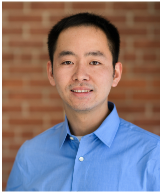
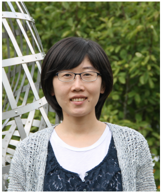
Recent Advances in Nonconvex Methods for High-Dimensional Estimation





Yuxin Chen
Princeton

Yuejie Chi
CMU

Yue M. Lu
Harvard

ICASSP 2018 Tutorial
Calgary, Canada

Slides available at: **https://goo.gl/TndZoW**

Acknowledgement

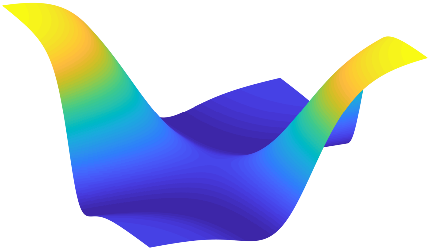
Collaborators: Emmanuel Candès, Jianqing Fan, Hong Hu, Gen Li, Yuanxin Li, Yingbin Liang, Wangyu Luo, Cong Ma, Jonathan Mattingly, Chuang Wang, Kaizheng Wang, Huishuai Zhang

Sponsors: This work is supported in part by AFOSR FA9550-15-1-0205, ONR N00014-18-1-2142, NSF ECCS-1818571, CCF-1806154, ARO W911NF-16-1-0265, NSF CCF-1319140, and NSF CCF-1718698

Nonconvex estimation problems are everywhere

Empirical risk minimization is usually nonconvex

$$\text{minimize}_x \qquad f(\boldsymbol{x};\boldsymbol{y}) \qquad \longrightarrow$$ Loss function may be nonconvex

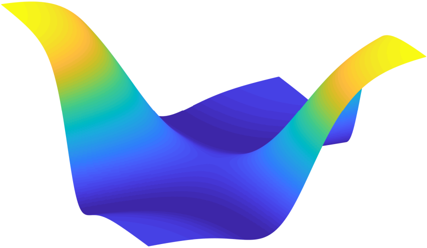


Nonconvex estimation problems are everywhere

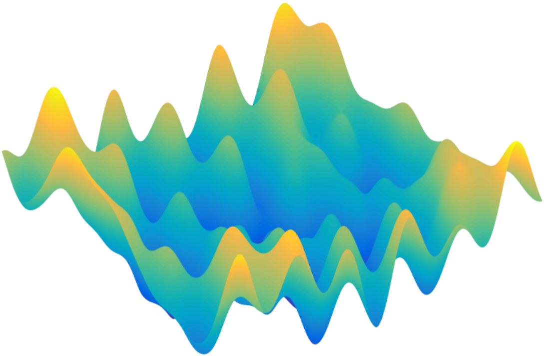
Empirical risk minimization is usually nonconvex

$$\text{minimize}_x \qquad f(\boldsymbol{x}; \boldsymbol{y}) \qquad \longrightarrow$$ Loss function may be nonconvex

- nonlinear regression
- low-rank matrix completion
- blind deconvolution
- dictionary learning
- learning mixture models
- deep learning
- generative adversarial networks
- …



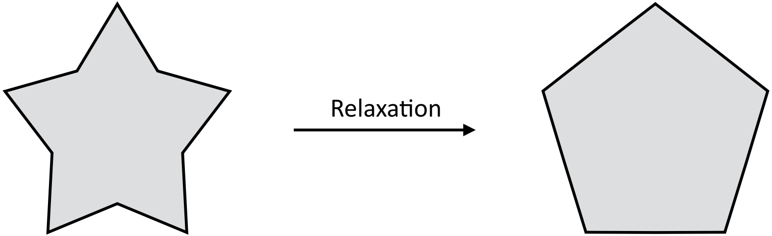
Nonconvex optimization may be super scary



There may be bumps everywhere and exponentially many local optima

e.g. 1-layer neural net [Auer, Herbster, Warmuth '96; Vu '98]

Convex relaxation



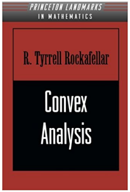
Examples:

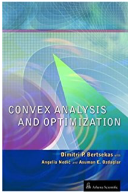
- sparse recovery (ℓ_1-minimization) [Donoho ’06], [Candès, Romberg, Tao, ’16]
- phase retrieval and low-rank matrix estimation (lifting and SDP) [Candès et al., ’13], [Jaganathan et al., ’13], [Waldspurger et al., ’15]
- subspace clustering (SSC) [Elhamifar & Vidal, ’12]
- MAXCUT (SDP relaxation) [Goemans & Williamson ’95]

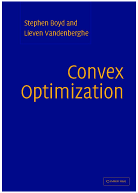
Convex optimization

Pros:

- mature theory + efficient algorithms
- strong performance guarantees



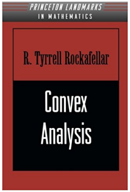


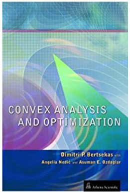


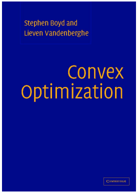
Convex optimization

Pros:

- mature theory + efficient algorithms
- strong performance guarantees







Cons:

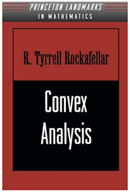
- much higher ***computation/memory cost*** (e.g. lifting)

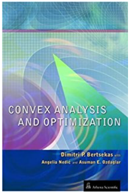
$$y_i = |\boldsymbol{a}_i^T \boldsymbol{x}|^2 = \boldsymbol{a}_i^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{a}_i$$

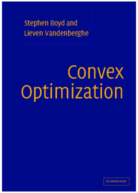
Convex optimization

Pros:

- mature theory + efficient algorithms
- strong performance guarantees







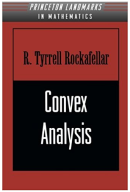
Cons:

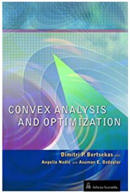
- much higher ***computation/memory cost*** (e.g. lifting)

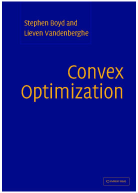
$$y_i = |\boldsymbol{a}_i^T \boldsymbol{x}|^2 = \boldsymbol{a}_i^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{a}_i \quad \Rightarrow \quad \begin{array}{ll} \text{find} & \boldsymbol{X} \\ \text{s.t.} & y_i = \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i, \qquad i = 1, \ldots, m \\ & \boldsymbol{X} \succeq 0 \end{array}$$

Convex optimization

Pros:







- mature theory + efficient algorithms
- strong performance guarantees

Cons:

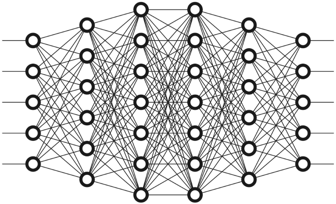
- much higher ***computation/memory cost*** (e.g. lifting)

$$y_i = |\boldsymbol{a}_i^T \boldsymbol{x}|^2 = \boldsymbol{a}_i^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{a}_i \quad \Rightarrow \quad \begin{array}{ll} \text{find} & \boldsymbol{X} \\ \text{s.t.} & y_i = \boldsymbol{a}_i^T \boldsymbol{X} \boldsymbol{a}_i, \qquad i = 1, \ldots, m \\ & \boldsymbol{X} \succeq 0 \end{array}$$

- many problems have no effective convex relaxation

Nonconvex problems are solved on a daily basis …

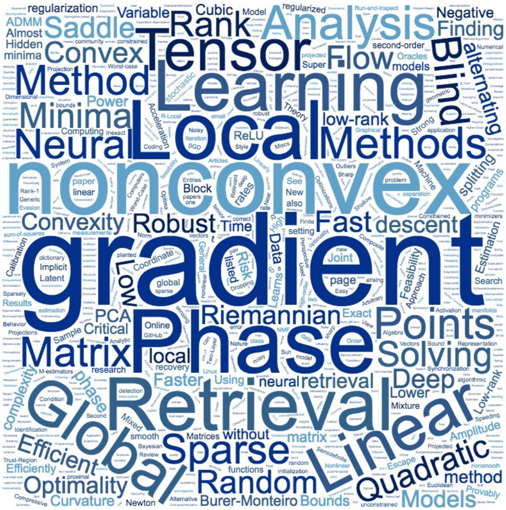
- Fineup algorithm for phase retrieval
- Gradient descent for robust regression
- EM-algorithm for parameter estimation
- alternating minimization for dictionary learning
- "back propagation" for training deep neural nets
- Simulated annealing and MCMC



Simple algorithms (such as ***gradient descent***) are often remarkably successful for solving nonconvex problems ***in practice***

Why?

Nonconvex optimization with performance guarantees



Phase retrieval: [Gerchberg-Saxton, '72], [Netrapalli et al. '13], [Candes, Li, Soltanolkotabi, '15], [Wei, '15], [Chen & Candes, '16], [Waldspurger, '16], [Wang et al. '18], and many others …

Matrix completion: [Keshavan et al., '09], [Jain et al. '12], [Hardt, '13], [Jin et al., '16], [Wei, '16], [Zheng & Lafferty, '16], [Sun & Luo, '16], [Ding & Chen, '18], and many others …

Landscape analysis: [Sun et al. '15], [Ge et al., '16], [Mei, Bai & Montanari, '16], [Li et al. '18], [Soltanolkotabi et al., '17], [Davis et al., '17], [Ge & Ma, '17], [Ge et al., '17], [Ballard et al., '17]

Blind deconvolution: [Li et al. '16], [Lee et al., '16], [Ling & Strohmer, '16], [Huang & Hand, '17], …

Blind calibration: [Cambareri & Jacques, '16], [Ling & Strohmer, '16], [Li, Lee & Bresler, '17]

Dictionary learning: [Arora et al., '14], [Sun et al., '15], [Chatterji & Bartlett, '17]

Spectral initialization: [Keshavan et al., '09], [Netrapalli et al. '13], [Sun et al., '15], [Lu & Li, '17]

Stochastic gradient methods: [Ghadimi & Lan, '13], [De Sa et al., '14], [Rong, '15], [Jin et al., '16], [Wang, Mattingly & Lu, '17], [Tripuraneri et al., '18]

See http://sunju.org/research/nonconvex/ for a detailed list of references.

Tutorial outline

Part I: Overview

Part II: Phase retrieval: a case study

- Spectral initialization
- Local refinement: algorithm and analysis

Part III: Low-rank matrix estimation

Part IV: Closing remarks

Tutorial outline

Part I: Overview

Part II: Phase retrieval: a case study

- Spectral initialization
- Local refinement: algorithm and analysis

Part III: Low-rank matrix estimation

Part IV: Closing remarks

Signal estimation from nonlinear measurements

Model:

$$\boldsymbol{a}_i^T\boldsymbol{x}^\natural \longrightarrow \boxed{\mathbf{P}(y_i \mid \boldsymbol{a}_i^T\boldsymbol{x}^\natural)} \longrightarrow y_i$$

arbitrary "channel"

- Unknown vector: $\boldsymbol{x}^\natural \in \mathbb{R}^n$
- Sensing vectors: $\{a_i\}_{i=1}^m \subset \mathbb{R}^n$

Signal estimation from nonlinear measurements

Model:

$$\boldsymbol{a}_i^T \boldsymbol{x}^\natural \longrightarrow \boxed{\mathbf{P}(y_i \mid \boldsymbol{a}_i^T \boldsymbol{x}^\natural)} \longrightarrow y_i$$

arbitrary "channel"

- Unknown vector: $\boldsymbol{x}^\natural \in \mathbb{R}^n$
- Sensing vectors: $\{a_i\}_{i=1}^m \subset \mathbb{R}^n$

Examples:

- Nonlinear sensors: $y_i = f(\boldsymbol{a}_i^T \boldsymbol{x}^\natural) + w_i$
- Imaging: $y_i \sim \mathrm{Poisson}(\boldsymbol{a}_i^T \boldsymbol{x}^\natural)$
- Logistic regression: $y_i \sim \mathrm{Bernoulli}\big[\mathrm{Logit}(\boldsymbol{a}_i^T \boldsymbol{x}^\natural)\big]$

Example: Phase Retrieval

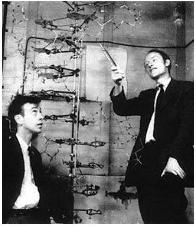
Reconstruct $\boldsymbol{x}^\natural \in \mathbb{C}^n$ ***without the phase information***

$$y_1 = \left| \langle \boldsymbol{a}_1, \boldsymbol{x}^\natural \rangle \right|^2$$

$$y_2 = \left| \langle \boldsymbol{a}_2, \boldsymbol{x}^\natural \rangle \right|^2$$

$$\vdots$$

$$y_m = \left| \langle \boldsymbol{a}_m, \boldsymbol{x}^\natural \rangle \right|^2$$



Nobel Prize for Watson, Crick, and Wilkins in 1962 based on work by Rosalind Franklin

Applications:

- Phase retrieval (X-ray crystallography, diffractive imaging, …)
- Blind deconvolution
- Channel estimation
- Spectral factorization

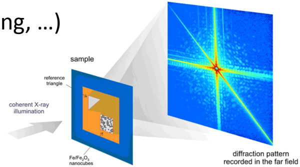


Fig credit: Stanford SLAC

Empirical risk minimization

M-estimator:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{m}\sum_i \mathrm{Loss}(y_i, \boldsymbol{a}_i^T\boldsymbol{x}) + \Phi(\boldsymbol{x})$$

Empirical risk minimization

M-estimator:

data fidelity

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{m}\sum_i \boxed{\mathrm{Loss}(y_i, \boldsymbol{a}_i^T\boldsymbol{x})} + \Phi(\boldsymbol{x})$$

Empirical risk minimization

M-estimator:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{m}\sum_i \underbrace{\boxed{\mathrm{Loss}(y_i, \boldsymbol{a}_i^T\boldsymbol{x})}}_{\text{data fidelity}} + \underbrace{\boxed{\Phi(\boldsymbol{x})}}_{\text{prior}}$$

Empirical risk minimization

M-estimator:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{m}\sum_i \underbrace{\boxed{\mathrm{Loss}(y_i, \boldsymbol{a}_i^T\boldsymbol{x})}}_{\text{data fidelity}} + \underbrace{\boxed{\Phi(\boldsymbol{x})}}_{\text{prior}}$$

Challenges:

- Nonconvex loss functions (e.g. phase retrieval)

$$\mathrm{minimize}_{\boldsymbol{x}} \ \frac{1}{m}\sum_i (y_i - (\boldsymbol{a}_i^T\boldsymbol{x})^2)^2$$

- Nonconvex regularizers

$$\Phi(\boldsymbol{x}) = \|\boldsymbol{x}\|_p^p \quad \text{for} \quad 0 < p < 1$$

Empirical risk minimization

M-estimator:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{m}\sum_i \overbrace{\boxed{\mathrm{Loss}(y_i, \boldsymbol{a}_i^T\boldsymbol{x})}}^{\text{data fidelity}} + \overbrace{\boxed{\Phi(\boldsymbol{x})}}^{\text{prior}}$$

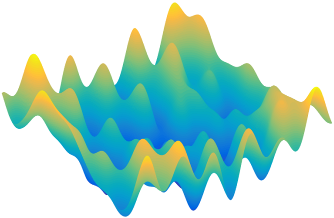
Challenges:

- Nonconvex loss functions (e.g. phase retrieval)

$$\mathrm{minimize}_{\boldsymbol{x}} \ \frac{1}{m}\sum_i (y_i - (\boldsymbol{a}_i^T\boldsymbol{x})^2)^2$$

- Nonconvex regularizers

$$\Phi(\boldsymbol{x}) = \|\boldsymbol{x}\|_p^p \quad \text{for} \quad 0 < p < 1$$



Nonconvex optimization with ***performance guarantee***?

Where is hope?

PCA: a classical success story of nonconvex optimization

Find the best ***rank-one*** approximation of a symmetric PSD matrix $\boldsymbol{M}$

$$\text{minimize}_{\boldsymbol{x}} \quad f(x) = \left\| \boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{M} \right\|_{\mathrm{F}}^2$$

Nonconvex, but ***global optimal*** solution is well-known.

PCA: a classical success story of nonconvex optimization

Find the best ***rank-one*** approximation of a symmetric PSD matrix $\boldsymbol{M}$

$$\mathrm{minimize}_{\boldsymbol{x}} \quad f(x) = \left\| \boldsymbol{x}\boldsymbol{x}^T - \boldsymbol{M} \right\|_{\mathrm{F}}^2$$

Nonconvex, but ***global optimal*** solution is well-known.

Eckart-Young Theorem:

1. Eigenvalue decomposition:

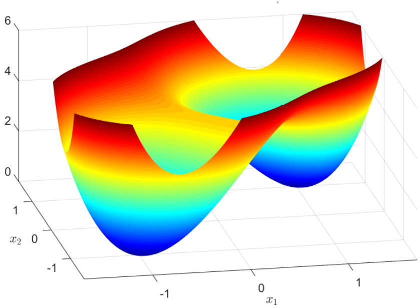
$$\boldsymbol{M} = \boldsymbol{U}\,\mathrm{diag}\,\{\sigma_1, \sigma_2, \ldots, \sigma_n\}\;\boldsymbol{U}^T$$

2. Find the dominant eigenvector: $\boldsymbol{x}_{\mathrm{opt}} = \sqrt{\sigma_1}\,\boldsymbol{u}_1$

The optimization landscape of PCA

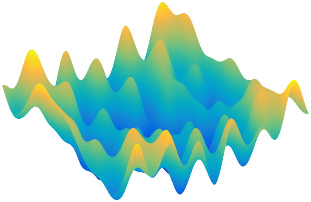
Example:

$$\text{minimize}_{\boldsymbol{x}} \quad f(x) = \left\| \boldsymbol{x}\boldsymbol{x}^T - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\|_{\mathrm{F}}^2$$

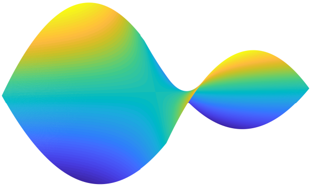


Critical points are either ***global optima*** or ***strict saddles*** [see Part III for details]

In many problems: nonconvex but benign landscapes



intractable (***worst-case***)



tractable (***typical case***)

Under certain ***statistical models***, we see benign global geometry: critical points are either global optima or strict saddles

Empirical risk and population risk

Example: phase retrieval with Gaussian designs $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$

$$\text{minimize}_{\boldsymbol{x}} \ f_m(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}(y_i - (\boldsymbol{a}_i^T\boldsymbol{x})^2)^2 \quad \text{with} \quad y_i = (\boldsymbol{a}_i^T\boldsymbol{x}^\natural)^2$$

Empirical risk and population risk

Example: phase retrieval with Gaussian designs $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$

$$\text{minimize}_{\boldsymbol{x}} \; f_m(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - (\boldsymbol{a}_i^T\boldsymbol{x})^2\right)^2 \quad \text{with} \quad y_i = (\boldsymbol{a}_i^T\boldsymbol{x}^\natural)^2$$

$$\Bigg\Downarrow$$

"law of large numbers"

$m \to \infty$

$$\text{minimize}_{\boldsymbol{x}} \; f(\boldsymbol{x}) = \mathbb{E}\left(y_1 - (\boldsymbol{a}_1^T\boldsymbol{x})^2\right)^2$$

Empirical risk and population risk

Example: phase retrieval with Gaussian designs $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$

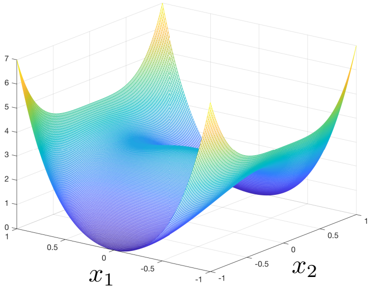
$$\text{minimize}_{\boldsymbol{x}} \; f_m(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - (\boldsymbol{a}_i^T \boldsymbol{x})^2\right)^2 \quad \text{with} \quad y_i = (\boldsymbol{a}_i^T \boldsymbol{x}^\natural)^2$$

"law of large numbers"

$m \to \infty$

$$\Big\Downarrow$$

$$\text{minimize}_{\boldsymbol{x}} \; f(\boldsymbol{x}) = \mathbb{E}\left(y_1 - (\boldsymbol{a}_1^T \boldsymbol{x})^2\right)^2$$

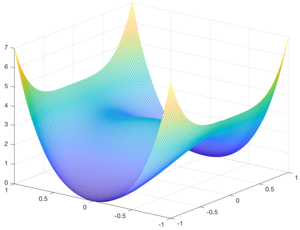


$$f(x_1, x_2) = 3 + 3(x_1^2 + x_2^2)^2 - 6x_1^2 - 2x_2^2$$

Sample complexity:

how large m needs to be?

Landscape analysis for phase retrieval



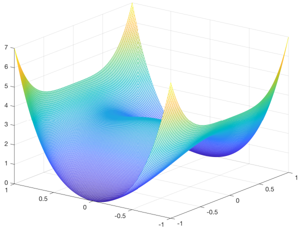
$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{i=1}^{m}\left[y_i - (\boldsymbol{a}_i^T\boldsymbol{x})^2\right]^2$$

Theorem: (informal) [Sun, Qu, Wright, '16]

Let $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\boldsymbol{I})$. When $m \gtrsim n\log^3 n$, w.h.p.,

- All local (and global) minimizers are of the form $\boldsymbol{x}^\natural, -\boldsymbol{x}^\natural$
- All other critical points of $f(\boldsymbol{x})$ are strict saddles (i.e. there exist escape directions)

Landscape analysis for phase retrieval



$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{i=1}^{m}\left[y_i - (\boldsymbol{a}_i^T\boldsymbol{x})^2\right]^2$$

Notation:
$f(n) \gtrsim g(n)$ means $\lim_{n\to\infty}\frac{|f(n)|}{|g(n)|} \geq \text{const}$

Theorem: (informal) [Sun, Qu, Wright, '16]

Let $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I})$. When $m \gtrsim n\log^3 n$, w.h.p.,

- All local (and global) minimizers are of the form $\boldsymbol{x}^\natural, -\boldsymbol{x}^\natural$
- All other critical points of $f(\boldsymbol{x})$ are strict saddles (i.e. there exist escape directions)

More general results on the landscapes of empirical risk

empirical risk: $\mathrm{minimize}_{\boldsymbol{x}} \ f_m(x) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i; \boldsymbol{x})$

$$\Big\Downarrow \quad \text{“law of large numbers”} \quad m \to \infty$$

population risk: $\mathrm{minimize}_{\boldsymbol{x}} \ f(x) = \mathbb{E}_{\mathrm{model}}\, \ell(y; \boldsymbol{x})$

More general results on the landscapes of empirical risk

empirical risk: $$\text{minimize}_{\boldsymbol{x}} \; f_m(x) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i; \boldsymbol{x})$$

$$\Downarrow$$

"law of large numbers"

$m \to \infty$

population risk: $$\text{minimize}_{\boldsymbol{x}} \; f(x) = \mathbb{E}_{\text{model}}\, \ell(y; \boldsymbol{x})$$

Theorem: (informal) [Mei, Bai, Montanari, '17]

Under technical assumptions on the loss function $\ell(\cdot;\cdot)$, w.h.p.,

1. $\sup_{\boldsymbol{x}} \left\| \nabla f_m(\boldsymbol{x}) - \nabla f(\boldsymbol{x}) \right\|_2 \lesssim \sqrt{n \log m / m}$
2. $\sup_{\boldsymbol{x}} \left\| \nabla^2 f_m(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x}) \right\|_{\text{op}} \lesssim \sqrt{n \log m / m}$

Uniform convergence of gradient and hessian

Example: binary linear classification

Model: $y_i \in \{0,1\}$ with $\mathbb{P}(Y = 1 \mid R = \boldsymbol{a}_i) = \sigma(\boldsymbol{a}_i^T \boldsymbol{x}^\natural)$

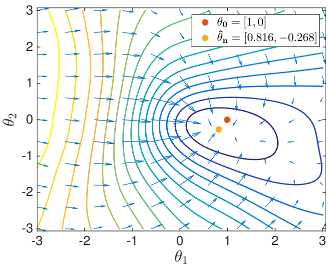
Nonlinear least-squares: $\text{minimize}_{\boldsymbol{x}} \ f_m(x) = \frac{1}{m}\sum_{i=1}^{m}[y_i - \sigma(\boldsymbol{a}_i^T \boldsymbol{x})]^2$

Example: binary linear classification

Model: $y_i \in \{0, 1\}$ with $\mathbb{P}(Y = 1 \mid R = \boldsymbol{a}_i) = \sigma(\boldsymbol{a}_i^T \boldsymbol{x}^\natural)$

Nonlinear least-squares: $\text{minimize}_{\boldsymbol{x}} \ f_m(x) = \frac{1}{m} \sum_{i=1}^{m} [y_i - \sigma(\boldsymbol{a}_i^T \boldsymbol{x})]^2$

empirical risk $\approx$ population risk



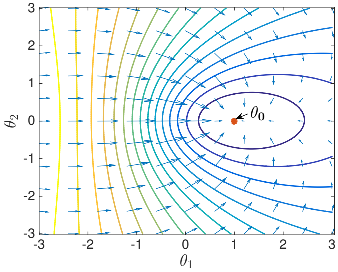


Fig credit: Mei, Bai and Montanari

Benign landscapes lead to efficient algorithms with polynomial complexity

Generic results and algorithms for benign landscapes

- Gradient decent with random initialization escapes saddles almost surely [Lee et al., ’16]

- Saddle escaping algorithms with polynomial complexity:
  - Trust-region [Sun et al. ’16]
  - Perturbed GD [Jin et al. ’17]
  - Perturbed accelerated GD [Jin et al. ’17]
  - Natasha [Allen-Zhu ’17]
  - Cubic-regularized method [Agarwal et al., ’17]

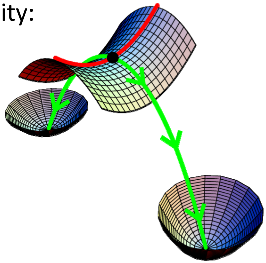


Fig. credit: Turnhout et al.

Generic results and algorithms for benign landscapes

- Gradient decent with random initialization escapes saddles almost surely [Lee et al., '16]

- Saddle escaping algorithms with polynomial complexity:
  - Trust-region [Sun et al. '16]
  - Perturbed GD [Jin et al. '17]
  - Perturbed accelerated GD [Jin et al. '17]
  - Natasha [Allen-Zhu '17]
  - Cubic-regularized method [Agarwal et al., '17]

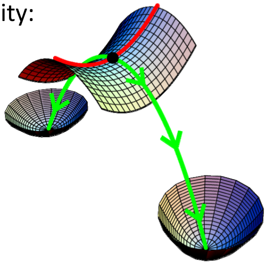


Fig. credit: Turnhout et al.

Cons: computational complexity is $\mathrm{Poly}(n)$

$\longrightarrow$ ***Ideally***: linear complexity (proportional to the time to load the data)

Much ***stronger guarantees*** are possible
by studying ***specific problems***!

Tutorial outline

Part I: Overview

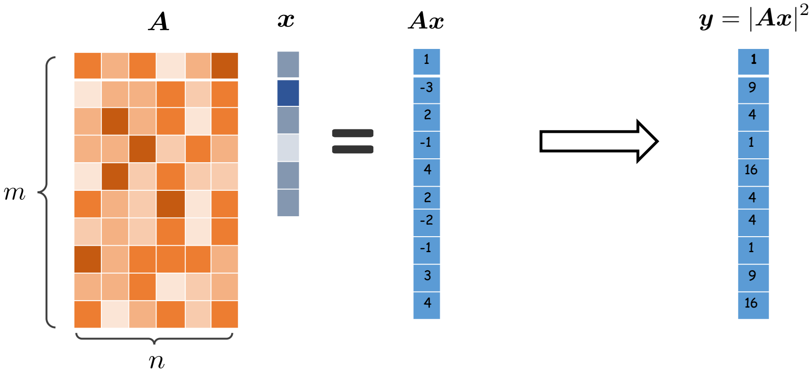
Part II: Phase retrieval: a case study

- Spectral initialization
- Local refinement: algorithm and analysis

Part III: Low-rank matrix estimation

Part IV: Closing remarks

Phase retrieval: solving quadratic systems of equations



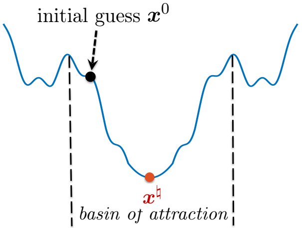
Recover $\boldsymbol{x}^\natural \in \mathbb{R}^n$ from m random quadratic measurements

$$y_i = \left| \boldsymbol{a}_i^T \boldsymbol{x}^\natural \right|^2, \qquad i = 1, \ldots, m$$

assume w.l.o.g. $\|\boldsymbol{x}^\natural\|_2 = 1$

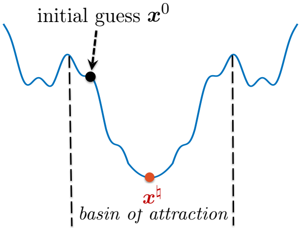
Common theme: two-stage approach

1. ***Initialization***: find an initial point within a local basin close to $\boldsymbol{x}^\natural$

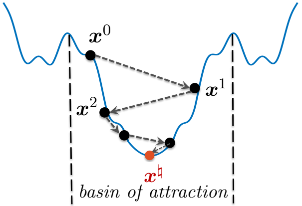


Common theme: two-stage approach

1. ***Initialization***: find an initial point within a local basin close to $\boldsymbol{x}^\natural$

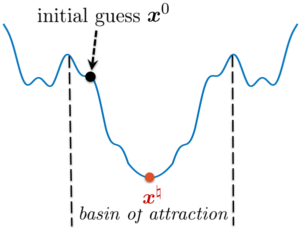


2. Careful iterative ***local refinement*** (e.g. gradient descent)



Common theme: two-stage approach

1. ***Initialization***: find an initial point within a local basin close to $\boldsymbol{x}^\natural$



2. Careful iterative ***local refinement*** (e.g. gradient descent)

A spectral method for initialization

Spectral Initialization

Model:

$$y_i \approx f(\boldsymbol{a}_i^T \boldsymbol{x}^\natural), \qquad i = 1, 2, \ldots, m$$

Spectral Initialization

Model:

$$y_i \approx f(\boldsymbol{a}_i^T \boldsymbol{x}^\natural), \qquad i = 1, 2, \ldots, m$$

Spectral initialization:

1. $\boldsymbol{D}_m = \frac{1}{m} \sum_{i=1}^{m} \mathcal{T}(y_i) \boldsymbol{a}_i \boldsymbol{a}_i^T$

2. $\boldsymbol{x}_1 =$ top eigenvector$(\boldsymbol{D}_m)$

PHD: principal Hessian direction [Li '92], [Keshavan et al. '10], [Netrapalli et al. '13]

Why doe it work?

The model:

$$y_i \approx f(\boldsymbol{a}_i^T \boldsymbol{x}^\natural), \qquad i = 1, 2, \ldots, m$$

The data matrix:

"Law of large numbers"

$$\boldsymbol{D}_m = \frac{1}{m}\sum_{i=1}^{m} \mathcal{T}(y_i)\boldsymbol{a}_i\boldsymbol{a}_i^T \Longrightarrow \mathbb{E}\left[\mathcal{T}(y)\boldsymbol{a}\boldsymbol{a}^T\right]$$

Why doe it work?

The model:

$$y_i \approx f(\boldsymbol{a}_i^T \boldsymbol{x}^\natural), \qquad i = 1, 2, \ldots, m$$

The data matrix:

"Law of large numbers"

$$\boldsymbol{D}_m = \frac{1}{m} \sum_{i=1}^{m} \mathcal{T}(y_i) \boldsymbol{a}_i \boldsymbol{a}_i^T \Longrightarrow \mathbb{E}\left[\mathcal{T}(y) \boldsymbol{a}\boldsymbol{a}^T\right] = \beta_1 \boldsymbol{I} + (\beta_2 - \beta_1) \boldsymbol{x}^\natural (\boldsymbol{x}^\natural)^T$$

Why doe it work?

The model:

$$y_i \approx f(\boldsymbol{a}_i^T \boldsymbol{x}^\natural), \qquad i = 1, 2, \ldots, m$$

The data matrix:

"Law of large numbers"

$$\boldsymbol{D}_m = \frac{1}{m}\sum_{i=1}^{m} \mathcal{T}(y_i)\boldsymbol{a}_i\boldsymbol{a}_i^T \Longrightarrow \mathbb{E}\left[\mathcal{T}(y)\boldsymbol{a}\boldsymbol{a}^T\right] = \beta_1 \boldsymbol{I} + (\beta_2 - \beta_1)\boldsymbol{x}^\natural(\boldsymbol{x}^\natural)^T$$

with $\beta_1 = \mathbb{E}\,\mathcal{T}(y)$, $\beta_2 = \mathbb{E}\left[\mathcal{T}(y)(\boldsymbol{a}^T\boldsymbol{x}^\natural)^2\right]$

Similar approaches used in matrix completion, blind deconvolution, …

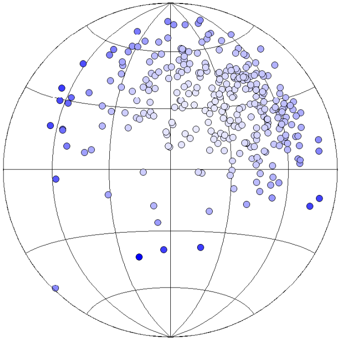
Why does it work? The deterministic case

The data matrix:

$$\boldsymbol{D}_m = \frac{1}{m}\left\{(\boldsymbol{a}_1^T\boldsymbol{x}^\natural)^2\boldsymbol{a}_1\boldsymbol{a}_1^T + (\boldsymbol{a}_2^T\boldsymbol{x}^\natural)^2\boldsymbol{a}_2\boldsymbol{a}_2^T + (\boldsymbol{a}_3^T\boldsymbol{x}^\natural)^2\boldsymbol{a}_3\boldsymbol{a}_3^T + \ldots(\boldsymbol{a}_m^T\boldsymbol{x}^\natural)^2\boldsymbol{a}_m\boldsymbol{a}_m^T\right\}$$

Correlated patterns: higher weights

Uncorrelated patterns: lower weights

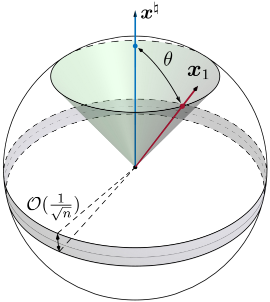


Pattern matching: $\max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}^T\boldsymbol{D}_m\boldsymbol{x}$

Performance Analysis

Cosine similarity:

$$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \stackrel{\text{def}}{=} \frac{(\boldsymbol{x}_1^T \boldsymbol{x}^\natural)^2}{\|\boldsymbol{x}_1\|^2 \|\boldsymbol{x}^\natural\|^2}$$



Performance guarantees:

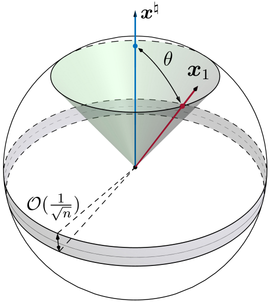
[Gaussian measurements]

$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) > 1 - \delta$ *w. high prob. if*

Performance Analysis

Cosine similarity:

$$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \stackrel{\text{def}}{=} \frac{(\boldsymbol{x}_1^T \boldsymbol{x}^\natural)^2}{\|\boldsymbol{x}_1\|^2 \|\boldsymbol{x}^\natural\|^2}$$



Performance guarantees:

[Gaussian measurements]

$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) > 1 - \delta$ *w. high prob. if*

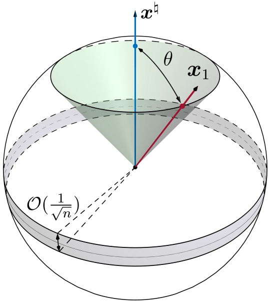
[Netrapalli et al, ’13]

$$m \gtrsim n \log^3 n$$

Performance Analysis

Cosine similarity:

$$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \stackrel{\text{def}}{=} \frac{(\boldsymbol{x}_1^T \boldsymbol{x}^\natural)^2}{\|\boldsymbol{x}_1\|^2 \|\boldsymbol{x}^\natural\|^2}$$



Performance guarantees:

[Gaussian measurements]

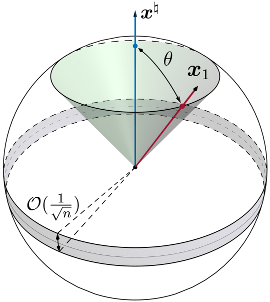
$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) > 1 - \delta$ *w. high prob. if*

[Netrapalli et al, '13]	[Candes et al., '15]
$m \gtrsim n \log^3 n$	$m \gtrsim n \log n$

Performance Analysis

Cosine similarity:

$$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \stackrel{\text{def}}{=} \frac{(\boldsymbol{x}_1^T \boldsymbol{x}^\natural)^2}{\|\boldsymbol{x}_1\|^2 \|\boldsymbol{x}^\natural\|^2}$$



Performance guarantees:

[Gaussian measurements]

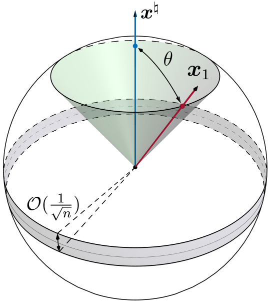
$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) > 1 - \delta$ *w. high prob. if*

[Netrapalli et al, '13]	[Candes et al., '15]	[Chen & Candes, '15]
$m \gtrsim n \log^3 n$	$m \gtrsim n \log n$	$m \gtrsim n$

Performance Analysis

Cosine similarity:

$$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \overset{\text{def}}{=} \frac{(\boldsymbol{x}_1^T \boldsymbol{x}^\natural)^2}{\|\boldsymbol{x}_1\|^2 \|\boldsymbol{x}^\natural\|^2}$$



Performance guarantees:

[Gaussian measurements]

$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) > 1 - \delta$ *w. high prob. if*

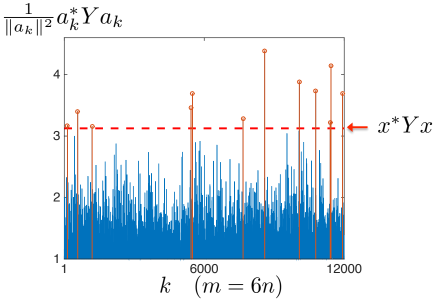
[Netrapalli et al, '13]	[Candes et al., '15]	[Chen & Candes, '15]
$m \gtrsim n \log^3 n$	$m \gtrsim n \log n$	$\boxed{m \gtrsim n}$

Truncation: $\mathcal{T}(y) = y\,\mathbb{1}_{\{|y| \le t\}}$

Truncated spectral initialization

$$\mathbb{E}[\boldsymbol{D}] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T\right]$$

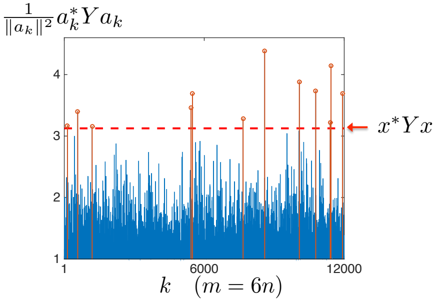
$$= \boldsymbol{I} + 2\boldsymbol{x}^\natural (\boldsymbol{x}^\natural)^T$$



Truncated spectral initialization

$$\mathbb{E}[\boldsymbol{D}] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T\right]$$

$$= \boldsymbol{I} + 2\boldsymbol{x}^\natural (\boldsymbol{x}^\natural)^T$$

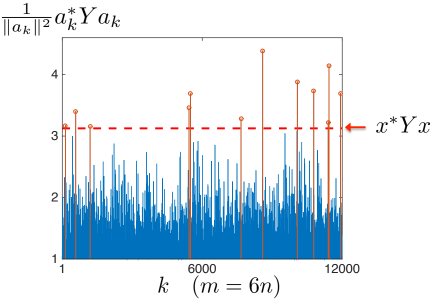


Problem: Unless $m \gg n$, dangerous to use empirical average as large observations $y_i = (\boldsymbol{a}_i^T \boldsymbol{x}^\natural)^2$ bear too much influence

Truncated spectral initialization

$$\mathbb{E}[\boldsymbol{D}] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T\right]$$

$$= \boldsymbol{I} + 2\boldsymbol{x}^\natural (\boldsymbol{x}^\natural)^T$$



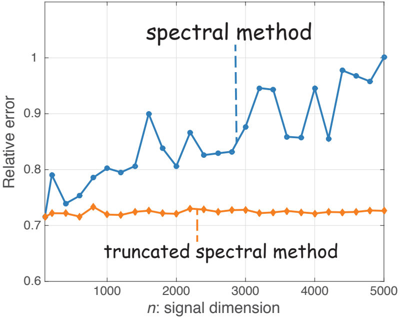
Problem: Unless $m \gg n$, dangerous to use empirical average as large observations $y_i = (\boldsymbol{a}_i^T \boldsymbol{x}^\natural)^2$ bear too much influence

Solution: Discard high leverage samples and consider a ***truncated sum***

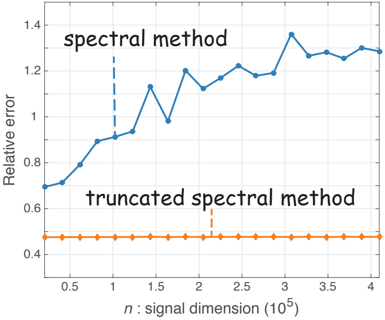
$$\frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^T \cdot \mathbb{1}_{\{|y| \le t\}}$$

[Chen & Candes, '15]

Importance of truncated spectral initialization



real Gaussian $m = 6n$

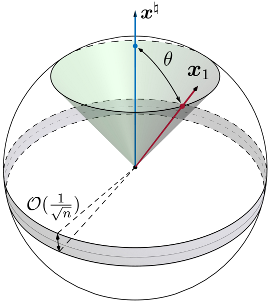


complex CDP $m = 12n$

Performance Analysis

Cosine similarity:

$$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \stackrel{\text{def}}{=} \frac{(\boldsymbol{x}_1^T \boldsymbol{x}^\natural)^2}{\|\boldsymbol{x}_1\|^2 \|\boldsymbol{x}^\natural\|^2}$$



Performance guarantees:

[Gaussian measurements]

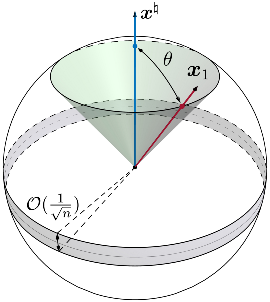
$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) > 1 - \delta$ *w. high prob. if*

[Netrapalli et al, ’13]	[Candes et al., ’15]	[Chen & Candes, ’15]
$m \gtrsim n \log^3 n$	$m \gtrsim n \log n$	$m \gtrsim n$

Performance Analysis

Cosine similarity:

$$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \stackrel{\text{def}}{=} \frac{(\boldsymbol{x}_1^T \boldsymbol{x}^\natural)^2}{\|\boldsymbol{x}_1\|^2 \|\boldsymbol{x}^\natural\|^2}$$



Performance guarantees:

[Gaussian measurements]

$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) > 1 - \delta$ *w. high prob. if*

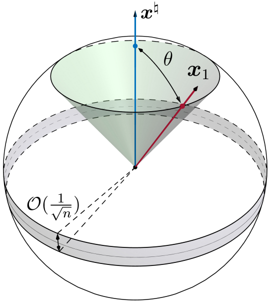
[Netrapalli et al, '13]	[Candes et al., '15]	[Chen & Candes, '15]
$m \gtrsim n \log^3 n$	$m \gtrsim n \log n$	$\boxed{m \gtrsim n}$

order-optimal, but ***unknown*** *constant*

Performance Analysis

Cosine similarity:

$$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \stackrel{\text{def}}{=} \frac{(\boldsymbol{x}_1^T \boldsymbol{x}^\natural)^2}{\|\boldsymbol{x}_1\|^2 \|\boldsymbol{x}^\natural\|^2}$$



Performance guarantees:

[Gaussian measurements]

$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) > 1 - \delta$ *w. high prob. if*

[Netrapalli et al, '13]	[Candes et al., '15]	[Chen & Candes, '15]	***[Lu & Li, '17]***
$m \gtrsim n \log^3 n$	$m \gtrsim n \log n$	$\boxed{m \gtrsim n}$	***Precise analysis***

order-optimal, but ***unknown*** *constant*

Why do we care about a precise analysis?

1. Order-wise estimates are not good enough for practitioners

Vehicle for commute	***Energy consumption***
Bike	$\mathcal{O}(\text{distance})$

Credit: Yoram Bresler

Why do we care about a precise analysis?

1. Order-wise estimates are not good enough for practitioners

Vehicle for commute	***Energy consumption***
Bike	$\mathcal{O}(\text{distance})$
Tractor	$\mathcal{O}(\text{distance})$

Credit: Yoram Bresler

Why do we care about a precise analysis?

1. Order-wise estimates are not good enough for practitioners

Vehicle for commute	***Energy consumption***
Bike	$\mathcal{O}(\text{distance})$
Tractor	$\mathcal{O}(\text{distance})$

Credit: Yoram Bresler

2. From precise analysis to ***optimal designs***

Precise Asymptotic Characterizations

Setting:

- High-dimensional $m, n \to \infty$, linear sample complexity $\frac{m}{n} \to \alpha > 0$
- i.i.d. Gaussian sensing ensemble

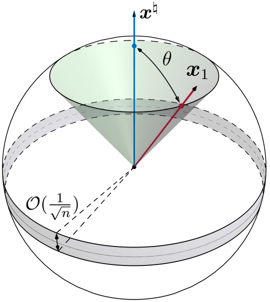
Proposition: [Lu and Li '17] Under a few ***technical conditions****:

$$\boxed{\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \xrightarrow{\mathcal{P}} \begin{cases} 0, & \text{if } \alpha < \alpha_{c,\min}, \\ \rho(\alpha), & \text{if } \alpha > \alpha_{c,\max}, \end{cases}}$$

where ***analytical formulas*** are given for $\rho(\alpha)$, $\alpha_{c,\min}$ and $\alpha_{c,\max}$

*These results were recently extended in [Mondelli & Montanari, '17], with some technical conditions relaxed

Phase transitions



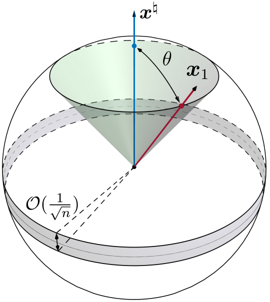
Recall $\alpha = m/n$

Uncorrelated phase: $\alpha < \alpha_{c,\min}$

$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \xrightarrow{\mathcal{P}} 0$ uninformative

$\lambda_1 - \lambda_2 \xrightarrow{\mathcal{P}} 0$ slow convergence

Phase transitions



Recall $\alpha = m/n$

Uncorrelated phase: $\alpha < \alpha_{c,\min}$

$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \xrightarrow{\mathcal{P}} 0$ uninformative

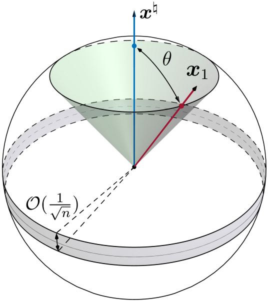
$\lambda_1 - \lambda_2 \xrightarrow{\mathcal{P}} 0$ slow convergence

Correlated phase: $\alpha > \alpha_{c,\max}$

$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \xrightarrow{\mathcal{P}} \rho(\alpha) > 0$ concentration on the surface of a cone

$\lambda_1 - \lambda_2 \xrightarrow{\mathcal{P}} \zeta(\alpha) > 0$ rapid convergence in $\mathcal{O}(\log\, n)$ steps

Phase transitions



Recall $\alpha = m/n$

Uncorrelated phase: $\alpha < \alpha_{c,\min}$

$$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \xrightarrow{\mathcal{P}} 0$$ uninformative

$$\lambda_1 - \lambda_2 \xrightarrow{\mathcal{P}} 0$$ slow convergence

Correlated phase: $\alpha > \alpha_{c,\max}$

$$\rho(\boldsymbol{x}^\natural, \boldsymbol{x}_1) \xrightarrow{\mathcal{P}} \rho(\alpha) > 0$$ concentration on the surface of a cone

$$\lambda_1 - \lambda_2 \xrightarrow{\mathcal{P}} \zeta(\alpha) > 0$$ rapid convergence in $\mathcal{O}(\log\, n)$ steps

Related phenomena: spiked model [Baik, Ben Arous & Peche, '05]
low-rank perturbation of random matrices [Benaych-Georges & Nadakuditi, '11]

Is the asymptotic prediction useful?

Theoretical predictions vs. simulations

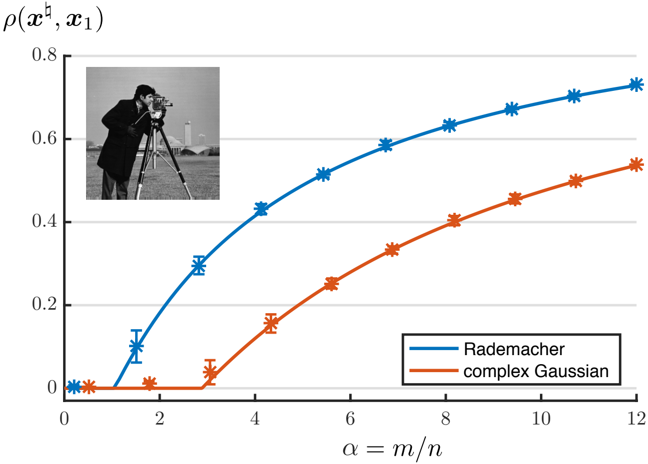
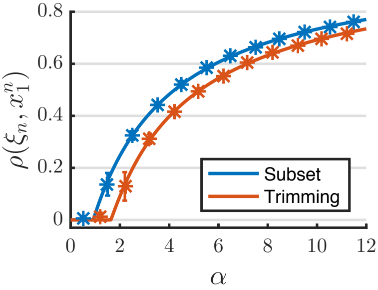


Image size: 64×64

Designing the pre-processing function



Quadratic measurements: $y_i = (\boldsymbol{a}_i^T \boldsymbol{x}^\natural)^2$

$$\boldsymbol{D}_m = \frac{1}{m} \sum_{i=1}^{m} \mathcal{T}(y_i) \boldsymbol{a}_i \boldsymbol{a}_i^T$$

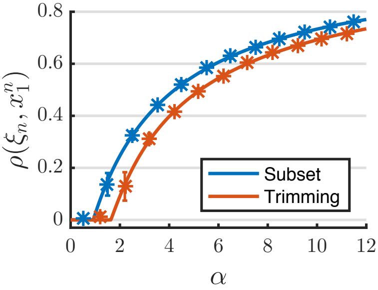
1. ***Trimming*** [Chen & Candes '15]

$$\mathcal{T}(y) = y\, \mathbb{1}_{[0,t]}(y)$$

2. ***Subset*** [Wang, Eldar, Giannakis '16]

$$\mathcal{T}(y) = \mathbb{1}(y_i > t)$$

Designing the pre-processing function



Quadratic measurements: $y_i = (\boldsymbol{a}_i^T \boldsymbol{x}^\natural)^2$

$$\boldsymbol{D}_m = \frac{1}{m} \sum_{i=1}^{m} \mathcal{T}(y_i) \boldsymbol{a}_i \boldsymbol{a}_i^T$$

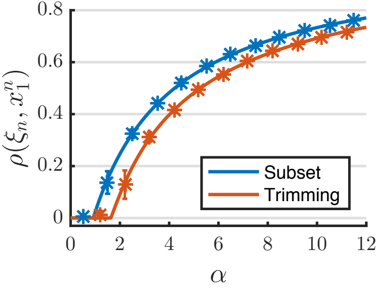
1. ***Trimming*** [Chen & Candes '15]

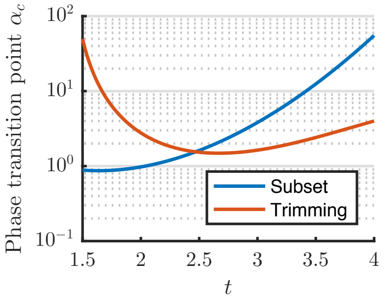
$$\mathcal{T}(y) = y\, \mathbb{1}_{[0,t]}(y)$$

2. ***Subset*** [Wang, Eldar, Giannakis '16]

$$\mathcal{T}(y) = \mathbb{1}(y_i > t)$$

Designing the pre-processing function





Quadratic measurements: $y_i = (\boldsymbol{a}_i^T \boldsymbol{x}^\natural)^2$

$$\boldsymbol{D}_m = \frac{1}{m}\sum_{i=1}^{m} \mathcal{T}(y_i)\boldsymbol{a}_i\boldsymbol{a}_i^T$$

1. ***Trimming*** [Chen & Candes ’15]

$$\mathcal{T}(y) = y\,\mathbb{1}_{[0,t]}(y)$$

2. ***Subset*** [Wang, Eldar, Giannakis ’16]

$$\mathcal{T}(y) = \mathbb{1}(y_i > t)$$

From Sharp Predictions to Optimal Design

For any fixed α, what is the ***optimal*** pre-processing function $\mathcal{T}^*_\alpha(y)$?

$$\boldsymbol{D}_m = \frac{1}{m}\sum_{i=1}^{m}\mathcal{T}(y_i)\boldsymbol{a}_i\boldsymbol{a}_i^T$$

Challenge: functional optimization

[Mondell & Montanari, 2017]: optimal function to minimize phase transition threshold

From Sharp Predictions to Optimal Design

For any fixed α, what is the ***optimal*** pre-processing function $\mathcal{T}_\alpha^*(y)$?

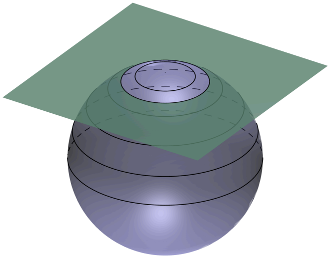
$$\boldsymbol{D}_m = \frac{1}{m}\sum_{i=1}^{m} \mathcal{T}(y_i)\boldsymbol{a}_i\boldsymbol{a}_i^T$$

Challenge: functional optimization

[Mondell & Montanari, 2017]: optimal function to minimize phase transition threshold

Uniformly optimal solution:

$$\boxed{\mathcal{T}^*(y) = 1 - \frac{\mathbb{E}_s[p(y|s)]}{\mathbb{E}_s[s^2 p(y|s)]}}$$



Finding a minimum norm solution in an affine subspace of finite co-dimension

Uniformly Optimal Pre-Processing

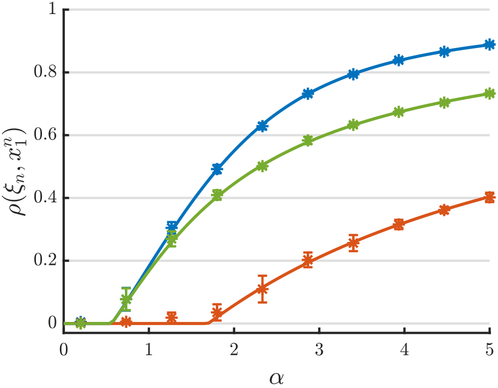
Example:

$$y_i \sim \mathrm{Poisson}[(\boldsymbol{a}_i^T \boldsymbol{x}^\natural)^2] \quad \xrightarrow{\text{optimal}} \quad \mathcal{T}^*(y) = \frac{y-1}{2y+1}$$

Uniformly Optimal Pre-Processing

Example:

$$y_i \sim \mathrm{Poisson}[(\boldsymbol{a}_i^T \boldsymbol{x}^\natural)^2] \xrightarrow{\text{optimal}} \mathcal{T}^*(y) = \frac{y-1}{2y+1}$$



Uniformly optimal solution

[Mondell & Montanari, 2017]

Trimming scheme: [Candes & Chen, '15]

Beyond the Gaussian assumption

Towards physical setups: coded diffraction

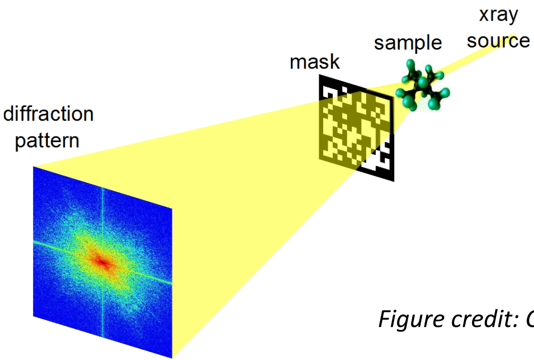


Figure credit: Candes et al. '11

random mask + diffraction

Coded diffraction

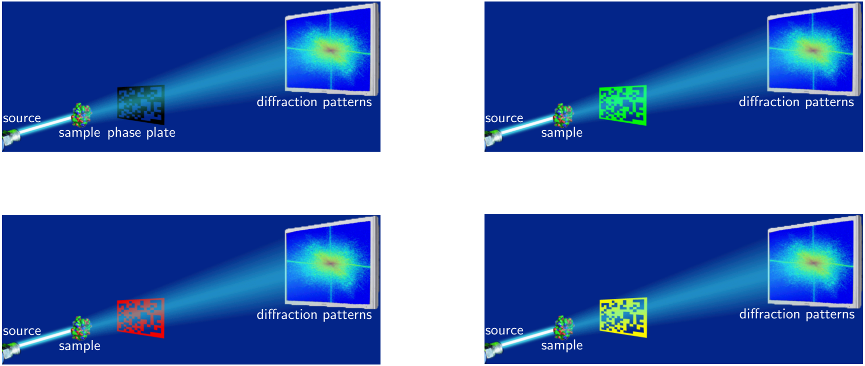
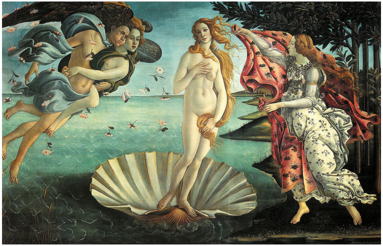


Figure credit: Candes et al. '11

Measurements: Fourier transform of randomly modulated samples

$$\left|\mathcal{F}(\boldsymbol{w} \circ \boldsymbol{x})\right|^2, \qquad \boldsymbol{w} \in \text{Patterns}$$

Performance of spectral method for coded diffraction

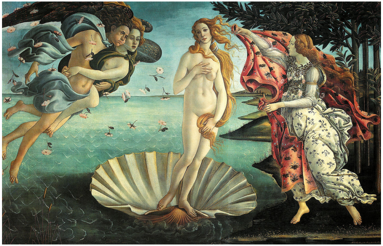


Original image

Figure credit: Mondelli & Montanari, '17

$\alpha = 6$; trimming $\mathcal{T}(\cdot)$

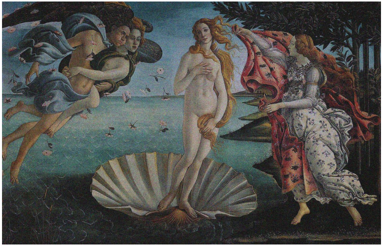
Performance of spectral method for coded diffraction



Original image

Figure credit: Mondelli & Montanari, '17

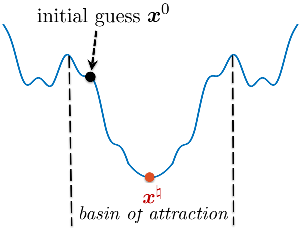
$\alpha = 6$; trimming $\mathcal{T}(\cdot)$



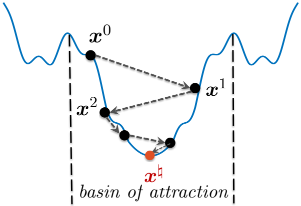
$\alpha = 6$; optimized $\mathcal{T}(\cdot)$

Common theme: two-stage approach

1. ***Initialization***: find an initial point within a local basin close to $\boldsymbol{x}^\natural$



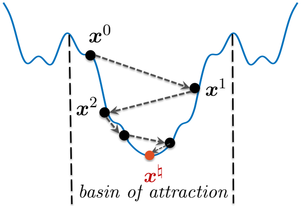
2. Careful iterative ***local refinement*** (e.g. gradient descent) to stay within the local basin



Common theme: two-stage approach

1. ***Initialization***: find an initial point within a local basin close to $x^\natural$

2. Careful iterative ***local refinement*** (e.g. gradient descent) to stay within the local basin



A nonlinear least squares formulation

given: $$y_i = \left| \boldsymbol{a}_i^T \boldsymbol{x}^\natural \right|^2, \qquad i = 1, \ldots, m$$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{i=1}^{m} \left[y_i - (\boldsymbol{a}_i^T \boldsymbol{x})^2 \right]^2$$

A nonlinear least squares formulation

given: $y_i = \left| \boldsymbol{a}_i^T \boldsymbol{x}^\natural \right|^2, \qquad i = 1, \ldots, m$

$\Downarrow$

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{i=1}^{m} \left[y_i - (\boldsymbol{a}_i^T \boldsymbol{x})^2 \right]^2$$

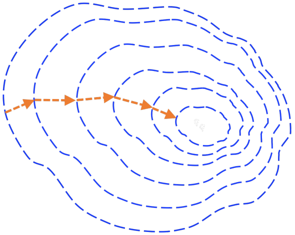
pros: often exact as long as sample size is sufficiently large

cons: $f(\boldsymbol{x})$ is nonconvex

$\longrightarrow$ ***computationally challenging***!

Wirtinger flow (Candès, Li, Soltanolkotabi ’14)

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{i=1}^{m}\left[y_i - (\boldsymbol{a}_i^T\boldsymbol{x})^2\right]^2$$



- ***spectral initialization***: $\boldsymbol{x}^0 \leftarrow$ leading eigenvector of the ***data matrix***

- ***gradient descent***:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta\nabla f(\boldsymbol{x}^t), \qquad t = 0, 1, \ldots$$

Computational cost

$$\boldsymbol{A} := \left[\boldsymbol{a}_i^T \boldsymbol{x}\right]_{1 \leq i \leq m}$$

- ***Spectral initialization***: leading eigenvector $\longrightarrow$ a few applications of $\boldsymbol{A}$ and $\boldsymbol{A}^T$

$$\frac{1}{m}\sum_{i=1}^{m} \mathcal{T}(y_i)\boldsymbol{a}_i\boldsymbol{a}_i^T = \frac{1}{m}\boldsymbol{A}^T \mathrm{diag}\left\{\mathcal{T}(y_i)\right\}\boldsymbol{A}$$

- ***Gradient descent***: one application of $\boldsymbol{A}$ and $\boldsymbol{A}^T$ per iteration

Gradient descent: performance guarantees?

Asymptotic notation

- $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \;\le\; \text{const}$$

- $f(n) \gtrsim g(n)$ means

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \;\ge\; \text{const}$$

- $f(n) \asymp g(n)$ means

$$\text{const}_1 \;\le\; \lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} \;\le\; \text{const}_2$$

First theory of WF

$$\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) := \min\{\|\boldsymbol{x}^t \pm \boldsymbol{x}^\natural\|_2\}$$

Theorem 1 (Candès, Li, Soltanolkotabi '14)

Under i.i.d. Gaussian design, WF with spectral initialization achieves

$$\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \lesssim \left(1 - \frac{\eta}{4}\right)^{t/2} \|\boldsymbol{x}^\natural\|_2,$$

with high prob., provided that step size $\eta \lesssim 1/n$ *and sample size:* $m \gtrsim n \log n$

First theory of WF

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) := \min\{\|\boldsymbol{x}^t \pm \boldsymbol{x}^\natural\|_2\}$$

Theorem 1 (Candès, Li, Soltanolkotabi '14)

Under i.i.d. Gaussian design, WF with spectral initialization achieves

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \lesssim \left(1 - \frac{\eta}{4}\right)^{t/2} \|\boldsymbol{x}^\natural\|_2,$$

with high prob., provided that step size $\eta \lesssim 1/n$ *and sample size:* $m \gtrsim n \log n$

- Iteration complexity: $O(n \log \frac{1}{\epsilon})$

First theory of WF

$$\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) := \min\{\|\boldsymbol{x}^t \pm \boldsymbol{x}^\natural\|_2\}$$

Theorem 1 (Candès, Li, Soltanolkotabi '14)

Under i.i.d. Gaussian design, WF with spectral initialization achieves

$$\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \lesssim \left(1 - \frac{\eta}{4}\right)^{t/2} \|\boldsymbol{x}^\natural\|_2,$$

with high prob., provided that step size $\eta \lesssim 1/n$ *and sample size:* $m \gtrsim n \log n$

- Iteration complexity: $O(n \log \frac{1}{\epsilon})$
- Sample complexity: $O(n \log n)$

First theory of WF

$$\text{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) := \min\{\|\boldsymbol{x}^t \pm \boldsymbol{x}^\natural\|_2\}$$

Theorem 1 (Candès, Li, Soltanolkotabi '14)

Under i.i.d. Gaussian design, WF with spectral initialization achieves

$$\text{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \lesssim \left(1 - \frac{\eta}{4}\right)^{t/2} \|\boldsymbol{x}^\natural\|_2,$$

with high prob., provided that step size and sample size:

- Iteration complexity: $O(n \log \frac{1}{\epsilon})$
- Sample complexity: $O(n \log n)$
- Derived based on (worst-case) local geometry

Improved theory of WF

$$\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) := \min\{\|\boldsymbol{x}^t \pm \boldsymbol{x}^\natural\|_2\}$$

Theorem 2 (Ma, Wang, Chi, Chen '17)

Under i.i.d. Gaussian design, WF with spectral initialization achieves

$$\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\natural\|_2$$

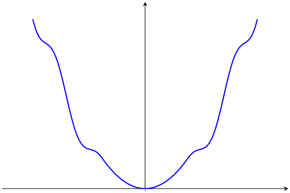
with high prob., provided that step size $\eta \asymp 1/\log n$ *and sample size* $m \gtrsim n \log n$.

- Iteration complexity: $O(n \log \frac{1}{\epsilon}) \searrow O(\log n \log \frac{1}{\epsilon})$
- Sample complexity: $O(n \log n)$
- Derived based on finer analysis of GD trajectory

Gradient descent theory revisited

Consider unconstrained optimization problem

$$\text{minimize}_{\boldsymbol{x}} \qquad f(\boldsymbol{x})$$

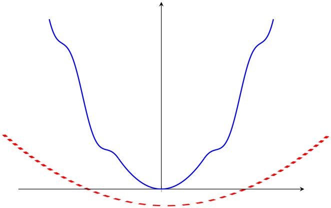


Two standard conditions that enable geometric convergence of GD

Gradient descent theory revisited

Consider unconstrained optimization problem

$$\text{minimize}_{\boldsymbol{x}} \qquad f(\boldsymbol{x})$$



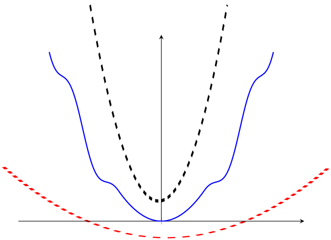
Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity (or regularity condition)

Gradient descent theory revisited

Consider unconstrained optimization problem

$$\text{minimize}_{\boldsymbol{x}} \qquad f(\boldsymbol{x})$$



Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity (or regularity condition)
- (local) smoothness

$$\nabla^2 f(\boldsymbol{x}) \succ \mathbf{0} \quad \text{and} \quad \text{is well-conditioned}$$

Gradient descent theory revisited

f is said to be α-strongly convex and β-smooth if

$$\mathbf{0} \preceq \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{x}$$

ℓ_2 **error contraction:** GD with $\eta = 1/\beta$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^\natural\|_2 \leq \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{x}^t - \boldsymbol{x}^\natural\|_2$$

Gradient descent theory revisited

f is said to be α-strongly convex and β-smooth if

$$\mathbf{0} \preceq \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{x}$$

ℓ_2 **error contraction:** GD with $\eta = 1/\beta$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{\natural}\|_2 \leq \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{x}^{t} - \boldsymbol{x}^{\natural}\|_2$$

- Condition number β/α determines rate of convergence

Gradient descent theory revisited

f is said to be α-strongly convex and β-smooth if

$$\mathbf{0} \preceq \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{x}$$

ℓ_2 **error contraction:** GD with $\eta = 1/\beta$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^{\natural}\|_2 \le \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{x}^{t} - \boldsymbol{x}^{\natural}\|_2$$

- Condition number β/α determines rate of convergence
- Attains ε-accuracy within $O(\frac{\beta}{\alpha} \log \frac{1}{\varepsilon})$ iterations

What does this optimization theory say about WF?

Gaussian designs: $\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \quad 1 \le k \le m$

What does this optimization theory say about WF?

$$\textit{Gaussian designs: } \boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \quad 1 \le k \le m$$

Finite-sample level ($m \asymp n \log n$**)**

$$\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0}$$

What does this optimization theory say about WF?

$$\textit{Gaussian designs: } \boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \quad 1 \le k \le m$$

Finite-sample level ($m \asymp n \log n$)

$$\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0} \quad \underbrace{\text{but ill-conditioned}}_{\text{condition number } \asymp n} \text{ (even locally)}$$

What does this optimization theory say about WF?

$$\textit{Gaussian designs: } \boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \quad 1 \le k \le m$$

Finite-sample level ($m \asymp n \log n$)

$$\nabla^2 f(\boldsymbol{x}) \succ \boldsymbol{0} \quad \underbrace{\text{but ill-conditioned}}_{\text{condition number } \asymp n} \text{ (even locally)}$$

Consequence (Candès et al '14): WF attains ε-accuracy within $O(n \log \frac{1}{\varepsilon})$ iterations if $m \asymp n \log n$

Generic optimization theory gives pessimistic bounds

WF converges in $O(n)$ iterations

Generic optimization theory gives pessimistic bounds

WF converges in $O(n)$ iterations

$$\Uparrow$$

Step size taken to be $\eta = O(1/n)$

Generic optimization theory gives pessimistic bounds

WF converges in $O(n)$ iterations

$\Uparrow$

Step size taken to be $\eta = O(1/n)$

$\Uparrow$

This choice is suggested by worst-case optimization theory

Generic optimization theory gives pessimistic bounds

WF converges in $O(n)$ iterations

$$\Uparrow$$

Step size taken to be $\eta = O(1/n)$

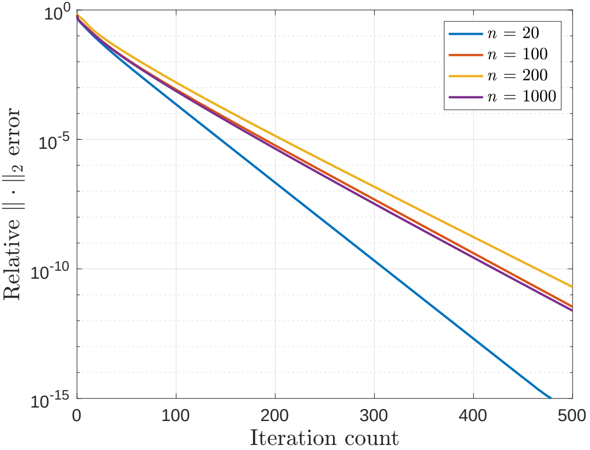
$$\Uparrow$$

This choice is suggested by worst-case optimization theory

$$\Uparrow$$

Does it capture what really happens?

Numerical efficiency with $\eta_t = 0.1$



Vanilla GD (WF) converges fast for a constant step size!

A second look at gradient descent theory

Which local region enjoys both strong convexity and smoothness?

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m}\left[3(\boldsymbol{a}_k^\top \boldsymbol{x})^2 - (\boldsymbol{a}_k^\top \boldsymbol{x}^\natural)^2\right]\boldsymbol{a}_k\boldsymbol{a}_k^\top$$

A second look at gradient descent theory

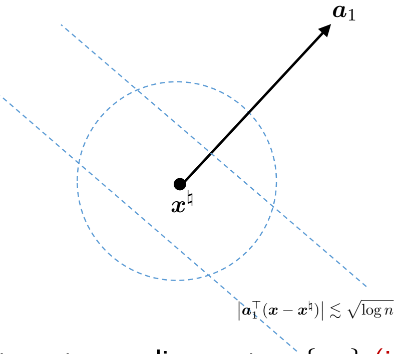
Which local region enjoys both strong convexity and smoothness?

$$\nabla^2 f(\boldsymbol{x}) = \frac{1}{m}\sum_{k=1}^{m}\left[3(\boldsymbol{a}_k^\top \boldsymbol{x})^2 - (\boldsymbol{a}_k^\top \boldsymbol{x}^\natural)^2\right]\boldsymbol{a}_k\boldsymbol{a}_k^\top$$

- Not sufficiently smooth if $\boldsymbol{x}$ and $\boldsymbol{a}_k$ are too close (coherent)

A second look at gradient descent theory

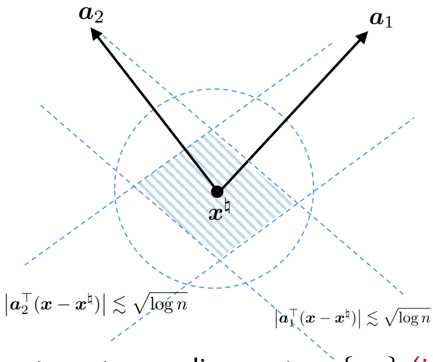
Which local region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors $\{\boldsymbol{a}_k\}$ (incoherence region)

A second look at gradient descent theory

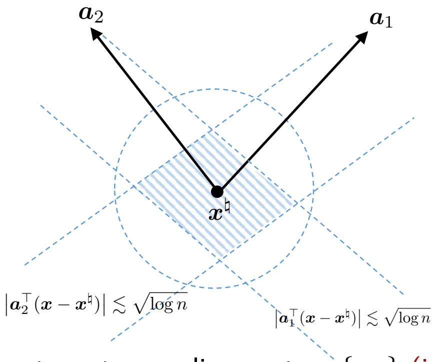
Which local region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors $\{\boldsymbol{a}_k\}$ (incoherence region)

A second look at gradient descent theory

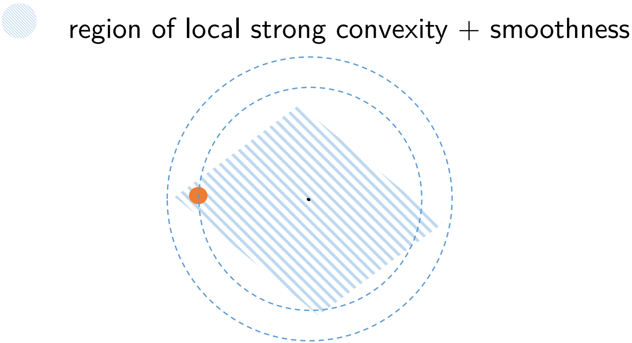
Which local region enjoys both strong convexity and smoothness?



- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors $\{\boldsymbol{a}_k\}$ (incoherence region)

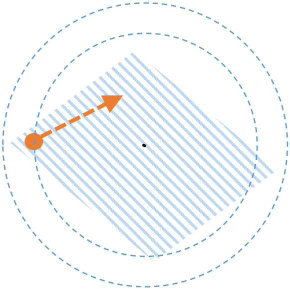
Prior works suggest enforcing regularization (e.g. truncation, projection, regularized loss) to promote incoherence

Encouraging message: GD is implicitly regularized

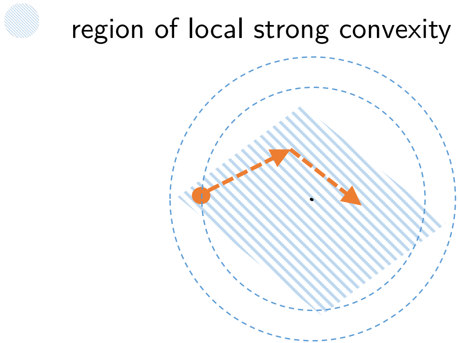


Encouraging message: GD is implicitly regularized

region of local strong convexity + smoothness

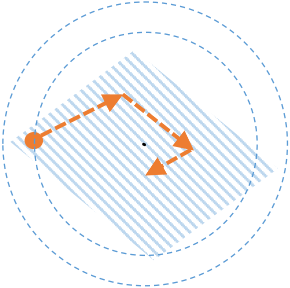


Encouraging message: GD is implicitly regularized



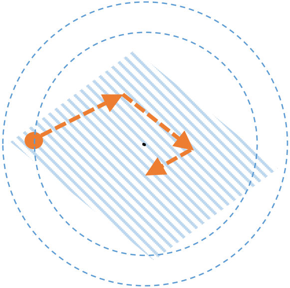
Encouraging message: GD is implicitly regularized

region of local strong convexity + smoothness



Encouraging message: GD is implicitly regularized

region of local strong convexity + smoothness



GD implicitly forces iterates to remain incoherent with $\{\boldsymbol{a}_k\}$

$$\max_k \left|\boldsymbol{a}_k^\top (\boldsymbol{x}^t - \boldsymbol{x}^\natural)\right| \lesssim \sqrt{\log n}\, \|\boldsymbol{x}^\natural\|_2, \quad \forall t$$

— cannot be derived from generic optimization theory; relies on finer statistical analysis for entire trajectory of GD

Theoretical guarantees for local refinement stage

Theorem 3 (Ma, Wang, Chi, Chen '17)

Under i.i.d. Gaussian design, WF with spectral initialization achieves

- $\max_k \left|\boldsymbol{a}_k^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log n}\,\|\boldsymbol{x}^\natural\|_2$ *(incoherence)*

Theoretical guarantees for local refinement stage

Theorem 3 (Ma, Wang, Chi, Chen '17)

Under i.i.d. Gaussian design, WF with spectral initialization achieves

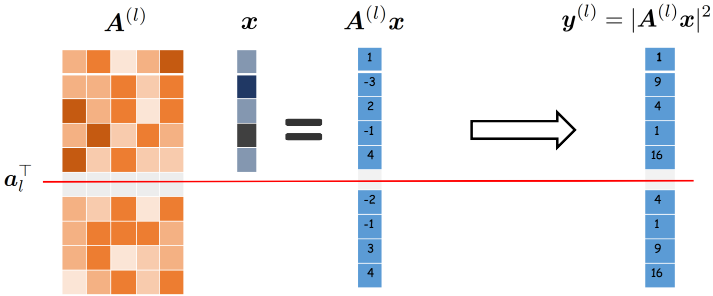
- $\max_k \left|\boldsymbol{a}_k^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log n}\, \|\boldsymbol{x}^\natural\|_2$ *(incoherence)*
- $\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\natural\|_2$ *(linear convergence)*

provided that step size $\eta \asymp 1/\log n$ *and sample size* $m \gtrsim n \log n$.

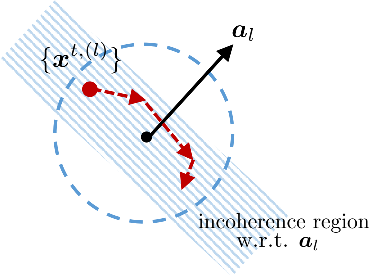
- Attains ε accuracy within $O\big(\log n \, \log \frac{1}{\varepsilon}\big)$ iterations

Key proof idea: leave-one-out analysis

For each $1 \leq l \leq m$, introduce leave-one-out iterates $\boldsymbol{x}^{t,(l)}$ by dropping lth measurement

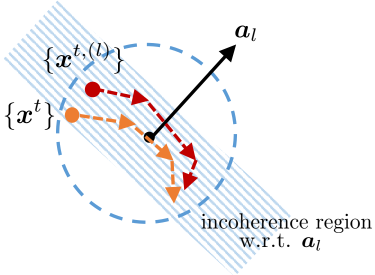


Key proof idea: leave-one-out analysis



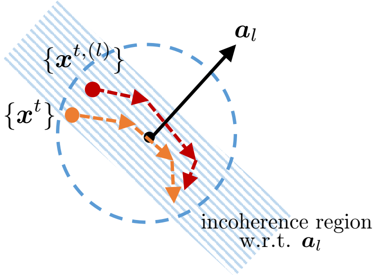
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$

Key proof idea: leave-one-out analysis



- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)} \approx$ true iterate $\boldsymbol{x}^t$

Key proof idea: leave-one-out analysis

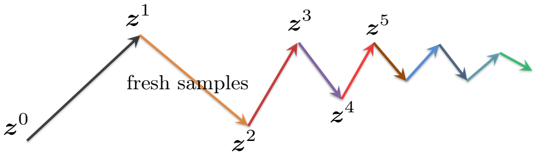


- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)} \approx$ true iterate $\boldsymbol{x}^t$

$$\Longrightarrow \boldsymbol{x}^t \text{ is } \underbrace{\text{nearly independent of}}_{\text{nearly orthogonal to}} \boldsymbol{a}_l$$

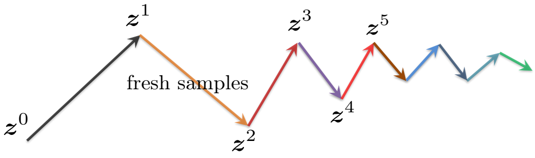
No need of sample splitting

- Several prior works use sample-splitting: require fresh samples at each iteration; not practical but helps analysis

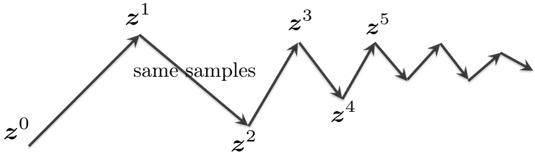


No need of sample splitting

- Several prior works use sample-splitting: require fresh samples at each iteration; not practical but helps analysis



- **This tutorial:** reuses all samples in all iterations



Questions

So far we have presented theory for

$$\text{spectral initialization} \;+\; \text{vanilla gradient descent (WF)}$$

Questions

So far we have presented theory for

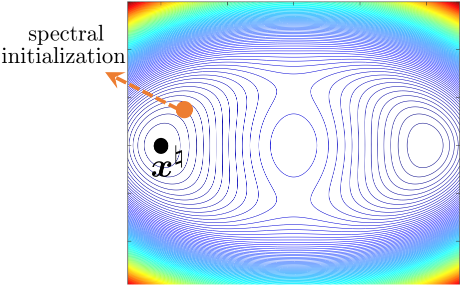
$$\text{spectral initialization} \; + \; \text{vanilla gradient descent (WF)}$$

Questions:

- Is carefully-designed initialization necessary for fast convergence?
- Can we further improve sample complexity?
- Robustness vis a vis noise and outliers?

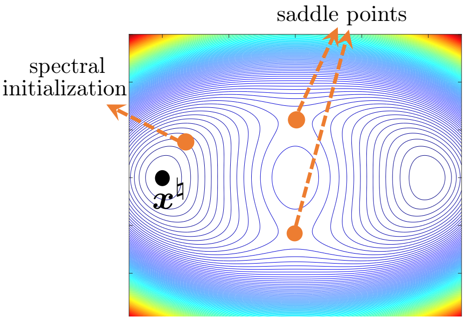
Is carefully-designed initialization necessary for fast convergence?

Initialization



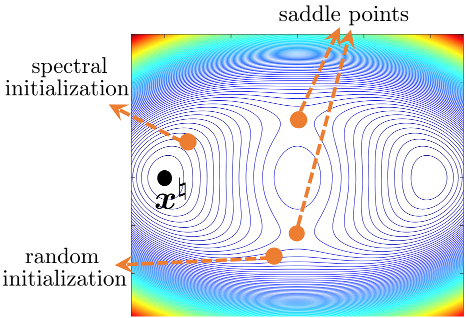
- Spectral initialization gets us reasonably close to truth

Initialization



- Spectral initialization gets us reasonably close to truth
- Cannot initialize GD from anywhere, e.g. it might get stucked at local stationary points (e.g. saddle points)

Initialization

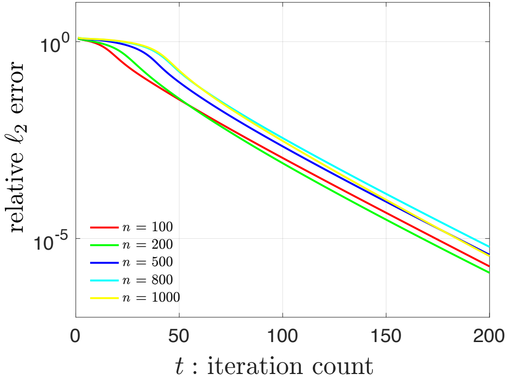


- Spectral initialization gets us reasonably close to truth
- Cannot initialize GD from anywhere, e.g. it might get stucked at local stationary points (e.g. saddle points)

Can we initialize GD randomly, which is simpler and model-agnostic?

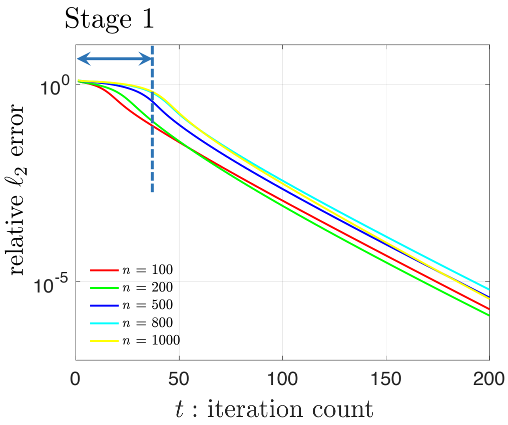
Numerical efficiency of randomly initialized GD

$$\eta_t = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$$



Numerical efficiency of randomly initialized GD

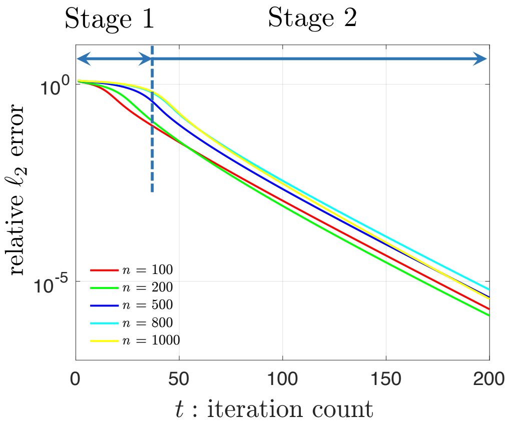
$$\eta_t = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$$



Randomly initialized GD enters local basin within a few iterations

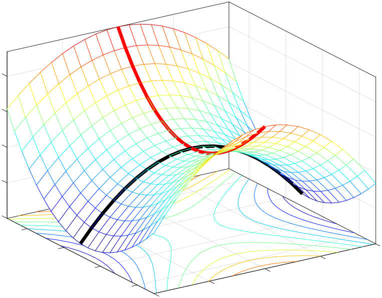
Numerical efficiency of randomly initialized GD

$$\eta_t = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$$



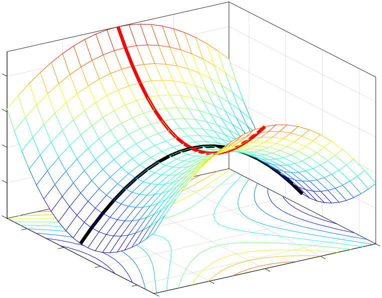
Randomly initialized GD enters local basin within **a few iterations**

A geometric analysis



- if $m \gtrsim n \log^3 n$, then (Sun et al. '16)
    - there is no spurious local mins
    - all saddle points are strict (i.e. associated Hessian matrices have at least one sufficiently negative eigenvalue)

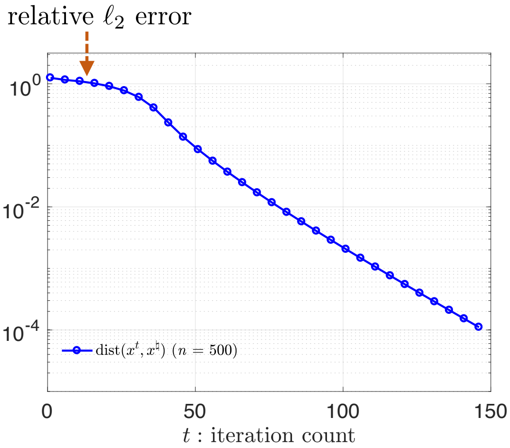
A geometric analysis



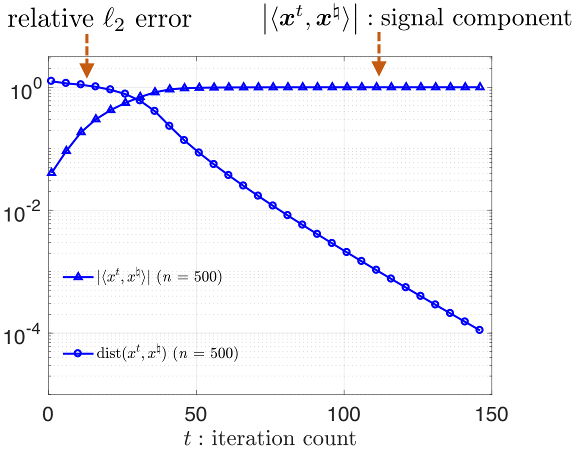
- With such benign landscape, GD with random initialization converges to global min almost surely (Lee et al. '16)

No convergence rate guarantees for vanilla GD!

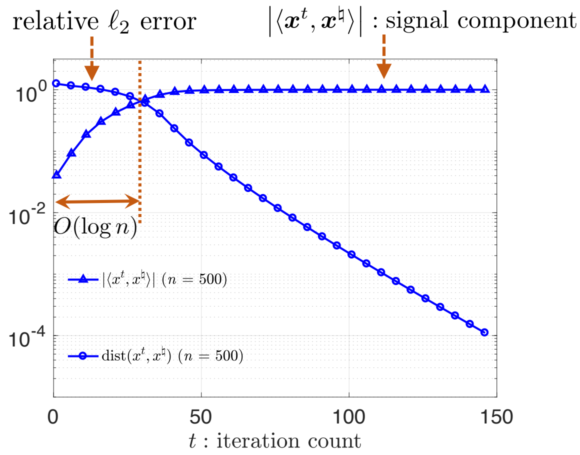
Exponential growth of signal strength in Stage 1



Exponential growth of signal strength in Stage 1

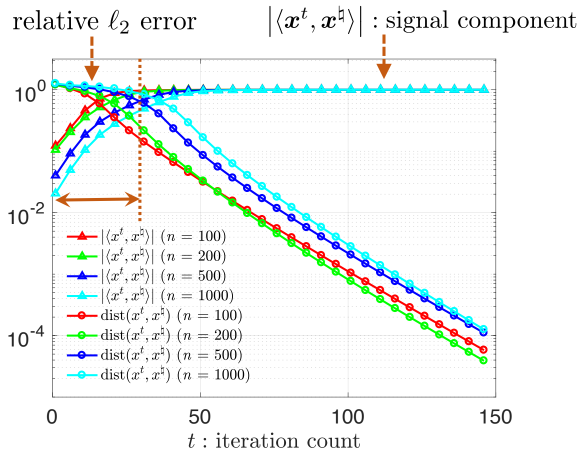


Exponential growth of signal strength in Stage 1



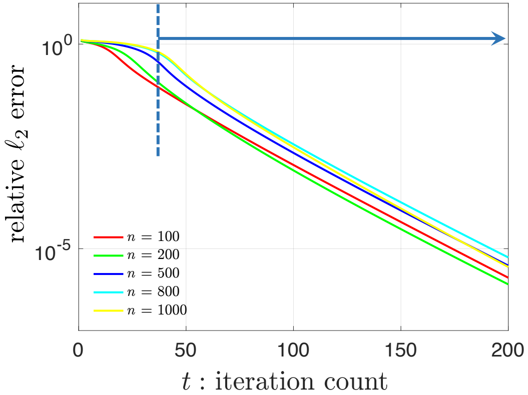
Numerically, $O(\log n)$ iterations are enough to enter local region

Exponential growth of signal strength in Stage 1

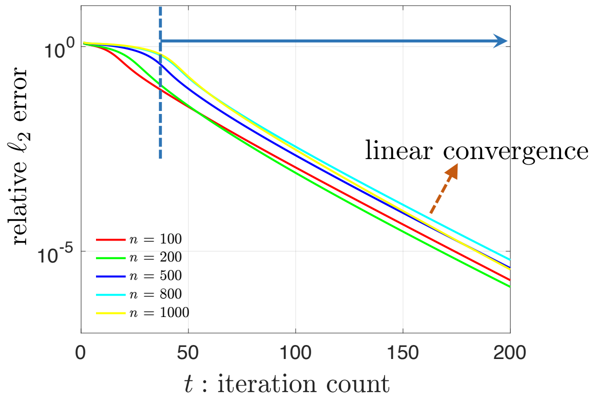


Numerically, $O(\log n)$ iterations are enough to enter local region

Linear / geometric convergence in Stage 2



Linear / geometric convergence in Stage 2



Numerically, GD converges linearly within local region

Theoretical guarantees for randomly initialized GD

These numerical findings can be formalized when $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$:

Theorem 4 (Chen, Chi, Fan, Ma '18)

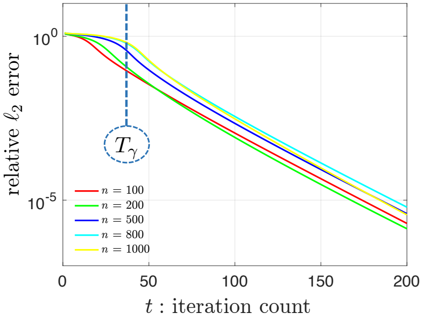
Under i.i.d. Gaussian design, GD with $\boldsymbol{x}^0 \sim \mathcal{N}(\mathbf{0}, n^{-1}\boldsymbol{I}_n)$ achieves

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma(1-\rho)^{t-T_\gamma} \|\boldsymbol{x}^\natural\|_2, \qquad t \geq T_\gamma$$

for $T_\gamma \lesssim \log n$ and some constants $\gamma, \rho > 0$, provided that step size $\eta \asymp 1$ and sample size $m \gtrsim n\,\mathrm{poly}\log m$

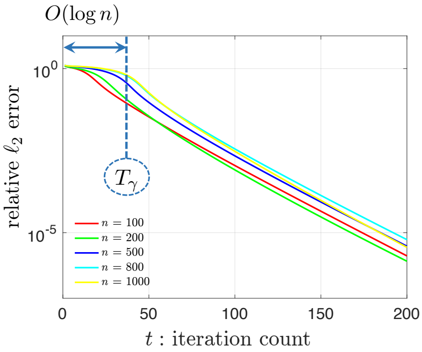
Theoretical guarantees for randomly initialized GD

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \quad t \geq T_\gamma \asymp \log n$$



Theoretical guarantees for randomly initialized GD

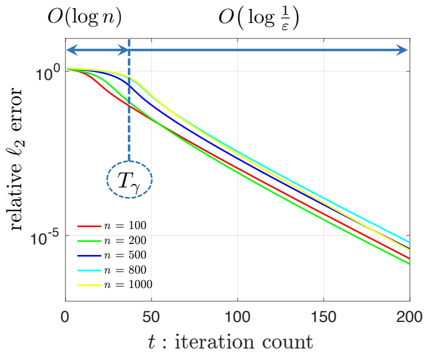
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \quad t \ge T_\gamma \asymp \log n$$



- *Stage 1:* takes $O(\log n)$ iterations to reach $\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \le \gamma$

Theoretical guarantees for randomly initialized GD

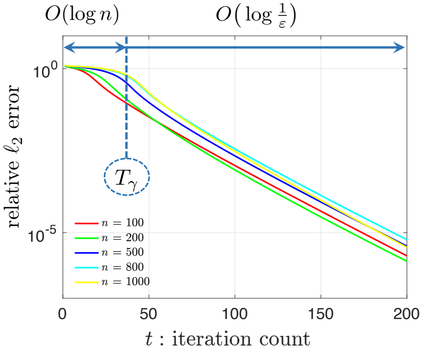
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \quad t \geq T_\gamma \asymp \log n$$



- *Stage 1:* takes $O(\log n)$ iterations to reach $\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma$
- *Stage 2:* linear convergence

Theoretical guarantees for randomly initialized GD

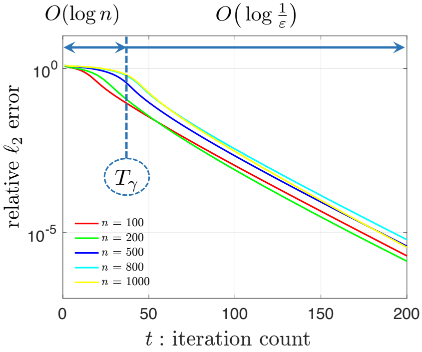
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \quad t \geq T_\gamma \asymp \log n$$



- *near-optimal compututational cost*:
  — $O\left(\log n + \log \frac{1}{\varepsilon}\right)$ iterations to yield ε accuracy

Theoretical guarantees for randomly initialized GD

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \leq \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\natural\|_2, \quad t \geq T_\gamma \asymp \log n$$



- *near-optimal computututational cost*:

  — $O\left(\log n + \log \frac{1}{\varepsilon}\right)$ iterations to yield ε accuracy

- *near-optimal sample size*: $m \gtrsim n \mathrm{poly}\log m$

Experiments on images

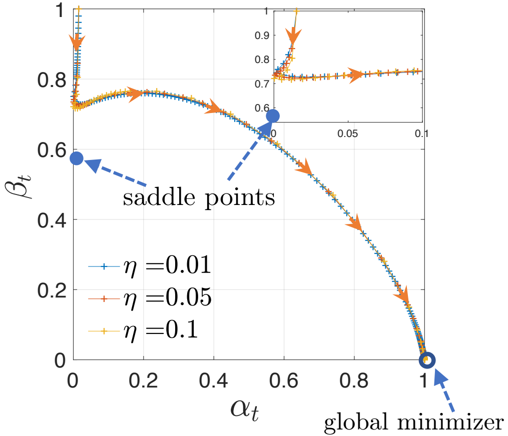
- coded diffraction patterns
- $\boldsymbol{x}^\natural \in \mathbb{R}^{256\times 256}$
- $m/n = 12$

GD with random initialization

$\boldsymbol{x}^t$	$\langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle \boldsymbol{x}^\natural$	$\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\natural \rangle \boldsymbol{x}^\natural$
GD iterate	signal component	perpendicular component

use Adobe Acrobat to see animation

Saddle-escaping schemes?



Randomly initialized GD never hits saddle points in phase retrieval!

Other saddle-escaping schemes

	iteration complexity	**num of iterations needed to escape saddles**	**local iteration complexity**
Trust-region (Sun et al. '16)	$n^7 + \log\log\frac{1}{\varepsilon}$	n^7	$\log\log\frac{1}{\varepsilon}$
Perturbed GD (Jin et al. '17)	$n^3 + n\log\frac{1}{\varepsilon}$	n^3	$n\log\frac{1}{\varepsilon}$
Perturbed accelerated GD (Jin et al. '17)	$n^{2.5} + \sqrt{n}\log\frac{1}{\varepsilon}$	$n^{2.5}$	$\sqrt{n}\log\frac{1}{\varepsilon}$
GD (Chen et al. '18)	$\log n + \log\frac{1}{\varepsilon}$	$\log n$	$\log\frac{1}{\varepsilon}$

Generic optimization theory yields highly suboptimal convergence guarantees

Even simplest possible nonconvex methods are quite efficient for phase retrieval

smart initialization	sample splitting	saddle escaping
		

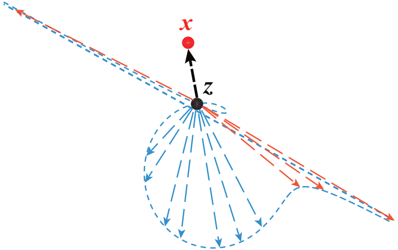
Can we further improve sample complexity?

Improving search directions

$$\text{WF (GD):} \quad \boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m} \sum_k \nabla f_k(\boldsymbol{x}^t)$$

Improving search directions

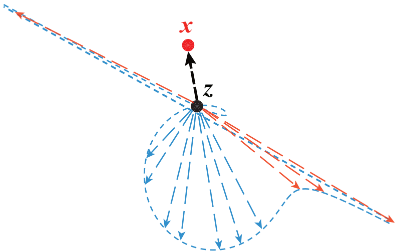
$$\text{WF (GD):} \quad \boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m} \sum_k \nabla f_k(\boldsymbol{x}^t)$$



locus of $\{\nabla f_k(\boldsymbol{z})\}$

Improving search directions

WF (GD): $\quad \boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m} \sum_k \nabla f_k(\boldsymbol{x}^t)$

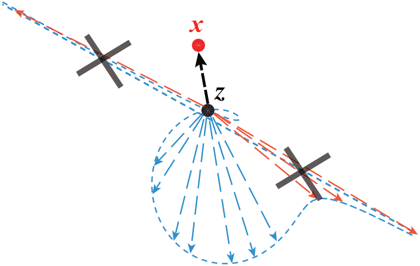


locus of $\{\nabla f_k(\boldsymbol{z})\}$

Problem: descent direction might have large variability

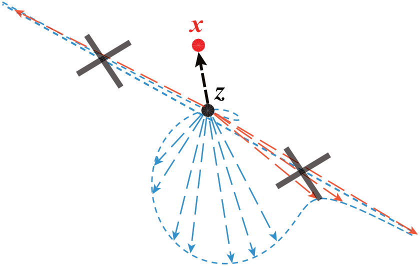
Solution: variance reduction via trimming

More adaptive rule: $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m} \sum_{k \in \mathcal{T}_t} \nabla f_k(\boldsymbol{x}^t)$



Solution: variance reduction via trimming

More adaptive rule: $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m} \sum_{k \in \mathcal{T}_t} \nabla f_k(\boldsymbol{x}^t)$



- $\mathcal{T}_t$ trims away excessively large grad components

$$\mathcal{T}_t := \left\{ k : \quad \left\| \nabla f_k(\boldsymbol{x}^t) \right\|_2 \lesssim \text{typical-size} \left\{ \left\| \nabla f_l(\boldsymbol{x}^t) \right\|_2 \right\}_{1 \le l \le m} \right\}$$

Slight bias $\quad + \quad$ much reduced variance

Summary: truncated Wirtinger flow

(1) **Regularized spectral initialization:** $\boldsymbol{x}^0 \leftarrow$ principal component of

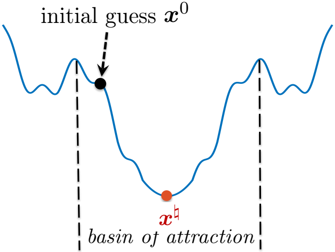
$$\frac{1}{m}\sum\nolimits_{k\in\mathcal{T}_0} y_k\, \boldsymbol{a}_k\boldsymbol{a}_k^*$$

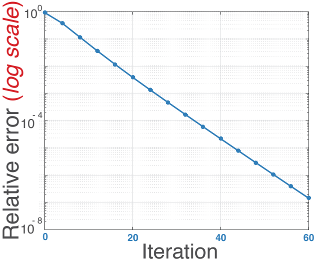
(2) Follow **adaptive gradient descent**

$$\boldsymbol{x}^t = \boldsymbol{x}^t - \frac{\eta_t}{m}\sum\nolimits_{k\in\mathcal{T}_t} \nabla f_k(\boldsymbol{x}^t)$$

Adaptive and iteration-varying rules: discard high-leverage data $\{y_k : k \notin \mathcal{T}_t\}$

Theoretical guarantees (noiseless data)





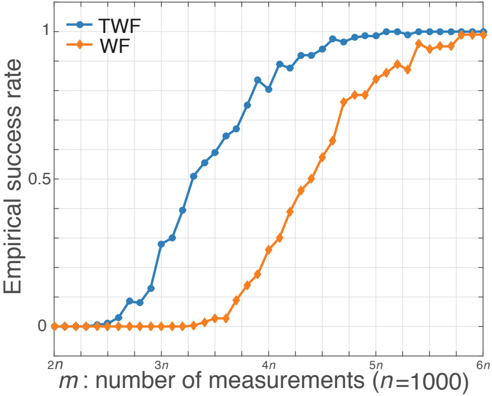
Theorem 5 (Chen, Candès '15)

Suppose $\boldsymbol{a}_k \overset{i.i.d.}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$ *and sample size* $m \gtrsim n$*. With high prob.,*

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) := \min \|\boldsymbol{x}^t \pm \boldsymbol{x}^\natural\|_2 \leq \nu (1-\rho)^t \|\boldsymbol{x}\|_2$$

where $0 < \nu, \rho < 1$ *are universal constants*

Empirical success rate (noiseless data)



Empirical success rate vs. sample size

Stability vis a vis noise and outliers?

Stability under noisy data

- Noisy data: $y_k = |\boldsymbol{a}_k^* \boldsymbol{x}^\natural|^2 + \eta_k$
- Signal-to-noise ratio:

$$\mathsf{SNR} := \frac{\sum_k |\boldsymbol{a}_k^* \boldsymbol{x}^\natural|^4}{\sum_k \eta_k^2} \approx \frac{3m\|\boldsymbol{x}^\natural\|^4}{\|\boldsymbol{\eta}\|^2}$$

- i.i.d. Gaussian design $\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$

Stability under noisy data

- Noisy data: $y_k = |\boldsymbol{a}_k^* \boldsymbol{x}^\natural|^2 + \eta_k$
- Signal-to-noise ratio:

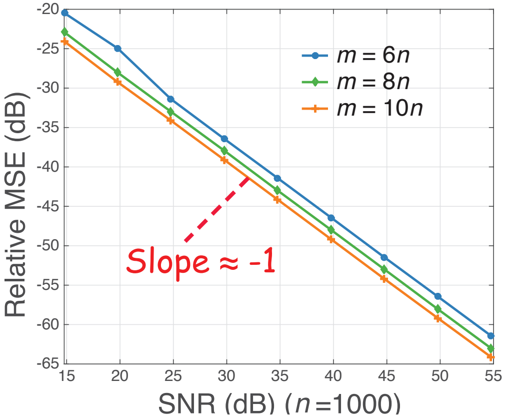
$$\mathsf{SNR} := \frac{\sum_k |\boldsymbol{a}_k^* \boldsymbol{x}^\natural|^4}{\sum_k \eta_k^2} \approx \frac{3m\|\boldsymbol{x}^\natural\|^4}{\|\boldsymbol{\eta}\|^2}$$

- i.i.d. Gaussian design $\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$

Theorem 6 (Chen, Candès '15)

Relative error of TWF converges to $O(\frac{1}{\sqrt{\mathsf{SNR}}})$

Relative MSE vs. SNR (Poisson data)

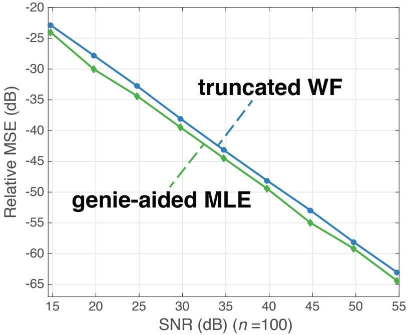


Empirical evidence: relative MSE scales inversely with SNR

This accuracy is nearly un-improvable (empirically)

Comparison with ideal MLE (with phase info. revealed)

ideal knowledge: $y_k \sim \text{Poisson}\left(\left| \boldsymbol{a}_k^* \boldsymbol{x}^\natural \right|^2 \right)$ and $\varepsilon_k = \text{sign}(\boldsymbol{a}_k^* \boldsymbol{x}^\natural)$



Little loss due to missing phases!

This accuracy is nearly un-improvable (theoretically)

- Poisson data: $y_k \overset{\text{ind.}}{\sim} \mathsf{Poisson}(\,|\boldsymbol{a}_k^* \boldsymbol{x}^\natural|^2\,)$
- Signal-to-noise ratio:

$$\mathsf{SNR} \;\approx\; \frac{\sum_k |\boldsymbol{a}_k^* \boldsymbol{x}^\natural|^4}{\sum_k \mathsf{Var}(y_k)} \;\approx\; 3\|\boldsymbol{x}^\natural\|_2^2$$

This accuracy is nearly un-improvable (theoretically)

- Poisson data: $y_k \overset{\text{ind.}}{\sim} \text{Poisson}(\,|\boldsymbol{a}_k^* \boldsymbol{x}^\natural|^2\,)$
- Signal-to-noise ratio:

$$\text{SNR} \;\approx\; \frac{\sum_k |\boldsymbol{a}_k^* \boldsymbol{x}^\natural|^4}{\sum_k \text{Var}(y_k)} \;\approx\; 3\|\boldsymbol{x}^\natural\|_2^2$$

Theorem 7 (Chen, Candès '15)

. *Under i.i.d. Gaussian design, for any estimator* $\widehat{\boldsymbol{x}}$*,*

$$\inf_{\widehat{\boldsymbol{x}}} \sup_{\boldsymbol{x}:\ \|\boldsymbol{x}\|_2 \geq \log^{1.5} m} \frac{\mathbb{E}\left[\text{dist}\left(\widehat{\boldsymbol{x}}, \boldsymbol{x}\right) \mid \{\boldsymbol{a}_k\}\right]}{\|\boldsymbol{x}\|_2} \;\gtrsim\; \frac{1}{\sqrt{\text{SNR}}},$$

provided that sample size $m \asymp n$

Robust recovery vis a vis outliers

Consider now two sources of corruption: *sparse outliers* and *bounded noise*

$$y_i = |\boldsymbol{a}_i^\top \boldsymbol{x}^\natural|^2 + \eta_i + w_i, \quad i = 1, \ldots, m,$$

- $\|\boldsymbol{\eta}\|_0 \leq s \cdot m$: sparse outlier, where $0 \leq s < 1$ is fraction of outliers
- $\boldsymbol{w}$: bounded noise

Motivation: outliers happen with sensor failures, malicious attacks ...

Robust recovery vis a vis outliers

Goal: develop algorithms that are *oblivious* to outliers, and statistically and computationally efficient

- performs equally well regardless of existence of outliers
- small sample size: ideally $m \asymp n$
- large fraction of outliers: ideally $s \asymp 1$
- low computational complexity and easy to implement

Existing approaches are not robust in the presence of arbitrary outliers

- **Spectral initialization would fail:** leading eigenvector of $\boldsymbol{Y}$ can be arbitrarily perturbed

$$\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \qquad \text{(WF)}$$

$$\text{or} \quad \boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \mathbb{1}_{\{|y_i| \lesssim \mathsf{mean}(\{y_i\})\}} \qquad \text{(TWF)}$$

Existing approaches are not robust in the presence of arbitrary outliers

- **Spectral initialization would fail:** leading eigenvector of $\boldsymbol{Y}$ can be arbitrarily perturbed

$$\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \qquad \text{(WF)}$$

$$\text{or} \quad \boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \mathbb{1}_{\{|y_i| \lesssim \mathsf{mean}(\{y_i\})\}} \qquad \text{(TWF)}$$

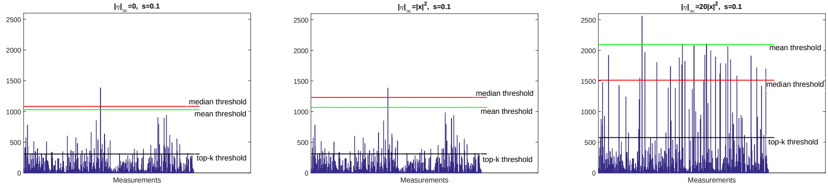
- **GD would fail:** search directions can be arbitrarily perturbed

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m}\sum_{i=1}^{m} \nabla f_k(\boldsymbol{x}^t)$$

Solution: median truncation

Median is often more stable for various levels of outliers

- well-known in robust statistics to be outlier-resilient

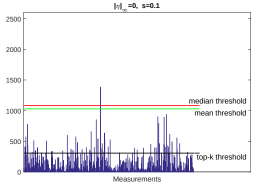


no outliers small outlier magnitudes large outlier magnitudes

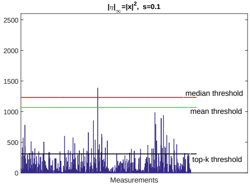
Solution: median truncation

Median is often more stable for various levels of outliers

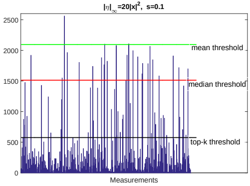
- well-known in robust statistics to be outlier-resilient



no outliers



small outlier magnitudes



large outlier magnitudes

Key idea: "median-truncation" — discard samples *adaptively* based on how large sample gradients/values deviate from median

Median-truncated gradient descent

(1) **Median-truncated spectral initialization:** $\boldsymbol{x}^0 \leftarrow$ leading eigenvector of

$$\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \mathbb{1}_{\{|y_i| \lesssim \text{median}(\{y_i\})\}}$$

(2) **Median-truncated gradient descent:**

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m}\sum_{k \in \mathcal{T}_t} \nabla f_k(\boldsymbol{x}^t),$$

where

$$\mathcal{T}_t = \left\{k : \left|y_k - |\boldsymbol{a}_k^\top \boldsymbol{x}^t|\right| \lesssim \text{median}\left(\left\{\left|y_k - |\boldsymbol{a}_k^\top \boldsymbol{x}^t|\right|\right\}\right)\right\}$$

Performance guarantees

Theorem 8 (Zhang, Chi and Liang '16)

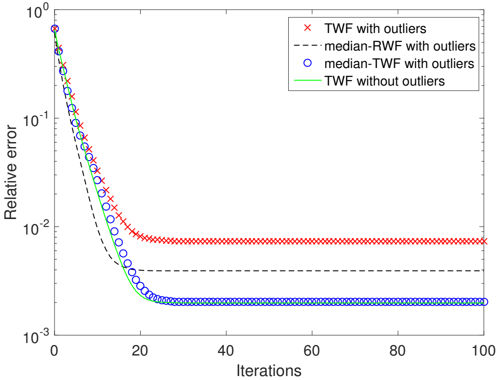
Assume $\|\boldsymbol{w}\|_\infty \le c_1 \|\boldsymbol{x}^\natural\|_2^2$, *and* $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. *If* $m \gtrsim n \log n$ *and* $s \lesssim s_0$, *then with high prob., median-TWF/RWF yields*

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\natural) \lesssim \frac{\|\boldsymbol{w}\|_\infty}{\|\boldsymbol{x}^\natural\|_2} + (1-\rho)^t \|\boldsymbol{x}^\natural\|_2, \quad t = 0, 1, \cdots$$

for some constants $0 < \rho, s_0 < 1$

- **Exact recovery** when $\boldsymbol{w} = \boldsymbol{0}$ but with a constant fraction of outliers $s \asymp 1$
- **Stable recovery** with additional bounded noise
- Resist outliers **obliviously**: no prior knowledge of outliers (except sparsity)

Numerical experiment with both dense noise and sparse outliers



Median-TWF with outliers achieves almost identical accuracy as TWF without outliers

Tutorial outline

- Part I: Overview
- Part II: Phase retrieval: a case study
  - Spectral initialization
  - Local refinement: algorithm and analysis
- Part III: Low-rank matrix estimation
- Part IV: Closing remarks

Motivation

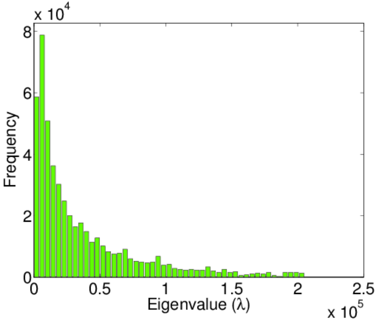
Low-rank matrix estimation problems arise in many applications

A popular example is **recommendation systems**: how to predict unseen user ratings for movies?

figure credit: E. Candes

Low-rank modeling

figure credit: E. Candes



A few factors explain most of the data

Low-rank modeling

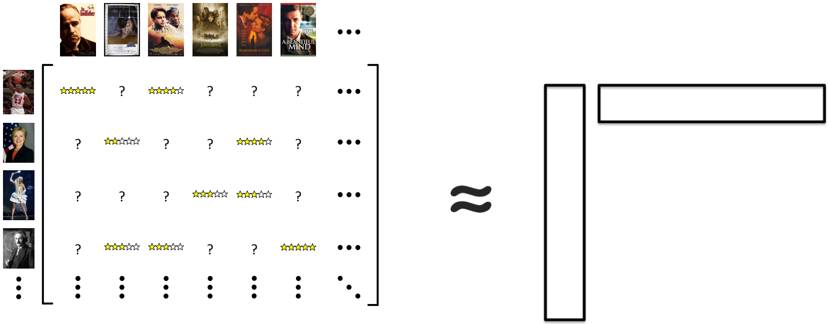


figure credit: E. Candes

A few factors explain most of the data $\longrightarrow$ low-rank approximation

How to exploit (approx.) low-rank structure in prediction?

Other problems with low-rank matrices

- sensor network localization
- structure from motion
- system identification and time series analysis
- spatial-temporal data modeling, e.g. video, network traffic, ..
- face recognition
- quantum state tomography
- community detection
- ...

Rank-constrained optimization

Rank-constrained optimization:

$$\text{minimize}_{\boldsymbol{M} \in \mathbb{R}^{n \times n}} \quad F(\boldsymbol{M}) \quad \text{s.t.} \quad \text{rank}(\boldsymbol{M}) \leq r,$$

where $F(\boldsymbol{M})$ is convex in $\boldsymbol{M}$, and $r \ll n$

- useful model for many low-rank estimation problems;
- computationally intractable.

Convex relaxation

Convex relaxation:

$$\text{minimize}_{\boldsymbol{M}\in\mathbb{R}^{n\times n}} \quad F(\boldsymbol{M}) \quad \text{s.t.} \quad \|\boldsymbol{M}\|_* \le \zeta$$

where $\|\cdot\|_*$ is nuclear norm — convex relaxation of rank

- **Pros:** mature theory; versatile to incorporate other constraints
- **Cons:** run-time in $O(n^3)$; even $\boldsymbol{M}$ itself takes $O(n^2)$ storage

Question: can we develop algorithms that work with computational and memory complexities nearly linear in n?

Burer-Monteiro factorization

Matrix factorization:

$$\text{minimize}_{\boldsymbol{U},\boldsymbol{V}} \; f(\boldsymbol{U},\boldsymbol{V}) := F(\boldsymbol{U}\boldsymbol{V}^\top)$$

where $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U},\boldsymbol{V} \in \mathbb{R}^{n\times r}$.

- pioneered by Burer, Monteiro '03
- highly non-convex
- global ambiguity: for any orthonormal $\boldsymbol{R} \in \mathbb{R}^{r\times r}$ and $\alpha \neq 0$,

$$\boldsymbol{U}\boldsymbol{V}^\top = (\alpha \boldsymbol{U}\boldsymbol{R})(\alpha^{-1}\boldsymbol{V}\boldsymbol{R})^\top$$

i.e. if $(\boldsymbol{U},\boldsymbol{V})$ is a global minimizer, so does $(\alpha\boldsymbol{U}\boldsymbol{R}, \alpha^{-1}\boldsymbol{V}\boldsymbol{R})$

Revisiting PCA

Given PSD $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ (not necessarily low-rank), solve *low-rank approximation problem* (best rank-r approximation):

$$\underbrace{\widehat{\boldsymbol{M}} = \text{argmin}_{\boldsymbol{Z}} \ \|\boldsymbol{Z}-\boldsymbol{M}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \text{rank}(\boldsymbol{Z}) \leq r}_{\text{nonconvex optimization!}}$$

Solution is truncated eigen-decomposition (Eckart-Young theorem)

- let $\boldsymbol{M} = \sum_{i=1}^n \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ be EVD of $\boldsymbol{M}$ ($\sigma_1 \geq \cdots \geq \sigma_n$), then

$$\widehat{\boldsymbol{M}} = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$$

— *nonconvex, but tractable*

Optimization viewpoint

Factorize $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{X}^\top$ with $\boldsymbol{X} \in \mathbb{R}^{n \times r}$. We're interested in the landscape of

$$f(\boldsymbol{X}) := \frac{1}{4}\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$$

Optimization viewpoint

Factorize $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{X}^\top$ with $\boldsymbol{X} \in \mathbb{R}^{n\times r}$. We're interested in the landscape of

$$f(\boldsymbol{X}) := \frac{1}{4}\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$$

To simplify exposition: set $r = 1$.

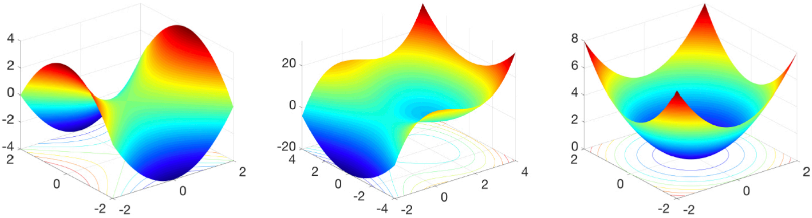
$$f(\boldsymbol{x}) = \frac{1}{4}\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$$

Definition 9 (critical points)

A first-order critical point (stationary point) of f satisfies

$$\nabla f(\boldsymbol{x}) = \boldsymbol{0}$$

Several types of critical points



(a) strict saddle (b) local minimum (c) global minimum

Figure credit: Li et al. '16

Critical points of $f(\boldsymbol{x})$

$$\boldsymbol{x} \text{ is critical point, i.e. } \nabla f(\boldsymbol{x}) = (\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{M})\boldsymbol{x} = \mathbf{0}$$

$$\Updownarrow$$

$$\boldsymbol{M}\boldsymbol{x} = \|\boldsymbol{x}\|_2^2 \boldsymbol{x}$$

$$\Updownarrow$$

$$\boldsymbol{x} \text{ aligns with eigenvectors of } \boldsymbol{M} \quad \text{or} \quad \boldsymbol{x} = \mathbf{0}$$

Since $\boldsymbol{M}\boldsymbol{u}_i = \sigma_i \boldsymbol{u}_i$, set of critical points is given by

$$\{\mathbf{0}\} \cup \{\sqrt{\sigma_i}\boldsymbol{u}_i, \ i = 1, \ldots, n\}$$

Categorization of critical points

Critical points can be further categorized based on **Hessians**:

$$\nabla^2 f(\boldsymbol{x}) := 2\boldsymbol{x}\boldsymbol{x}^\top + \|\boldsymbol{x}\|_2^2 \boldsymbol{I} - \boldsymbol{M}$$

- For any non-zero critical points $\boldsymbol{x}_k := \sqrt{\sigma_k}\boldsymbol{u}_k$:

$$\begin{aligned}
\nabla^2 f(\boldsymbol{x}_k) &= 2\sigma_k \boldsymbol{u}_k \boldsymbol{u}_k^\top + \sigma_k \boldsymbol{I} - \boldsymbol{M} \\
&= 2\sigma_k \boldsymbol{u}_k \boldsymbol{u}_k^\top + \sigma_k \left(\sum_{i=1}^n \boldsymbol{u}_i \boldsymbol{u}_i^\top \right) - \sum_{i=1}^n \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top \\
&= \sum_{i:i\neq k} (\sigma_k - \sigma_i) \boldsymbol{u}_i \boldsymbol{u}_i^\top + 2\sigma_k \boldsymbol{u}_k \boldsymbol{u}_k^\top
\end{aligned}$$

Categorization of critical points

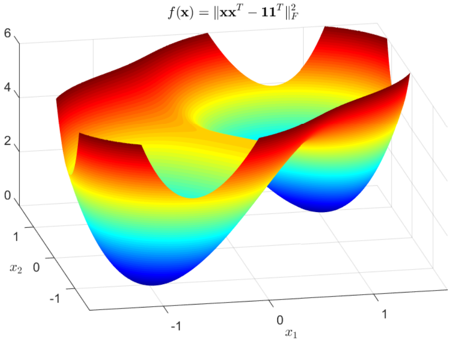
Critical points can be further categorized based on **Hessians**:

$$\nabla^2 f(\boldsymbol{x}) := 2\boldsymbol{x}\boldsymbol{x}^\top + \|\boldsymbol{x}\|_2^2 \boldsymbol{I} - \boldsymbol{M}$$

- If $\sigma_1 > \sigma_2 \geq \ldots \geq \sigma_n \geq 0$, then
  - $k = 1$: $\nabla^2 f(\boldsymbol{x}_1) \succ \boldsymbol{0} \quad \rightarrow \quad$ local minima
  - $1 < k \leq n$: $\lambda_{\min}(\nabla^2 f(\boldsymbol{x}_k)) < 0$, $\lambda_{\max}(\nabla^2 f(\boldsymbol{x}_k)) > 0$
    $\rightarrow \quad$ strict saddle
  - $\boldsymbol{x} = \boldsymbol{0}$: $\nabla^2 f(\boldsymbol{0}) \preceq \boldsymbol{0} \quad \rightarrow \quad$ local maxima (or strict saddle)

Good news: benign landscape

For example, for 2-dimensional case $f(\boldsymbol{x}) = \left\| \boldsymbol{x}\boldsymbol{x}^\top - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\|_{\mathrm{F}}^2$



global minima $\boldsymbol{x} = \pm \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ & strict saddle $\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\pm \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

— *No "spurious" local minima!*

Key messages from landscape analysis

$$f(\boldsymbol{X}) := \frac{1}{4}\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2, \qquad \boldsymbol{X} \in \mathbb{R}^{n\times r}$$

If $\sigma_r > \sigma_{r+1}$:

- **all local minima are global:** $\boldsymbol{X}$ contains top-r eigenvectors (up to orthonormal transformation)
- **strict saddle points:** all stationary points are saddle points except global optimum

Low-rank recovery with few measurements

Consider linear measurements:

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{M}), \quad \boldsymbol{y} \in \mathbb{R}^m, \quad m \ll n^2$$

where $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ is rank-r $(r \ll n)$ and PSD (for simplicity).

- Consider least-squares loss function:

$$f(\boldsymbol{X}) := \frac{1}{4}\|\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M})\|_{\mathrm{F}}^2$$

- If $\mathcal{A}$ is isotropic (i.e. $\mathbb{E}[\mathcal{A}^*\mathcal{A}] = \mathcal{I}$), then

$$\mathbb{E}[f(\boldsymbol{X})] = \frac{1}{4}\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2$$

- Does $f(\boldsymbol{X})$ inherit benign landscape?

Landscape preserving under RIP

Definition 10

Rank-r restricted isometry constants δ_r is smallest quantity obeying

$$(1-\delta_r)\|\boldsymbol{M}\|_{\mathsf{F}}^2 \le \|\mathcal{A}(\boldsymbol{M})\|_{\mathsf{F}}^2 \le (1+\delta_r)\|\boldsymbol{M}\|_{\mathsf{F}}^2, \ \forall \boldsymbol{M} : \mathsf{rank}(\boldsymbol{M}) \le r$$

Landscape preserving under RIP

Definition 10

Rank-r restricted isometry constants δ_r is smallest quantity obeying

$$(1-\delta_r)\|\boldsymbol{M}\|_{\mathsf{F}}^2 \le \|\mathcal{A}(\boldsymbol{M})\|_{\mathsf{F}}^2 \le (1+\delta_r)\|\boldsymbol{M}\|_{\mathsf{F}}^2, \ \forall \boldsymbol{M} : \mathsf{rank}(\boldsymbol{M}) \le r$$

Key message: benign landscape is preserved when $\mathcal{A}$ satisfies RIP e.g., when $\mathcal{A}$ follows the Gaussian design

Landscape preserving under RIP

Definition 10

Rank-r restricted isometry constants δ_r is smallest quantity obeying

$$(1-\delta_r)\|\boldsymbol{M}\|_{\mathsf{F}}^2 \le \|\mathcal{A}(\boldsymbol{M})\|_{\mathsf{F}}^2 \le (1+\delta_r)\|\boldsymbol{M}\|_{\mathsf{F}}^2, \ \forall \boldsymbol{M} : \mathsf{rank}(\boldsymbol{M}) \le r$$

Theorem 11 (Bhojanapalli et al. '16, Ge et al. '17)

If $\mathcal{A}$ satisfies RIP with $\delta_{2r} < \frac{1}{10}$, then

- *all local min are global: any local minimum $\boldsymbol{X}$ of $f(\cdot)$ satisfies $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{M}$*
- *strict saddle points: any non-local min critical point $\boldsymbol{X}$ of $f(\cdot)$ satisfies $\lambda_{\min}[\nabla^2 f(\boldsymbol{X})] \le -\frac{2}{5}\sigma_r$*

Landscape without RIP

Matrix completion:

$$\text{Complete } \boldsymbol{M} \text{ from partial entries } M_{i,j},\ (i,j) \in \Omega$$

where (i,j) is included in Ω independently with prob. p

$$\text{find low-rank } \widehat{\boldsymbol{M}} \quad \text{s.t.} \quad \mathcal{P}_{\Omega}(\widehat{\boldsymbol{M}}) = \mathcal{P}_{\Omega}(\boldsymbol{M})$$

In matrix completion, RIP does not hold
$\quad\rightarrow\quad$ need to regularize loss function by promoting **incoherent** solutions

Incoherence for matrix completion

Definition 12 (Incoherence for matrix completion)

A rank-r matrix $\boldsymbol{M}$ with eigendecomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$ is said to be μ-incoherent if

$$\|\boldsymbol{U}\|_{2,\infty} \leq \sqrt{\frac{\mu}{n}}\|\boldsymbol{U}\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{n}}.$$

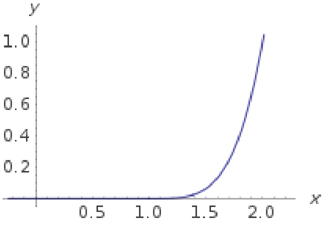
e.g. $\underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}}_{\text{hard } \mu=n}$ vs. $\underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}}_{\text{easy } \mu=1}$

Regularization

One possible regularizer:

$$Q(\boldsymbol{X}) = \sum_{i=1}^{n} (\underbrace{\|\boldsymbol{e}_i^\top \boldsymbol{X}\|_2}_{\text{row norm}} - \alpha)_+^4 := \sum_{i=1}^{n} Q_i(\|\boldsymbol{e}_i^\top \boldsymbol{X}\|_2)$$

where α is regularization parameter, and $z_+ = \max\{z, 0\}$



(x from −0.2 to 2.2)

MC has no spurious local minima under proper regularization

Consider *regularized* loss function

$$f_{\mathsf{reg}}(\boldsymbol{X}) = \frac{1}{p}\|\mathcal{P}_{\Omega}(\boldsymbol{X}\boldsymbol{X}^{\top} - \boldsymbol{M})\|_{\mathrm{F}}^2 + \underbrace{\lambda Q(\boldsymbol{X})}_{\text{promote incoherence}}$$

where λ: regularization parameter

Theorem 13 (Ge et al, 2016)

If sample size $n^2 p \gtrsim \mu^4 n r^6 \log n$ and if α and λ are chosen properly, then with high prob.,

- *all local min are global: any local minimum $\boldsymbol{X}$ of $f_{\mathsf{reg}}(\cdot)$ satisfies $\boldsymbol{X}\boldsymbol{X}^{\top} = \boldsymbol{M}$*
- *saddle points that are not local minima are strict saddles*

Initialization-free theory

Implications:

- Under benign landscape, local search algorithms that can find local minima are often sufficient, *regardless of initialization*
- Key algorithm issue: how to escape saddle points

Saddle-point escaping algorithms

- *Vanilla GD with random initialization*: converges to global minimizers almost surely, but no rates are known (Lee et al. '16)
- *Second-order algorithms (Hessian-based)*: trust-region methods, ... (Sun et al. '16)
- *First-order algorithms*: (perturbed) gradient descent, stochastic gradient descent, ... (Jin et al. '17)

Open problem: does MC converge fast with random initialization?

Gradient descent for matrix completion

Let $\boldsymbol{M} = \boldsymbol{X}^{\natural}\boldsymbol{X}^{\natural\top}$. Observe

$$Y_{i,j} = M_{i,j} + E_{i,j}, \quad (i,j) \in \Omega$$

where $\mathbb{P}\left((i,j) \in \Omega\right) = p$ and $E_{i,j} \sim \mathcal{N}(0, \sigma^2)$[1].

$$\text{minimize} \left\| \mathcal{P}_{\Omega}\left(\widehat{\boldsymbol{M}} - \boldsymbol{Y}\right) \right\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \text{rank}(\widehat{\boldsymbol{M}}) \le r$$

[1] can be relaxed to sub-Gaussian noise and asymmetric case.

Gradient descent for matrix completion

Let $\boldsymbol{M} = \boldsymbol{X}^{\natural}\boldsymbol{X}^{\natural\top}$. Observe

$$Y_{i,j} = M_{i,j} + E_{i,j}, \quad (i,j) \in \Omega$$

where $\mathbb{P}\left((i,j) \in \Omega\right) = p$ and $E_{i,j} \sim \mathcal{N}(0, \sigma^2)$[1].

$$\text{minimize} \left\| \mathcal{P}_{\Omega}\left(\widehat{\boldsymbol{M}} - \boldsymbol{Y}\right) \right\|_{\mathrm{F}}^{2} \quad \text{s.t.} \quad \text{rank}(\widehat{\boldsymbol{M}}) \leq r$$

$$\text{minimize}_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \quad \underbrace{f(\boldsymbol{X}) = \sum_{(j,k) \in \Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - Y_{j,k} \right)^2}_{\text{unregularized least-squares loss}}$$

[1] can be relaxed to sub-Gaussian noise and asymmetric case.

Gradient descent for matrix completion

(1) **Spectral initialization**: let $\boldsymbol{U}^0\boldsymbol{\Sigma}^0\boldsymbol{U}^{0\top}$ be rank-r eigendecomposition of

$$\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{Y}).$$

and set $\boldsymbol{X}^0 = \boldsymbol{U}^0\left(\boldsymbol{\Sigma}^0\right)^{1/2}$

(2) **Gradient descent updates:**

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla f\big(\boldsymbol{X}^t\big), \qquad t = 0, 1, \cdots$$

Gradient descent for matrix completion

Define optimal transform from the tth iterate $\boldsymbol{X}^t$ to $\boldsymbol{X}^\natural$ as

$$\boldsymbol{Q}^t := \mathrm{argmin}_{\boldsymbol{R} \in \mathcal{O}^{r\times r}} \left\| \boldsymbol{X}^t \boldsymbol{R} - \boldsymbol{X}^\natural \right\|_{\mathrm{F}}$$

Theorem 14 (Noiseless MC, Ma, Wang, Chi, Chen '17)

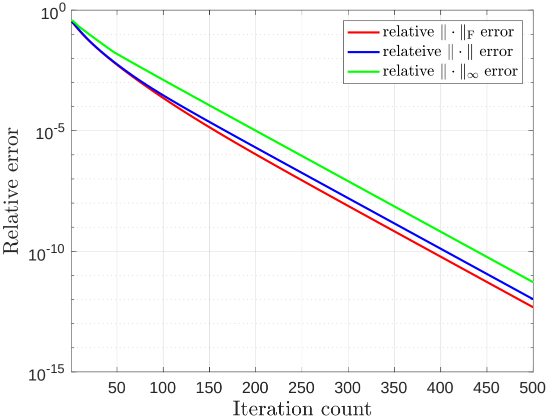
Suppose $\boldsymbol{M} = \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}$ is rank-r, incoherent and well-conditioned. Vanilla GD (with spectral initialization) achieves

- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}} \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|\boldsymbol{X}^\natural\|_{\mathrm{F}}$,
- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\| \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|\boldsymbol{X}^\natural\|$, *(spectral)*
- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\|_{2,\infty} \lesssim \rho^t \mu r \sqrt{\frac{\log n}{np}} \|\boldsymbol{X}^\natural\|_{2,\infty}$, *(incoherence)*

where $0 < \rho < 1$, if step size $\eta \asymp 1/\sigma_{max}$ and sample complexity $n^2 p \gtrsim \mu^3 n r^3 \log^3 n$

- vanilla gradient descent converges linearly for matrix completion!

Numerical evidence for noiseless data



Relative error of $\boldsymbol{X}^t\boldsymbol{X}^{t\top}$ (measured by $\|\cdot\|_{\mathrm{F}}, \|\cdot\|, \|\cdot\|_{\infty}$) vs. iteration count for MC, where $n = 1000$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$

Noisy matrix completion

Theorem 15 (Noisy MC, Ma, Wang, Chi, Chen '17)

Under sample complexity of Theorem 14, if noise satisfies $\sigma\sqrt{\frac{n}{p}} \ll \frac{\sigma_{\min}}{\sqrt{\kappa^3 \mu r \log^3 n}}$, *then GD iterates satisfy*

$$\left\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right\|_{\mathrm{F}} \lesssim \left(\rho^t \mu r \frac{1}{\sqrt{np}} + \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\natural\right\|_{\mathrm{F}},$$

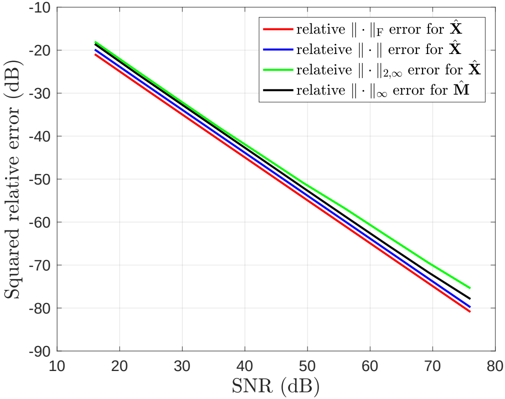
$$\left\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right\|_{2,\infty} \lesssim \left(\rho^t \mu r \sqrt{\frac{\log n}{np}} + \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n\log n}{p}}\right)\left\|\boldsymbol{X}^\natural\right\|_{2,\infty},$$

$$\left\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right\| \lesssim \left(\rho^t \mu r \frac{1}{\sqrt{np}} + \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n}{p}}\right)\left\|\boldsymbol{X}^\natural\right\|$$

- *minimax entrywise error control in* $\left\|\boldsymbol{X}^t\boldsymbol{X}^{t\top} - \boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}\right\|_\infty$

Numerical evidence for noisy data

Set SNR $:= \frac{\|\boldsymbol{M}\|_{\mathrm{F}}^2}{n^2\sigma^2}$



Squared relative error of the estimate $\widehat{\boldsymbol{X}}$ (measured by $\|\cdot\|_{\mathrm{F}}, \|\cdot\|, \|\cdot\|_{2,\infty}$) and $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{X}}\widehat{\boldsymbol{X}}^\top$ (measured by $\|\cdot\|_{\infty}$) vs. SNR, where $n = 500$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$

Related theory

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - Y_{j,k}\right)^2$$

Related theory promotes incoherence explicitly:

Related theory

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - Y_{j,k} \right)^2$$

Related theory promotes incoherence explicitly:

- regularized loss (solve $\min_{\boldsymbol{X}} \ f(\boldsymbol{X}) + Q(\boldsymbol{X})$ instead)
  - e.g. Keshavan, Montanari, Oh '10, Sun, Luo '14, Ge, Lee, Ma '16

Related theory

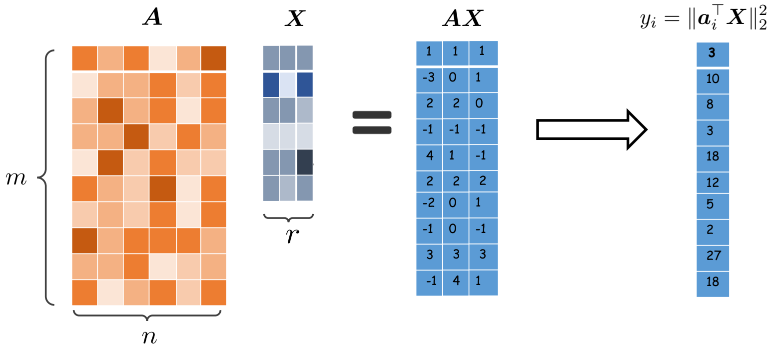
$$\text{minimize}_{\boldsymbol{X} \in \mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - Y_{j,k} \right)^2$$

Related theory promotes incoherence explicitly:

- regularized loss (solve $\min_{\boldsymbol{X}} \;\; f(\boldsymbol{X}) + Q(\boldsymbol{X})$ instead)
  - e.g. Keshavan, Montanari, Oh '10, Sun, Luo '14, Ge, Lee, Ma '16
- projection onto set of incoherent matrices
  - e.g. Chen, Wainwright '15, Zheng, Lafferty '16

$$\boldsymbol{X}^{t+1} = \mathcal{P}_{\mathcal{C}} \left(\boldsymbol{X}^t - \eta_t \nabla f \left(\boldsymbol{X}^t \right) \right), \qquad t = 0, 1, \cdots$$

Quadratic sampling



Recover $X^\natural \in \mathbb{R}^{n\times r}$ from m random quadratic measurements

$$y_i = \|a_i^\top X^\natural\|_2^2, \qquad i = 1, \ldots, m$$

Applications: quantum state tomography, covariance sketching, ...

Gradient descent with spectral initialization

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \frac{1}{4m}\sum_{k=1}^{m}\left(\left\|\boldsymbol{a}_k^\top \boldsymbol{X}\right\|_2^2 - y_k\right)^2$$

Gradient descent with spectral initialization

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \frac{1}{4m}\sum_{k=1}^{m}\left(\|\boldsymbol{a}_k^\top \boldsymbol{X}\|_2^2 - y_k\right)^2$$

Theorem 16 (Quadratic sampling)

Under i.i.d. Gaussian designs $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, *GD (with spectral initialization) achieves*

- $\max_l \left\|\boldsymbol{a}_l^\top(\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural)\right\|_2 \lesssim \sqrt{\log n}\,\frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_\mathrm{F}}$ *(incoherence)*
- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\|_\mathrm{F} \lesssim \left(1 - \frac{\sigma_r^2(\boldsymbol{X}^\natural)\eta}{2}\right)^t \|\boldsymbol{X}^\natural\|_\mathrm{F}$ *(linear convergence)*

provided that $\eta \asymp \frac{1}{(\log n \vee r)^2\sigma_r^2(\boldsymbol{X}^\natural)}$ *and* $m \gtrsim nr^4\log n$

Demixing sparse and low-rank matrices

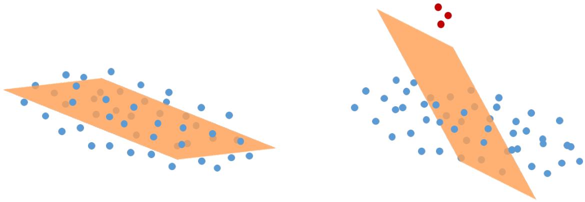
Suppose we are given a matrix

$$\boldsymbol{M} = \underbrace{\boldsymbol{L}}_{\text{low-rank}} + \underbrace{\boldsymbol{S}}_{\text{sparse}} \in \mathbb{R}^{n \times n}$$

Question: can we hope to recover both $\boldsymbol{L}$ and $\boldsymbol{S}$ from $\boldsymbol{M}$?

Applications

- Robust PCA



- Video surveillance: separation of background and foreground

Nonconvex approach

- $\text{rank}(\boldsymbol{L}) \le r$; if we write the SVD of $\boldsymbol{L} = \boldsymbol{U\Sigma V}^\top$, set

$$\boldsymbol{X}^\star = \boldsymbol{U}_L \boldsymbol{\Sigma}^{1/2}; \quad \boldsymbol{Y}^\star = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$$

- non-zero entries of $\boldsymbol{S}$ are "spread out" (no more than s fraction of non-zeros per row/column), but otherwise arbitrary

$$\mathcal{S}_s = \left\{ \boldsymbol{S} \in \mathbb{R}^{n\times n} : \quad \|\boldsymbol{S}_{i,:}\|_0 \le s \cdot n; \ \|\boldsymbol{S}_{:,j}\|_0 \le s \cdot n \right\}$$

$$\underset{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}\in\mathcal{S}_s}{\text{minimize}} \ F(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}) := \underbrace{\|\boldsymbol{M} - \boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{S}\|_{\mathrm{F}}^2}_{\text{least-squares loss}} + \underbrace{\frac{1}{4}\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\|_{\mathrm{F}}^2}_{\text{fix scaling ambiguity}}$$

where $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}$.

Gradient descent and hard thresholding

$$\text{minimize}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}\in\mathcal{S}_s} \quad F(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S})$$

- **Spectral initialization**: Set $\boldsymbol{S}^0 = \mathcal{H}_{\gamma s}(\boldsymbol{M})$. Let $\boldsymbol{U}^0\boldsymbol{\Sigma}^0\boldsymbol{V}^{0\top}$ be rank-r SVD of $\boldsymbol{M}^0 := \mathcal{P}_\Omega(\boldsymbol{M}-\boldsymbol{S}^0)$; set $\boldsymbol{X}^0 = \boldsymbol{U}^0\left(\boldsymbol{\Sigma}^0\right)^{1/2}$ and $\boldsymbol{Y}^0 = \boldsymbol{V}^0\left(\boldsymbol{\Sigma}^0\right)^{1/2}$

Gradient descent and hard thresholding

$$\text{minimize}_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S}\in\mathcal{S}_s} \quad F(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{S})$$

- **Spectral initialization**: Set $\boldsymbol{S}^0 = \mathcal{H}_{\gamma s}(\boldsymbol{M})$. Let $\boldsymbol{U}^0\boldsymbol{\Sigma}^0\boldsymbol{V}^{0\top}$ be rank-r SVD of $\boldsymbol{M}^0 := \mathcal{P}_\Omega(\boldsymbol{M}-\boldsymbol{S}^0)$; set $\boldsymbol{X}^0 = \boldsymbol{U}^0\left(\boldsymbol{\Sigma}^0\right)^{1/2}$ and $\boldsymbol{Y}^0 = \boldsymbol{V}^0\left(\boldsymbol{\Sigma}^0\right)^{1/2}$
- **for** $t = 0, 1, 2, \cdots$
  - **Hard thresholding**: $\boldsymbol{S}^{t+1} = \mathcal{H}_{\gamma s}(\boldsymbol{M} - \boldsymbol{X}^t\boldsymbol{Y}^{t\top})$
  - **Gradient updates**:

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta\nabla_{\boldsymbol{X}} F\left(\boldsymbol{X}^t, \boldsymbol{Y}^t, \boldsymbol{S}^{t+1}\right)$$
$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta\nabla_{\boldsymbol{Y}} F\left(\boldsymbol{X}^t, \boldsymbol{Y}^t, \boldsymbol{S}^{t+1}\right)$$

Efficient nonconvex recovery

Theorem 17 (Nonconvex RPCA, Yi et al. '16)

Set $\gamma = 2$ *and* $\eta = 1/(36\sigma_{\max})$. *Suppose that*

$$s \lesssim \min\left\{\frac{1}{\mu\sqrt{\kappa r^3}}, \frac{1}{\mu\kappa^2 r}\right\}$$

Then GD+HT satisfies

$$\|\boldsymbol{X}^t\boldsymbol{Y}^{t\top} - \boldsymbol{L}\|_{\mathrm{F}}^2 \lesssim \left(1 - \frac{1}{288\kappa}\right)^t \mu^2\kappa r^3 s^2 \sigma_{\max}$$

- $O(\kappa \log 1/\epsilon)$ iterations to reach ϵ-accuracy
- For adversarial outliers, optimal fraction is $s = O(1/\mu r)$; Theorem 17 is suboptimal by a factor of $\sqrt{r}$
- extendable to partial observation models

Tutorial outline

- Part I: Overview
- Part II: Phase retrieval: a case study
  - Spectral initialization
  - Local refinement: algorithm and analysis
- Part III: Low-rank matrix estimation
- Part IV: Closing remarks

A growing list of "benign" nonconvex problems

- blind deconvolution / self-calibration
- dictionary learning
- tensor decomposition
- robust PCA
- mixture linear regression
- Gaussian mixture models
- etc...

Topics we did not cover

- **other algorithms:** alternating minimization, stochastic gradient descent, mirror descent, singular value projection, etc...
- **additional structures:** e.g. sparsity, piece-wise smoothness
- **saddle-point escaping algorithms**

Reference

[1] Dr. Ju Sun's webpage: "`http://sunju.org/research/nonconvex/`".

[2] "*Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation*," Y. Chen, and Y. Chi, *arXiv preprint arXiv:1802.08397*, IEEE Signal Processing Magazine, to appear.

[3] "*Phase retrieval via Wirtinger flow: Theory and algorithms*," E. Candes, X. Li, M. Soltanolkotabi, *IEEE Transactions on Information Theory*, 2015.

[4] "*Solving random quadratic systems of equations is nearly as easy as solving linear systems*," Y. Chen, E. Candes, Communications on Pure and Applied Mathematics, 2017.

[5] "*Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow*," H. Zhang, Y. Chi, and Y. Liang, ICML 2016.

[6] "*Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval, Matrix Completion and Blind Deconvolution*," C. Ma, K. Wang, Y. Chi and Y. Chen, *arXiv preprint arXiv:1711.10467*, 2017.

Reference

[7] "*Gradient Descent with Random Initialization: Fast Global Convergence for Nonconvex Phase Retrieval*," Y. Chen, Y. Chi, J. Fan, C. Ma, *arXiv preprint arXiv:1803.07726*, 2018.

[8] "*Solving systems of random quadratic equations via truncated amplitude flow*," G. Wang, G. Giannakis, and Y. Eldar, *IEEE Transactions on Information Theory*, 2017.

[9] "*Matrix completion from a few entries*," R. Keshavan, A. Montanari, and S. Oh, *IEEE Transactions on Information Theory*, 2010.

[10] "*Guaranteed matrix completion via non-convex factorization*," R. Sun, T. Luo, *IEEE Transactions on Information Theory*, 2016.

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

[12] "*Fast Algorithms for Robust PCA via Gradient Descent*," X. Yi, D. Park, Y. Chen, and C. Caramanis, *NIPS*, 2016.

Reference

[13] "*Matrix completion has no spurious local minimum*," R. Ge, J. Lee, and T. Ma, *NIPS*, 2016.

[14] "*No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis*," R. Ge, C. Jin, and Y. Zheng, *ICML*, 2017.

[15] "*Symmetry, Saddle Points, and Global Optimization Landscape of Nonconvex Matrix Factorization*," X. Li *et al.*, *arXiv preprint arxiv:1612.09296*, 2016.

[16] "*Phase Transitions of Spectral Initialization for High-Dimensional Nonconvex Estimation*," Y. M. Lu and G. Li, *Information and Inference*, to appear, *arXiv:1702.06435*, 2018.

[17] "*Kaczmarz Method for Solving Quadratic Equations*," Y. Chi and Y. M. Lu, *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1183-1187, 2016.

[18] "*Scaling Limit: Exact and Tractable Analysis of Online Learning Algorithms with Applications to Regularized Regression and PCA*," C. Wang, J. Mattingly and Y. M. Lu, *arXiv preprint arXiv:1712.04332*, 2017.

[19] "*A Geometric Analysis of Phase Retrieval*," S. Ju, Q. Qu, and J. Wright, to appear, *Foundations of Computational Mathematics*, 2016.

Reference

[20] "*Gradient descent converges to minimizers*," J. Lee, M. Simchowitz, M. Jordan, B. Recht, *Conference on Learning Theory*, 2016.

[21] "*Fundamental limits of weak recovery with applications to phase retrieval*," M. Mondelli, and A. Montanari, arXiv:1708.05932, 2017.

[22] "*Phase retrieval using alternating minimization*," P. Netrapalli, P. Jain, and S. Sanghavi, *NIPS*, 2013.

[23] "*Optimization-based AMP for Phase Retrieval: The Impact of Initialization and ℓ_2-regularization*," J. Ma, J. Xu, and A. Maleki, arXiv:1801.01170, 2018.

[24] "*How to escape saddle points efficiently*," C. Jin, R. Ge, P. Netrapalli, S. Kakade, M. Jordan, arXiv:1703.00887, 2017.

[25] "*Complete dictionary recovery over the sphere*," J. Sun, Q. Qu, J. Wright, *IEEE Transactions on Information Theory*, 2017.

[26] "*A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*," S. Burer, and R. Monteiro, *Mathematical Programming*, 2003.

Reference

[27] "*Memory-efficient Kernel PCA via Partial Matrix Sampling and Nonconvex Optimization: a Model-free Analysis of Local Minima*," J. Chen, X. Li, arXiv:1711.01742, 2017.

[28] "*Rapid, robust, and reliable blind deconvolution via nonconvex optimization*," X. Li, S. Ling, T. Strohmer, K. Wei, arXiv:1606.04933, 2016.

[29] "*Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming*," E. Candes, T. Strohmer, V. Voroninski, *Communications on Pure and Applied Mathematics*, 2012.

[30] "*Exact matrix completion via convex optimization*," E. Candes, B. Recht, *Foundations of Computational mathematics*, 2009.

[31] "*Low-rank solutions of linear matrix equations via procrustes flow,*" S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, B. Recht, arXiv:1507.03566, 2015.

[32] "*Global optimality of local search for low rank matrix recovery,*" S. Bhojanapalli, B. Neyshabur, and N. Srebro, NIPS, 2016.

Reference

[33] "*Phase retrieval via matrix completion,*" E. Candes, Y. Eldar, T. Strohmer, and V. Voroninski, *SIAM Journal on Imaging Sciences*, 2013.

[34] "*Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow,*" T. Cai, X. Li, Z. Ma, *The Annals of Statistics*, 2016.

[35] "*The landscape of empirical risk for non-convex losses,*" S. Mei, Y. Bai, and A. Montanari, arXiv:1607.06534, 2016.

[36] "*Non-convex robust PCA,*" P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, NIPS, 2014.

[37] "*Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach,*" D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, arXiv:1609.03240, 2016.

[38] "*Solving almost all systems of random quadratic equations*" G. Wang, G. Giannakis, Y. Saad, and J. Chen, arXiv:1705.10407, 2017.

[39] "*A Nonconvex Approach for Phase Retrieval: Reshaped Wirtinger Flow and Incremental Algorithms,*" H. Zhang, Y. Zhou, Y. Liang and Y. Chi, *Journal of Machine Learning Research*, 2017.

Reference

[40] "*Nonconvex Matrix Factorization from Rank-One Measurements,*" Y. Li, C. Ma, Y. Chen and Y. Chi, arXiv:1802.06286, 2018.

[41] "*Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization,*", M. Soltanolkotabi, arXiv:1702.06175, 2017.

Thanks!