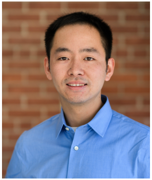
Geometry and (Implicit) Regularization in Nonconvex Low-Rank Estimation

Yuejie Chi

Carnegie Mellon University

SPARS 2019
Jul. 2019

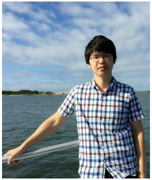
Acknowledgements



Yuxin Chen
Princeton

Jianqing Fan
Princeton

Yingbin Liang
Ohio State



Cong Ma
Princeton

Kaizheng Wang
Princeton

Yuanxin Li
CMU

Huishuai Zhang
MSRA

Empirical risk minimization

Given data $\boldsymbol{z}$, estimate parameters $\boldsymbol{x} \in \mathbb{R}^n$:

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) := \frac{1}{m} \sum_{i=1}^{m} \ell(z_i; \boldsymbol{x})$$

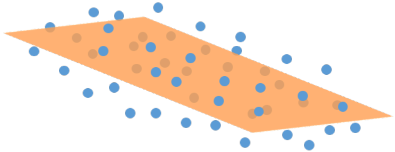
where $\ell(z_i; \boldsymbol{x})$ is the sample loss.

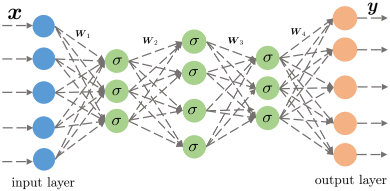
Empirical risk minimization

Given data $\boldsymbol{z}$, estimate parameters $\boldsymbol{x} \in \mathbb{R}^n$:

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) := \frac{1}{m} \sum_{i=1}^{m} \ell(z_i; \boldsymbol{x})$$

where $\ell(z_i; \boldsymbol{x})$ is the sample loss.



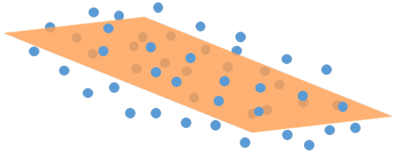


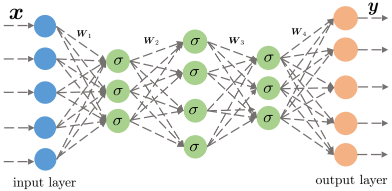
Empirical risk minimization

Given data $\boldsymbol{z}$, estimate parameters $\boldsymbol{x} \in \mathbb{R}^n$:

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) := \frac{1}{m}\sum_{i=1}^{m} \ell(z_i; \boldsymbol{x})$$

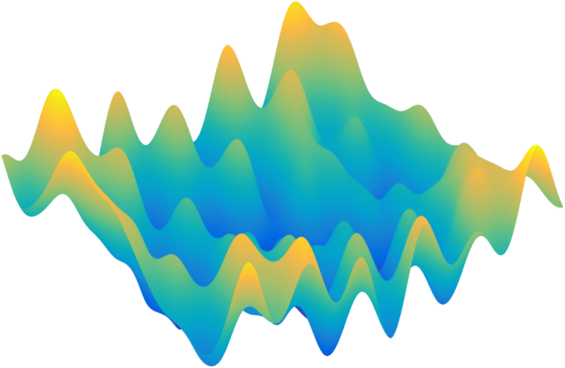
where $\ell(z_i; \boldsymbol{x})$ is the sample loss.





Often lead to nonconvex problems that are deemed intractable!

Nonconvex problems are hard!



"...in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity.

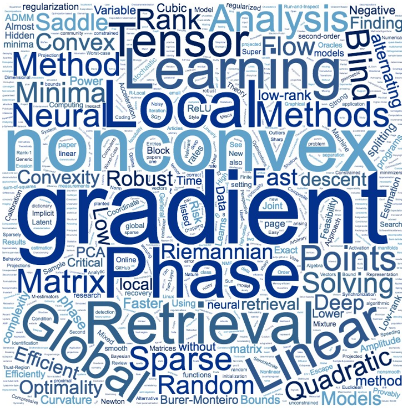
R. T. Rockafellar, in SIAM Review, 1993

Nonconvex problems are hard!

"...in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity.

R. T. Rockafellar, in SIAM Review, 1993

Recent developments: provable nonconvex optimization



Only an incomplete list...

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, Qu, Wright '16, Ma et al. '17, Chen et al. '18, Soltani, Hegde '18, Ruan and Duchi, '18, ...

Matrix sensing/completion: Keshavan et al. '09, Jain et al. '09, Hardt '13, Jain et al. '13, Sun, Luo '15, Chen, Wainwright '15, Tu et al. '15, Zheng, Lafferty '15, Bhojanapalli et al. 16, Ge, Lee, Ma '16, Jin et al. '16, Ma et al. '17, Chen and Li '17, Cai et al. '18, Li, Zhu, Tang, Wakin '18, Charisopoulos et al. '19, ...

Blind deconvolution/demixing: Li et al. '16, Lee et al. '16, Cambareri, Jacques '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, Bai et al. '18, Gilboa et al. '18, Rambhatla et al. '19, ...

Robust principal component analysis: Netrapalli et al. '14, Yi et al. '16, Gu et al. '16, Ge et al. '17, Cherapanamjeri et al. '17, Vaswani et al. '18, Maunu et al. '19, ...

Deep learning: Zhong et al. '17, Bai, Mei, Montanari '17, Du et al. '17, Ge, Lee, Ma '17, Du et al. '18, Soltanolkotabi and Oymak, '18...

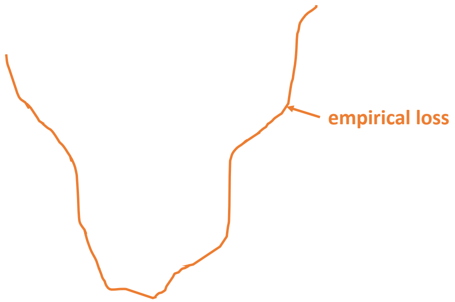
Statistical thinking in nonconvex optimization

Data/measurements follow certain **statistical models** and hence are not worst-case instances.

Statistical thinking in nonconvex optimization

Data/measurements follow certain **statistical models** and hence are not worst-case instances.

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



Statistical thinking in nonconvex optimization

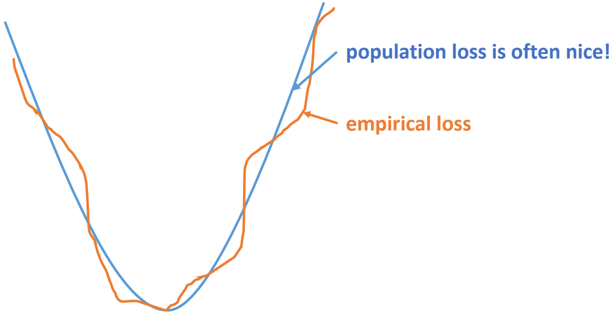
Data/measurements follow certain **statistical models** and hence are not worst-case instances.

$$\text{minimize}_{\boldsymbol{x}} \; f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} \ell(y_i; \boldsymbol{x}) \quad \overset{m\to\infty}{\Longrightarrow} \quad \mathbb{E}[\ell(y; \boldsymbol{x})]$$

Statistical thinking in nonconvex optimization

Data/measurements follow certain **statistical models** and hence are not worst-case instances.

$$\text{minimize}_{\boldsymbol{x}}\ f(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\ell(y_i;\boldsymbol{x}) \quad \stackrel{m\to\infty}{\Longrightarrow} \quad \mathbb{E}[\ell(y;\boldsymbol{x})]$$

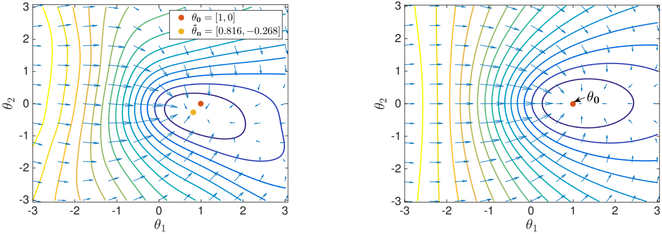


We will detail an example of "nice" population landscape later.

From population to empirical risk: a geometric perspective

Geometric analysis: uniform concentration of Hessians and gradients, along some descent directions;

- one-to-one correspondence between critical points;
- preservation of geometric curvatures.



empirical risk $\approx$ population risk

Bai et al. '16, Sun et al. '15, Sun et al. '16, Ge et al. '16; Figure credit: Bai, Mei, and Montanari

This talk: sample-starved regime

Sample-starved regime:

sample size $\gtrsim$ *O(number of unknowns)*

Sample-starved regime:

sample size $\gtrsim$ O(number of unknowns)

Even when $\mathbb{E}[f(\boldsymbol{x})]$ is locally strongly convex and smooth,

- $f(\boldsymbol{x})$ *may be much more ill-conditioned than* $\mathbb{E}[f(\boldsymbol{x})]$*;*

  smaller step size and more computation

This talk: sample-starved regime

Sample-starved regime:

sample size $\gtrsim$ O(number of unknowns)

Even when $\mathbb{E}[f(\boldsymbol{x})]$ is locally strongly convex and smooth,

- $f(\boldsymbol{x})$ *may be much more ill-conditioned than* $\mathbb{E}[f(\boldsymbol{x})]$*;*

  smaller step size and more computation

- $f(\boldsymbol{x})$ *may lack curvatures in certain regions.*

  complicated regularization

This talk: sample-starved regime

Sample-starved regime:

sample size $\gtrsim$ *O(number of unknowns)*

Even when $\mathbb{E}[f(\boldsymbol{x})]$ is locally strongly convex and smooth,

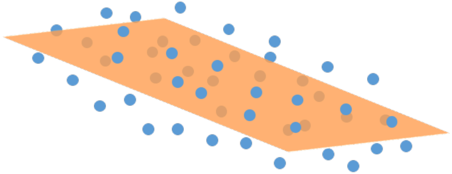
- $f(\boldsymbol{x})$ *may be much more ill-conditioned than* $\mathbb{E}[f(\boldsymbol{x})]$*;*
  smaller step size and more computation
- $f(\boldsymbol{x})$ *may lack curvatures in certain regions.*
  complicated regularization

$$\Downarrow$$

Does the geometric gap between $f(x)$ vs $\mathbb{E}[f(x)]$ hurt optimization efficacy?

a case study with low-rank matrix completion

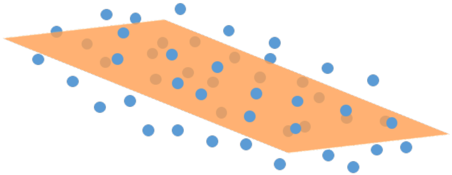
Revisiting PCA: in search of low-rank representation



Given $\boldsymbol{M} \succeq 0 \in \mathbb{R}^{n \times n}$ (e.g. sample covariance matrix), find its best rank-r approximation:

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

Revisiting PCA: in search of low-rank representation



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 $\boldsymbol{M}$ $(\lambda_1 \geq \cdots \lambda_r > \lambda_{r+1} \cdots \geq \lambda_n)$, then

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

— *nonconvex, but tractable*

An optimization viewpoint

Low-rank factorization: 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:

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

An optimization viewpoint

Low-rank factorization: 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:

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

Theorem (Baldi and Hornik, 1989)

Suppose $\boldsymbol{M}$ has a strict eigen-gap between λ_r and λ_{r+1}, the critical points of $f(\boldsymbol{X})$ can be categorized into

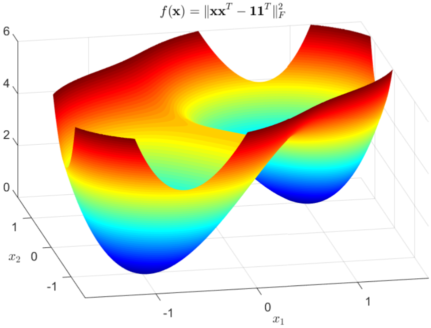
- *global minima;*
- *strict saddle points, from which there exist directions to strictly decrease $f(\boldsymbol{X})$.*

In other words, all local minima are global minima!

Baldi and Hornik. "Neural networks and principal component analysis: Learning from examples without local minima." Neural networks 2.1 (1989): 53-58.

Benign landscape of PCA

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

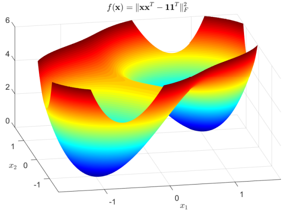


global minima: $\boldsymbol{x} = \pm \begin{bmatrix} 1 \\ 1 \end{bmatrix}$; strict saddles: $\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, and $\pm \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

— *No "spurious" local minima!*

Parameter recovery via gradient descent

a two-step recovery strategy:



- **Find an initial point that falls into a "basin of attraction"**

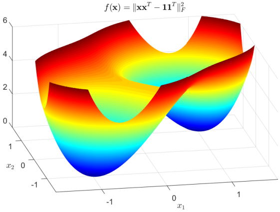
- **Gradient iterations:**

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

for $t = 0, 1, \ldots$

Parameter recovery via gradient descent

a two-step recovery strategy:



- **Find an initial point that falls into a "basin of attraction"**
- **Gradient iterations:**

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

for $t = 0, 1, \ldots$

- The spectral method can be used for initialization;
- Low-complexity local refinements via gradient descent.

Low-rank matrix completion: dealing with missing data

$$\begin{bmatrix} \checkmark & ? & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \\ \checkmark & ? & ? & \checkmark & ? & ? \\ ? & ? & \checkmark & ? & ? & \checkmark \\ \checkmark & ? & ? & ? & ? & ? \\ ? & \checkmark & ? & ? & \checkmark & ? \\ ? & ? & \checkmark & \checkmark & ? & ? \end{bmatrix}$$

Given partial samples of a *low-rank* matrix $\boldsymbol{M}$ in an index set Ω, fill in missing entries.

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

Applications: recommendation systems, ...

Incoherence

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}} \quad \text{vs.} \quad \underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}}$$

Incoherence

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

Definition (Incoherence for matrix completion)

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

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

Note: $\|\boldsymbol{U}\|_{2,\infty} = \max_i \|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2$.

Lower bound [Candès and Tao]: $p \gtrsim \mu r \log n / n$.

Incoherence

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

Definition (Incoherence for matrix completion)

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

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

Note: $\|\boldsymbol{U}\|_{2,\infty} = \max_i \|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2$.

Lower bound [Candès and Tao]: $p \gtrsim \mu r \log n / n$.

A natural least-squares formulation

given: $\mathcal{P}_\Omega(\boldsymbol{M})$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \left\| \mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}) \right\|_{\mathrm{F}}^2$$

A natural least-squares formulation

$$\text{given:} \quad \mathcal{P}_\Omega(\boldsymbol{M})$$

$$\Downarrow$$

$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \left\| \mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}) \right\|_{\mathrm{F}}^2$$

- **Bernoulli sampling:** *Assume every entry is observed i.i.d. with* $0 < p \leq 1$*:*

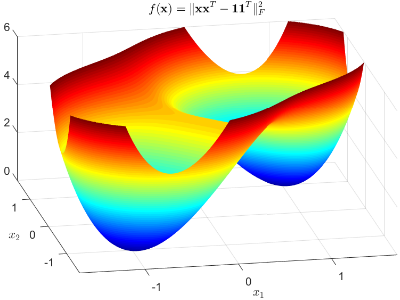
$$\mathbb{E}[f(\boldsymbol{X})] = p \left\| \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M} \right\|_{\mathrm{F}}^2 .$$

What does the population level look like?

Population level ($p = 1$): this is PCA.

$$f(\boldsymbol{X}) \text{ restricted strongly convex and smooth}$$

along descent direction $\boldsymbol{V}$ when $\boldsymbol{X}$ is close to $\boldsymbol{X}^\natural$.

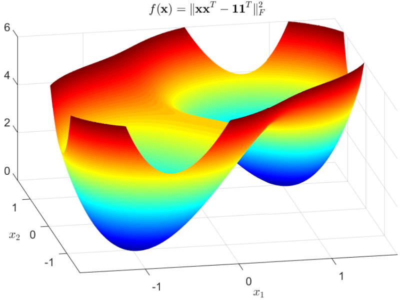


What does the population level look like?

Population level ($p = 1$): this is PCA.

$$f(\boldsymbol{X}) \text{ restricted strongly convex and smooth}$$

along descent direction $\boldsymbol{V}$ when $\boldsymbol{X}$ is close to $\boldsymbol{X}^\natural$.



Consequence: GD converges within $O\left(\log \frac{1}{\varepsilon}\right)$ iterations if $p = 1$.

What does the finite-sample level look like?

Assume every entry is observed i.i.d. with probability $0 < p \leq 1$.

$$f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - M_{j,k}\right)^2$$

What does the finite-sample level look like?

Assume every entry is observed i.i.d. with probability $0 < p \leq 1$.

$$f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - M_{j,k}\right)^2$$

Finite-sample level ($p \asymp \frac{\mathrm{polylog} n}{n}$**)**

$f\left(\boldsymbol{X}\right)$ restricted strongly convex and smooth

along descent direction $\boldsymbol{V}$ only when $\boldsymbol{X}$ is incoherent:

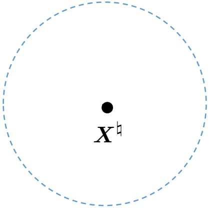
$$\|\boldsymbol{X} - \boldsymbol{X}^\natural\|_{2,\infty} \ll \|\boldsymbol{X}^\natural\|_{2,\infty}$$

Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

Incoherence region

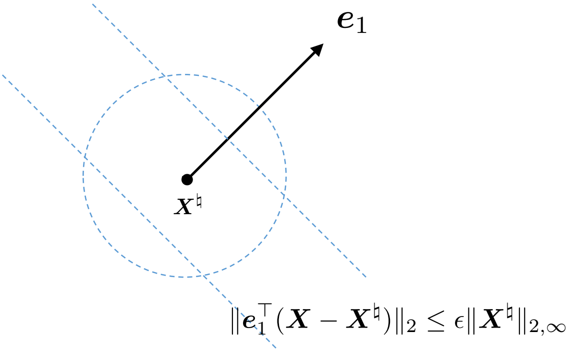
Which region enjoys both restricted strong convexity and smoothness?



- $\boldsymbol{X}$ is not far away from $\boldsymbol{X}^\natural$

Incoherence region

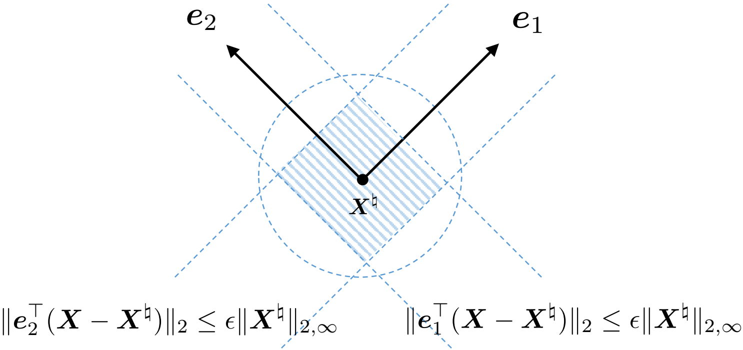
Which region enjoys both restricted strong convexity and smoothness?



- $\boldsymbol{X}$ is not far away from $\boldsymbol{X}^\natural$
- $\boldsymbol{X}$ is incoherent w.r.t. coordinates (incoherence region)

Incoherence region

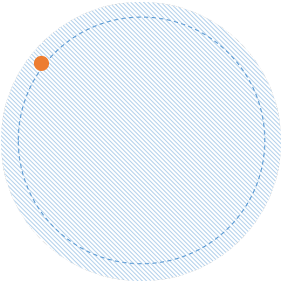
Which region enjoys both restricted strong convexity and smoothness?



- X is not far away from $X^\natural$
- X is incoherent w.r.t. coordinates (incoherence region)

Vanilla gradient descent is at risk

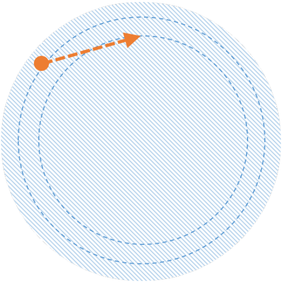
region of local strong convexity + smoothness



GD on the pop. loss

Vanilla gradient descent is at risk

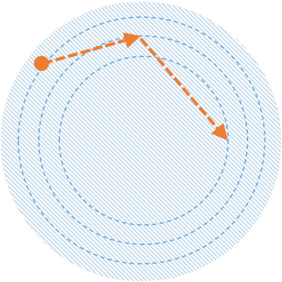
region of local strong convexity + smoothness



GD on the pop. loss

Vanilla gradient descent is at risk

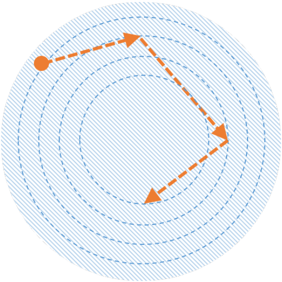
region of local strong convexity + smoothness



GD on the pop. loss

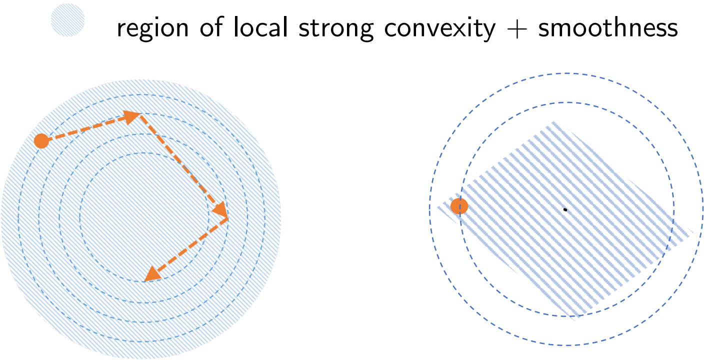
Vanilla gradient descent is at risk

region of local strong convexity + smoothness



GD on the pop. loss

Vanilla gradient descent is at risk



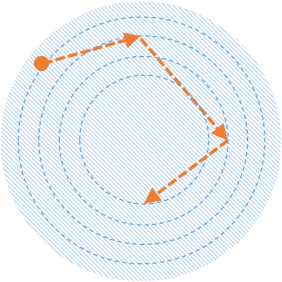
GD on the pop. loss

GD on the emp. loss

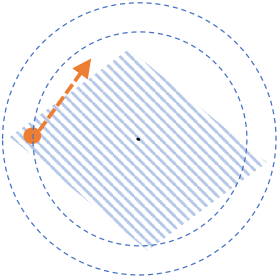
- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

Vanilla gradient descent is at risk

region of local strong convexity + smoothness



GD on the pop. loss

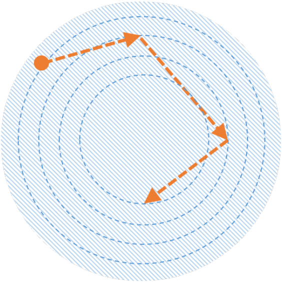


GD on the emp. loss

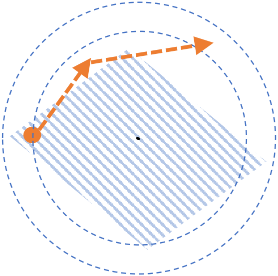
- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

Vanilla gradient descent is at risk

 region of local strong convexity + smoothness



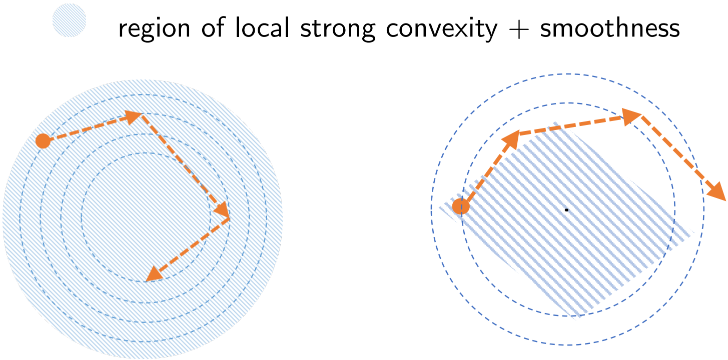
GD on the pop. loss



GD on the emp. loss

- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

Vanilla gradient descent is at risk

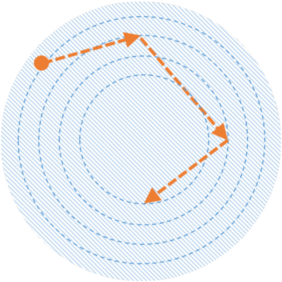


GD on the pop. loss *GD on the emp. loss*

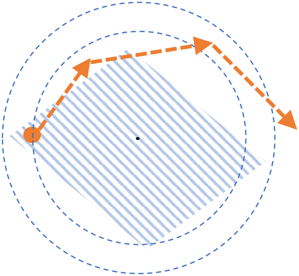
- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region

Vanilla gradient descent is at risk

region of local strong convexity + smoothness



GD on the pop. loss

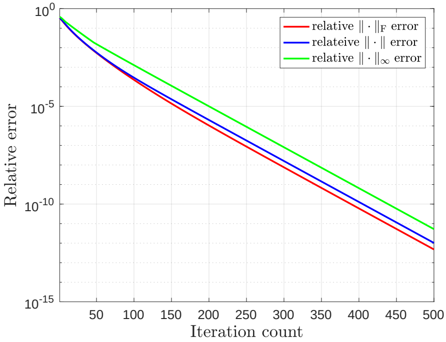


GD on the emp. loss

- Generic optimization theory only ensures that iterates remain in ℓ_2 ball but not incoherence region
- *Existing algorithms enforce regularization, or apply sample splitting to promote incoherence*

Matrix completion via vanilla GD

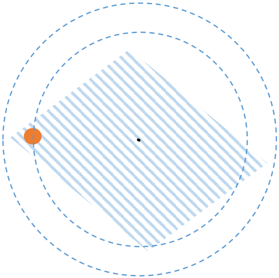
$$\text{minimize}_{\boldsymbol{X}\in\mathbb{R}^{n\times r}} \quad f(\boldsymbol{X}) = \sum_{(j,k)\in\Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{e}_k - M_{j,k}\right)^2$$



Vanilla GD converges fast without regularization!

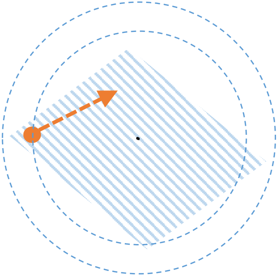
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



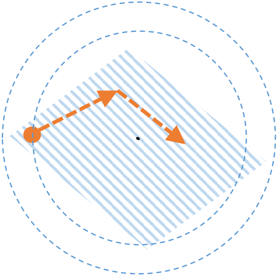
Our findings: GD is implicitly regularized

 region of local strong convexity + smoothness



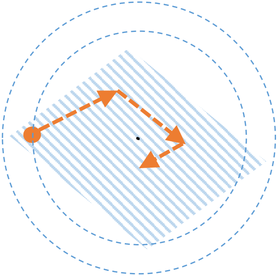
Our findings: GD is implicitly regularized

region of local strong convexity + smoothness

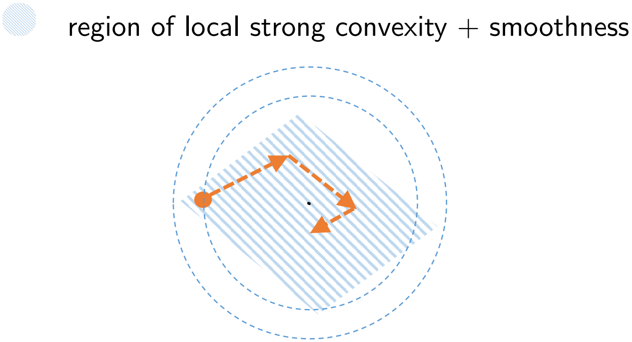


Our findings: GD is implicitly regularized

region of local strong convexity + smoothness



Our findings: GD is implicitly regularized



GD implicitly forces iterates to remain incoherent
even without regularization

Theoretical guarantees - noise-free case

Theorem (Ma, Wang, Chi, Chen, FoCM 2019+)

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

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

where $\rho = 1 - \frac{\sigma_{\min}\eta}{5} < 1$, *if step size* $\eta \asymp 1/\sigma_{\max}$ *and sample complexity* $n^2 p \gtrsim \mu^3 n r^3 \log^3 n$.

Theoretical guarantees - noise-free case

Theorem (Ma, Wang, Chi, Chen, FoCM 2019+)

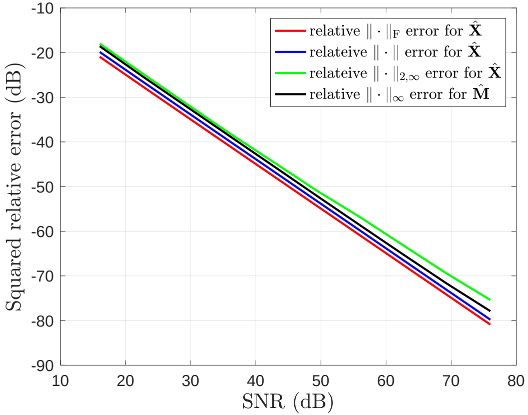
Suppose $\boldsymbol{M} = \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}$ *is rank-*r*, incoherent and well-conditioned. Vanilla GD (with spectral initialization) achieves*

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

where $\rho = 1 - \frac{\sigma_{\min}\eta}{5} < 1$*, if step size* $\eta \asymp 1/\sigma_{\max}$ *and sample complexity* $n^2 p \gtrsim \mu^3 n r^3 \log^3 n$.

- A recent follow-up by Xiaodong Li studied the rectangular case and improved the sample complexity to $O(\mu^2 n r^2 \log n)$.

Noisy matrix completion via vanilla GD

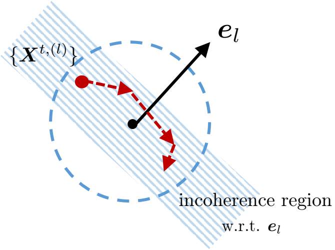


Near-optimal entry-wise error control:

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

Key ingredient: leave-one-out analysis

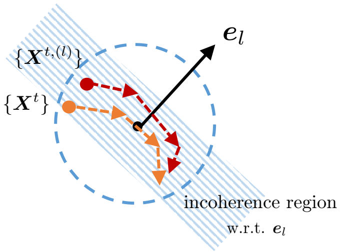
How to establish $\left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^t - \boldsymbol{X}^\natural)\right\|_2 \ll \|\boldsymbol{X}^\natural\|_{2,\infty}$?



- Create auxiliary leave-one-out iterates $\{\boldsymbol{X}^{t,(l)}\}$ are incoherent in the lth row;

Key ingredient: leave-one-out analysis

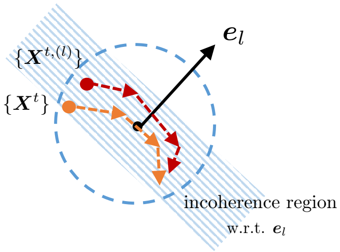
How to establish $\left\|e_l^\top(X^t - X^\natural)\right\|_2 \ll \|X^\natural\|_{2,\infty}$?



- Create auxiliary leave-one-out iterates $\{X^{t,(l)}\}$ are incoherent in the lth row;
- Leave-one-out iterates $X^{t,(l)} \approx$ true iterates X^t

Key ingredient: leave-one-out analysis

How to establish $\left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^t - \boldsymbol{X}^\natural)\right\|_2 \ll \|\boldsymbol{X}^\natural\|_{2,\infty}$?



- Create auxiliary leave-one-out iterates $\{\boldsymbol{X}^{t,(l)}\}$ are incoherent in the lth row;
- Leave-one-out iterates $\boldsymbol{X}^{t,(l)} \approx$ true iterates $\boldsymbol{X}^t$
- $\left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^t - \boldsymbol{X}^\natural)\right\|_2 \le \left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^{t,(l)} - \boldsymbol{X}^\natural)\right\|_2 + \left\|\boldsymbol{e}_l^\top(\boldsymbol{X}^t - \boldsymbol{X}^{t,(l)})\right\|_2$

An aside: stability of nuclear norm minimization

convex

nonconvex

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times n}} \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_*$$

$$\min_{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}} \sum_{(i,j)\in\Omega} \left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{i,j} - M_{i,j}\right]^2 + \frac{\lambda}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{\lambda}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2$$

Theorem (Chen, Chi, Fan, Ma, Yan '19)

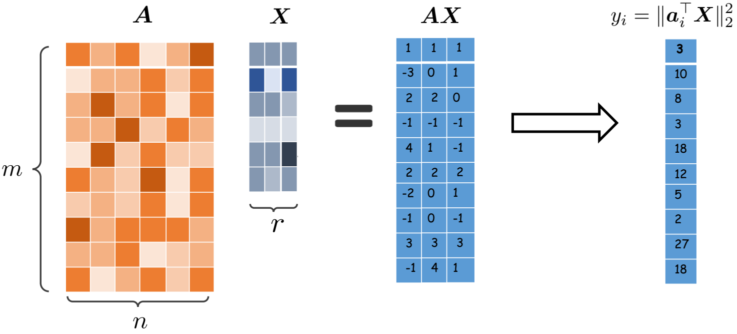
With high prob., any minimizer $\widehat{\boldsymbol{M}}_{\mathsf{cvx}}$ of convex program is nearly rank-r and is minimax near-optimal:

$$\left\|\widehat{\boldsymbol{M}}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}, \quad \left\|\widehat{\boldsymbol{M}}_{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_\infty \lesssim \sigma\sqrt{\frac{n\log n}{p}} \cdot \frac{1}{n}$$

Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization. arXiv:1902.07698.

The phenomenon is quite general

Generalized phase retrieval



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

$$y_i \quad = \quad \left\|\boldsymbol{a}_i^\top \boldsymbol{X}^\natural\right\|_2^2 = \langle \boldsymbol{a}_i\boldsymbol{a}_i^\top, \boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}\rangle, \qquad i = 1,\ldots,m$$

where $\boldsymbol{a}_i$'s are i.i.d. Gaussian entries.

Applications: optical imaging, phase space tomography ...

Implicit regularization for generalized phase retrieval

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

Implicit regularization for generalized phase retrieval

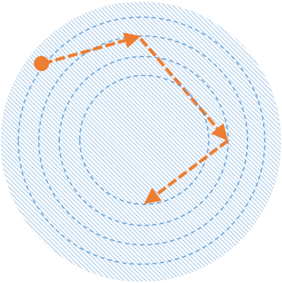
$$\text{minimize}_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \quad f(\boldsymbol{X}) = \frac{1}{4m} \sum_{k=1}^{m} \left(\left\| \boldsymbol{a}_k^\top \boldsymbol{X} \right\|^2 - y_k \right)^2$$

region of local strong convexity + smoothness

Implicit regularization for generalized phase retrieval

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

region of local strong convexity + smoothness

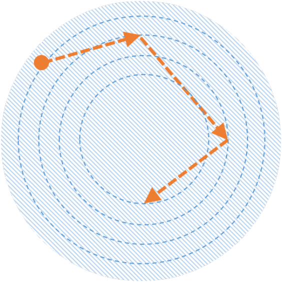


$$O(1) \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(n)$$

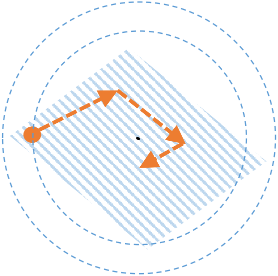
Implicit regularization for generalized phase retrieval

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

region of local strong convexity + smoothness



$O(1) \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(n)$



$O(1) \preceq \nabla^2 f(\boldsymbol{x}) \preceq O(\log n)$

Theoretical guarantees

Theorem (Li, Ma, Chen, Chi, AISTATS 2019)

Under i.i.d. Gaussian design, GD achieves linear convergence

- $\max_k \left\| \boldsymbol{a}_k^\top (\boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\natural) \right\| \lesssim \sqrt{\log n}\, \frac{\sigma_r^2(\boldsymbol{X}^\natural)}{\|\boldsymbol{X}^\natural\|_\mathrm{F}}$ *(incoherence)*

Theoretical guarantees

Theorem (Li, Ma, Chen, Chi, AISTATS 2019)

Under i.i.d. Gaussian design, GD achieves linear convergence

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

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

Theorem (Li, Ma, Chen, Chi, AISTATS 2019)

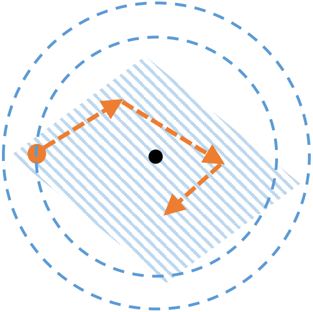
Under i.i.d. Gaussian design, GD achieves linear convergence

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

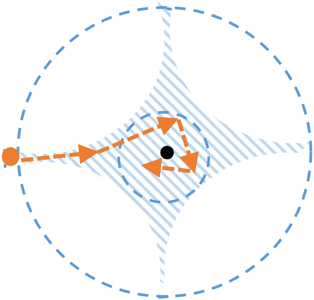
provided that $\eta \asymp \frac{1}{(\log n \vee r)^2 \sigma_r^2(\boldsymbol{X}^\natural)}$ *and* $m \gtrsim nr^4 \log n$.

Big computational saving: GD attains ε-accuracy within $O\big((\log n \vee r)^2 \log \frac{1}{\varepsilon}\big)$ iterations if $m \asymp nr^4 \log n$.

Incoherence region in high dimensions



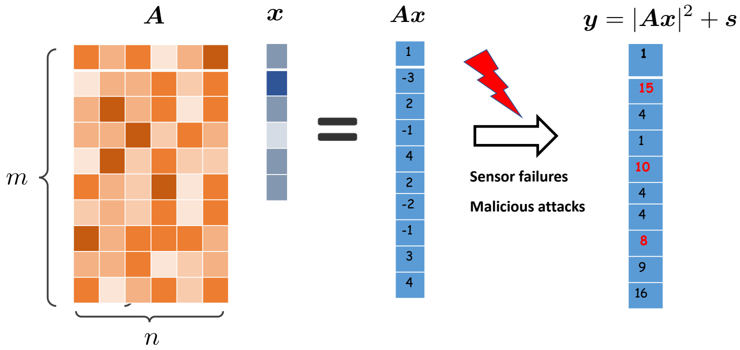
2-dimensional



$\underbrace{\text{high-dimensional}}_{\text{incoherence region is vanishingly small}}$

Towards robust nonconvex statistical estimation

Outlier-corrupted phase retrieval

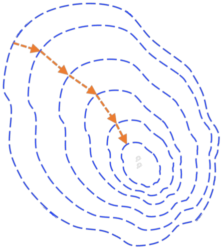


Recover $\boldsymbol{x}^\natural \in \mathbb{R}^n$ from m corrupted measurements

$$y_i \quad = \quad \left| \boldsymbol{a}_i^\top \boldsymbol{x}^\natural \right|_2^2 + s_i, \qquad i = 1, \ldots, m$$

where $\|\boldsymbol{s}\|_0 \leq \alpha \cdot m$, $0 \leq \alpha < 1$ is fraction of outliers.

Existing approaches fail



- **Initialization would fail:** $\boldsymbol{x}^0 \leftarrow$ leading eigenvector of

$$\boldsymbol{Y} = \frac{1}{m}\sum_{i=1}^{m} y_i \boldsymbol{a}_i \boldsymbol{a}_i^\top$$

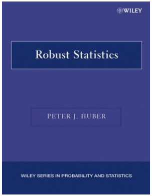
- **Gradient iterations would fail:**

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

for $t = 0, 1, \ldots$

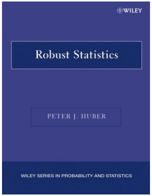
Even a single outlier can fail the algorithm!

Median-truncated gradient descent



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

Median-truncated gradient descent



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

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

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

- **Robustifying gradient descent:**

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \frac{\eta}{m}\sum_{i \in \mathcal{T}_t} \nabla \ell_i(y_i; \boldsymbol{x}^t), \quad t = 0, 1, \ldots$$

where $\mathcal{T}_t = \left\{ i : \ \left| y_i - |\boldsymbol{a}_i^\top \boldsymbol{x}^t| \right| \lesssim \mathsf{median} \left\{ \left| y_i - |\boldsymbol{a}_i^\top \boldsymbol{x}^t| \right| \right\} \right\}$.

Theoretical guarantees

Theorem (Zhang, Chi and Liang, TIT 2019)

Under i.i.d. Gaussian design, median-truncated GD achieves linear convergence

- $\|\boldsymbol{x}^t - \boldsymbol{x}^\natural\|_2 \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\natural\|_2$ *(linear convergence)*

for $\eta \asymp 1$*, provided that* $m \gtrsim n \log n$ *and* $\alpha \lesssim \alpha_0$ *for some constant* α_0*.*

Theoretical guarantees

Theorem (Zhang, Chi and Liang, TIT 2019)

Under i.i.d. Gaussian design, median-truncated GD achieves linear convergence

- $\|\boldsymbol{x}^t - \boldsymbol{x}^\natural\|_2 \lesssim \left(1 - \frac{\eta}{2}\right)^t \|\boldsymbol{x}^\natural\|_2$ *(linear convergence)*

for $\eta \asymp 1$, *provided that* $m \gtrsim n \log n$ *and* $\alpha \lesssim \alpha_0$ *for some constant* α_0.

Add-on robustness: GD attains ε-accuracy within $O\left(\log \frac{1}{\varepsilon}\right)$ iterations if $m \gtrsim n \log n$ *even with a constant fraction of arbitrary outliers.*

Extension to low-rank matrix recovery

Similar idea for compressive low-rank matrix recovery:

$$y_i = \langle \boldsymbol{A}_i, \boldsymbol{X}^\natural \rangle + s_i, \quad i = 1, \ldots, m$$

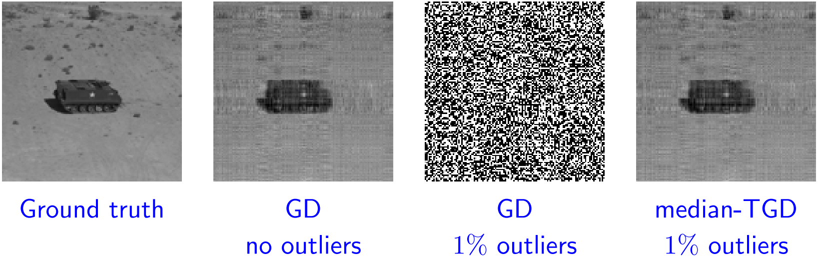
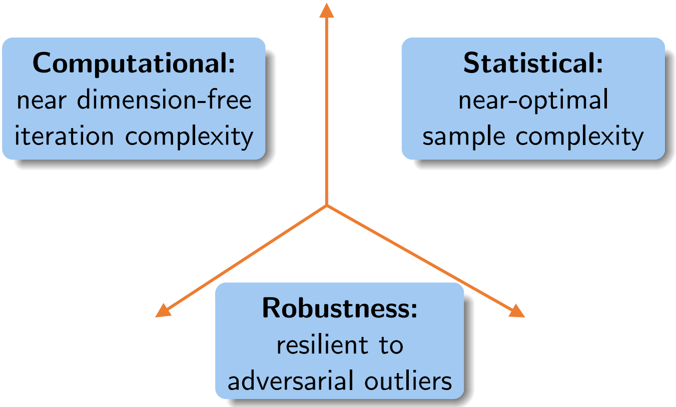


Figure: Recovery performance comparisons for compressive recovery of a 128×128 image from $m = 4600$ measurements and assumed rank $r = 8$.

Li, Chi, Zhang and Liang, "Non-convex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent", Information and Inference: A Journal of the IMA, 2019+.

Final remarks

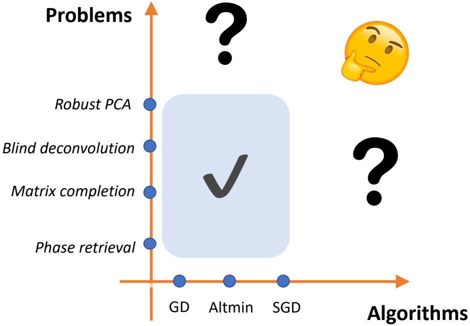
Bridging the theory-practice gap



Fusing statistical thinkings into nonconvex optimization:

- identification and exploitation of benign geometric properties;
- analyzing iterate trajectories beyond black-box optimization.

Limitations



- current analysis is largely case-by-case: lengthy proofs, somewhat similar recipes;
- Is there a unified framework? E.g., RIP for sparsity.
- Can we relax strong randomness assumptions, e.g. Gaussian (phase retrieval), and uniform sampling (matrix completion)?

References

Survey, tutorial articles:

1. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview, **Y. Chi**, Y. M. Lu and Y. Chen, overview article, *IEEE Trans. on Signal Processing*, accepted, arXiv:1809.09573.
2. Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation, Y. Chen and **Y. Chi**, *IEEE Signal Processing Magazine*, 2018.

Selected articles:

1. Implicit Regularization for Nonconvex Statistical Estimation, C. Ma, K. Wang, **Y. Chi** and Y. Chen, *Foundations of Computational Mathematics*, accepted.
2. Nonconvex Matrix Factorization from Rank-One Measurements, Y. Li, C. Ma, Y. Chen, and **Y. Chi**, AISTATS 2019.
3. Non-convex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent, Y. Li, Y. Chi, H. Zhang and Y. Liang, *Information and Inference: A Journal of the IMA*, 2019+.
4. Median-Truncated Nonconvex Approach for Phase Retrieval with Outliers, H. Zhang, **Y. Chi** and Y. Liang, *IEEE Trans. on Information Theory*, 2019.
5. Gradient Descent with Random Initialization: Fast Global Convergence for Nonconvex Phase Retrieval, Y. Chen, **Y. Chi**, J. Fan and C. Ma, *Mathematical Programming*, 2019.

Thanks!

Our research is supported by NSF, ONR and ARO.

https://users.ece.cmu.edu/~yuejiec/