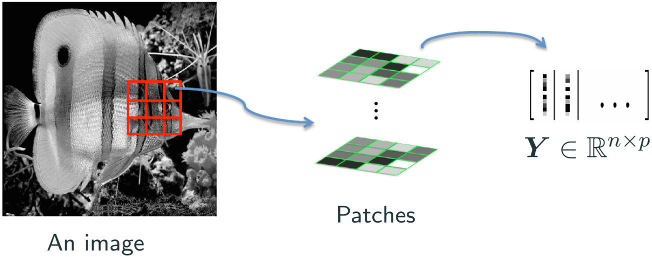
Columbia University

Joint with **Qing Qu**, **John Wright** (Columbia U.)

A curious experiment

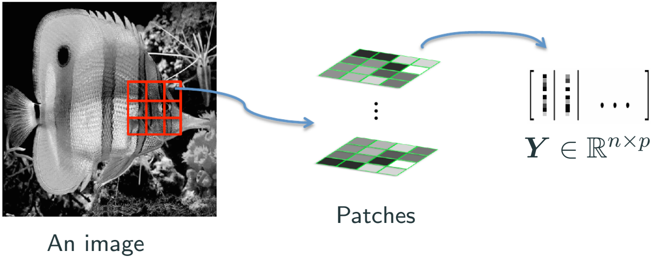


A curious experiment



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

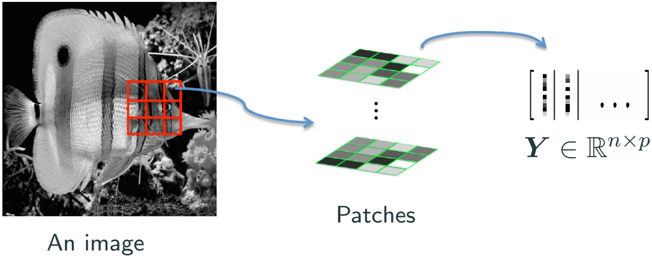
A curious experiment



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

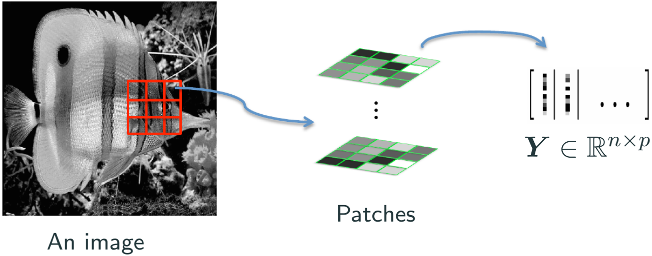
A curious experiment



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

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

A curious experiment



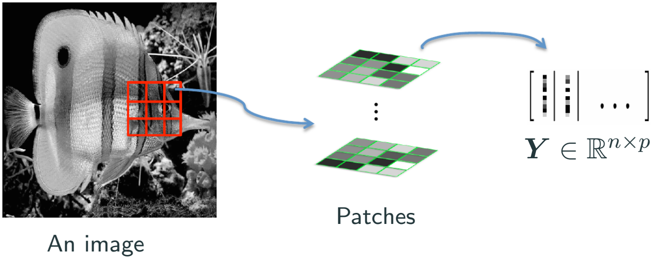
$$\min \quad f(\boldsymbol{Q}, \boldsymbol{X}) \doteq \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_1 , \quad \text{s.t. } \boldsymbol{Q} \in O_n$$

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

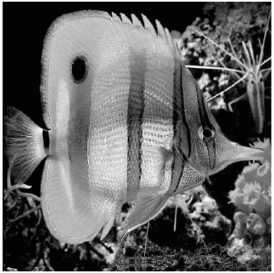


$$\min \quad f(\boldsymbol{Q}, \boldsymbol{X}) \doteq \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_1, \quad \text{s.t. } \boldsymbol{Q} \in O_n$$

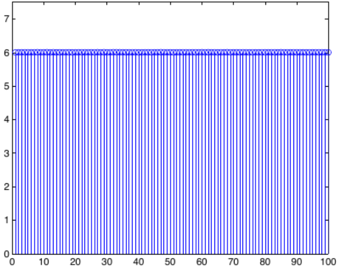
Naive **alternating directions**: starting from a random $\boldsymbol{Q}_0 \in O_n$

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

A curious experiment



An image



Final $f(\boldsymbol{Q}_\infty, \boldsymbol{X}_\infty)$, varying $\boldsymbol{Q}_0$.

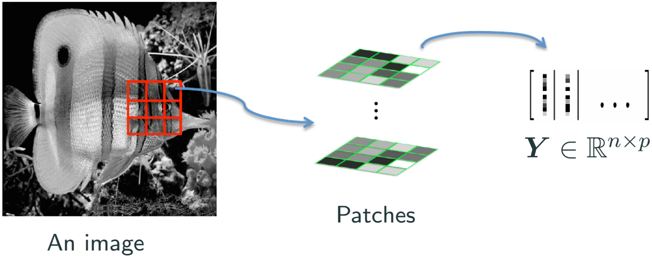
$$\min \quad f(\boldsymbol{Q}, \boldsymbol{X}) \doteq \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_1, \quad \text{s.t. } \boldsymbol{Q} \in O_n$$

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

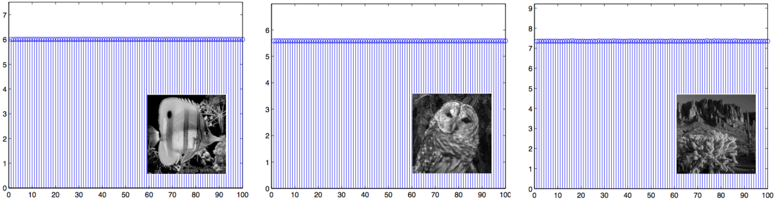
$$\boldsymbol{X}_k = \mathcal{S}_\lambda\left[\boldsymbol{Q}_{k-1}^* \boldsymbol{Y}\right]$$

$$\boldsymbol{Q}_k = \boldsymbol{U}\boldsymbol{V}^*, \text{ where } \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \texttt{SVD}\left(\boldsymbol{Y}\boldsymbol{X}^*\right).$$

Global solutions to feature learning on real images?



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

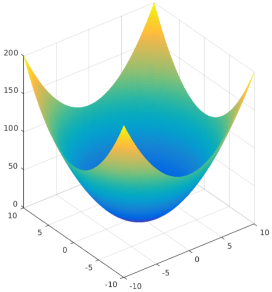


Nonconvex optimization

Many problems in modern **signal processing**, **machine learning**, **statistics**, ..., are most naturally formulated as **nonconvex** optimization problems.

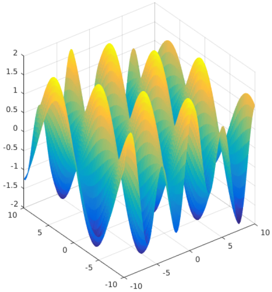
$$\min f(\boldsymbol{x}) \quad \text{s.t. } \boldsymbol{x} \in \mathcal{D}.$$

$f(x, y) = x^2 + y^2$



"easy"

$g(x, y) = \sum_{i=1}^{2} a_i \sin(b_i x + c_i y) + d_i \cos(e_i x + f_i y)$



"hard"

Nonconvex: Even computing a local minimizer is NP-hard!

This work

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 **random/generic**.
- This benign structure allows "**initialization-free**" iterative methods to **efficiently** find a "global" minimizer.

Outline

$\mathcal{X}$ functions

Examples from practical problems

- Sparse (complete) dictionary learning [S., Qu, Wright, '15]
- Generalized phase retrieval [S., Qu, Wright, '16]
- Other examples in the literature

Algorithms: Riemannian trust-region method

Comparison with alternatives

Outline

$\mathcal{X}$ functions

Examples from practical problems

- Sparse (complete) dictionary learning [S., Qu, Wright, '15]
- Generalized phase retrieval [S., Qu, Wright, '16]
- Other examples in the literature

Algorithms: Riemannian trust-region method

Comparison with alternatives

A classical example: the Rayleigh quotient

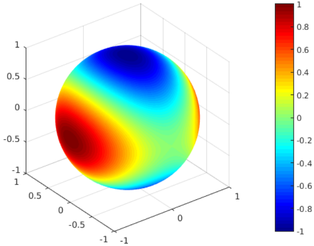
For a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$,

$$\min \ \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} \quad \text{s.t.} \quad \|\boldsymbol{x}\|_2 = 1.$$

Let $\boldsymbol{v}_i$ the eigenvectors of $\boldsymbol{A}$, λ_i the eigenvalues. Suppose

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

- Only **global minimizers** are $\pm \boldsymbol{v}_n$
- 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**.



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

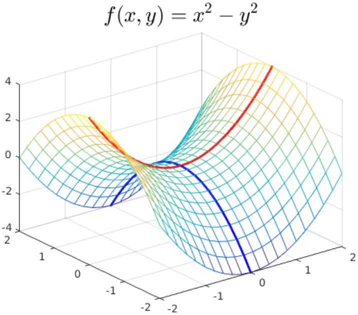
$\mathcal{X}$ functions

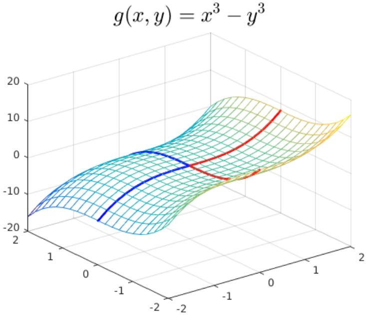
$\mathcal{X}$ **functions** (qualitative version):

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

- (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

$\mathcal{X}$ functions

Examples from practical problems

- Sparse (complete) dictionary learning [S., Qu, Wright, '15]
- Generalized phase retrieval [S., Qu, Wright, '16]
- Other examples in the literature

Algorithms: Riemannian trust-region method

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

$\mathcal{X}$ functions

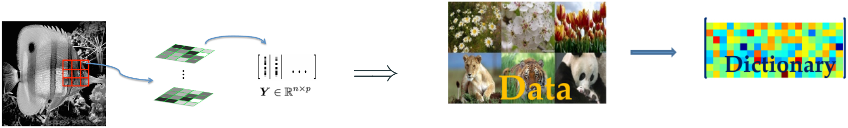
Examples from practical problems

- Sparse (complete) dictionary learning [S., Qu, Wright, '15]
- Generalized phase retrieval [S., Qu, Wright, '16]
- Other examples in the literature

Algorithms: Riemannian trust-region method

Comparison with alternatives

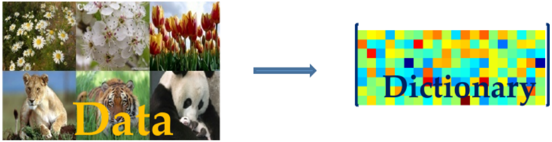
Example I: Sparse Dictionary Learning



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

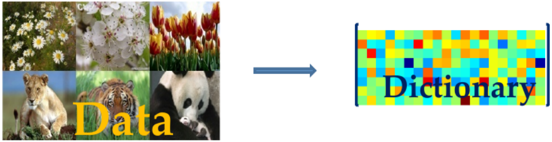
- Algorithmic study initiated in neuroscience [Olshausen and Field, 1996].
- Important algorithmic contributions from many researchers: [Lewicki and Sejnowski, 2000, Engan et al., 1999, Aharon et al., 2006], many others
- Widely used in image processing, visual recognition, compressive signal acquisition, deep architecture for signal classification (see, e.g., [Mairal et al., 2014])

Dictionary recovery - the complete case



Given $\boldsymbol{Y}$ generated as $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

Dictionary recovery - the complete case



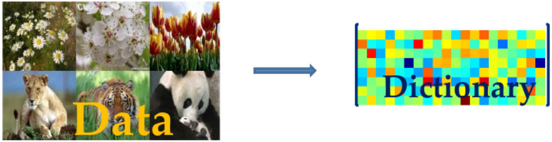
Given $\boldsymbol{Y}$ generated as $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

Random Data Model

$\boldsymbol{Q}_0$ complete (square and invertible)

$\boldsymbol{X}_0$ Bernoulli(θ)-Gaussian: $\boldsymbol{X}_0 = \boldsymbol{\Omega} \odot \boldsymbol{G}, \boldsymbol{\Omega} \sim_{iid} \mathrm{Ber}(\theta), \boldsymbol{G} \sim_{iid} \mathcal{N}(0,1)$.

Dictionary recovery - the complete case



Given $\boldsymbol{Y}$ generated as $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

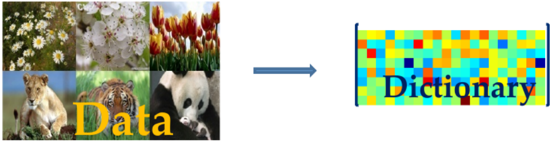
Random Data Model

$\boldsymbol{Q}_0$ complete (square and invertible)

$\boldsymbol{X}_0$ Bernoulli(θ)-Gaussian: $\boldsymbol{X}_0 = \boldsymbol{\Omega} \odot \boldsymbol{G}, \boldsymbol{\Omega} \sim_{iid} \mathrm{Ber}(\theta), \boldsymbol{G} \sim_{iid} \mathcal{N}(0, 1)$.

- $\boldsymbol{Q}_0$ complete $\Longrightarrow$ $\boxed{\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)}$

Dictionary recovery - the complete case



Given $\boldsymbol{Y}$ generated as $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

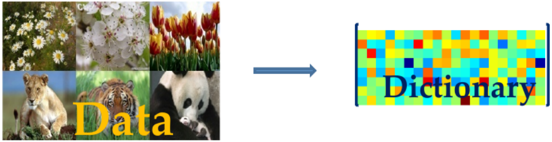
Random Data Model

$\boldsymbol{Q}_0$ complete (square and invertible)

$\boldsymbol{X}_0$ Bernoulli(θ)-Gaussian: $\boldsymbol{X}_0 = \boldsymbol{\Omega} \odot \boldsymbol{G}, \boldsymbol{\Omega} \sim_{iid} \mathrm{Ber}(\theta), \boldsymbol{G} \sim_{iid} \mathcal{N}(0,1)$.

- $\boldsymbol{Q}_0$ complete $\Longrightarrow$ $\boxed{\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)}$
- Rows of $\boldsymbol{X}_0$ are sparse vectors in $\mathrm{row}(\boldsymbol{Y})$

Dictionary recovery - the complete case



Given $\boldsymbol{Y}$ generated as $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

Random Data Model

$\boldsymbol{Q}_0$ complete (square and invertible)

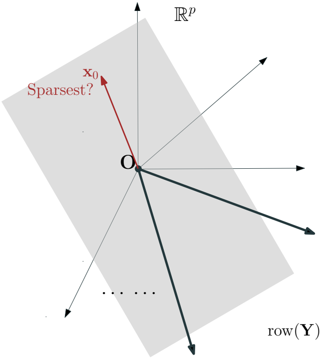
$\boldsymbol{X}_0$ Bernoulli(θ)-Gaussian: $\boldsymbol{X}_0 = \boldsymbol{\Omega} \odot \boldsymbol{G}, \boldsymbol{\Omega} \sim_{iid} \mathrm{Ber}(\theta), \boldsymbol{G} \sim_{iid} \mathcal{N}(0,1)$.

- $\boldsymbol{Q}_0$ complete $\Longrightarrow$ $\boxed{\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)}$
- 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 $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

$\boldsymbol{Q}_0$ square, invertible: $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)$

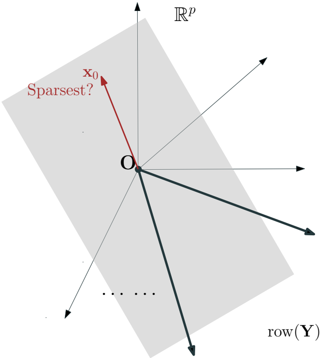


Find the sparsest vectors in $\mathrm{row}(\boldsymbol{Y})$:

Dictionary recovery - the complete case

Dictionary recovery: Given $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

$\boldsymbol{Q}_0$ square, invertible: $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)$



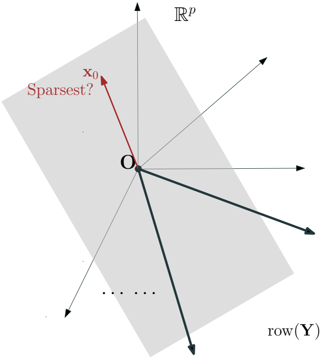
Find the sparsest vectors in $\mathrm{row}(\boldsymbol{Y})$:

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

Dictionary recovery - the complete case

Dictionary recovery: Given $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

$\boldsymbol{Q}_0$ square, invertible: $\mathrm{row}(\boldsymbol{Y}) = \mathrm{row}(\boldsymbol{X}_0)$



Find the sparsest vectors in $\mathrm{row}(\boldsymbol{Y})$:

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

Nonconvex "relaxation":

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

Many precedents, e.g.,

[Zibulevsky and Pearlmutter, 2001] in blind source separation.

Towards geometric understanding

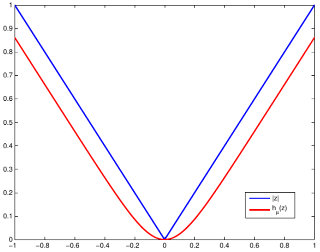
Model problem

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

Smoothed model problem

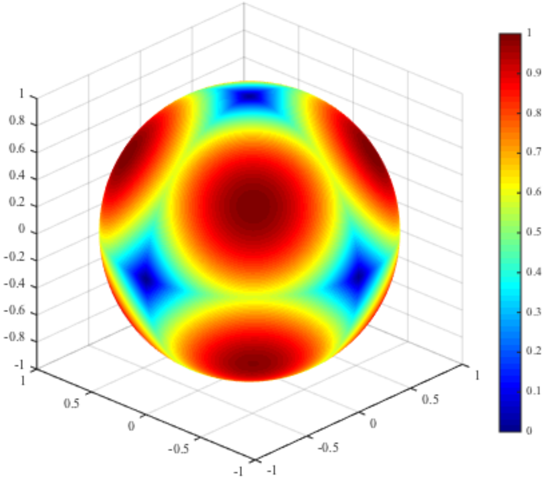
$$\min_{\boldsymbol{q}} \quad f(\boldsymbol{q}) \doteq \frac{1}{p}\sum_{i=1}^{p} h_\mu\left(\boldsymbol{q}^*\boldsymbol{y}_i\right) \quad \text{s.t.}\ \|\boldsymbol{q}\|_2^2 = 1. \quad \boldsymbol{Y} \in \mathbb{R}^{n\times p}$$

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

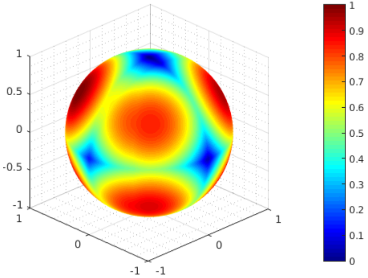


An $\mathcal{X}$ function!

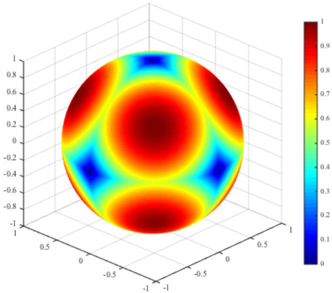
A low-dimensional example ($n = 3$) of the landscape when the target dictionary $\boldsymbol{Q}_0$ is orthogonal and $p \to \infty$



From finite samples



$p = 100$



$p \to \infty$

When $p \sim n^3$ (suppressing $\log$ factors, dependence on μ), the finite sample version is also "nice".

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, S., Qu, Wright '15)

When p is reasonably large, and $\theta \le 1/3$, 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 et al., 2012] $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$
  - [Agarwal et al., 2013b] $\boldsymbol{Q} \in \mathbb{R}^{m \times n}\,(m \le n)$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$
  - [Arora et al., 2013] $\boldsymbol{Q} \in \mathbb{R}^{m \times n}\,(m \le n)$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$
  - [Arora et al., 2015] $\boldsymbol{Q} \in \mathbb{R}^{m \times n}\,(m \le n)$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$
- **Quasipolynomial algorithms** with better guarantees
  - [Arora et al., 2014] different model, $\theta = O\left(1/\mathrm{polylog}\left(n\right)\right)$
  - [Barak et al., 2014] sum-of-squares, $\theta = \tilde{O}\left(1\right)$ polytime for $\theta = O(n^{-\varepsilon})$.
- Other theoretical work on **local geometry**:

  [Gribonval and Schnass, 2010], [Geng and Wright, 2011], [Schnass, 2014], etc

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

Outline

$\mathcal{X}$ functions

Examples from practical problems

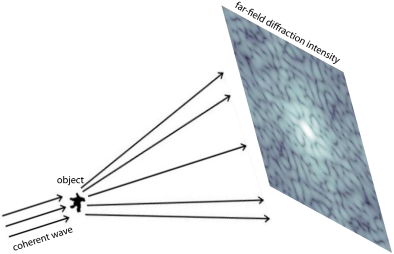
- Sparse (complete) dictionary learning [S., Qu, Wright, '15]
- Generalized phase retrieval [S., Qu, Wright, '16]
- Other examples in the literature

Algorithms: Riemannian trust-region method

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]

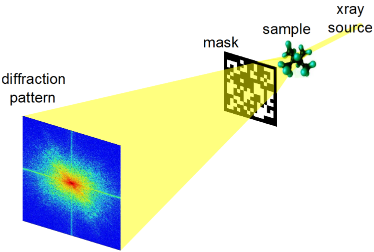
Generalized phase retrieval

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., 2015b]

A nonconvex formulation

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

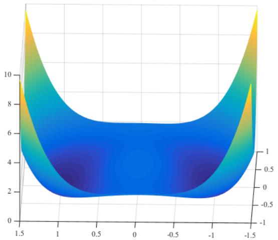
$$\boxed{\min_{\boldsymbol{z} \in \mathbb{C}^n} f(\boldsymbol{z}) \doteq \frac{1}{2m} \sum_{k=1}^{m} (y_k^2 - |\boldsymbol{a}_k^* \boldsymbol{z}|^2)^2.}$$

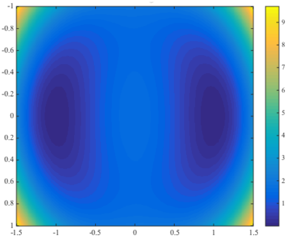
A nonconvex formulation

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

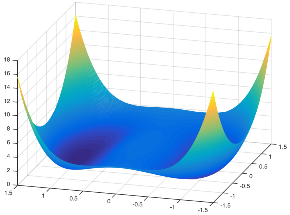
$$\min_{\boldsymbol{z} \in \mathbb{C}^n} f(\boldsymbol{z}) \doteq \frac{1}{2m} \sum_{k=1}^{m} (y_k^2 - |\boldsymbol{a}_k^* \boldsymbol{z}|^2)^2.$$

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





The results



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

Theorem (Informal, S., Qu, Wright '16)

Let $\boldsymbol{a}_k \sim_{\text{iid}} \mathcal{CN}(0,1)$. *When* $m \geq \Omega(n \log^3(n))$, *w.h.p.,*

- *All local (and global) minimizers are of the form* $\boldsymbol{x} e^{i\phi}$.
- f *is* $(c, c/(n \log m), c, c/(n \log m))$*-ridable over* $\mathbb{C}^n$ *for some* $c > 0$.

Comparison with the literature

- **SDP relaxations and their analysis**:

[Candès et al., 2013a]	SDP relaxation
[Candès et al., 2013b]	Guarantees for $m \sim n \log n$, adaptive
[Candès and Li, 2014]	Guarantees for $m \sim n$, non-adaptive
[Candès et al., 2015a]	Coded diffraction patterns
[Waldspurger et al., 2015]	SDP relaxation in phase space

- **Nonconvex methods** (spectral init. + local refinement):

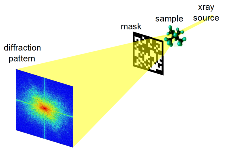
[Netrapalli et al., 2013]	Spectral init. sample splitting
[Candès et al., 2015b]	Spectral init. + gradient descent, $m \sim n \log n$.
[White et al., 2015]	Spectral init. + gradient descent
[Chen and Candès, 2015]	Spectral init. + truncation, $m \sim n$.

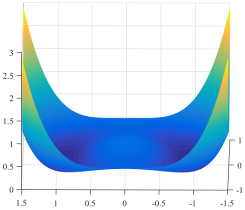
This work: a global characterization of the geometry of the problem. Algorithms succeed independent of initialization, $m \sim n \log^3 n$.

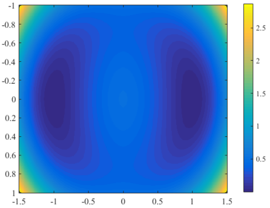
Other measurement models for GPR

Other measurements

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







- Convolutional model (with Yonina Eldar): $\boldsymbol{y} = |\boldsymbol{a} \circledast \boldsymbol{x}|$

Outline

$\mathcal{X}$ functions

Examples from practical problems

- Sparse (complete) dictionary learning [S., Qu, Wright, '15]
- Generalized phase retrieval [S., Qu, Wright, '16]
- Other examples in the literature

Algorithms: Riemannian trust-region method

Comparison with alternatives

Example III: Orthogonal tensor decomposition

... generalizes eigen-decomposition of matrices $\boldsymbol{M} = \sum_{i=1}^{r} \lambda_i \boldsymbol{a}_i \otimes \boldsymbol{a}_i$

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.

Example III: Orthogonal tensor decomposition

... generalizes eigen-decomposition of matrices $\boldsymbol{M} = \sum_{i=1}^{r} \lambda_i \boldsymbol{a}_i \otimes \boldsymbol{a}_i$

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 and permutations).

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

One component each time

Focus on OD tensors of the form

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

One component each time

Focus on OD tensors of the form

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

Consider

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

[Ge et al., 2015] proved that

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

All components in one shot

Focus on OD tensors of the form

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

Consider the "contrast" formulation

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

[Ge et al., 2015] proved that

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

Example IV: Phase synchronization

Synchronization: recovery from **noisy/incomplete** pairwise relative measurements

- angles/phases – from $e^{i(\theta_i - \theta_j)} + \Delta_{ij}$;
- rotations – from $\boldsymbol{R}_i \boldsymbol{R}_j^{-1} + \boldsymbol{\Delta}_{ij}$, $\boldsymbol{R}_i, \boldsymbol{R}_j \in \mathrm{SO}(3)$
- group elements – from $g_i g_j^{-1} + \Delta_{ij}$ for g_i, g_j over a compact group $\mathcal{G}$

Applications: signal reconstruction, computer vision (structure from motion, surface reconstruction), cryo-electron microscopy, digital communications, ranking, ... (see, e.g., [Bandeira et al., 2014, Bandeira et al., 2015])

Example IV: Phase synchronization

Phase synchronization: Let $\boldsymbol{z} \in \mathbb{C}^n$ and $|z_1| = \cdots = |z_n| = 1$.
Given measurements $C_{ij} = z_i \overline{z_j} + \Delta_{ij}$, recover $\boldsymbol{z}$.

In matrix form, $\boldsymbol{C} = \boldsymbol{z}\boldsymbol{z}^* + \boldsymbol{\Delta}$ and assume $\boldsymbol{\Delta}$ Hermitian.

Least-squares formulation:

$$\min_{\boldsymbol{x} \in \mathbb{C}^n} \|\boldsymbol{x}\boldsymbol{x}^* - \boldsymbol{C}\|_F^2, \quad \mathrm{s.\,t.} \quad |x_1| = \cdots = |x_n| = 1.$$

Equivalent to

$$\boxed{\min_{\boldsymbol{x} \in \mathbb{C}^N : |x_1| = \cdots = |x_n| = 1} f(\boldsymbol{u}) \doteq -\boldsymbol{x}^* \boldsymbol{C} \boldsymbol{x}}$$

Quadratic over products of circles

$\boldsymbol{C} = \boldsymbol{z}\boldsymbol{z}^* + \boldsymbol{\Delta}$ and assume $\boldsymbol{\Delta}$ Hermitian

$$\boxed{\min_{\boldsymbol{x} \in \mathbb{C}^N : |x_1| = \cdots = |x_n| = 1} f(\boldsymbol{u}) \doteq -\boldsymbol{x}^* \boldsymbol{C} \boldsymbol{x}}$$

Quadratic over products of circles

$\boldsymbol{C} = \boldsymbol{z}\boldsymbol{z}^* + \boldsymbol{\Delta}$ and assume $\boldsymbol{\Delta}$ Hermitian

$$\boxed{\min_{\boldsymbol{x} \in \mathbb{C}^N : |x_1| = \cdots = |x_n| = 1} f(\boldsymbol{u}) \doteq -\boldsymbol{x}^* \boldsymbol{C} \boldsymbol{x}}$$

[Boumal, 2016] showed when $\boldsymbol{\Delta}$ is "small",

> *second-order necessary conditions for optimality is also sufficient and the global minimizers recover $\boldsymbol{z}$.*

This implies

> *all local minimizers are global; all saddles are ridable.*

Quadratic over products of circles

$\boldsymbol{C} = \boldsymbol{z}\boldsymbol{z}^* + \boldsymbol{\Delta}$ and assume $\boldsymbol{\Delta}$ Hermitian

$$\min_{\boldsymbol{x} \in \mathbb{C}^N : |x_1| = \cdots = |x_n| = 1} f(\boldsymbol{u}) \doteq -\boldsymbol{x}^* \boldsymbol{C} \boldsymbol{x}$$

[Boumal, 2016] showed when $\boldsymbol{\Delta}$ is "small",

> *second-order necessary conditions for optimality is also sufficient and the global minimizers recover $\boldsymbol{z}$.*

This implies

> *all local minimizers are global; all saddles are ridable.*

Analogous results obtained on synchronization over signs and two-block community detection [Bandeira et al., 2016].

Outline

$\mathcal{X}$ functions

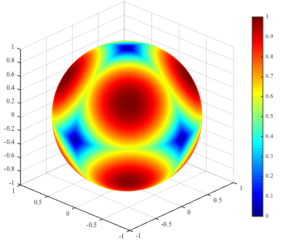
Examples from practical problems

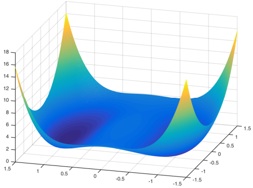
- Sparse (complete) dictionary learning [S., Qu, Wright, '15]
- Generalized phase retrieval [S., Qu, Wright, '16]
- Other examples in the literature

Algorithms: Riemannian trust-region method

Comparison with alternatives

Benign structure

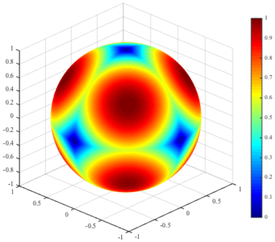


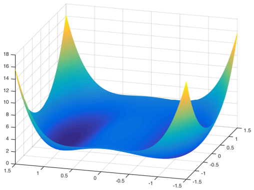


- (P-1) All local minimizers are also global,
- (P-2A) For all local minimizers, $\operatorname{Hess} f \succ \mathbf{0}$, and
- (P-2B) For all other critical points, $\lambda_{\min}(\operatorname{Hess} f) < 0$.

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

Algorithmic possibilities





- **Second-order trust-region method** (described here, [Conn et al., 2000], [Nesterov and Polyak, 2006])
- Curvilinear search [Goldfarb, 1980]
- Noisy/stochastic gradient descent [Ge et al., 2015]
- ...

Second-order methods 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

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

Guaranteed decrease in f when **movement is small** such that the **approximation is reasonably good**.

Trust-region method - Euclidean Space

Generate iterates $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ by

- Forming a second order approximation of the objective $f(\boldsymbol{x})$ about $\boldsymbol{x}_k$:

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

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

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

- Next iterate is $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{\delta}_\star$.

Can choose $\boldsymbol{B}_k = \nabla^2 f(\boldsymbol{x}^{(k)})$ or an approximation.

The trust-region subproblem

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

- QCQP, but can be solved in polynomial time by:

  Root finding [Moré and Sorensen, 1983]

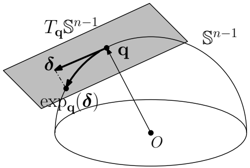
  SDP relaxation [Rendl and Wolkowicz, 1997].

- In practice, only need an approximate solution (with controllable quality) to ensure convergence.

Trust-region method - Riemannian Manifold

Local quadratic approximation:

$$f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) = \underbrace{f(\boldsymbol{q}) + \boldsymbol{\delta}^* \operatorname{grad} f(\boldsymbol{q}) + \frac{1}{2}\boldsymbol{\delta}^* \operatorname{Hess} f(\boldsymbol{q})\boldsymbol{\delta}}_{\doteq \widehat{f}(\boldsymbol{\delta};\boldsymbol{q})} + O(\|\boldsymbol{\delta}\|^3).$$



Basic **Riemannian trust-region method**:

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

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

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

Proof of convergence

- Strong gradient or negative curvature
  $\Longrightarrow$ at least a fixed reduction in $f(\boldsymbol{x})$ at each iteration
- Strong convexity near a local minimizer
  $\Longrightarrow$ quadratic convergence $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\| \le c \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|^2$.

Theorem (Very informal)

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

Worked out examples in [Sun et al., 2015, Sun et al., 2016];

See also promise of 1-st order method [Ge et al., 2015, Lee et al., 2016].

Outline

$\mathcal{X}$ functions

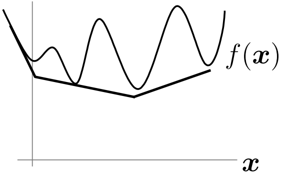
Examples from practical problems

- Sparse (complete) dictionary learning [S., Qu, Wright, '15]
- Generalized phase retrieval [S., Qu, Wright, '16]
- Other examples in the literature

Algorithms: Riemannian trust-region method

Comparison with alternatives

Convexification – a recipe



Find a **tractable convex surrogate** $\widehat{f}$ for f.
Minimize the surrogate $\widehat{f}$
Prove that for well structured instances
the solution is accurate.

Separates formulations/analysis from algorithms

Beautiful mathematical results, substantial applied impact:

- 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 ...

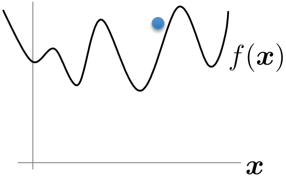
Tensor recovery	[Hillar and Lim, 2013]
Nonnegative low-rank approximation	[Vavasis, 2009]

... or may not work as well as we might hope.

Simultaneous structure estimation	[Oymak et al., 2012]
Tensor recovery	[Mu et al., 2014]
Sparse PCA	[Berthet and Rigollet, 2013]
Dictionary learning	[Spielman et al., 2012]

Substantial and provable gaps between the performance of known convex relaxations and the information theoretic optimum.

Prior work: proving nonconvex recovery



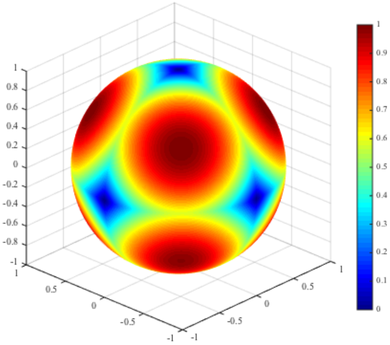
Use problem structure to find a
clever initial guess.

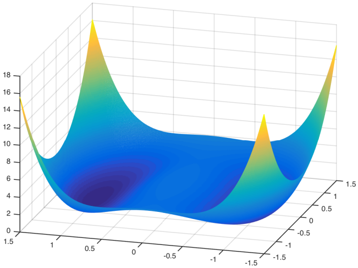
Analyze iteration-by-iteration
in the vicinity of the optimum.

- **Matrix completion/recovery**: [Keshavan et al., 2010], [Jain et al., 2013], [Hardt, 2014], [Hardt and Wootters, 2014], [Netrapalli et al., 2014], [Jain and Netrapalli, 2014], [Sun and Luo, 2014], [Zheng and Lafferty, 2015], [Tu et al., 2015], [Chen and Wainwright, 2015], [Sa et al., 2015], [Wei et al., 2015]. Also [Jain et al., 2010]
- **Dictionary learning**: [Agarwal et al., 2013a], [Arora et al., 2013], [Agarwal et al., 2013b], [Arora et al., 2015]
- **Tensor recovery**: [Jain and Oh, 2014], [Anandkumar et al., 2014c], [Anandkumar et al., 2014b], [Anandkumar et al., 2015]
- **Phase retrieval**: [Netrapalli et al., 2013], [Candès et al., 2015b], [Chen and Candès, 2015], [White et al., 2015]
- **More on the webpage:** `http://sunju.org/research/nonconvex/`

See also [Loh and Wainwright, 2011]

This work





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

Thanks to ...

John Wright

Columbia

Qing Qu

Columbia

A Geometric Analysis of Phase Retrieval, **S.**, Qu, Wright, '16

Complete Dictionary Recovery over the Sphere, **S.**, Qu, Wright, '15

When are Nonconvex Optimization Problems Not Scary, **S.**, Qu, Wright, NIPS Workshop, '15

Finding a Sparse Vector in a Subspace: Linear Sparsity Using Alternating Directions, Qu, **S.**, Wright, '15

Webpage on provable nonconvex heuristics:

`http://sunju.org/research/nonconvex/`

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.

[Agarwal et al., 2013a] Agarwal, A., Anandkumar, A., Jain, P., Netrapalli, P., and Tandon, R. (2013a). **Learning sparsely used overcomplete dictionaries via alternating minimization.** *arXiv preprint arXiv:1310.7991*.

[Agarwal et al., 2013b] Agarwal, A., Anandkumar, A., and Netrapalli, P. (2013b). **Exact recovery of sparsely used overcomplete dictionaries.** *arXiv preprint arXiv:1309.1952*.

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

[Anandkumar et al., 2014b] Anandkumar, A., Ge, R., and Janzamin, M. (2014b). **Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models.** *arXiv preprint arXiv:1411.1488*.

[Anandkumar et al., 2014c] Anandkumar, A., Ge, R., and Janzamin, M. (2014c). **Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates.** *arXiv preprint arXiv:1402.5180*.

[Anandkumar et al., 2015] Anandkumar, A., Jain, P., Shi, Y., and Niranjan, U. N. (2015). **Tensor vs matrix methods: Robust tensor decomposition under block sparse perturbations.** *arXiv preprint arXiv:1510.04747*.

[Arora et al., 2014] Arora, S., Bhaskara, A., Ge, R., and Ma, T. (2014). **More algorithms for provable dictionary learning.** *arXiv preprint arXiv:1401.0579*.

References II

[Arora et al., 2015] Arora, S., Ge, R., Ma, T., and Moitra, A. (2015). **Simple, efficient, and neural algorithms for sparse coding.** *arXiv preprint arXiv:1503.00778*.

[Arora et al., 2013] Arora, S., Ge, R., and Moitra, A. (2013). **New algorithms for learning incoherent and overcomplete dictionaries.** *arXiv preprint arXiv:1308.6273*.

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

[Bandeira et al., 2014] Bandeira, A. S., Boumal, N., and Singer, A. (2014). **Tightness of the maximum likelihood semidefinite relaxation for angular synchronization.** *arXiv preprint arXiv:1411.3272*.

[Bandeira et al., 2016] Bandeira, A. S., Boumal, N., and Voroninski, V. (2016). **On the low-rank approach for semidefinite programs arising in synchronization and community detection.** *arXiv preprint arXiv:1602.04426*.

[Bandeira et al., 2015] Bandeira, A. S., Chen, Y., and Singer, A. (2015). **Non-unique games over compact groups and orientation estimation in cryo-em.** *arXiv preprint arXiv:1505.03840*.

[Barak et al., 2014] Barak, B., Kelner, J. A., and Steurer, D. (2014). **Dictionary learning and tensor decomposition via the sum-of-squares method.** *arXiv preprint arXiv:1407.1543*.

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

[Boumal, 2016] Boumal, N. (2016). **Nonconvex phase synchronization.** *arXiv preprint arXiv:1601.06114*.

[Candès et al., 2013a] Candès, E. J., Eldar, Y. C., Strohmer, T., and Voroninski, V. (2013a). **Phase retrieval via matrix completion.** *SIAM Journal on Imaging Sciences*, 6(1).

[Candès and Li, 2014] Candès, E. J. and Li, X. (2014). **Solving quadratic equations via phaselift when there are about as many equations as unknowns.** *Foundations of Computational Mathematics*, 14(5):1017–1026.

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

References III

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

[Candès et al., 2013b] Candès, E. J., Strohmer, T., and Voroninski, V. (2013b). **Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming.** *Communications on Pure and Applied Mathematics*, 66(8):1241–1274.

[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.

[Chen and Candès, 2015] Chen, Y. and Candès, E. J. (2015). **Solving random quadratic systems of equations is nearly as easy as solving linear systems.** *arXiv preprint arXiv:1505.05114*.

[Chen and Wainwright, 2015] Chen, Y. and Wainwright, M. J. (2015). **Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees.** *arXiv preprint arXiv:1509.03025*.

[Conn et al., 2000] Conn, A. R., Gould, N. I. M., and Toint, P. L. (2000). **Trust-region Methods.** Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[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.

[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.

[Geng and Wright, 2011] Geng, Q. and Wright, J. (2011). **On the local correctness of ℓ^1-minimization for dictionary learning.** Submitted to *IEEE Transactions on Information Theory*. Preprint: `http://www.columbia.edu/~jw2966`.

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

References IV

[Gribonval and Schnass, 2010] Gribonval, R. and Schnass, K. (2010). **Dictionary identification - sparse matrix-factorization via ℓ^1-minimization.** *IEEE Transactions on Information Theory*, 56(7):3523–3539.

[Hardt, 2014] Hardt, M. (2014). **Understanding alternating minimization for matrix completion.** In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 651–660. IEEE.

[Hardt and Wootters, 2014] Hardt, M. and Wootters, M. (2014). **Fast matrix completion without the condition number.** In *Proceedings of The 27th Conference on Learning Theory*, pages 638–678.

[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.

[Jain et al., 2010] Jain, P., Meka, R., and Dhillon, I. S. (2010). **Guaranteed rank minimization via singular value projection.** In *Advances in Neural Information Processing Systems*, pages 937–945.

[Jain and Netrapalli, 2014] Jain, P. and Netrapalli, P. (2014). **Fast exact matrix completion with finite samples.** *arXiv preprint arXiv:1411.1087*.

[Jain et al., 2013] Jain, P., Netrapalli, P., and Sanghavi, S. (2013). **Low-rank matrix completion using alternating minimization.** In *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, pages 665–674. ACM.

[Jain and Oh, 2014] Jain, P. and Oh, S. (2014). **Provable tensor factorization with missing data.** In *Advances in Neural Information Processing Systems*, pages 1431–1439.

[Keshavan et al., 2010] Keshavan, R. H., Montanari, A., and Oh, S. (2010). **Matrix completion from a few entries.** *Information Theory, IEEE Transactions on*, 56(6):2980–2998.

[Lee et al., 2016] Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016). **Gradient descent converges to minimizers.** *arXiv preprint arXiv:1602.04915*.

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

References V

[Loh and Wainwright, 2011] Loh, P.-L. and Wainwright, M. J. (2011). **High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity.** In *Advances in Neural Information Processing Systems*, pages 2726–2734.

[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., 2014] Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2014). **Square deal: Lower bounds and improved convex relaxations for tensor recovery.** *Journal of Machine Learning Research*, 1:1–48.

[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.

[Nesterov and Polyak, 2006] Nesterov, Y. and Polyak, B. T. (2006). **Cubic regularization of newton method and its global performance.** *Mathematical Programming*, 108(1):177–205.

[Netrapalli et al., 2013] Netrapalli, P., Jain, P., and Sanghavi, S. (2013). **Phase retrieval using alternating minimization.** In *Advances in Neural Information Processing Systems*, pages 2796–2804.

[Netrapalli et al., 2014] Netrapalli, P., Niranjan, U. N., Sanghavi, S., Anandkumar, A., and Jain, P. (2014). **Non-convex robust PCA.** In *Advances in Neural Information Processing Systems*, pages 1107–1115.

[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.

References VI

[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.

[Sa et al., 2015] Sa, C. D., Re, C., and Olukotun, K. (2015). **Global convergence of stochastic gradient descent for some non-convex matrix problems.** In *The 32nd International Conference on Machine Learning*, volume 37, pages 2332–2341.

[Schnass, 2014] Schnass, K. (2014). **Local identification of overcomplete dictionaries.** *arXiv preprint arXiv:1401.6354*.

[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.

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

[Sun et al., 2016] Sun, J., Qu, Q., and Wright, J. (2016). **A geometric analysis of phase retreival.** *arXiv preprint arXiv:1602.06664*.

[Sun and Luo, 2014] Sun, R. and Luo, Z.-Q. (2014). **Guaranteed matrix completion via non-convex factorization.** *arXiv preprint arXiv:1411.8003*.

[Tu et al., 2015] Tu, S., Boczar, R., Soltanolkotabi, M., and Recht, B. (2015). **Low-rank solutions of linear matrix equations via procrustes flow.** *arXiv preprint arXiv:1507.03566*.

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

References VII

[Waldspurger et al., 2015] Waldspurger, I., d`Aspremont, A., and Mallat, S. (2015). **Phase recovery, maxcut and complex semidefinite programming.** *Mathematical Programming*, 149(1-2):47–81.

[Wei et al., 2015] Wei, K., Cai, J.-F., Chan, T. F., and Leung, S. (2015). **Guarantees of Riemannian optimization for low rank matrix recovery.** *arXiv preprint arXiv:1511.01562*.

[White et al., 2015] White, C. D., Ward, R., and Sanghavi, S. (2015). **The local convexity of solving quadratic equations.** *arXiv preprint arXiv:1506.07868*.

[Zheng and Lafferty, 2015] Zheng, Q. and Lafferty, J. (2015). **A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements.** *arXiv preprint arXiv:1506.06081*.

[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.