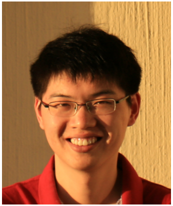
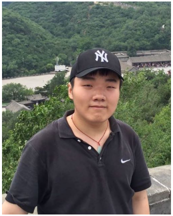
Bridging convex and nonconvex optimization in noisy matrix completion: Stability and uncertainty quantification

Yuxin Chen

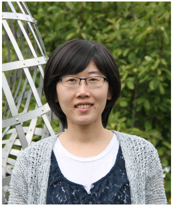
EE, Princeton University



Cong Ma
Princeton ORFE



Yuling Yan
Princeton ORFE



Yuejie Chi
CMU ECE

Jianqing Fan
Princeton ORFE

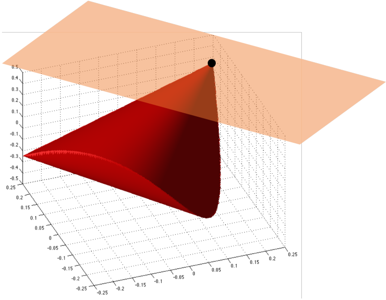
Convex relaxation for low-rank structure

$$\underset{\boldsymbol{Z}}{\text{minimize}} \quad \|\boldsymbol{Z}\|_* \triangleq \sum_i \sigma_i(\boldsymbol{Z})$$

$$\text{subj. to} \quad \text{noiseless data constraints}$$

low-rank matrix

Composition C by Piet Mondrian



semidefinite relaxation

Convex relaxation for low-rank structure

$$\begin{array}{ll} \underset{\boldsymbol{Z}}{\text{minimize}} & \|\boldsymbol{Z}\|_* \triangleq \sum_i \sigma_i(\boldsymbol{Z}) \\ \text{subj. to} & \text{noiseless data constraints} \end{array}$$

- ✓ matrix sensing (Recht, Fazel, Parrilo '07)
- ✓ phase retrieval (Candès, Strohmer, Voroninski '11, Candès, Li '12)
- ✓ matrix completion (Candès, Recht '08, Candès, Tao '08, Gross '09)
- ✓ robust PCA (Chandrasekaran et al. '09, Candès et al. '09)
- ✓ Hankel matrix completion (Fazel et al. '13, Chen, Chi '13, Cai et al. '15)
- ✓ blind deconvolution (Ahmed, Recht, Romberg '12, Ling, Strohmer '15)
- ✓ joint alignment / matching (Chen, Huang, Guibas '14)

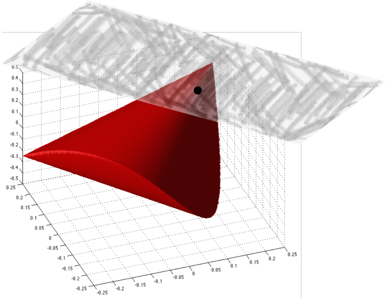
...

Stability of convex relaxation against noise

$$\begin{array}{ll} \underset{Z}{\text{minimize}} & \|Z\|_* \\ \text{subj. to} & \text{noisy data constraints} \end{array}$$

low-rank matrix

Composition C by Piet Mondrian



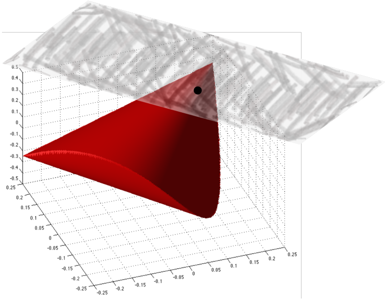
semidefinite relaxation

Stability of convex relaxation against noise

$$\underset{Z}{\text{minimize}} \qquad \underbrace{f(Z;\text{data})}_{\text{empirical loss}} + \lambda \|Z\|_*$$

low-rank matrix

Composition C by Piet Mondrian



semidefinite relaxation

Stability of convex relaxation against noise

$$\underset{Z}{\text{minimize}} \qquad \underbrace{f(Z; \text{data})}_{\text{empirical loss}} + \lambda \|Z\|_*$$

- ✓ matrix sensing (RIP measurements) (Candès, Plan '10)
- ✓ phase retrieval (Gaussian measurements) (Candès et al. '11)
- ? matrix completion (Candès, Plan '09, Negahban, Wainwright '10, Koltchinskii et al. '10)
- ? robust PCA (Zhou, Li, Wright, Candès, Ma '10)
- ? Hankel matrix completion (Chen, Chi '13)
- ? blind deconvolution (Ahmed, Recht, Romberg '12, Ling, Strohmer '15)
- ? joint alignment / matching

...

Stability of convex relaxation against noise

$$\underset{\boldsymbol{Z}}{\text{minimize}} \qquad \underbrace{f(\boldsymbol{Z};\text{data})}_{\text{empirical loss}} + \lambda \|\boldsymbol{Z}\|_*$$

- ✓ matrix sensing (RIP measurements) (Candès, Plan ’10)
- ✓ phase retrieval (Gaussian measurements) (Candès et al. ’11)
- ? **this talk: matrix completion**
  (Candès, Plan ’09, Negahban, Wainwright ’10, Koltchinskii et al. ’10)
- ? robust PCA (Zhou, Li, Wright, Candès, Ma ’10)
- ? Hankel matrix completion (Chen, Chi ’13)
- ? blind deconvolution (Ahmed, Recht, Romberg ’12, Ling, Strohmer ’15)
- ? joint alignment / matching

...

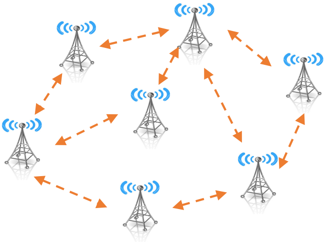
Low-rank matrix completion

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

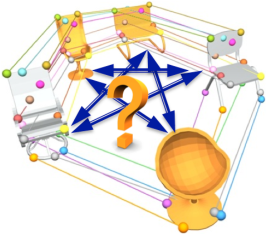
figure credit: E. J. Candès

Given partial samples of a low-rank matrix $M^{\star}$, fill in missing entries

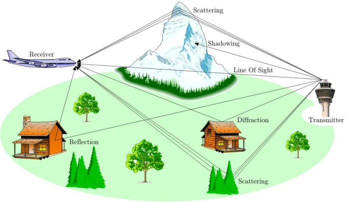
recommendation systems



localization



shape matching



channel estimation

Noisy low-rank matrix completion

observations: $M_{i,j} = M^\star_{i,j} + \text{noise}, \quad (i,j) \in \Omega$

goal: estimate $\boldsymbol{M}^\star$

unknown rank-r matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$

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

sampling set Ω

Noisy low-rank matrix completion

observations: $\quad M_{i,j} = M^{\star}_{i,j} + \text{noise}, \quad (i,j) \in \Omega$

goal: $\quad$ estimate $\boldsymbol{M}^{\star}$

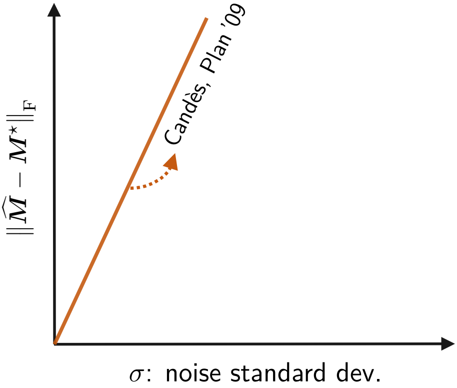
convex relaxation:

$$\operatorname*{minimize}_{\boldsymbol{Z} \in \mathbb{R}^{n \times n}} \quad \underbrace{\sum_{(i,j) \in \Omega} \left(Z_{i,j} - M_{i,j}\right)^2}_{\text{squared loss}} + \lambda \|\boldsymbol{Z}\|_*$$

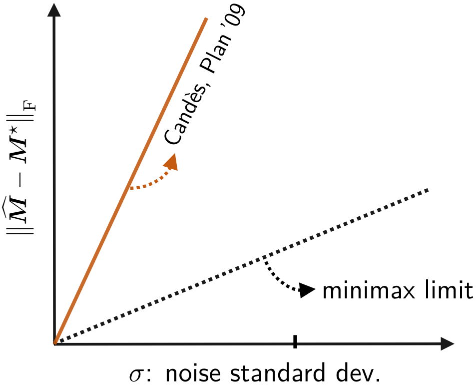
Prior statistical guarantees for convex relaxation

- **random sampling:** each $(i, j) \in \Omega$ with prob. p
- **random noise:** i.i.d. sub-Gaussian noise with variance σ^2
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: rank $r = O(1)$, incoherent, $\ldots$

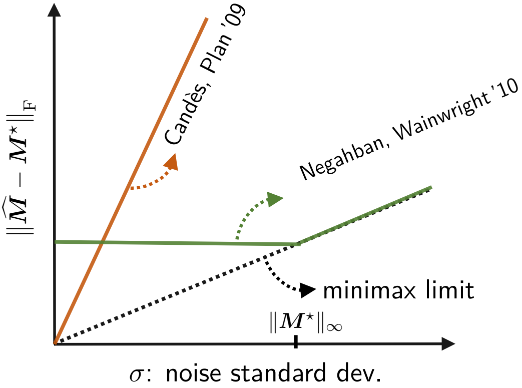
Candès, Plan '09	$\sigma n^{1.5}$



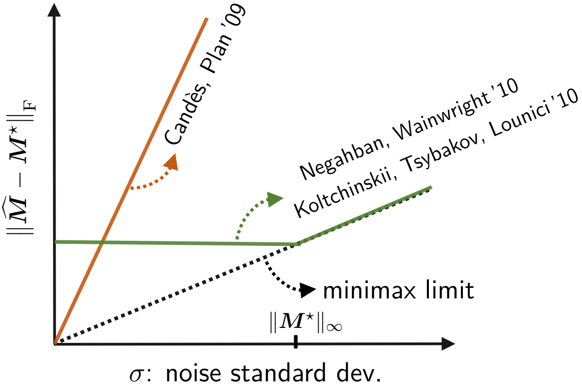
minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$



minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$
Negahban, Wainwright '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$

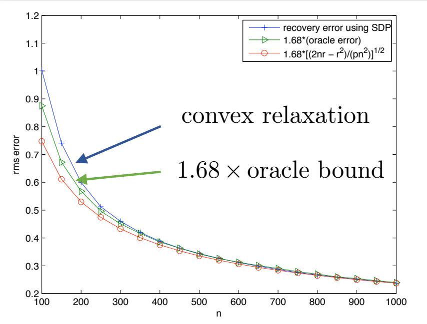


minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$
Negahban, Wainwright '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$
Koltchinskii, Tsybakov, Lounici '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$



Matrix Completion with Noise

Emmanuel J. Candès and Yaniv Plan



Existing theory for convex relaxation does not match practice . . .

Matrix Completion with Noise

Emmanuel J. Candès and Yaniv Plan

with adversarial noise. Consequently, our analysis looses a $\sqrt{n}$ factor vis a vis an optimal bound that is achievable via the help of an oracle.

Existing theory for convex relaxation does not match practice . . .

What are the roadblocks?

Strategy: $\boldsymbol{M}^{\mathsf{cvx}}$ is optimizer if $\underbrace{\text{there exists } \boldsymbol{W}}_{\text{dual certificate}}$ s.t.

$(\boldsymbol{M}^{\mathsf{cvx}}, \boldsymbol{W})$ obeys KKT optimality condition

What are the roadblocks?

Strategy: $\boldsymbol{M}^{\mathsf{cvx}}$ is optimizer if $\underbrace{\text{there exists } \boldsymbol{W}}_{\text{dual certificate}}$ s.t.

$(\boldsymbol{M}^{\mathsf{cvx}}, \boldsymbol{W})$ obeys KKT optimality condition

David Gross

- **noiseless case**: $\underbrace{\boldsymbol{M}^{\mathsf{cvx}} \leftarrow \boldsymbol{M}^{\star}}_{\text{exact recovery}}$; $\boldsymbol{W} \leftarrow$ golfing scheme

What are the roadblocks?

Strategy: $\boldsymbol{M}^{\mathsf{cvx}}$ is optimizer if $\underbrace{\text{there exists } \boldsymbol{W}}_{\text{dual certificate}}$ s.t.

$(\boldsymbol{M}^{\mathsf{cvx}}, \boldsymbol{W})$ obeys KKT optimality condition

David Gross

- **noiseless case**: $\underbrace{\boldsymbol{M}^{\mathsf{cvx}} \leftarrow \boldsymbol{M}^{\star}}_{\text{exact recovery}}$; $\boldsymbol{W} \leftarrow$ golfing scheme
- **noisy case**: $\boldsymbol{M}^{\mathsf{cvx}}$ is very complicated, hard to construct $\boldsymbol{W}$ ...

dual certification (golfing scheme)

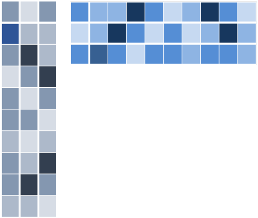
dual certification (golfing scheme)

nonconvex optimization

A detour: nonconvex optimization

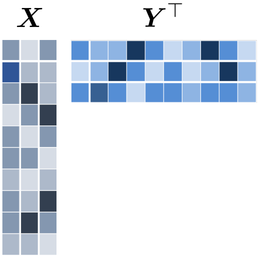
Burer–Monteiro: represent $\boldsymbol{Z}$ by $\boldsymbol{X}\boldsymbol{Y}^\top$ with $\underbrace{\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}}_{\text{low-rank factors}}$

$\boldsymbol{X}$ $\quad$ $\boldsymbol{Y}^\top$



A detour: nonconvex optimization

Burer–Monteiro: represent $\boldsymbol{Z}$ by $\boldsymbol{X}\boldsymbol{Y}^\top$ with $\underbrace{\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}}_{\text{low-rank factors}}$



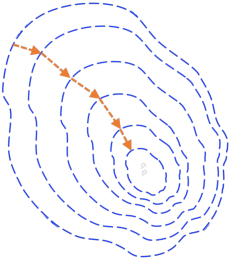
$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \underbrace{\sum_{(i,j)\in\Omega} \Big[\big(\boldsymbol{X}\boldsymbol{Y}^\top\big)_{i,j} - M_{i,j} \Big]^2}_{\text{squared loss}} + \text{reg}(\boldsymbol{X},\boldsymbol{Y})$$

A detour: nonconvex optimization

- Burer, Monteiro '03
- Rennie, Srebro '05
- Keshavan, Montanari, Oh '09 '10
- Jain, Netrapalli, Sanghavi '12
- Hardt '13
- Sun, Luo '14
- Chen, Wainwright '15
- Tu, Boczar, Simchowitz, Soltanolkotabi, Recht '15
- Zhao, Wang, Liu '15
- Zheng, Lafferty '16
- Yi, Park, Chen, Caramanis '16
- Ge, Lee, Ma '16
- Ge, Jin, Zheng '17
- Ma, Wang, Chi, Chen '17
- Chen, Li '18
- Chen, Liu, Li '19
- ...

A detour: nonconvex optimization

$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \sum_{(i,j)\in\Omega}\left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{i,j} - M_{i,j}\right]^2 + \text{reg}(\boldsymbol{X},\boldsymbol{Y})$$



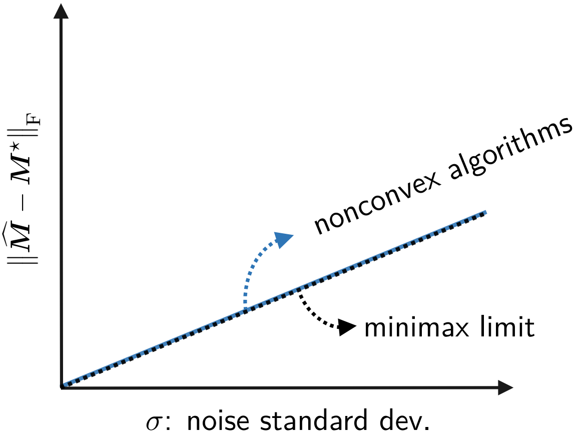
- **suitable initialization:** $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$
- **gradient descent:** for $t = 0, 1, \ldots$

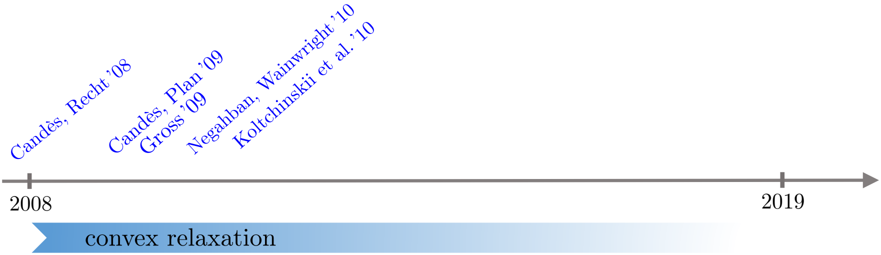
$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$
$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta_t \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$

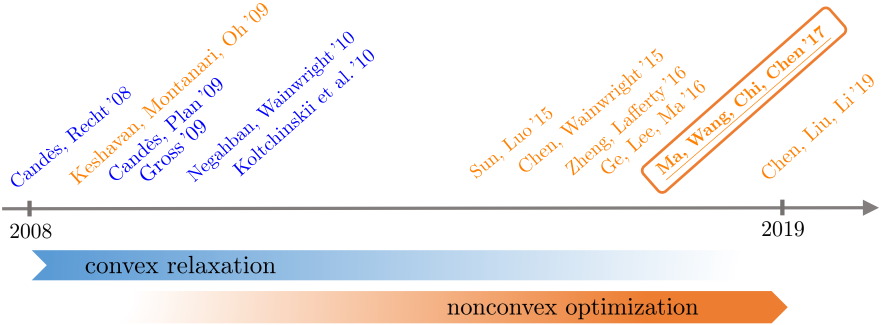
A detour: nonconvex optimization

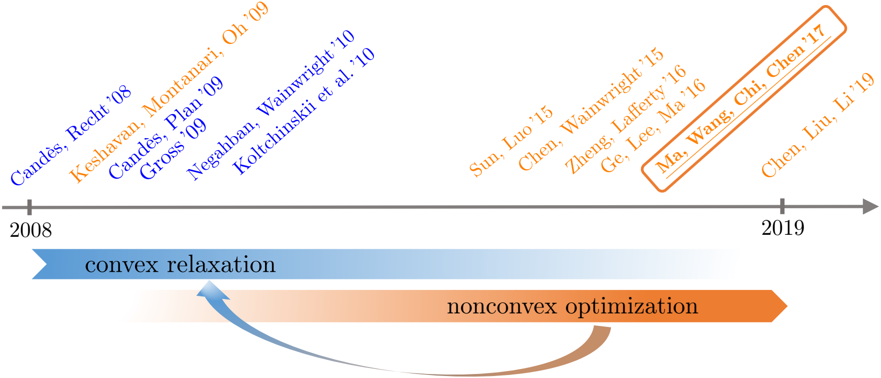
- **random sampling:** each $(i, j) \in \Omega$ with prob. p
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, ...

minimax limit	$\sigma\sqrt{n/p}$
nonconvex algorithms	$\sigma\sqrt{n/p}$ (optimal!)




Candès, Recht '08
Candès, Plan '09
Gross '09
Negahban, Wainwright '10
Koltchinskii et al. '10
2008
2019
convex relaxation


Candès, Recht '08
Keshavan, Montanari, Oh '09
Candès, Plan '09
Gross '09
Negahban, Wainwright '10
Koltchinskii et al. '10
Sun, Luo '15
Chen, Wainwright '15
Zheng, Lafferty '16
Ge, Lee, Ma '16
Ma, Wang, Chi, Chen '17
Chen, Liu, Li '19
2008
2019
convex relaxation
nonconvex optimization



Candès, Recht '08
Keshavan, Montanari, Oh '09
Candès, Plan '09
Gross '09
Negahban, Wainwright '10
Koltchinskii et al. '10
Sun, Luo '15
Chen, Wainwright '15
Zheng, Lafferty '16
Ge, Lee, Ma '16
Ma, Wang, Chi, Chen '17
Chen, Liu, Li '19
2008
2019
convex relaxation
nonconvex optimization

A motivating experiment

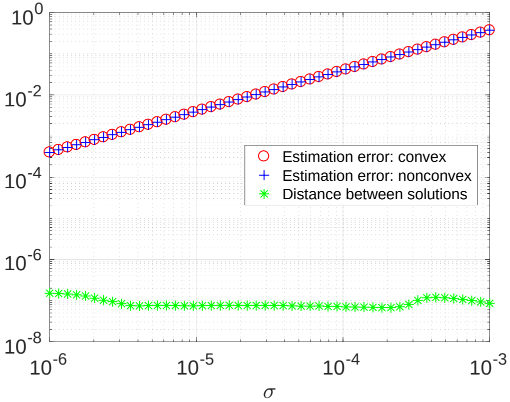
convex: $$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_*$$

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

— $\|\boldsymbol{Z}\|_* = \min_{\boldsymbol{Z}=\boldsymbol{X}\boldsymbol{Y}^\top} \frac{1}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \frac{1}{2}\|\boldsymbol{Y}\|_{\mathrm{F}}^2$

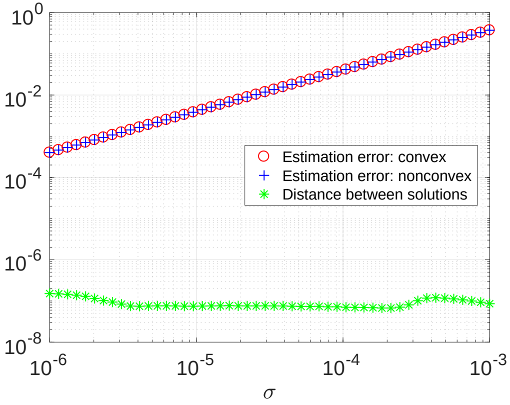
A motivating experiment

$$n = 1000,\ r = 5,\ p = 0.2,\ \lambda = 5\sigma\sqrt{np}$$



A motivating experiment

$$n = 1000, \; r = 5, \; p = 0.2, \; \lambda = 5\sigma\sqrt{np}$$



Convex and nonconvex solutions are exceedingly close!

convex  nonconvex

$$\Downarrow$$

$$\text{stability}\left(\boxed{\text{convex}}\right) \approx \text{stability}\left(\boxed{\text{nonconvex}}\right)$$

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, well-conditioned

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2 + \lambda \|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2 + \lambda \|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

Theorem 1 (Chen, Chi, Fan, Ma, Yan '19)

With high prob., any minimizer $\boldsymbol{M}^{\mathsf{cvx}}$ of convex program obeys

1. $\boldsymbol{M}^{\mathsf{cvx}}$ *is nearly rank-r*

$$\big\|\boldsymbol{M}^{\mathsf{cvx}} - \mathsf{proj}_{\textit{rank-}r}(\boldsymbol{M}^{\mathsf{cvx}})\big\|_{\mathrm{F}} \ll \frac{1}{n^5} \cdot \sigma\sqrt{\frac{n}{p}}$$

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2 + \lambda \|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

Theorem 1 (Chen, Chi, Fan, Ma, Yan '19)

With high prob., any minimizer $\boldsymbol{M}^{\mathsf{cvx}}$ of convex program obeys

1. $\boldsymbol{M}^{\mathsf{cvx}}$ *is nearly rank-r*
2. $$\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}$$

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

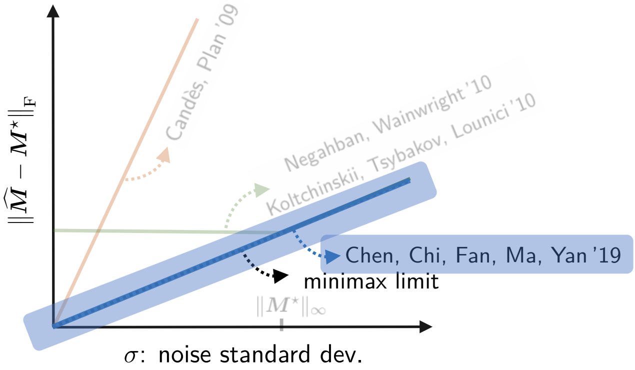
$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

Theorem 1 (Chen, Chi, Fan, Ma, Yan '19)

With high prob., any minimizer $\boldsymbol{M}^{\mathsf{cvx}}$ of convex program obeys

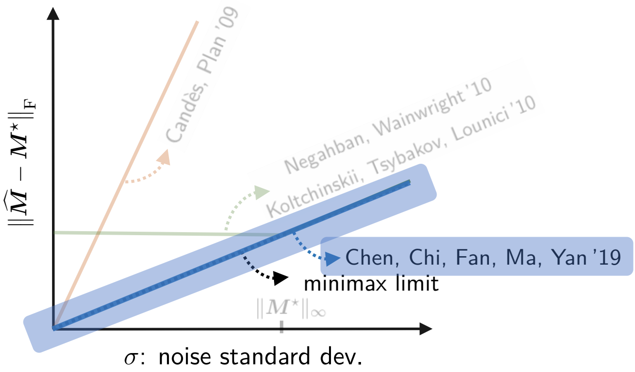
1. $\boldsymbol{M}^{\mathsf{cvx}}$ *is nearly rank-r*
2. $$\left\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}$$
$$\left\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\infty} \lesssim \sigma\sqrt{\frac{n\log n}{p}} \cdot \frac{1}{n}$$

$$\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^{\star}\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}$$



- minimax optimal when $r = O(1)$

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



- minimax optimal when $r = O(1)$
- estimation errors are spread out across all entries

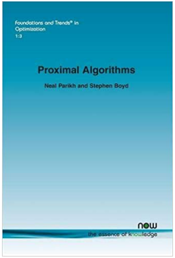
Implicit regularization

No need to enforce spikiness constraint as in Negahban & Wainwright

$$\underset{\|\boldsymbol{Z}\|_\infty \le \alpha}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_* \qquad \text{(Negahban et al.)}$$

- convex programming automatically controls spikiness of solutions

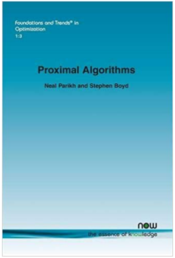
Statistical guarantees for iterative algorithms



$$\underset{\boldsymbol{Z}}{\text{minimize}} \quad g(\boldsymbol{Z}) := \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_* \quad (1)$$

Many algorithms (e.g. SVT, SOFT-IMPUTE, FPC, FISTA) have been proposed to solve (1), typically without statistical guarantees

Statistical guarantees for iterative algorithms



$$\underset{\boldsymbol{Z}}{\text{minimize}} \quad g(\boldsymbol{Z}) := \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda \|\boldsymbol{Z}\|_* \quad (1)$$

Many algorithms (e.g. SVT, SOFT-IMPUTE, FPC, FISTA) have been proposed to solve (1), typically without statistical guarantees

We provide statistical guarantees for any $\boldsymbol{Z}$ with $g(\boldsymbol{Z}) \leq g(\boldsymbol{Z}_{\mathsf{opt}}) + \varepsilon$ for some sufficiently small $\varepsilon > 0$

Main results: general case

- **random sampling:** each $(i, j) \in \Omega$ with prob. $p \gtrsim \frac{r^2 \log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: incoherent, well-conditioned

Main results: general case

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{r^2 \log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: incoherent, well-conditioned

Theorem 2 (Chen, Chi, Fan, Ma, Yan '19)

With high prob., any minimizer $\boldsymbol{M}^{\mathsf{cvx}}$ of convex program obeys

1. $\boldsymbol{M}^{\mathsf{cvx}}$ *is nearly rank-r*
2. $$\begin{aligned} \left\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\mathrm{F}} &\lesssim \frac{\sigma}{\sigma_{\min}(\boldsymbol{M}^\star)}\sqrt{\frac{n}{p}}\,\left\|\boldsymbol{M}^\star\right\|_{\mathrm{F}} \\ \left\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\right\|_{\infty} &\lesssim \sqrt{r}\frac{\sigma}{\sigma_{\min}(\boldsymbol{M}^\star)}\sqrt{\frac{n\log n}{p}}\,\left\|\boldsymbol{M}^\star\right\|_{\infty} \\ \left\|\boldsymbol{M}^{\mathsf{cvx}} - \boldsymbol{M}^\star\right\| &\lesssim \frac{\sigma}{\sigma_{\min}(\boldsymbol{M}^\star)}\sqrt{\frac{n}{p}}\,\left\|\boldsymbol{M}^\star\right\| \end{aligned}$$

Main results: general case

- **random sampling:** each $(i, j) \in \Omega$ with prob. $p \gtrsim \frac{r^2 \log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian with variance σ^2 (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: incoherent, well-conditioned

sample complexity bound $O(nr^2 \log^3 n)$ is suboptimal in r!

A little analysis:
connection between convex and nonconvex solutions

Link between convex and nonconvex optimizers

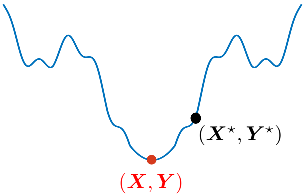
$(\boldsymbol{X}, \boldsymbol{Y})$ is nonconvex optimizer

Link between convex and nonconvex optimizers

$$(\boldsymbol{X}, \boldsymbol{Y}) \text{ is nonconvex optimizer} \quad \stackrel{?}{\Longrightarrow} \quad \boldsymbol{X}\boldsymbol{Y}^\top \text{ is convex solution}$$

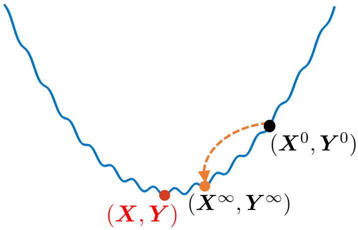
Link between convex and nonconvex optimizers

- $(\boldsymbol{X}, \boldsymbol{Y})$ is close to truth (in $\ell_{2,\infty}$ sense)
- a little condition on noise size



$(\boldsymbol{X}, \boldsymbol{Y})$ is nonconvex optimizer $\quad \overset{\checkmark}{\Longrightarrow} \quad \boldsymbol{X}\boldsymbol{Y}^\top$ is convex solution

Approximate nonconvex optimizers



Issue: we do NOT know properties of nonconvex optimizers

- It is unclear whether nonconvex algorithms converge to optimizers (due to lack of strong convexity)

Approximate nonconvex optimizers

Strategy: resort to "$\underbrace{\text{approximate stationary points}}_{\nabla f(\boldsymbol{X},\boldsymbol{Y}) \approx \boldsymbol{0}}$" instead

Approximate nonconvex optimizers

Strategy: resort to "$\underbrace{\text{approximate stationary points}}_{\nabla f(\boldsymbol{X},\boldsymbol{Y})\approx \boldsymbol{0}}$" instead

starting from $(\boldsymbol{X}^0, \boldsymbol{Y}^0) =$ truth or spectral initialization:

$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \end{aligned} \qquad t = 0, 1, \cdots, T$$

Approximate nonconvex optimizers

Strategy: resort to "$\underbrace{\text{approximate stationary points}}_{\nabla f(\boldsymbol{X},\boldsymbol{Y})\approx \boldsymbol{0}}$" instead

starting from $(\boldsymbol{X}^0, \boldsymbol{Y}^0) =$ truth or spectral initialization:

$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \end{aligned} \qquad t = 0, 1, \cdots, T$$

- when T is large: there exists point with $\underbrace{\text{very small gradient}}_{\|\nabla f(\boldsymbol{X},\boldsymbol{Y})\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{\eta T}}}$
- hopefully not far from $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$

Analyzing nonconvex GD: leave-one-out analysis

Leave out a small amount of information from data and run GD

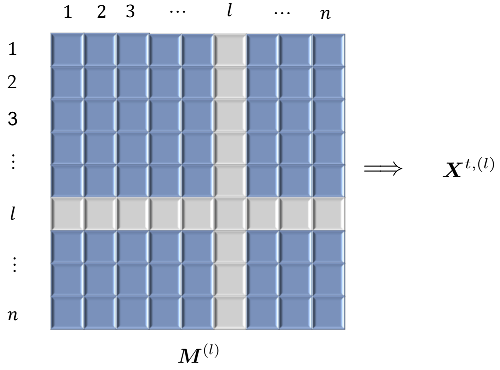
Analyzing nonconvex GD: leave-one-out analysis

Leave out a small amount of information from data and run GD

- Stein '72
- El Karoui, Bean, Bickel, Lim, Yu '13
- El Karoui '15
- Javanmard, Montanari '15
- Zhong, Boumal '17
- Lei, Bickel, El Karoui '17
- Sur, Chen, Candès '17
- Abbe, Fan, Wang, Zhong '17
- Chen, Fan, Ma, Wang '17
- Ma, Wang, Chi, Chen '17
- Chen, Chi, Fan, Ma '18
- Ding, Chen '18
- Dong, Shi '18
- Chen, Liu, Li '19

Analyzing nonconvex GD: leave-one-out analysis

For each $1 \leq l \leq n$, introduce leave-one-out iterates $\boldsymbol{X}^{t,(l)}$ by replacing l^{th} row and column with true values



- exploit partial statistical independence
- exploit leave-one-out stability

Inference and uncertainty quantification

Reasoning about uncertainty

	2		2	
		6		
3	1		4	
		4		1
	0			

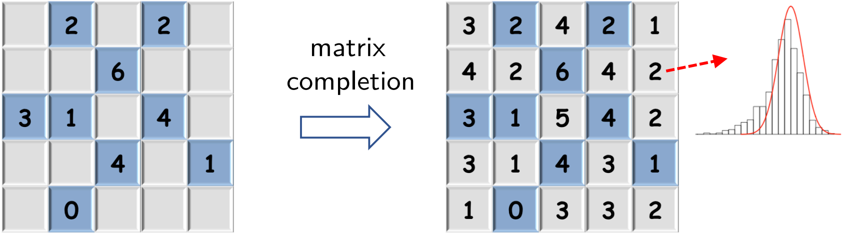
Reasoning about uncertainty

	2		2	
		6		
3	1		4	
		4		1
	0			

matrix completion

3	2	4	2	1
4	2	6	4	2
3	1	5	4	2
3	1	4	3	1
1	0	3	3	2

Reasoning about uncertainty



How to assess uncertainty, or "confidence", of obtained estimates?

INFERENCE IN HIGH DIMENSIONAL REGRESSION

organized by
Peter Buehlmann, Andrea Montanari, and Jonathan Taylor

American Institute of Mathematics

(3) *Confidence intervals for matrix completion.* In matrix completion, the data analyst is given a large data matrix with a number of missing entries. In many interesting applications (e.g. to collaborative filtering) it is indeed the case that the vast majority of entries is missing. In order to fill the missing entries, the assumption is made that the underlying –unknown– matrix has a low-rank structure.

Substantial work has been devoted to methods for computing point estimates of the missing entries. In applications, it would be very interesting to compute confidence intervals as well. This requires developing distributional characterizations of standard matrix completion methods.

Challenges

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\arg\min} \sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2 + \lambda\|\boldsymbol{Z}\|_*$$

- convex estimate $\boldsymbol{M}^{\mathsf{cvx}}$ is biased towards small norm

Challenges

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \arg\min_{\boldsymbol{Z}\in\mathbb{R}^{n\times n}} \sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2 + \lambda\|\boldsymbol{Z}\|_*$$

- convex estimate $\boldsymbol{M}^{\mathsf{cvx}}$ is biased towards small norm
- very challenging to pin down distributions of obtained estimates

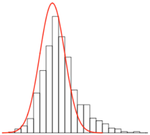
Challenges

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\arg\min} \sum_{(i,j)\in\Omega} \big(Z_{i,j} - M_{i,j}\big)^2 + \lambda\|\boldsymbol{Z}\|_*$$

- convex estimate $\boldsymbol{M}^{\mathsf{cvx}}$ is biased towards small norm
- very challenging to pin down distributions of obtained estimates
- existing orderwise bounds come with unspecified (but huge) pre-constants

  $\longrightarrow$ overly wide confidence intervals

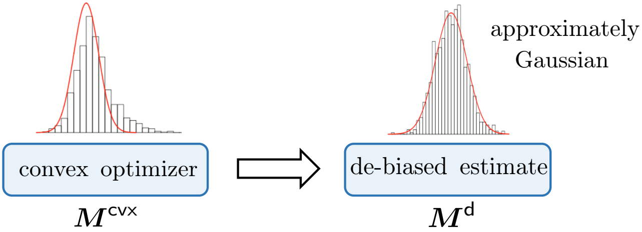
— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*



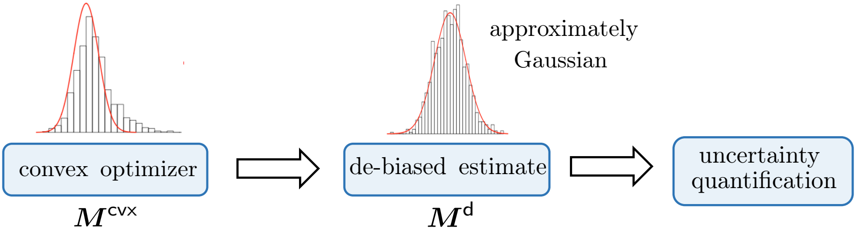
convex optimizer

M^{cvx}

— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*



— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*



De-biasing convex estimate

$$\boldsymbol{M}^{\mathsf{cvx}} \quad \xrightarrow{\text{de-biasing}} \quad \underbrace{\boldsymbol{M}^{\mathsf{cvx}} + \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E} - \boldsymbol{M}^{\mathsf{cvx}})}_{\text{(nearly) unbiased estimate of } \boldsymbol{M}^{\star}}$$

De-biasing convex estimate

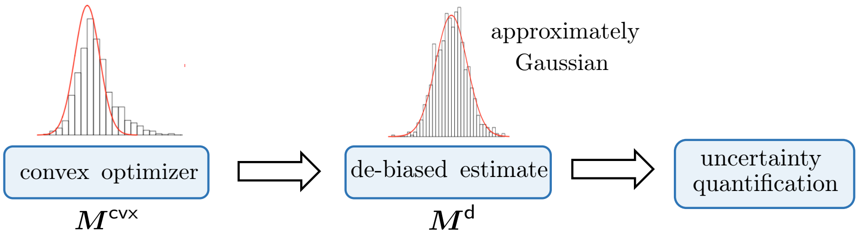
$$\boldsymbol{M}^{\mathsf{cvx}} \quad \xrightarrow{\text{de-biasing}} \quad \underbrace{\boldsymbol{M}^{\mathsf{cvx}} + \frac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E} - \boldsymbol{M}^{\mathsf{cvx}})}_{\text{(nearly) unbiased estimate of } \boldsymbol{M}^{\star}}$$

- **issue:** high-rank after de-biasing; statistical accuracy suffers

De-biasing convex estimate

$$\boldsymbol{M}^{\mathsf{cvx}} \overset{\text{de-biasing}}{\longrightarrow} \underbrace{\mathsf{proj}_{\mathsf{rank}\text{-}r}\Big(\boldsymbol{M}^{\mathsf{cvx}} + \tfrac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E} - \boldsymbol{M}^{\mathsf{cvx}}) \Big)}_{\text{1 iteration of singular value projection (Jain, Meka, Dhillon '10)}} =: \boldsymbol{M}^{\mathrm{d}}$$

- **issue:** high-rank after de-biasing; statistical accuracy suffers
- **solution:** low-rank projection



approximately Gaussian
convex optimizer
M^{cvx}
de-biased estimate
M^{d}
uncertainty quantification

Distributional guarantees for low-rank factors

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. $\mathcal{N}(0, \sigma^2)$ (not too large)
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, well-conditioned
- regularization parameter: $\lambda \asymp \sigma\sqrt{np}$

$$\boldsymbol{X}^{\mathsf{d}}\boldsymbol{Y}^{\mathsf{d}\top} \leftarrow \underbrace{\text{balanced}}_{\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}} = \boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}} \text{ rank-}r \text{ decomp. of } \boldsymbol{M}^{\mathsf{d}}$$

$$\boldsymbol{X}^\star\boldsymbol{Y}^{\star\top} \leftarrow \underbrace{\text{balanced}}_{\boldsymbol{X}^{\star\top}\boldsymbol{X}^\star = \boldsymbol{Y}^{\star\top}\boldsymbol{Y}^\star} \text{ rank-}r \text{ decomp. of } \boldsymbol{M}^\star$$

Distributional guarantees for low-rank factors

$$\boldsymbol{X}^{\mathsf{d}}\boldsymbol{Y}^{\mathsf{d}\top} \leftarrow \underbrace{\text{balanced}}_{\boldsymbol{X}^{\mathsf{d}\top}\boldsymbol{X}^{\mathsf{d}}=\boldsymbol{Y}^{\mathsf{d}\top}\boldsymbol{Y}^{\mathsf{d}}} \text{ rank-}r \text{ approx. of } \boldsymbol{M}^{\mathsf{d}}$$

$$\boldsymbol{X}^{\star}\boldsymbol{Y}^{\star\top} \leftarrow \underbrace{\text{balanced}}_{\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star}=\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star}} \text{ rank-}r \text{ decomp. of } \boldsymbol{M}^{\star}$$

Theorem 3 (Chen, Fan, Ma, Yan '19)

With high prob., there exists global rotation matrix $\boldsymbol{H} \in \mathbb{R}^{r\times r}$ s.t.

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \;\approx\; \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \;\approx\; \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

- estimation errors for different rows of $\boldsymbol{X}^{\star}$ are nearly independent

$$\boldsymbol{X}^{\mathrm{d}}_{i,\cdot}\boldsymbol{H} - \boldsymbol{X}^{\star}_{i,\cdot} \quad \text{nearly ind. of} \quad \boldsymbol{X}^{\mathrm{d}}_{j,\cdot}\boldsymbol{H} - \boldsymbol{X}^{\star}_{j,\cdot}$$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \;\approx\; \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \;\approx\; \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

- accurate uncertainty quantification for low-rank factors, e.g.

$$\boldsymbol{X}^{\mathsf{d}}_{i,\cdot}\boldsymbol{H} - \boldsymbol{X}^{\star}_{i,\cdot} \;\sim\; \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big) + \text{negligible term}$$

$$\boldsymbol{Y}^{\mathsf{d}}_{i,\cdot}\boldsymbol{H} - \boldsymbol{Y}^{\star}_{i,\cdot} \;\sim\; \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big) + \text{negligible term}$$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\right)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\right)$$

- accurate uncertainty quantification for low-rank factors, e.g.

$$\boldsymbol{X}^{\mathsf{d}}_{i,\cdot} - \boldsymbol{X}^{\star}_{i,\cdot}\boldsymbol{H}^{\top} \sim \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\mathrm{d}\top}\boldsymbol{Y}^{\mathrm{d}})^{-1}\right) + \text{negligible term}$$

$$\boldsymbol{Y}^{\mathsf{d}}_{i,\cdot} - \boldsymbol{Y}^{\star}_{i,\cdot}\boldsymbol{H}^{\top} \sim \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\mathrm{d}\top}\boldsymbol{X}^{\mathrm{d}})^{-1}\right) + \text{negligible term}$$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

- accurate uncertainty quantification for low-rank factors, e.g.

$$\boldsymbol{X}^{\mathsf{d}}_{i,\cdot} - \boldsymbol{X}^{\star}_{i,\cdot}\boldsymbol{H}^{\top} \sim \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\mathrm{d}\top}\boldsymbol{Y}^{\mathrm{d}})^{-1}\big) + \text{negligible term}$$

$$\boldsymbol{Y}^{\mathsf{d}}_{i,\cdot} - \boldsymbol{Y}^{\star}_{i,\cdot}\boldsymbol{H}^{\top} \sim \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\mathrm{d}\top}\boldsymbol{X}^{\mathrm{d}})^{-1}\big) + \text{negligible term}$$

— *asymptotically optimal*

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\right)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\right)$$

- accurate uncertainty quantification for matrix entries: if $\|\boldsymbol{X}^{\star}_{i,\cdot}\|_2 + \|\boldsymbol{Y}^{\star}_{j,\cdot}\|_2$ is not too small, then

$$M^{\mathsf{d}}_{i,j} - M^{\star}_{i,j} \sim \mathcal{N}\left(0, v^{\star}_{i,j}\right) + \text{negligible term}$$

where $v^{\star}_{i,j} \triangleq \frac{\sigma^2}{p}\left\{\boldsymbol{X}^{\star}_{i,\cdot}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\boldsymbol{X}^{\star\top}_{i,\cdot} + \boldsymbol{Y}^{\star}_{j,\cdot}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\boldsymbol{Y}^{\star\top}_{j,\cdot}\right\}$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\big)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\big(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\big)$$

- accurate uncertainty quantification for matrix entries: if $\|\boldsymbol{X}^{\star}_{i,\cdot}\|_2 + \|\boldsymbol{Y}^{\star}_{j,\cdot}\|_2$ is not too small, then

$$M^{\mathsf{d}}_{i,j} - M^{\star}_{i,j} \sim \mathcal{N}\big(0, \widehat{v}_{i,j}\big) + \text{negligible term}$$

where $\widehat{v}_{i,j} \triangleq \frac{\sigma^2}{p}\Big\{\boldsymbol{X}^{\mathrm{d}}_{i,\cdot}(\boldsymbol{X}^{\mathrm{d}\top}\boldsymbol{X}^{\mathrm{d}})^{-1}\boldsymbol{X}^{\mathrm{d}\top}_{i,\cdot} + \boldsymbol{Y}^{\mathrm{d}}_{j,\cdot}(\boldsymbol{Y}^{\mathrm{d}\top}\boldsymbol{Y}^{\mathrm{d}})^{-1}\boldsymbol{Y}^{\mathrm{d}\top}_{j,\cdot}\Big\}$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{X}^{\star} \approx \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{Y}^{\star\top}\boldsymbol{Y}^{\star})^{-1}\right)$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{H} - \boldsymbol{Y}^{\star} \approx \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \tfrac{\sigma^2}{p}(\boldsymbol{X}^{\star\top}\boldsymbol{X}^{\star})^{-1}\right)$$

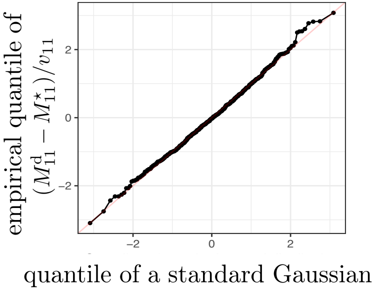
- accurate uncertainty quantification for matrix entries: if $\|\boldsymbol{X}^{\star}_{i,\cdot}\|_2 + \|\boldsymbol{Y}^{\star}_{j,\cdot}\|_2$ is not too small, then

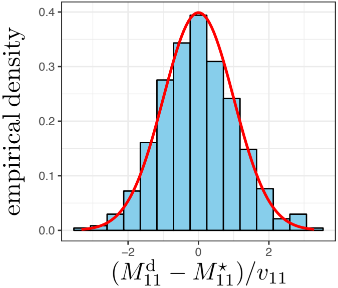
$$M^{\mathsf{d}}_{i,j} - M^{\star}_{i,j} \sim \mathcal{N}\big(0, \widehat{v}_{i,j}\big) + \text{negligible term}$$

where $\widehat{v}_{i,j} \triangleq \frac{\sigma^2}{p}\Big\{\boldsymbol{X}^{\mathrm{d}}_{i,\cdot}(\boldsymbol{X}^{\mathrm{d}\top}\boldsymbol{X}^{\mathrm{d}})^{-1}\boldsymbol{X}^{\mathrm{d}\top}_{i,\cdot} + \boldsymbol{Y}^{\mathrm{d}}_{j,\cdot}(\boldsymbol{Y}^{\mathrm{d}\top}\boldsymbol{Y}^{\mathrm{d}})^{-1}\boldsymbol{Y}^{\mathrm{d}\top}_{j,\cdot}\Big\}$

— *asymptotically optimal*

Numerical experiments





$$n = 1000,\ p = 0.2,\ r = 5,\ \|\boldsymbol{M}^{\star}\| = 1,\ \kappa = 1,\ \sigma = 10^{-3}$$

convex

nonconvex

convex

nonconvex

$$\Downarrow$$

$$\text{inference}\left(\boxed{\text{convex}}\right) \approx \text{inference}\left(\boxed{\text{nonconvex}}\right)$$

Same inference procedures work for both cvx & noncvx estimates!

A bit of intuition

Consider rank-1 PSD case $\boldsymbol{M}^\star = \boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$, $p = 1$ (no missing data)

$$\text{minimize}_{\boldsymbol{x}} \qquad \frac{1}{2}\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{x}\|_2^2$$

A bit of intuition

Consider rank-1 PSD case $\boldsymbol{M}^\star = \boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$, $p = 1$ (no missing data)

$$\text{minimize}_{\boldsymbol{x}} \qquad \frac{1}{2}\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star \boldsymbol{x}^{\star\top} - \boldsymbol{E}\|_{\mathrm{F}}^2 + \lambda\|\boldsymbol{x}\|_2^2$$

- first-order optimality condition

$$(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star \boldsymbol{x}^{\star\top} - \boldsymbol{E})\boldsymbol{x} + \lambda\boldsymbol{x} = \mathbf{0}$$

A bit of intuition

$$(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E})\boldsymbol{x} \quad \underbrace{+\lambda\boldsymbol{x}}_{\text{causes bias}} \quad = \boldsymbol{0}$$

A bit of intuition

$$(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E})\boldsymbol{x} \quad \underbrace{+\lambda\boldsymbol{x}}_{\text{causes bias}} \quad = \boldsymbol{0}$$

$$\Updownarrow$$

$$(\boldsymbol{x}^{\mathrm{d}}\boldsymbol{x}^{\mathrm{d}\top} - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E})\boldsymbol{x}^{\mathrm{d}} = \boldsymbol{0}, \qquad \boldsymbol{x}^{\mathrm{d}} = \sqrt{\frac{\lambda + \|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}}\,\boldsymbol{x}$$

A bit of intuition

$$(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E})\boldsymbol{x} \underbrace{+\lambda\boldsymbol{x}}_{\text{causes bias}} = \boldsymbol{0}$$

$$\Updownarrow$$

$$(\boldsymbol{x}^{\mathrm{d}}\boldsymbol{x}^{\mathrm{d}\top} - \boldsymbol{x}^\star\boldsymbol{x}^{\star\top} - \boldsymbol{E})\boldsymbol{x}^{\mathrm{d}} = \boldsymbol{0}, \qquad \boldsymbol{x}^{\mathrm{d}} = \sqrt{\tfrac{\lambda + \|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}}\,\boldsymbol{x}$$

$$\Updownarrow$$

$$\boldsymbol{x}^{\mathrm{d}} - \boldsymbol{x}^\star = \underbrace{\frac{1}{\|\boldsymbol{x}^{\mathrm{d}}\|_2^2}\boldsymbol{E}\boldsymbol{x}^{\mathrm{d}}}_{\text{nearly Gaussian}} + \underbrace{\frac{\left(\boldsymbol{x}^\star - \boldsymbol{x}^{\mathrm{d}}\right)^\top \boldsymbol{x}^{\mathrm{d}}}{\|\boldsymbol{x}^{\mathrm{d}}\|_2^2}\boldsymbol{x}^\star}_{\text{hopefully small}}$$

An alternative de-biased estimator

— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \arg\min_{\boldsymbol{Z} \in \mathbb{R}^{n \times n}} \ \left\| \mathcal{P}_{\Omega}(\boldsymbol{Z} - \boldsymbol{M}) \right\|_{\mathrm{F}}^2 + \lambda \|\boldsymbol{Z}\|_*$$

An alternative de-biased estimator

— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \arg\min_{\boldsymbol{Z} \in \mathbb{R}^{n \times n}} \left\| \mathcal{P}_{\Omega}(\boldsymbol{Z} - \boldsymbol{M}) \right\|_{\mathrm{F}}^{2} + \lambda \|\boldsymbol{Z}\|_{*}$$

$$\Downarrow$$

$$\widehat{\boldsymbol{M}}^{\mathrm{d}} = \boldsymbol{M}^{\mathsf{cvx}} + \textsf{linear-map}\Big(\mathcal{P}_{\Omega}\big(\boldsymbol{M} - \boldsymbol{M}^{\mathsf{cvx}}\big) \Big)$$

An alternative de-biased estimator

— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*

$$\boldsymbol{M}^{\mathsf{cvx}} \triangleq \underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\arg\min} \ \left\|\mathcal{P}_{\Omega}(\boldsymbol{Z}-\boldsymbol{M})\right\|_{\mathrm{F}}^{2} + \lambda\|\boldsymbol{Z}\|_{*}$$

$$\Downarrow$$

$$\widehat{\boldsymbol{M}}^{\mathrm{d}} = \boldsymbol{M}^{\mathsf{cvx}} + \frac{1}{p}\mathcal{P}_{T}\Big(\mathcal{P}_{\Omega}(\boldsymbol{M}-\boldsymbol{M}^{\mathsf{cvx}})\Big)$$

— $\mathcal{P}_T$: projection onto T (tangent space at $\mathrm{proj}_{\mathsf{rank\text{-}r}}(\boldsymbol{M}^{\mathsf{cvx}})$)

$$\widehat{\boldsymbol{M}}^{\mathrm{d}} \approx \boldsymbol{M}^{\mathrm{d}} \text{ !}$$

Back to estimation: de-biased estimator is optimal

Distributional theory in turn allows us to track estimation accuracy

Back to estimation: de-biased estimator is optimal

Distributional theory in turn allows us to track estimation accuracy

Theorem 4 (Chen, Fan, Ma, Yan '19)

$$\|\boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^{\star}\|_{\mathrm{F}}^{2} = \underbrace{\frac{(2+o(1))nr\sigma^{2}}{p}}_{\text{Cramer-Rao lower bound}} \qquad \textit{with high prob.}$$

Back to estimation: de-biased estimator is optimal

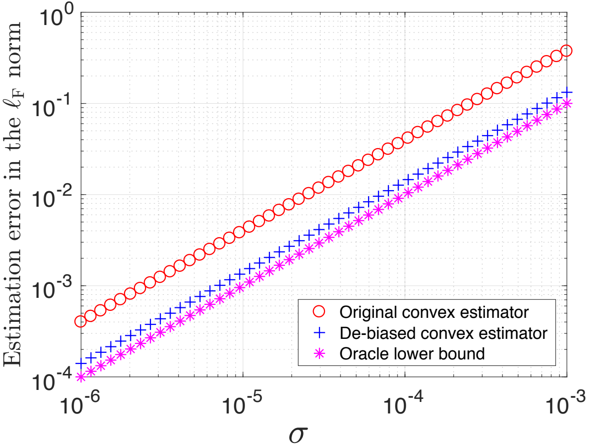
Distributional theory in turn allows us to track estimation accuracy

Theorem 4 (Chen, Fan, Ma, Yan '19)

$$\|\boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^{\star}\|_{\mathrm{F}}^2 = \underbrace{\frac{(2+o(1))nr\sigma^2}{p}}_{\textit{Cramer-Rao lower bound}} \quad \textit{with high prob.}$$

- precise characterization of estimation accuracy
- achieves full statistical efficiency (including pre-constant)

Numerical evidence ($r = 5$, $p = 0.2$, $n = 1000$)



Euclidean estimation error vs. noise standard deviation σ

Concluding remarks

near-optimal
stability guarantees

convex

nonconvex

asymptotically optimal
uncertainty quantification

Concluding remarks

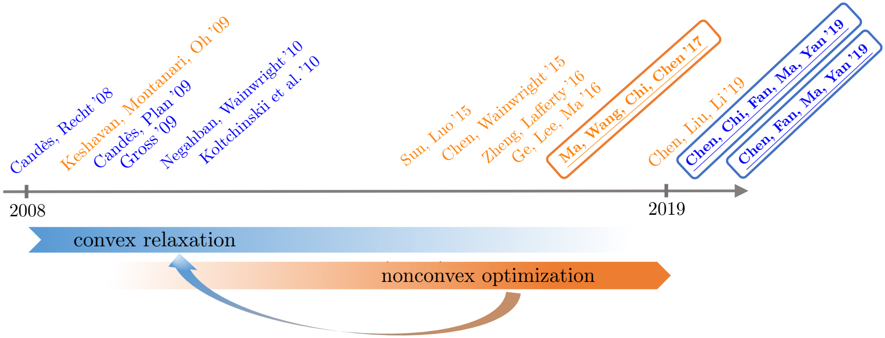
near-optimal
stability guarantees

convex

nonconvex

asymptotically optimal
uncertainty quantification

- improve dependency on rank & cond. number
- what if $M^\star$ is only approximately low-rank?
- more general sampling patterns

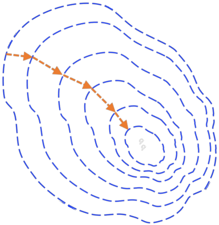


"Noisy matrix completion: understanding statistical guarantees for convex relaxation via nonconvex optimization," Y. Chen, Y. Chi, J. Fan, C. Ma, Y. Yan, 2019

"Inference and uncertainty quantification for noisy matrix completion," Y. Chen, J. Fan, C. Ma, Y. Yan, 2019

Backup slides: gradient descent for nonconvex matrix completion

Gradient descent for nonconvex matrix completion



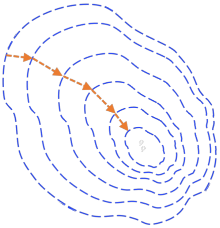
$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^{t} - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^{t}, \boldsymbol{Y}^{t}) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^{t} - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^{t}, \boldsymbol{Y}^{t}) \end{aligned}$$

Prior works analyze regularized GD

- not guaranteed to return small-gradient solutions
- no $\ell_{2,\infty}$ error control

— *Keshavan et al. '09, Sun, Luo '15, Chen, Wainwright '15, Zheng, Lafferty '16*

Gradient descent for nonconvex matrix completion



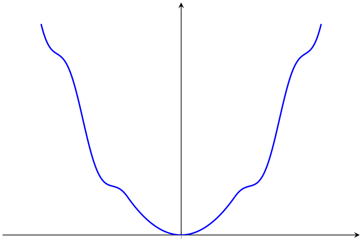
$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \end{aligned}$$

Our work and Chen et al. analyze vanilla GD

- regularization-free
- optimal $\ell_{2,\infty}$ error control

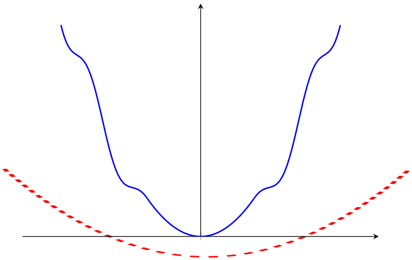
— *Ma, Wang, Chi, Chen '17, Chen, Liu, Li '19*

Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

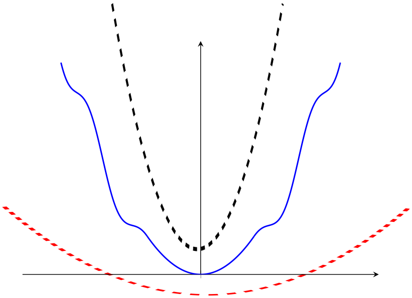
Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity

Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

- (local) restricted strong convexity
- (local) smoothness

Gradient descent theory revisited

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

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

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

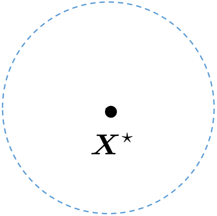
$$\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^\star\|_{\mathrm{F}} \le \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{X}^t - \boldsymbol{X}^\star\|_{\mathrm{F}}$$

Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

Incoherence region

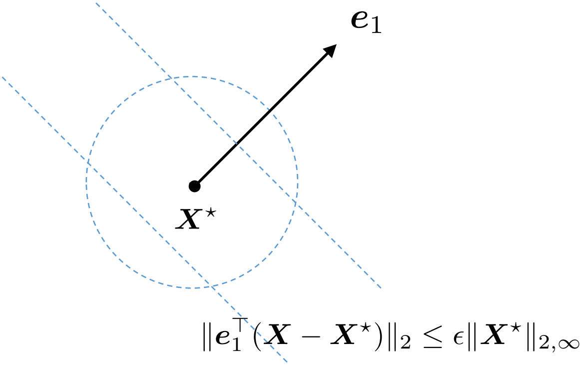
Which region enjoys both restricted strong convexity and smoothness?



- X is not far away from $X^\star$

Incoherence region

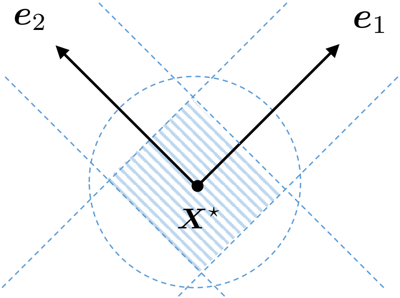
Which region enjoys both restricted strong convexity and smoothness?



- X is not far away from $X^\star$
- X is incoherent w.r.t. standard basis vectors (incoherence region)

Incoherence region

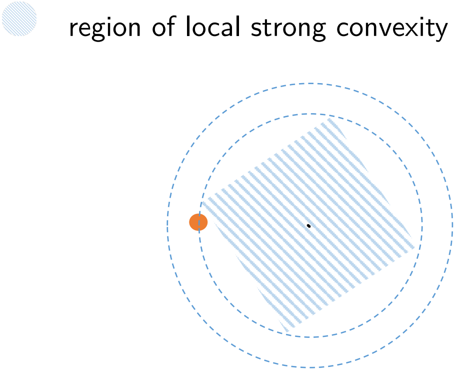
Which region enjoys both restricted strong convexity and smoothness?



$\|e_2^\top (X - X^\star)\|_2 \le \epsilon \|X^\star\|_{2,\infty}$ $\|e_1^\top (X - X^\star)\|_2 \le \epsilon \|X^\star\|_{2,\infty}$

- X is not far away from $X^\star$
- X is incoherent w.r.t. standard basis vectors (incoherence region)

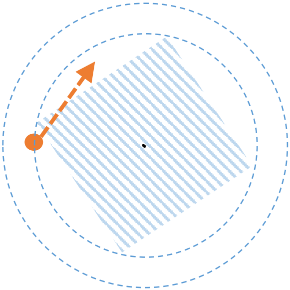
Inadequacy of generic gradient descent theory



- Generic optimization theory does NOT ensure GD stays in incoherence region

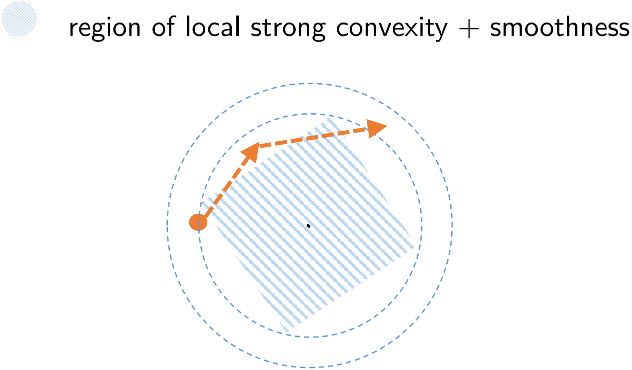
Inadequacy of generic gradient descent theory

region of local strong convexity + smoothness



- Generic optimization theory does NOT ensure GD stays in incoherence region

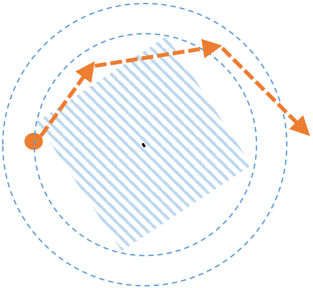
Inadequacy of generic gradient descent theory



- Generic optimization theory does NOT ensure GD stays in incoherence region

Inadequacy of generic gradient descent theory

region of local strong convexity + smoothness



- Generic optimization theory does NOT ensure GD stays in incoherence region
- Calls for new analysis tools

Key proof idea: leave-one-out analysis

Leave out a small amount of information from data and run GD

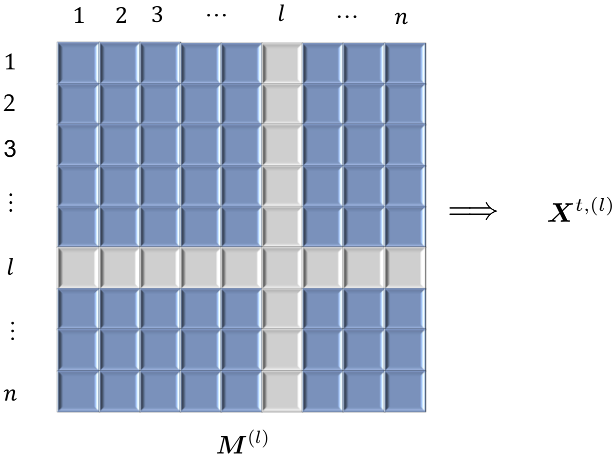
Key proof idea: leave-one-out analysis

Leave out a small amount of information from data and run GD

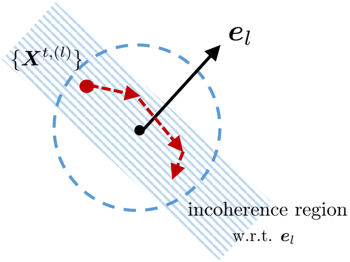
- Stein '72
- El Karoui, Bean, Bickel, Lim, Yu '13
- El Karoui '15
- Javanmard, Montanari '15
- Zhong, Boumal '17
- Lei, Bickel, El Karoui '17
- Sur, Chen, Candès '17
- Abbe, Fan, Wang, Zhong '17
- Chen, Fan, Ma, Wang '17
- Ma, Wang, Chi, Chen '17
- Chen, Chi, Fan, Ma '18
- Ding, Chen '18
- Dong, Shi '18
- Chen, Liu, Li '19

Key proof idea: leave-one-out analysis

For each $1 \leq l \leq n$, introduce leave-one-out iterates $\boldsymbol{X}^{t,(l)}$ by replacing l^{th} row and column with true values

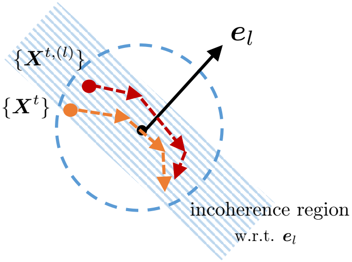


Key proof idea: leave-one-out analysis



- Leave-one-out iterates $\{\boldsymbol{X}^{t,(l)}\}$ contains more information of l^{th} row of truth; indep. of randomness in l^{th} row

Key proof idea: leave-one-out analysis



- Leave-one-out iterates $\{\boldsymbol{X}^{t,(l)}\}$ contains more information of l^{th} row of truth; indep. of randomness in l^{th} row
- Leave-one-out iterates $\{\boldsymbol{X}^{t,(l)}\}$ $\approx$ true iterates $\{\boldsymbol{X}^{t}\}$