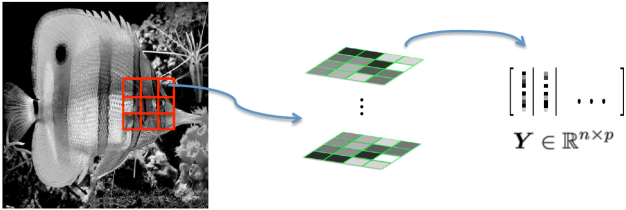
When Nonconvex Optimization Meets Big Data

Ju Sun

[Advisor: Prof. John Wright]

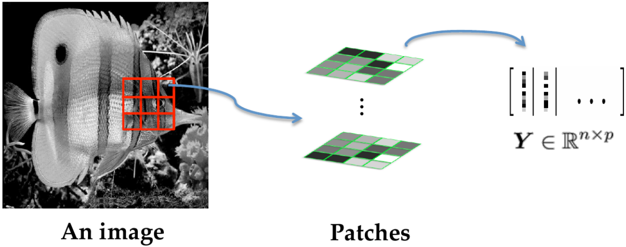
November 21, 2014

A curious experiment



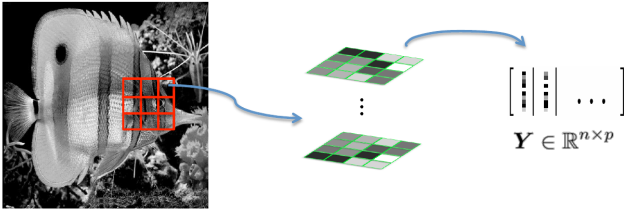
An image **Patches**

A curious experiment



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

A curious experiment

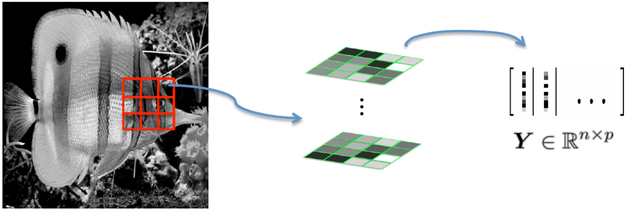


An image

Patches

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

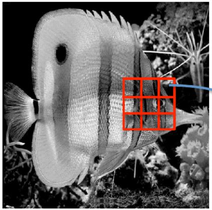


An image **Patches**

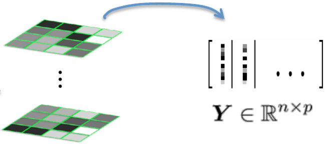
$$\min \quad f(\boldsymbol{Q}, \boldsymbol{X}) \doteq \frac{1}{2}\left\|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\right\|_F^2 + \lambda \left\|\boldsymbol{X}\right\|_1, \quad \text{s.t. } \boldsymbol{Q} \in O(n).$$

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

A curious experiment



An image



Patches

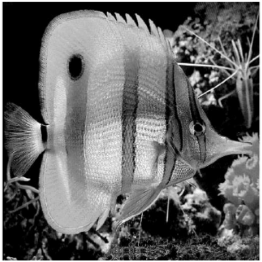
$$\min \quad f(\boldsymbol{Q}, \boldsymbol{X}) \doteq \frac{1}{2} \|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_1, \quad \text{s.t. } \boldsymbol{Q} \in O(n)$$

Apply the naive **alternating directions**: starting from a random $\boldsymbol{Q}_0 \in O(n)$

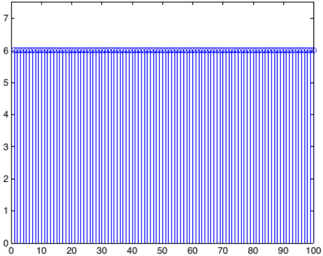
$$\boldsymbol{X}_k = \arg\min_{\boldsymbol{X}} f\left(\boldsymbol{Q}_{k-1}, \boldsymbol{X}\right)$$

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

A curious experiment



An image



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

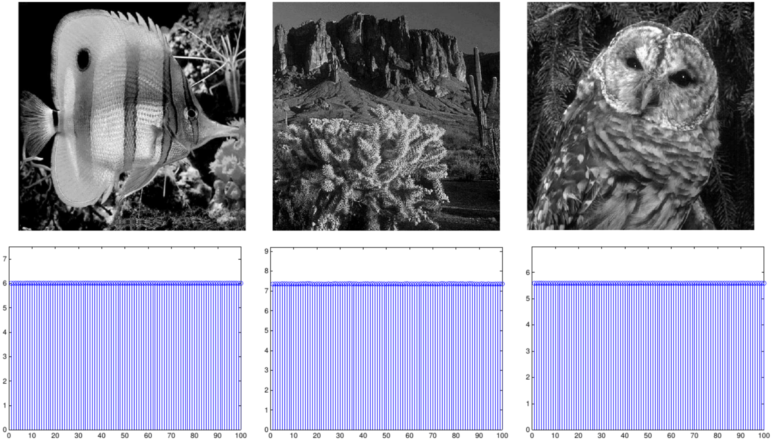
$$\min \quad f(\boldsymbol{Q}, \boldsymbol{X}) \doteq \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Q}\boldsymbol{X}\|_F^2 + \lambda \|\boldsymbol{X}\|_1, \quad \text{s.t. } \boldsymbol{Q} \in O(n)$$

Apply the naive **alternating directions**: starting from a random $\boldsymbol{Q}_0 \in O(n)$

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

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

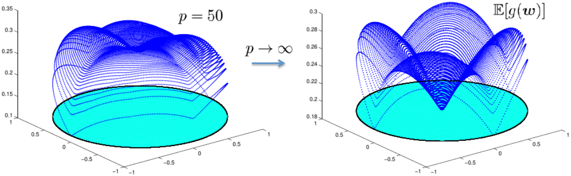
What is going on here?



... You can find me and see thousands more!

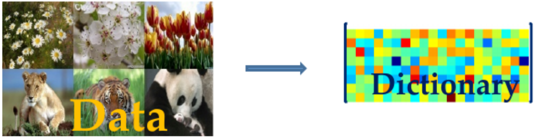
Theme of this thesis

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



... starting with *sparse dictionary learning*!

Sparse dictionary learning



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

- 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, vision, audio, and machine learning

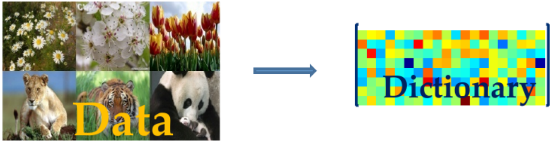
Dictionary learning - the complete case

$$\boldsymbol{Y} \approx \boldsymbol{Q}\boldsymbol{X} \qquad \boldsymbol{X} = \boldsymbol{\Omega} \odot \boldsymbol{V},\ \boldsymbol{\Omega} \sim \mathrm{Ber}(\theta),\ \boldsymbol{V} \sim \mathcal{N}(0,1).$$

Dictionary learning - the complete case

$$Y \approx QX \qquad X = \Omega \odot V,\ \Omega \sim \mathrm{Ber}(\theta),\ V \sim \mathcal{N}(0,1).$$

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



$$\boldsymbol{Y} \approx \boldsymbol{Q}\boldsymbol{X} \qquad \boldsymbol{X} = \boldsymbol{\Omega} \odot \boldsymbol{V},\ \boldsymbol{\Omega} \sim \mathrm{Ber}(\theta),\ \boldsymbol{V} \sim \mathcal{N}(0,1).$$

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

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

if we recover one row (up to scaling) of $\boldsymbol{X}$, then we use *deflation* to find the others.

Dictionary learning: the complete case

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

- Convex relaxation:

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

  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].

- Nonconvex relaxation:

Model problem

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

many precedents, e.g., [Zibulevsky-Perlmutter, '01] in source separation.

The model problem

Model problem

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

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

Towards geometric understanding

Model problem

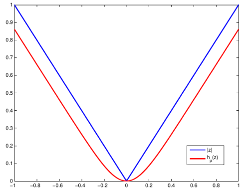
$$\min \quad \frac{1}{p}\left\|\boldsymbol{q}^*\boldsymbol{Y}\right\|_1 = \frac{1}{p}\sum_{i=1}^{p}\left|\boldsymbol{q}^*\boldsymbol{y}_i\right| \qquad \text{s.t.} \quad \left\|\boldsymbol{q}\right\|_2^2 = 1. \quad \boldsymbol{Y} \in \mathbb{R}^{n\times p}$$

Slightly modified model problem

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

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

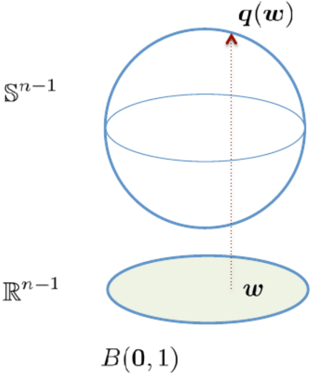
$$h_\mu\left(z\right) = \mu \log\left(\frac{e^{z/\mu} + e^{-z/\mu}}{2}\right)$$



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

Why might this work?

W.l.o.g., assume for analysis that $\boldsymbol{Y} = \boldsymbol{X}$ (i.e., $\boldsymbol{A} = \boldsymbol{I}$); we correctly recover a row iff an algorithm produces $\boldsymbol{q}_\diamond = \pm \boldsymbol{e}_i$, $i = 1, \dots, n$

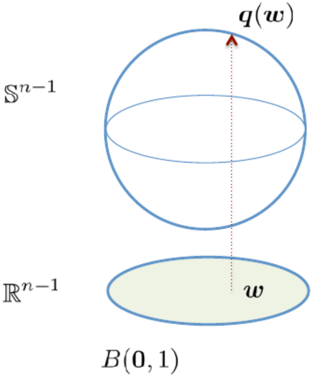


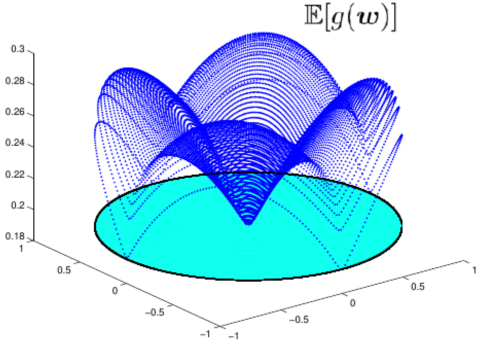
- $f(\boldsymbol{q}) = \frac{1}{p} \sum_{i=1}^{p} h_\mu(\boldsymbol{q}^* \boldsymbol{x}_i)$
- $\boldsymbol{q}(\boldsymbol{w}) = \left(w_1, \dots, w_{n-1}, \sqrt{1 - \|\boldsymbol{w}\|_2^2}\right)$
- $g(\boldsymbol{w}) = f(g(\boldsymbol{w}))$

Target points are $\boldsymbol{w}_\diamond = \boldsymbol{0}, \pm \boldsymbol{e}_1, \dots, \pm \boldsymbol{e}_{n-1}$

Why might this work?

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

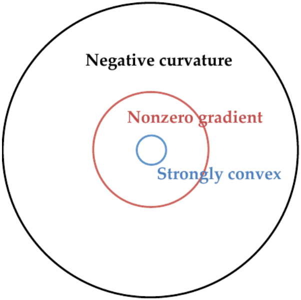




Target points are
$\boldsymbol{w}_\diamond = \mathbf{0}, \pm \boldsymbol{e}_1, \ldots, \pm \boldsymbol{e}_{n-1}$

Formally

Lemma: Suppose $\theta \in \left[\frac{1}{n}, \frac{1}{2}\right)$, and $\mu < cn^{-5/4}$. Then

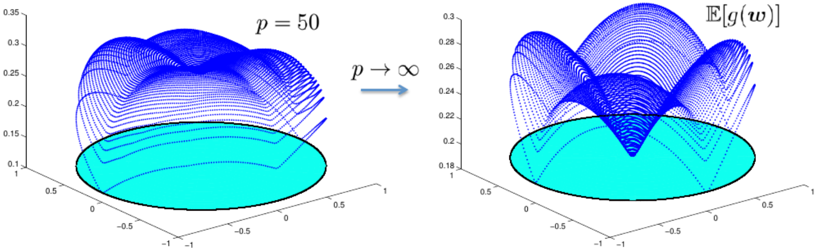


- $\nabla^2 \mathbb{E}\left[g\left(\boldsymbol{w}\right)\right] \preceq \frac{c'\theta}{\mu}\boldsymbol{I}$, for $\|\boldsymbol{w}\| \le \frac{\mu}{4\sqrt{2}}$
- $\frac{\boldsymbol{w}^* \nabla \mathbb{E}[g(\boldsymbol{w})]}{\|\boldsymbol{w}\|} \ge c'\theta$, for $\frac{\mu}{4\sqrt{2}} \le \|\boldsymbol{w}\| \le R$
- $\frac{\boldsymbol{w}^* \nabla \mathbb{E}[g(\boldsymbol{w})] \boldsymbol{w}}{\|\boldsymbol{w}\|^2} \le -c'\theta$, for $\|\boldsymbol{w}\| \ge R$

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

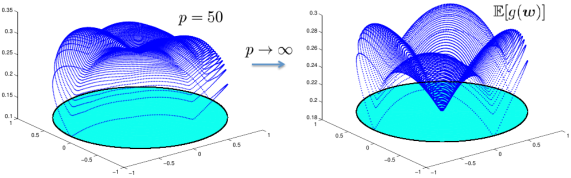
and so, **every local optimizer of $\mathbb{E}\left[g\left(w\right)\right]$ is a target point.**

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 but 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 concentration-expectation path

Lemma

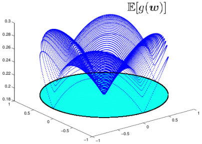
Suppose $\theta \in \left[\frac{1}{n}, \frac{1}{2}\right)$, if $\mu < cn^{-5/4}$, and $p \geq \frac{Cn^3}{\mu^2\theta^2}\log n$, it ***holds uniformly w.h.p.*** *that*

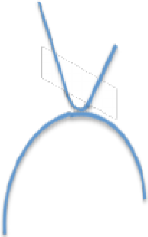
- $\nabla^2 \mathbb{E}\left[g(\boldsymbol{w})\right] \preceq \frac{c''\theta}{\mu}\boldsymbol{I}$, *for* $\|\boldsymbol{w}\| \leq \frac{\mu}{4\sqrt{2}}$
- $\frac{\boldsymbol{w}^* \nabla \mathbb{E}[g(\boldsymbol{w})]}{\|\boldsymbol{w}\|} \geq c''\theta$, *for* $\frac{\mu}{4\sqrt{2}} \leq \|\boldsymbol{w}\|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

Algorithm design

... following intuition we build from the geometry:

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





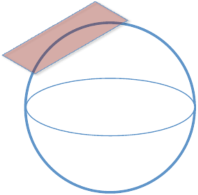
Trust-region on manifolds [Absil, Baker, Gallivan. '07], also [Absil, Mahoney, Sepulchre. '08]

Trust region method

Consider $\boldsymbol{q} \in \mathbb{S}^{n-1}$; for $\boldsymbol{\delta} \perp \boldsymbol{q}$, calculus gives

$$\begin{aligned} f(\exp_{\boldsymbol{q}}(\boldsymbol{\delta})) &= f(\boldsymbol{q}) + \langle \boldsymbol{\delta}, \nabla f(\boldsymbol{q}) \rangle + \tfrac{1}{2} \boldsymbol{\delta}^* \left(\nabla^2 f(\boldsymbol{q}) - \langle \boldsymbol{q}, \nabla f(\boldsymbol{q}) \rangle \right) \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|_2^3) \\ &\doteq \widehat{f}(\boldsymbol{q}; \boldsymbol{\delta}) + O(\|\boldsymbol{\delta}\|_2^3) \end{aligned}$$

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



Basic **Riemannian trust region method**:

$$\begin{aligned} &\boldsymbol{\delta}_\star \in \arg \min_{\delta \in T_{\boldsymbol{q}_k} \mathbb{S}^{n-1}, \ \|\boldsymbol{\delta}\|_2 \le \Delta} \widehat{f}(\boldsymbol{q}_k; \boldsymbol{\delta}) \\ &\boldsymbol{q}_{k+1} = \exp_{\boldsymbol{q}_k}(\boldsymbol{\delta}_\star). \end{aligned}$$

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

Pull things together

Theorem (informal)

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

$$\|\widehat{\boldsymbol{q}} - \boldsymbol{q}_\diamond\| \leq C \frac{\mu}{\theta} \sqrt{\frac{n \log p}{p}}$$

for some target solution $\boldsymbol{q}_\diamond$ *satisfying* $\boldsymbol{q}_\diamond^* \boldsymbol{Y} = \pm e_i^* \boldsymbol{X}$.

- Using **linear programming rounding** + **deflation**, one can recover all of $\boldsymbol{X}$, and subsequently $\boldsymbol{Q}$.
- If $\boldsymbol{Q}$ is not an orthobasis, apply preconditioning, but need $p \geq \mathrm{poly}\left(n, \sigma_{\min}(\boldsymbol{Q})^{-1}\right)$.

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]	$\boldsymbol{Q} \in \mathbb{R}^{m \times n}$ $(m \le n)$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$

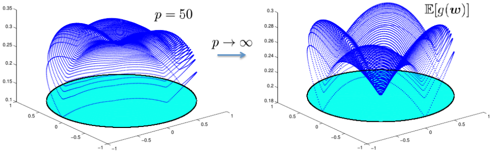
- **Quasipolynomial algorithms** with better guarantees

[Spielman, Wang, Wright,'12]	$\boldsymbol{Q} \in \mathbb{R}^{n \times n}$, $\theta = \tilde{O}\left(1/\sqrt{n}\right)$
[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]

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

What we have done so far...



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

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

Related publications:

- Sun, Qu, Wright. Complete dictionary recovery over the sphere. In preparation.
- Qu, Sun, Wright. Finding a sparse vector in a subspace: Linear sparsity using alternating directions. NIPS'14.

Future Directions

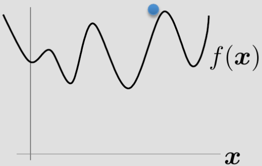
For sparse dictionary learning:

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

Proving nonconvex recovery

Theme of this thesis

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



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

- **Matrix completion**: [Keshevan, Oh, Montanari.'09], [Jain, Netrapali, Sanghavi. '13], [Hardt'13], [Hardt, Wooters. '14]. Also [Meta, Jain, Dhillon.'09]
- **Dictionary learning**: [Agarwal, Anandkumar, Netrapali. '13], [Arora, Ge, Moitra. '13], [Agarwal, Anandkumar, Jain, Netrapali.'13]
- **Tensor recovery**: [Jain, Oh. '13], [Anandkumar, Ge, Janzamin. '14]
- **Phase retrieval**: [Netrapali, Jain, Sanghavi.'13], [Candes, Li, Soltanokoltabi. '14]

Future Directions

For the analytic strategy:

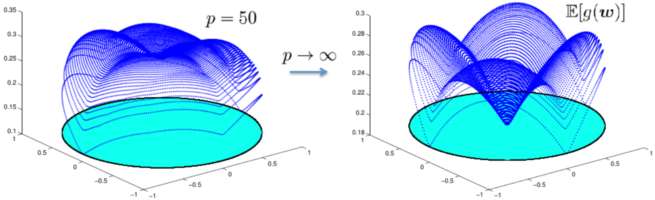
Generalized Model Problem

$$\min \quad \frac{1}{p}\sum_{i=1}^{p} f_i\left(\boldsymbol{q}\right) \quad \text{s.t. } \boldsymbol{q} \in \mathcal{M}.$$

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

- Other problems: phase retrieval, matrix/tensor recovery, recovery of signal with simultenous structures, **blind deconvolution**, etc

THANKS to Prof. John Wright and Mr. Qing Qu.



Questions?