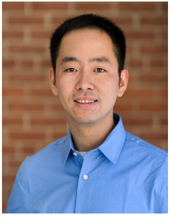
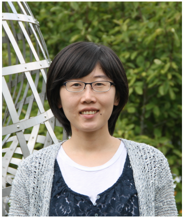
Nonconvex Optimization Meets Low-Rank Matrix Factorization



Yuxin Chen
Princeton



Yuejie Chi
CMU

Acknowledgement

- Our collaborators: Emmanuel J. Candès, Jianqing Fan, Yuanxin Li, Yingbin Liang, Yue M. Lu, Cong Ma, Kaizheng Wang, Huishuai Zhang
- This work is supported in part by ARO W911NF-18-1-0303, AFOSR FA9550-19-1-0030 and FA9550-15-1-0205, ONR N00014-18-1-2142, and NSF ECCS-1818571 and CCF-1806154

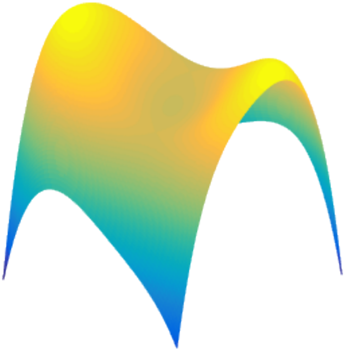
Outline

- Part I: Overview
- Part II: Preliminaries and rank-one matrix factorization
- Part IV: Two-stage approaches
  - Spectral initialization
  - Local refinement: algorithm and analysis
- Part I: Global landscape and initialization-free algorithms
  - Landscape analysis
  - Saddle-point escaping algorithms
  - Random initialization?
- Part VI: Closing remarks

Nonconvex estimation problems are everywhere

Empirical risk minimization is usually nonconvex

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}; \text{data}) \quad \rightarrow \quad \text{loss function may be nonconvex}$$

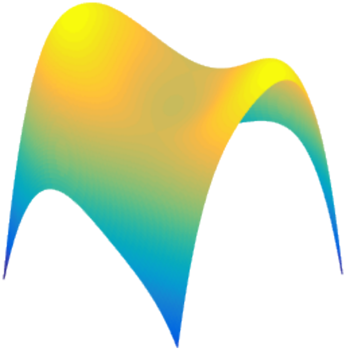


Nonconvex estimation problems are everywhere

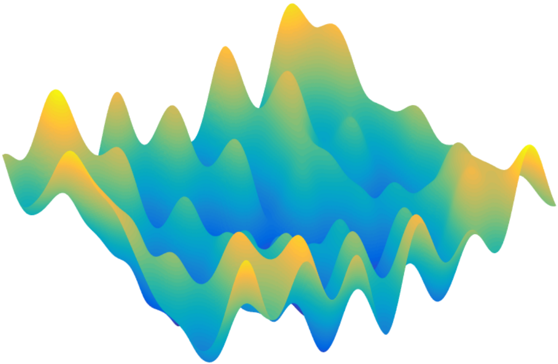
Empirical risk minimization is usually nonconvex

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x};\text{data}) \quad \rightarrow \quad \text{loss function may be nonconvex}$$

- low-rank matrix completion
- blind deconvolution
- dictionary learning
- mixture models
- deep learning
- ...



Nonconvex optimization may be super scary



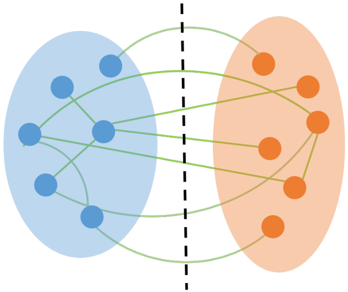
There may be bumps everywhere and exponentially many local optima

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

Example: solving quadratic programs is hard

Finding maximum cut in a graph is about solving a quadratic program

$$
\begin{array}{ll}
\text{maximize}_{\boldsymbol{x}} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\
\text{subj. to} & x_i^2 = 1, \quad i = 1, \cdots, n
\end{array}
$$



Example: solving quadratic programs is hard

"I can't find an efficient algorithm, but neither can all these people."

figure credit: coding horror

\$1,000,000 question

One strategy: convex relaxation

Can relax into convex problems by

- finding convex surrogates (e.g. matrix completion)
- lifting into higher dimensions (e.g. Max-Cut)

Example of convex surrogate: matrix completion

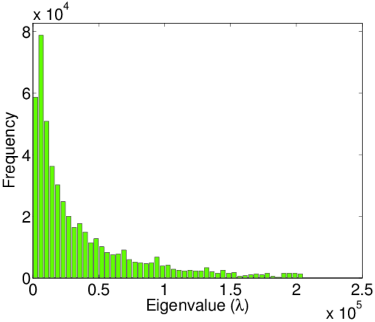
figure credit: Candès et al.

Netflix challenge

Predict unseen ratings

Low-rank modeling

figure credit: E. Candès



A few factors explain most of the data

Low-rank modeling

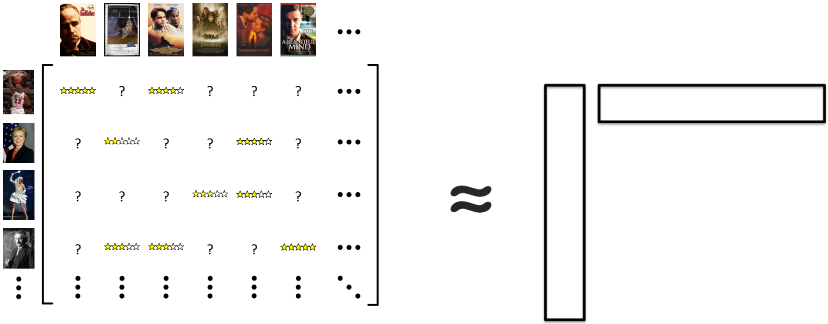


figure credit: E. Candès

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

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

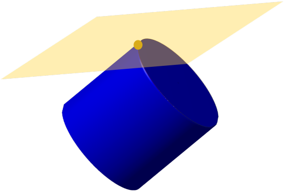
Example of convex surrogate: matrix completion

— Fazel '02, Recht, Parrilo, Fazel '10, Candès, Recht '09

$$\text{minimize}_{\boldsymbol{M}} \ \text{rank}(\boldsymbol{M}) \ \text{subj. to data constraints}$$

$$\Downarrow \ \text{cvx surrogate}$$

$$\text{minimize}_{\boldsymbol{M}} \ \text{nuc-norm}(\boldsymbol{M}) \qquad \text{subj. to data constraints}$$



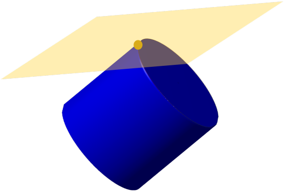
Example of convex surrogate: matrix completion

— Fazel '02, Recht, Parrilo, Fazel '10, Candès, Recht '09

$$\text{minimize}_{\boldsymbol{M}} \;\; \text{rank}(\boldsymbol{M}) \;\; \text{subj. to data constraints}$$

$$\Downarrow \;\; \text{cvx surrogate}$$

$$\text{minimize}_{\boldsymbol{M}} \;\; \text{nuc-norm}(\boldsymbol{M}) \qquad \text{subj. to data constraints}$$



robust variation used by Netflix

— Candès, Li, Ma, Wright '10

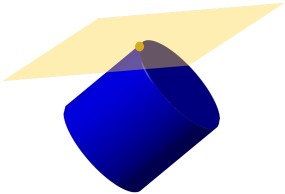
Example of convex surrogate: matrix completion

— Fazel '02, Recht, Parrilo, Fazel '10, Candès, Recht '09

$$\text{minimize}_{\boldsymbol{M}} \;\; \text{rank}(\boldsymbol{M}) \;\; \text{subj. to data constraints}$$

$\Downarrow$ cvx surrogate

$$\text{minimize}_{\boldsymbol{M}} \;\; \text{nuc-norm}(\boldsymbol{M}) \qquad \text{subj. to data constraints}$$

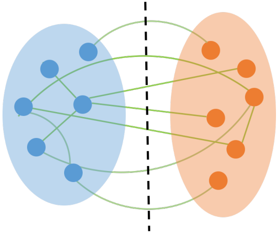


robust variation used by Netflix
— Candès, Li, Ma, Wright '10

Problem: operate in *full* matrix space even though $\boldsymbol{X}$ is low-rank

Example of lifting: Max-Cut

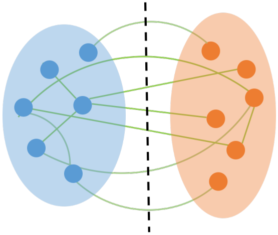
— Goemans, Williamson '95



$$\begin{aligned} \text{maximize}_{\boldsymbol{x}} \quad & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subj. to} \quad & x_i^2 = 1, \quad i = 1, \cdots, n \end{aligned}$$

Example of lifting: Max-Cut

— Goemans, Williamson '95



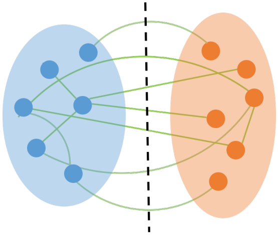
$$\begin{array}{ll} \text{maximize}_{\boldsymbol{x}} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subj. to} & x_i^2 = 1, \quad i = 1, \cdots, n \end{array}$$

$\Downarrow$ let $\boldsymbol{X}$ be $\boldsymbol{x}\boldsymbol{x}^\top$

$$\begin{array}{ll} \text{maximize}_{\boldsymbol{X}} & \langle \boldsymbol{X}, \boldsymbol{W} \rangle \\ \text{subj. to} & \boldsymbol{X}_{i,i} = 1, \quad i = 1, \cdots, n \\ & \boldsymbol{X} \succeq \boldsymbol{0} \\ & \text{rank}(\boldsymbol{X}) = 1 \end{array}$$

Example of lifting: Max-Cut

— Goemans, Williamson '95



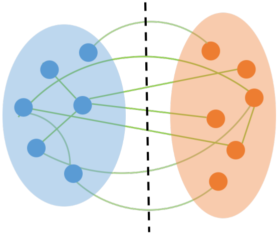
$$\begin{array}{ll} \text{maximize}_{\boldsymbol{x}} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subj. to} & x_i^2 = 1, \quad i = 1, \cdots, n \end{array}$$

$$\Downarrow \quad \text{let } \boldsymbol{X} \text{ be } \boldsymbol{x}\boldsymbol{x}^\top$$

$$\begin{array}{ll} \text{maximize}_{\boldsymbol{X}} & \langle \boldsymbol{X}, \boldsymbol{W} \rangle \\ \text{subj. to} & \boldsymbol{X}_{i,i} = 1, \quad i = 1, \cdots, n \\ & \boldsymbol{X} \succeq \boldsymbol{0} \\ & \cancel{\text{rank}(\boldsymbol{X}) = 1} \end{array}$$

Example of lifting: Max-Cut

— Goemans, Williamson '95



$$\begin{aligned} \text{maximize}_{\boldsymbol{x}} \quad & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} \\ \text{subj. to} \quad & x_i^2 = 1, \quad i = 1, \cdots, n \end{aligned}$$

$\Downarrow$ let $\boldsymbol{X}$ be $\boldsymbol{x}\boldsymbol{x}^\top$

$$\begin{aligned} \text{maximize}_{\boldsymbol{X}} \quad & \langle \boldsymbol{X}, \boldsymbol{W} \rangle \\ \text{subj. to} \quad & \boldsymbol{X}_{i,i} = 1, \quad i = 1, \cdots, n \\ & \boldsymbol{X} \succeq \boldsymbol{0} \\ & \cancel{\text{rank}(\boldsymbol{X}) = 1} \end{aligned}$$

Problem: explosion in dimensions ($\mathbb{R}^n \to \mathbb{R}^{n \times n}$)

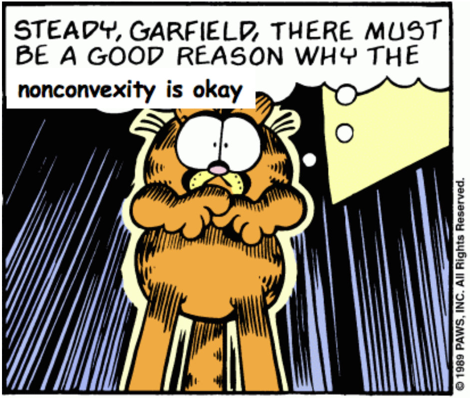
How about optimizing nonconvex problems directly without lifting?

Nonconvex optimization

Complicated nonconvex problems are solved on a daily basis via simple algorithms such as stochastic gradient descent

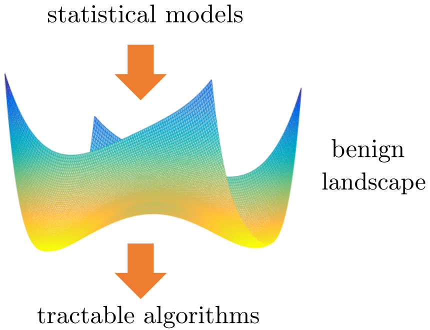
Nonconvex optimization

Complicated nonconvex problems are solved on a daily basis via simple algorithms such as stochastic gradient descent



- How come simple nonconvex algorithms work so well in practice?

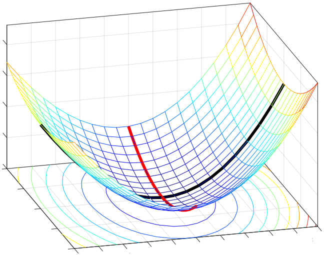
Statistical models come to rescue



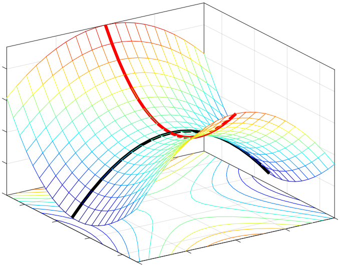
When data are generated by certain statistical models, problems are often much nicer than worst-case instances

Sometimes they are much nicer than we think

Under certain statistical models,
we see benign global geometry: no spurious local optima



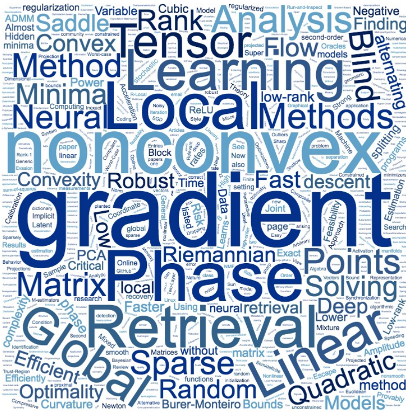
global minimum



saddle point

Even the simplest possible nonconvex methods
might be remarkably efficient under suitable statistical models

Nonconvex optimization with guarantees



Phase retrieval: Gerchberg-Saxton '72, Netrapalli et al. '13, Candès, Li, Soltanolkotabi '14, Chen, Candès '15, Cai, Li, Ma '15, Zhang et al. 16, Wang et al. '16, Sun et al. '16, Ma et al. '17, Chen et al. '18, ...

Matrix completion: Keshavan et al. '09, Jain et al. '09, Hardt '13, Sun, Luo '15, Chen, Wainwright '15, Zheng, Lafferty '16, Ge et al. '16, Jin et al. '16, Ma et al. '17, ...

Matrix sensing: Jain et al. '13, Tu et al. '15, Zheng, Lafferty '15, Bhojanapalli et al. 16, Li, Zhu, Tang '18, ...

Blind deconvolution / demixing: Li et al. '16, Lee et al. '16, Ling, Strohmer '16, Huang, Hand '16, Ma et al. '17, Zhang et al. '18, Li, Bresler '18, Dong, Shi '18, ...

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

Robust principal component analysis: Netrapalli et al. '14, Yi et al. '16, Gu et al. '16, Ge et al. '17, Cherapanamjeri et al. '17, ...

- *http://sunju.org/research/nonconvex/*
- *"Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview," Chi, Lu, Chen '18*

Outline

- Part I: Overview
- Part II: Preliminaries and rank-one matrix factorization
- Part IV: Two-stage approaches
  - Spectral initialization
  - Local refinement: algorithm and analysis
- Part I: Global landscape and initialization-free algorithms
  - Landscape analysis
  - Saddle-point escaping algorithms
  - Random initialization?
- Part VI: Closing remarks

Unconstrained optimization

Consider an unconstrained optimization problem

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

Definition 1 (first-order critical points)

A first-order critical point of f satisfies

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

Unconstrained optimization

Consider an unconstrained optimization problem

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

Definition 2 (second-order critical points)

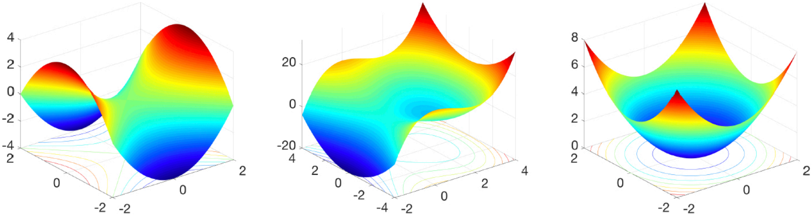
A second-order critical point $\boldsymbol{x}$ satisfies

$$\nabla f(\boldsymbol{x}) = \mathbf{0} \quad \text{and} \quad \nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$$

Several types of critical points

For any first-order critical point $\boldsymbol{x}$:

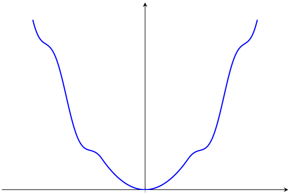
- $\nabla^2 f(\boldsymbol{x}) \prec \mathbf{0}$ $\quad\rightarrow\quad$ local maximum
- $\nabla^2 f(\boldsymbol{x}) \succ \mathbf{0}$ $\quad\rightarrow\quad$ local minimum
- $\lambda_{\min}(\nabla^2 f(\boldsymbol{x})) < 0$ $\quad\rightarrow\quad$ *strict* saddle point



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

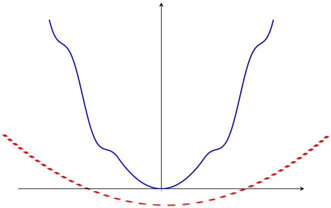
figure credit: Li et al. '16

Gradient descent theory



Two standard conditions that enable geometric convergence of GD

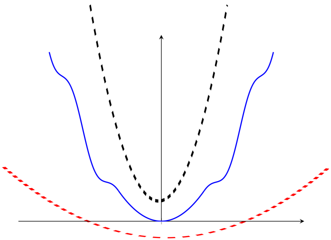
Gradient descent theory



Two standard conditions that enable geometric convergence of GD

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

Gradient descent theory



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 ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$) with $\eta = 1/\beta$ obeys

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

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 ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$) with $\eta = 1/\beta$ obeys

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

- Condition number β/α determines rate of convergence

Gradient descent theory revisited

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

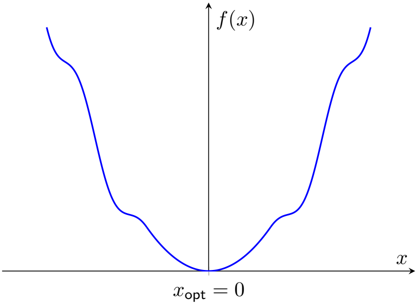
$$\mathbf{0} \preceq \alpha \boldsymbol{I} \preceq \nabla^2 f(\boldsymbol{x}) \preceq \beta \boldsymbol{I}, \qquad \forall \boldsymbol{x}$$

ℓ_2 **error contraction:** GD ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$) with $\eta = 1/\beta$ obeys

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

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

Regularity Condition (RC)



Definition 3 (Regularity Condition (RC))

$\boldsymbol{g}(\cdot)$ is said to obey RC(μ, λ, ζ) for some $\mu, \lambda, \zeta > 0$ if

$$2\langle\, \boldsymbol{g}(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_{\text{opt}} \rangle \geq \mu \|\boldsymbol{g}(\boldsymbol{x})\|_2^2 + \lambda \, \|\boldsymbol{x} - \boldsymbol{x}_{\text{opt}}\|_2^2 \quad \forall \boldsymbol{x}$$

Convergence under RC

ℓ_2 **error contraction:** The update rule ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \boldsymbol{g}(\boldsymbol{x}^t)$) with $\eta = \mu$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq (1 - \mu\lambda)\, \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2$$

- $\boldsymbol{g}(\cdot)$: more general search directions
  - example: in vanilla GD, $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$

Convergence under RC

ℓ_2 **error contraction:** The update rule ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \boldsymbol{g}(\boldsymbol{x}^t)$) with $\eta = \mu$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \le (1 - \mu\lambda) \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2$$

- $\boldsymbol{g}(\cdot)$: more general search directions
  - example: in vanilla GD, $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$
- The product $\mu\lambda$ determines the rate of convergence

Convergence under RC

ℓ_2 **error contraction:** The update rule ($\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \boldsymbol{g}(\boldsymbol{x}^t)$) with $\eta = \mu$ obeys

$$\|\boldsymbol{x}^{t+1} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq (1 - \mu\lambda) \|\boldsymbol{x}^t - \boldsymbol{x}_{\mathsf{opt}}\|_2$$

- $\boldsymbol{g}(\cdot)$: more general search directions
  - example: in vanilla GD, $\boldsymbol{g}(\boldsymbol{x}) = \nabla f(\boldsymbol{x})$
- The product $\mu\lambda$ determines the rate of convergence
- Attains ε-accuracy within $O(\frac{1}{\mu\lambda} \log \frac{1}{\varepsilon})$ iterations

RC = one-point strong convexity + smoothness

- One-point α-strong convexity:

$$f(\boldsymbol{x}_{\mathsf{opt}}) - f(\boldsymbol{x}) \geq \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}_{\mathsf{opt}} - \boldsymbol{x} \rangle + \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 \qquad (1)$$

- β-moothness:

$$\begin{aligned} f(\boldsymbol{x}_{\mathsf{opt}}) - f(\boldsymbol{x}) &\leq f\Big(\boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})\Big) - f(\boldsymbol{x}) \\ &\leq \Big\langle \nabla f(\boldsymbol{x}), -\frac{1}{\beta}\nabla f(\boldsymbol{x}) \Big\rangle + \frac{\beta}{2}\Big\| \frac{1}{\beta}\nabla f(\boldsymbol{x}) \Big\|_2^2 \\ &= -\frac{1}{2\beta} \|\nabla f(\boldsymbol{x})\|_2^2 \qquad (2) \end{aligned}$$

RC = one-point strong convexity + smoothness

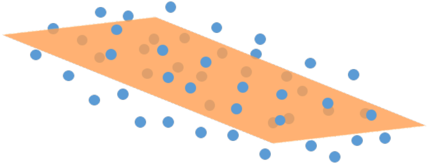
Combining (1) and (2) yields

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}} \rangle \geq \frac{\alpha}{2} \|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2^2 + \frac{1}{2\beta} \|\nabla f(\boldsymbol{x})\|_2^2 \qquad (3)$$

— *RC holds with $\mu = 1/\beta$ and $\lambda = \alpha$*

A toy example: rank-1 matrix factorization

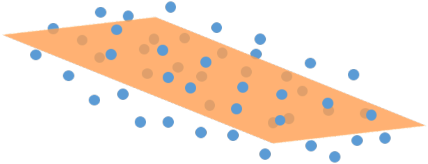
Revisiting PCA



Given $\boldsymbol{M} \succeq \mathbf{0} \in \mathbb{R}^{n \times n}$ (not necessarily low-rank), find its best rank-r approximation:

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

Revisiting PCA



This problem admits a closed-form solution

- let $\boldsymbol{M} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$ be eigen-decomposition of $\boldsymbol{M}$ $(\lambda_1 \geq \cdots \geq \lambda_n)$, then

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

— *nonconvex, but tractable*

Optimization viewpoint

If we factorize $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{X}^\top$ with $\boldsymbol{X} \in \mathbb{R}^{n\times r}$, then it leads to a nonconvex problem:

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

To simplify exposition, set $r = 1$:

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

Questions

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

- Where / what are the critical points?
- What does the curvature behave like, at least locally around the global minimizer?

Critical points of $f(\cdot)$

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

$$\Updownarrow$$

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

$$\Updownarrow$$

$$\boldsymbol{x} \text{ aligns with an eigenvector of } \boldsymbol{M} \quad \text{or} \quad \boldsymbol{x} = \boldsymbol{0}$$

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

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

Categorization of critical points

The critical points can be further categorized based on the **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 point $\boldsymbol{x}_k = \pm\sqrt{\lambda_k}\boldsymbol{u}_k$:

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

Categorization of critical points

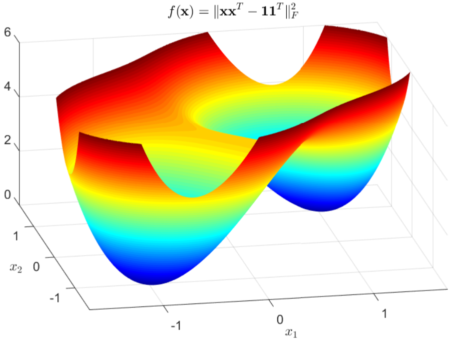
The critical points can be further categorized based on the **Hessians**:

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

- If $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n \geq 0$, then
  - $\nabla^2 f(\boldsymbol{x}_1) \succ \mathbf{0}$ $\rightarrow$ local minima
  - $1 < k \leq n$: $\lambda_{\min}(\nabla^2 f(\boldsymbol{x}_k)) < 0$, $\lambda_{\max}(\nabla^2 f(\boldsymbol{x}_k)) > 0$ $\rightarrow$ strict saddle
  - $\boldsymbol{x} = \mathbf{0}$: $\nabla^2 f(\mathbf{0}) \preceq \mathbf{0}$ $\rightarrow$ local maxima

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 saddles: $\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\pm \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

— *No "spurious" local minima!*

Local strong convexity and local linear convergence

- The global minimizers: $\boldsymbol{x}_{\mathsf{opt}} = \pm\sqrt{\lambda_1}\boldsymbol{u}_1$
- For all $\boldsymbol{x}$ obeying $\underbrace{\|\boldsymbol{x} - \boldsymbol{x}_{\mathsf{opt}}\|_2 \leq \frac{\lambda_1 - \lambda_2}{15\sqrt{\lambda_1}}}_{\text{basin of attraction}}$, one has

$$0.25(\lambda_1 - \lambda_2)\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq 4.5\lambda_1 \boldsymbol{I}_n$$

Local strong convexity and local linear convergence

- The global minimizers: $\boldsymbol{x}_{\text{opt}} = \pm\sqrt{\lambda_1}\boldsymbol{u}_1$
- For all $\boldsymbol{x}$ obeying $\underbrace{\|\boldsymbol{x} - \boldsymbol{x}_{\text{opt}}\|_2 \le \frac{\lambda_1 - \lambda_2}{15\sqrt{\lambda_1}}}_{\text{basin of attraction}}$, one has

$$0.25(\lambda_1 - \lambda_2)\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq 4.5\lambda_1 \boldsymbol{I}_n$$

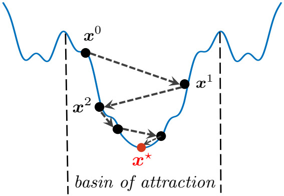
ℓ_2 **error contraction:** The GD iterates obey

$$\left\|\boldsymbol{x}^t - \sqrt{\lambda_1}\boldsymbol{u}_1\right\|_2 \le \left(1 - \frac{\lambda_1 - \lambda_2}{18\lambda_1}\right)^t \left\|\boldsymbol{x}^0 - \sqrt{\lambda_1}\boldsymbol{u}_1\right\|_2, \; t \ge 0,$$

as long as $\|\boldsymbol{x}^0 - \sqrt{\lambda_1}\boldsymbol{u}_1\|_2 \le \frac{\lambda_1 - \lambda_2}{15\sqrt{\lambda_1}}$

Outlook: two vignettes

Two-stage approach:



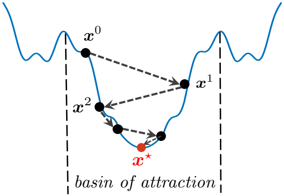
smart initialization

$+$

local refinement

Outlook: two vignettes

Two-stage approach:

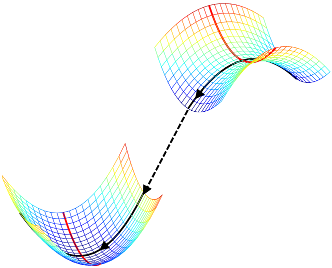


smart initialization

$+$

local refinement

Global landscape:



benign landscape

$+$

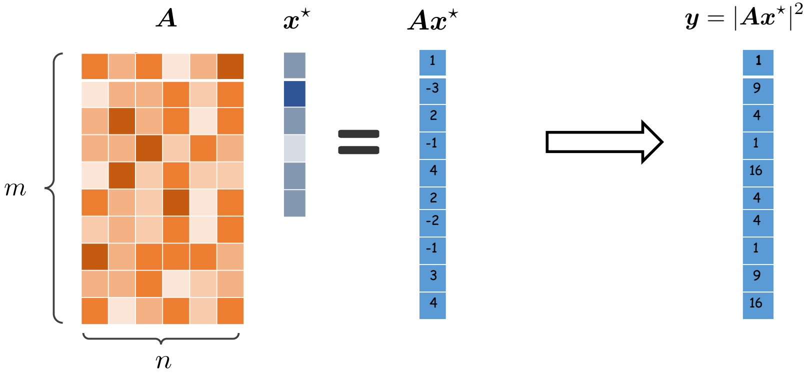
saddle-point escaping

Outline

- Part I: Overview
- Part II: Preliminaries and rank-one matrix factorization
- Part IV: Two-stage approaches
  - Spectral initialization
  - Local refinement: algorithm and analysis
- Part I: Global landscape and initialization-free algorithms
  - Landscape analysis
  - Saddle-point escaping algorithms
  - Random initialization?
- Part VI: Closing remarks

A case study: solving quadratic systems of equations

Solving quadratic systems of equations



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

$$y_k \;=\; (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2, \qquad k = 1, \ldots, m$$

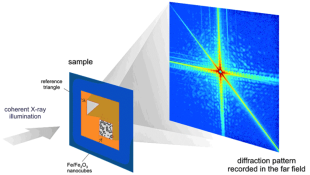
assume w.l.o.g. $\|\boldsymbol{x}^\star\|_2 = 1$

Motivation: phase retrieval

Detectors record intensities of diffracted rays

- electric field $x(t_1, t_2) \longrightarrow$ Fourier transform $\widehat{x}(f_1, f_2)$

figure credit: Stanford SLAC



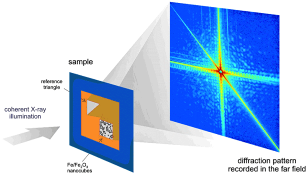
intensity of electrical field: $\left|\widehat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

Motivation: phase retrieval

Detectors record intensities of diffracted rays

- electric field $x(t_1, t_2) \longrightarrow$ Fourier transform $\widehat{x}(f_1, f_2)$

figure credit: Stanford SLAC



intensity of electrical field: $\left|\widehat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

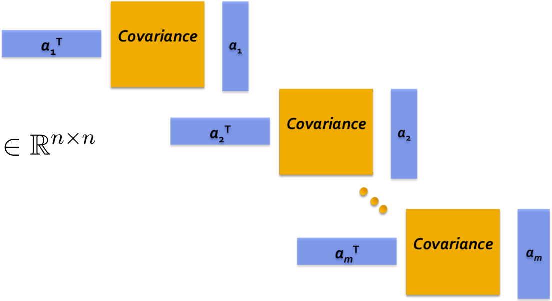
Phase retrieval: recover signal $x(t_1, t_2)$ from intensity $\left|\widehat{x}(f_1, f_2)\right|^2$

Motivation: covariance estimation from quadratic sketches

— *Chen, Chi, Goldsmith '13, Cai, Zhang '13*

- **Data**: m quadratic measurements about *low-rank* covariance matrix $\boldsymbol{\Sigma}$

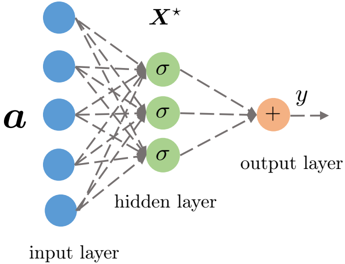
$$y_i = \boldsymbol{a}_i^\top \boldsymbol{\Sigma} \boldsymbol{a}_i + \text{noise}, \qquad i = 1, \cdots, m$$



- **Goal:** recover $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$

Motivation: learning neural nets with quadratic activation

— *Soltanolkotabi, Javanmard, Lee '17, Li, Ma, Zhang '17*

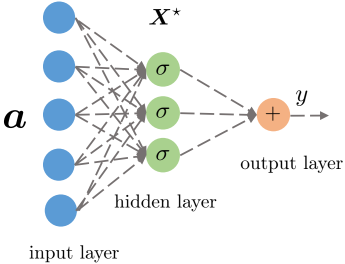


input features: $\boldsymbol{a}$; weights: $\boldsymbol{X}^\star = [\boldsymbol{x}_1^\star, \cdots, \boldsymbol{x}_r^\star]$

$$\text{output:} \quad y = \sum_{i=1}^{r} \sigma(\boldsymbol{a}^\top \boldsymbol{x}_i^\star)$$

Motivation: learning neural nets with quadratic activation

— *Soltanolkotabi, Javanmard, Lee '17, Li, Ma, Zhang '17*

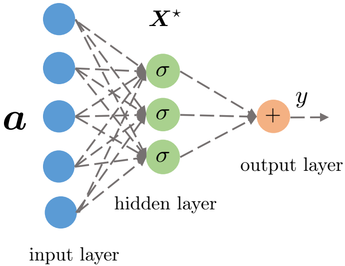


input features: $\boldsymbol{a}$; weights: $\boldsymbol{X}^\star = [\boldsymbol{x}_1^\star, \cdots, \boldsymbol{x}_r^\star]$

$$\text{output:} \quad y = \sum_{i=1}^{r} \sigma(\boldsymbol{a}^\top \boldsymbol{x}_i^\star) \overset{\sigma(z)=z^2}{:=} \sum_{i=1}^{r} (\boldsymbol{a}^\top \boldsymbol{x}_i^\star)^2$$

Motivation: learning neural nets with quadratic activation

— *Soltanolkotabi, Javanmard, Lee '17, Li, Ma, Zhang '17*



input features: $\boldsymbol{a}$; weights: $\boldsymbol{X}^\star = [\boldsymbol{x}_1^\star, \cdots, \boldsymbol{x}_r^\star]$

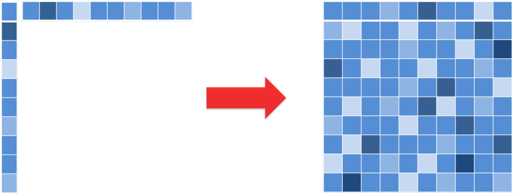
$$\text{output:} \quad y = \sum_{i=1}^{r} \sigma(\boldsymbol{a}^\top \boldsymbol{x}_i^\star) \overset{\sigma(z)=z^2}{:=} \sum_{i=1}^{r} (\boldsymbol{a}^\top \boldsymbol{x}_i^\star)^2$$

We consider simplest model when $r = 1$ (higher r is similar)

An equivalent view: low-rank factorization

Introduce $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ to linearize constraints

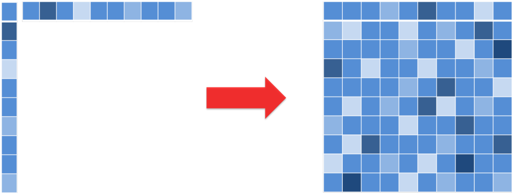
$$y_k = (\boldsymbol{a}_k^\top \boldsymbol{x})^2 = \boldsymbol{a}_k^\top (\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{a} \qquad \Longrightarrow \qquad y_k = \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k$$



An equivalent view: low-rank factorization

Introduce $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ to linearize constraints

$$y_k \;= (\boldsymbol{a}_k^\top \boldsymbol{x})^2 = \boldsymbol{a}_k^\top (\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{a} \qquad \Longrightarrow \qquad y_k = \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k$$

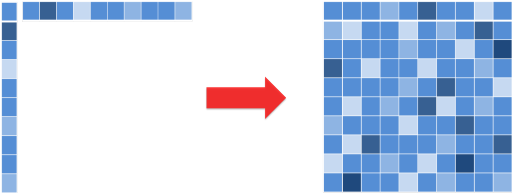


$$\begin{array}{lll} \text{find} & \boldsymbol{X} & \\ \text{s.t.} & y_k \;=\; \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k, & k = 1, \cdots, m \\ & \mathrm{rank}(\boldsymbol{X}) = 1 & \end{array}$$

An equivalent view: low-rank factorization

Introduce $\boldsymbol{X} = \boldsymbol{x}\boldsymbol{x}^\top$ to linearize constraints

$$y_k \; = (\boldsymbol{a}_k^\top \boldsymbol{x})^2 = \boldsymbol{a}_k^\top (\boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{a} \qquad \Longrightarrow \qquad y_k = \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k$$



$$\begin{array}{ll} \text{find} & \boldsymbol{X} \\ \text{s.t.} & y_k \; = \; \boldsymbol{a}_k^\top \boldsymbol{X} \boldsymbol{a}_k, \qquad k = 1, \cdots, m \\ & \text{rank}(\boldsymbol{X}) = 1 \end{array}$$

Solving quadratic systems is essentially low-rank matrix completion

A natural least-squares formulation

$$\text{given:} \qquad y_k \;=\; (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2, \quad 1 \le k \le m$$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{k=1}^{m} \Big[\big(\boldsymbol{a}_k^\top \boldsymbol{x} \big)^2 - y_k \Big]^2$$

A natural least-squares formulation

$$\text{given:} \qquad y_k \;=\; (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2, \quad 1 \le k \le m$$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\Big[\big(\boldsymbol{a}_k^\top \boldsymbol{x}\big)^2 - y_k\Big]^2$$

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

A natural least-squares formulation

$$\text{given:} \qquad y_k \;=\; (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2, \quad 1 \le k \le m$$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\Big[\big(\boldsymbol{a}_k^\top \boldsymbol{x}\big)^2 - y_k\Big]^2$$

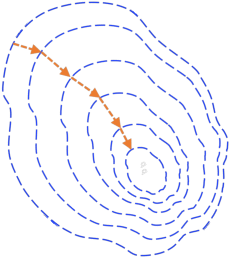
- **pros:** often exact as long as sample size is sufficiently large
- **cons:** $f(\cdot)$ is highly nonconvex
  $\longrightarrow$ *computationally challenging!*

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\left[\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - y_k\right]^2$$

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

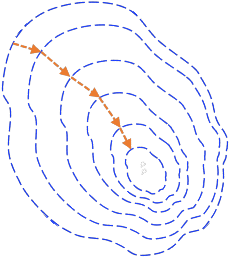
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{k=1}^{m} \left[\left(\boldsymbol{a}_k^\top \boldsymbol{x} \right)^2 - y_k \right]^2$$



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

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\left[\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - y_k\right]^2$$



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

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

Spectral initialization

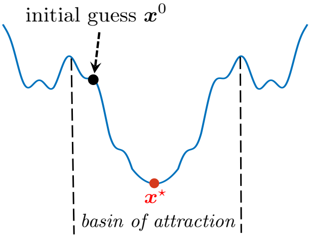
$\boldsymbol{x}^0 \longleftarrow$ leading eigenvector of

$$\boldsymbol{Y} := \frac{1}{m}\sum_{k=1}^{m} y_k\, \boldsymbol{a}_k \boldsymbol{a}_k^\top$$

Rationale: under random Gaussian design $\boldsymbol{a}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$,

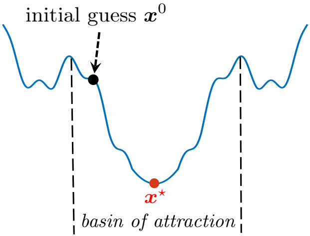
$$\mathbb{E}[\boldsymbol{Y}] := \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m} \boldsymbol{y}_k \boldsymbol{a}_k \boldsymbol{a}_k^\top\right] = \underbrace{\|\boldsymbol{x}^\star\|_2^2\, \boldsymbol{I} + 2\boldsymbol{x}^\star \boldsymbol{x}^{\star\top}}_{\text{leading eigenvector: } \pm\boldsymbol{x}^\star}$$

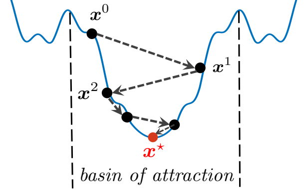
Rationale of two-stage approach



1. initialize within $\underbrace{\text{local basin sufficiently close to } \boldsymbol{x}^\star}_{\text{(restricted) strongly convex; no saddles / spurious local mins}}$

Rationale of two-stage approach





1. initialize within $\underbrace{\text{local basin sufficiently close to } \boldsymbol{x}^\star}_{\text{(restricted) strongly convex; no saddles / spurious local mins}}$

2. iterative refinement

A highly incomplete list of two-stage methods

phase retrieval:

- Netrapalli, Jain, Sanghavi '13
- Candès, Li, Soltanolkotabi '14
- Chen, Candès '15
- Cai, Li, Ma '15
- Wang, Giannakis, Eldar '16
- Zhang, Zhou, Liang, Chi '16
- Kolte, Ozgur '16
- Zhang, Chi, Liang '16
- Soltanolkotabi '17
- Vaswani, Nayer, Eldar '16
- Chi, Lu '16
- Wang, Zhang, Giannakis, Akcakaya, Chen '16
- Tan, Vershynin '17
- Ma, Wang, Chi, Chen '17
- Duchi, Ruan '17
- Jeong, Gunturk '17
- Yang, Yang, Fang, Zhao, Wang, Neykov '17
- Qu, Zhang, Wright '17
- Goldstein, Studer '16
- Bahmani, Romberg '16
- Hand, Voroninski '16
- Wang, Giannakis, Saad, Chen '17
- Barmherzig, Sun '17
- ...

other problems:

- Keshavan, Montanari, Oh '09
- Sun, Luo '14
- Chen, Wainwright '15
- Tu, Boczar, Simchowitz, Soltanolkotabi, Recht '15
- Zheng, Lafferty '15
- Balakrishnan, Wainwright, Yu '14
- Chen, Suh '15
- Chen, Candès '16
- Li, Ling, Strohmer, Wei '16
- Yi, Park, Chen, Caramanis '16
- Jin, Kakade, Netrapalli '16
- Huang, Kakade, Kong, Valiant '16
- Ling, Strohmer '17
- Li, Ma, Chen, Chi '18
- Aghasi, Ahmed, Hand '17
- Lee, Tian, Romberg '17
- Li, Chi, Zhang, Liang '17
- Cai, Wang, Wei '17
- Abbe, Bandeira, Hall '14
- Chen, Kamath, Suh, Tse '16
- Zhang, Zhou '17
- Boumal '16
- Zhong, Boumal '17
- ...

Computational cost

$$\boldsymbol{A}\boldsymbol{x} := \left[\boldsymbol{a}_k^\top \boldsymbol{x}\right]_{1\le k\le m}$$

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

$$\frac{1}{m}\sum_{k=1}^{m} y_k\, \boldsymbol{a}_k\boldsymbol{a}_k^\top = \frac{1}{m}\boldsymbol{A}^\top \,\mathrm{diag}\{y_k\}\, \boldsymbol{A}$$

Computational cost

$$\boldsymbol{A}\boldsymbol{x} := \left[\boldsymbol{a}_k^\top \boldsymbol{x}\right]_{1\leq k\leq m}$$

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

$$\frac{1}{m}\sum_{k=1}^{m} y_k\, \boldsymbol{a}_k \boldsymbol{a}_k^\top = \frac{1}{m}\boldsymbol{A}^\top \, \mathrm{diag}\{y_k\}\, \boldsymbol{A}$$

- **Iterations:** one application of $\boldsymbol{A}$ and $\boldsymbol{A}^\top$ per iteration

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

Asymptotic notation

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

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

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

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

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

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

First theory of WF

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

Theorem 4 (Candès, Li, Soltanolkotabi '14)

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

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

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

First theory of WF

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

Theorem 4 (Candès, Li, Soltanolkotabi '14)

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

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

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

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

First theory of WF

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

Theorem 4 (Candès, Li, Soltanolkotabi '14)

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

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

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

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

Theorem 4 (Candès, Li, Soltanolkotabi '14)

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

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

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

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

Theorem 5 (Ma, Wang, Chi, Chen '17)

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

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

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

What does optimization theory say about WF?

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

What does optimization theory say about WF?

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

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

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

What does optimization theory say about WF?

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

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

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

What does optimization theory say about WF?

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

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

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

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 $\eta = O(1/n)$

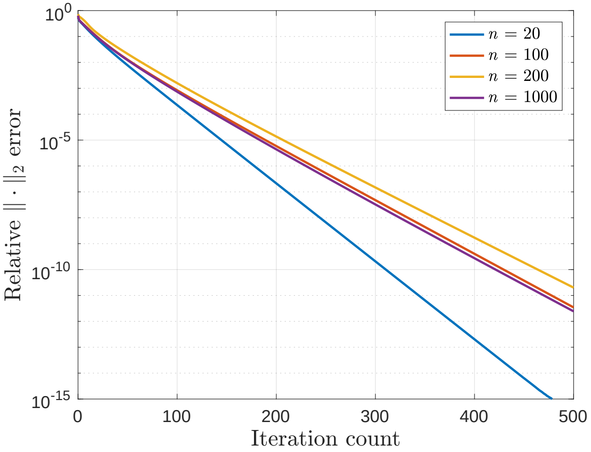
$\Uparrow$

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?

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

A second look at gradient descent theory

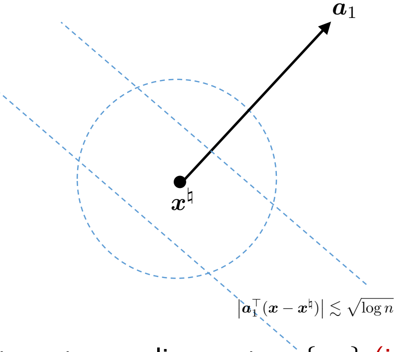
Which local region enjoys both strong convexity and smoothness?

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

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

A second look at gradient descent theory

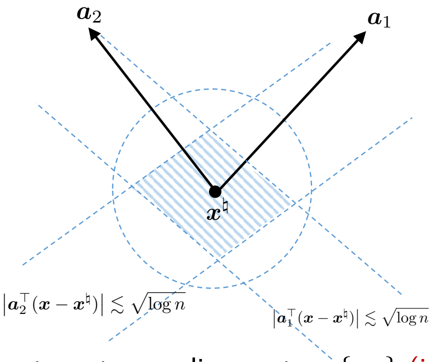
Which local region enjoys both strong convexity and smoothness?



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

A second look at gradient descent theory

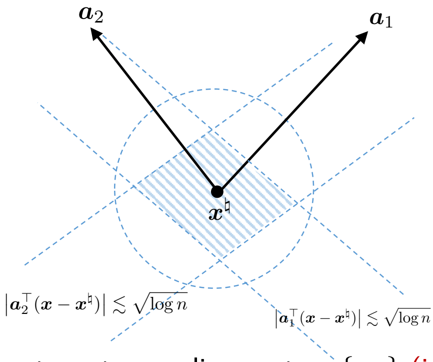
Which local region enjoys both strong convexity and smoothness?



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

A second look at gradient descent theory

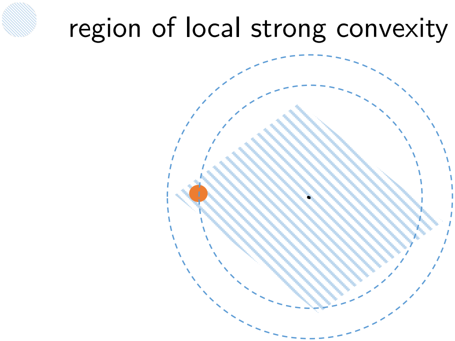
Which local region enjoys both strong convexity and smoothness?



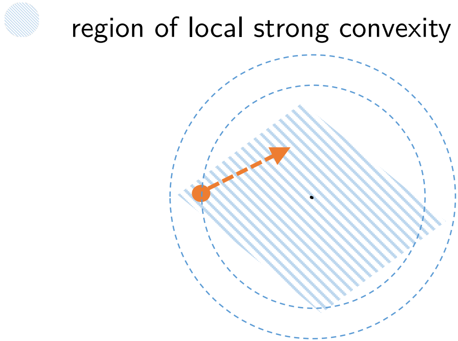
- $\boldsymbol{x}$ is incoherent w.r.t. sampling vectors $\{\boldsymbol{a}_k\}$ (incoherence region)

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

Encouraging message: GD is implicitly regularized

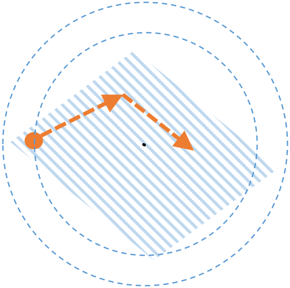


Encouraging message: GD is implicitly regularized

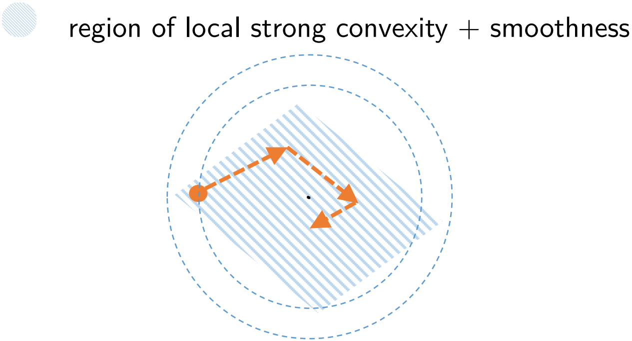


Encouraging message: GD is implicitly regularized

region of local strong convexity + smoothness

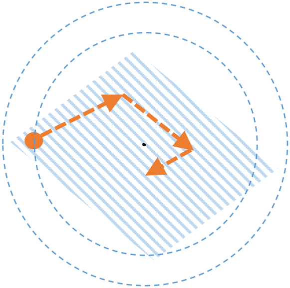


Encouraging message: GD is implicitly regularized



Encouraging message: GD is implicitly regularized

region of local strong convexity + smoothness



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

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

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

Theoretical guarantees for local refinement stage

Theorem 6 (Ma, Wang, Chi, Chen '17)

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

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

Theoretical guarantees for local refinement stage

Theorem 6 (Ma, Wang, Chi, Chen '17)

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

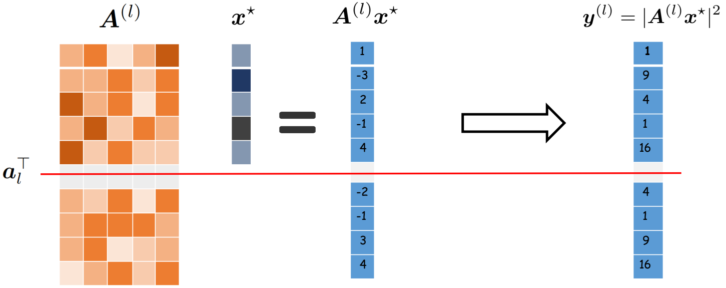
- $\max_k \left|\boldsymbol{a}_k^\top \boldsymbol{x}^t\right| \lesssim \sqrt{\log n}\, \|\boldsymbol{x}^\star\|_2$ *(incoherence)*
- $\mathsf{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\star\|_2$ *(linear convergence)*

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

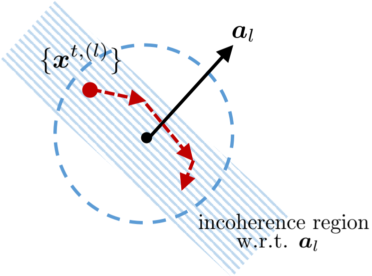
- Attains ε accuracy within $O\big(\log n \,\log \frac{1}{\varepsilon}\big)$ iterations

Key proof idea: leave-one-out analysis

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

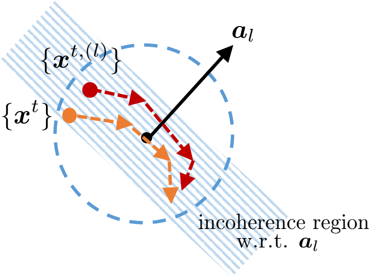


Key proof idea: leave-one-out analysis



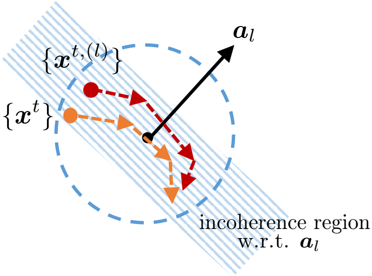
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$

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

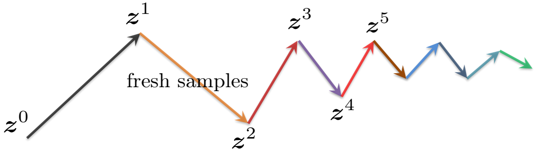


- Leave-one-out iterate $\boldsymbol{x}^{t,(l)}$ is independent of $\boldsymbol{a}_l$
- Leave-one-out iterate $\boldsymbol{x}^{t,(l)} \approx$ true iterate $\boldsymbol{x}^t$

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

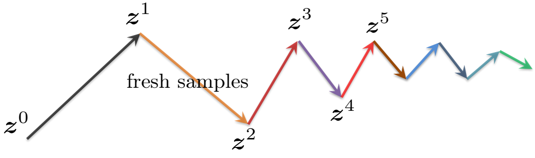
No need of sample splitting

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

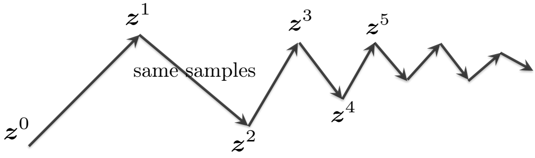


No need of sample splitting

- 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

spectral initialization $+$ vanilla gradient descent (WF)

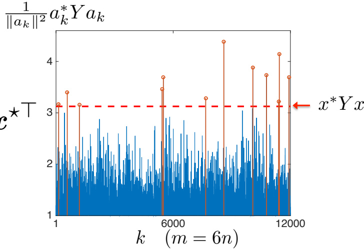
Questions:

- Can we further improve sample complexity?
- Robustness vis a vis noise and outliers?

Can we further improve sample complexity?

Truncated spectral initialization

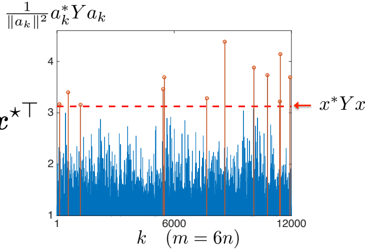
$$\mathbb{E}[\boldsymbol{Y}] := \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m} \boldsymbol{y}_k \boldsymbol{a}_k \boldsymbol{a}_k^\top\right] = \|\boldsymbol{x}^\star\|_2^2\, \boldsymbol{I} + 2\boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$$



problem: unless $m \gg n$, dangerous to use empirical average because large observations $y_k = (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2$ bear too much influence

Truncated spectral initialization

$$\mathbb{E}[\boldsymbol{Y}] := \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m} \boldsymbol{y}_k \boldsymbol{a}_k \boldsymbol{a}_k^\top\right] = \|\boldsymbol{x}^\star\|_2^2\, \boldsymbol{I} + 2\boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$$

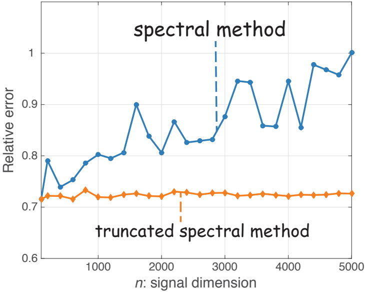


problem: unless $m \gg n$, dangerous to use empirical average because large observations $y_k = (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2$ bear too much influence

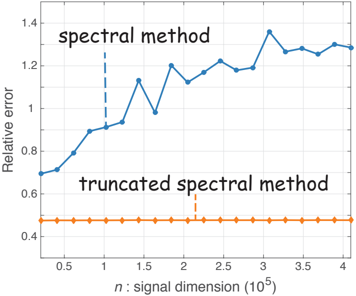
solution: discard high leverage samples and compute leading eigenvector of truncated sum

$$\frac{1}{m}\sum_{k=1}^{m} y_k \boldsymbol{a}_k \boldsymbol{a}_k^\top \cdot \mathbf{1}_{\{|y_k| \le \alpha^2 \operatorname{Avg}(|y_j|)\}}$$

Importance of truncated spectral initialization



real Gaussian $m = 6n$

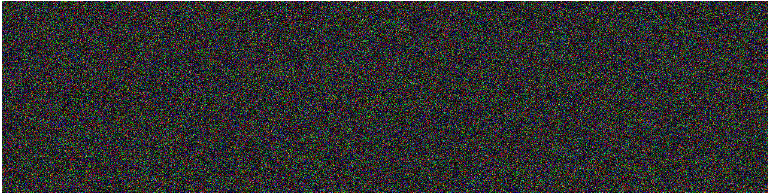


complex CDP $m = 12n$

Importance of truncated spectral initialization

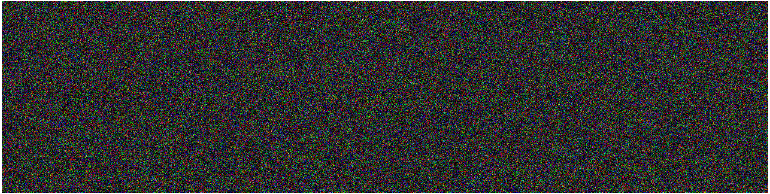
Original image

Importance of truncated spectral initialization



Spectral initialization

Importance of truncated spectral initialization



Spectral initialization

Truncated spectral initialization

Precise asymptotic characterization (Lu, Li '17)

- $m/n \asymp 1$
- i.i.d. Gaussian design

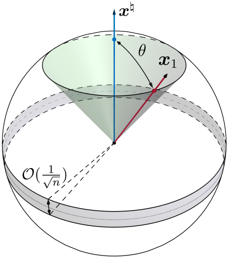


Fig. credit: Lu, Li '17

Theorem 7 (Lu, Li '17, Mondelli, Montanari '17)

There exist analytical formulas $\rho(\cdot)$ *and constants* $\alpha_{\min}$ *and* $\alpha_{\max}$ *s.t.*

$$\underbrace{\frac{\left(\boldsymbol{x}^{\star\top}\boldsymbol{x}^0\right)^2}{\|\boldsymbol{x}^\star\|_2^2\|\boldsymbol{x}^0\|_2^2}}_{\text{cosine similarity}} \longrightarrow \begin{cases} 0, & \text{if } m/n < \alpha_{\min} \\ \rho(m/n), & \text{if } m/n > \alpha_{\max} \end{cases}$$

Theoretical prediction vs. simulations

image size: 64×64

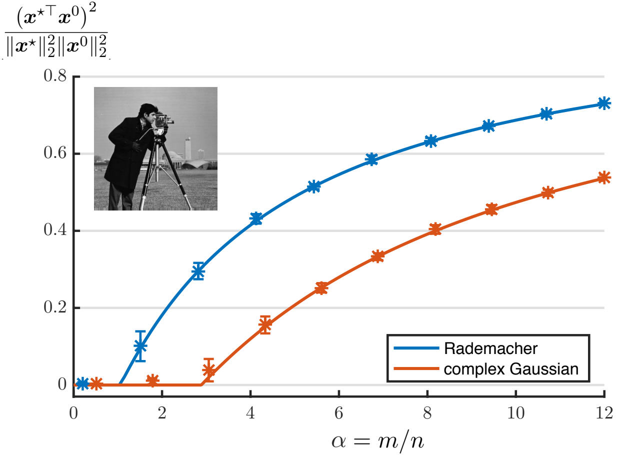


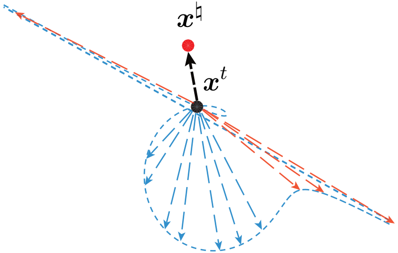
Fig. credit: Lu, Li '17

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

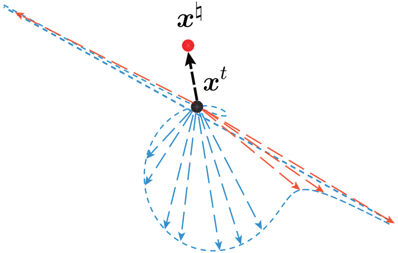
$$\text{WF (GD):} \quad \boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m} \sum_k \nabla f_k(\boldsymbol{x}^t)$$



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

Improving search directions

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

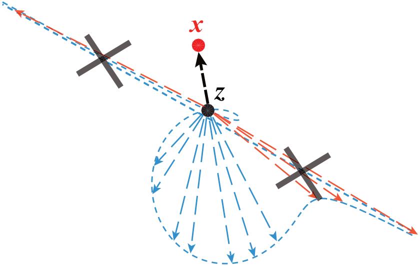


locus of $\{\nabla f_k(\boldsymbol{z})\}$

Problem: descent direction might have large variability

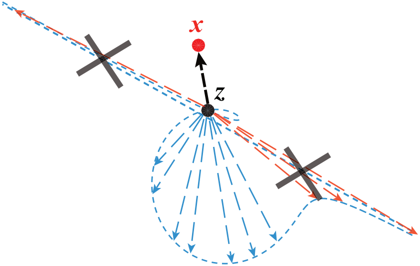
Solution: variance reduction via trimming

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



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

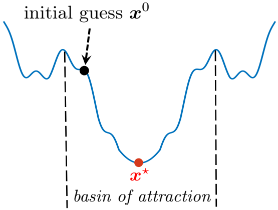
$$\frac{1}{m}\sum\nolimits_{k\in\mathcal{T}_0} y_k\,\boldsymbol{a}_k\boldsymbol{a}_k^\top$$

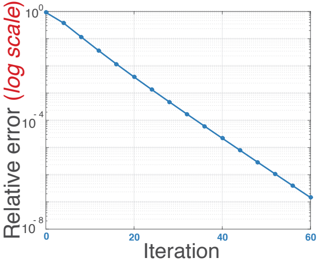
(2) Follow **adaptive gradient descent**

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

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

Theoretical guarantees (noiseless data)





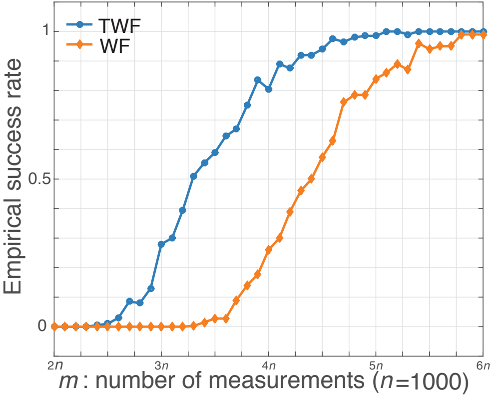
Theorem 8 (Chen, Candès '15)

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

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

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: $y_k = (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2 + \eta_k$
- Signal-to-noise ratio:

$$\mathsf{SNR} := \frac{\sum_k (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^4}{\sum_k \eta_k^2} \approx \frac{3m\|\boldsymbol{x}^\star\|_2^4}{\|\boldsymbol{\eta}\|_2^2}$$

- 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: $y_k = (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2 + \eta_k$
- Signal-to-noise ratio:

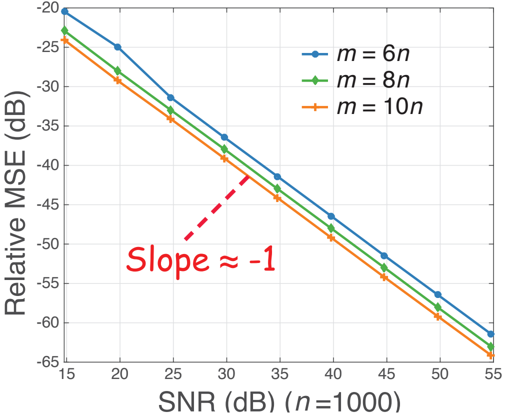
$$\mathsf{SNR} := \frac{\sum_k (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^4}{\sum_k \eta_k^2} \approx \frac{3m\|\boldsymbol{x}^\star\|_2^4}{\|\boldsymbol{\eta}\|_2^2}$$

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

Theorem 9 (Chen, Candès '15)

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

Relative MSE vs. SNR (Poisson data)

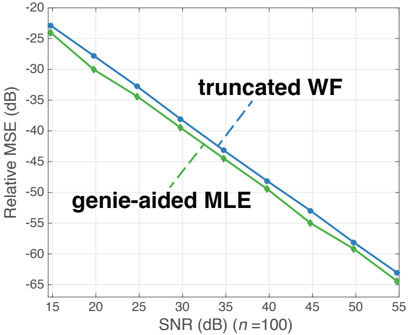


Empirical evidence: relative MSE scales inversely with SNR

This accuracy is nearly un-improvable (empirically)

Comparison with ideal MLE (with phase info. revealed)

ideal knowledge: $y_k \sim \text{Poisson}(\,|\boldsymbol{a}_k^\top \boldsymbol{x}^\star|^2\,)$ and $\varepsilon_k = \text{sign}(\boldsymbol{a}_k^\top \boldsymbol{x}^\star)$



Little loss due to missing phases!

This accuracy is nearly un-improvable (theoretically)

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

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

This accuracy is nearly un-improvable (theoretically)

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

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

Theorem 10 (Chen, Candès '15)

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

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

provided that sample size $m \asymp n$

Robust recovery vis a vis outliers

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

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

- $\|\boldsymbol{\eta}\|_0 \leq s \cdot m$: 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

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

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

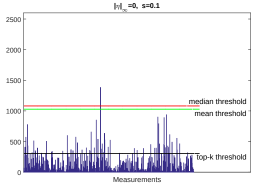
- **GD would fail:** search directions can be arbitrarily perturbed

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

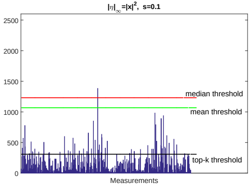
Solution: median truncation

Median is often more stable for various levels of outliers

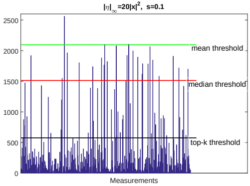
- well-known in robust statistics to be outlier-resilient



no outliers



small outlier magnitudes

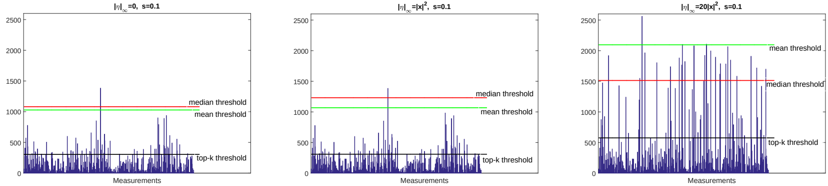


large outlier magnitudes

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 11 (Zhang, Chi and Liang '16)

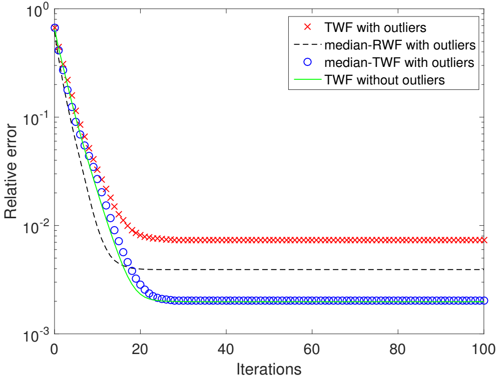
Assume $\|\boldsymbol{w}\|_\infty \leq c_1 \|\boldsymbol{x}^\star\|_2^2$, *and* $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. *If* $m \gtrsim n \log n$ *and* $s \lesssim s_0$, *then with high prob., median-TWF / RWF yields*

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

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

Other examples: low-rank matrix estimation

Low-rank matrix completion

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

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

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

In matrix completion, strong convexity and smoothness do not hold in general

$\rightarrow$ need to regularize the 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

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

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

Gradient descent for matrix completion

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

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

where $(i,j) \in \Omega$ independently with prob. p, and $E_{i,j} \sim \mathcal{N}(0, \sigma^2)$[1]

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

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

Gradient descent for matrix completion

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

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

where $(i,j) \in \Omega$ independently with prob. p, and $E_{i,j} \sim \mathcal{N}(0, \sigma^2)$[1]

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

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

[1] can be relaxed to sub-Gaussian noise and the 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 the optimal transform from the tth iterate $\boldsymbol{X}^t$ to $\boldsymbol{X}^\star$ as

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

where $\mathcal{O}^{r\times r}$ is the set of $r \times r$ orthonormal matrices

- orthogonal Procrustes problem

Gradient descent for matrix completion

Theorem 13 (Noiseless MC, Ma, Wang, Chi, Chen '17)

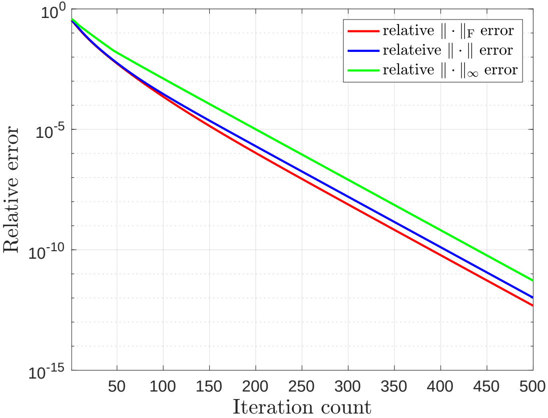
Suppose $\boldsymbol{M} = \boldsymbol{X}^\star \boldsymbol{X}^{\star\top}$ *is rank-*r*, incoherent and well-conditioned. Vanilla GD (with spectral initialization) achieves*

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

where $0 < \rho < 1$, *if the step size* $\eta \asymp 1/\sigma_{max}$ *and the 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 14 (Noisy MC, Ma, Wang, Chi, Chen '17)

Under the sample complexity of Theorem 13, if the noise satisfies $\sigma\sqrt{\frac{n}{p}} \ll \frac{\sigma_{\min}}{\sqrt{\kappa^3 \mu r \log^3 n}}$, *then the GD iterates satisfy*

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

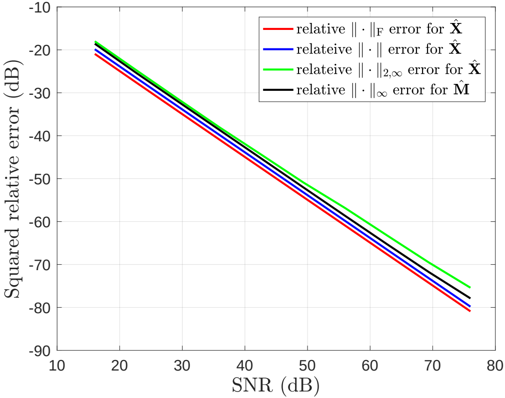
$$\left\|\boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\star\right\|_{2,\infty} \lesssim \left(\rho^t \mu r \sqrt{\frac{\log n}{np}} + \frac{\sigma}{\sigma_{\min}}\sqrt{\frac{n \log n}{p}}\right)\left\|\boldsymbol{X}^\star\right\|_{2,\infty},$$

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

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

Numerical evidence for noisy data

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



$n = 500, r = 10, p = 0.1,$ 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

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

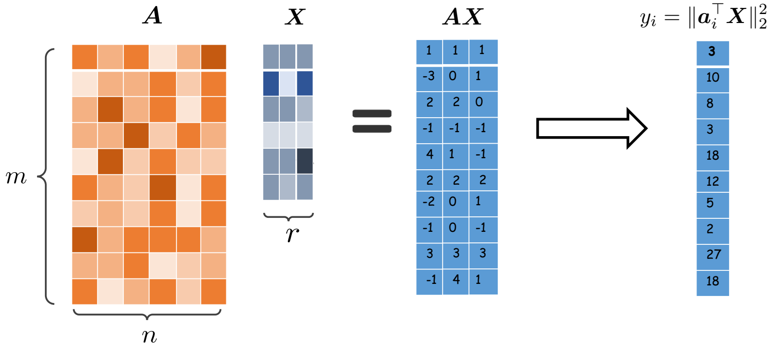
Related theory promotes incoherence explicitly:

- 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

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

Quadratic sampling



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

$$y_i \quad = \quad \|\boldsymbol{a}_i^\top \boldsymbol{X}^\star\|_2^2, \qquad i = 1, \ldots, m$$

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

Gradient descent with spectral initialization

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

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 15 (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*

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

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

Demixing sparse and low-rank matrices

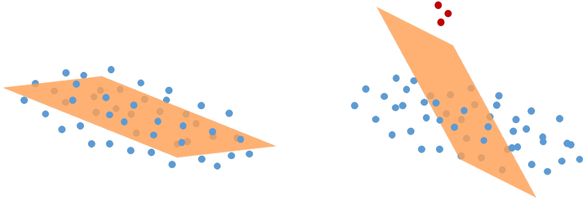
Suppose we are given a matrix

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

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

- $\text{rank}(\boldsymbol{L}) \leq r$; 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

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

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

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 16 (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

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

- $O(\kappa \log \frac{1}{\varepsilon})$ iterations to reach ε accuracy
- for adversarial outliers, optimal fraction is $s = O(1/\mu r)$; Theorem 16 is suboptimal by a factor of $\sqrt{r}$
- extendable to partial observation models

Outline

- Part I: Overview
- Part II: Preliminaries and rank-one matrix factorization
- Part IV: Two-stage approaches
  - Spectral initialization
  - Local refinement: algorithm and analysis
- Part I: Global landscape and initialization-free algorithms
  - Landscape analysis
  - Saddle-point escaping algorithms
  - Random initialization?
- Part VI: Closing remarks

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 $r \ll n$.

Rank-constrained optimization

Rank-constrained optimization:

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

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

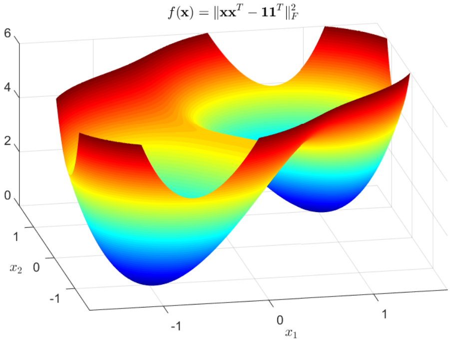
$$\Downarrow$$

Burer-Monteiro 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 $\boldsymbol{U},\boldsymbol{V} \in \mathbb{R}^{n\times r}$.

Characteristics of "benign" landscape



- all local minima are global minima
- non-local-minima critical points are strict saddles

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 17

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 17

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 17

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 18 (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 $\lambda_{\min}[\nabla^2 f(\boldsymbol{X})] \le -\frac{4}{5}\sigma_r$*

What about matrix completion?

Consider the loss function

$$f(\boldsymbol{X}) = \frac{1}{p}\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M})\|_\mathrm{F}^2$$

- The matrix RIP does not hold
- Need incoherence!

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

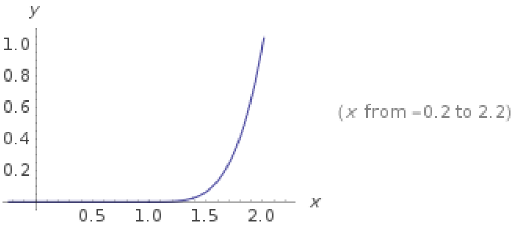
$$\|\boldsymbol{U}\|_{2,\infty} \le \sqrt{\frac{\mu}{n}}\|\boldsymbol{U}\|_\mathrm{F} = \sqrt{\frac{\mu r}{n}}.$$

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 $z_+ = \max\{z, 0\}$



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 19 (Ge et al, 2016, Chen and Li, 2017)

If sample size $n^2 p \gtrsim n \max\{\mu r \log n, \mu^2 r^2\}$ 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 of benign landscape:

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

active research area in machine learning and optimization

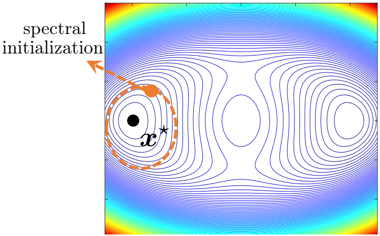
Saddle-point escaping algorithms

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

Question: Does GD converge fast with random initialization?

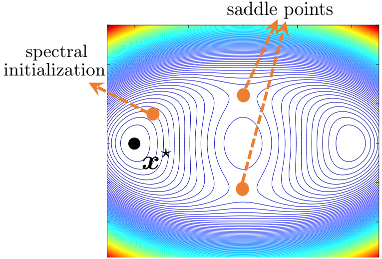
Are carefully-designed initialization or saddle-point escaping schemes necessary for fast convergence?

Initialization



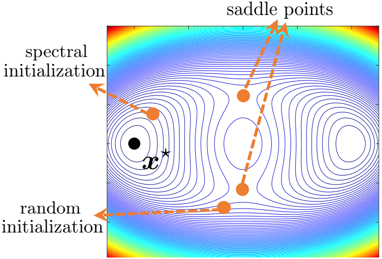
- Spectral initialization gets us reasonably close to truth

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

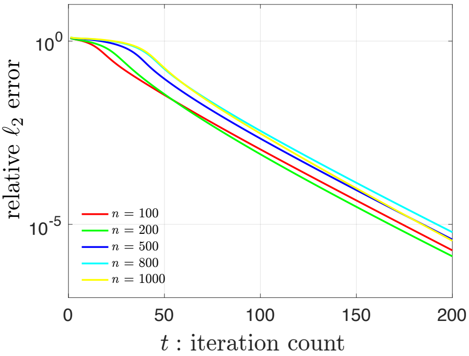


- Spectral initialization gets us reasonably close to truth
- Cannot initialize GD from anywhere, e.g. it might get stucked at local stationary points (e.g. saddle points)

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

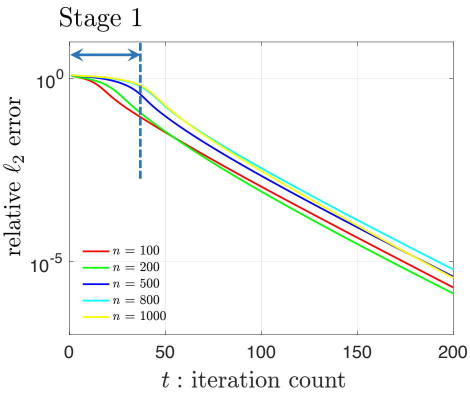
Numerical efficiency of randomly initialized GD

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



Numerical efficiency of randomly initialized GD

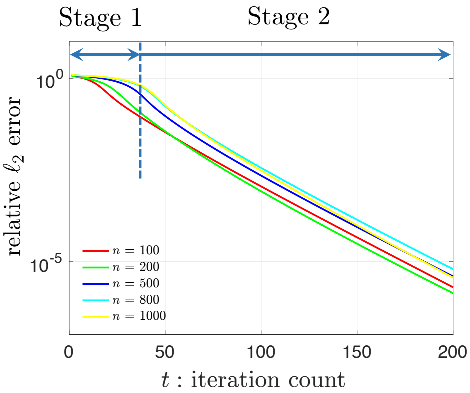
$$\eta_t = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\mathbf{0}, n^{-1}\boldsymbol{I}_n)$$



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

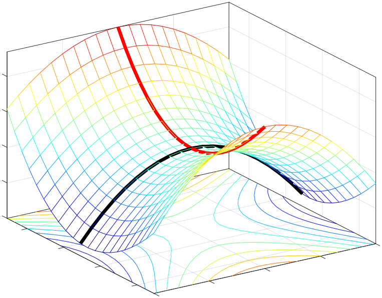
Numerical efficiency of randomly initialized GD

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



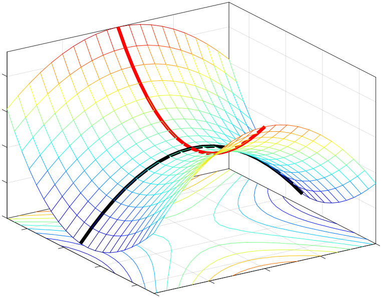
Randomly initialized GD enters local basin within **a few iterations**

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)

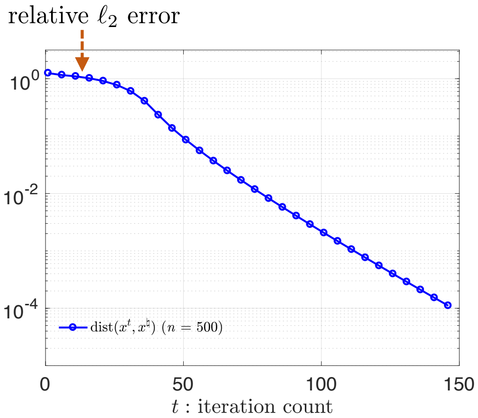
A geometric analysis



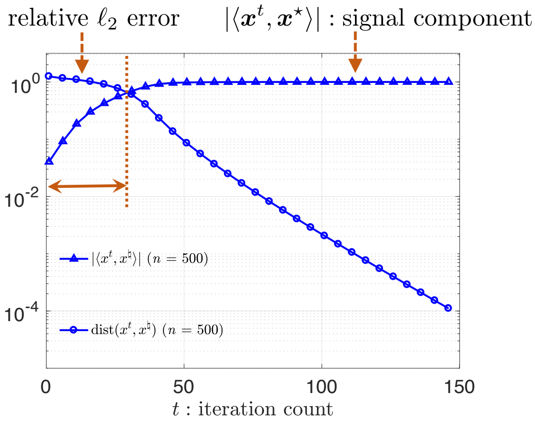
- With such benign landscape, GD with random initialization converges to global min almost surely (Lee et al. '16)

No convergence rate guarantees for vanilla GD!

Exponential growth of signal strength in Stage 1

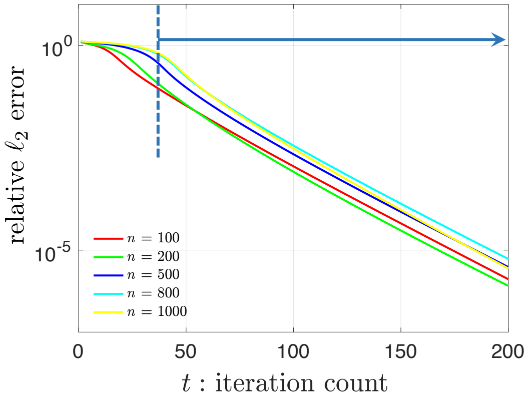


Exponential growth of signal strength in Stage 1

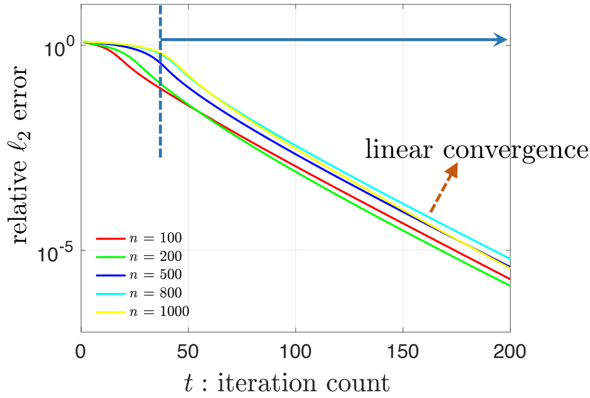


Numerically, $O(\log n)$ iterations are enough to enter local region

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 20 (Chen, Chi, Fan, Ma '18)

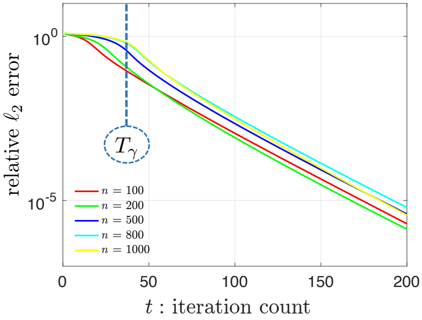
Under i.i.d. Gaussian design, GD with $\boldsymbol{x}^0 \sim \mathcal{N}(\mathbf{0}, n^{-1}\boldsymbol{I}_n)$ achieves

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \qquad t \ge T_\gamma$$

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

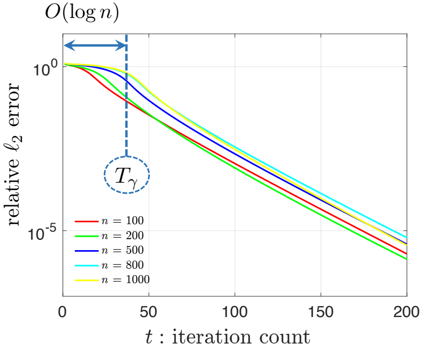
Theoretical guarantees for randomly initialized GD

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



Theoretical guarantees for randomly initialized GD

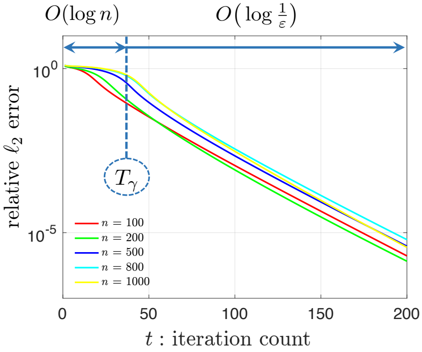
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$



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

Theoretical guarantees for randomly initialized GD

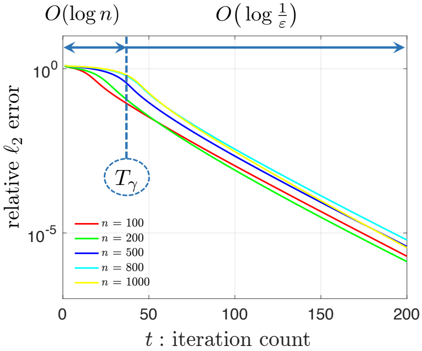
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$



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

Theoretical guarantees for randomly initialized GD

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

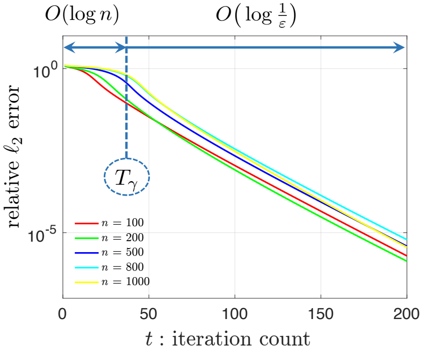


- *near-optimal compututational cost*:

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

Theoretical guarantees for randomly initialized GD

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

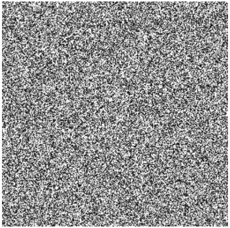


- *near-optimal compututational cost*:
  — $O\big(\log n + \log \frac{1}{\varepsilon}\big)$ iterations to yield ε accuracy
- *near-optimal sample size*: $m \gtrsim n\text{poly}\log m$

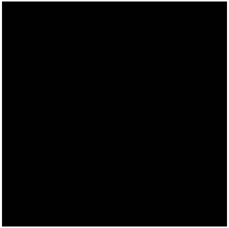
Experiments on images

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

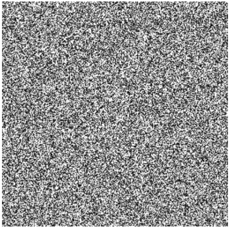
GD with random initialization



$\boldsymbol{x}^t$
GD iterate



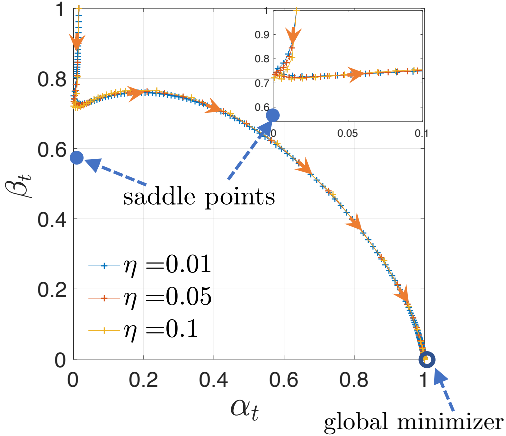
$\langle \boldsymbol{x}^t, \boldsymbol{x}^\star \rangle \boldsymbol{x}^\star$
signal component



$\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\star \rangle \boldsymbol{x}^\star$
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

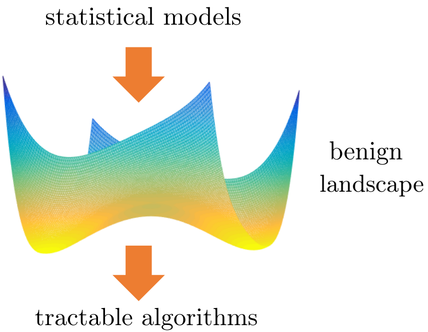
	iteration complexity	num of iterations needed to escape saddles	local iteration complexity
Trust-region (Sun et al. '16)	$n^7 + \log\log\frac{1}{\varepsilon}$	n^7	$\log\log\frac{1}{\varepsilon}$
Perturbed GD (Jin et al. '17)	$n^3 + n\log\frac{1}{\varepsilon}$	n^3	$n\log\frac{1}{\varepsilon}$
Perturbed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

Outline

- Part I: Overview
- Part II: Preliminaries and rank-one matrix factorization
- Part IV: Two-stage approaches
  - Spectral initialization
  - Local refinement: algorithm and analysis
- Part I: Global landscape and initialization-free algorithms
  - Landscape analysis
  - Saddle-point escaping algorithms
  - Random initialization?
- Part VI: Closing remarks

Statistical thinking + Optimization efficiency



When data are generated by certain statistical models, problems are often much nicer than worst-case instances

A growing list of "benign" nonconvex problems

- phase retrieval
- matrix sensing
- matrix completion
- blind deconvolution / self-calibration
- dictionary learning
- tensor decomposition
- robust PCA
- mixed linear regression
- learning one-layer neural networks
- ...

Open problems

- characterize generic landscape properties that enable fast convergence of gradient methods from random initialization
- relax the stringent assumptions on the statistical models underlying the data
- develop robust and scalable nonconvex methods that can handle distributed data with strong statistical guarantees
- identify new classes of nonconvex problems that admit efficient optimization procedures
- ...

Advertisement

"Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview"

— Y. Chi, Y. Lu and Y. Chen, arXiv: 1809.09573

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!