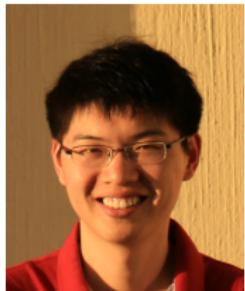


Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization



Yuxin Chen

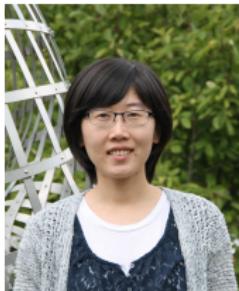
EE, Princeton University



Cong Ma
Princeton ORFE



Yuling Yan
Princeton ORFE



Yuejie Chi
CMU ECE

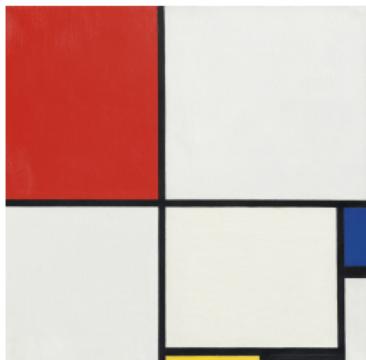


Jianqing Fan
Princeton ORFE

Convex relaxation for low-rank structure

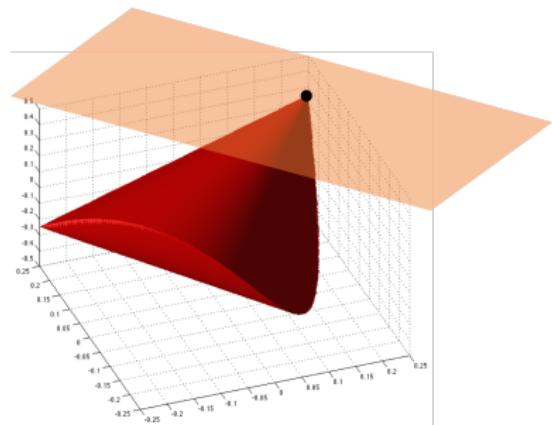
$$\underset{Z}{\text{minimize}} \quad \|Z\|_*$$

subj. to noiseless data constraints



low-rank matrix

figure credit: Piet Mondrian



semidefinite relaxation

Convex relaxation for low-rank structure

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

- ✓ 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

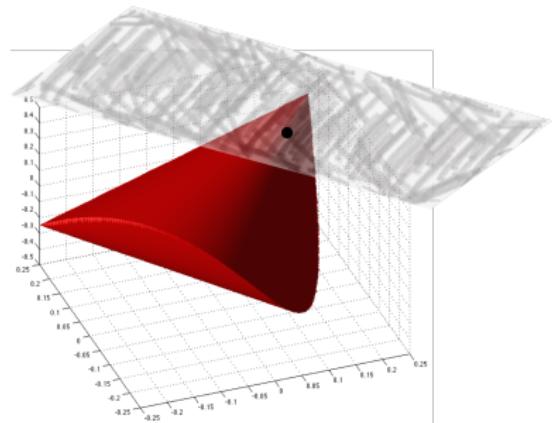
minimize _{Z} $\|Z\|_*$

subj. to **noisy** data constraints



low-rank matrix

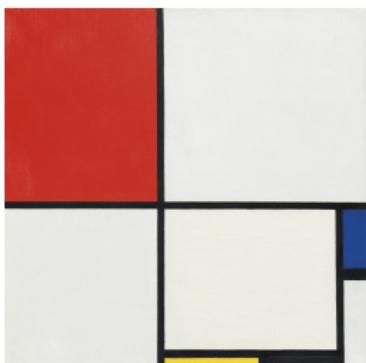
figure credit: Piet Mondrian



semidefinite relaxation

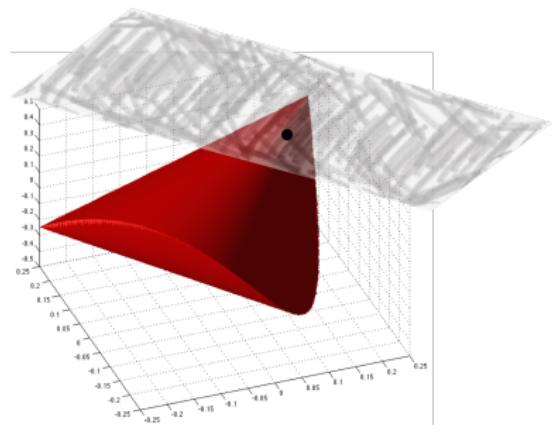
Stability of convex relaxation against noise

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



low-rank matrix

figure credit: Piet Mondrian



semidefinite relaxation

Stability of convex relaxation against noise

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

- ✓ 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

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

- ✓ 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

...

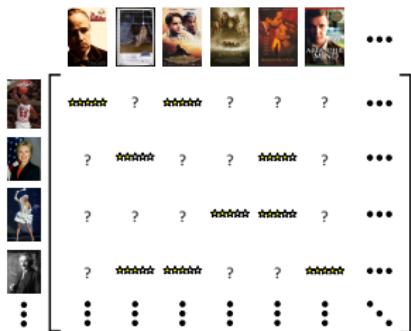
Low-rank matrix completion

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

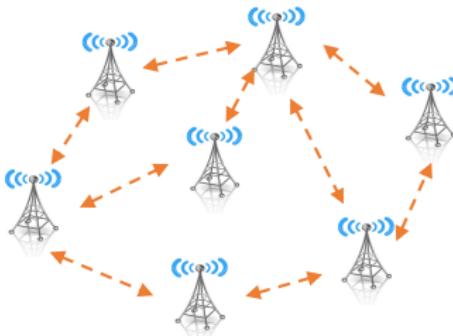


figure credit: E. J. Candès

Given partial samples of a low-rank matrix M^* , fill in missing entries



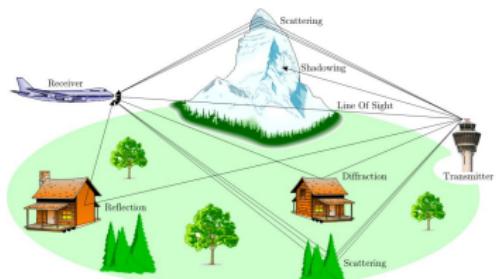
recommendation systems



localization



shape matching

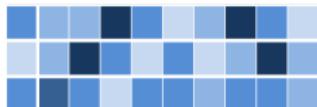
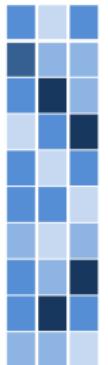


channel estimation

Noisy low-rank matrix completion

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

goal: estimate M^*



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

✓	?	?	?	✓	?
?	?	✓	✓	?	?
✓	?	?	✓	?	?
?	?	✓	?	?	✓
✓	?	?	?	?	?
?	✓	?	?	✓	?
?	?	✓	✓	?	?

sampling set Ω

Noisy low-rank matrix completion

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

goal: estimate M^*

convex relaxation:

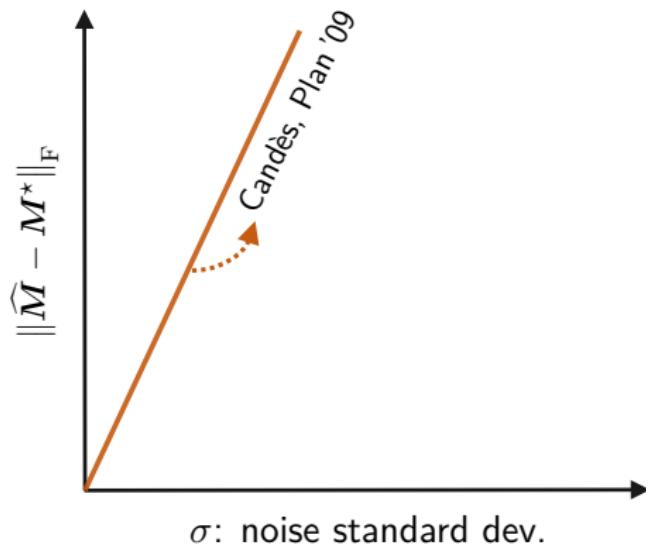
$$\underset{\mathbf{Z} \in \mathbb{R}^{n \times n}}{\text{minimize}} \quad \underbrace{\sum_{(i,j) \in \Omega} (Z_{i,j} - M_{i,j})^2}_{\text{squared loss}} + \lambda \|\mathbf{Z}\|_*$$

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 $M^* \in \mathbb{R}^{n \times n}$: rank $r = O(1)$, incoherent, ...

Candès, Plan '09

$\sigma n^{1.5}$

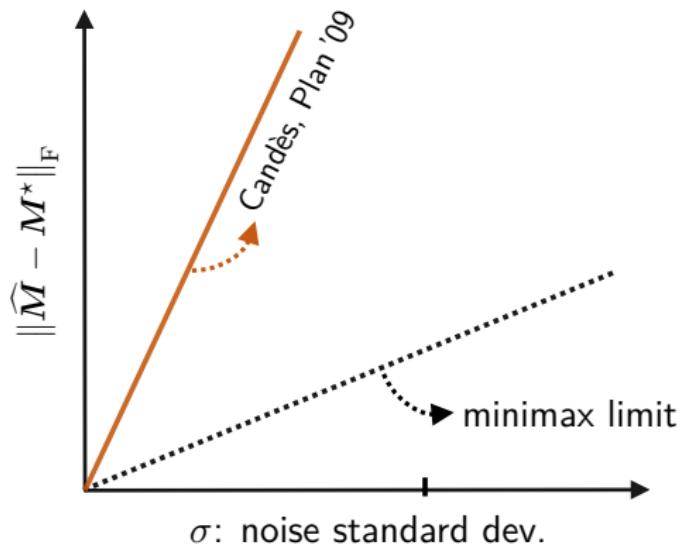


minimax limit

$$\sigma\sqrt{n/p}$$

Candès, Plan '09

$$\sigma n^{1.5}$$



minimax limit

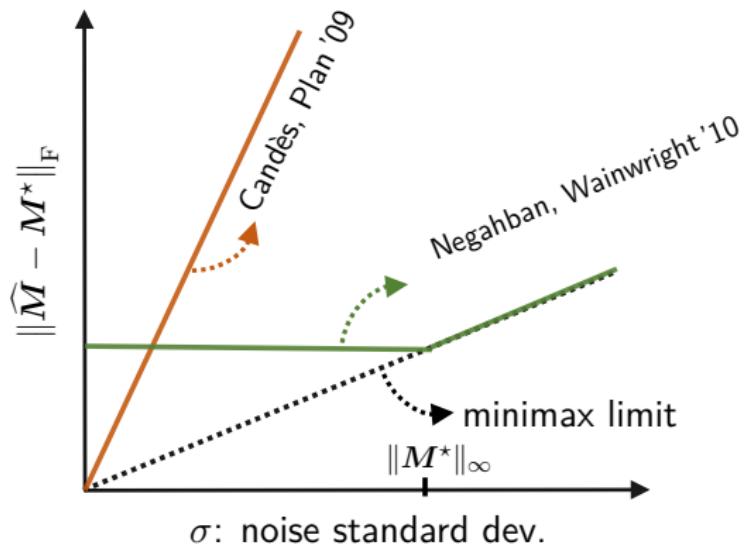
$$\sigma\sqrt{n/p}$$

Candès, Plan '09

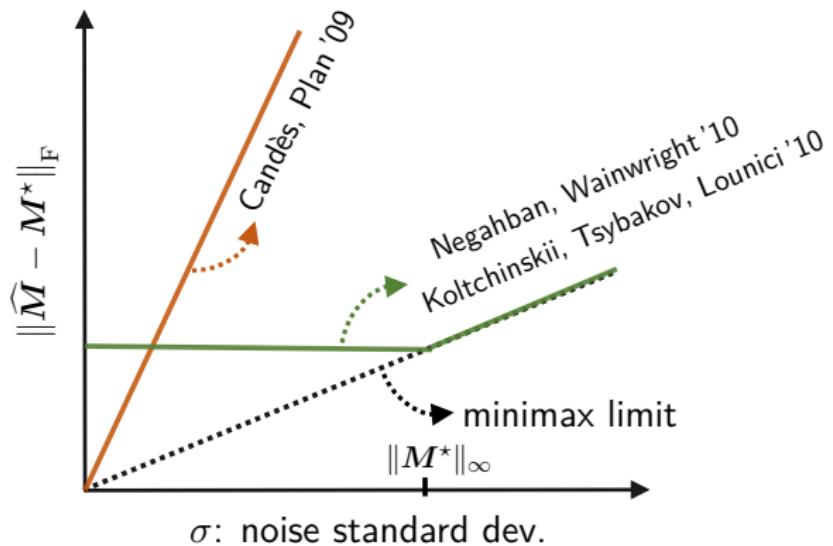
$$\sigma n^{1.5}$$

Negahban, Wainwright '10

$$\max\{\sigma, \|\mathbf{M}^*\|_\infty\} \sqrt{n/p}$$

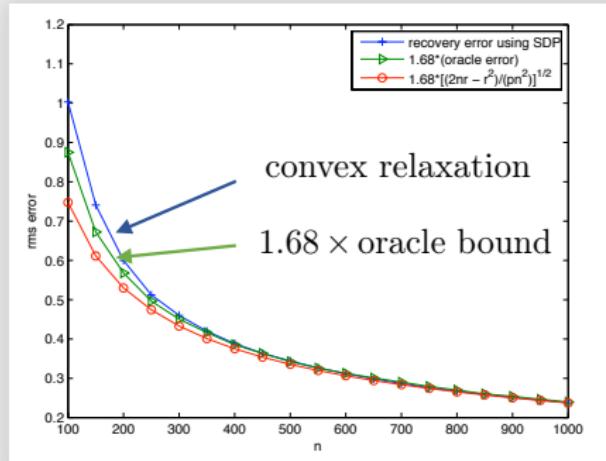


minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$
Negahban, Wainwright '10	$\max\{\sigma, \ \mathbf{M}^*\ _\infty\} \sqrt{n/p}$
Koltchinskii, Tsybakov, Lounici '10	$\max\{\sigma, \ \mathbf{M}^*\ _\infty\} \sqrt{n/p}$



Matrix Completion with Noise

Emmanuel J. Candès and Yannick Plan



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

Matrix Completion with Noise

Emmanuel J. Candès and Yannick Plan

with adversarial noise. Consequently, our analysis loses
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{M}_{cvx} is optimizer if $\underbrace{\text{there exists } \mathbf{W}}_{\text{dual certificate}}$ s.t.

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

What are the roadblocks?

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

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



David Gross

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

What are the roadblocks?

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

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



David Gross

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

dual certification (golfing scheme)



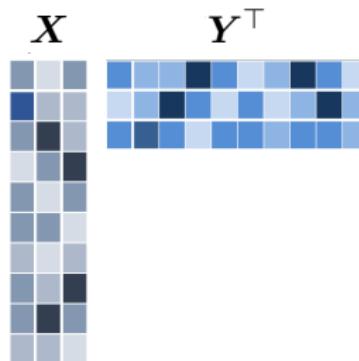
dual certification (golfing scheme)



nonconvex optimization

A detour: nonconvex optimization

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

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


A detour: nonconvex optimization

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

$$\begin{array}{c} \mathbf{X} \qquad \mathbf{Y}^\top \\ \begin{matrix} \text{---} & \text{---} \\ \text{---} & \text{---} \end{matrix} \end{array}$$

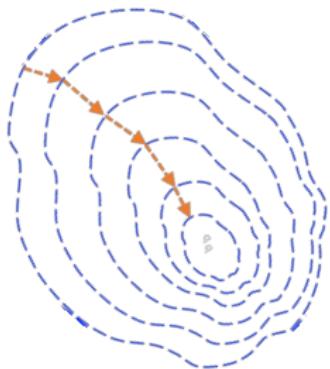
$$\underset{\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times r}}{\text{minimize}} \quad f(\mathbf{X}, \mathbf{Y}) = \underbrace{\sum_{(i,j) \in \Omega} \left[(\mathbf{X}\mathbf{Y}^\top)_{i,j} - M_{i,j} \right]^2}_{\text{squared loss}} + \text{reg}(\mathbf{X}, \mathbf{Y})$$

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

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



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

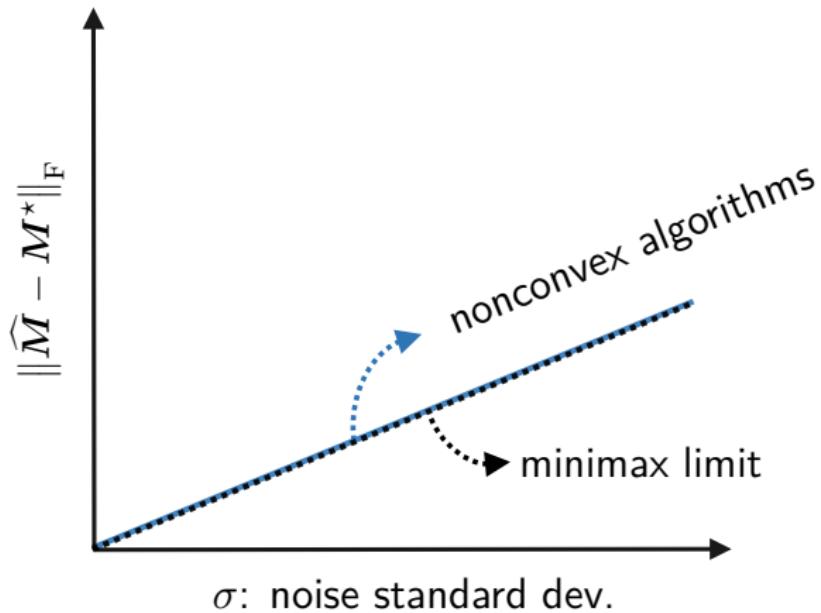
$$\mathbf{X}^{t+1} = \mathbf{X}^t - \eta_t \nabla_{\mathbf{X}} f(\mathbf{X}^t, \mathbf{Y}^t)$$

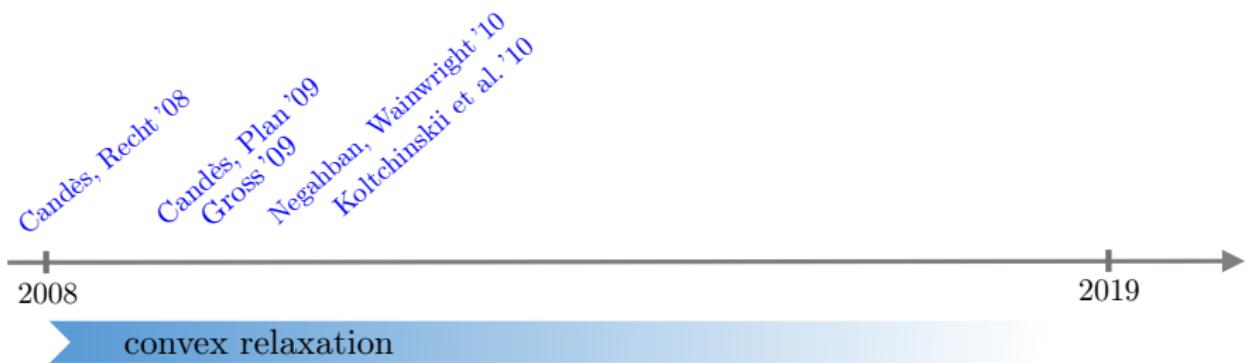
$$\mathbf{Y}^{t+1} = \mathbf{Y}^t - \eta_t \nabla_{\mathbf{Y}} f(\mathbf{X}^t, \mathbf{Y}^t)$$

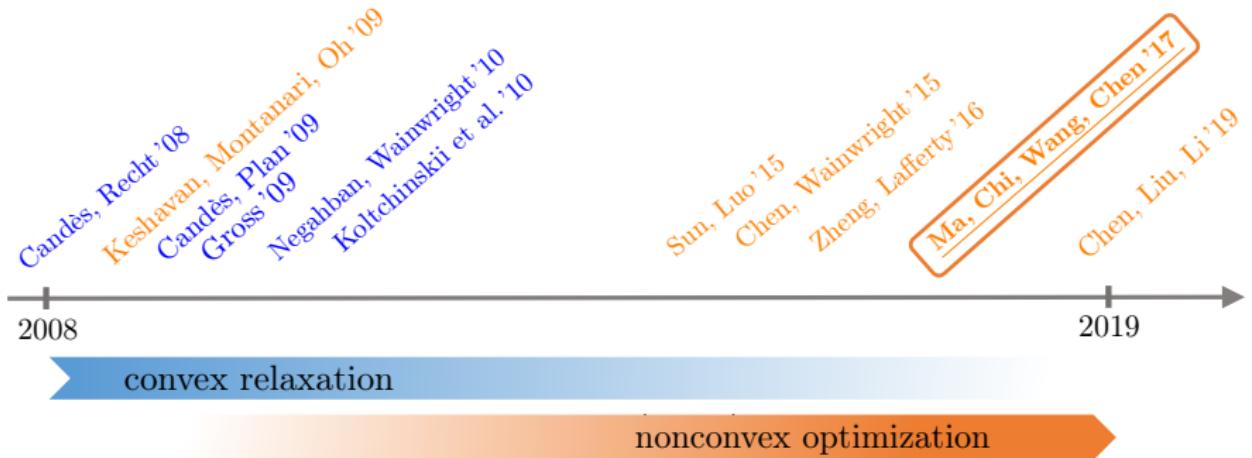
A detour: nonconvex optimization

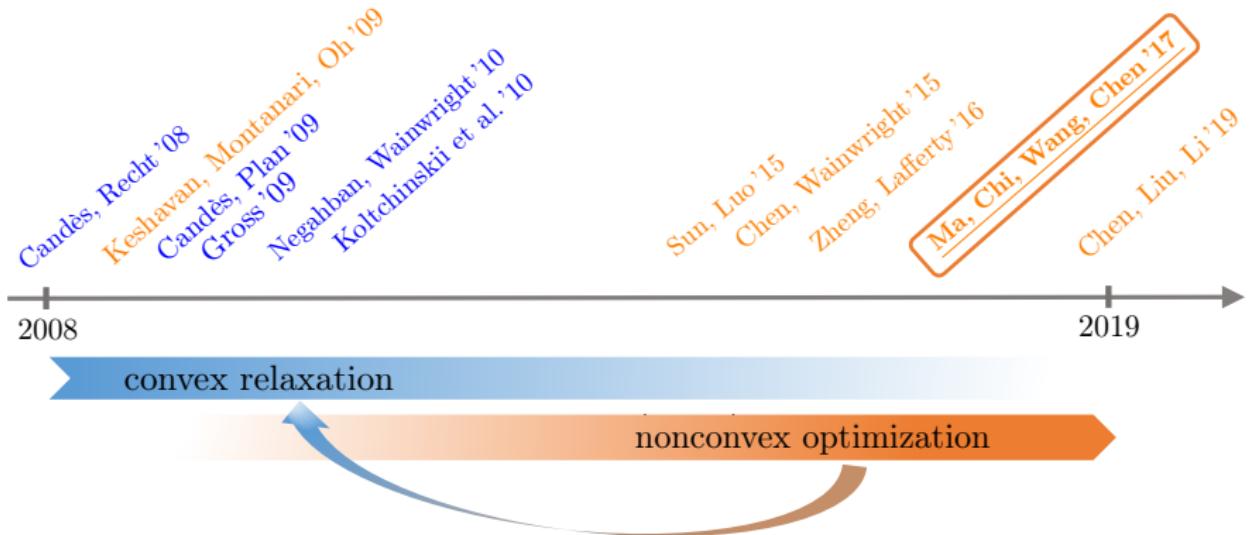
- **random sampling:** each $(i, j) \in \Omega$ with prob. p
- **random noise:** i.i.d. sub-Gaussian noise with variance σ^2
- true matrix $M^* \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, ...

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









A motivating experiment

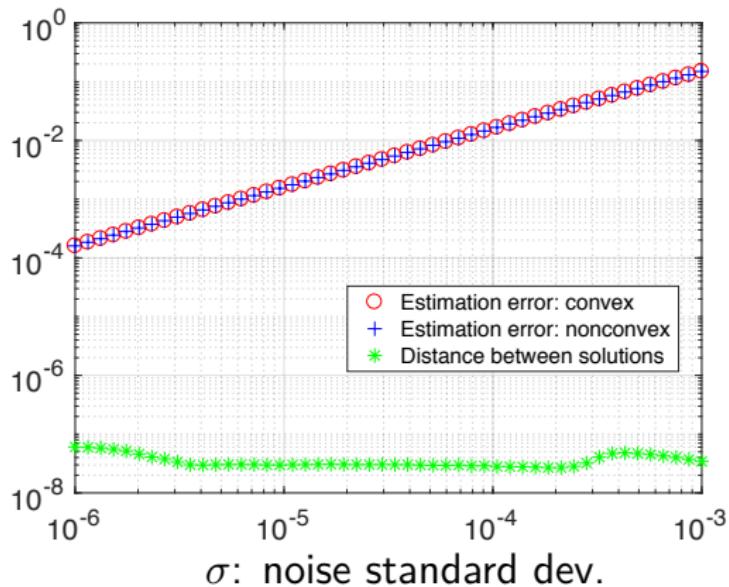
convex: $\underset{\mathbf{Z} \in \mathbb{R}^{n \times n}}{\text{minimize}} \sum_{(i,j) \in \Omega} (Z_{i,j} - M_{i,j})^2 + \lambda \|\mathbf{Z}\|_*$

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

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

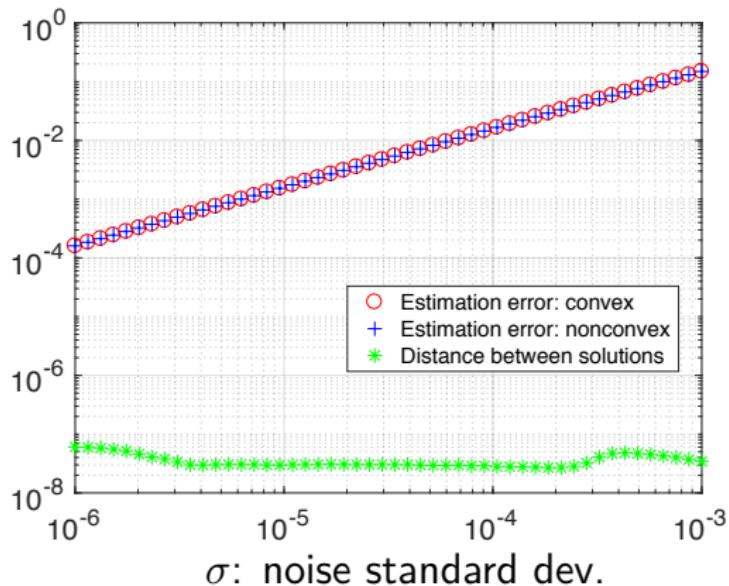
A motivating experiment

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



A motivating experiment

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



Convex and nonconvex solutions are exceedingly close!

convex



nonconvex



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

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

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

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 $M^* \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, well-conditioned

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

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

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

1. $\widehat{\mathbf{M}}_{\text{cvx}}$ is nearly rank- r

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

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 $M^* \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, well-conditioned

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

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

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

1. \widehat{M}_{cvx} is nearly rank- r

2. $\|\widehat{M}_{\text{cvx}} - M^*\|_{\text{F}} \lesssim \sigma \sqrt{\frac{n}{p}}$

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 $M^* \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, well-conditioned

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

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

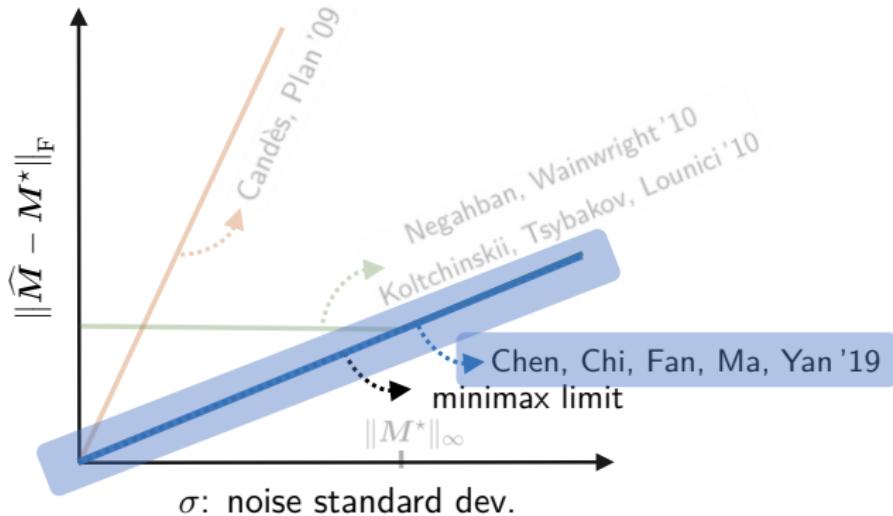
With high prob., any minimizer \widehat{M}_{cvx} of convex program obeys

1. \widehat{M}_{cvx} is nearly rank- r

2. $\|\widehat{M}_{\text{cvx}} - M^*\|_{\text{F}} \lesssim \sigma \sqrt{\frac{n}{p}}$

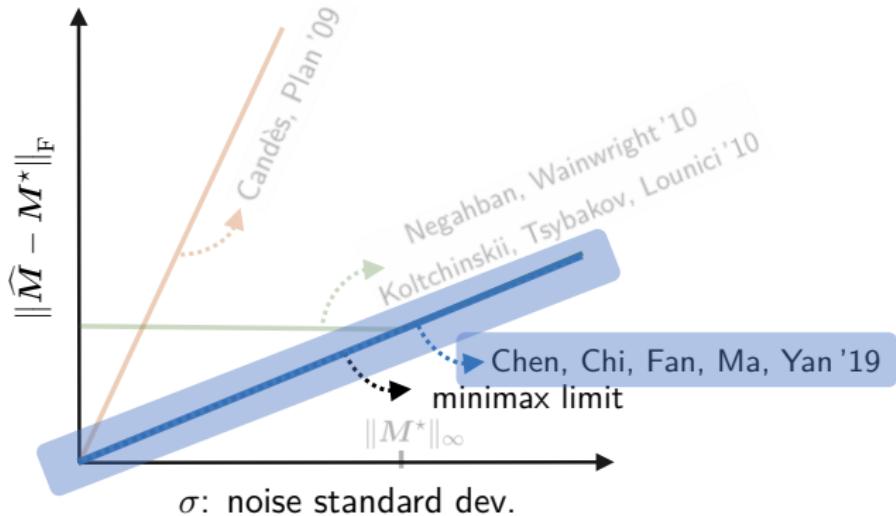
$$\|\widehat{M}_{\text{cvx}} - M^*\|_{\infty} \lesssim \sigma \sqrt{\frac{n \log n}{p}} \cdot \frac{1}{n}$$

$$\|\widehat{\boldsymbol{M}}_{\text{cvx}} - \boldsymbol{M}^*\|_{\text{F}} \lesssim \sigma \sqrt{\frac{n}{p}}$$



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

$$\|\widehat{\mathbf{M}}_{\text{cvx}} - \mathbf{M}^*\|_{\text{F}} \lesssim \sigma \sqrt{\frac{n}{p}} \quad \|\widehat{\mathbf{M}}_{\text{cvx}} - \mathbf{M}^*\|_{\infty} \lesssim \sigma \sqrt{\frac{n \log n}{p}} \cdot \frac{1}{n}$$



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

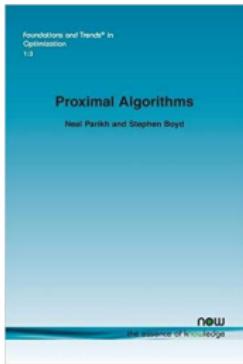
Implicit regularization

No need to enforce spikiness constraint as in Negahban & Wainwright

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

- convex programming automatically controls spikiness of solutions

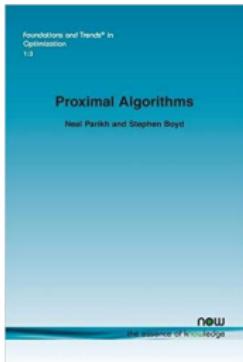
Statistical guarantees for iterative algorithms



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

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

Statistical guarantees for iterative algorithms



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

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 \mathbf{Z} with $g(\mathbf{Z}) \leq g(\mathbf{Z}_{\text{opt}}) + \varepsilon$ 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 $M^* \in \mathbb{R}^{n \times n}$: 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 $M^* \in \mathbb{R}^{n \times n}$: incoherent, well-conditioned

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

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

1. \widehat{M}_{cvx} is nearly rank- r

2.
$$\|\widehat{M}_{\text{cvx}} - M^*\|_{\text{F}} \lesssim \frac{\sigma}{\sigma_{\min}(M^*)} \sqrt{\frac{n}{p}} \|M^*\|_{\text{F}}$$

$$\|\widehat{M}_{\text{cvx}} - M^*\|_{\infty} \lesssim \sqrt{r} \frac{\sigma}{\sigma_{\min}(M^*)} \sqrt{\frac{n \log n}{p}} \|M^*\|_{\infty}$$

$$\|\widehat{M}_{\text{cvx}} - M^*\| \lesssim \frac{\sigma}{\sigma_{\min}(M^*)} \sqrt{\frac{n}{p}} \|M^*\|$$

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 $M^* \in \mathbb{R}^{n \times n}$: 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

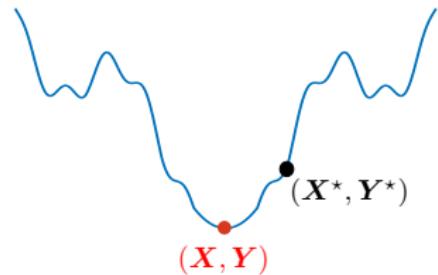
(X, Y) is nonconvex optimizer

Link between convex and nonconvex optimizers

(X, Y) is nonconvex optimizer $\xrightarrow{?}$ XY^\top is convex solution

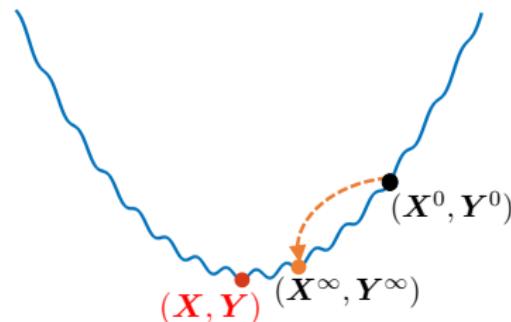
Link between convex and nonconvex optimizers

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



(\mathbf{X}, \mathbf{Y}) is nonconvex optimizer $\xrightarrow{\text{✓}}$ \mathbf{XY}^\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 “approximate stationary points” instead
 $\nabla f(\mathbf{X}, \mathbf{Y}) \approx 0$

Approximate nonconvex optimizers

Strategy: resort to “approximate stationary points” instead
 $\nabla f(\mathbf{X}, \mathbf{Y}) \approx \mathbf{0}$

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

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

Approximate nonconvex optimizers

Strategy: resort to “approximate stationary points” instead
 $\nabla f(\mathbf{X}, \mathbf{Y}) \approx 0$

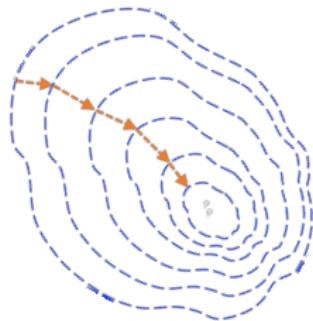
starting from $(\mathbf{X}^0, \mathbf{Y}^0) = \text{truth}$ or spectral initialization:

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

- when T is large: there exists point with very small gradient
 $\|\nabla f(\mathbf{X}, \mathbf{Y})\|_F \lesssim \frac{1}{\sqrt{\eta T}}$
- hopefully not far from $(\mathbf{X}^*, \mathbf{Y}^*)$

Gradient descent for nonconvex matrix completion

Gradient descent for nonconvex matrix completion



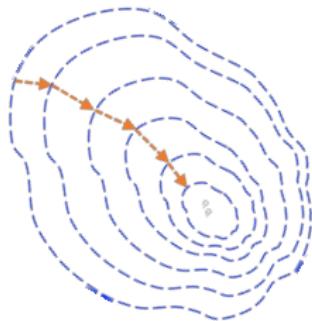
$$\begin{aligned}\mathbf{X}^{t+1} &= \mathbf{X}^t - \eta \nabla_{\mathbf{X}} f(\mathbf{X}^t, \mathbf{Y}^t) \\ \mathbf{Y}^{t+1} &= \mathbf{Y}^t - \eta \nabla_{\mathbf{Y}} f(\mathbf{X}^t, \mathbf{Y}^t)\end{aligned}$$

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



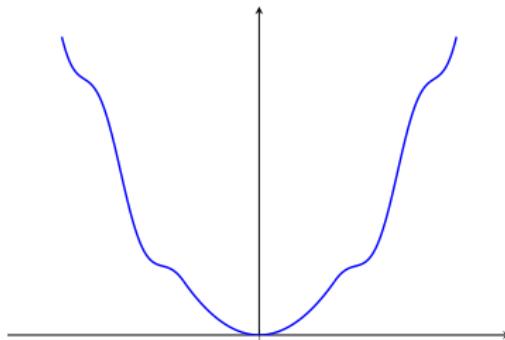
$$\begin{aligned}\mathbf{X}^{t+1} &= \mathbf{X}^t - \eta \nabla_{\mathbf{X}} f(\mathbf{X}^t, \mathbf{Y}^t) \\ \mathbf{Y}^{t+1} &= \mathbf{Y}^t - \eta \nabla_{\mathbf{Y}} f(\mathbf{X}^t, \mathbf{Y}^t)\end{aligned}$$

Our work and Chen et al. analyze **vanilla** GD

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

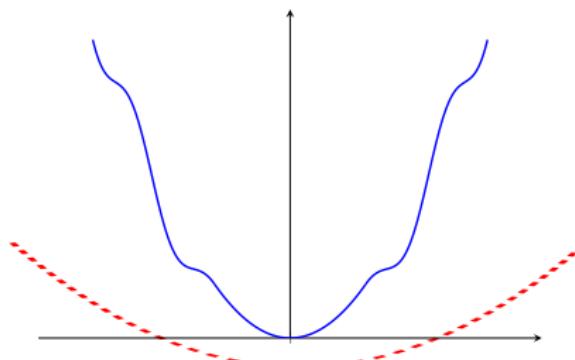
— Ma, Wang, Chi, Chen '17, Chen, Liu, Li '19

Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

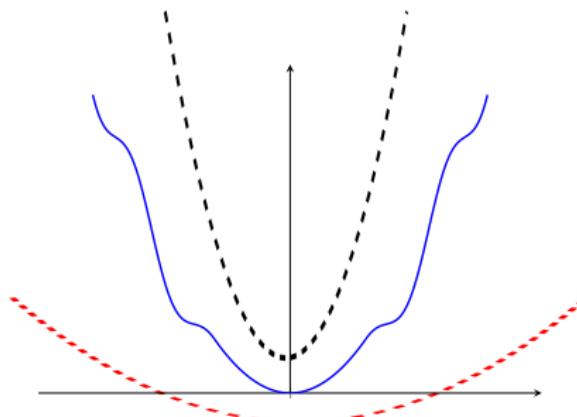
Gradient descent theory revisited



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

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

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

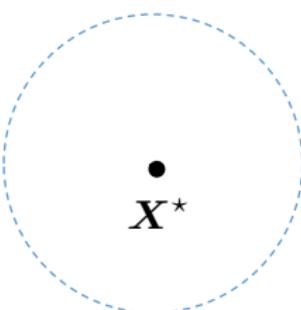
$$\|\mathbf{X}^{t+1} - \mathbf{X}^\star\|_{\text{F}} \leq \left(1 - \frac{\alpha}{\beta}\right) \|\mathbf{X}^t - \mathbf{X}^\star\|_{\text{F}}$$

Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

Incoherence region

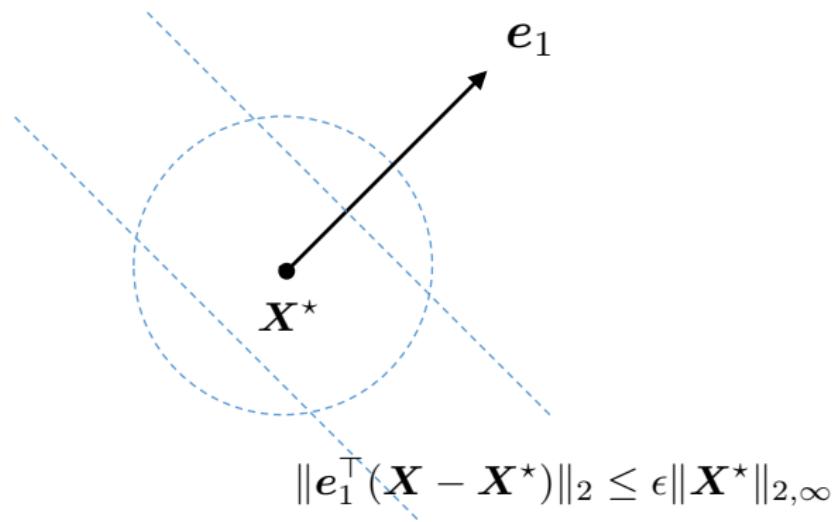
Which region enjoys both restricted strong convexity and smoothness?



- X is not far away from X^*

Incoherence region

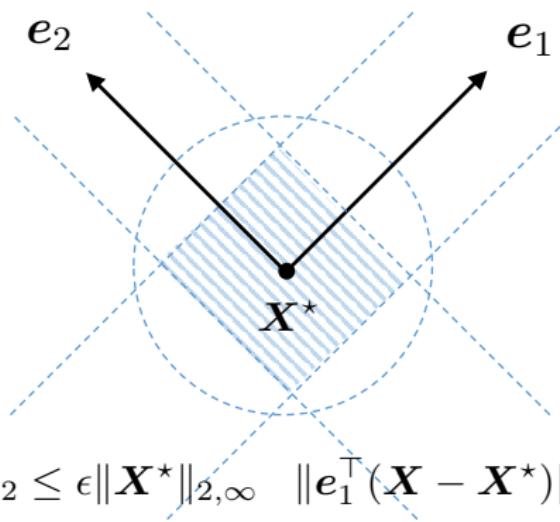
Which region enjoys both restricted strong convexity and smoothness?



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

Incoherence region

Which region enjoys both restricted strong convexity and smoothness?



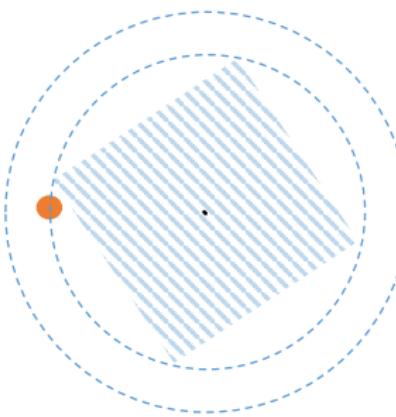
$$\|e_2^\top (X - X^*)\|_2 \leq \epsilon \|X^*\|_{2,\infty} \quad \|e_1^\top (X - X^*)\|_2 \leq \epsilon \|X^*\|_{2,\infty}$$

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

Inadequacy of generic gradient descent theory



region of local strong convexity + smoothness

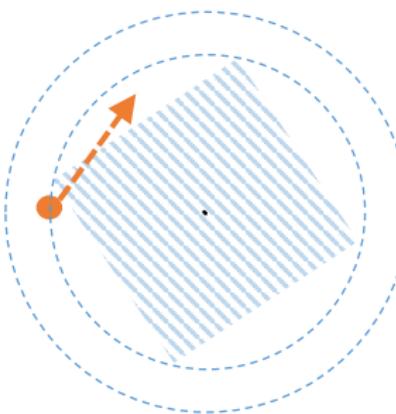


- Generic optimization theory does NOT ensure GD stays in incoherence region

Inadequacy of generic gradient descent theory



region of local strong convexity + smoothness

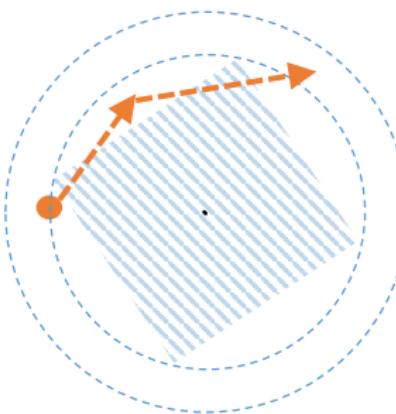


- Generic optimization theory does NOT ensure GD stays in incoherence region

Inadequacy of generic gradient descent theory



region of local strong convexity + smoothness

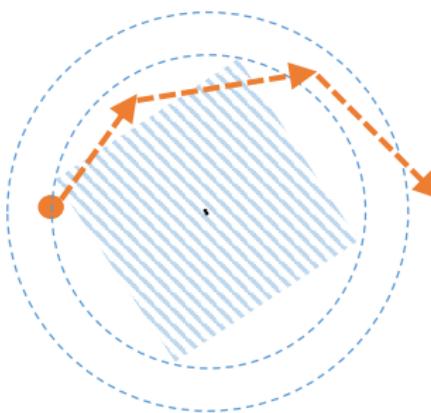


- Generic optimization theory does NOT ensure GD stays in incoherence region

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

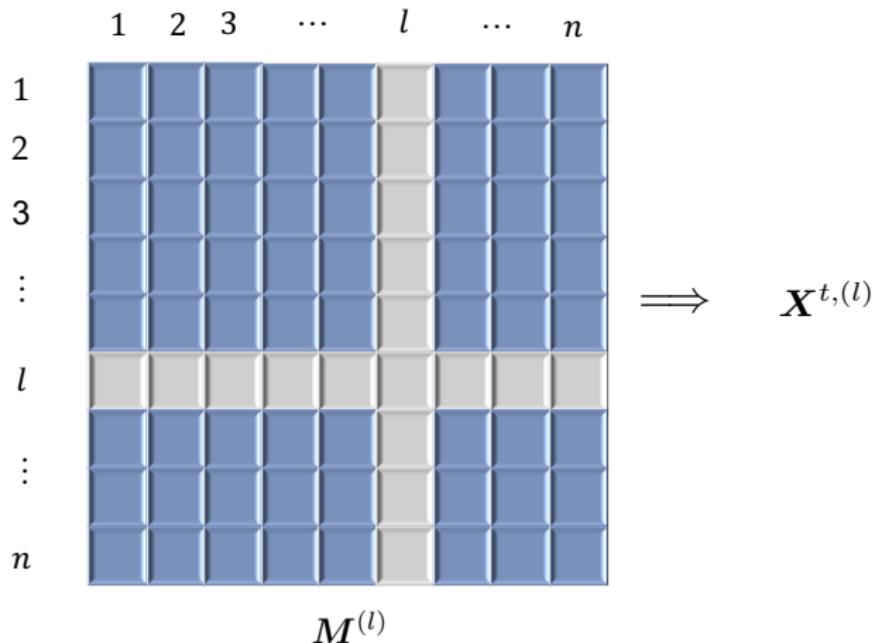
Key proof idea: leave-one-out analysis

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

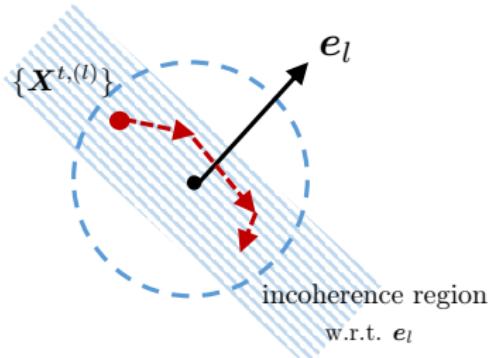
- Stein '72
- El Karoui, Bean, Bickel, Lim, Yu '13
- El Karoui '15
- Javanmard, Montanari '15
- Zhong, Boumal '17
- Lei, Bickel, El Karoui '17
- Sur, Chen, Candès '17
- Abbe, Fan, Wang, Zhong '17
- Chen, Fan, Ma, Wang '17
- Ma, Wang, Chi, Chen '17
- Chen, Chi, Fan, Ma '18
- Ding, Chen '18
- Dong, Shi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

Key proof idea: leave-one-out analysis

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

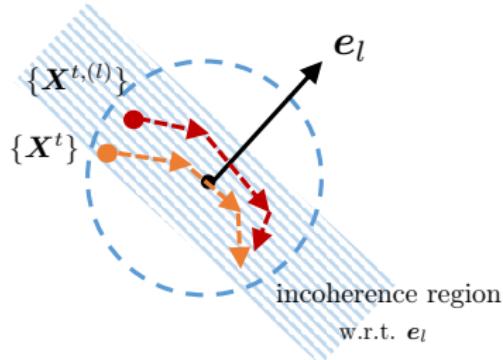


Key proof idea: leave-one-out analysis

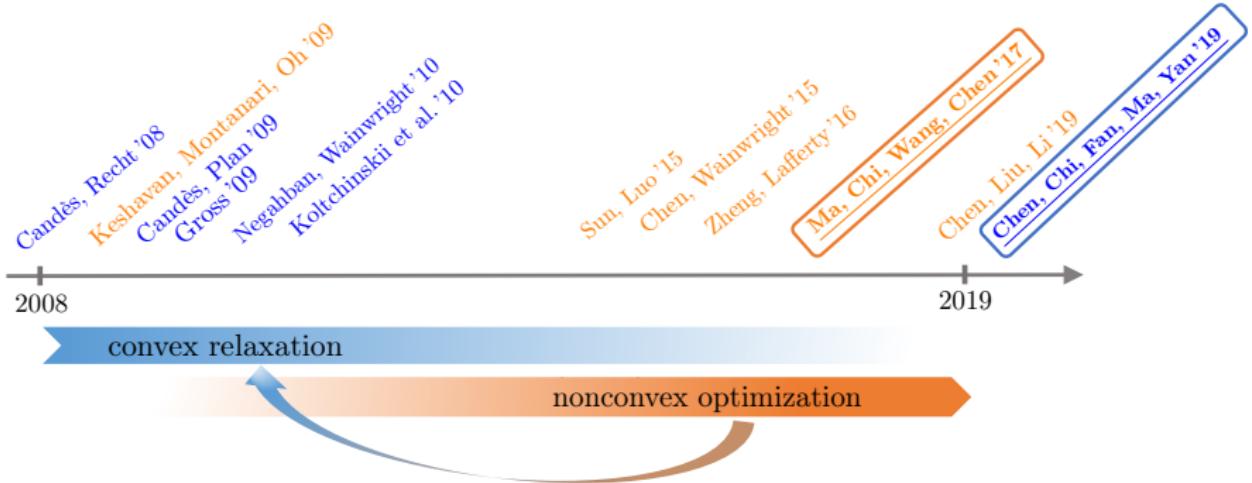


- Leave-one-out iterates $\{\mathbf{X}^{t,(l)}\}$ contains more information of l^{th} row of truth; indep. of randomness in l^{th} row

Key proof idea: leave-one-out analysis



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



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