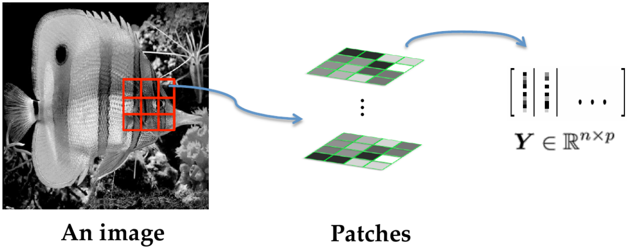
Complete Dictionary Recovery over the Sphere

Ju Sun

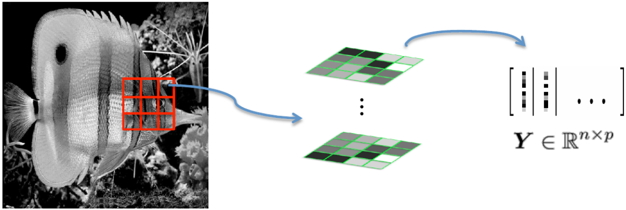
[joint with **Qing Qu** and **Prof. John Wright**]
Electrical Engineering, Columbia University

July 6th, 2015
SPARS 2015, Cambridge, UK

A curious experiment



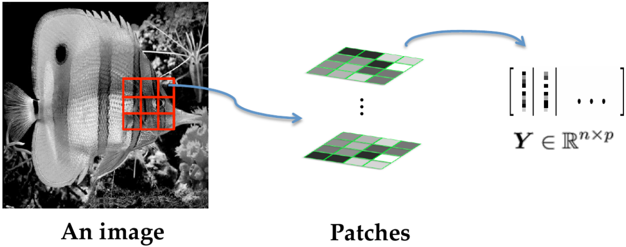
A curious experiment



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.

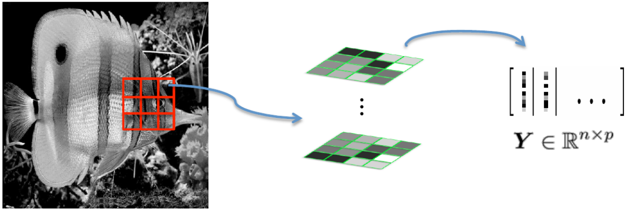
A curious experiment



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

A curious experiment

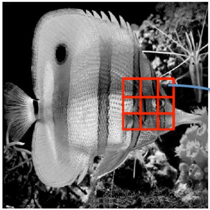


An image **Patches**

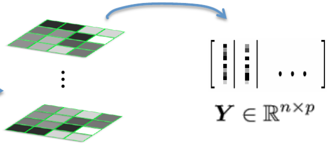
$$\min \quad f(\boldsymbol{Q}, \boldsymbol{X}) \doteq \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_1, \quad \text{s.t. } \boldsymbol{Q} \in O_n.$$

- 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

A curious experiment



An image



Patches

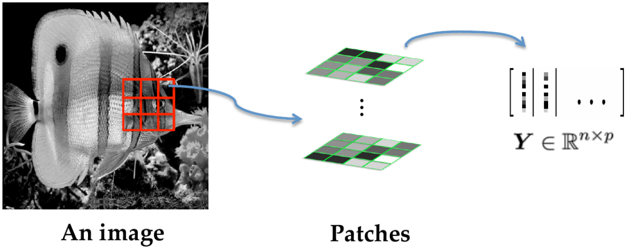
$$\min \quad f\left(\boldsymbol{Q}, \boldsymbol{X}\right) \doteq \frac{1}{2}\left\|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\right\|_F^2 + \lambda\left\|\boldsymbol{X}\right\|_1, \quad \text{s.t. } \boldsymbol{Q} \in O_n$$

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

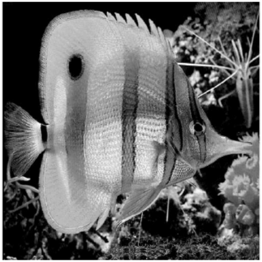


$$\min \quad f(\boldsymbol{Q}, \boldsymbol{X}) \doteq \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_1, \quad \text{s.t. } \boldsymbol{Q} \in O_n$$

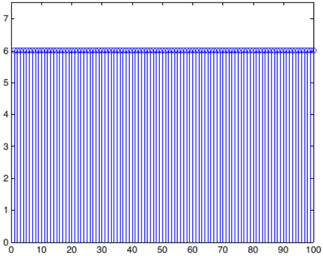
Apply the naive **alternating directions**: starting from a random $\boldsymbol{Q}_0 \in O_n$

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

A curious experiment



An image



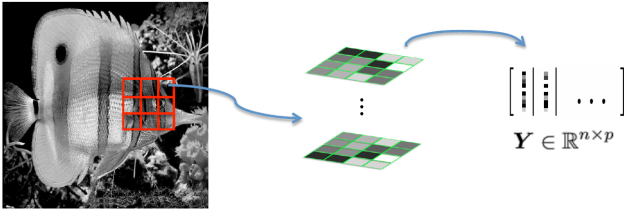
Final $f(\boldsymbol{Q}_\infty, \boldsymbol{X}_\infty)$, varying $\boldsymbol{Q}_0$.

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

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

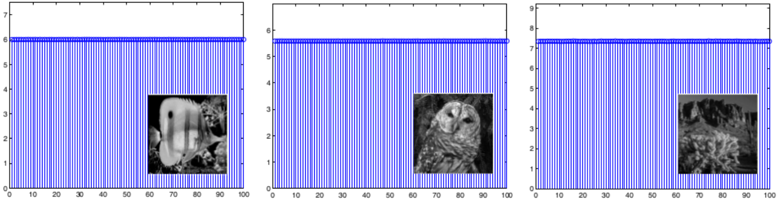
$$\boxed{\begin{aligned} \boldsymbol{X}_k &= \mathcal{S}_\lambda\left[\boldsymbol{Q}_{k-1}^*\boldsymbol{Y}\right] \\ \boldsymbol{Q}_k &= \boldsymbol{U}\boldsymbol{V}^*, \text{ where } \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \texttt{SVD}\left(\boldsymbol{Y}\boldsymbol{X}^*\right). \end{aligned}}$$

Global solutions of feature learning on real images?



An image **Patches**

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



Working hypothesis

Certain nonconvex optimization problems become tractable when the input data are large and random (generic).

A Geometric Approach

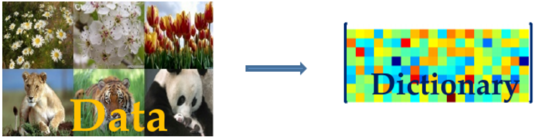
Geometry of the function landscape determines algorithm design and analysis.

... starting with *sparse dictionary learning*!

This work

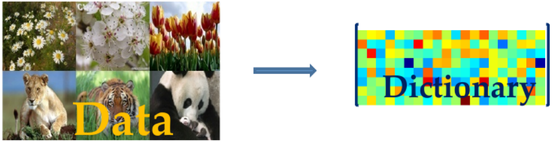
- Sparse dictionary learning: problem formulation
- Main result: **dictionary learning with proportional sparsity**
    - A nonconvex formulation
    - A glimpse into high-dimensional geometry
    - A Riemannian trust-region method and efficiency guarantees

Sparse dictionary learning



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

- Algorithmic study initialized with [Olshausen, Field. '96] in neuroscience.
- Important algorithmic contributions from many researchers: [Lewicki, Sejnowski.'99], [Engan et al. '99], [Aharon, Elad, Bruckstein. '06], many others
- Widely used in image processing, recently used in visual recognition, compressive signal acquisition, deep architecture for signal classification



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

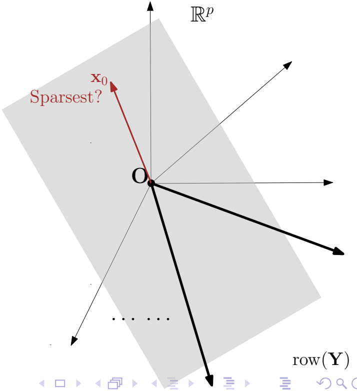
Dictionary recovery – given $\boldsymbol{Y}$ generated as $\boldsymbol{Y} = \boldsymbol{Q}_0 \boldsymbol{X}_0$, recover $\boldsymbol{Q}_0$ and $\boldsymbol{X}_0$.

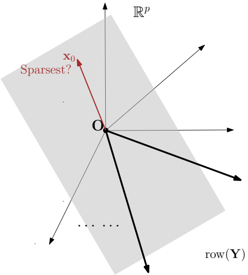
Our Model

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

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

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





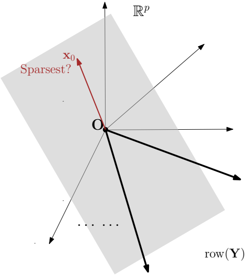
$$\boxed{\min \quad \|\boldsymbol{q}^* \boldsymbol{Y}\|_0 \quad \text{s.t. } \boldsymbol{q} \neq \mathbf{0}.}$$

- Convex relaxation:

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

  - Solve a sequence of convex (linear) programs.
  - Provably succeeds when $\theta n = O\left(\sqrt{n}\right)$, provably fails if $\theta n = \Omega\left(\sqrt{n \log n}\right)$ [Spielman, Wang, Wright.'12].

Dictionary learning: the complete case



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

- 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-Perlmutter, '01] in source separation.

Towards geometric understanding

Model problem

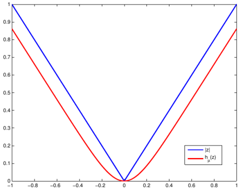
$$\min \quad \frac{1}{p}\left\|\boldsymbol{q}^*\boldsymbol{Y}\right\|_1 = \frac{1}{p}\sum_{i=1}^{p}\left|\boldsymbol{q}^*\boldsymbol{y}_i\right| \quad \text{s.t.} \quad \|\boldsymbol{q}\|_2^2 = 1. \quad \boldsymbol{Y} \in \mathbb{R}^{n\times p}$$

Slightly modified model problem

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

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

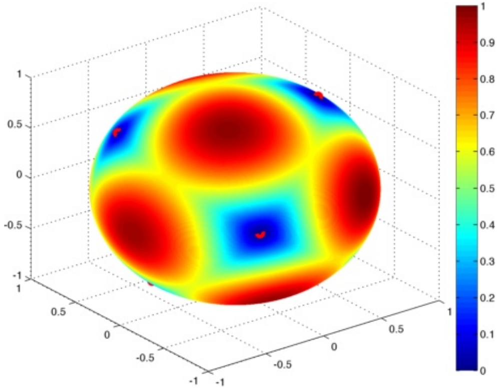
$$h_\mu\left(z\right) = \mu \log \frac{e^{z/\mu} + e^{-z/\mu}}{2} = \mu \log\cosh\frac{z}{\mu}$$

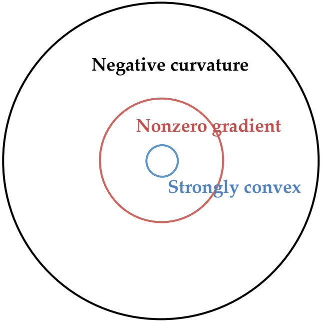


- Recognize the objective as a *normalized sum of independent random variables* $\longrightarrow$ expectation, asymptotically

Why might this work?

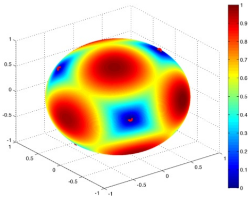
A low-dimensional example ($n = 3$) of the landscape when $p \to \infty$

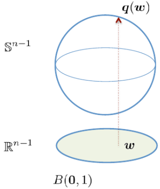


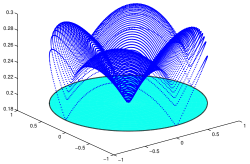


Every local minimizer is a target!

In high dimensions ... details





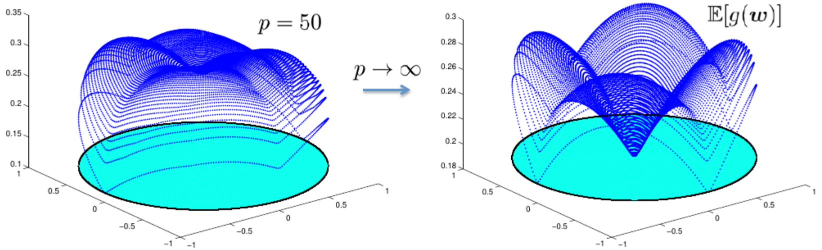


Assuming $A_0 = I$...
Proposition: Suppose $\theta \in \left(0, \frac{1}{2}\right)$, and $\mu < \min\left\{cn^{-5/4}, c'\theta n^{-1}\right\}$.
Then

- $\nabla^2 \mathbb{E}\left[g\left(w\right)\right] \succeq c_\star \frac{\theta}{\mu} I$, for $\|w\| \le \frac{\mu}{4\sqrt{2}}$
- $\frac{w^* \nabla \mathbb{E}[g(w)]}{\|w\|} \ge c_\star \theta$, for $\frac{\mu}{4\sqrt{2}} \le \|w\| \le R$
- $\frac{w^* \nabla \mathbb{E}[g(w)] w}{\|w\|^2} \le -c_\star \theta$, for $\|w\| \ge R \; (R = \Theta(1))$.

and so, **every local minimizer of** $\mathbb{E}\left[g\left(w\right)\right]$ **is a target point**.

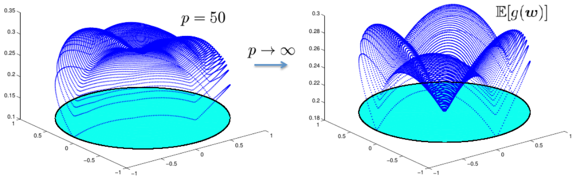
Convergence in function landscape



When does the finite-sample objective **converge** to the asymptotic one, in optimization sense?

...informally, is the function geometry “nice” for some large yet finite p?

Finite-sample result



- Objective $g(\boldsymbol{w}) = \frac{1}{p}\sum_{i=1}^{p} h_\mu\left(q(\boldsymbol{w})^* \boldsymbol{x}_i\right)$ is a sum of independent RVs.
- The proof follows a typical expectation-concentration path

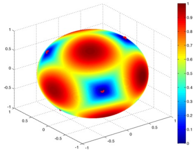
Theorem

Suppose $\theta \in \left(0, \frac{1}{2}\right)$, *if* $\mu < \min\{cn^{-5/4}, c'n^{-1}\}$, *and* $p \geq \frac{Cn^3}{\mu^2\theta^2} \log \frac{n}{\mu\theta}$, *it* ***holds uniformly w.h.p.*** *that*

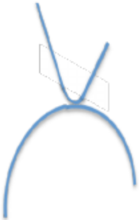
- $\nabla^2 g(\boldsymbol{w}) \succeq c_\star \frac{\theta}{\mu} \boldsymbol{I}$, *for* $\|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}$
- $\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w})}{\|\boldsymbol{w}\|} \geq c_\star \theta$, *for* $\frac{\mu}{4\sqrt{2}} \leq \|\boldsymbol{w}\| \leq R$
- $\frac{\boldsymbol{w}^* \nabla g(\boldsymbol{w}) \boldsymbol{w}}{\|\boldsymbol{w}\|^2} \leq -c_\star \theta$, *for* $\|\boldsymbol{w}\| \geq R$ $(R = \Theta(1))$.

Algorithm design

... following intuition we build from the geometry:



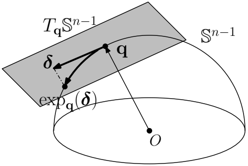
- No spurious local minimizers $\Longrightarrow$ Run **descent algorithm** with **an arbitrary initial point**



- Need to escape saddle points $\Longrightarrow$ Use second-order information. Here, via the **trust region method**.

Trust-region on manifolds [Absil, Baker, Gallivan. '07], also [Absil, Mahoney, Sepulchre. '08]

Trust region method - Riemannian Manifold



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

$$\exp_{\boldsymbol{q}}(\boldsymbol{\delta}) = \boldsymbol{q}\cos\|\boldsymbol{\delta}\| + \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|}\sin\|\boldsymbol{\delta}\|$$

- Consider a function $f : \mathbb{S}^{n-1} \mapsto \mathbb{R}$. For $\boldsymbol{q} \in \mathbb{S}^{n-1}$ and $\boldsymbol{\delta} \in T_{\boldsymbol{q}}\mathbb{S}^{n-1}$. Define the function $f_{\boldsymbol{q}} : T_{\boldsymbol{q}}\mathbb{S}^{n-1} \mapsto \mathbb{R}$ as $f_{\boldsymbol{q}} \doteq f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta}))$.
- Taylor's theorem implies

$$\begin{aligned} f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) &= f(\boldsymbol{q}) + \langle \boldsymbol{\delta}, \nabla f(\boldsymbol{q})\rangle + \frac{1}{2}\boldsymbol{\delta}^*(\nabla^2 f(\boldsymbol{q}) - \langle \boldsymbol{q}, \nabla f(\boldsymbol{q})\rangle \boldsymbol{I})\boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|^3) \\ &\doteq f_{\boldsymbol{q}}(\boldsymbol{\delta};\boldsymbol{q}) + O(\|\boldsymbol{\delta}\|^3). \end{aligned}$$

- Basic **Riemannian trust-region method**:

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

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

Pull things together

Theorem (informal, exact recovery for complete dictionaries)

For any $\theta \in (0, 1/3)$, *given* $\boldsymbol{Y} = \boldsymbol{A}_0 \boldsymbol{X}_0$ *with* $\boldsymbol{A}_0$ *a complete dictionary and* $\boldsymbol{X}_0 \sim_{i.i.d.} \mathrm{BG}(\theta)$, *there is a polynomial-time algorithm that recovers* $\boldsymbol{A}_0$ *and* $\boldsymbol{X}_0$ *with high probability (at least* $1 - O(p^{-6})$*) whenever* $p \geq p_\star(n, 1/\theta, \kappa(\boldsymbol{A}_0), 1/\mu)$ *for a fixed polynomial* $p_\star(\cdot)$*, where* $\kappa(\boldsymbol{A}_0)$ *is the condition number of* $\boldsymbol{A}_0$ *and* μ *is the smoothing parameter.*

Comparison with the 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]	$\boldsymbol{Q} \in \mathbb{R}^{m \times n}$ $(m \le n)$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$
[Arora, Ge, Moitra,'13]	$\boldsymbol{Q} \in \mathbb{R}^{m \times n}$ $(m \le n)$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$
[Arora, Ge, Ma, Moitra,'15]	$\boldsymbol{Q} \in \mathbb{R}^{m \times n}$ $(m \le n)$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$

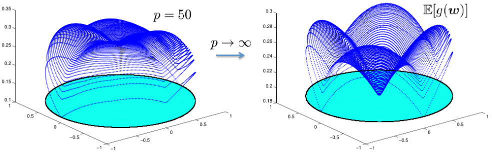
- **Quasipolynomial algorithms** with better guarantees

[Arora, Bhaskara, Ge, Ma,'14]	different prob. model, $\theta = O\left(1/\mathrm{polylog}\left(n\right)\right)$
[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], [Schnass'15]

This work: a polynomial algorithm for squared $\boldsymbol{Q}$, $\theta = O(1)$.

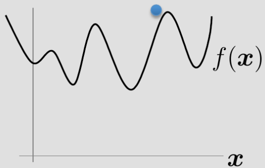
What we have done ...



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

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

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]. 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]

Questions?