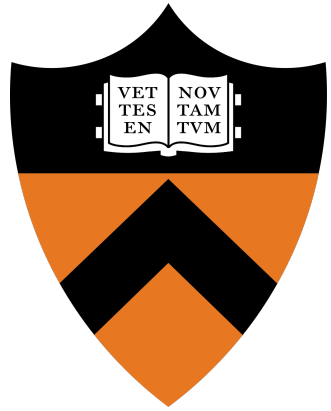
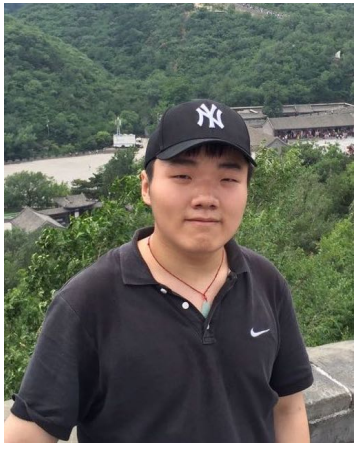
# Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization

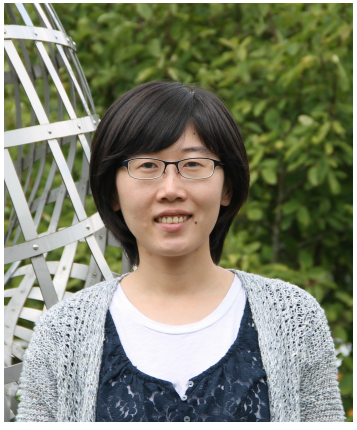



Cong Ma

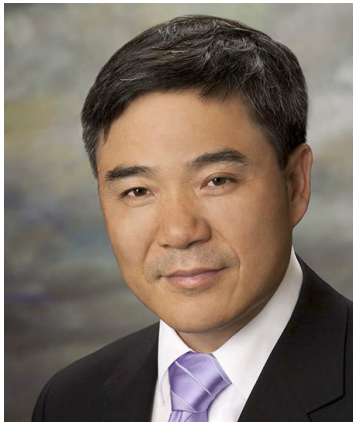
ORFE, Princeton University



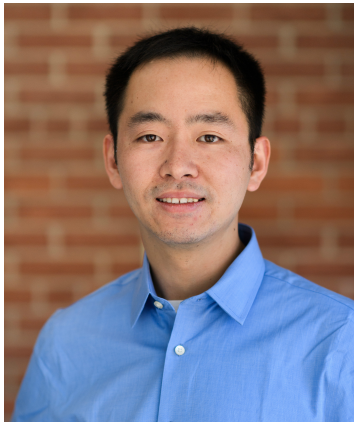
Yuling Yan
Princeton ORFE



Yuejie Chi
CMU ECE



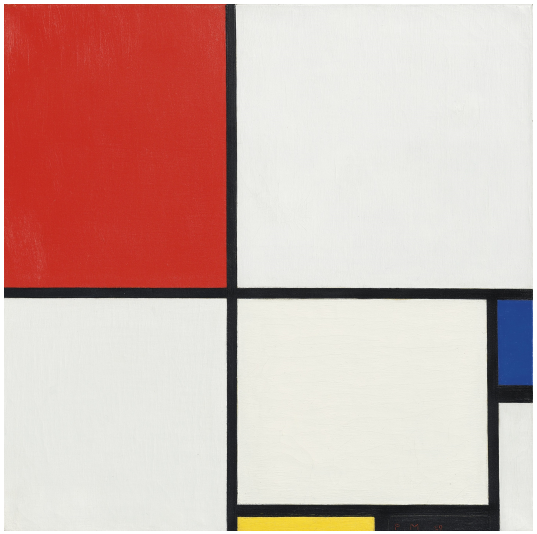
Jianqing Fan
Princeton ORFE



Yuxin Chen
Princeton EE

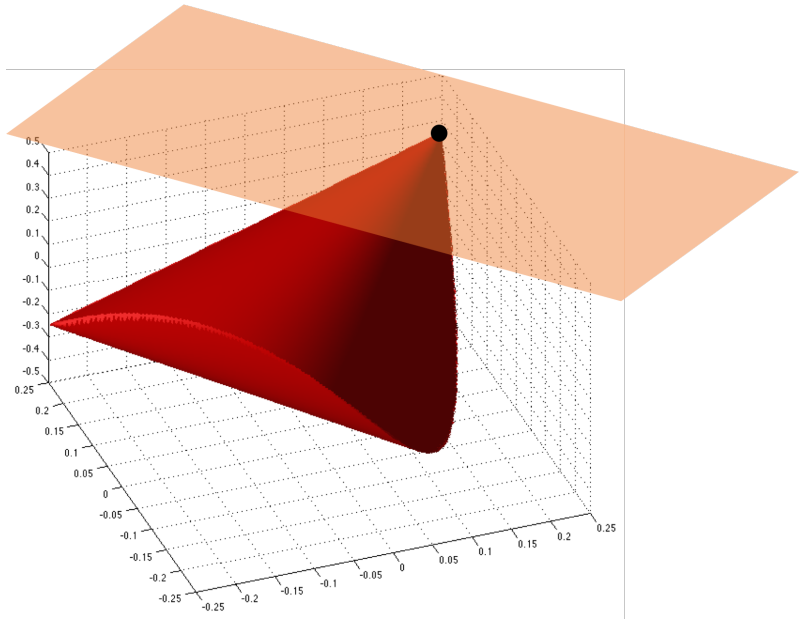
# Convex relaxation for low-rank structure

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



low-rank matrix

*figure credit: Piet Mondrian*



semidefinite relaxation

# Convex relaxation for low-rank structure

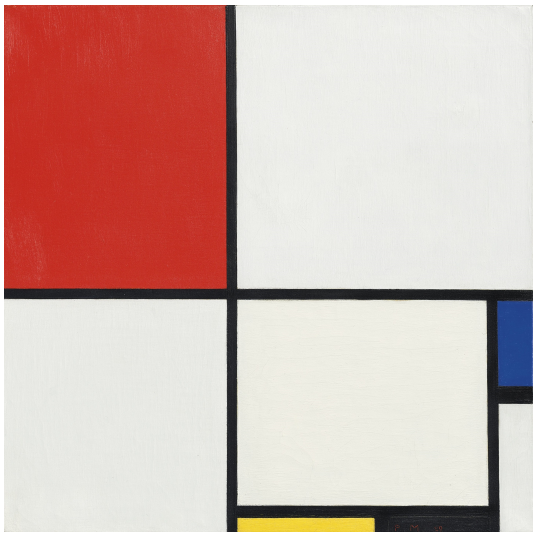
$$\begin{aligned} \underset{Z}{\text{minimize}} \quad & \|Z\|_* \\ \text{subject to} \quad & \text{noiseless data constraints} \end{aligned}$$

- ✓ 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)

...

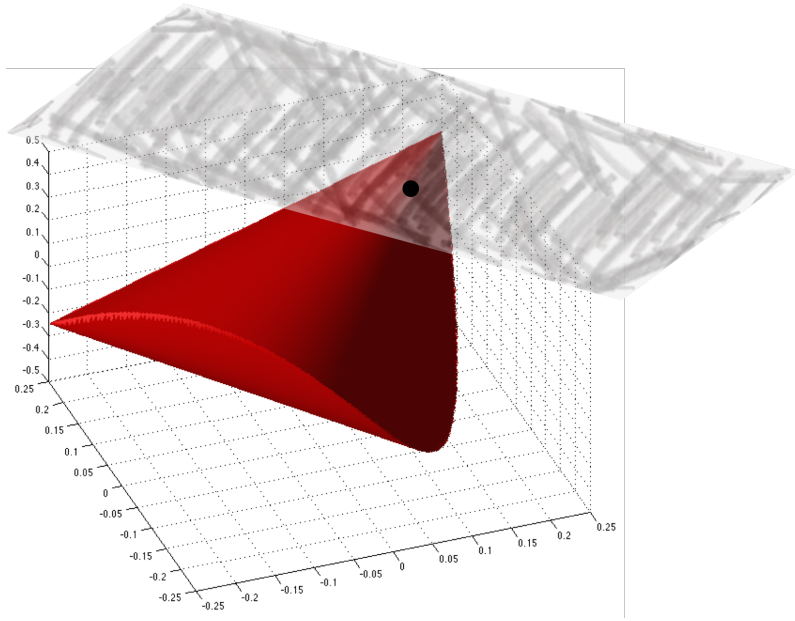
# Stability of convex relaxation against noise

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



low-rank matrix

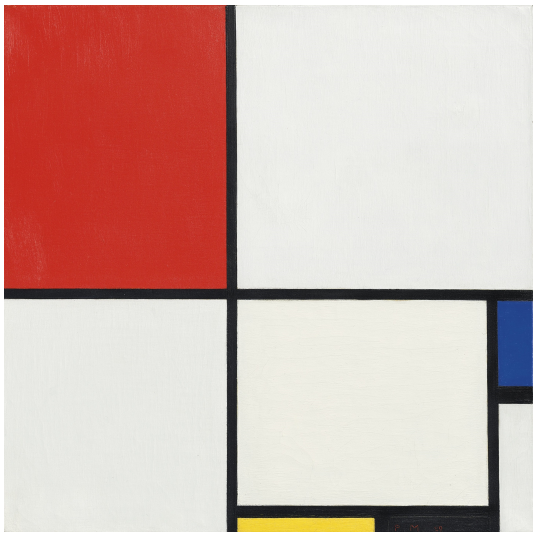
*figure credit: Piet Mondrian*



semidefinite relaxation

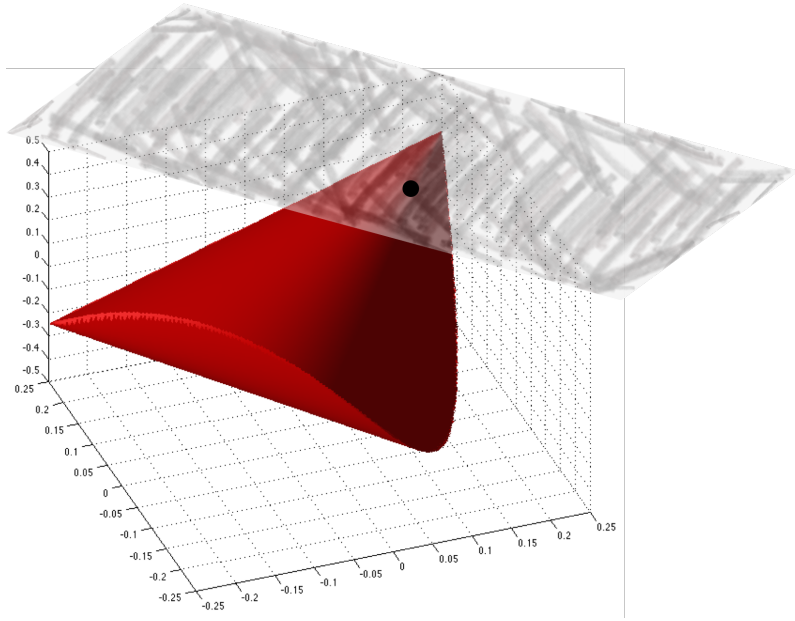
# Stability of convex relaxation against noise

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



low-rank matrix

*figure credit: Piet Mondrian*



semidefinite relaxation

# Stability of convex relaxation against noise

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

- ✓ 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

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

- ✓ 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

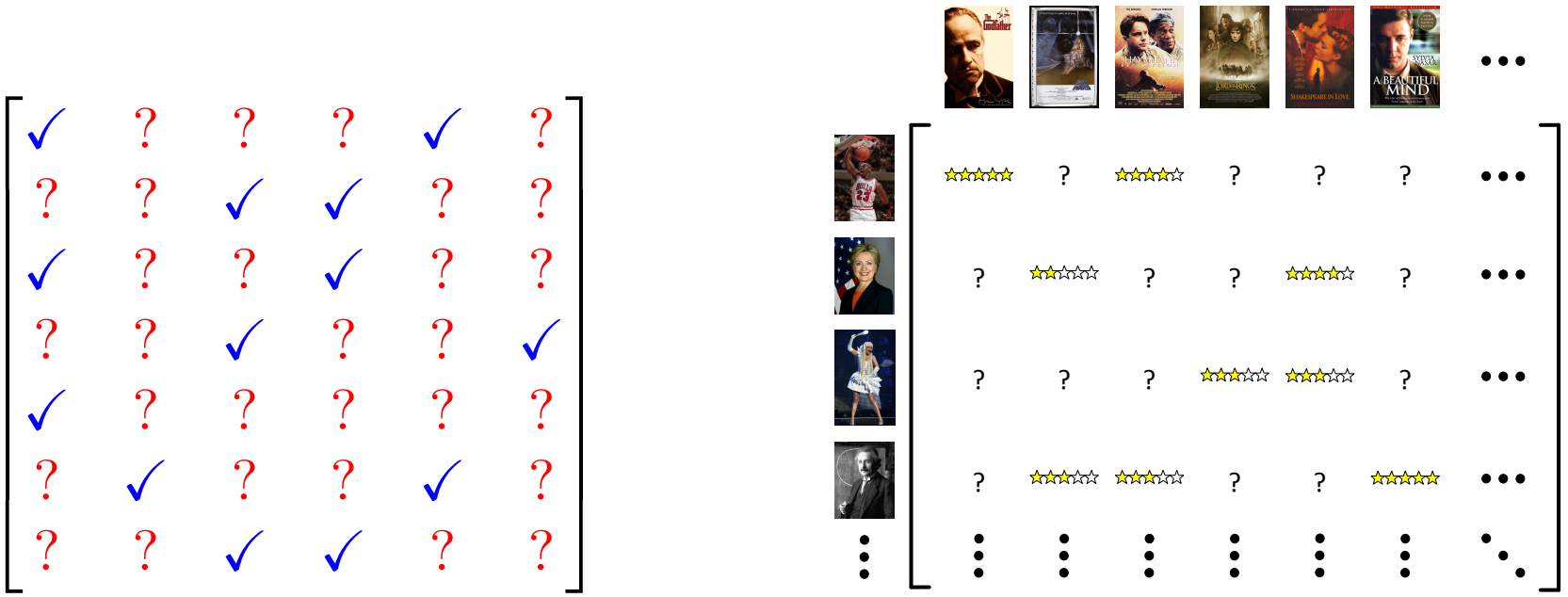


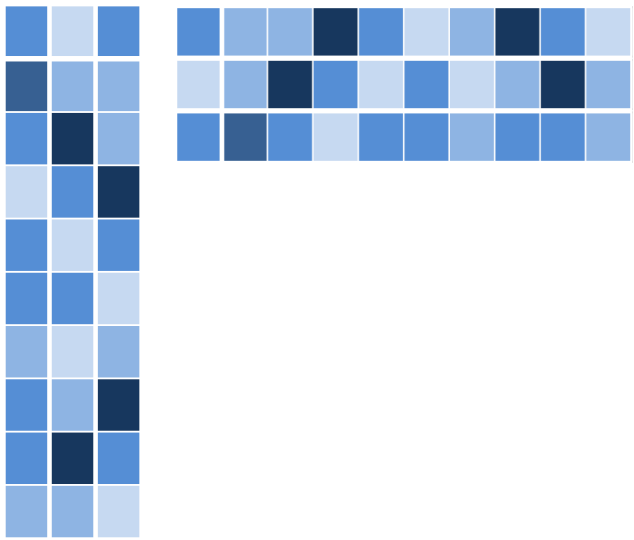
*figure credit: E. J. Candès*

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

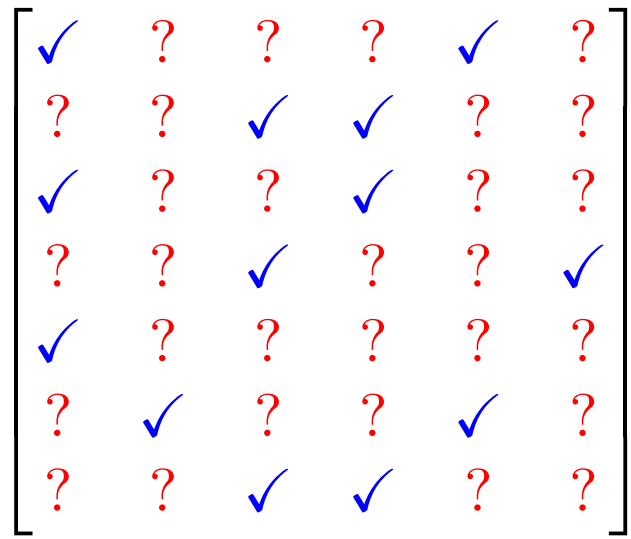
# Noisy low-rank matrix completion

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 $\Omega$

# Noisy low-rank matrix completion

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

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

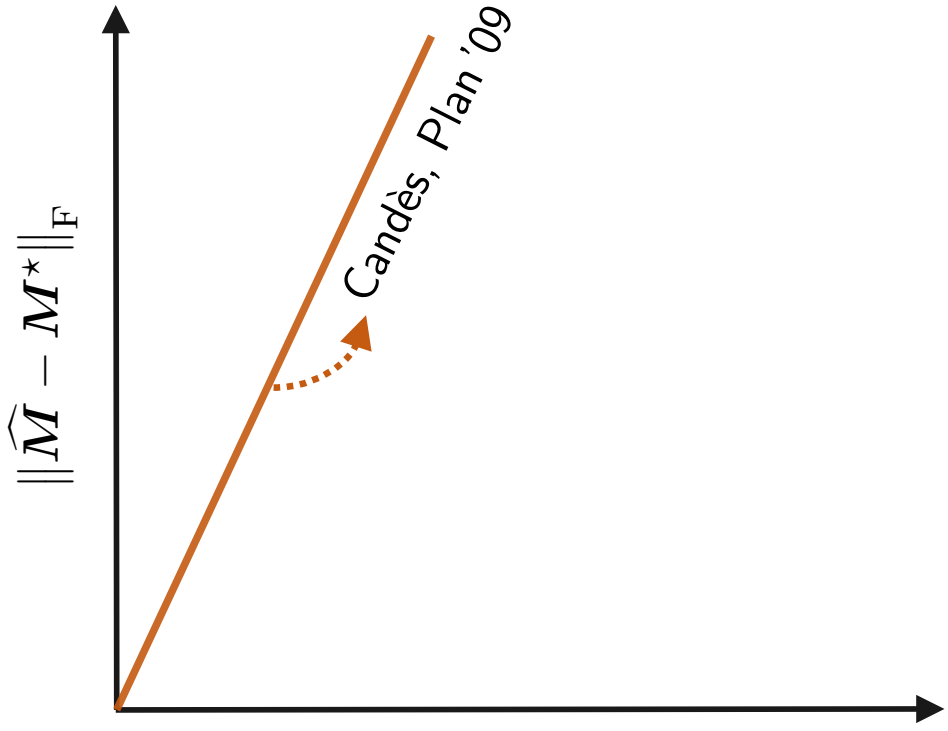
**convex relaxation:**

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

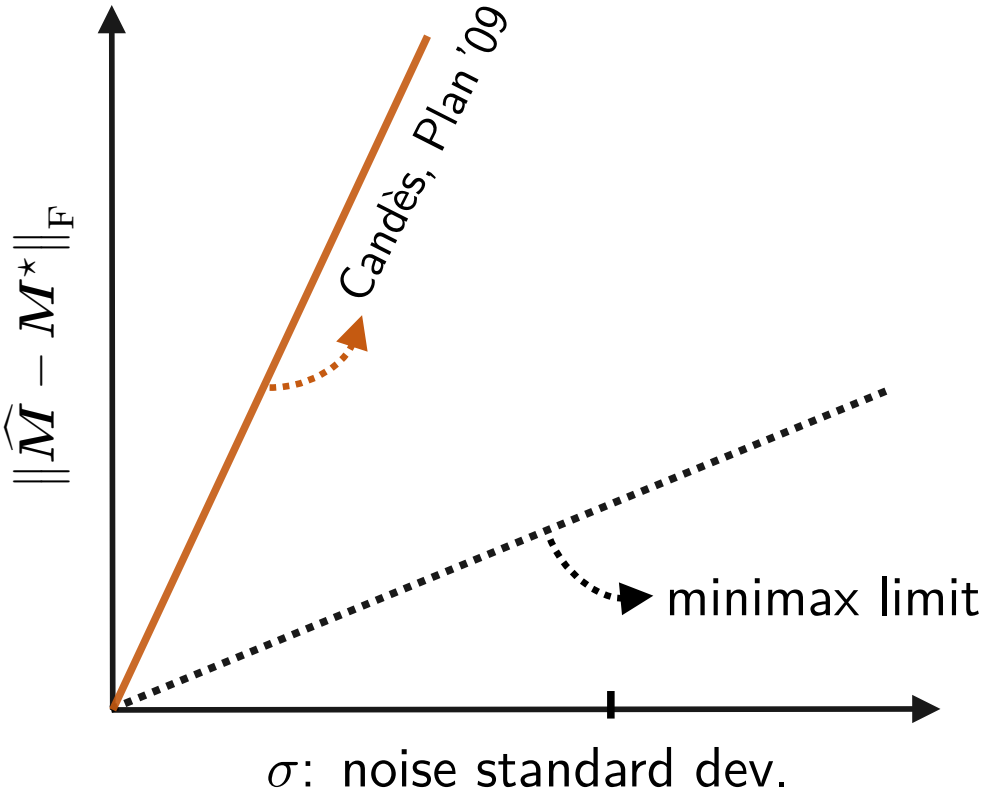
# Prior statistical guarantees for convex relaxation

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

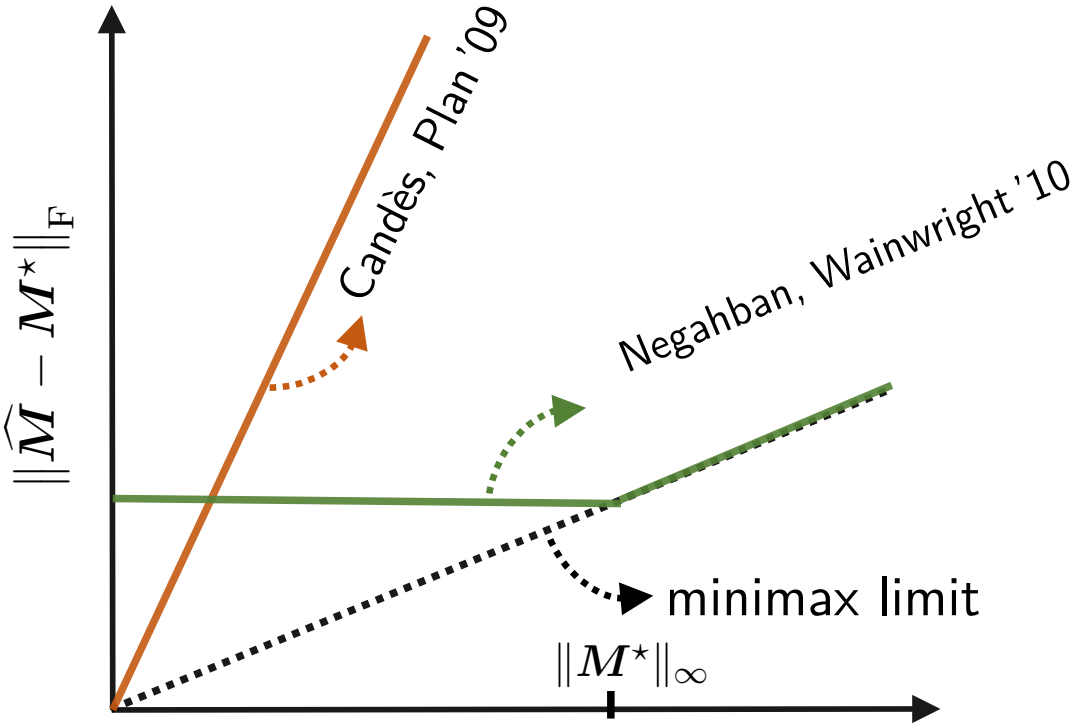
| | |
|---|---|
| Candès, Plan '09 | $\sigma n^{1.5}$ |



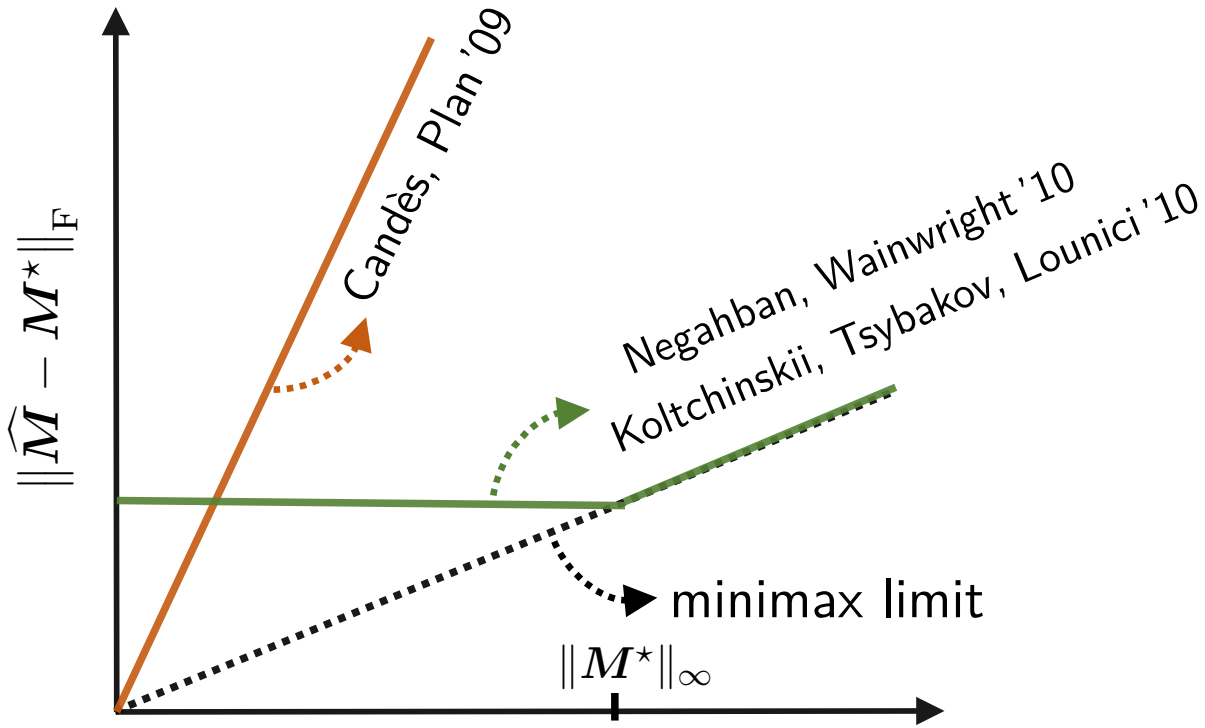
| minimax limit | $\sigma\sqrt{n/p}$ |
|---|---|
| Candès, Plan '09 | $\sigma n^{1.5}$ |



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

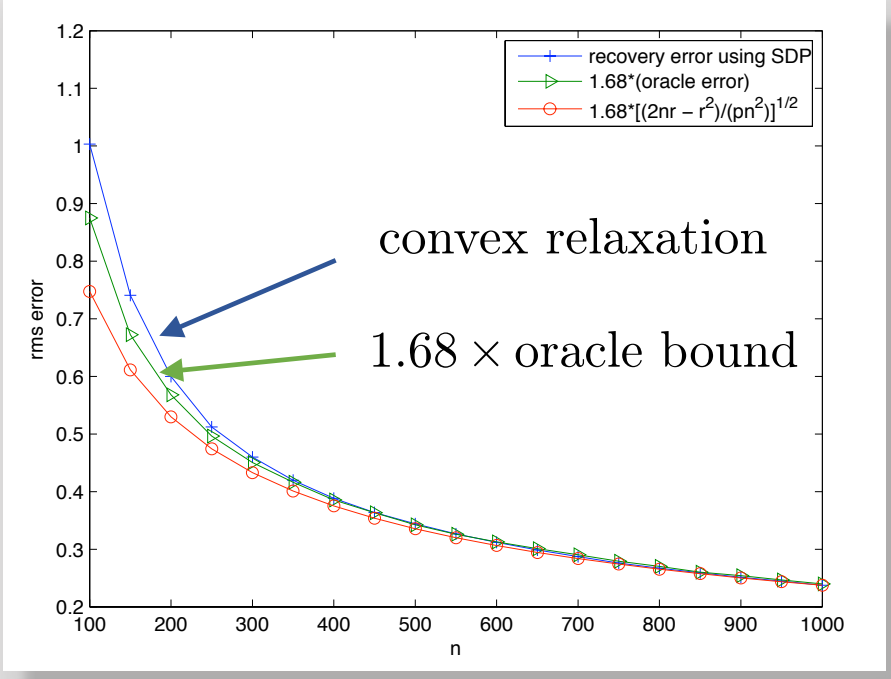


| minimax limit | $\sigma\sqrt{n/p}$ |
|---|---|
| Candès, Plan '09 | $\sigma n^{1.5}$ |
| Negahban, Wainwright '10 | $\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$ |
| Koltchinskii, Tsybakov, Lounici '10 | $\max\{\sigma, \|\boldsymbol{M}^\star\|_\infty\}\sqrt{n/p}$ |



# 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: $\widehat{\boldsymbol{M}}_{\text{cvx}}$ 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

# 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: $\widehat{\boldsymbol{M}}_{\mathsf{cvx}}$ 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

David Gross

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

dual certification (golfing scheme)

dual certification (golfing scheme)

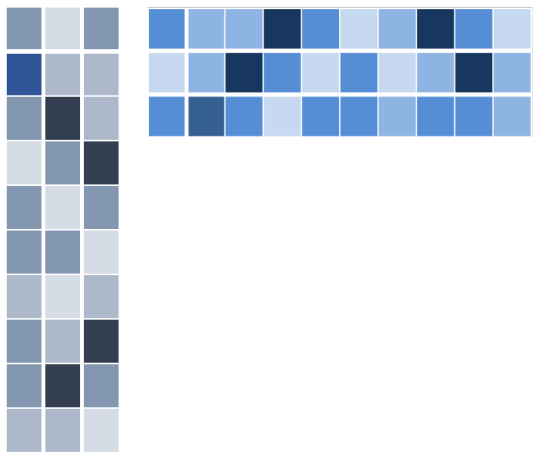


nonconvex optimization

# A detour: nonconvex optimization

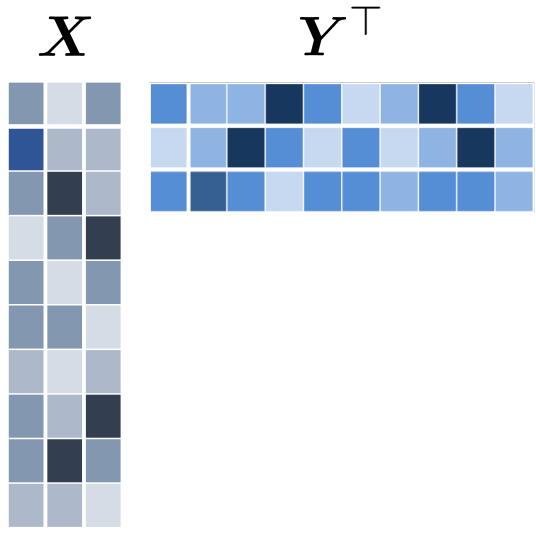
**Burer–Monteiro:** represent $\boldsymbol{Z}$ by $\boldsymbol{X}\boldsymbol{Y}^\top$ with $\underbrace{\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times r}}_{\text{low-rank factors}}$

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



# A detour: nonconvex optimization

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




**nonconvex approach:**

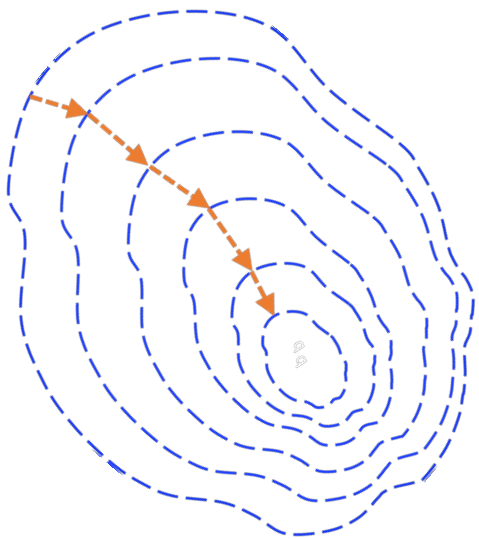
$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \underbrace{\sum_{(i,j)\in\Omega} \left[\left(\boldsymbol{X}\boldsymbol{Y}^\top\right)_{i,j} - M_{i,j}\right]^2}_{\text{squared loss}} + \text{reg}(\boldsymbol{X},\boldsymbol{Y})$$

# 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

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



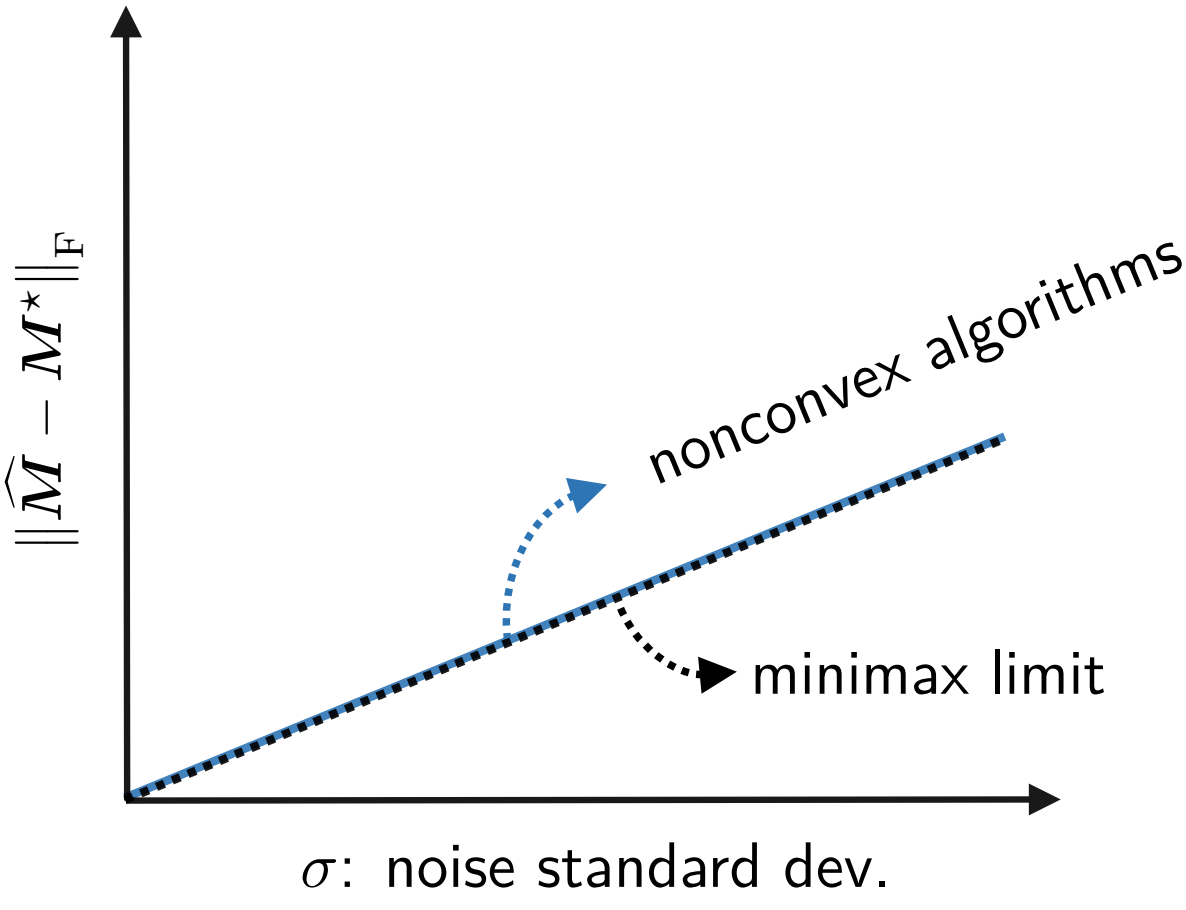
- **suitable initialization:** $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$
- **gradient descent:** for $t = 0, 1, \ldots$

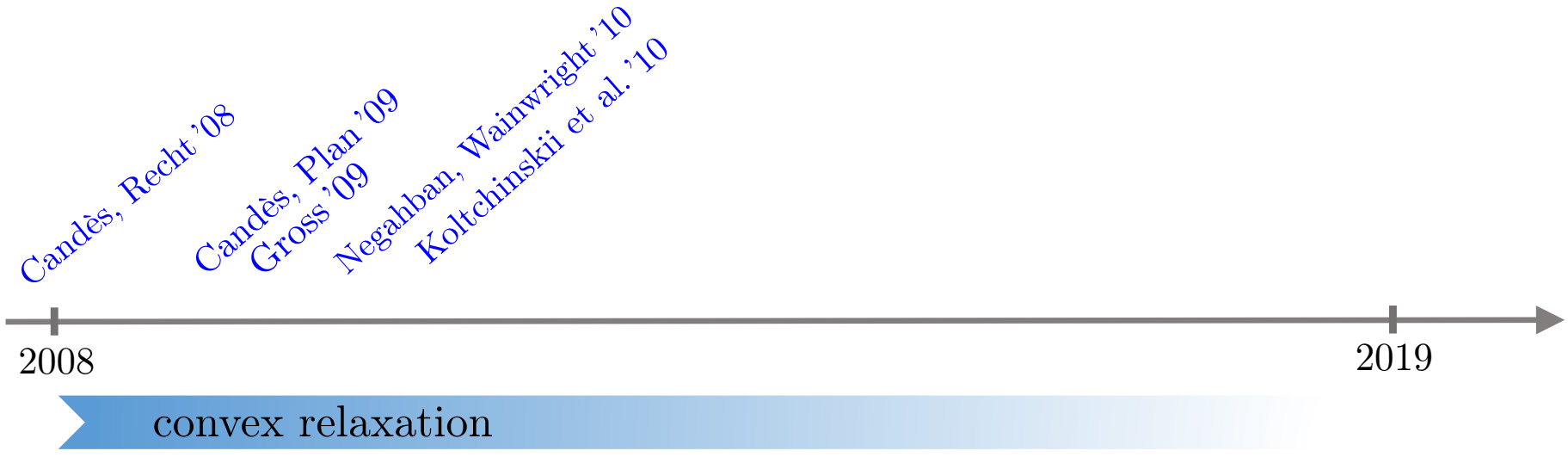
$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$
$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta_t \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$

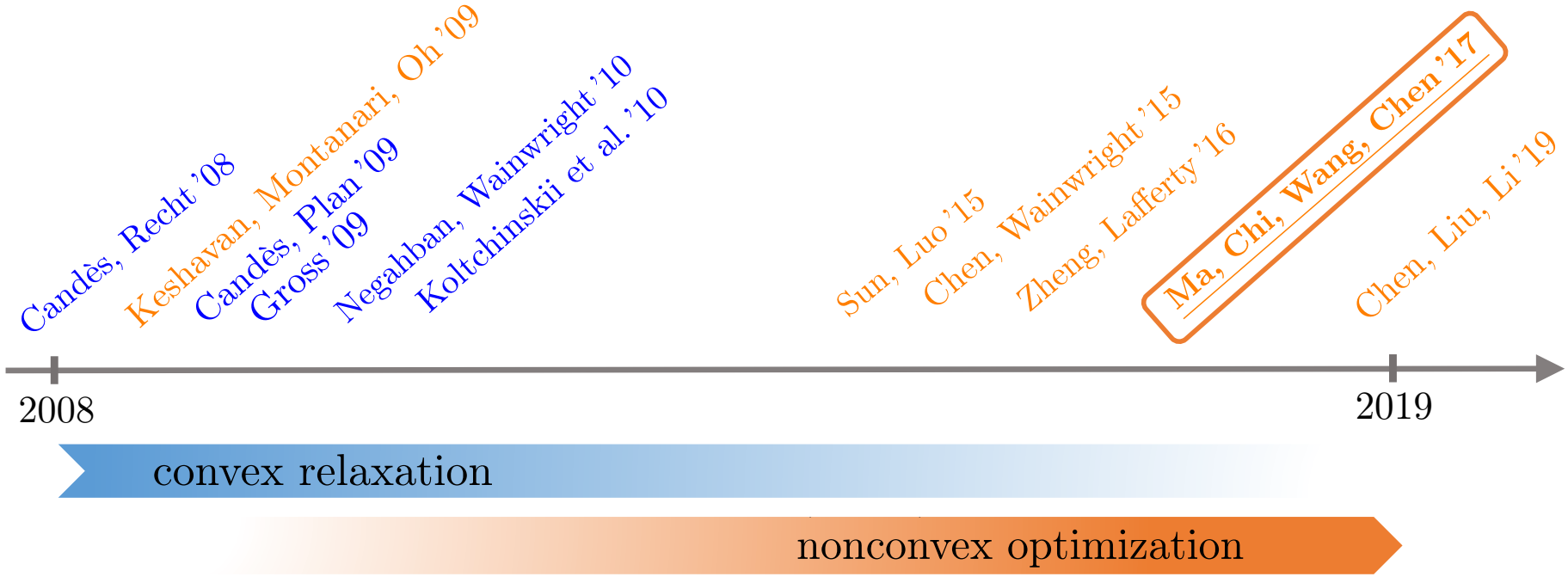
# A detour: nonconvex optimization

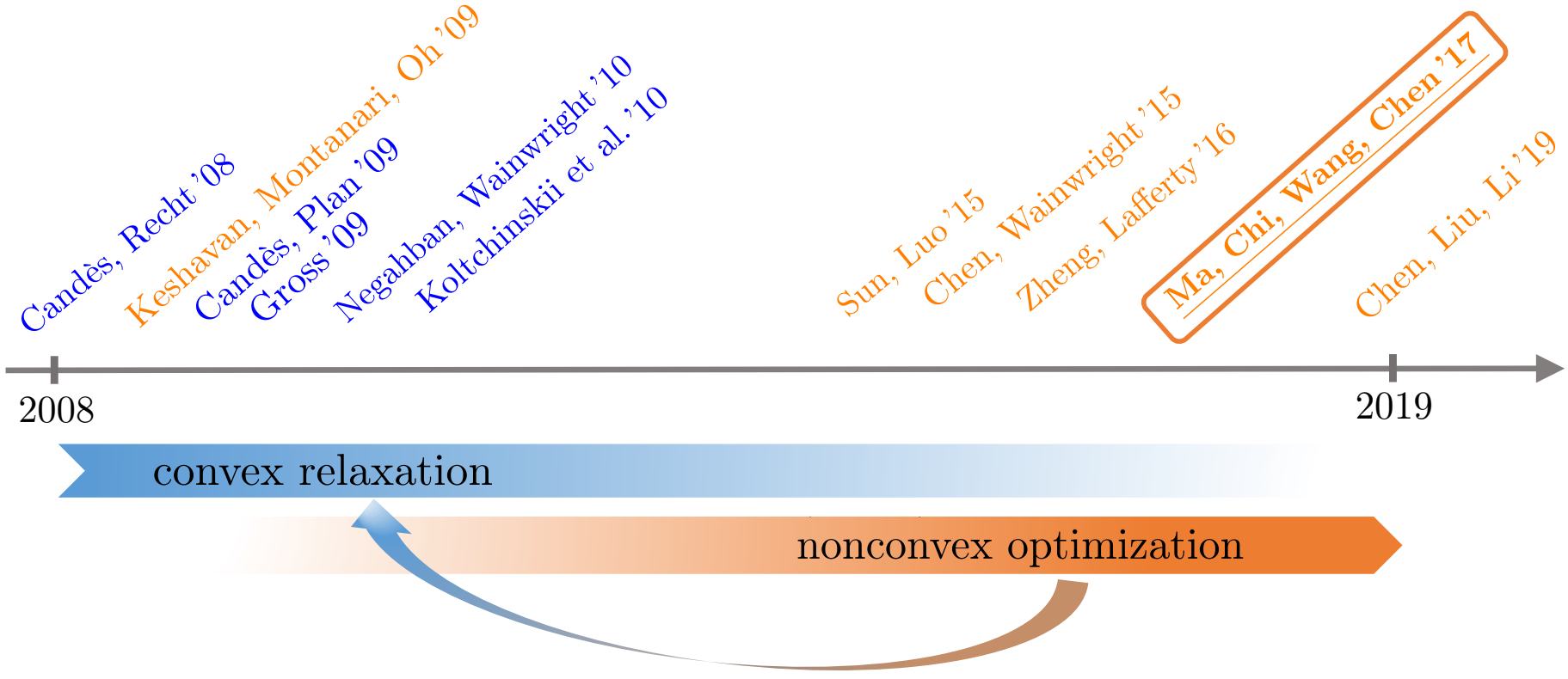
- **random sampling:** each $(i, j) \in \Omega$ with prob. $p$
- **random noise:** i.i.d. sub-Gaussian noise with variance $\sigma^2$
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n \times n}$: $r = O(1)$, incoherent, ...

| 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

# An interesting experiment

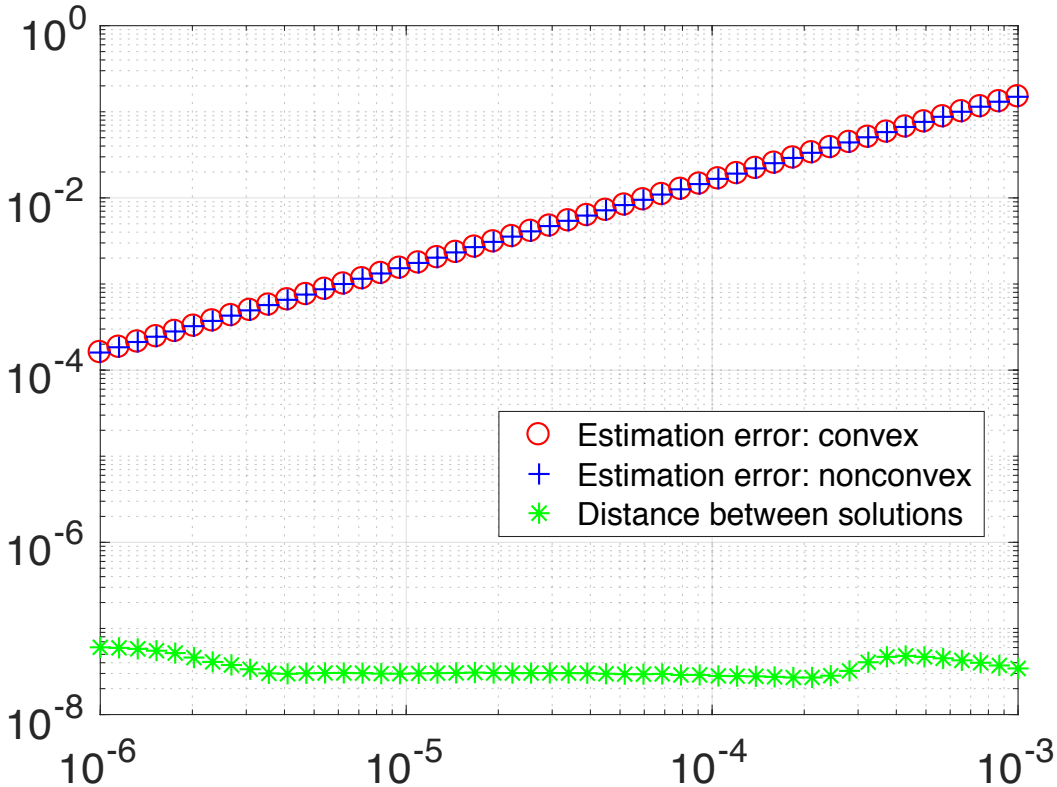
**convex:** $\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_*$

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

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

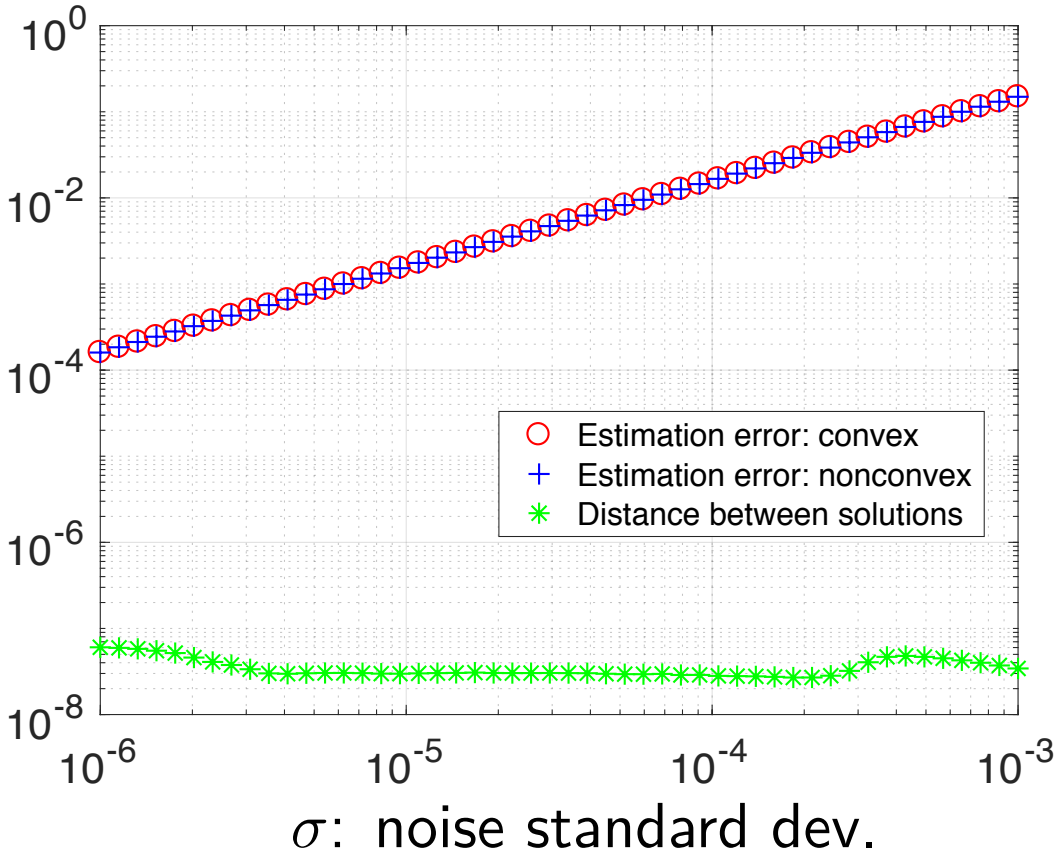
# A motivating experiment

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



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

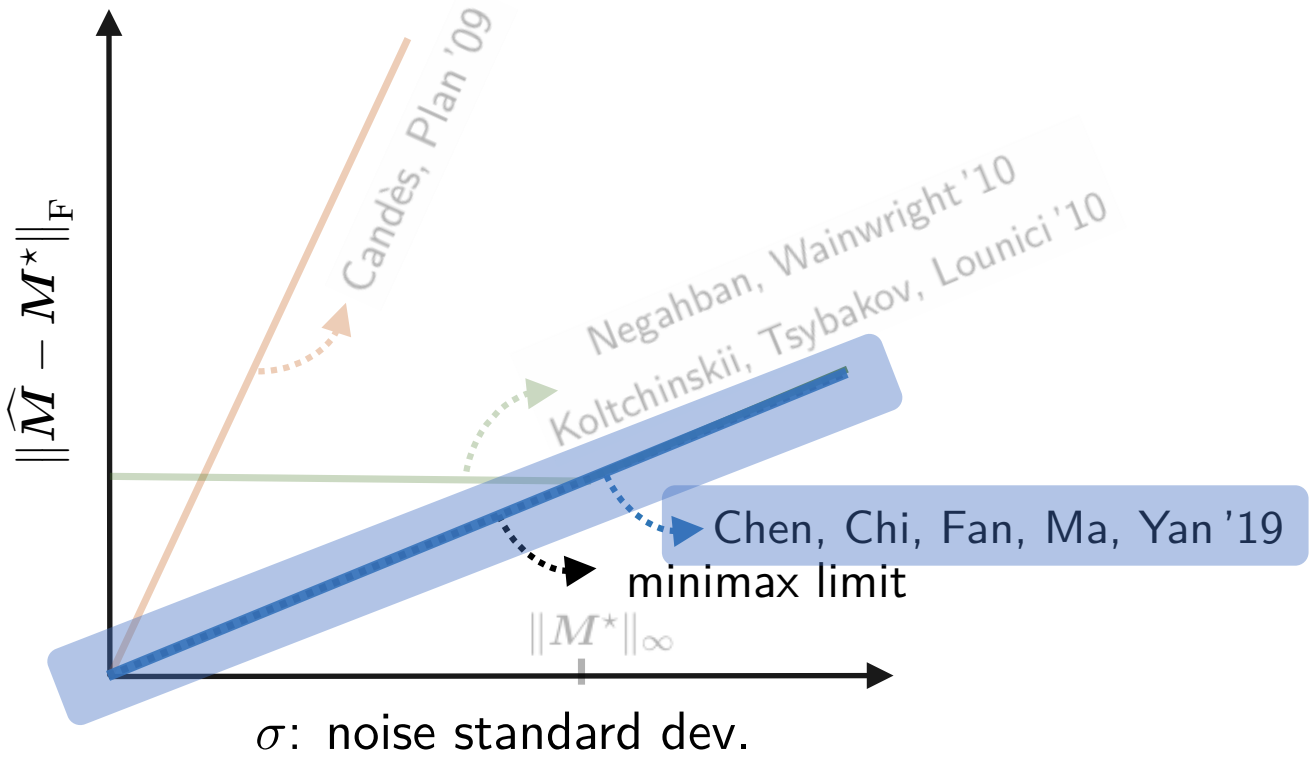
$$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \quad \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_* \qquad (\lambda \asymp \sigma\sqrt{np})$$

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

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

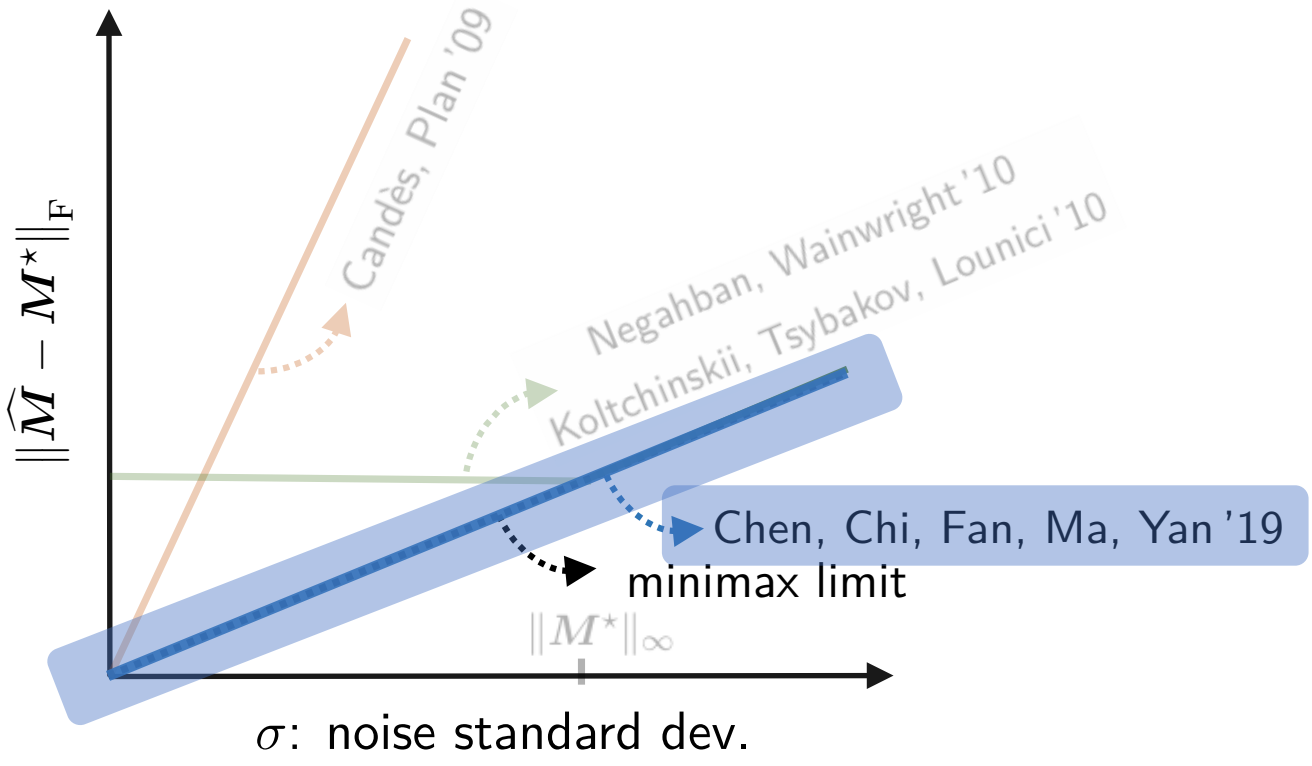
1. *$\widehat{\boldsymbol{M}}_{\mathsf{cvx}}$ is nearly rank-$r$*
2. $$\|\widehat{\boldsymbol{M}}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}$$
$$\|\widehat{\boldsymbol{M}}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\infty} \lesssim \sigma\sqrt{\frac{n\log n}{p}}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

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




- minimax optimal estimation error

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




- minimax optimal estimation error
- estimation errors are spread out across all entries

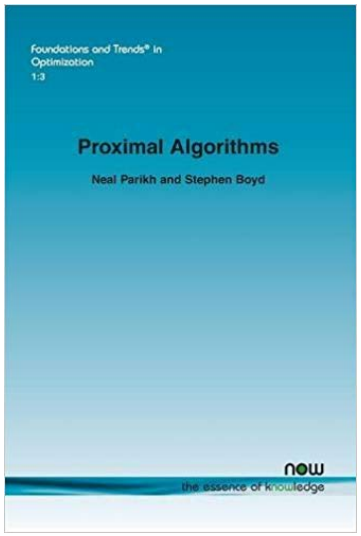
# Implicit regularization

No need to enforce spikiness constraint as in Negahban & Wainwright

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

- convex programming automatically controls spikiness of solutions

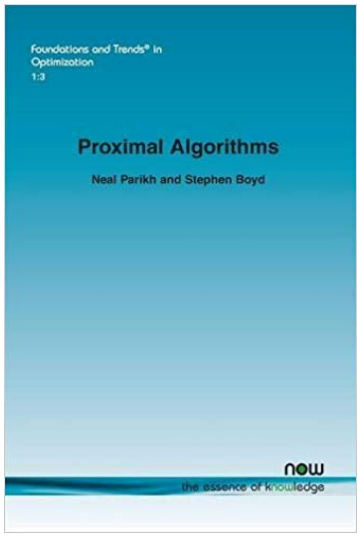
# Statistical guarantees for iterative algorithms




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

Many algorithms (e.g. SVT, SOFT-IMPUTE, FPC, FISTA) have been proposed to minimize $g(\boldsymbol{Z})$, typically without statistical guarantees

# Statistical guarantees for iterative algorithms




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

Many algorithms (e.g. SVT, SOFT-IMPUTE, FPC, FISTA) have been proposed to minimize $g(\boldsymbol{Z})$, 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 $\sigma^2$
- 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 $\sigma^2$
- 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 $\sigma^2$
- 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

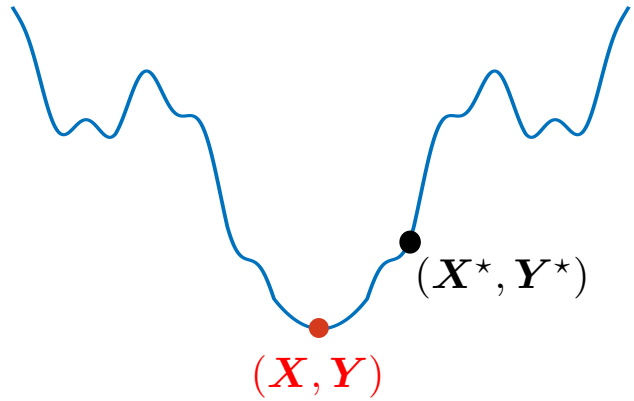
$(\boldsymbol{X}, \boldsymbol{Y})$ is nonconvex optimizer

# 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

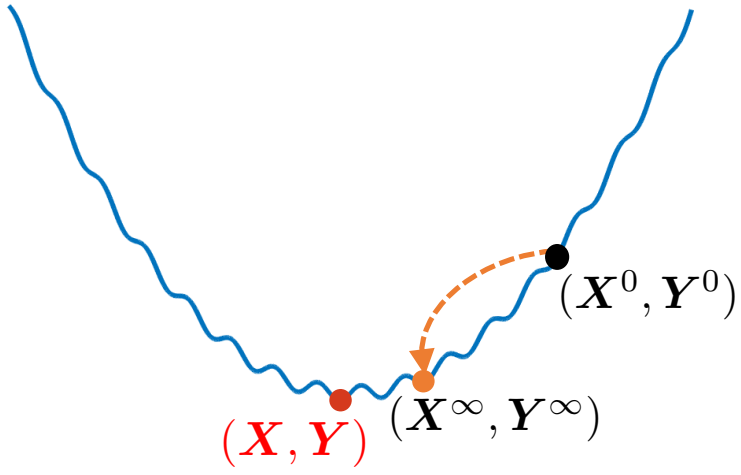
- $\lambda$ is properly chosen
- $(\boldsymbol{X}, \boldsymbol{Y})$ is close to truth (in $\ell_{2,\infty}$ sense)




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

i.e. dist(convex solution, nonconvex solution) $= 0$

# 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

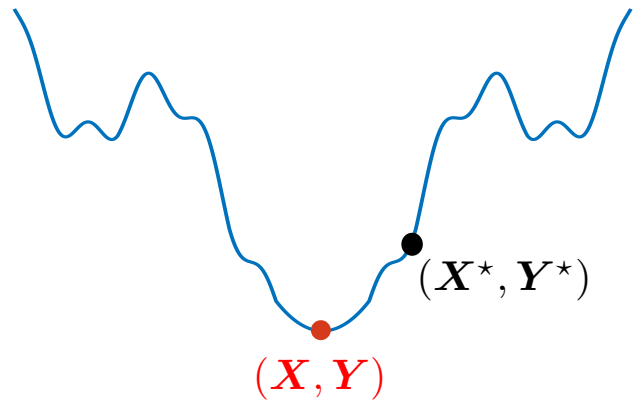
**Strategy:** resort to "$\underbrace{\text{approximate stationary points}}_{\nabla f(\boldsymbol{X},\boldsymbol{Y}) \approx \mathbf{0}}$" instead

# Approximate nonconvex optimizers

**Strategy:** resort to "approximate stationary points" instead

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

- $\lambda$ is properly chosen
- $(\boldsymbol{X}, \boldsymbol{Y})$ is close to truth (in $\ell_{2,\infty}$ sense)




$$\nabla f(\boldsymbol{X}, \boldsymbol{Y}) \approx \boldsymbol{0} \quad \overset{\checkmark}{\Longrightarrow} \quad \text{dist}(\boldsymbol{X}\boldsymbol{Y}^\top, \text{convex solutions}) \approx 0$$

# Construct approximate nonconvex optimizers via GD

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

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

# Construct approximate nonconvex optimizers via GD

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

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

- when $T$ is large: there exists point with $\underbrace{\text{very small gradient}}_{\|\nabla f(\boldsymbol{X}, \boldsymbol{Y})\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{\eta T}}}$

# Construct approximate nonconvex optimizers via GD

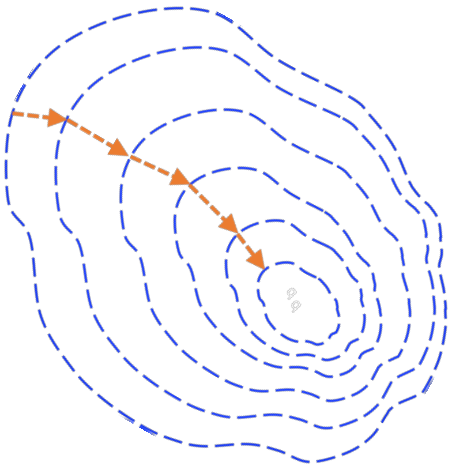
starting from $(\boldsymbol{X}^0, \boldsymbol{Y}^0) =$ truth or spectral initialization:

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

- when $T$ is large: there exists point with $\underbrace{\text{very small gradient}}_{\|\nabla f(\boldsymbol{X}, \boldsymbol{Y})\|_{\mathrm{F}} \lesssim \frac{1}{\sqrt{\eta T}}}$
- hopefully not far from $(\boldsymbol{X}^\star, \boldsymbol{Y}^\star)$ (in $\ell_{2,\infty}$ sense in particular)

*Gradient descent for nonconvex matrix completion*

# Gradient descent for nonconvex matrix completion



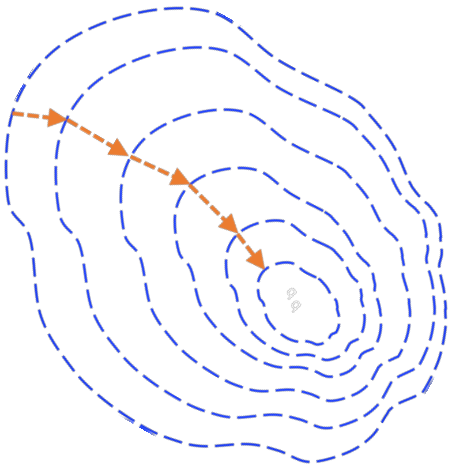
$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \end{aligned}$$

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



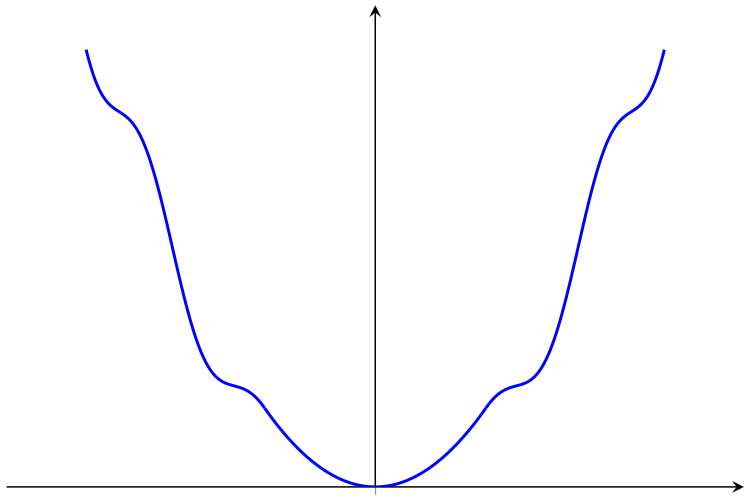
$$\begin{aligned} \boldsymbol{X}^{t+1} &= \boldsymbol{X}^t - \eta \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \\ \boldsymbol{Y}^{t+1} &= \boldsymbol{Y}^t - \eta \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t) \end{aligned}$$

Our work and Chen et al. analyze vanilla GD

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

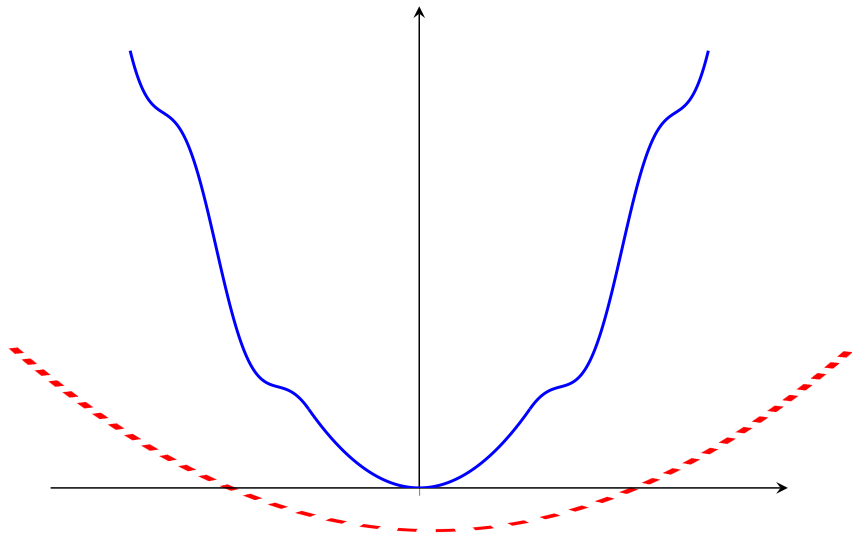
— *Ma, Wang, Chi, Chen '17, Chen, Liu, Li '19*

# Gradient descent theory revisited



Two standard conditions that enable geometric convergence of GD

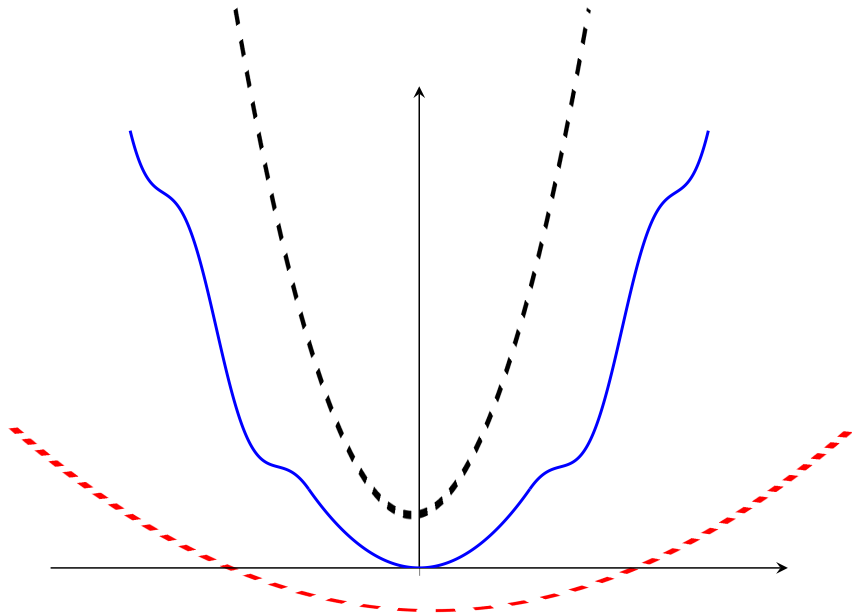
# Gradient descent theory revisited



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 $\alpha$-strongly convex and $\beta$-smooth if

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

**$\ell_2$ error contraction:** GD with $\eta = 1/\beta$ obeys

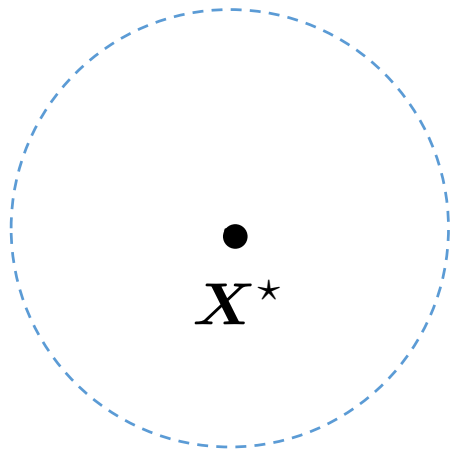
$$\|\boldsymbol{X}^{t+1} - \boldsymbol{X}^\star\|_{\mathrm{F}} \leq \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{X}^t - \boldsymbol{X}^\star\|_{\mathrm{F}}$$

# Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

# Incoherence region

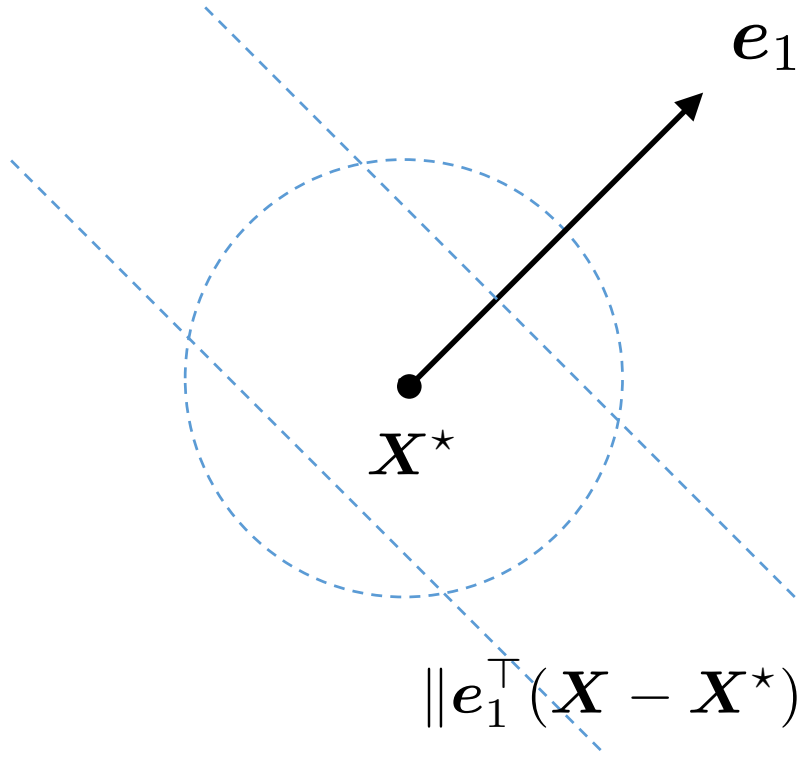
Which region enjoys both restricted strong convexity and smoothness?




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

# Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

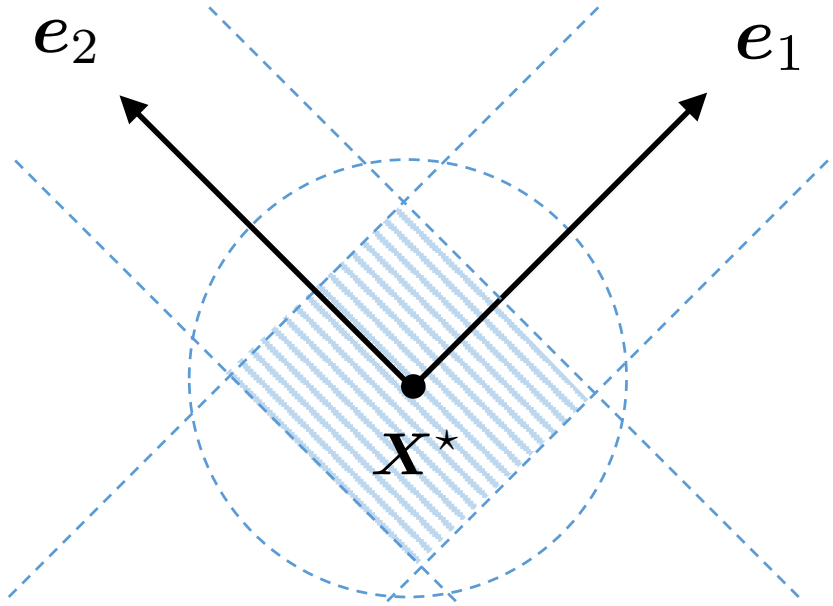



$$\|\boldsymbol{e}_1^\top(\boldsymbol{X}-\boldsymbol{X}^\star)\|_2 \le \epsilon\|\boldsymbol{X}^\star\|_{2,\infty}$$

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

# Incoherence region

Which region enjoys both restricted strong convexity and smoothness?

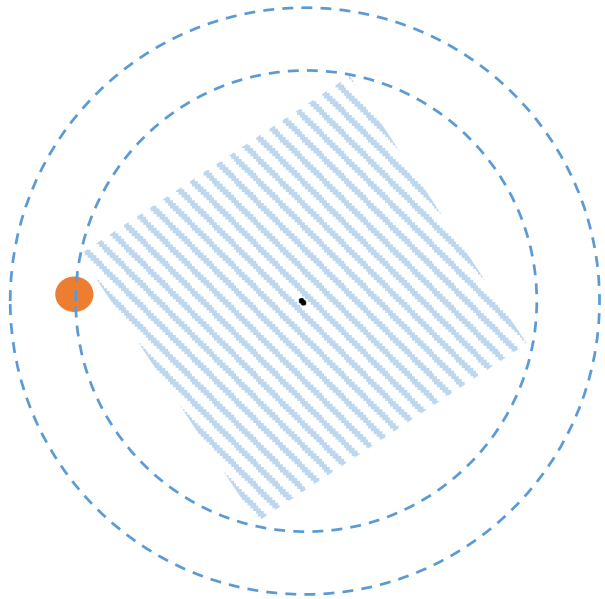



$$\|\boldsymbol{e}_2^\top(\boldsymbol{X}-\boldsymbol{X}^\star)\|_2 \le \epsilon\|\boldsymbol{X}^\star\|_{2,\infty} \quad \|\boldsymbol{e}_1^\top(\boldsymbol{X}-\boldsymbol{X}^\star)\|_2 \le \epsilon\|\boldsymbol{X}^\star\|_{2,\infty}$$

- $\boldsymbol{X}$ 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

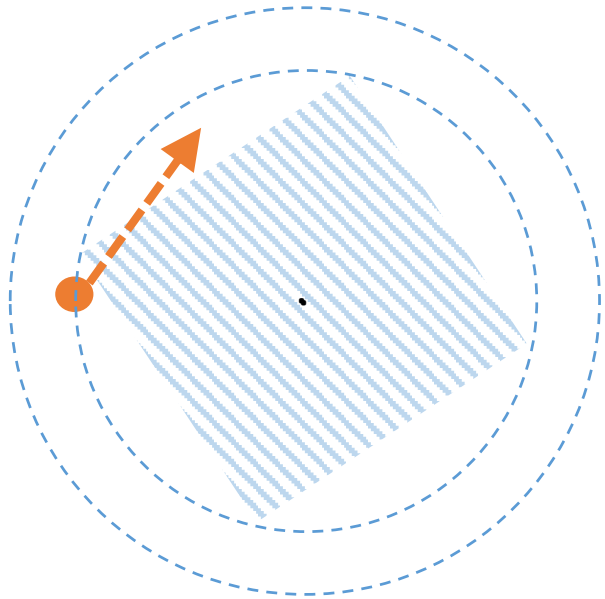
region of local strong convexity + smoothness



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

# Inadequacy of generic gradient descent theory

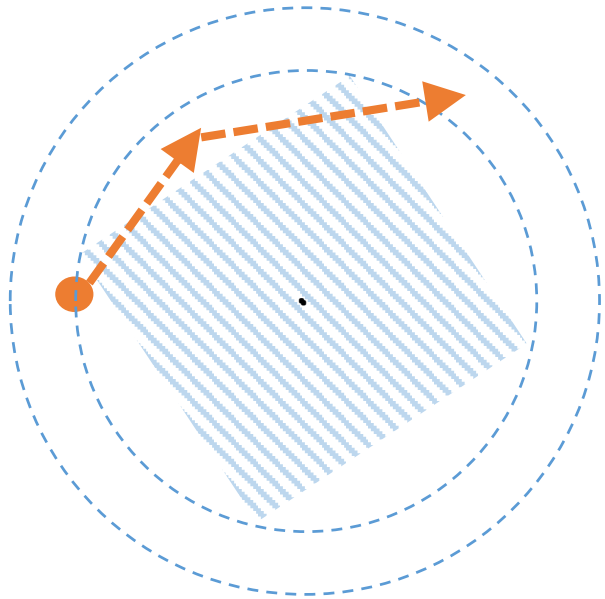
region of local strong convexity + smoothness



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

# Inadequacy of generic gradient descent theory

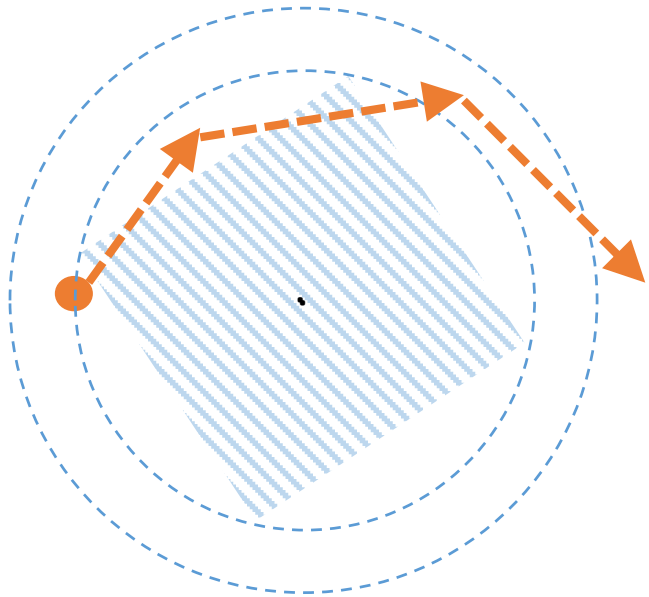
region of local strong convexity + smoothness



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

# Inadequacy of generic gradient descent theory

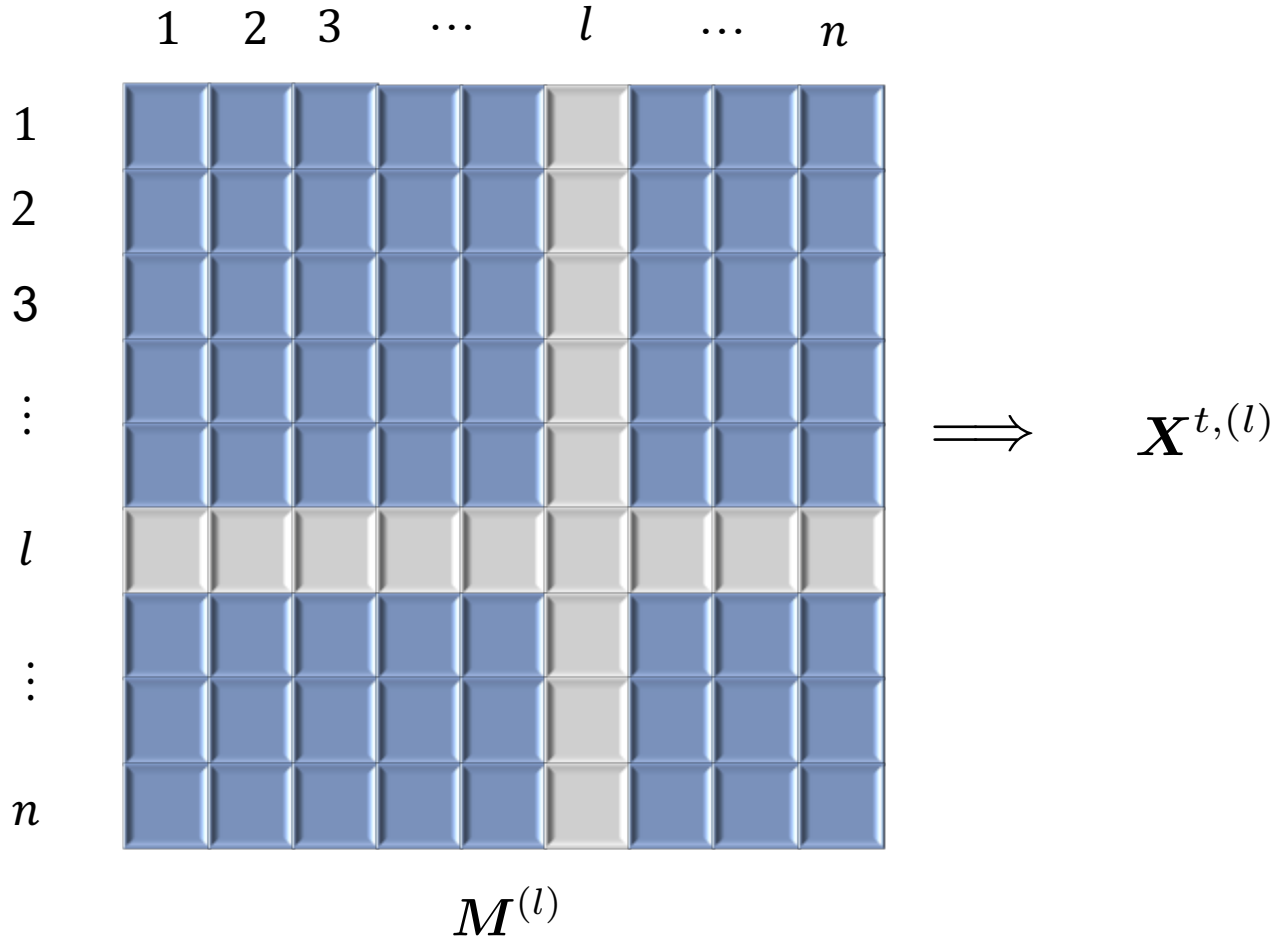
region of local strong convexity + smoothness



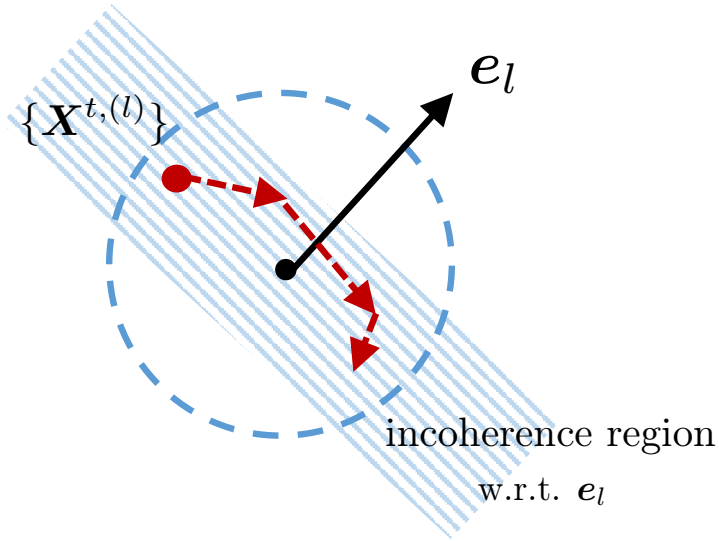
- Generic optimization theory does NOT ensure GD stays in incoherence region
- Demonstrating incoherence calls for new analysis tools

# 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 $\{l^{\text{th}}\}$ row and column with true values

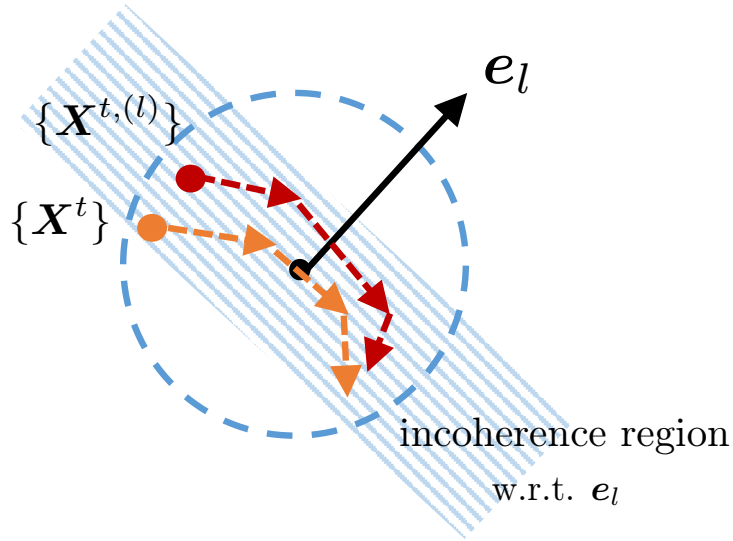


# Key proof idea: leave-one-out analysis

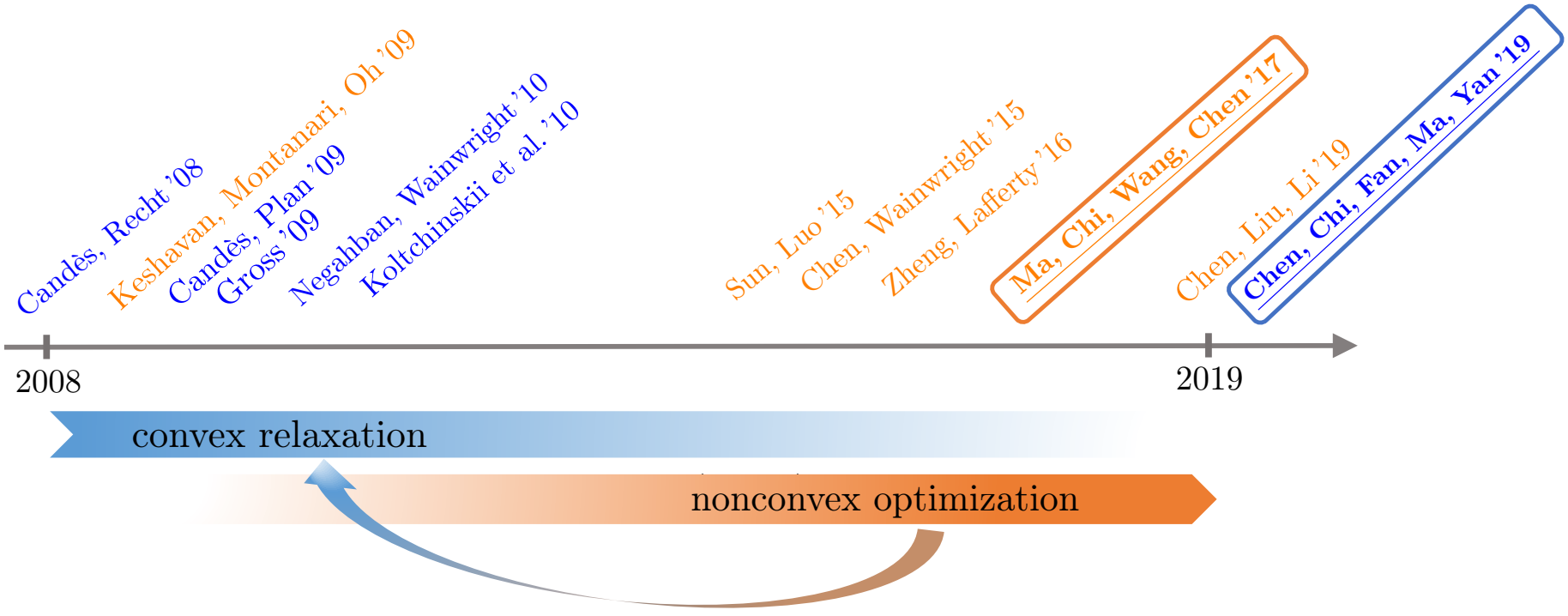



- Leave-one-out iterates $\{\boldsymbol{X}^{t,(l)}\}$ contains more information of $l^{\text{th}}$ row of truth; indep. of randomness in $l^{\text{th}}$ row

# Key proof idea: leave-one-out analysis




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




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