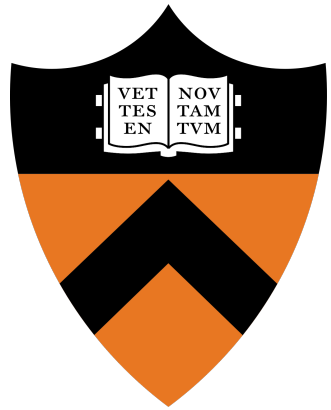
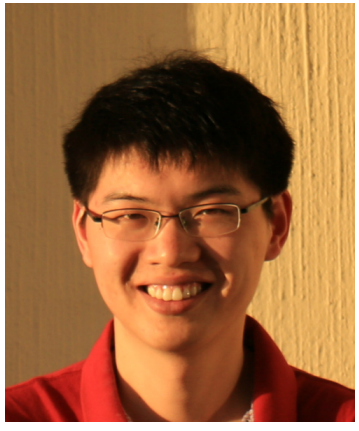
Stability, nonconvex optimization & asymmetry in low-rank matrix estimation



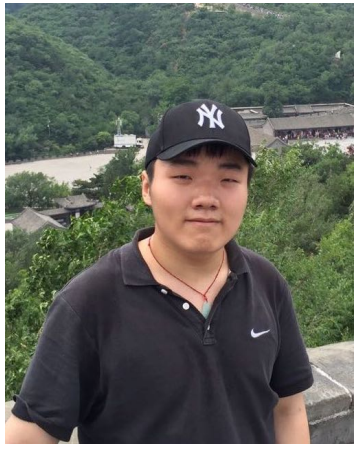
Yuxin Chen

EE, Princeton University

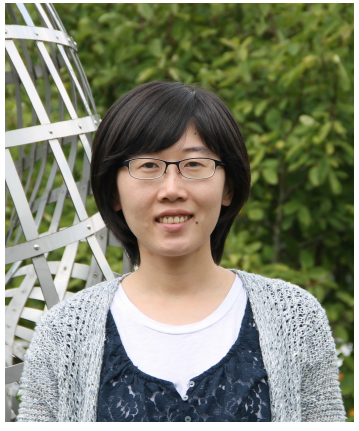
Noisy matrix completion: understanding stability of convex relaxation via nonconvex optimization



Cong Ma
Princeton ORFE



Yuling Yan
Princeton ORFE

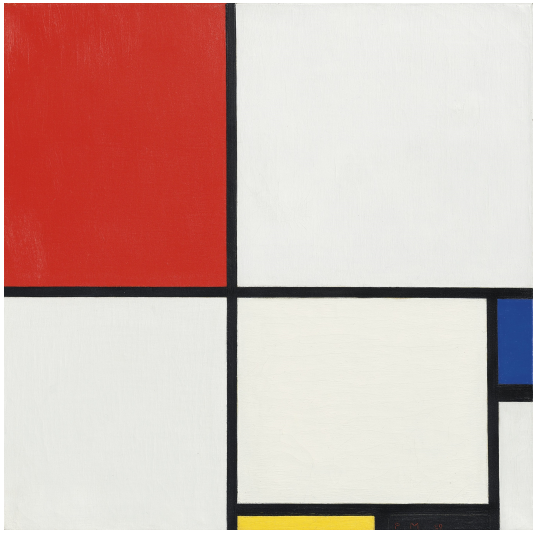


Yuejie Chi
CMU ECE

Jianqing Fan
Princeton ORFE

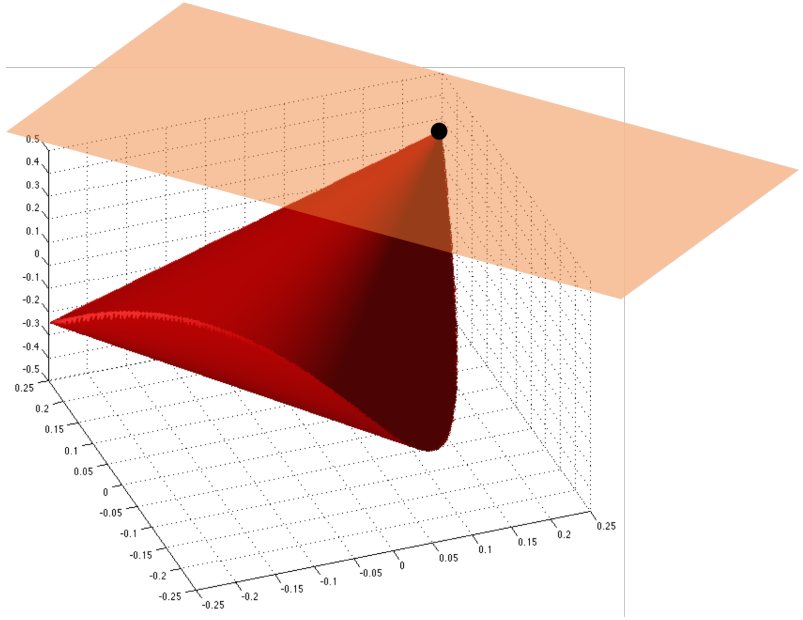
Convex relaxation for low-rank structure

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



low-rank matrix

figure credit: Piet Mondrian



semidefinite relaxation

Convex relaxation for low-rank structure

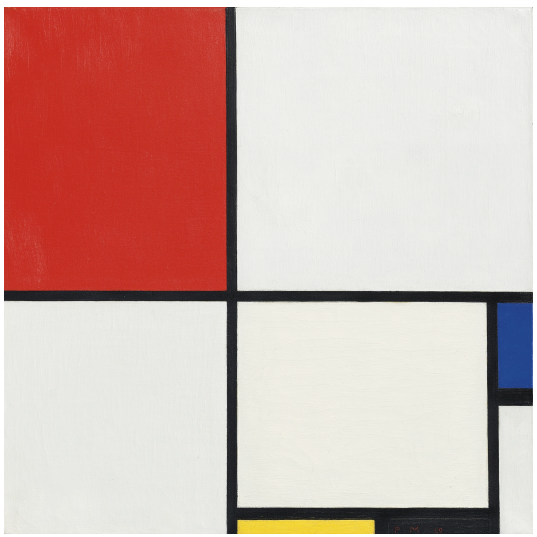
$$\begin{aligned} \underset{Z}{\text{minimize}} \quad & \|Z\|_* \\ \text{subj. to} \quad & \text{noiseless data constraints} \end{aligned}$$

- ✓ 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)

...

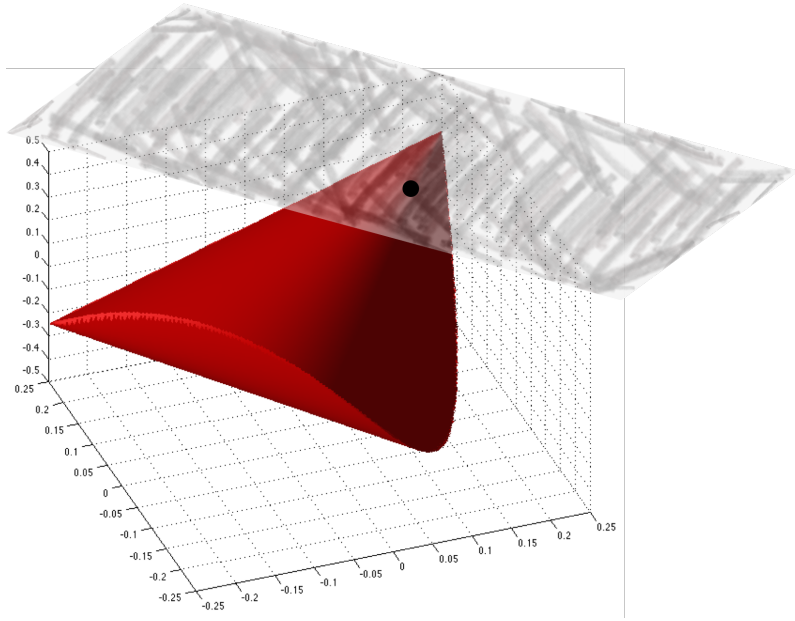
Stability of convex relaxation against noise

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



low-rank matrix

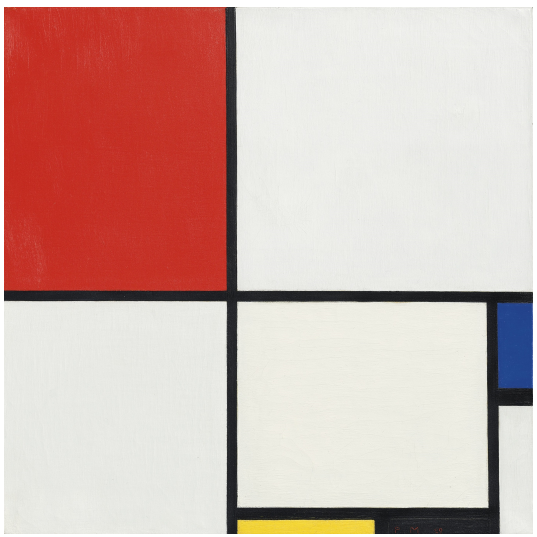
figure credit: Piet Mondrian



semidefinite relaxation

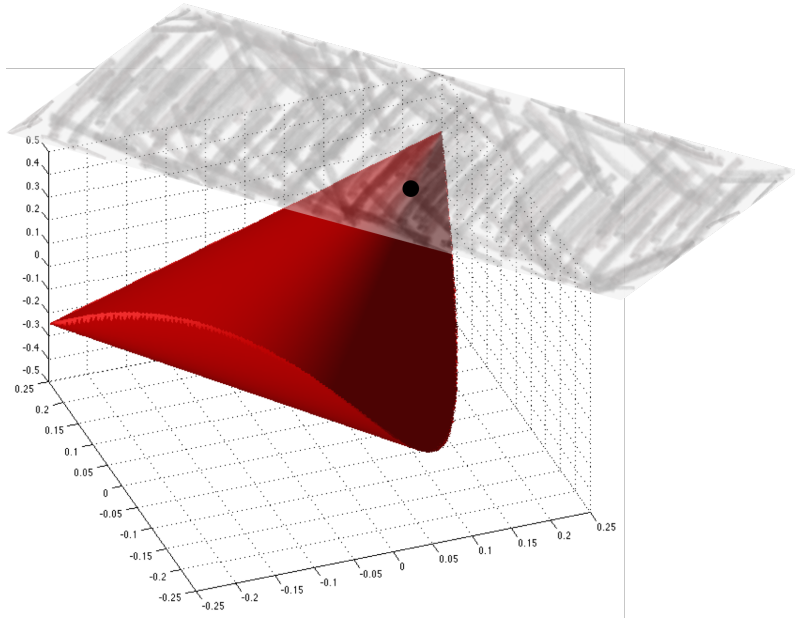
Stability of convex relaxation against noise

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



low-rank matrix

figure credit: Piet Mondrian



semidefinite relaxation

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)
- ? 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

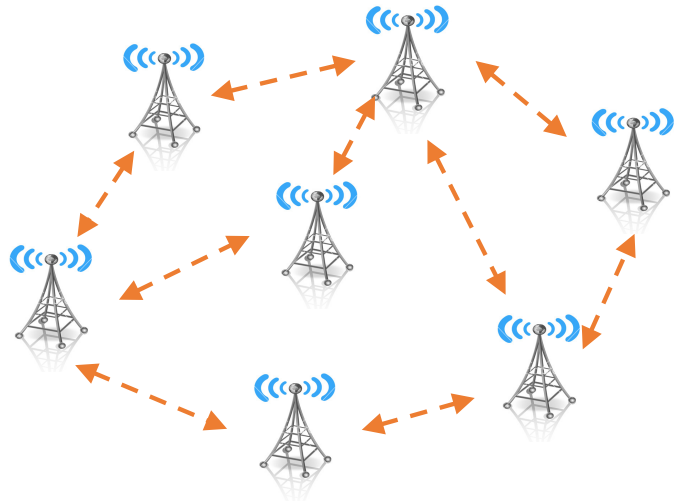
...

Low-rank matrix completion

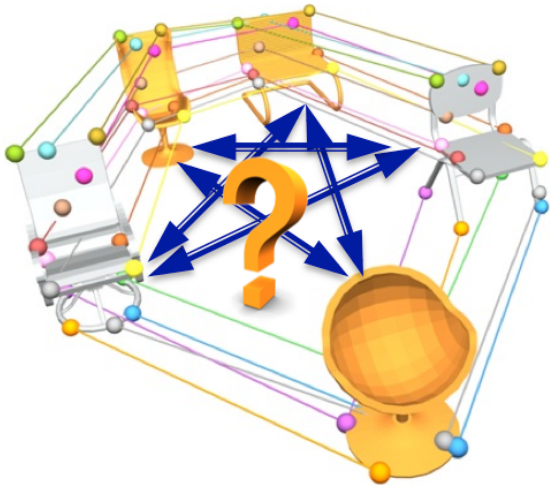
figure credit: E. J. Candès

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

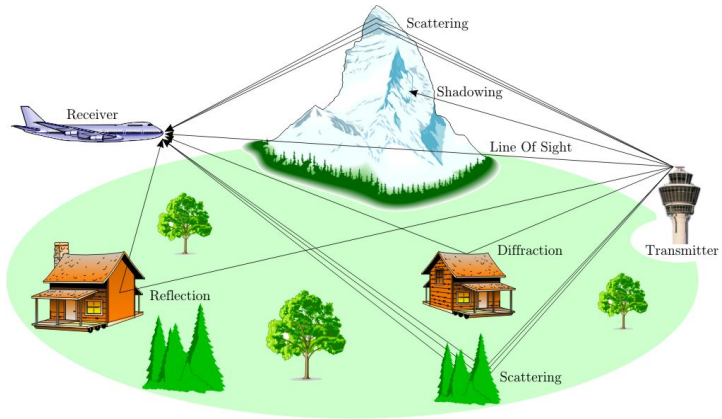
recommendation systems



localization



shape matching

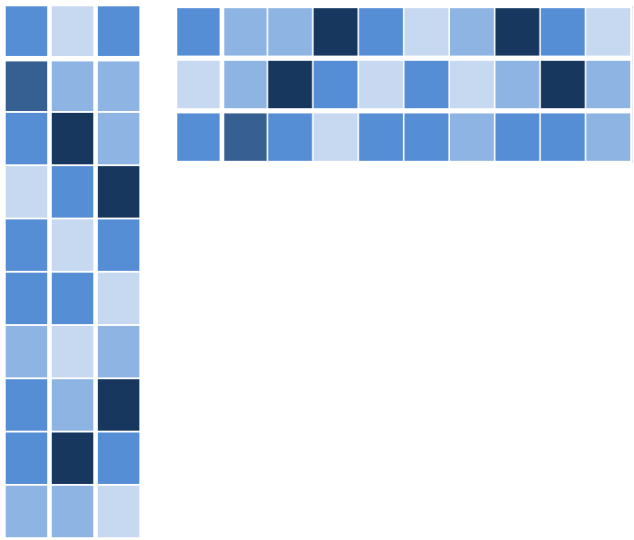


channel estimation

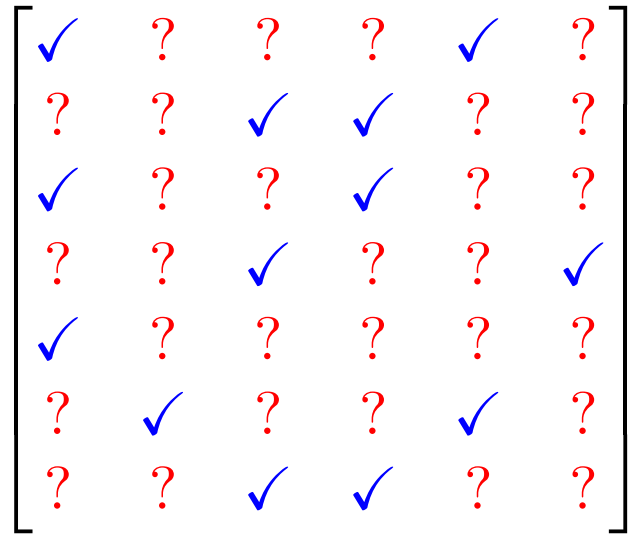
Noisy low-rank matrix completion

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

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



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



sampling set Ω

Noisy low-rank matrix completion

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

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

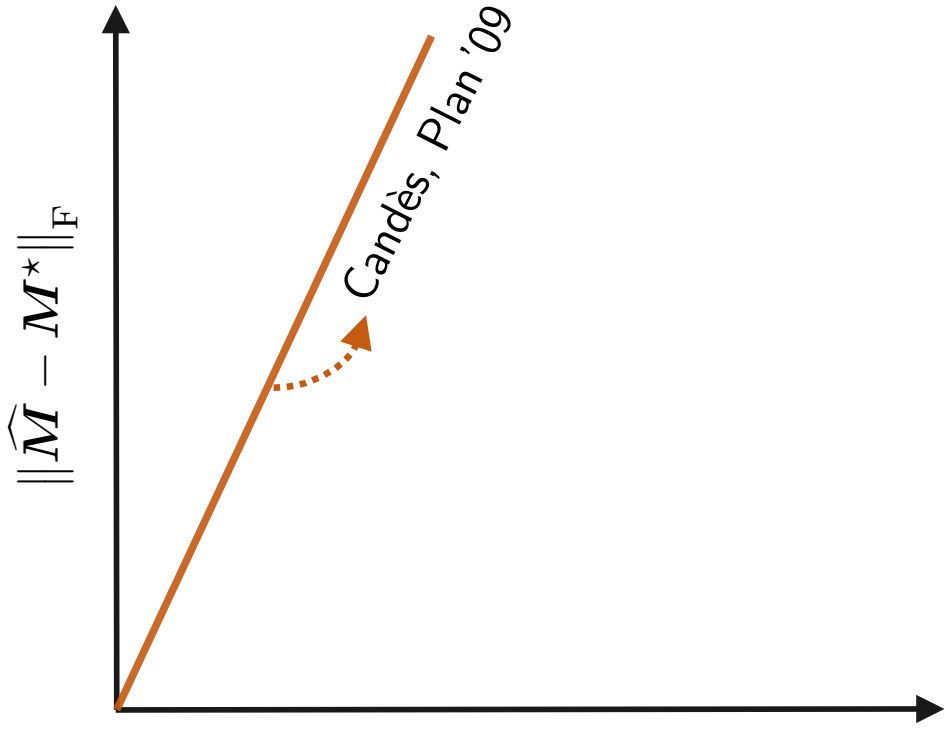
convex relaxation:

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

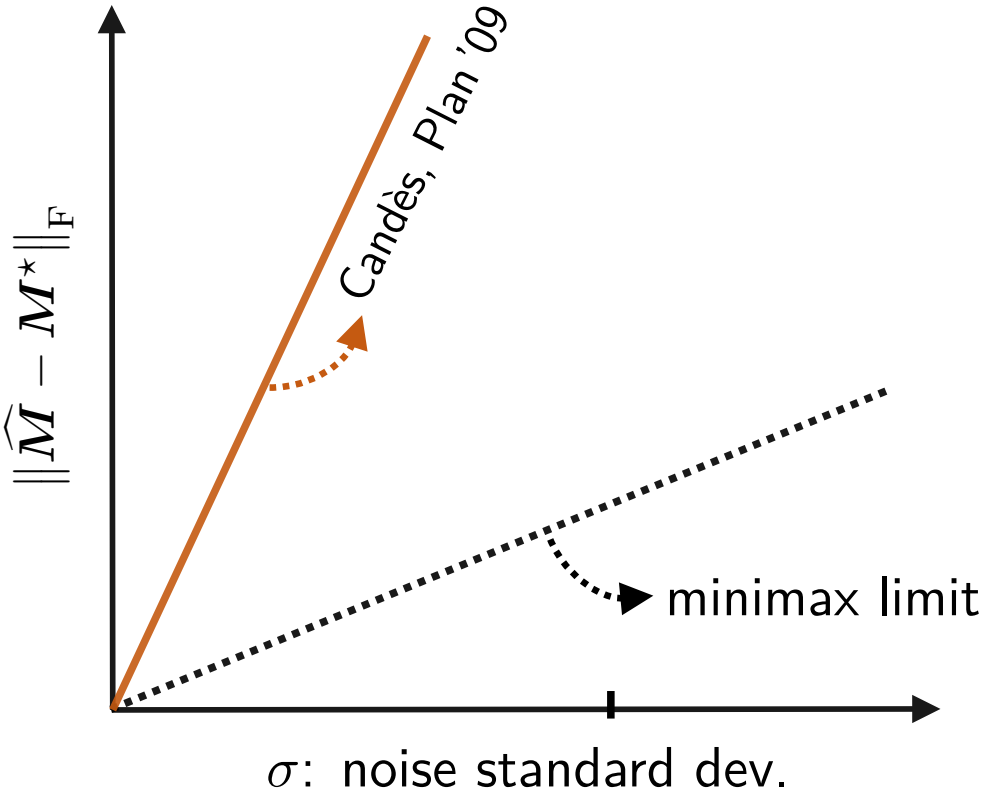
Prior statistical guarantees for convex relaxation

- **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, ...

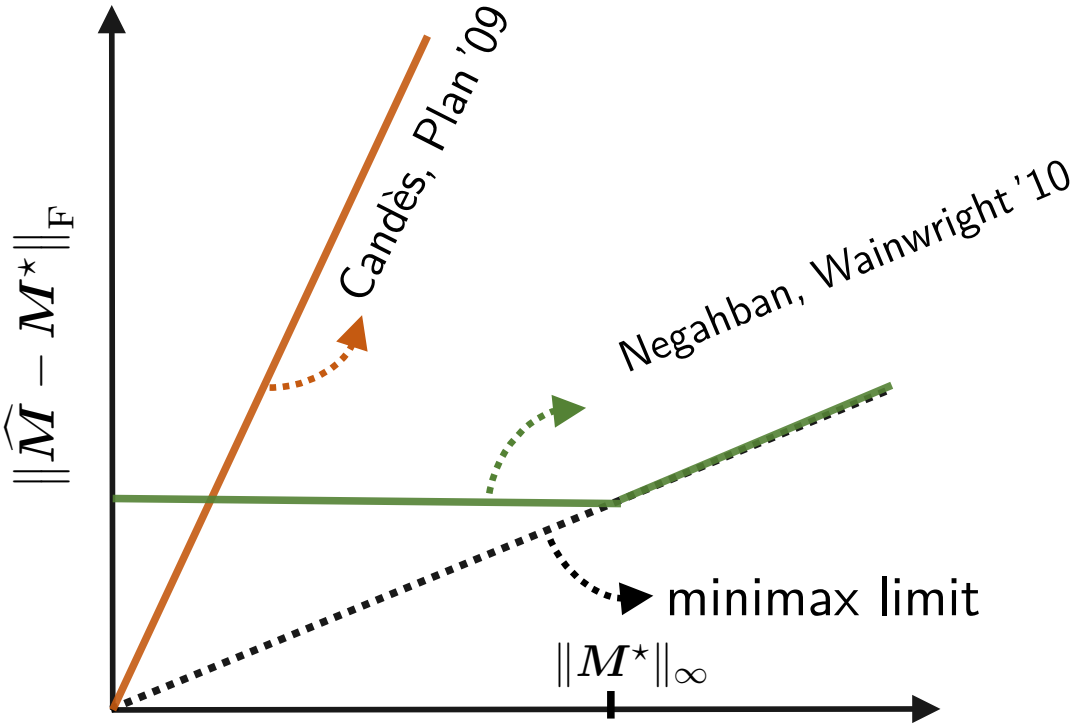
Candès, Plan '09	$\sigma n^{1.5}$



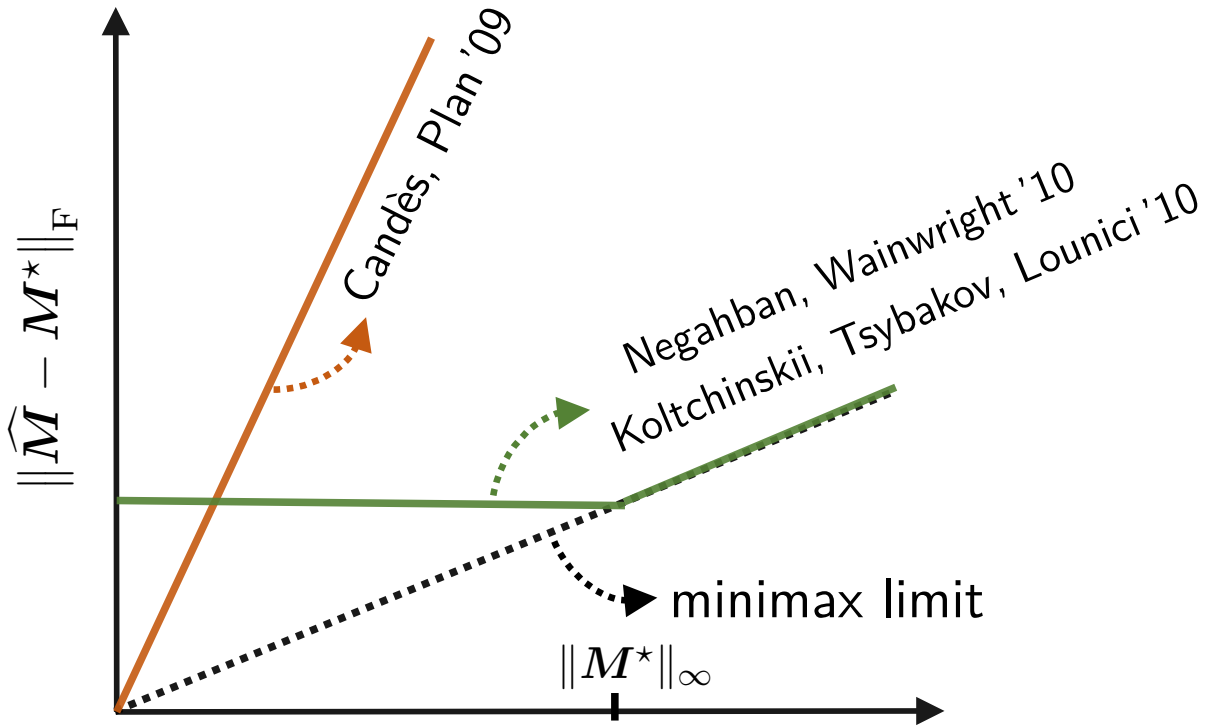
minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$



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

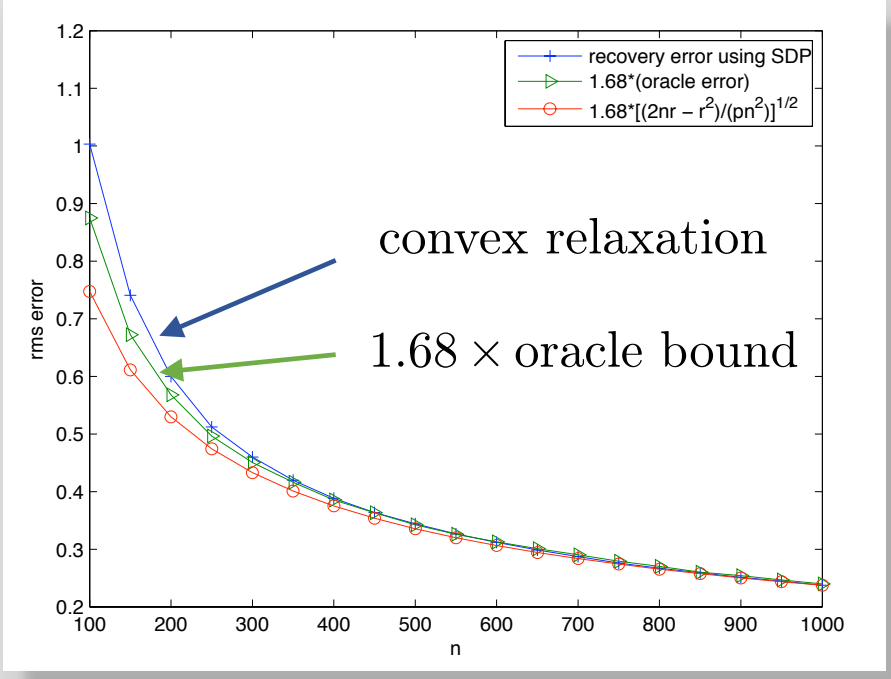


minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$
Negahban, Wainwright '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$
Koltchinskii, Tsybakov, Lounici '10	$\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$



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

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

What are the roadblocks?

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

$(\widehat{\boldsymbol{M}}_{\text{cvx}}, \boldsymbol{W})$ obeys KKT optimality condition

David Gross

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

What are the roadblocks?

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

$(\widehat{\boldsymbol{M}}_{\text{cvx}}, \boldsymbol{W})$ obeys KKT optimality condition

David Gross

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

dual certification (golfing scheme)

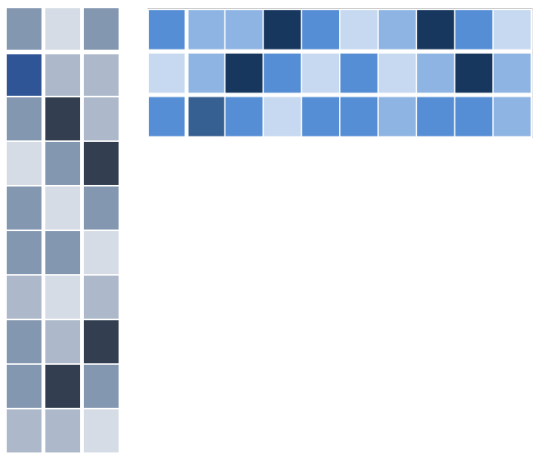
dual certification (golfing scheme)

nonconvex optimization

A detour: nonconvex optimization

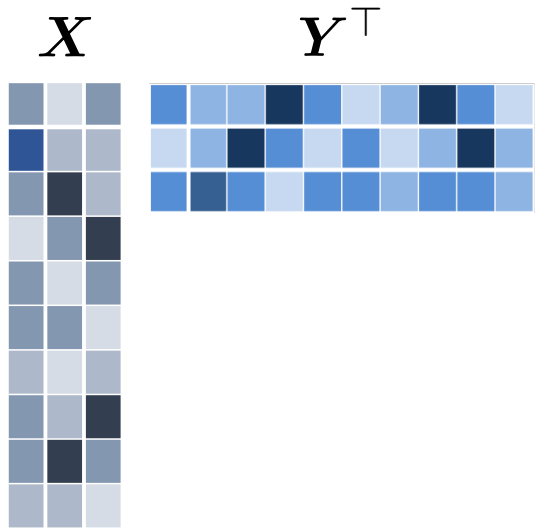
Burer–Monteiro: represent $\boldsymbol{Z}$ by $\boldsymbol{X}\boldsymbol{Y}^\top$ with $\underbrace{\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n \times r}}_{\text{low-rank factors}}$

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



A detour: nonconvex optimization

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



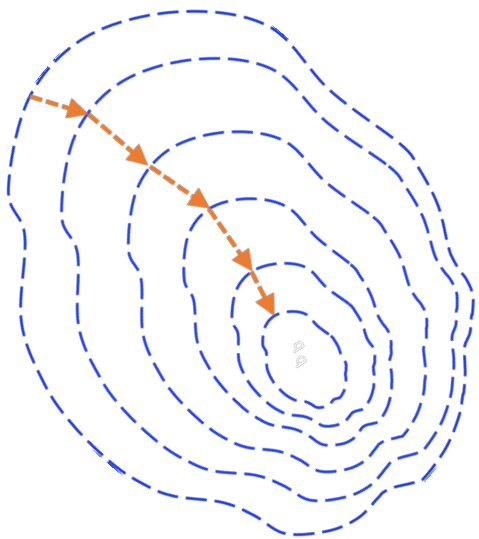
$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \underbrace{\sum_{(i,j)\in\Omega} \left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{i,j} - M_{i,j}\right]^2}_{\text{squared loss}} + \text{reg}(\boldsymbol{X},\boldsymbol{Y})$$

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

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



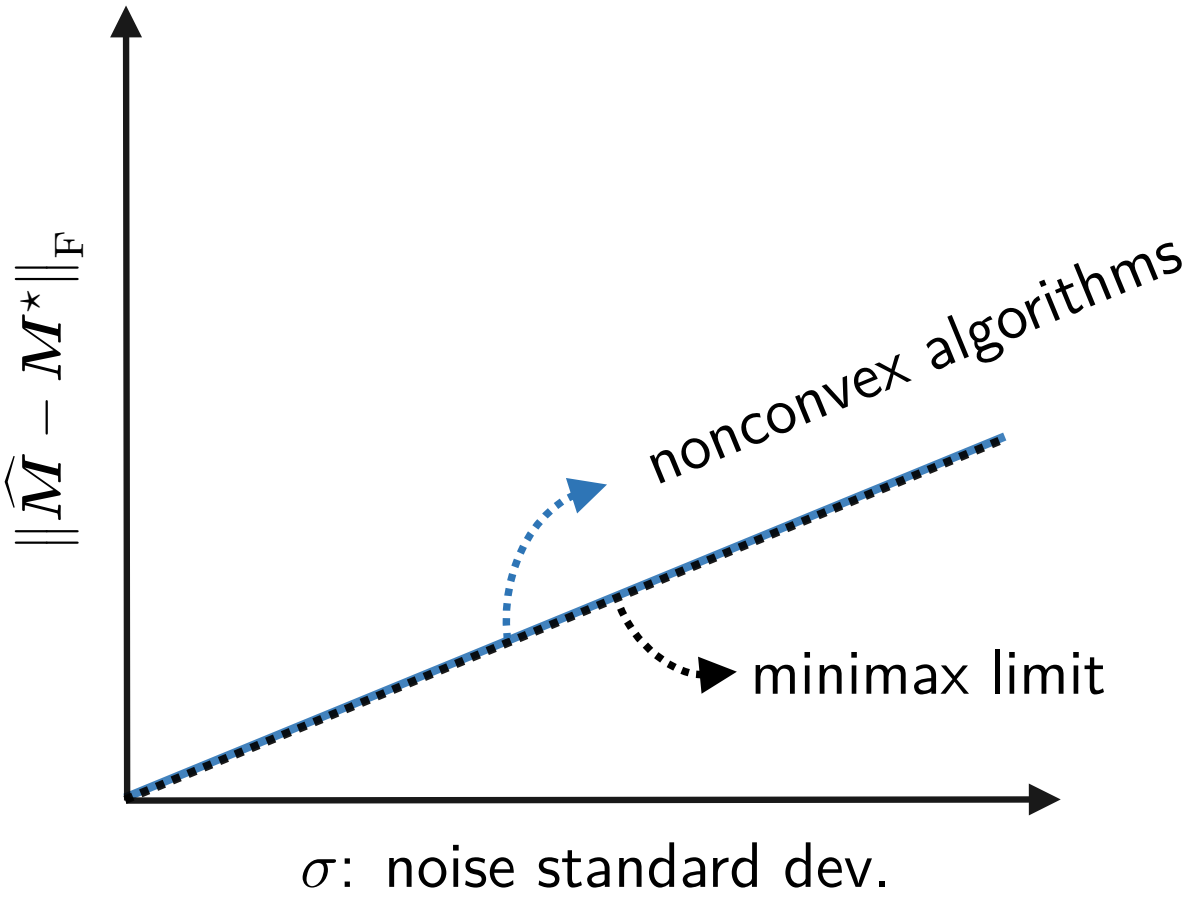
- **suitable initialization:** $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$
- **gradient descent:** for $t = 0, 1, \ldots$

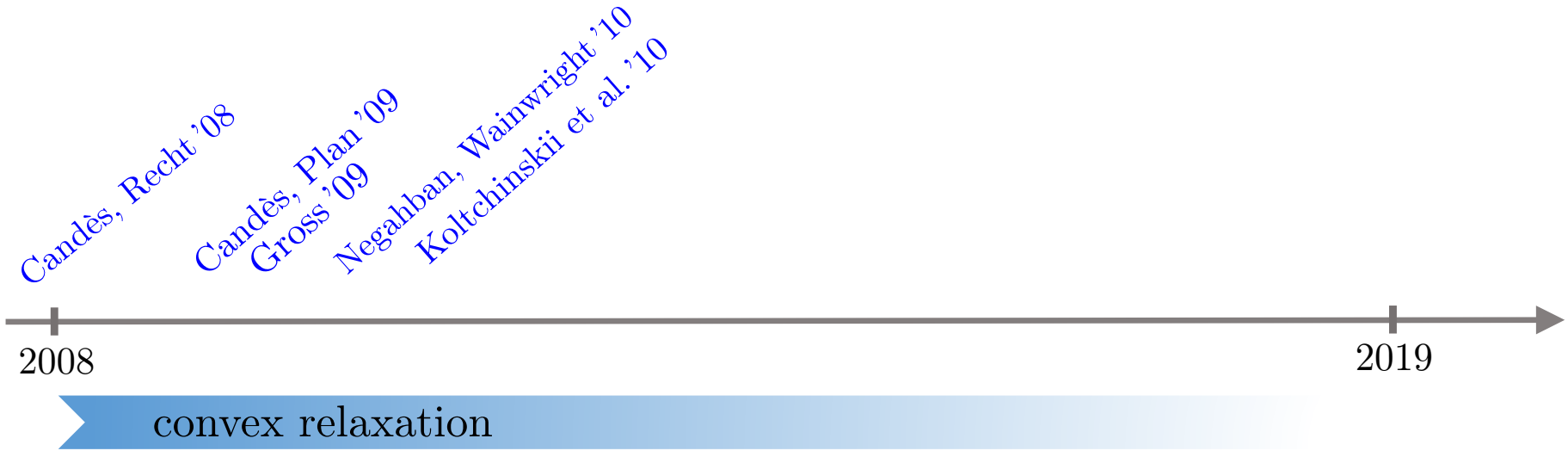
$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$
$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta_t \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$

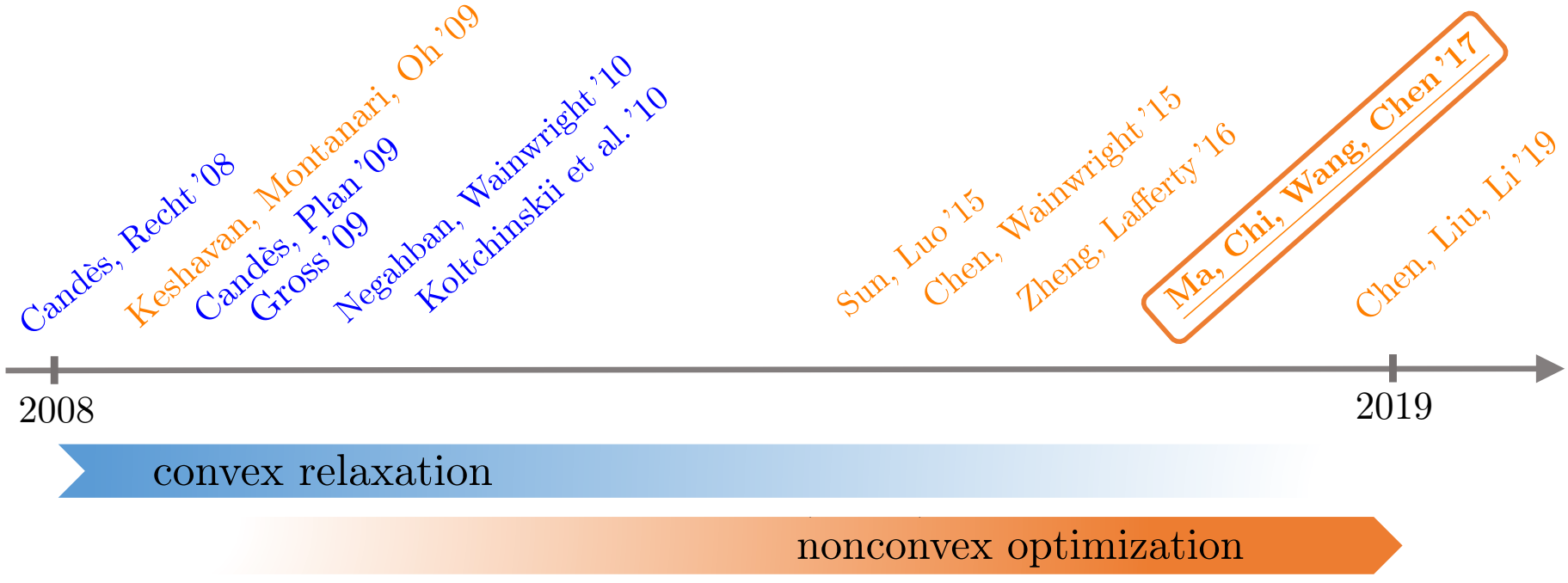
A detour: nonconvex optimization

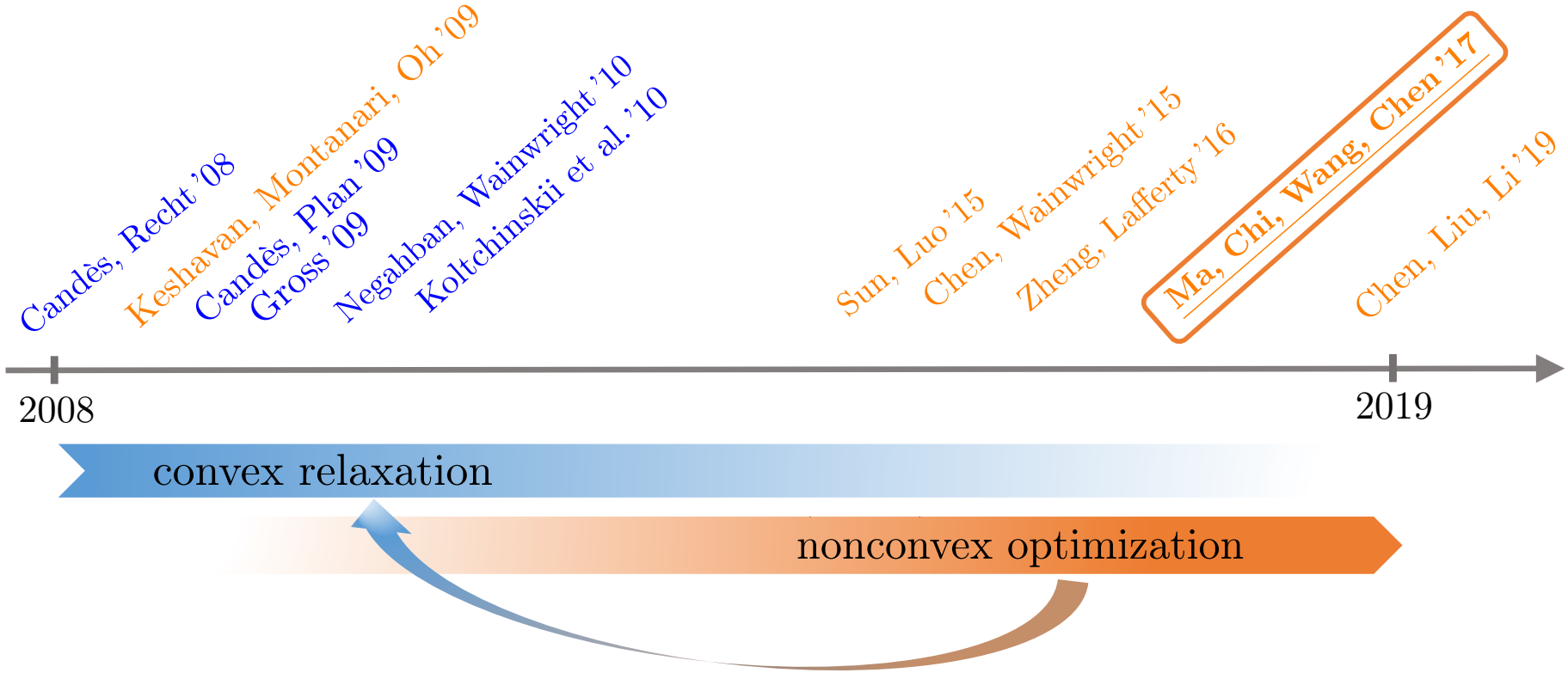
- **random sampling:** each $(i, j) \in \Omega$ with prob. p
- **random noise:** i.i.d. sub-Gaussian noise with variance σ^2
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, ...

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
Ma, Chi, Wang, 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
Ma, Chi, Wang, Chen '17
Chen, Liu, Li '19
2008
2019
convex relaxation
nonconvex optimization

A motivating experiment

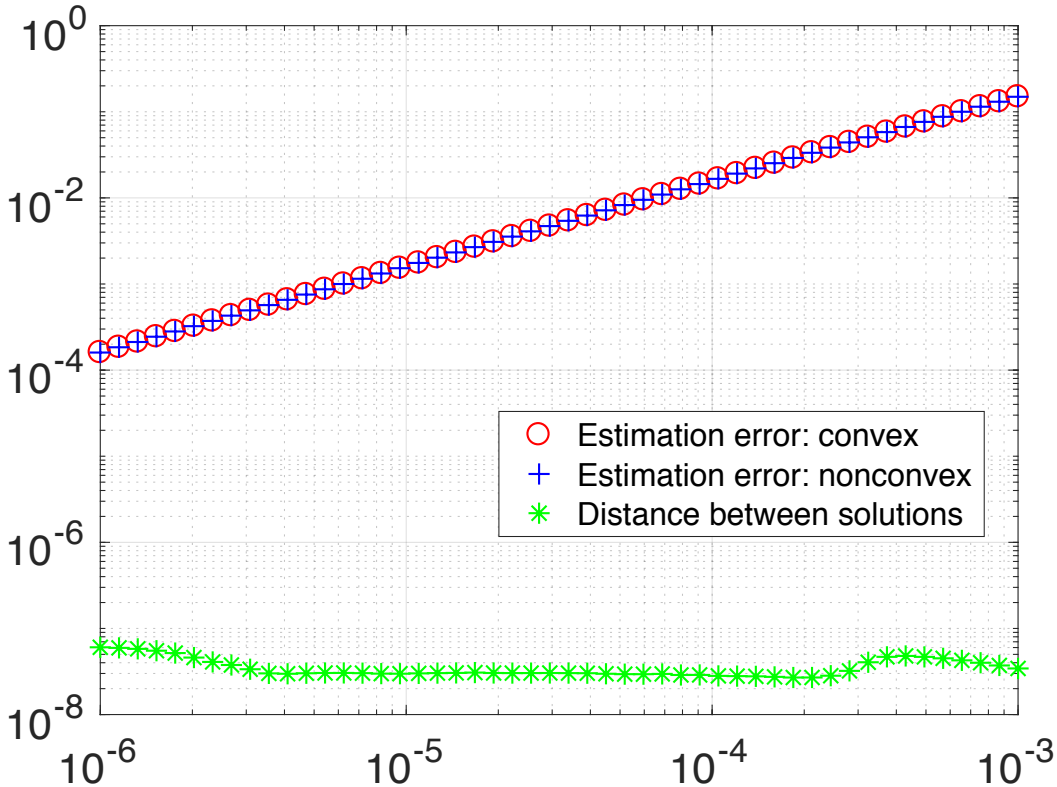
convex: $\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_*$

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

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

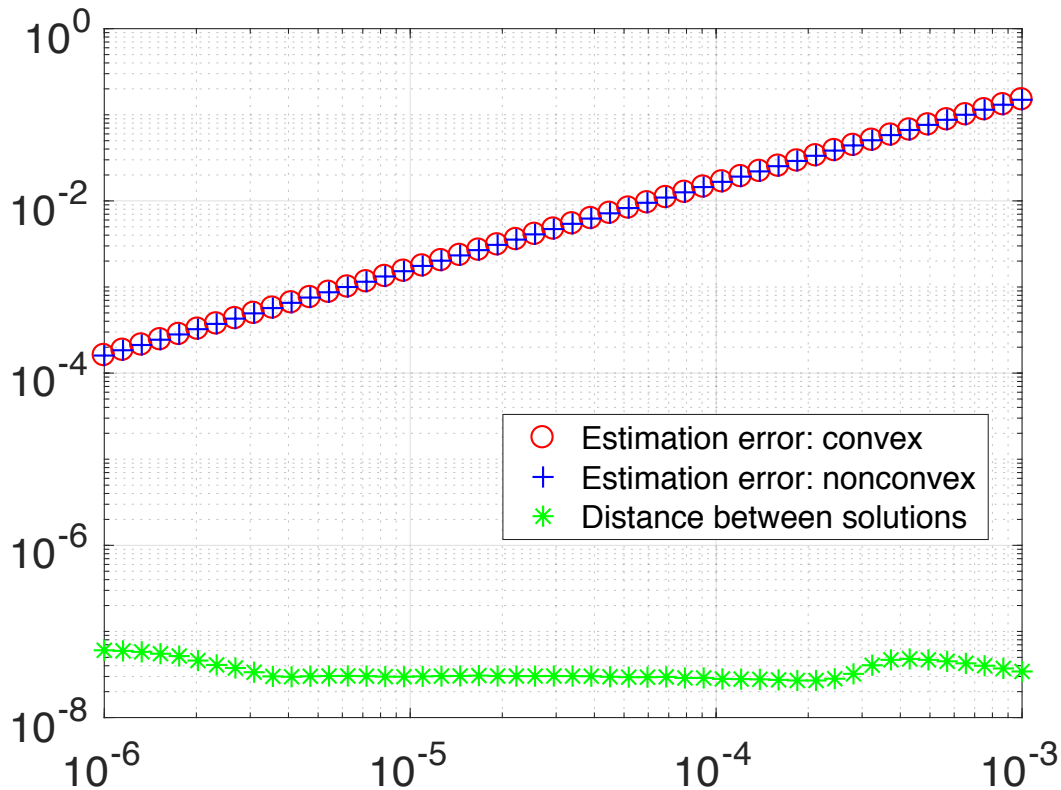
A motivating experiment

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



A motivating experiment

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



Convex and nonconvex solutions are exceedingly close!

convex

nonconvex

$$\Downarrow$$

$$\text{stability}\Big(\text{convex}\Big) \approx \text{stability}\Big(\text{nonconvex}\Big)$$

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 noise with variance σ^2
- 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 noise with variance σ^2
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, well-conditioned

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

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 noise with variance σ^2
- 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 $\widehat{\boldsymbol{M}}_{\mathsf{cvx}}$ of convex program obeys

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

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

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 noise with variance σ^2
- 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 $\widehat{\boldsymbol{M}}_{\mathsf{cvx}}$ of convex program obeys

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

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 noise with variance σ^2
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

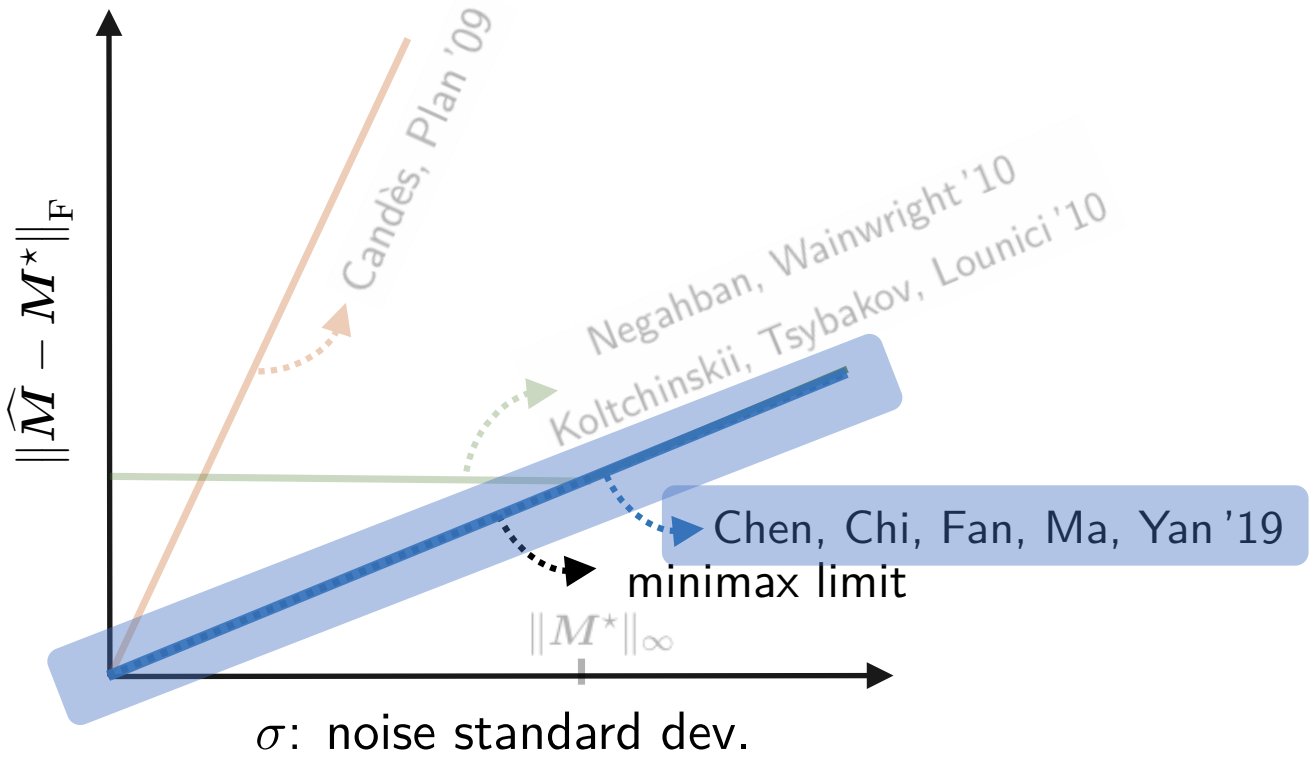
$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

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

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

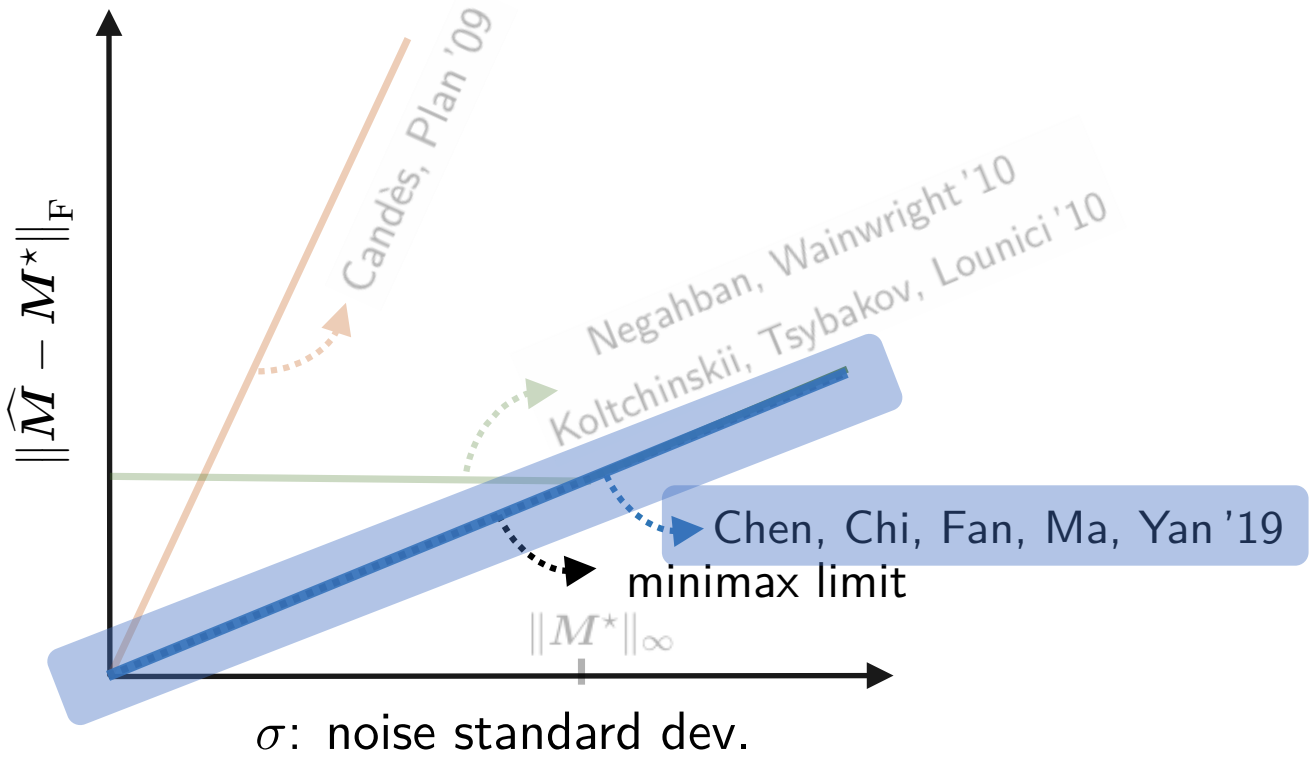
1. *$\widehat{\boldsymbol{M}}_{\mathsf{cvx}}$ is nearly rank-r*
2. $$\|\widehat{\boldsymbol{M}}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}$$
$$\|\widehat{\boldsymbol{M}}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\infty} \lesssim \sigma\sqrt{\frac{n\log n}{p}}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

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



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

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



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

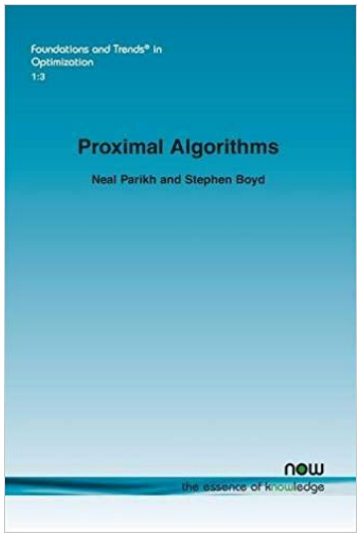
Implicit regularization

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

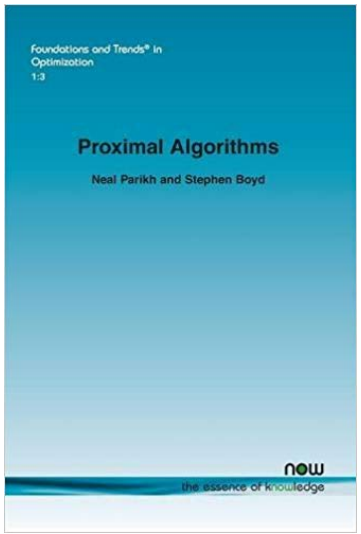
Statistical guarantees for iterative algorithms



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

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

Statistical guarantees for iterative algorithms



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

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 $g(\boldsymbol{Z}) \leq g(\boldsymbol{Z}_{\text{opt}}) + \varepsilon$ 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 noise with variance σ^2
- 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 noise with variance σ^2
- 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 $\widehat{\boldsymbol{M}}_{\mathsf{cvx}}$ of convex program obeys

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

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 noise with variance σ^2
- 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

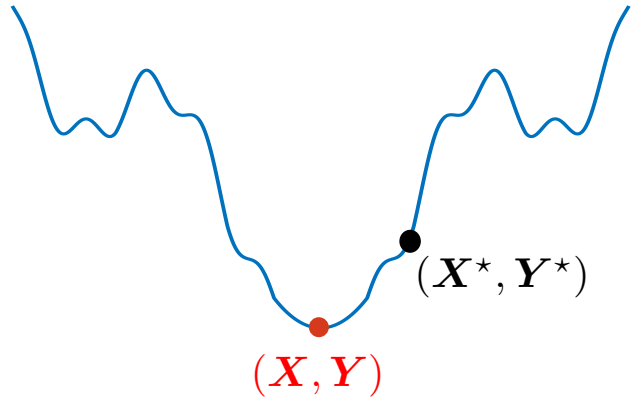
$(\boldsymbol{X}, \boldsymbol{Y})$ is nonconvex optimizer

Link between convex and nonconvex optimizers

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

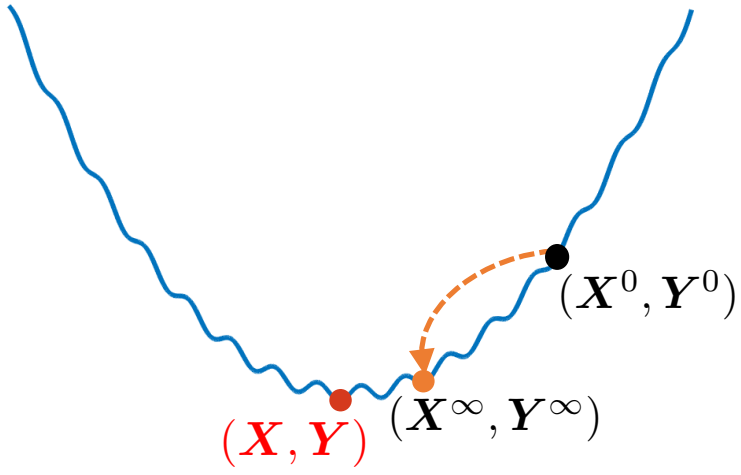
Link between convex and nonconvex optimizers

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



$(\boldsymbol{X}, \boldsymbol{Y})$ is nonconvex optimizer $\overset{\checkmark}{\Longrightarrow}$ $\boldsymbol{X}\boldsymbol{Y}^\top$ 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 "$\underbrace{\text{approximate stationary points}}_{\nabla f(\boldsymbol{X},\boldsymbol{Y})\approx \mathbf{0}}$" 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 "approximate stationary points" instead

$$\nabla f(\boldsymbol{X},\boldsymbol{Y}) \approx \boldsymbol{0}$$

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 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

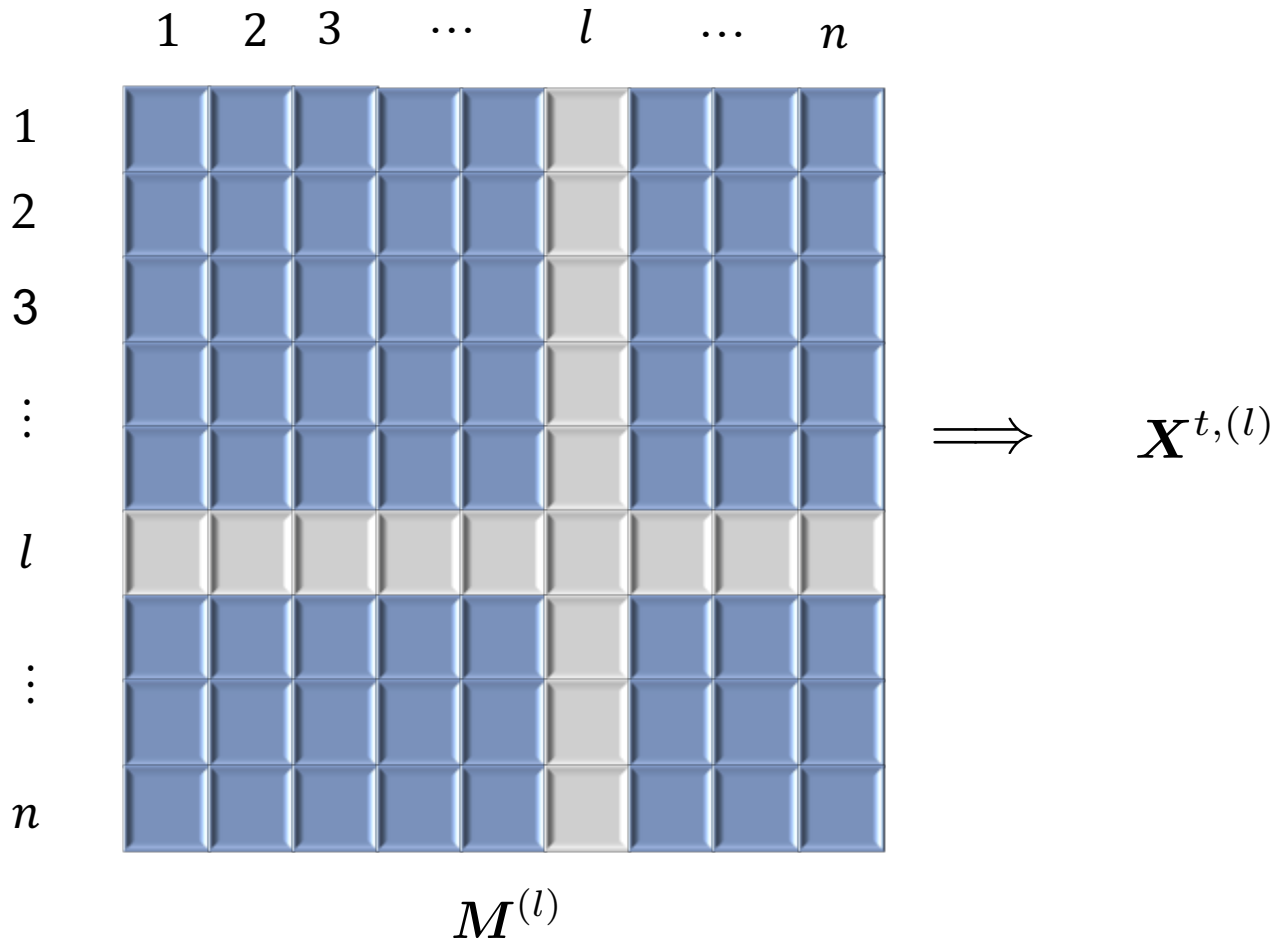
Analyzing nonconvex GD: leave-one-out analysis

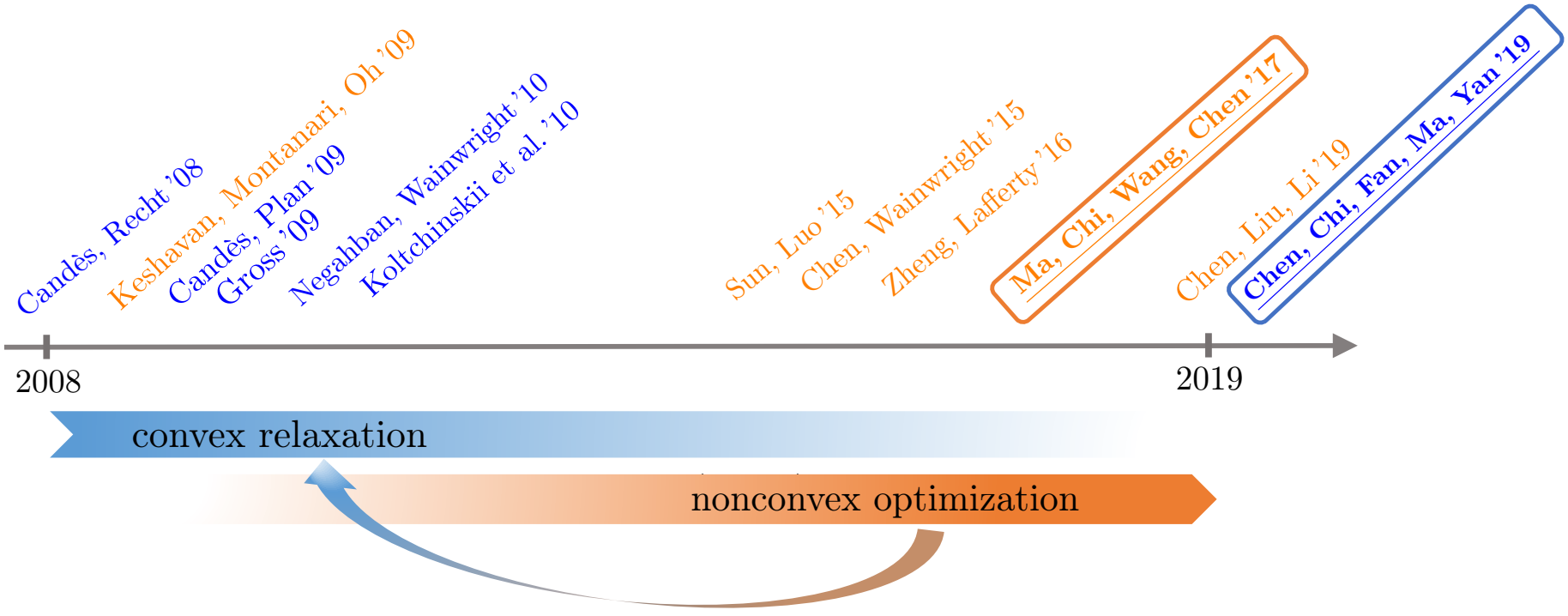
Leave out a small amount of information from data and run GD

- 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





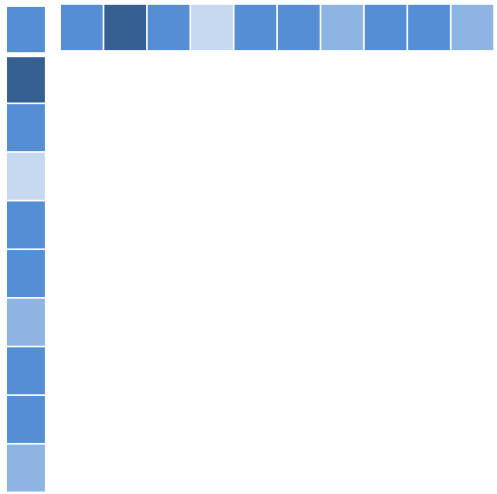
"Noisy matrix completion: understanding statistical guarantees for convex relaxation via nonconvex optimization", Y. Chen, Y. Chi, J. Fan, C. Ma, Y. Yan, 2019

Asymmetry helps: eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices

Chen Cheng
PKU Math & Stanford Stats

Jianqing Fan
Princeton ORFE

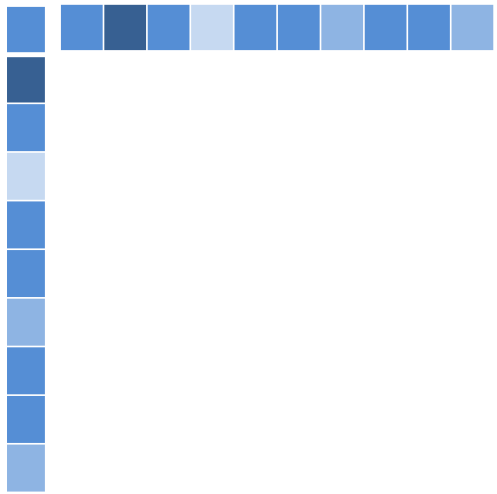
Eigenvalue / eigenvector estimation



$\boldsymbol{M}^\star$: truth

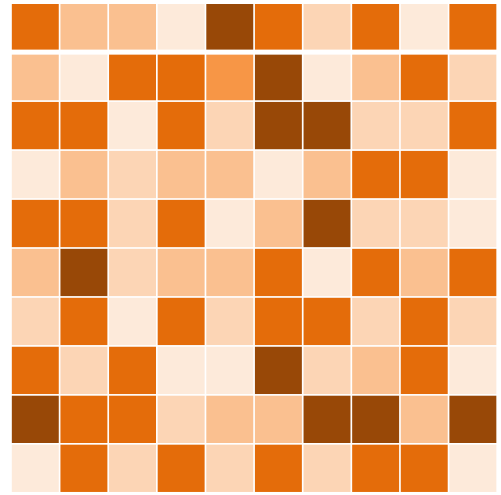
- A rank-1 matrix: $\boldsymbol{M}^\star = \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top} \in \mathbb{R}^{n\times n}$

Eigenvalue / eigenvector estimation



$\boldsymbol{M}^\star$: truth

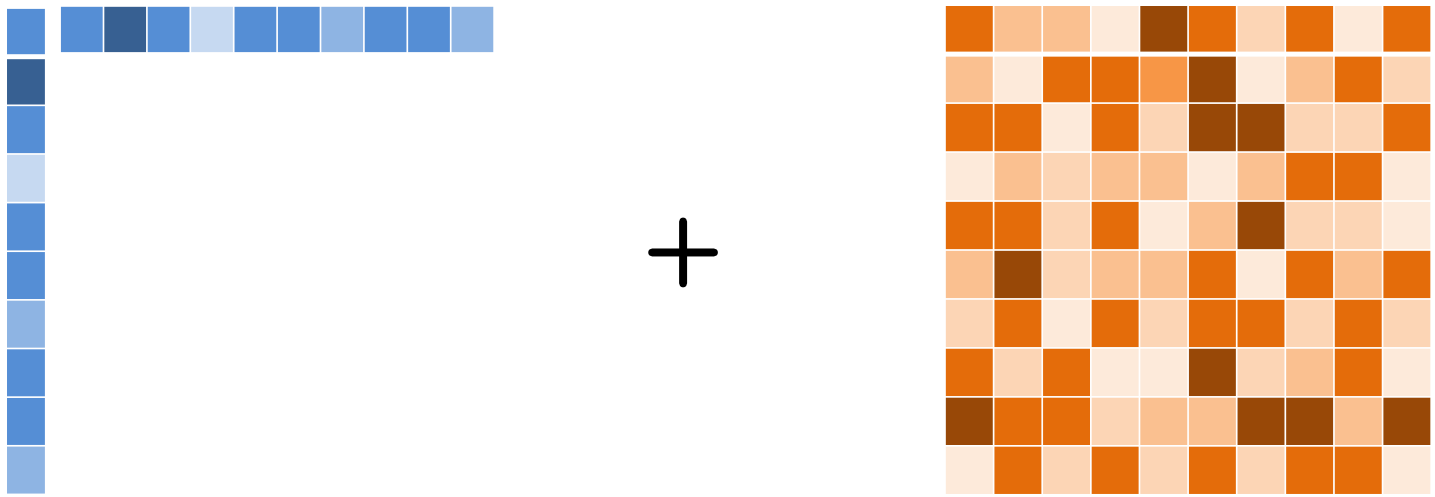
+



$\boldsymbol{H}$: noise

- A rank-1 matrix: $\boldsymbol{M}^\star = \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top} \in \mathbb{R}^{n\times n}$
- Observed noisy data: $\boldsymbol{M} = \boldsymbol{M}^\star + \boldsymbol{H}$

Eigenvalue / eigenvector estimation



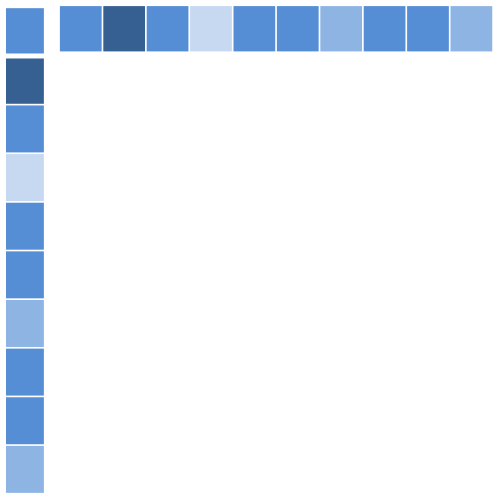
$\boldsymbol{M}^\star$: truth

$\boldsymbol{H}$: noise

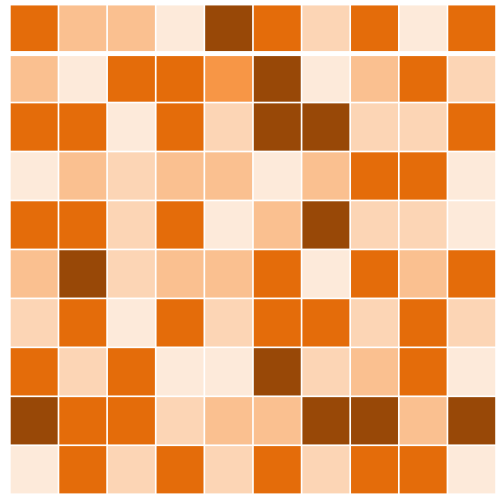
- A rank-1 matrix: $\boldsymbol{M}^\star = \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top} \in \mathbb{R}^{n\times n}$
- Observed noisy data: $\boldsymbol{M} = \boldsymbol{M}^\star + \boldsymbol{H}$
- **Goal:** estimate eigenvalue $\lambda^\star$ and eigenvector $\boldsymbol{u}^\star$

Non-symmetric noise matrix

$$\boldsymbol{M} = \boldsymbol{M}^\star + \boldsymbol{H}$$



$\boldsymbol{M}^\star = \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top}$

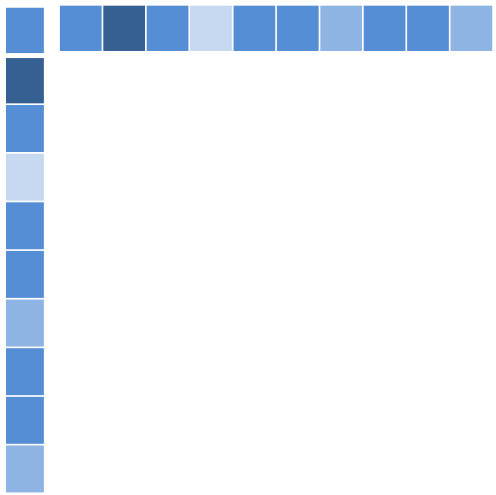


$\boldsymbol{H}$: *asymmetric* matrix

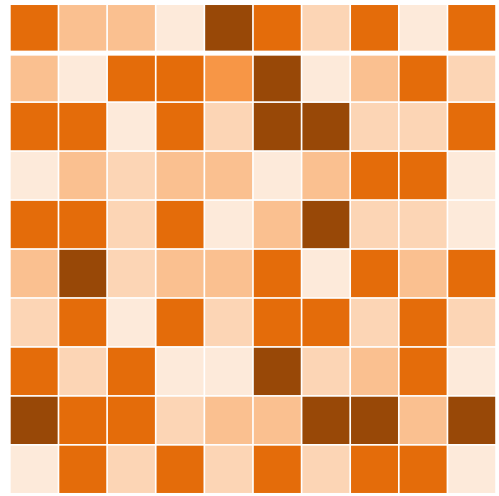
This may arise when, e.g., we have 2 samples for each entry of $\boldsymbol{M}^\star$ and arrange them in an asymmetric manner

A natural estimation strategy: SVD

$$\boldsymbol{M} = \quad + \quad \boldsymbol{H}\text{: } \textit{asymmetric} \text{ matrix}$$



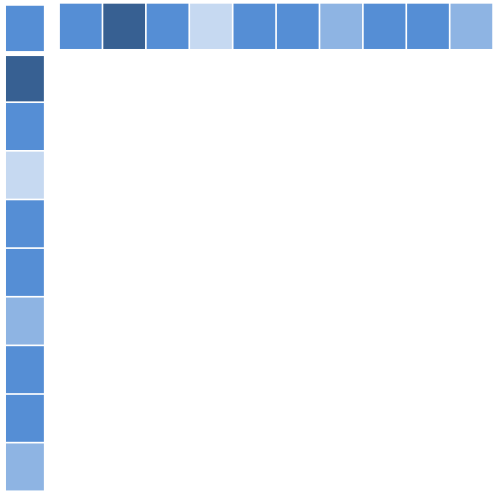
$$\boldsymbol{M}^\star = \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top}$$



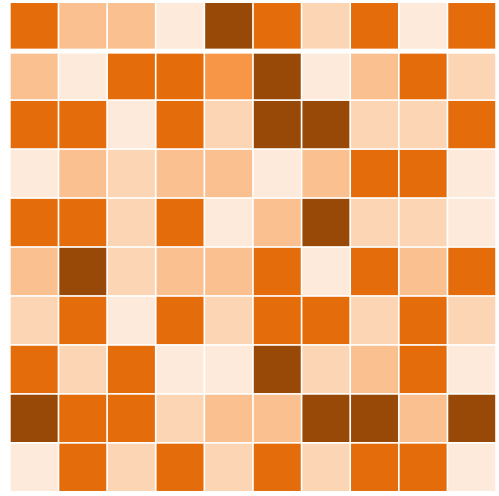
- Use leading singular value λ^{svd} of $\boldsymbol{M}$ to estimate $\lambda^\star$
- Use leading left singular vector of $\boldsymbol{M}$ to estimate $\boldsymbol{u}^\star$

A less popular strategy: eigen-decomposition

$$\boldsymbol{M} = \quad + \quad \boldsymbol{H}\text{: } \textit{asymmetric} \text{ matrix}$$



$$\boldsymbol{M}^\star = \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top}$$



- Use leading ~~singular value λ^{svd}~~ eigenvalue λ^{eigs} of $\boldsymbol{M}$ to estimate $\lambda^\star$
- Use leading ~~singular vector~~ eigenvector of $\boldsymbol{M}$ to estimate $\boldsymbol{u}^\star$

SVD vs. eigen-decomposition

For *asymmetric* matrices:

- Numerical stability

$$\text{SVD} \quad > \quad \text{eigen-decomposition}$$

SVD vs. eigen-decomposition

For *asymmetric* matrices:

- Numerical stability

$$\text{SVD} \quad > \quad \text{eigen-decomposition}$$

- **(Folklore?)** Statistical accuracy

$$\text{SVD} \quad \asymp \quad \text{eigen-decomposition}$$

SVD vs. eigen-decomposition

For *asymmetric* matrices:

- Numerical stability

$$\text{SVD} \quad > \quad \text{eigen-decomposition}$$

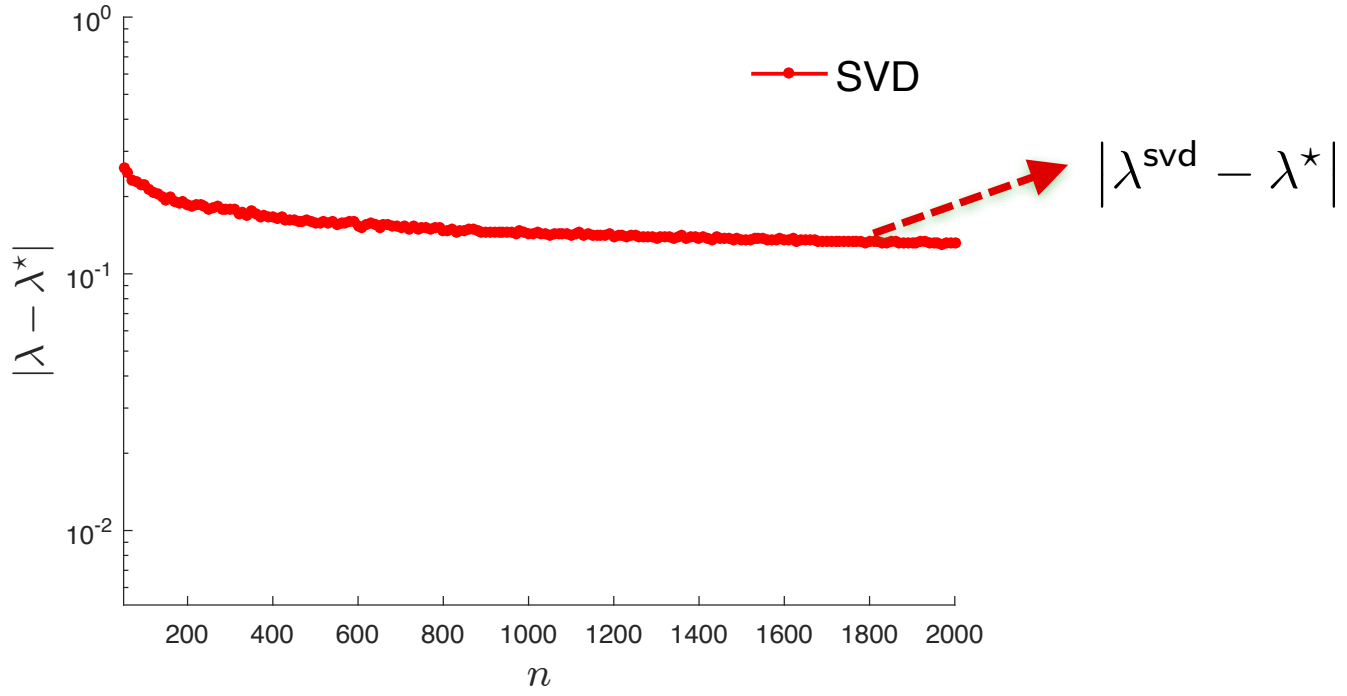
- **(Folklore?)** Statistical accuracy

$$\text{SVD} \quad \asymp \quad \text{eigen-decomposition}$$

Shall we always prefer SVD over eigen-decomposition?

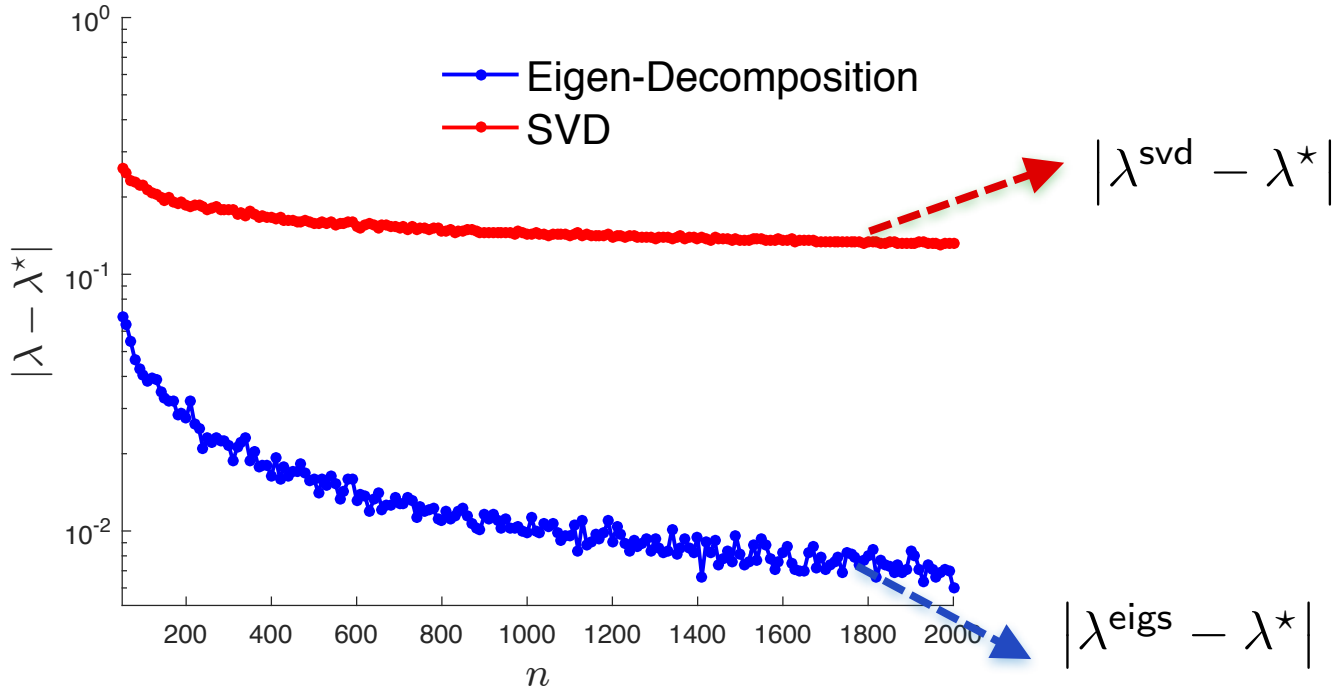
A curious numerical experiment: Gaussian noise

$$\boldsymbol{M} = \underbrace{\boldsymbol{u}^\star \boldsymbol{u}^{\star\top}}_{\boldsymbol{M}^\star} + \boldsymbol{H}; \qquad \{H_{i,j}\}: \text{ i.i.d. } \mathcal{N}(0, \sigma^2), \sigma = \frac{1}{\sqrt{n \log n}}$$



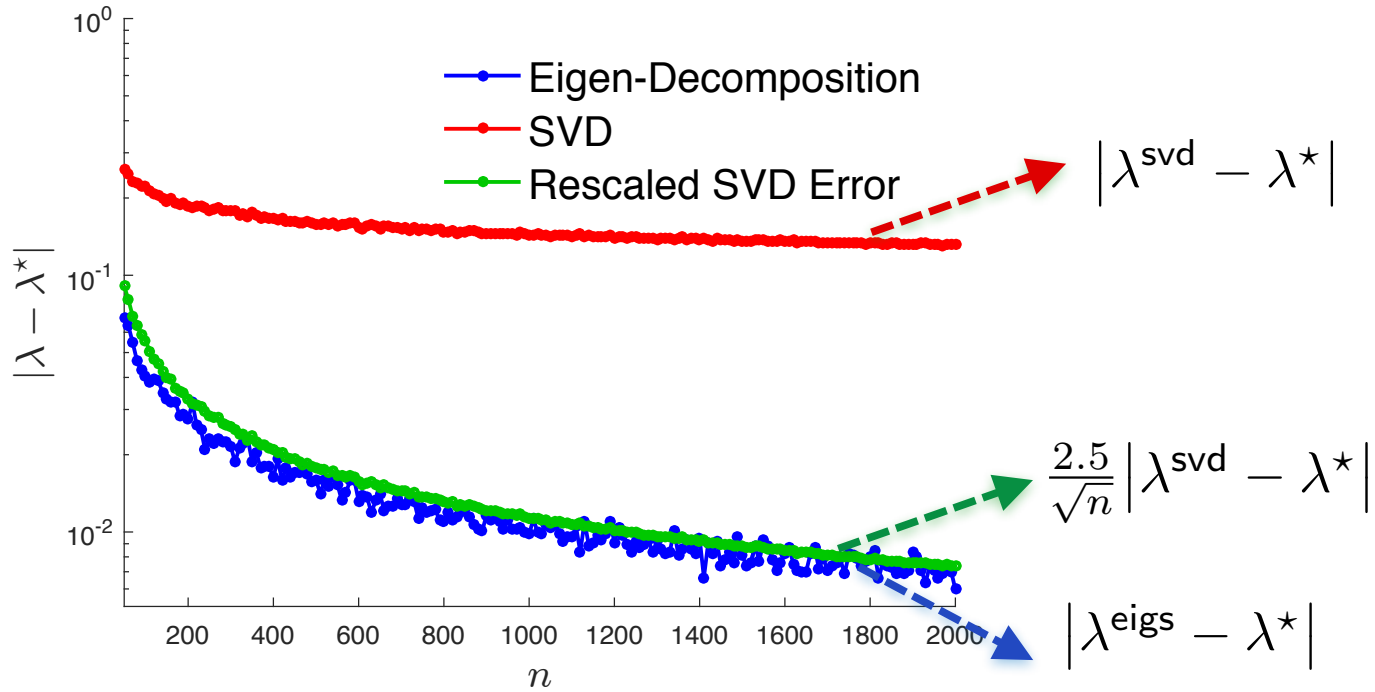
A curious numerical experiment: Gaussian noise

$$\boldsymbol{M} = \underbrace{\boldsymbol{u}^\star \boldsymbol{u}^{\star\top}}_{\boldsymbol{M}^\star} + \boldsymbol{H}; \qquad \{H_{i,j}\}: \text{ i.i.d. } \mathcal{N}(0, \sigma^2), \sigma = \frac{1}{\sqrt{n \log n}}$$



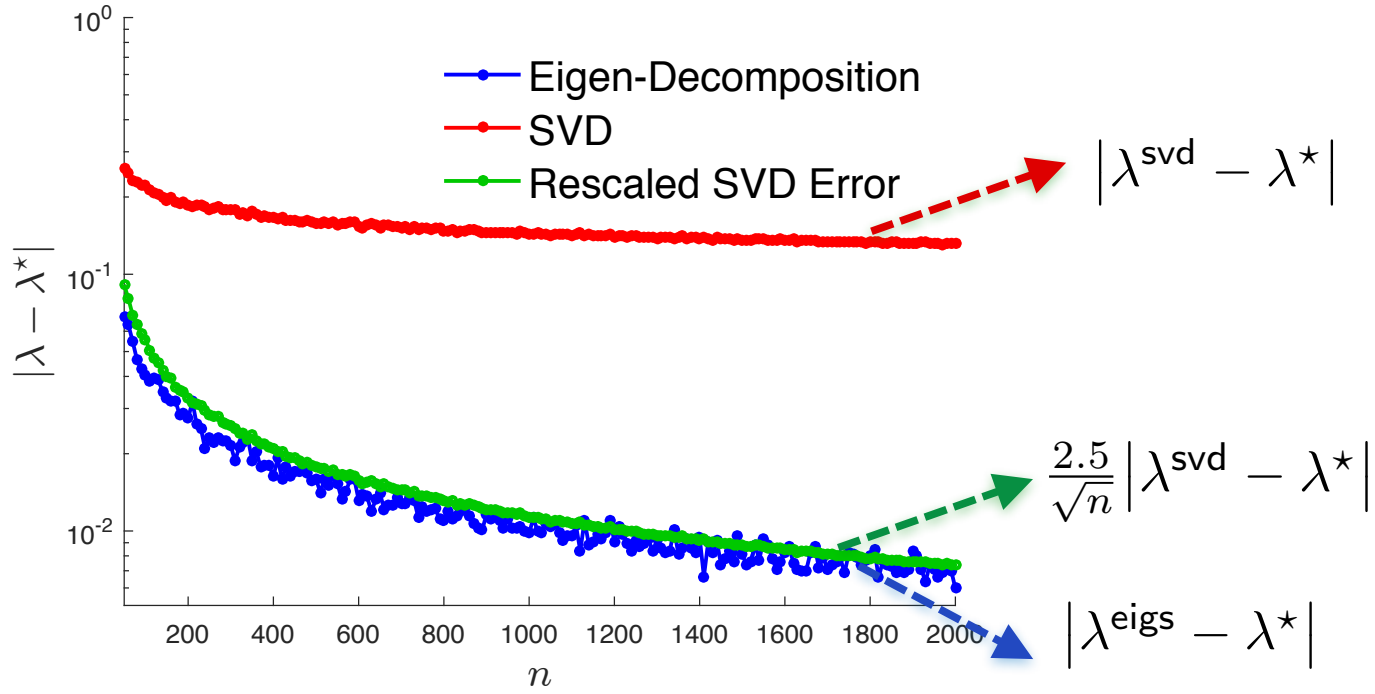
A curious numerical experiment: Gaussian noise

$$\boldsymbol{M} = \underbrace{\boldsymbol{u}^\star \boldsymbol{u}^{\star\top}}_{\boldsymbol{M}^\star} + \boldsymbol{H}; \qquad \{H_{i,j}\}: \text{ i.i.d. } \mathcal{N}(0, \sigma^2), \sigma = \frac{1}{\sqrt{n \log n}}$$



A curious numerical experiment: Gaussian noise

$$\boldsymbol{M} = \underbrace{\boldsymbol{u}^\star \boldsymbol{u}^{\star\top}}_{\boldsymbol{M}^\star} + \boldsymbol{H}; \qquad \{H_{i,j}\}: \text{ i.i.d. } \mathcal{N}(0, \sigma^2), \sigma = \frac{1}{\sqrt{n \log n}}$$



empirically, $\left|\lambda^{\text{eigs}} - \lambda^\star\right| \approx \frac{2.5}{\sqrt{n}}\left|\lambda^{\text{svd}} - \lambda^\star\right|$

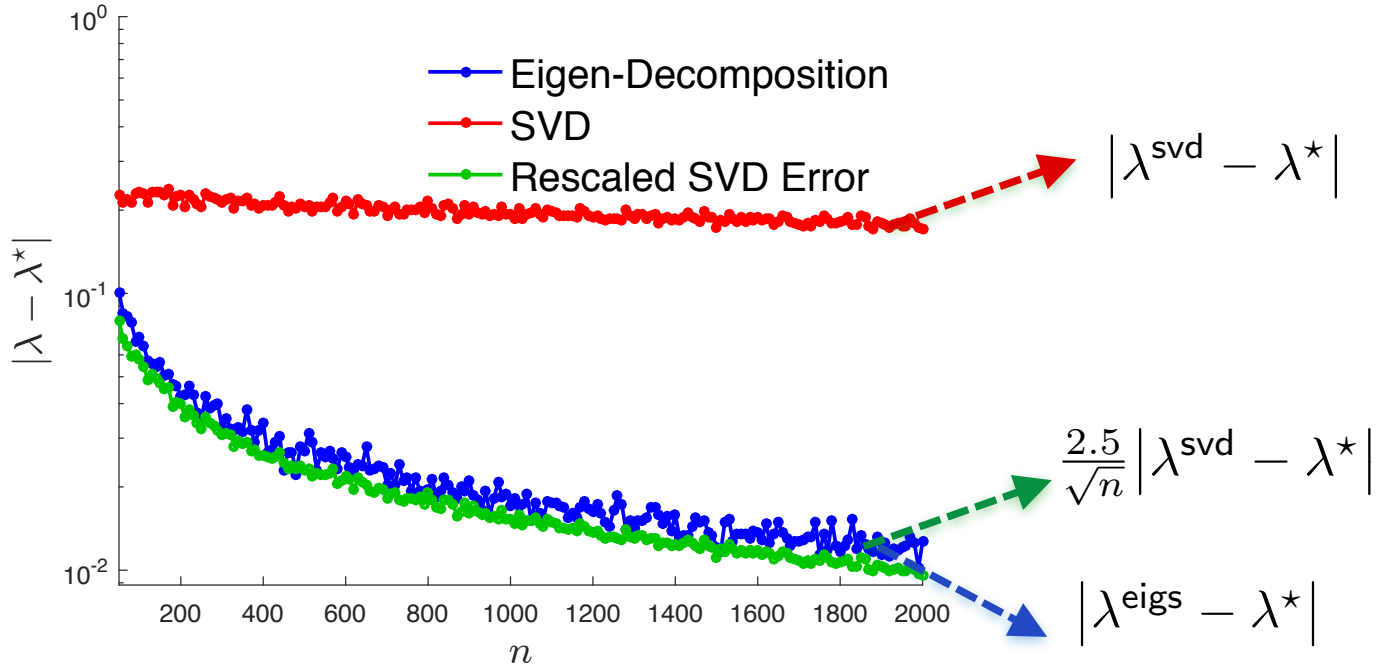
Another numerical experiment: matrix completion

$$\boldsymbol{M}^\star = \boldsymbol{u}^\star \boldsymbol{u}^{\star\top}; \qquad M_{i,j} = \begin{cases} \frac{1}{p} M^\star_{i,j} & \text{with prob. } p, \\ 0, & \text{else,} \end{cases} \qquad p = \frac{3\log n}{n}$$

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

Another numerical experiment: matrix completion

$$\boldsymbol{M}^\star = \boldsymbol{u}^\star \boldsymbol{u}^{\star\top}; \qquad M_{i,j} = \begin{cases} \frac{1}{p} M^\star_{i,j} & \text{with prob. } p, \\ 0, & \text{else,} \end{cases} \qquad p = \frac{3\log n}{n}$$



empirically, $|\lambda^{\text{eigs}} - \lambda^\star| \approx \frac{2.5}{\sqrt{n}}|\lambda^{\text{svd}} - \lambda^\star|$

Why does eigen-decomposition work so much better than SVD?

Problem setup

$$\boldsymbol{M} = \underbrace{\boldsymbol{u}^\star \boldsymbol{u}^{\star\top}}_{\boldsymbol{M}^\star} + \boldsymbol{H} \in \mathbb{R}^{n \times n}$$

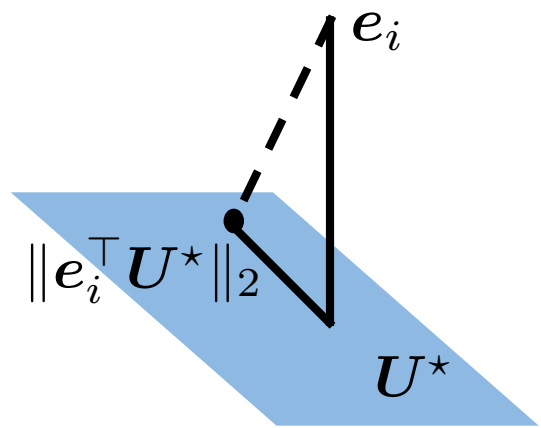
- $\boldsymbol{H}$: noise matrix
  - **independent entries:** $\{H_{i,j}\}$ are independent
  - **zero mean:** $\mathbb{E}[H_{i,j}] = 0$
  - **variance:** $\mathsf{Var}(H_{i,j}) \le \sigma^2$
  - **magnitudes:** $\mathbb{P}\{|H_{i,j}| \ge B\} \lesssim n^{-12}$

Problem setup

$$\boldsymbol{M} = \underbrace{\boldsymbol{u}^\star \boldsymbol{u}^{\star\top}}_{\boldsymbol{M}^\star} + \boldsymbol{H} \in \mathbb{R}^{n\times n}$$

- $\boldsymbol{H}$: noise matrix
  - **independent entries:** $\{H_{i,j}\}$ are independent
  - **zero mean:** $\mathbb{E}[H_{i,j}] = 0$
  - **variance:** $\mathsf{Var}(H_{i,j}) \le \sigma^2$
  - **magnitudes:** $\mathbb{P}\{|H_{i,j}| \ge B\} \lesssim n^{-12}$
- $\boldsymbol{M}^\star$ obeys incoherence condition

$$\max_{1\le i\le n} \left|\boldsymbol{e}_i^\top \boldsymbol{u}^\star\right| \le \sqrt{\frac{\mu}{n}}$$



Classical linear algebra results

$$\left|\lambda^{\text{svd}} - \lambda^{\star}\right| \leq \|\boldsymbol{H}\| \qquad \text{(Weyl)}$$

$$\left|\lambda^{\text{eigs}} - \lambda^{\star}\right| \leq \|\boldsymbol{H}\| \qquad \text{(Bauer-Fike)}$$

Classical linear algebra results

$$\left|\lambda^{\mathsf{svd}} - \lambda^\star\right| \leq \|\boldsymbol{H}\| \qquad \text{(Weyl)}$$
$$\left|\lambda^{\mathsf{eigs}} - \lambda^\star\right| \leq \|\boldsymbol{H}\| \qquad \text{(Bauer-Fike)}$$

$\Downarrow$ matrix Bernstein inequality

$$\left|\lambda^{\mathsf{svd}} - \lambda^\star\right| \lesssim \sigma\sqrt{n\log n} + B\log n$$
$$\left|\lambda^{\mathsf{eigs}} - \lambda^\star\right| \lesssim \sigma\sqrt{n\log n} + B\log n$$

Classical linear algebra results

$$|\lambda^{\text{svd}} - \lambda^\star| \le \|\boldsymbol{H}\| \qquad \text{(Weyl)}$$
$$|\lambda^{\text{eigs}} - \lambda^\star| \le \|\boldsymbol{H}\| \qquad \text{(Bauer-Fike)}$$

$$\Downarrow \quad \text{matrix Bernstein inequality}$$

$$|\lambda^{\text{svd}} - \lambda^\star| \lesssim \sigma\sqrt{n\log n} + B\log n \qquad \text{(reasonably tight if } \|\boldsymbol{H}\| \text{ is large)}$$
$$|\lambda^{\text{eigs}} - \lambda^\star| \lesssim \sigma\sqrt{n\log n} + B\log n$$

Classical linear algebra results

$$|\lambda^{\text{svd}} - \lambda^\star| \le \|\boldsymbol{H}\| \qquad \text{(Weyl)}$$
$$|\lambda^{\text{eigs}} - \lambda^\star| \le \|\boldsymbol{H}\| \qquad \text{(Bauer-Fike)}$$

$\Downarrow$ matrix Bernstein inequality

$$|\lambda^{\text{svd}} - \lambda^\star| \lesssim \sigma\sqrt{n\log n} + B\log n \qquad \text{(reasonably tight if } \|\boldsymbol{H}\| \text{ is large)}$$
$$|\lambda^{\text{eigs}} - \lambda^\star| \lesssim \sigma\sqrt{n\log n} + B\log n \qquad \text{(}\textit{can be significantly improved}\text{)}$$

Main results: eigenvalue perturbation

Theorem 3 (Chen, Cheng, Fan '18)

With high prob., leading eigenvalue λ^{eigs} *of* $\boldsymbol{M}$ *obeys*

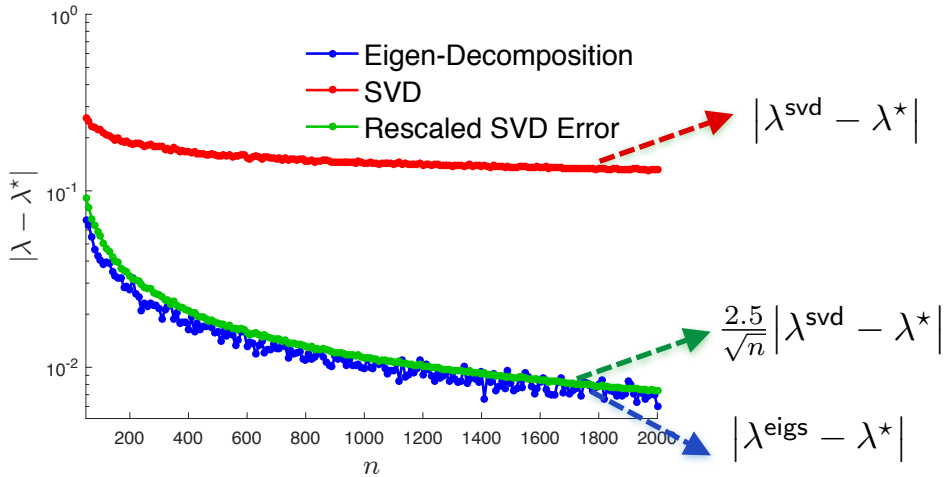
$$|\lambda^{\text{eigs}} - \lambda^\star| \lesssim \sqrt{\frac{\mu}{n}}(\sigma\sqrt{n\log n} + B\log n)$$

Main results: eigenvalue perturbation

Theorem 3 (Chen, Cheng, Fan '18)

With high prob., leading eigenvalue λ^{eigs} *of* $\boldsymbol{M}$ *obeys*

$$\left|\lambda^{\text{eigs}} - \lambda^\star\right| \lesssim \sqrt{\frac{\mu}{n}}\left(\sigma\sqrt{n\log n} + B\log n\right)$$



- Eigen-decomposition is $\sqrt{\frac{n}{\mu}}$ *times better* than SVD!

  — recall $\left|\lambda^{\text{svd}} - \lambda^\star\right| \lesssim \sigma\sqrt{n\log n} + B\log n$

Main results: entrywise eigenvector perturbation

Theorem 4 (Chen, Cheng, Fan '18)

With high prob., leading eigenvector $\boldsymbol{u}$ of $\boldsymbol{M}$ obeys

$$\min \|\boldsymbol{u} \pm \boldsymbol{u}^\star\|_\infty \lesssim \sqrt{\frac{\mu}{n}}(\sigma\sqrt{n\log n} + B\log n)$$

Main results: entrywise eigenvector perturbation

Theorem 4 (Chen, Cheng, Fan '18)

With high prob., leading eigenvector $\boldsymbol{u}$ of $\boldsymbol{M}$ obeys

$$\min \left\|\boldsymbol{u} \pm \boldsymbol{u}^{\star}\right\|_{\infty} \lesssim \sqrt{\frac{\mu}{n}}(\sigma \sqrt{n \log n}+B \log n)$$

- if $\|\boldsymbol{H}\| \ll \left|\lambda^{\star}\right|$, then

$$\min \left\|\boldsymbol{u} \pm \boldsymbol{u}^{\star}\right\|_{2} \ll \left\|\boldsymbol{u}^{\star}\right\|_{2} \qquad \text{(classical bound)}$$

Main results: entrywise eigenvector perturbation

Theorem 4 (Chen, Cheng, Fan '18)

With high prob., leading eigenvector $\boldsymbol{u}$ of $\boldsymbol{M}$ obeys

$$\min \|\boldsymbol{u} \pm \boldsymbol{u}^\star\|_\infty \lesssim \sqrt{\frac{\mu}{n}}(\sigma\sqrt{n\log n} + B\log n)$$

- if $\|\boldsymbol{H}\| \ll |\lambda^\star|$, then

$$\min \|\boldsymbol{u} \pm \boldsymbol{u}^\star\|_2 \ll \|\boldsymbol{u}^\star\|_2 \qquad \text{(classical bound)}$$
$$\min \|\boldsymbol{u} \pm \boldsymbol{u}^\star\|_\infty \ll \|\boldsymbol{u}^\star\|_\infty \qquad \text{(our bound)}$$

Main results: entrywise eigenvector perturbation

Theorem 4 (Chen, Cheng, Fan '18)

With high prob., leading eigenvector $\boldsymbol{u}$ of $\boldsymbol{M}$ obeys

$$\min \|\boldsymbol{u} \pm \boldsymbol{u}^\star\|_\infty \lesssim \sqrt{\frac{\mu}{n}}(\sigma\sqrt{n\log n} + B\log n)$$

- if $\|\boldsymbol{H}\| \ll |\lambda^\star|$, then

$$\min \|\boldsymbol{u} \pm \boldsymbol{u}^\star\|_2 \ll \|\boldsymbol{u}^\star\|_2 \qquad \text{(classical bound)}$$
$$\min \|\boldsymbol{u} \pm \boldsymbol{u}^\star\|_\infty \ll \|\boldsymbol{u}^\star\|_\infty \qquad \text{(our bound)}$$

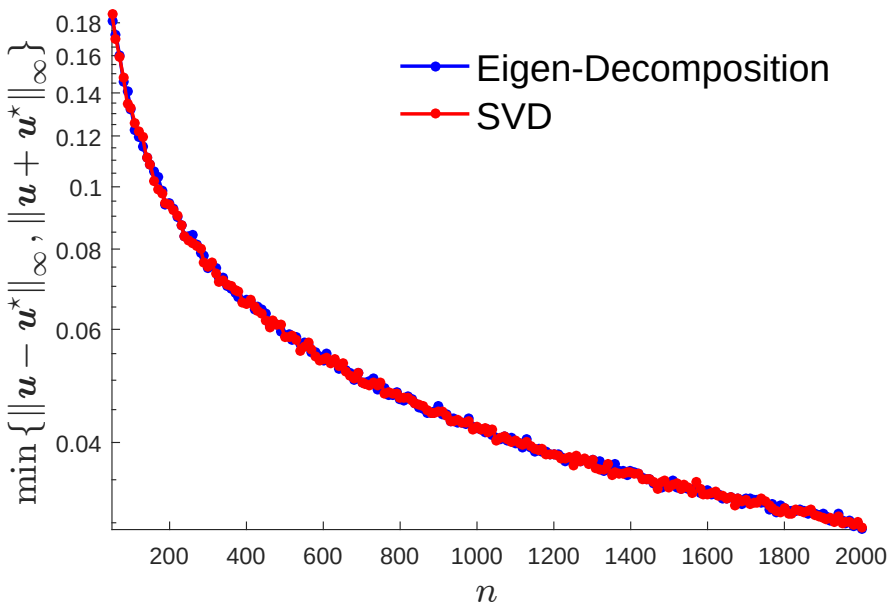
- entrywise eigenvector perturbation is well-controlled

Main results: entrywise eigenvector perturbation

Theorem 4 (Chen, Cheng, Fan '18)

With high prob., leading eigenvector $\boldsymbol{u}$ of $\boldsymbol{M}$ obeys

$$\min \left\|\boldsymbol{u} \pm \boldsymbol{u}^{\star}\right\|_{\infty} \lesssim \sqrt{\frac{\mu}{n}}(\sigma \sqrt{n \log n}+B \log n)$$



$\{H_{i,j}\}$: i.i.d. $\mathcal{N}(0, \sigma^2); \sigma^2 = \frac{1}{n \log n}$

Main results: perturbation of linear forms of eigenvectors

Theorem 5 (Chen, Cheng, Fan '18)

Fix any unit vector $\boldsymbol{a}$. With high prob., leading eigenvector $\boldsymbol{u}$ of $\boldsymbol{M}$ obeys

$$\min\left\{\left|\boldsymbol{a}^\top(\boldsymbol{u}\pm\boldsymbol{u}^\star)\right|\right\} \lesssim \max\left\{\left|\boldsymbol{a}^\top\boldsymbol{u}^\star\right|, \sqrt{\frac{\mu}{n}}\right\}(\sigma\sqrt{n\log n}+B\log n)$$

Main results: perturbation of linear forms of eigenvectors

Theorem 5 (Chen, Cheng, Fan '18)

Fix any unit vector $\boldsymbol{a}$. With high prob., leading eigenvector $\boldsymbol{u}$ of $\boldsymbol{M}$ obeys

$$\min\left\{\left|\boldsymbol{a}^\top(\boldsymbol{u}\pm\boldsymbol{u}^\star)\right|\right\} \lesssim \max\left\{\left|\boldsymbol{a}^\top\boldsymbol{u}^\star\right|, \sqrt{\frac{\mu}{n}}\right\}(\sigma\sqrt{n\log n}+B\log n)$$

- if $\|\boldsymbol{H}\| \ll |\lambda^\star|$, then

$$\min\left\{\left|\boldsymbol{a}^\top(\boldsymbol{u}\pm\boldsymbol{u}^\star)\right|\right\} \ll \max\left\{\left|\boldsymbol{a}^\top\boldsymbol{u}^\star\right|, \|\boldsymbol{u}^\star\|_\infty\right\}$$

- perturbation of an *arbitrary* linear form of leading eigenvector is well-controlled

Intuition: asymmetry reduces bias

From Neumann series one can verify

some sort of Taylor expansion

$$|\lambda - \lambda^\star| \asymp \left| \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H} \boldsymbol{u}^\star}{\lambda} + \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H}^2 \boldsymbol{u}^\star}{\lambda^2} + \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H}^3 \boldsymbol{u}^\star}{\lambda^3} + \cdots \right|$$

Intuition: asymmetry reduces bias

From Neumann series one can verify

some sort of Taylor expansion

$$|\lambda - \lambda^\star| \asymp \left| \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H} \boldsymbol{u}^\star}{\lambda} + \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H}^2 \boldsymbol{u}^\star}{\lambda^2} + \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H}^3 \boldsymbol{u}^\star}{\lambda^3} + \cdots \right|$$

To develop some intuition, let's look at 2nd order term

Intuition: asymmetry reduces bias

From Neumann series one can verify

some sort of Taylor expansion

$$|\lambda - \lambda^\star| \asymp \left| \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H} \boldsymbol{u}^\star}{\lambda} + \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H}^2 \boldsymbol{u}^\star}{\lambda^2} + \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H}^3 \boldsymbol{u}^\star}{\lambda^3} + \cdots \right|$$

To develop some intuition, let's look at 2nd order term

- if $\boldsymbol{H}$ is symmetric,

$$\mathbb{E}\big[\boldsymbol{u}^{\star\top} \boldsymbol{H}^2 \boldsymbol{u}^\star\big] = \mathbb{E}\big[\|\boldsymbol{H}\boldsymbol{u}^\star\|_2^2\big] = n\sigma^2$$

Intuition: asymmetry reduces bias

From Neumann series one can verify

some sort of Taylor expansion

$$|\lambda - \lambda^\star| \asymp \left| \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H} \boldsymbol{u}^\star}{\lambda} + \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H}^2 \boldsymbol{u}^\star}{\lambda^2} + \frac{\boldsymbol{u}^{\star\top} \boldsymbol{H}^3 \boldsymbol{u}^\star}{\lambda^3} + \cdots \right|$$

To develop some intuition, let's look at 2nd order term

- if $\boldsymbol{H}$ is symmetric,

$$\mathbb{E}\big[\boldsymbol{u}^{\star\top} \boldsymbol{H}^2 \boldsymbol{u}^\star\big] = \mathbb{E}\big[\|\boldsymbol{H}\boldsymbol{u}^\star\|_2^2\big] = n\sigma^2$$

- if $\boldsymbol{H}$ is asymmetric,

$$\underbrace{\mathbb{E}\big[\boldsymbol{u}^{\star\top} \boldsymbol{H}^2 \boldsymbol{u}^\star\big] = \mathbb{E}\big[\langle \boldsymbol{H}^\top \boldsymbol{u}^\star, \boldsymbol{H}\boldsymbol{u}^\star\rangle\big] = \sigma^2}$$

much smaller than symmetric case

What happens if $M^\star$ is also not symmetric?

- A rank-1 matrix: $\boldsymbol{M}^\star = \lambda^\star \boldsymbol{u}^\star \boldsymbol{v}^{\star\top} \in \mathbb{R}^{n_1 \times n_2}$
- Suppose we observe 2 independent noisy copies

$$\boldsymbol{M}_1 = \boldsymbol{M}^\star + \boldsymbol{H}_1, \qquad \boldsymbol{M}_2 = \boldsymbol{M}^\star + \boldsymbol{H}_2$$

- **Goal:** estimate $\lambda^\star$, $\boldsymbol{u}^\star$ and $\boldsymbol{v}^\star$

Asymmetrization + dilation

Compute leading eigenvalue / eigenvector of

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{M}_1 \\ \boldsymbol{M}_2^\top & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \boldsymbol{M}^\star + \boldsymbol{H}_1 \\ \boldsymbol{M}^{\star\top} + \boldsymbol{H}_2^\top & \mathbf{0} \end{bmatrix}$$

- Our findings (eigenvalue / eigenvector perturbation) continue to hold for this case!

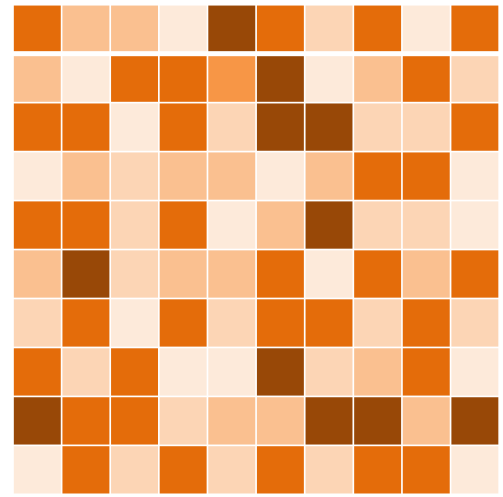
Rank-r case

$\boldsymbol{M}^\star$: truth

- A rank-r and well-conditioned matrix: $\boldsymbol{M}^\star = \sum_{i=1}^r \lambda_i^\star \boldsymbol{u}_i^\star \boldsymbol{u}_i^{\star\top}$
- Observed noisy data: $\boldsymbol{M} = \boldsymbol{M}^\star + \boldsymbol{H}$, where $\{H_{i,j}\}$ are independent
- **Goal:** estimate $\lambda^\star$

Rank-r case

$\boldsymbol{M}^\star$: truth



$\boldsymbol{H}$: noise

- A rank-r and well-conditioned matrix: $\boldsymbol{M}^\star = \sum_{i=1}^r \lambda_i^\star \boldsymbol{u}_i^\star \boldsymbol{u}_i^{\star\top}$
- Observed noisy data: $\boldsymbol{M} = \boldsymbol{M}^\star + \boldsymbol{H}$, where $\{H_{i,j}\}$ are independent
- **Goal:** estimate $\lambda^\star$

Eigenvalue perturbation: rank-r case

Theorem 6 (Chen, Cheng, Fan '18)

With high prob., ith largest eigenvalue λ_i $(1 \le i \le r)$ of $\boldsymbol{M}$ obeys

$$|\lambda_i - \lambda_j^\star| \lesssim \sqrt{\frac{\mu r^2}{n}} (\sigma \sqrt{n \log n} + B \log n)$$

for some $1 \le j \le r$

Eigenvalue perturbation: rank-r case

Theorem 6 (Chen, Cheng, Fan '18)

With high prob., ith largest eigenvalue λ_i $(1 \le i \le r)$ of $\boldsymbol{M}$ obeys

$$|\lambda_i - \lambda_j^\star| \lesssim \sqrt{\frac{\mu r^2}{n}}(\sigma\sqrt{n\log n} + B\log n)$$

for some $1 \le j \le r$

- Eigen-decomposition is $\sqrt{\frac{n}{\mu r^2}}$ times better than SVD!

Eigenvalue perturbation: rank-r case

Theorem 6 (Chen, Cheng, Fan '18)

With high prob., ith largest eigenvalue λ_i $(1 \le i \le r)$ of $\boldsymbol{M}$ obeys

$$|\lambda_i - \lambda_j^\star| \lesssim \sqrt{\frac{\mu r^2}{n}}(\sigma\sqrt{n\log n} + B\log n)$$

for some $1 \le j \le r$

- Eigen-decomposition is $\sqrt{\frac{n}{\mu r^2}}$ times better than SVD!
- Might be improvable to $\sqrt{\frac{\mu r}{n}}(\sigma\sqrt{n\log n} + B\log n)$?

Concluding remarks

Eigen-decomposition could be much more powerful than SVD when dealing with non-symmetric data matrices

Concluding remarks

Eigen-decomposition could be much more powerful than SVD when dealing with non-symmetric data matrices

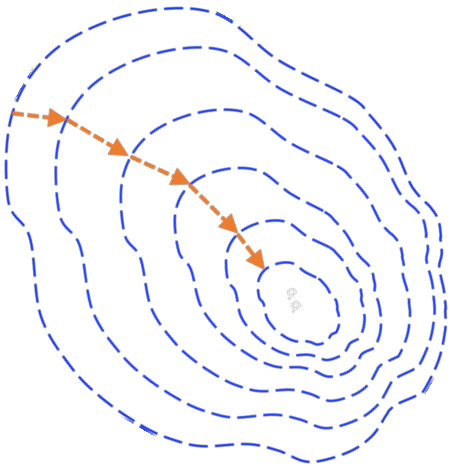
Future directions:

- Eigenvector perturbation for rank-r case
- Beyond i.i.d. noise

Y. Chen, C. Cheng, J. Fan, "Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices", arXiv:1811.12804, 2018

Backup slides: gradient descent for nonconvex matrix completion

Gradient descent for nonconvex matrix completion



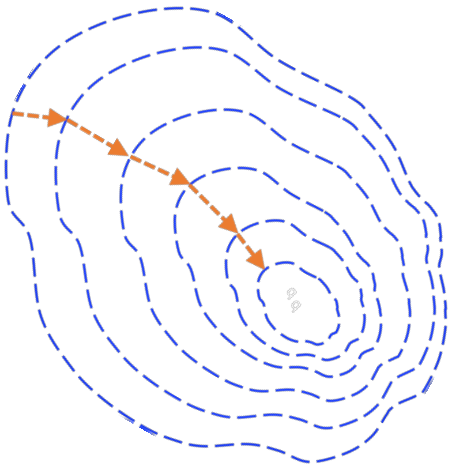
$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \end{aligned}$$

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



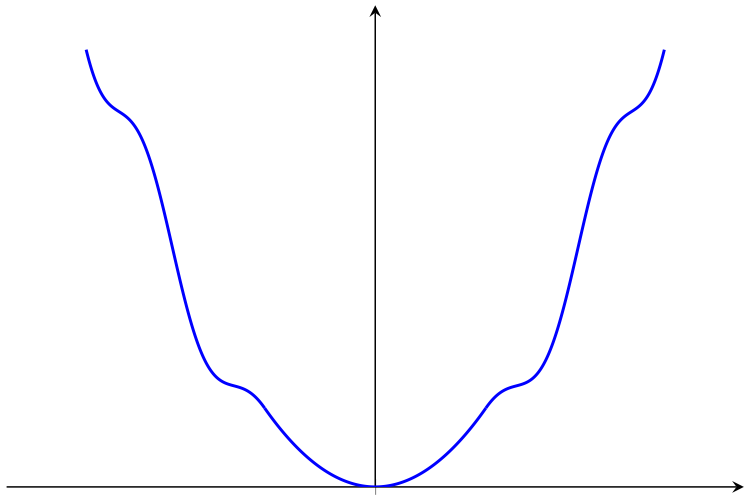
$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \end{aligned}$$

Our work and Chen et al. analyze vanilla GD

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

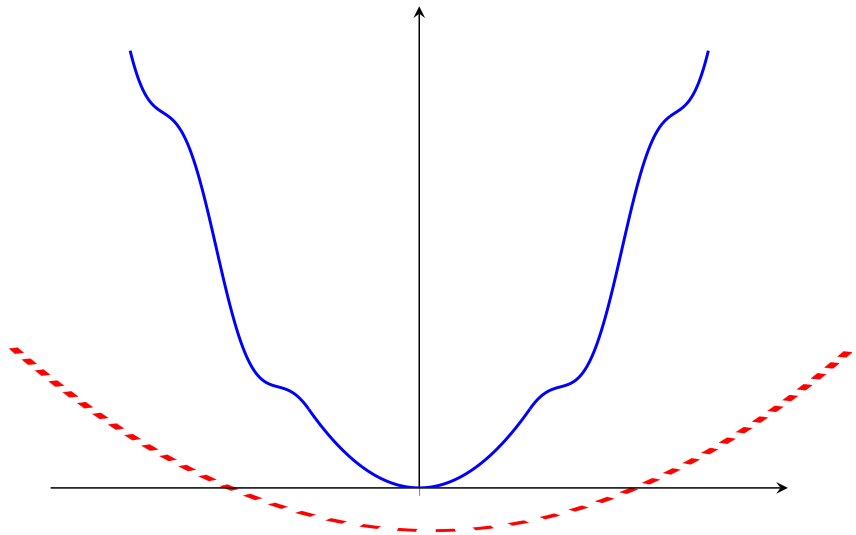
— *Ma, Wang, Chi, Chen '17, Chen, Liu, Li '19*

Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

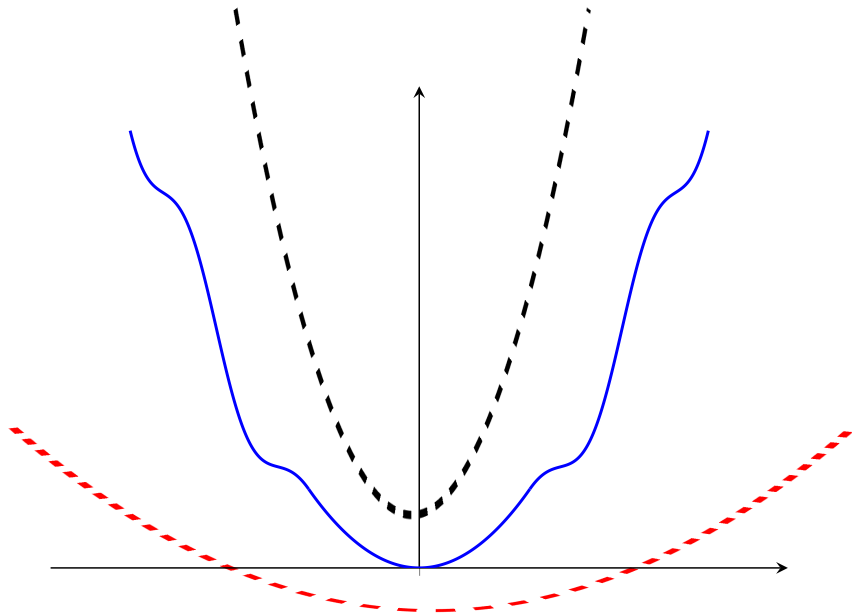
Gradient descent theory revisited



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

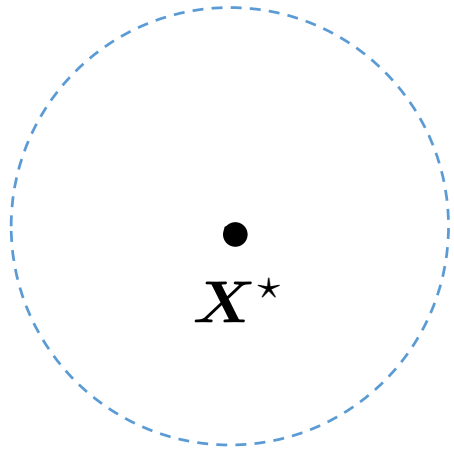
$$\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^\star\|_{\mathrm{F}} \leq \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{X}^t - \boldsymbol{X}^\star\|_{\mathrm{F}}$$

Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

Incoherence region

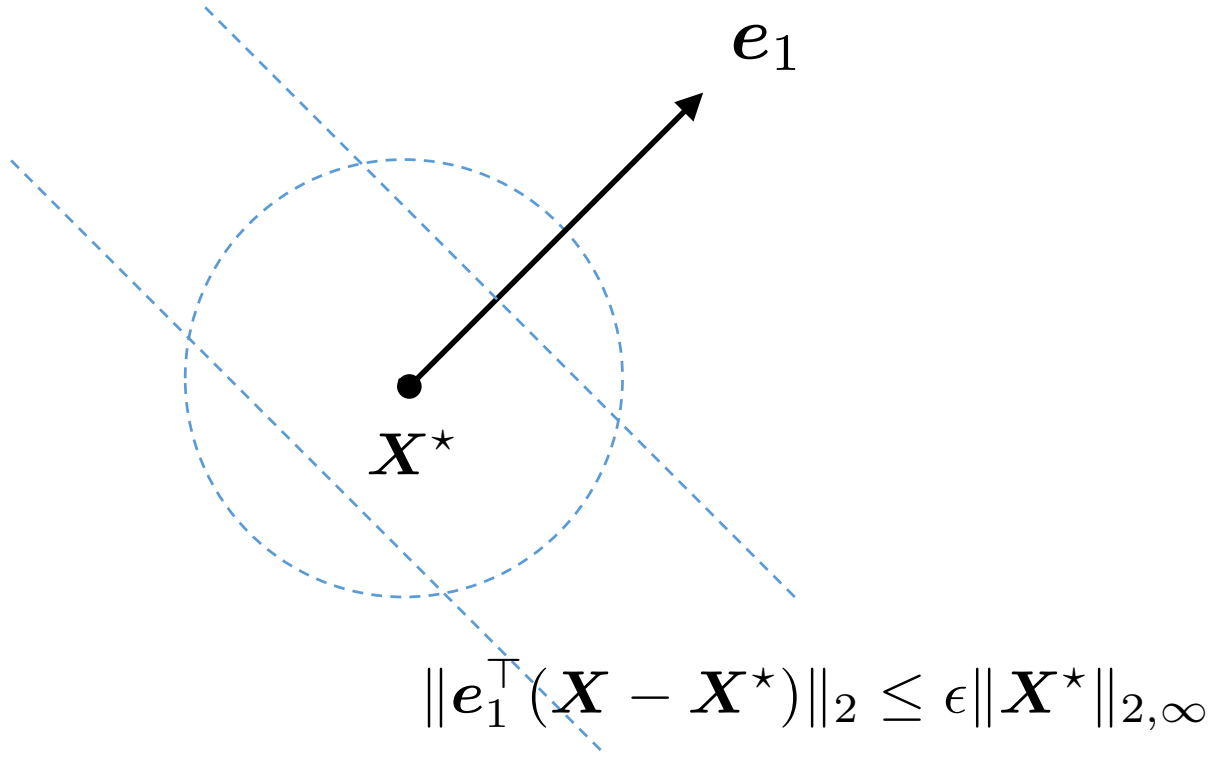
Which region enjoys both restricted strong convexity and smoothness?



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

Incoherence region

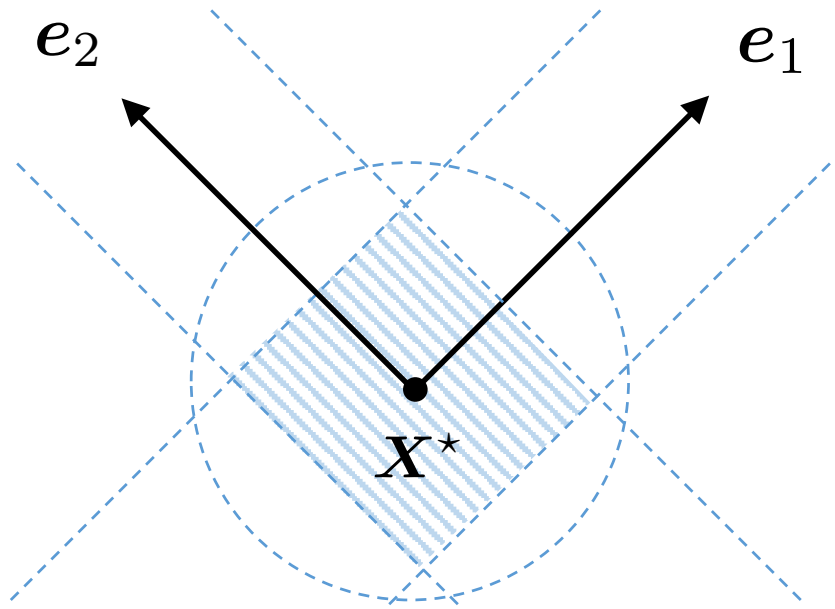
Which region enjoys both restricted strong convexity and smoothness?



- 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?

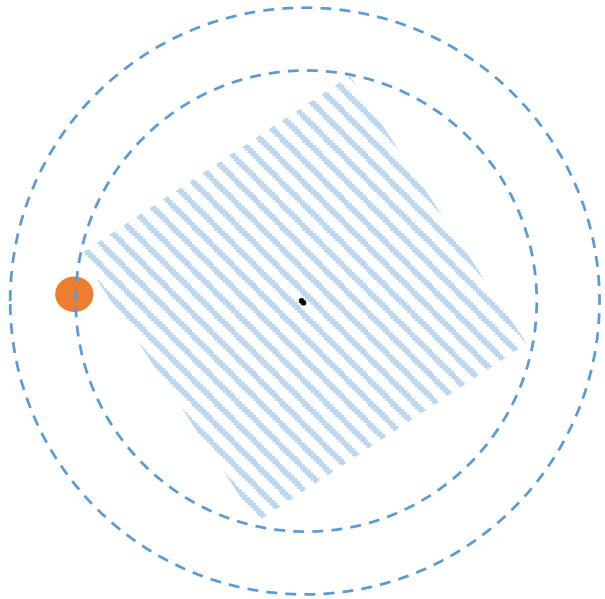


$$\|\boldsymbol{e}_2^\top(\boldsymbol{X}-\boldsymbol{X}^\star)\|_2 \le \epsilon\|\boldsymbol{X}^\star\|_{2,\infty} \quad \|\boldsymbol{e}_1^\top(\boldsymbol{X}-\boldsymbol{X}^\star)\|_2 \le \epsilon\|\boldsymbol{X}^\star\|_{2,\infty}$$

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

Inadequacy of generic gradient descent theory

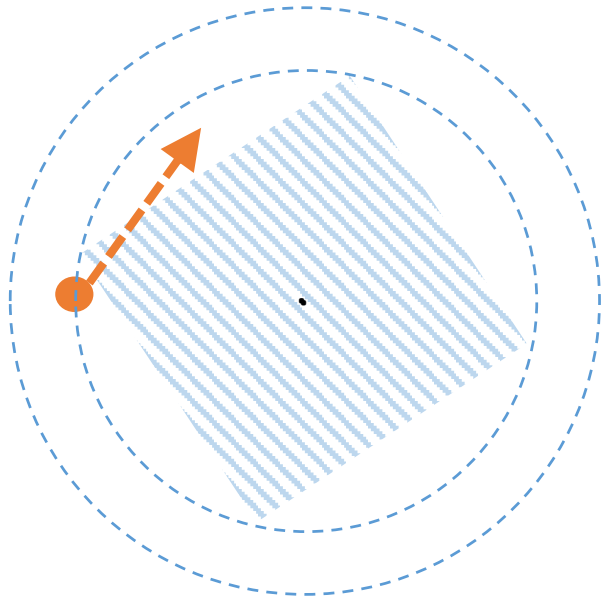
region of local strong convexity + smoothness



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

Inadequacy of generic gradient descent theory

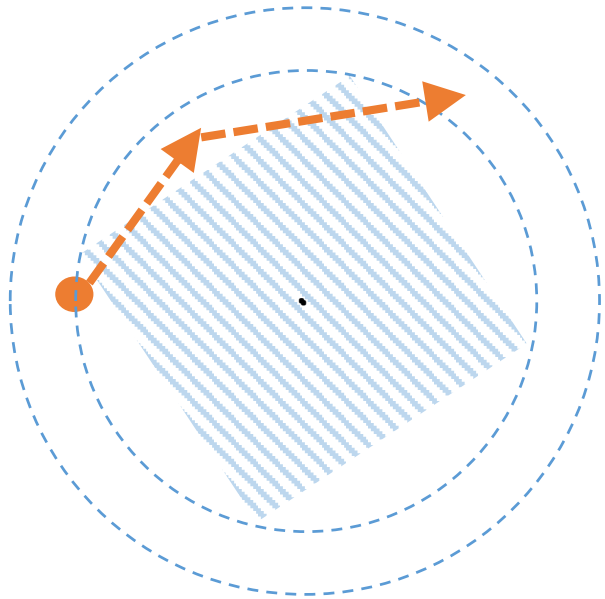
region of local strong convexity + smoothness



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

Inadequacy of generic gradient descent theory

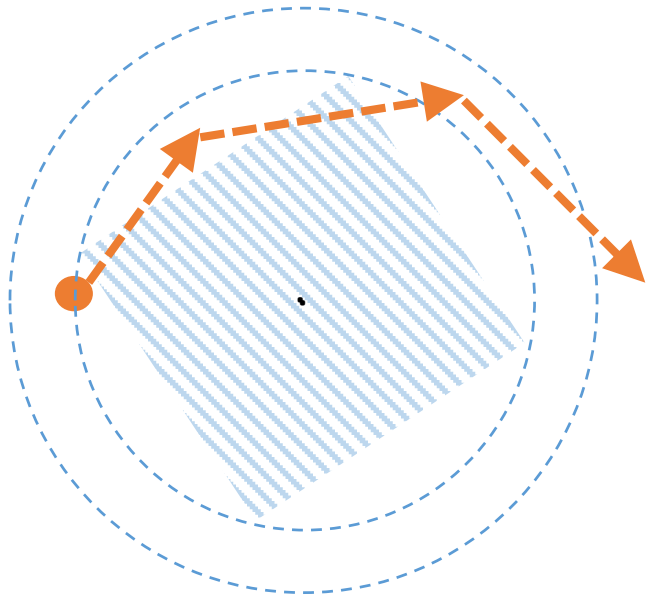
region of local strong convexity + smoothness



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

Inadequacy of generic gradient descent theory

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

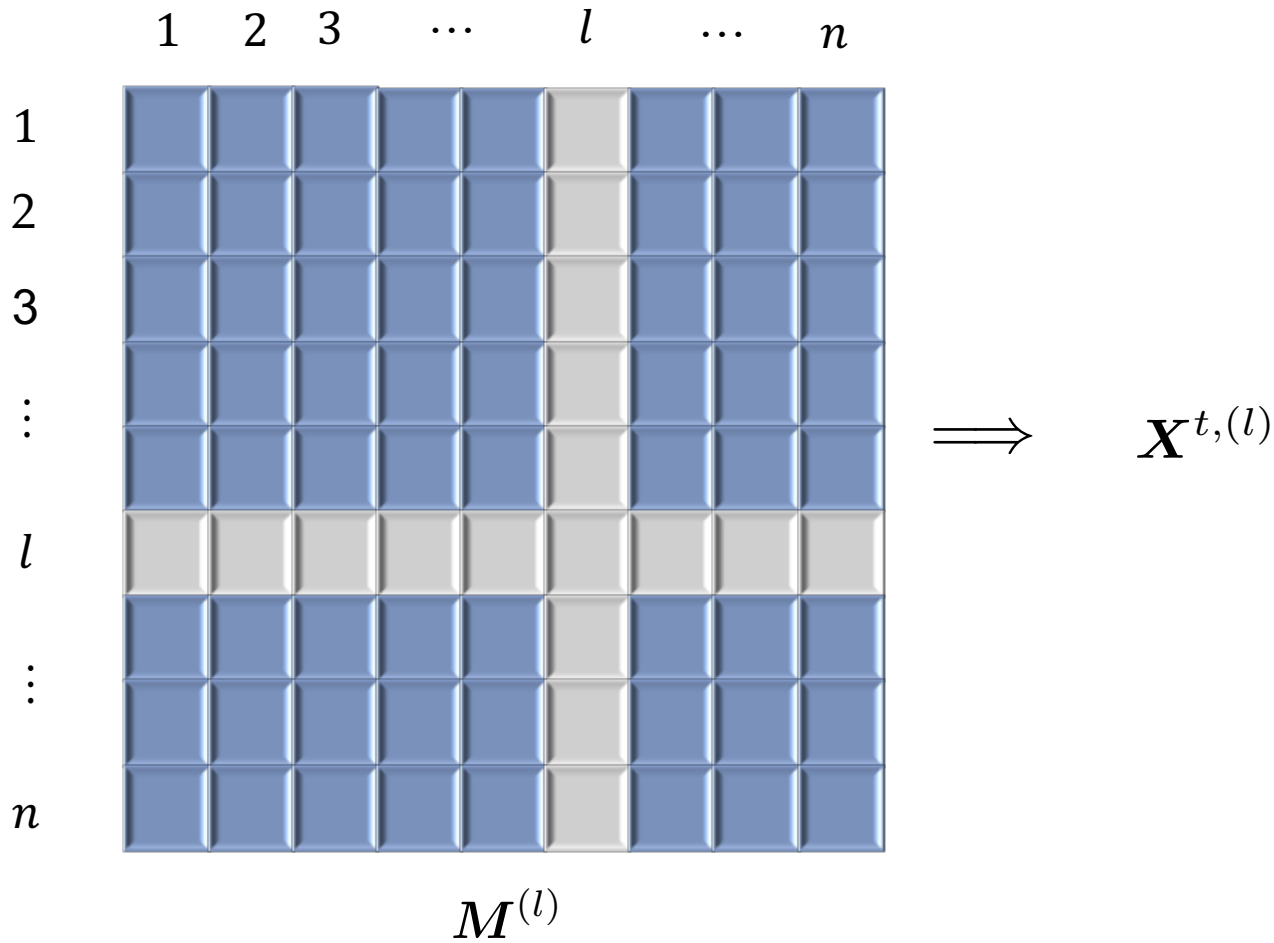
Key proof idea: leave-one-out analysis

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

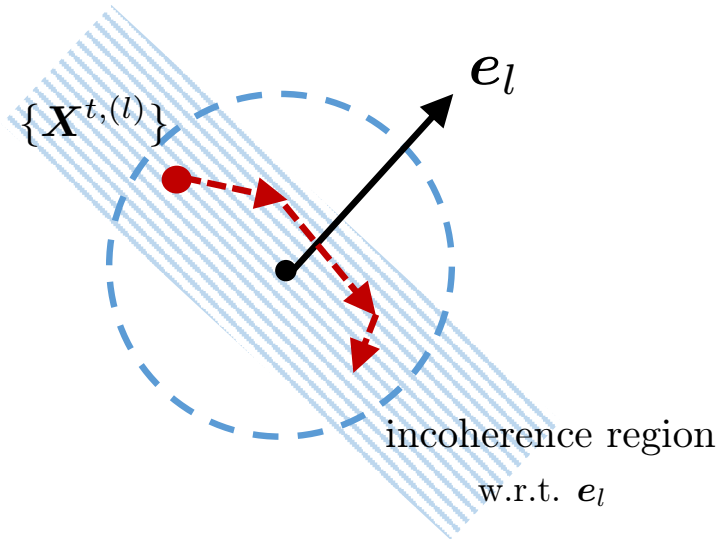
- Stein '72
- El Karoui, Bean, Bickel, Lim, Yu '13
- El Karoui '15
- Javanmard, Montanari '15
- Zhong, Boumal '17
- Lei, Bickel, El Karoui '17
- Sur, Chen, Candès '17
- Abbe, Fan, Wang, Zhong '17
- Chen, Fan, Ma, Wang '17
- Ma, Wang, Chi, Chen '17
- Chen, Chi, Fan, Ma '18
- Ding, Chen '18
- Dong, Shi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

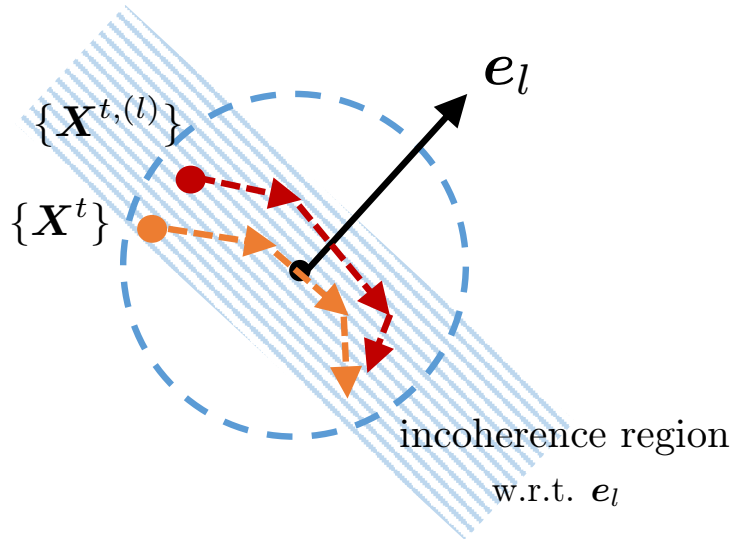


Key proof idea: leave-one-out analysis



- 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 $\{\boldsymbol{X}^{t}\}$