Nonconvex Low-Rank Matrix Estimation: Geometry, Robustness, and Acceleration

Yuejie Chi

Carnegie Mellon University

SIAM Conference on Imaging Science
July 9, 2020

Acknowledgements

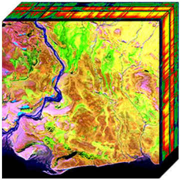
Our research is supported by National Science Foundation, Office of Naval Research and Army Research Office.

Sensing and imaging advances

New imaging/sensing modalities allow us to probe the nature in unprecedented manners.

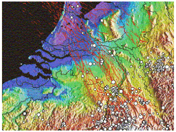
healthcare

Radio astronomy

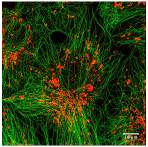


hyperspectral

Internet traffic



seismic imaging



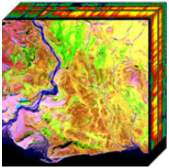
microscopy

The large amount of data brings exciting opportunities that call for new tools that are **scalable in computation and memory**.

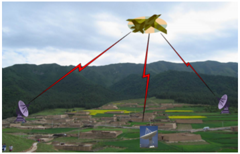
Low-rank matrices in imaging science

Why low-rank images?

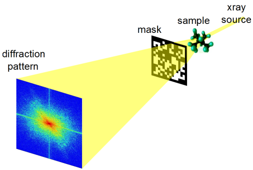
- *redundant representations of latent information;*
- *a small number of sources of interest;*
- *"lifting" of indirect correlation measurements.*



hyperspectral imaging



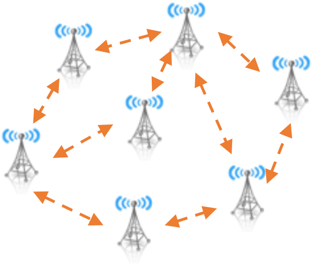
radar imaging



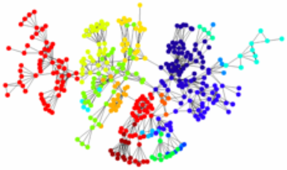
optical imaging

Beyond imaging science

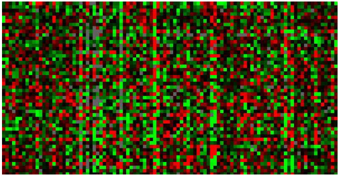
recommendation systems



localization



community detection



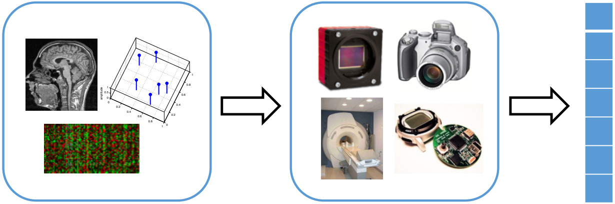
bioinformatics

Low-rank matrix sensing

$\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$
$\mathrm{rank}(\boldsymbol{M}) = r$

$\mathcal{A}(\cdot)$
linear map

$\boldsymbol{y} \in \mathbb{R}^m$



$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{M}) + \text{noise}$$

Recover M in the sample-starved regime:

$$\underbrace{(n_1+n_2)r}_{\text{degree of freedom}} \lesssim \underbrace{m}_{\text{sensing budget}} \ll \underbrace{n_1 n_2}_{\text{ambient dimension}}$$

Convex relaxation via nuclear norm minimization

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n_1\times n_2}} \quad \mathsf{rank}(\boldsymbol{Z}) \qquad\qquad \text{s.t.} \quad \boldsymbol{y}\approx\mathcal{A}(\boldsymbol{Z})$$

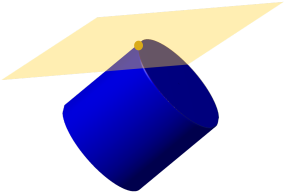
Convex relaxation via nuclear norm minimization

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n_1\times n_2}} \quad \mathsf{rank}(\boldsymbol{Z}) \qquad \text{s.t.} \quad \boldsymbol{y} \approx \mathcal{A}(\boldsymbol{Z})$$

$\Downarrow$ cvx surrogate

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n_1\times n_2}} \quad \|\boldsymbol{Z}\|_*$$
$$\text{s.t.} \quad \boldsymbol{y} \approx \mathcal{A}(\boldsymbol{Z})$$

where $\|\cdot\|_*$ is the nuclear norm.



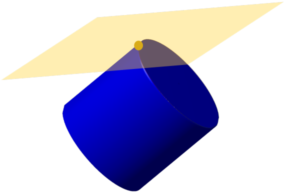
Convex relaxation via nuclear norm minimization

$$\min_{Z \in \mathbb{R}^{n_1 \times n_2}} \quad \text{rank}(Z) \qquad \text{s.t.} \quad y \approx \mathcal{A}(Z)$$

$\Downarrow$ cvx surrogate

$$\min_{Z \in \mathbb{R}^{n_1 \times n_2}} \quad \|Z\|_* \qquad \text{s.t.} \quad y \approx \mathcal{A}(Z)$$

where $\|\cdot\|_*$ is the nuclear norm.



Significant developments in the last decade:

Fazel '02, Recht, Parrilo, Fazel '10, Candès, Recht '09, Candès, Tao '10, Cai et al. '10, Gross '10, Negahban, Wainwright '11, Sanghavi et al. '13, Chen, Chi '14, ...

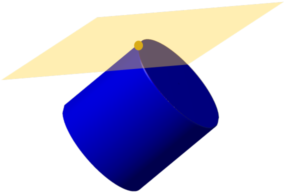
Convex relaxation via nuclear norm minimization

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n_1\times n_2}} \quad \text{rank}(\boldsymbol{Z}) \qquad \text{s.t.} \quad \boldsymbol{y}\approx\mathcal{A}(\boldsymbol{Z})$$

$\Downarrow$ cvx surrogate

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n_1\times n_2}} \quad \|\boldsymbol{Z}\|_*$$
$$\text{s.t.} \quad \boldsymbol{y}\approx\mathcal{A}(\boldsymbol{Z})$$

where $\|\cdot\|_*$ is the nuclear norm.



Significant developments in the last decade:

Fazel '02, Recht, Parrilo, Fazel '10, Candès, Recht '09, Candès, Tao '10, Cai et al. '10, Gross '10, Negahban, Wainwright '11, Sanghavi et al. '13, Chen, Chi '14, ...

Poor scalability: operate in the *ambient* matrix space

Low-rank matrix factorization

$$\min_{\boldsymbol{Z} \in \mathbb{R}^{n_1 \times n_2}} \quad \mathsf{rank}(\boldsymbol{Z}) \qquad \text{s.t.} \quad \boldsymbol{y} \approx \mathcal{A}(\boldsymbol{Z})$$

Low-rank matrix factorization

$$\min_{\boldsymbol{Z} \in \mathbb{R}^{n_1 \times n_2}} \quad \mathsf{rank}(\boldsymbol{Z}) \qquad \mathsf{s.t.} \quad \boldsymbol{y} \approx \mathcal{A}(\boldsymbol{Z})$$

$$\Downarrow$$

$$\min_{\operatorname{rank}(\boldsymbol{Z})=r} \quad \frac{1}{2} \left\| \boldsymbol{y} - \mathcal{A}(\boldsymbol{Z}) \right\|_2^2$$

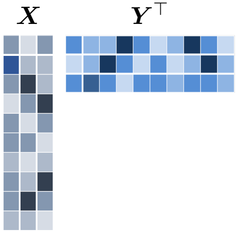
Low-rank matrix factorization

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n_1\times n_2}} \ \mathsf{rank}(\boldsymbol{Z}) \qquad \text{s.t.} \quad \boldsymbol{y}\approx\mathcal{A}(\boldsymbol{Z})$$

$$\Downarrow$$

$$\min_{\color{blue}{\mathrm{rank}(\boldsymbol{Z})=r}} \ \frac{1}{2}\|\boldsymbol{y}-\mathcal{A}(\boldsymbol{Z})\|_2^2$$

$$\Downarrow \qquad \boldsymbol{Z} =$$



$$\min_{\boldsymbol{X}\in\mathbb{R}^{n_1\times r},\boldsymbol{Y}\in\mathbb{R}^{n_2\times r}} \ f(\boldsymbol{X},\boldsymbol{Y})=\frac{1}{2}\left\|\boldsymbol{y}-\mathcal{A}(\boldsymbol{X}\boldsymbol{Y}^\top)\right\|_2^2$$

Low-rank matrix factorization

$$\min_{\boldsymbol{Z}\in\mathbb{R}^{n_1\times n_2}} \quad \text{rank}(\boldsymbol{Z}) \qquad \text{s.t.} \quad \boldsymbol{y}\approx\mathcal{A}(\boldsymbol{Z})$$

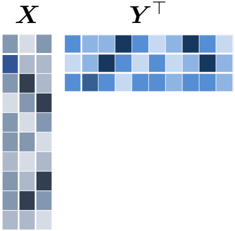
$$\Downarrow$$

$$\min_{\color{blue}{\text{rank}(\boldsymbol{Z})=r}} \quad \frac{1}{2}\left\|\boldsymbol{y}-\mathcal{A}(\boldsymbol{Z})\right\|_2^2$$

more scalable,
but nonconvex!

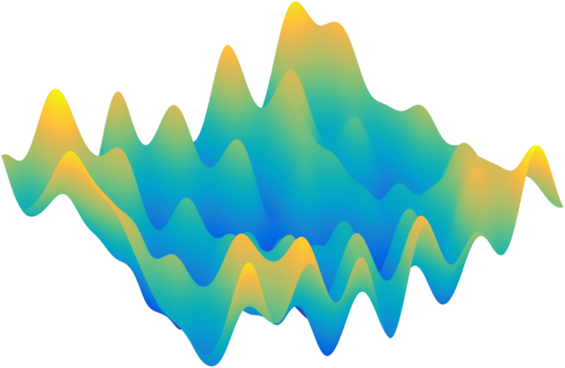
$$\Downarrow$$

$\boldsymbol{Z} =$



$$\min_{\boldsymbol{X}\in\mathbb{R}^{n_1\times r},\boldsymbol{Y}\in\mathbb{R}^{n_2\times r}} \quad f(\boldsymbol{X},\boldsymbol{Y})=\frac{1}{2}\left\|\boldsymbol{y}-\mathcal{A}(\boldsymbol{X}\boldsymbol{Y}^\top)\right\|_2^2$$

Nonconvex problems are hard (in theory)!



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

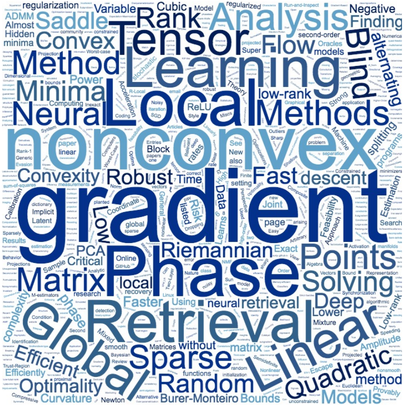
R. T. Rockafellar, in SIAM Review, 1993

Nonconvex problems are hard (in theory)!

"...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



"Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview," Chi, Lu, Chen TSP 2019

Phase retrieval: 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, Shi, Chi '19, Qu et al. '19...

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, Qu 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...

This talk: geometry, robustness, acceleration

Optimization geometry:

When and why does simple gradient descent work well for low-rank matrix estimation?

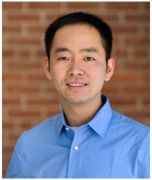
Robustness to adversarial outliers:

Can we design provably robust gradient algorithms that are oblivious to the presence of outliers?

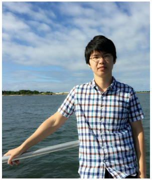
Acceleration for ill-conditioned matrix estimation:

Can we design provably fast gradient algorithms that are insensitive to the condition number of low-rank matrices?

Geometry and implicit regularization in nonconvex low-rank matrix estimation



Yuxin Chen
Princeton



Cong Ma
Princeton

Kaizheng Wang
Princeton

Yuanxin Li
CMU

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 $\boldsymbol{M} = \boldsymbol{X}_\star \boldsymbol{X}_\star^\top \in \mathbb{R}^{n \times n}$ in an index set Ω, fill in missing entries.

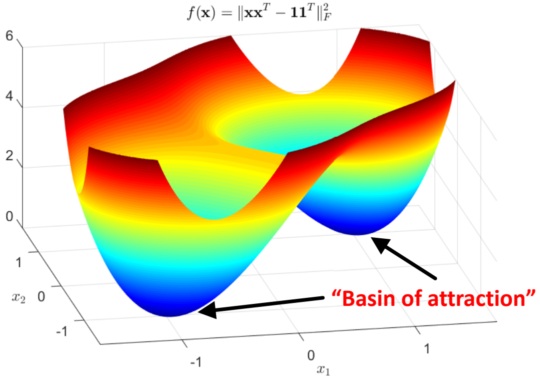
$$\min_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} f(\boldsymbol{X}) = \frac{1}{2} \left\| \mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M}) \right\|_\mathrm{F}^2$$

What might the loss function look like?

Full observation = PCA: $f(\boldsymbol{X}) = \left\| \boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{M} \right\|_\mathrm{F}^2.$

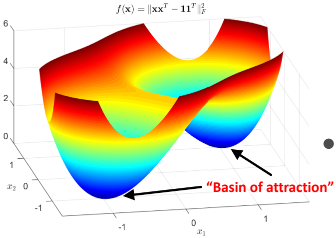
$f(\boldsymbol{X})$ restricted strongly convex and smooth

along descent direction $\boldsymbol{V}$ when $\boldsymbol{X}$ is close to $\boldsymbol{X}_\star$.



Parameter recovery via gradient descent (GD)

a two-step recovery strategy:



- **Spectral initialization:** find an initial point in the "basin of attraction".

$$\boldsymbol{X}_0 = \mathsf{SVD}_r(\mathcal{P}_\Omega(\boldsymbol{M}))$$

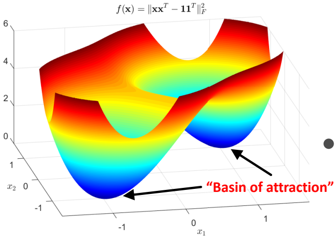
- **Gradient iterations:**

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

for $t = 0, 1, \ldots$

Parameter recovery via gradient descent (GD)

a two-step recovery strategy:



- **Spectral initialization:** find an initial point in the "basin of attraction".

$$\boldsymbol{X}_0 = \mathsf{SVD}_r(\mathcal{P}_\Omega(\boldsymbol{M}))$$

- **Gradient iterations:**

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

for $t = 0, 1, \ldots$

Question: Does vanilla GD still work with partial observations?

Which region has benign geometry?

Finite-sample level $\left(p \asymp \frac{\text{polylog} n}{n}\right)$ **:** *assume every entry is observed i.i.d. with probability* $0 < p \leq 1$.

Question: which matrix is easier to complete?

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

Which region has benign geometry?

Finite-sample level $\left(p \asymp \frac{\mathrm{poly}\log n}{n}\right)$ **:** *assume every entry is observed i.i.d. with probability* $0 < p \leq 1$.

Question: which matrix is easier to complete?

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

Low-rank matrix completion is only well-defined for "incoherent" matrices whose energies are spread evenly across the entries.

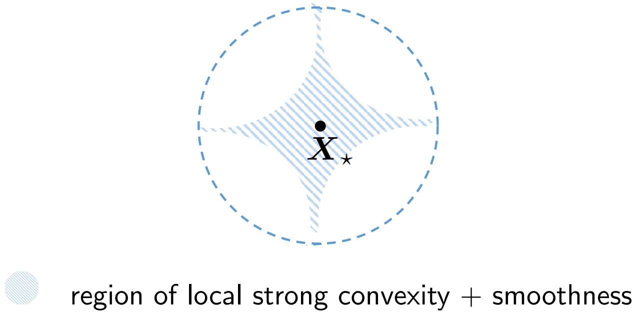
Which region has benign geometry?

Finite-sample level ($p \asymp \frac{\mathsf{polylog} n}{n}$) : *assume every entry is observed i.i.d. with probability* $0 < p \leq 1$.

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

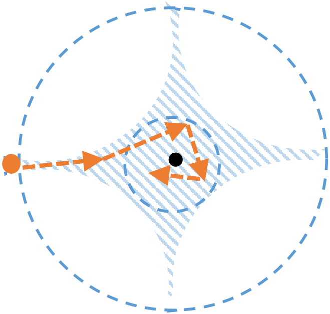
along descent direction $\boldsymbol{V}$ only when $\boldsymbol{X}$ is incoherent:

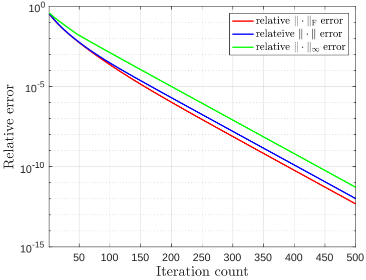
$$\|\boldsymbol{X} - \boldsymbol{X}_\star\|_{2,\infty} \ll \|\boldsymbol{X}_\star\|_{2,\infty}$$



Our findings: gradient descent is implicitly regularized

region of local strong convexity + smoothness





Gradient descent implicitly forces iterates to remain incoherent even without regularization

Theoretical guarantees - noise-free case

Theorem (Ma, Wang, Chi, Chen, FoCM 2020)

Suppose $\boldsymbol{M} = \boldsymbol{X}_\star \boldsymbol{X}_\star^\top$ is rank-r, μ-incoherent and has a condition number $\kappa = \sigma_{\max}(\boldsymbol{M})/\sigma_{\min}(\boldsymbol{M})$. Vanilla GD (with spectral initialization) achieves

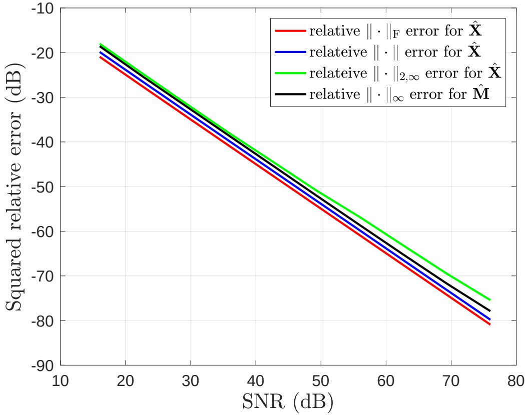
$$\|\boldsymbol{X}_t \boldsymbol{X}_t^\top - \boldsymbol{M}\|_\mathrm{F} \leq \varepsilon \cdot \sigma_{\min}(\boldsymbol{M})$$

- **Computational:** *within $O\big(\kappa \log \frac{1}{\varepsilon}\big)$ iterations;*
- **Statistical:** *as long as the sample complexity satisfies*

$$n^2 p \gtrsim n r^3 \mathsf{poly}(\mu, \kappa, \log n).$$

Key idea: the iterates are implicitly regularized

Noisy matrix completion via vanilla GD



Near-optimal entry-wise error control:

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

The phenomenon is quite general

	Prior theory		Our theory	
	sample complexity	iteration complexity	sample complexity	iteration complexity
Phase retrieval	$n \log n$	$n \log\left(\frac{1}{\varepsilon}\right)$	$n \log n$	$\log n \log\left(\frac{1}{\varepsilon}\right)$
Quadratic sensing	$nr^6 \log^2 n$	$n^4 r^2 \log\left(\frac{1}{\varepsilon}\right)$	$nr^4 \log n$	$r^2 \log\left(\frac{1}{\varepsilon}\right)$
Matrix completion	n/a	n/a	$nr^3 \text{poly} \log n$	$\log\left(\frac{1}{\varepsilon}\right)$
Blind deconvolution	n/a	n/a	$K \text{poly} \log m$	$\log\left(\frac{1}{\varepsilon}\right)$

Huge computational savings!

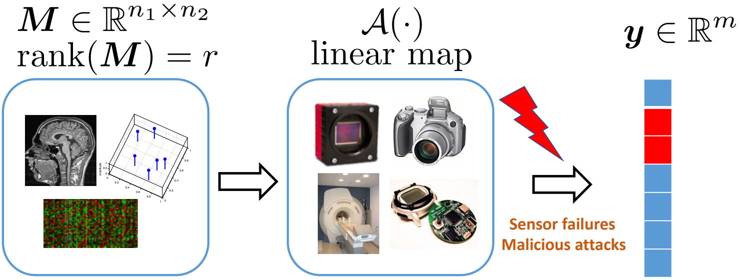
Towards robustness to adversarial outliers

Yuanxin Li
CMU

Yingbin Liang
OSU

Huishuai Zhang
MSRA

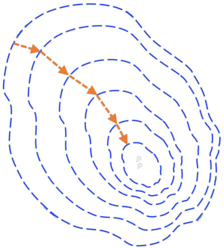
Outlier-corrupted low-rank matrix sensing



$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{M}) + \underbrace{\boldsymbol{s}}_{\text{outliers}}, \quad \mathcal{A}(\boldsymbol{M}) = \{\langle \boldsymbol{A}_i, \boldsymbol{M} \rangle\}_{i=1}^m$$

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

Existing approaches fail



- **Spectral initialization would fail:**
  $\boldsymbol{X}_0 \leftarrow$ top-r SVD of

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

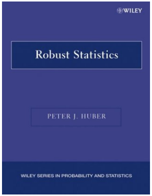
- **Gradient iterations would fail:**

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

for $t = 0, 1, \ldots$

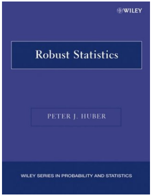
Even a single outlier can fail the algorithm!

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

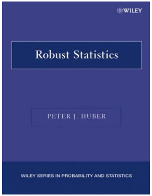


Key idea: "median-truncation" — discard samples *adaptively* based on how large sample gradients / values deviate from median

- **Robustify spectral initialization:** $\boldsymbol{X}_0 \leftarrow$ top-r SVD of

$$\boldsymbol{Y} = \frac{1}{m} \sum_{i:|y_i| \lesssim \mathsf{median}(|y_i|)} y_i \boldsymbol{A}_i$$

Median-truncated gradient descent



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

- **Robustify spectral initialization:** $\boldsymbol{X}_0 \leftarrow$ top-r SVD of

$$\boldsymbol{Y} = \frac{1}{m} \sum_{i:|y_i| \lesssim \text{median}(|y_i|)} y_i \boldsymbol{A}_i$$

- **Robustify gradient descent:**

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \frac{\eta}{m} \sum_{i:|r_t^i| \lesssim \text{median}(|r_t^i|)} \nabla \ell_i(y_i; \boldsymbol{X}_t), \quad t = 0, 1, \ldots$$

where $r_t^i := |y_i - \langle \boldsymbol{A}_i, \boldsymbol{X}_t \rangle|$ is the size of the gradient.

Theoretical guarantees

Theorem (Li, Chi, Zhang, and Liang, IMIAI 2020)

For low-rank matrix sensing with i.i.d. Gaussian design, median-truncated GD (with robust spectral initialization) achieves

$$\|\boldsymbol{X}_t \boldsymbol{X}_t^\top - \boldsymbol{M}\|_{\mathrm{F}} \leq \varepsilon \cdot \sigma_{\min}(\boldsymbol{M}),$$

- **Computational:** *within* $O\left(\kappa \log \frac{1}{\varepsilon}\right)$ *iterations;*
- **Statistical:** *the sample complexity satisfies*

$$m \gtrsim nr^2 \mathsf{poly}(\kappa, \log n);$$

- **Robustness:** *and the fraction of outliers*

$$\alpha \lesssim 1/\sqrt{r}.$$

Theoretical guarantees

Theorem (Li, Chi, Zhang, and Liang, IMIAI 2020)

For low-rank matrix sensing with i.i.d. Gaussian design, median-truncated GD (with robust spectral initialization) achieves

$$\|\boldsymbol{X}_t \boldsymbol{X}_t^\top - \boldsymbol{M}\|_{\mathrm{F}} \le \varepsilon \cdot \sigma_{\min}(\boldsymbol{M}),$$

- **Computational:** *within* $O\left(\kappa \log \frac{1}{\varepsilon}\right)$ *iterations;*
- **Statistical:** *the sample complexity satisfies*

$$m \gtrsim nr^2 \mathsf{poly}(\kappa, \log n);$$

- **Robustness:** *and the fraction of outliers*

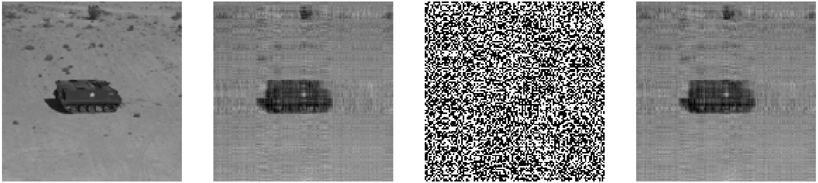
$$\alpha \lesssim 1/\sqrt{r}.$$

Median-truncated GD adds robustness to GD *obliviously*.

Numerical example

Low-rank matrix sensing:

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



Ground truth | GD no outliers | GD 1% outliers | median-TGD 1% outliers

Median-truncated GD achieves similar performance as if performing GD on the clean data.

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, 2020.

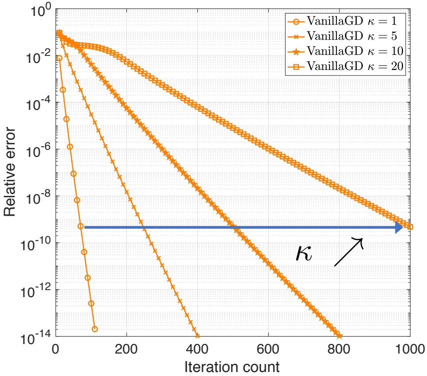
Accelerating ill-conditioned matrix estimation

Tian Tong
CMU

Cong Ma
Princeton

Convergence slows down for ill-conditioned matrices

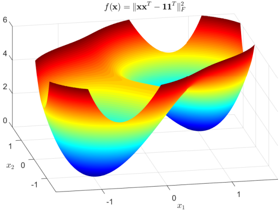
$$\min_{\boldsymbol{X},\boldsymbol{Y}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{2}\left\|\mathcal{P}_\Omega(\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M})\right\|_\mathrm{F}^2$$



Vanilla GD converges in $O\left(\kappa \log \frac{1}{\varepsilon}\right)$ iterations.

— *Can we accelerate the convergence to* $O\left(\log \frac{1}{\varepsilon}\right)$*?*

A new algorithm: scaled gradient descent (ScaledGD)

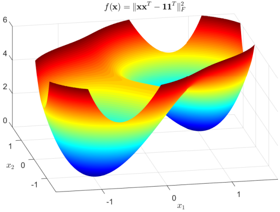


- **Spectral initialization.**
- **Scaled gradient iterations:**

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta \nabla f(\boldsymbol{X}_t) \underbrace{(\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1}}_{\texttt{preconditioner}}$$

for $t = 0, 1, \ldots$

A new algorithm: scaled gradient descent (ScaledGD)



- **Spectral initialization.**
- **Scaled gradient iterations:**

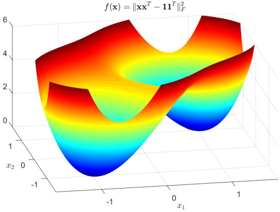
$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta \nabla f(\boldsymbol{X}_t) \underbrace{(\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1}}_{\texttt{preconditioner}}$$

for $t = 0, 1, \ldots$

For the asymmetric case:

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}_t, \boldsymbol{Y}_t)(\boldsymbol{Y}_t^\top \boldsymbol{Y}_t)^{-1}$$
$$\boldsymbol{Y}_{t+1} = \boldsymbol{Y}_t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}_t, \boldsymbol{Y}_t)(\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1}$$

A new algorithm: scaled gradient descent (ScaledGD)



- **Spectral initialization.**
- **Scaled gradient iterations:**

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \eta \nabla f(\boldsymbol{X}_t) \underbrace{(\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1}}_{\texttt{preconditioner}}$$

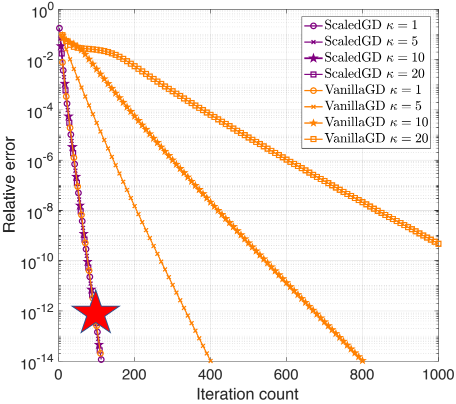
for $t = 0, 1, \ldots$

For the asymmetric case:

$$\begin{aligned} \boldsymbol{X}_{t+1} &= \boldsymbol{X}_t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}_t, \boldsymbol{Y}_t)(\boldsymbol{Y}_t^\top \boldsymbol{Y}_t)^{-1} \\ \boldsymbol{Y}_{t+1} &= \boldsymbol{Y}_t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}_t, \boldsymbol{Y}_t)(\boldsymbol{X}_t^\top \boldsymbol{X}_t)^{-1} \end{aligned}$$

ScaledGD is a *preconditioned* gradient method.

ScaledGD for low-rank matrix completion



Huge computational saving: ScaledGD converges in an κ-independent manner with a minimal overhead!

Theoretical guarantees of ScaledGD

Theorem (Tong, Ma and Chi, 2020)

For low-rank matrix sensing with i.i.d. Gaussian design, ScaledGD with spectral initialization achieves

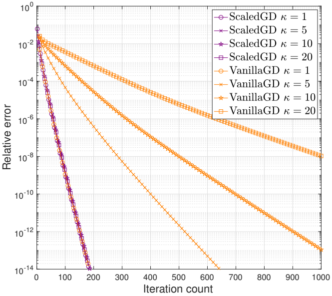
$$\|\boldsymbol{X}_t \boldsymbol{Y}_t^\top - \boldsymbol{M}\|_{\mathrm{F}} \lesssim \varepsilon \cdot \sigma_{\min}(\boldsymbol{M})$$

- **Computational:** *within* $O\big(\log \frac{1}{\varepsilon}\big)$ *iterations;*
- **Statistical:** *the sample complexity satisfies*

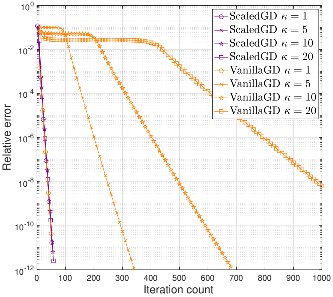
$$m \gtrsim nr^2\kappa^2.$$

Acceleration for ill-conditioning: ScaledGD provably accelerates vanilla GD for low-rank matrix sensing.

ScaledGD works more broadly



Robust PCA



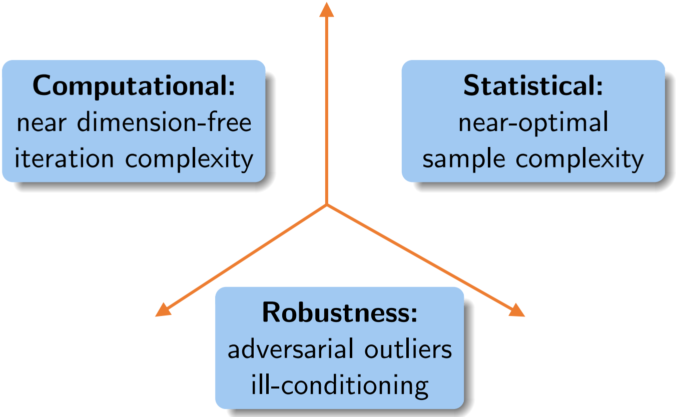
Hankel matrix completion

ScaledGD is more efficient when the low-rank matrix is ill-conditioned.

Code available at `https://github.com/Titan-Tong/ScaledGD`

Final remarks

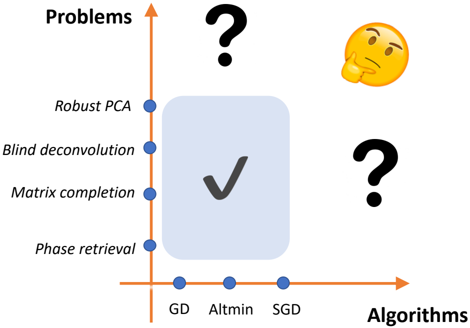
Bridging the theory-practice gap



Nonconvex low-rank matrix estimation:

- identification and exploitation of benign geometric properties;
- analyzing iterate trajectories beyond black-box optimization;
- simple variants of GD lead to robust and accelerated convergence.

Future directions



Limitations of current framework:

- largely case-by-case: lengthy proofs, somewhat similar recipes;
- somewhat strong assumptions, e.g. Gaussian measurements, uniformly sampling...

References

Overview:

1. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview, Y. Chi, Y. M. Lu and Y. Chen, *IEEE Trans. on Signal Processing*, 2019.
2. Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation, Y. Chen and Y. Chi, *IEEE Signal Processing Magazine*, 2018.

Geometry and implicit regularization:

1. Implicit Regularization for Nonconvex Statistical Estimation, C. Ma, K. Wang, Y. Chi and Y. Chen, *Foundations of Computational Mathematics*, 2020.
2. Nonconvex Matrix Factorization from Rank-One Measurements, Y. Li, C. Ma, Y. Chen, and Y. Chi, AISTATS 2019.

Accelerating ill-conditioned low-rank matrix estimation:

1. Accelerating Ill-Conditioned Low-Rank Matrix Estimation via Scaled Gradient Descent, T. Tong, C. Ma, and Y. Chi, preprint, 2020.

Robustness to adversarial outliers:

1. 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*, 2020.
2. Median-Truncated Nonconvex Approach for Phase Retrieval with Outliers, H. Zhang, Y. Chi and Y. Liang, *IEEE Trans. on Information Theory*, 2019.

Thanks!

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