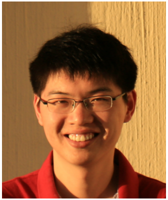
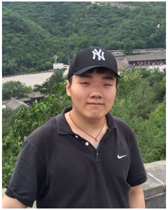
Nonconvex Optimization Meets Statistics: A Few Recent Stories

Yuxin Chen

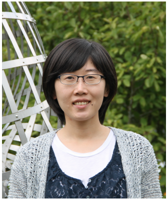
Electrical Engineering, Princeton University



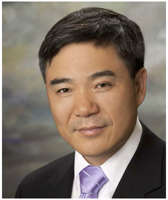
Cong Ma
Princeton ORFE



Yuling Yan
Princeton ORFE



Yuejie Chi
CMU ECE



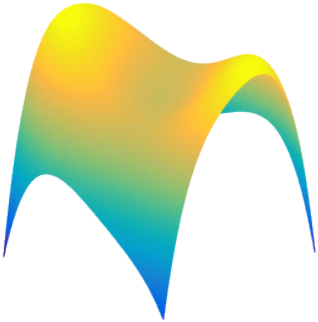
Jianqing Fan
Princeton ORFE

Nonconvex problems are everywhere

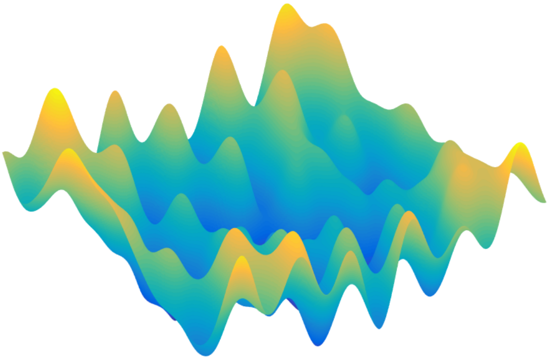
Empirical risk minimization is usually nonconvex

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}; \text{data})$$

- low-rank matrix completion
- blind deconvolution
- dictionary learning
- mixture models
- deep neural nets
- ...



Nonconvex optimization may be super scary

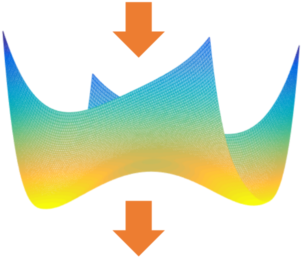


There may be bumps everywhere and exponentially many local optima

e.g. 1-layer neural net (Auer, Herbster, Warmuth '96; Vu '98)

Statistical models come to rescue

statistical models

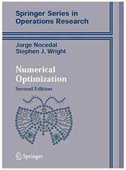


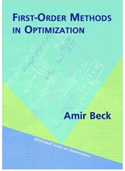
benign landscape

tractable algorithms

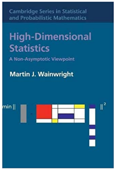
When data are generated by certain statistical models, problems are often much nicer than worst-case instances

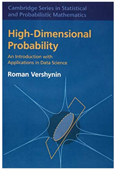
— *Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview*
Chi, Lu, Chen '18





nonconvex optimization





(high-dimensional) statistics

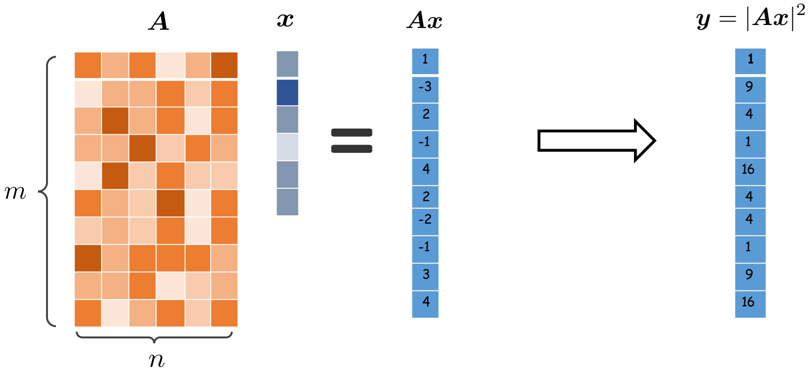
1. Random initialization when solving random quadratic systems
   — *optimal computational efficiency*

2. Inference and uncertainty quantification for matrix completion
   — *a distributional theory*

3. Bridging convex & nonconvex optimization in matrix completion
   — *an implicit gift*

Random initialization when solving random quadratic systems of equations

Solving quadratic systems of equations



Estimate $\boldsymbol{x}^\star \in \mathbb{R}^n$ from m random quadratic measurements

$$y_k = (\boldsymbol{a}_k^\top \boldsymbol{x}^\star)^2, \qquad k = 1, \ldots, m$$

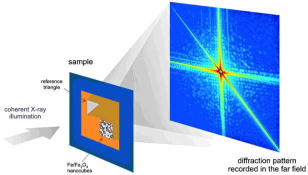
assume w.l.o.g. $\|\boldsymbol{x}^\star\|_2 = 1$

Motivation: phase retrieval

Detectors record intensities of diffracted rays

- electric field $x(t_1, t_2) \longrightarrow$ Fourier transform $\widehat{x}(f_1, f_2)$

Fig credit: Stanford SLAC

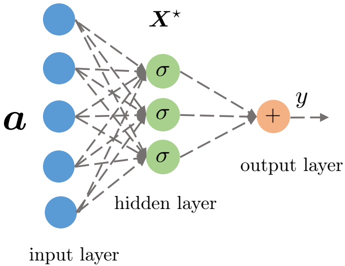


intensity of electrical field: $\left|\widehat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

Phase retrieval: recover signal $x(t_1, t_2)$ from intensity $\left|\widehat{x}(f_1, f_2)\right|^2$

Motivation: learning neural nets with quadratic activation

— *Soltanolkotabi, Javanmard, Lee '17, Li, Ma, Zhang '17*



input features: $\boldsymbol{a}$; weights: $\boldsymbol{X}^\star = [\boldsymbol{x}_1^\star, \cdots, \boldsymbol{x}_r^\star]$

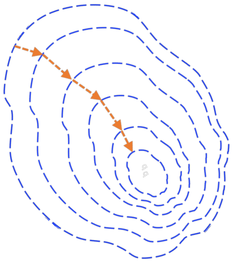
$$\text{output:} \quad y = \sum_{i=1}^{r} \sigma(\boldsymbol{a}^\top \boldsymbol{x}_i^\star) \overset{\sigma(z)=z^2}{=} \sum_{i=1}^{r} (\boldsymbol{a}^\top \boldsymbol{x}_i^\star)^2$$

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{k=1}^{m} \left[\left(\boldsymbol{a}_k^\top \boldsymbol{x} \right)^2 - y_k \right]^2$$

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

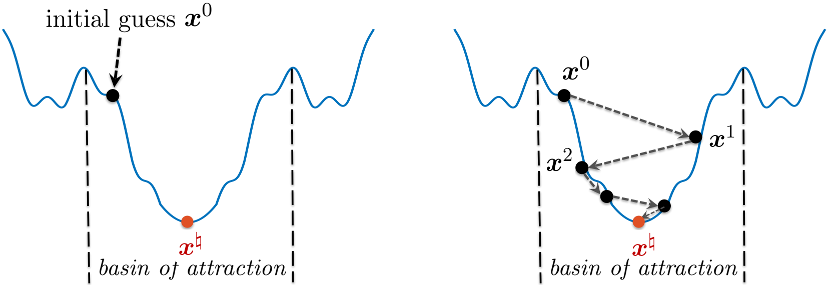
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\left[\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - y_k\right]^2$$



- **spectral initialization:** $x^0 \leftarrow$ leading eigenvector of certain data matrix
- **gradient descent:**

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta_t \nabla f(\boldsymbol{x}^t), \qquad t = 0, 1, \cdots$$

Rationale of two-stage approach



1. initialize within $\underbrace{\text{local basin sufficiently close to } \boldsymbol{x}^\star}_{\text{(restricted) strongly convex; no saddles / spurious local mins}}$

2. iterative refinement

A highly incomplete list of two-stage methods

phase retrieval:

- Netrapalli, Jain, Sanghavi '13
- Candès, Li, Soltanolkotabi '14
- Chen, Candès '15
- Cai, Li, Ma '15
- Wang, Giannakis, Eldar '16
- Zhang, Zhou, Liang, Chi '16
- Kolte, Ozgur '16
- Zhang, Chi, Liang '16
- Soltanolkotabi '17
- Vaswani, Nayer, Eldar '16
- Chi, Lu '16
- Wang, Zhang, Giannakis, Akcakaya, Chen '16
- Tan, Vershynin '17
- Ma, Wang, Chi, Chen '17
- Duchi, Ruan '17
- Jeong, Gunturk '17
- Yang, Yang, Fang, Zhao, Wang, Neykov '17
- Qu, Zhang, Wright '17
- Goldstein, Studer '16
- Bahmani, Romberg '16
- Hand, Voroninski '16
- Wang, Giannakis, Saad, Chen '17
- Barmherzig, Sun '17
- ...

other problems:

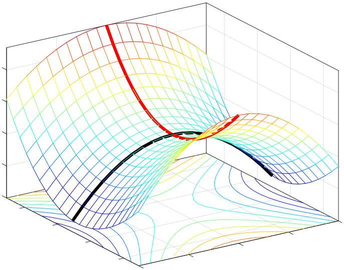
- Keshavan, Montanari, Oh '09
- Sun, Luo '14
- Chen, Wainwright '15
- Tu, Boczar, Simchowitz, Soltanolkotabi, Recht '15
- Zheng, Lafferty '15
- Balakrishnan, Wainwright, Yu '14
- Chen, Suh '15
- Chen, Candès '16
- Li, Ling, Strohmer, Wei '16
- Yi, Park, Chen, Caramanis '16
- Jin, Kakade, Netrapalli '16
- Huang, Kakade, Kong, Valiant '16
- Ling, Strohmer '17
- Aghasi, Ahmed, Hand '17
- Lee, Tian, Romberg '17
- Li, Chi, Zhang, Liang '17
- Cai, Wang, Wei '17
- Abbe, Bandeira, Hall '14
- Chen, Kamath, Suh, Tse '16
- Zhang, Zhou '17
- Boumal '16
- Zhong, Boumal '17
- Li, Ma, Chen, Chi '18
- Chen, Liu, Li '19
- Charisopoulos, Davis, Diaz, Drusvyatskiy '19
- Charisopoulos, Chen, Davis, Diaz, Ding, Drusvyatskiy '19
- ...

Is carefully-designed initialization necessary for fast convergence?

Is carefully-designed initialization necessary for fast convergence?

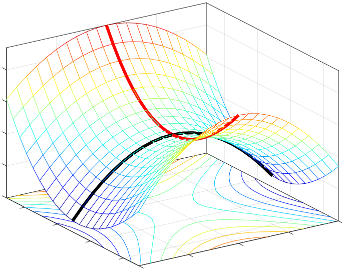
Can we initialize GD randomly, which is simpler and model-agnostic?

What does prior theory say?



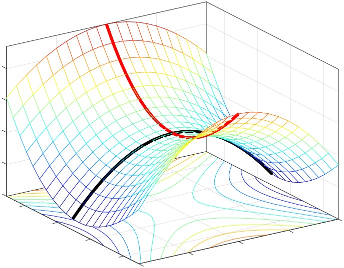
- **landscape:** no spurious local mins (Sun, Qu, Wright '16)

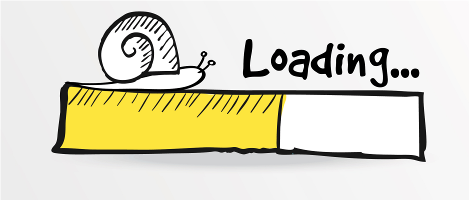
What does prior theory say?



- **landscape:** no spurious local mins (Sun, Qu, Wright '16)
- randomly initialized GD converges almost surely (Lee et al. '16)

What does prior theory say?



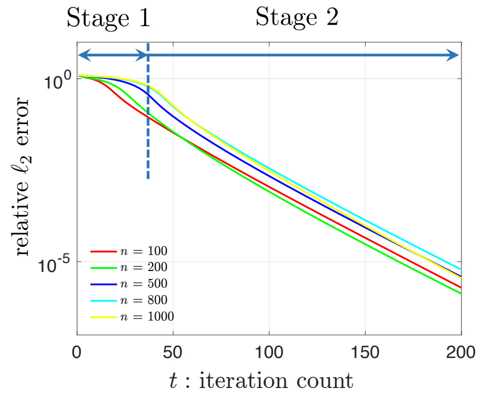


- **landscape:** no spurious local mins (Sun, Qu, Wright '16)
- randomly initialized GD converges almost surely (Lee et al. '16)

"almost surely" might mean "take forever"

Numerical efficiency of randomly initialized GD

$$\eta = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\mathbf{0}, n^{-1}\boldsymbol{I}_n)$$



Randomly initialized GD enters local basin within **tens of iterations**

Our theory

These numerical findings can be formalized when $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$:

Our theory

These numerical findings can be formalized when $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$:

$$\text{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) := \min\{\|\boldsymbol{x}^t \pm \boldsymbol{x}^\star\|_2\}$$

Theorem 1 (Chen, Chi, Fan, Ma '18)

Under i.i.d. Gaussian design, GD with $\boldsymbol{x}^0 \sim \mathcal{N}(\mathbf{0}, n^{-1}\boldsymbol{I}_n)$ *achieves*

Our theory

These numerical findings can be formalized when $\boldsymbol{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$:

$$\text{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) := \min\{\|\boldsymbol{x}^t \pm \boldsymbol{x}^\star\|_2\}$$

Theorem 1 (Chen, Chi, Fan, Ma '18)

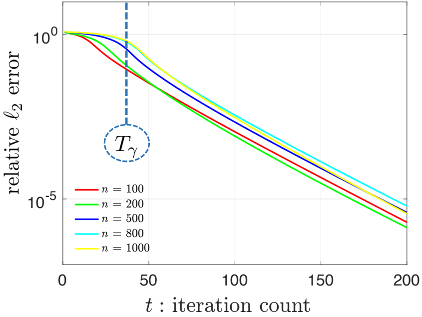
Under i.i.d. Gaussian design, GD with $\boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$ achieves

$$\text{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \qquad t \ge T_\gamma$$

with high prob. for $T_\gamma \lesssim \log n$ and some constants $\gamma, \rho > 0$, provided that step size $\eta \asymp 1$ and sample size $m \gtrsim n \,\text{poly}\log m$

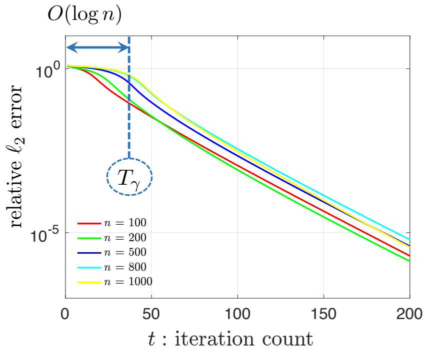
Our theory

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$



Our theory

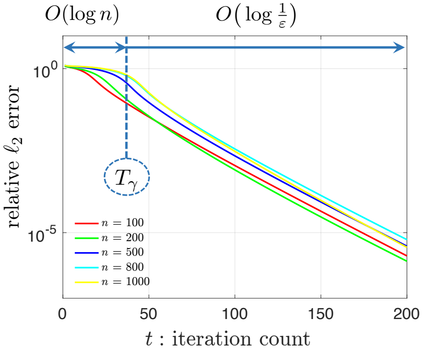
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \leq \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \geq T_\gamma \asymp \log n$$



- *Stage 1:* takes $O(\log n)$ iterations to reach $\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \leq \gamma$ (e.g. $\gamma = 0.1$)

Our theory

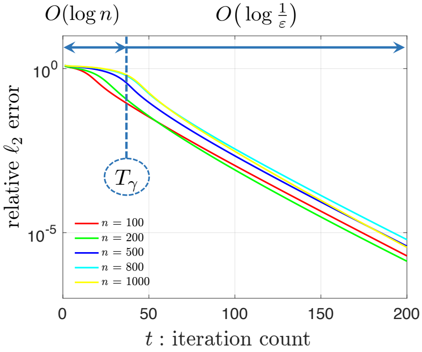
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma (1-\rho)^{t-T_\gamma} \|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$



- *Stage 1:* takes $O(\log n)$ iterations to reach $\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma$ (e.g. $\gamma = 0.1$)
- *Stage 2:* linear (geometric) convergence

Our theory

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$

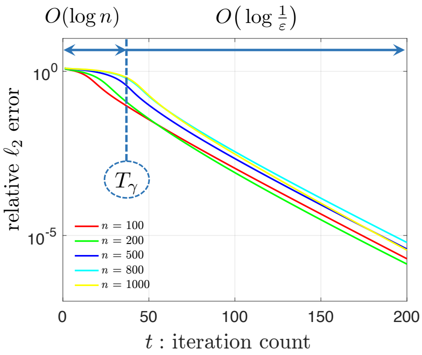


- *near-optimal computational cost*:

— $O\left(\log n + \log \frac{1}{\varepsilon}\right)$ iterations to yield ε accuracy

Our theory

$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \le \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \ge T_\gamma \asymp \log n$$

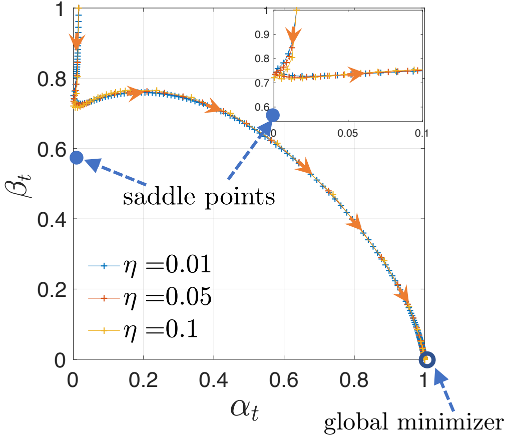


- *near-optimal computational cost*:

  — $O\left(\log n + \log \frac{1}{\varepsilon}\right)$ iterations to yield ε accuracy

- *near-optimal sample size*: $m \gtrsim n\mathrm{poly}\log m$

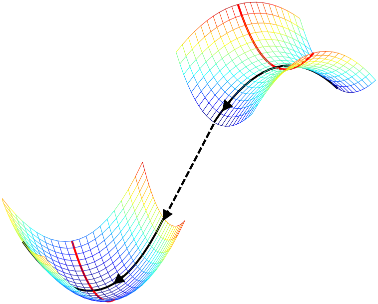
Automatic saddle avoidance



Randomly initialized GD never hits saddle points!

Other saddle-escaping schemes based on generic landscape analysis

	iteration complexity
trust-region (Sun et al. '16)	$n^7 + \log\log\frac{1}{\varepsilon}$
perturbed GD (Jin et al. '17)	$n^3 + n\log\frac{1}{\varepsilon}$
perturbed accelerated GD (Jin et al. '17)	$n^{2.5} + \sqrt{n}\log\frac{1}{\varepsilon}$
GD (ours) (Chen et al. '18)	$\log n + \log\frac{1}{\varepsilon}$



Generic optimization theory yields highly suboptimal convergence guarantees

Even simplest nonconvex methods
are remarkably efficient under suitable statistical models

smart initialization	extra regularization	sample splitting	saddle escaping
			

1. "Gradient Descent with Random Initialization: ...", Y. Chen, Y. Chi, J. Fan, C. Ma, *Mathematical Programming*, vol. 176, no. 1-2, pp. 5-37, 2019

2. "Implicit regularization in nonconvex statistical estimation: ...", C. Ma, K. Wang, Y. Chi, Y. Chen, accepted to *Foundations of Computational Mathematics*, 2019

3. "Nonconvex optimization meets low-rank matrix factorization: An overview", Y. Chi, Y. Lu, Y. Chen, *IEEE Trans. Signal Processing*, vol. 67, no. 20, pp. 5239-5269, 2019

Inference and uncertainty quantification for noisy matrix completion

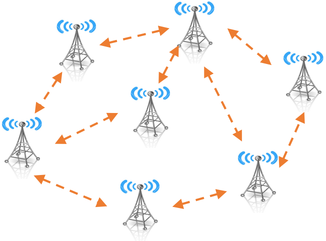
Low-rank matrix completion

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

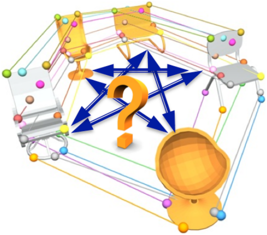
figure credit: E. J. Candès

Given partial samples of a <u>low-rank</u> matrix $M^\star$, fill in missing entries

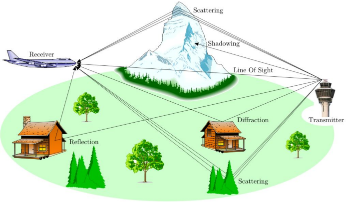
recommendation systems



localization



shape matching



channel estimation

Noisy low-rank matrix completion

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

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

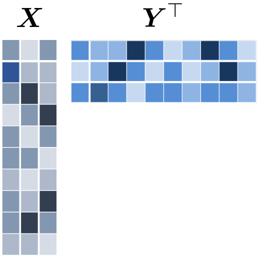
unknown rank-r matrix $\boldsymbol{M}^{\star} \in \mathbb{R}^{n \times n}$

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

sampling set Ω

Nonconvex matrix completion

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



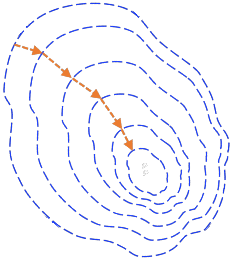
$$\underset{\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^{n\times r}}{\text{minimize}} \quad f(\boldsymbol{X},\boldsymbol{Y}) = \underbrace{\sum_{(i,j)\in\Omega}\Big[\big(\boldsymbol{X}\boldsymbol{Y}^\top\big)_{i,j} - M_{i,j}\Big]^2}_{\text{squared loss}} + \text{reg}(\boldsymbol{X},\boldsymbol{Y})$$

Nonconvex matrix completion

- 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
- Charisopoulos, Chen, Davis, Diaz, Ding, Drusvyatskiy '19
- ...

Nonconvex matrix completion

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



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

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

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

One step further: reasoning about uncertainty?

	2		2	
		6		
3	1		4	
		4		1
	0			

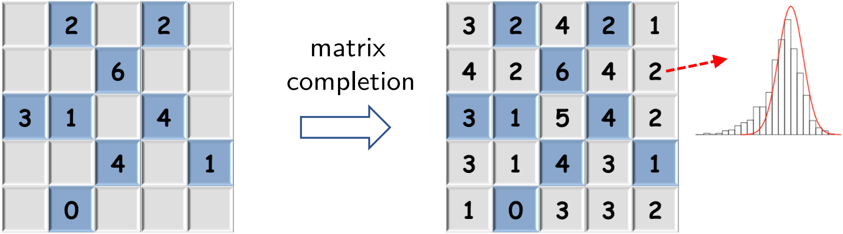
One step further: reasoning about uncertainty?

	2		2	
		6		
3	1		4	
		4		1
	0			

matrix completion

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

One step further: reasoning about uncertainty?



How to assess uncertainty, or “confidence”, of obtained estimates due to imperfect data acquisition?

- noise
- incomplete measurements
- $\cdots$

INFERENCE IN HIGH DIMENSIONAL REGRESSION

organized by
Peter Buehlmann, Andrea Montanari, and Jonathan Taylor

American Institute of Mathematics

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

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

Challenges

$$M^{\mathsf{ncvx}} \longleftarrow \arg\min_{X,Y} \quad \underbrace{f(X, Y; \mathsf{data})}_{\text{empirical loss}} + \mathsf{reg}(X, Y)$$

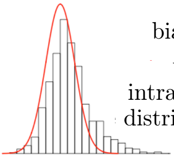
- very challenging to pin down distributions of obtained estimates
  $\longrightarrow$ due to nonconvexity

Challenges

$$\boldsymbol{M}^{\mathsf{ncvx}} \longleftarrow \arg\min_{\boldsymbol{X},\boldsymbol{Y}} \quad \underbrace{f(\boldsymbol{X},\boldsymbol{Y};\mathsf{data})}_{\text{empirical loss}} + \mathsf{reg}(\boldsymbol{X},\boldsymbol{Y})$$

- very challenging to pin down distributions of obtained estimates
  $\longrightarrow$ due to nonconvexity
- existing estimation error bounds are highly sub-optimal
  $\longrightarrow$ overly wide confidence intervals

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



biased

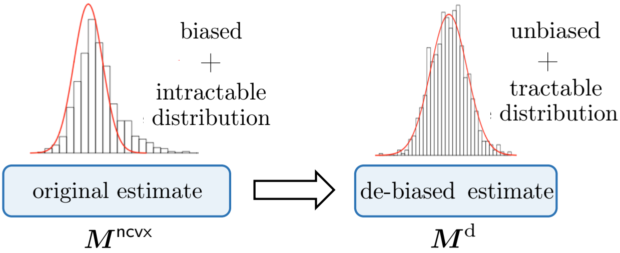
+

intractable distribution

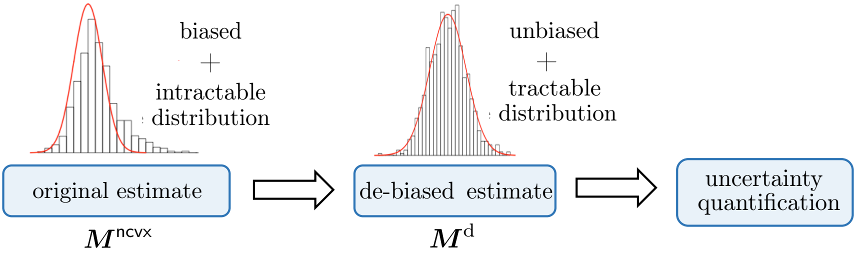
original estimate

$\boldsymbol{M}^{\mathsf{ncvx}}$

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



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



Model

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

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

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

De-biasing nonconvex estimate

$$\boldsymbol{M}^{\textsf{ncvx}} \xrightarrow{\text{de-biasing}} \underbrace{\boldsymbol{M}^{\textsf{ncvx}} + \underbrace{\frac{1}{p}\mathcal{P}_{\Omega}}_{\text{mean: } \mathcal{I}}\left(\boldsymbol{M}^{\star} + \textsf{noise} - \boldsymbol{M}^{\textsf{ncvx}}\right)}_{\text{(heuristically) unbiased estimate of } \boldsymbol{M}^{\star}}$$

De-biasing nonconvex estimate

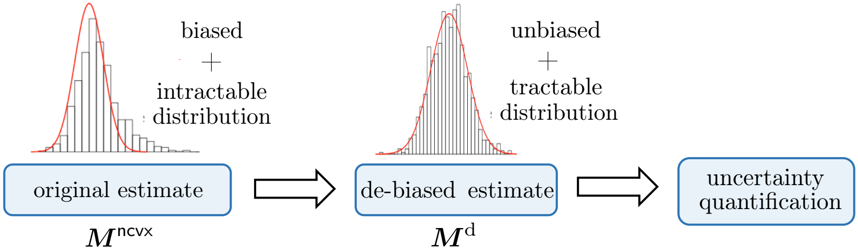
$$\boldsymbol{M}^{\mathsf{ncvx}} \overset{\text{de-biasing}}{\longrightarrow} \underbrace{\boldsymbol{M}^{\mathsf{ncvx}} + \underbrace{\tfrac{1}{p}\mathcal{P}_{\Omega}}_{\text{mean: } \mathcal{I}} \left(\boldsymbol{M}^{\star} + \text{noise} - \boldsymbol{M}^{\mathsf{ncvx}}\right)}_{\text{(heuristically) unbiased estimate of } \boldsymbol{M}^{\star}}$$

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

De-biasing nonconvex estimate

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

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


biased
+
intractable
distribution
unbiased
+
tractable
distribution
original estimate
$\boldsymbol{M}^{\mathsf{ncvx}}$
de-biased estimate
$\boldsymbol{M}^{\mathrm{d}}$
uncertainty
quantification

Distributional guarantees for low-rank factors

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

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

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

Distributional guarantees for low-rank factors

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

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

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

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

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{R} - \boldsymbol{X}^{\star} \;\approx\; \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \text{Cramer-Rao})$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{R} - \boldsymbol{Y}^{\star} \;\approx\; \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \text{Cramer-Rao})$$

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{R}-\boldsymbol{X}^{\star} \;\approx\; \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \text{Cramer-Rao})$$

$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{R}-\boldsymbol{Y}^{\star} \;\approx\; \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{0}, \text{Cramer-Rao})$$

- accurate uncertainty quantification for low-rank factors
  — *asymptotically optimal*

Implications

$$\boldsymbol{X}^{\mathrm{d}}\boldsymbol{R} - \boldsymbol{X}^{\star} \;\approx\; \boldsymbol{Z}^{X}, \qquad \boldsymbol{Z}^{X}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \text{Cramer-Rao})$$

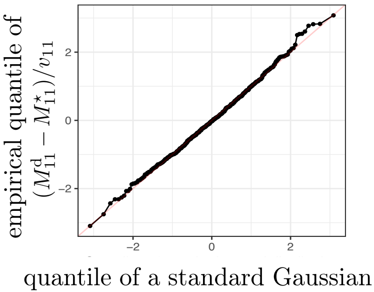
$$\boldsymbol{Y}^{\mathrm{d}}\boldsymbol{R} - \boldsymbol{Y}^{\star} \;\approx\; \boldsymbol{Z}^{Y}, \qquad \boldsymbol{Z}^{Y}_{i,\cdot} \overset{\text{ind.}}{\sim} \mathcal{N}(\mathbf{0}, \text{Cramer-Rao})$$

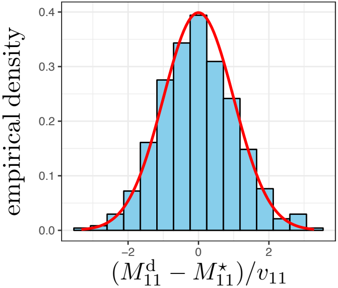
- accurate uncertainty quantification for matrix entries: if $\|\boldsymbol{X}^{\star}_{i,\cdot}\|_2 + \|\boldsymbol{Y}^{\star}_{j,\cdot}\|_2$ is not too small, then

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

— *asymptotically optimal*

Numerical experiments





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

Back to estimation: de-biased estimator is optimal

Distributional theory in turn allows us to track estimation accuracy

Back to estimation: de-biased estimator is optimal

Distributional theory in turn allows us to track estimation accuracy

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

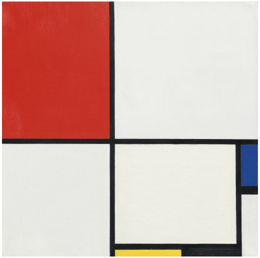
$$\|\boldsymbol{M}^{\mathsf{d}} - \boldsymbol{M}^{\star}\|_{\mathrm{F}}^2 = \underbrace{\frac{(2+o(1))nr\sigma^2}{p}}_{\textit{Cramer-Rao lower bound}} \qquad \textit{with high prob.}$$

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

Bridging convex and nonconvex optimization in noisy matrix completion

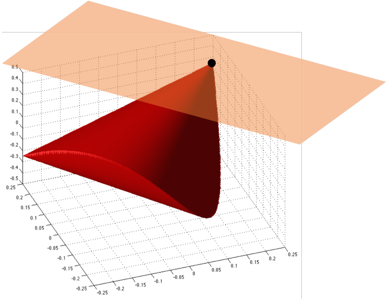
Convex relaxation for low-rank structure

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



low-rank matrix

figure credit: Piet Mondrian



semidefinite relaxation

Convex relaxation for low-rank structure

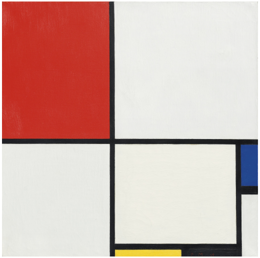
$$\begin{array}{ll} \underset{\boldsymbol{Z}}{\text{minimize}} & \|\boldsymbol{Z}\|_* := \sum_i \sigma_i(\boldsymbol{Z}) \\ \text{subj. to} & \text{noiseless data constraints} \end{array}$$

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

...

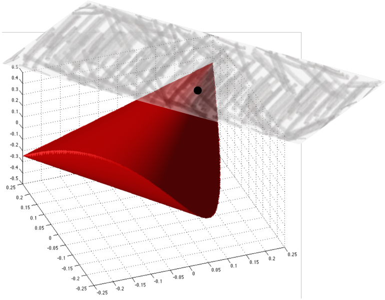
Stability of convex relaxation against noise

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



low-rank matrix

figure credit: Piet Mondrian



semidefinite relaxation

Stability of convex relaxation against noise

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

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

...

Noisy low-rank matrix completion

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

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

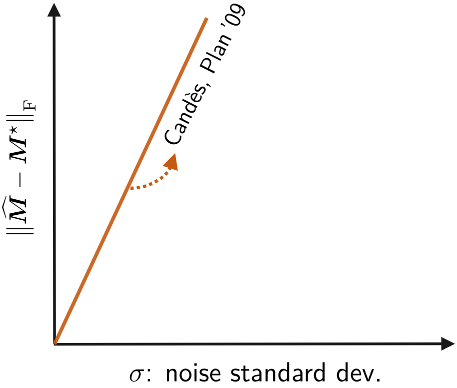
convex relaxation:

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

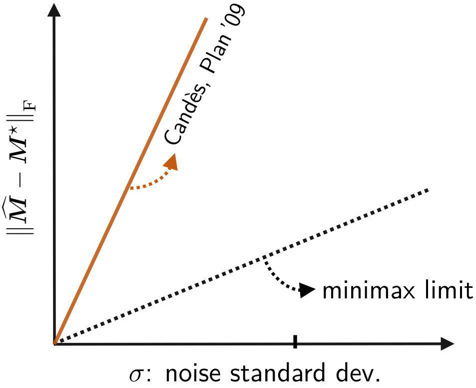
Prior statistical guarantees for convex relaxation

- **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, ...

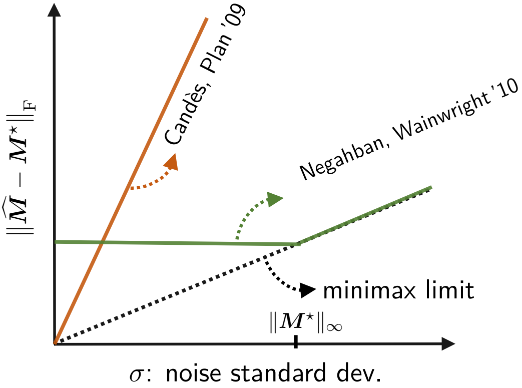
Candès, Plan '09	$\sigma n^{1.5}$



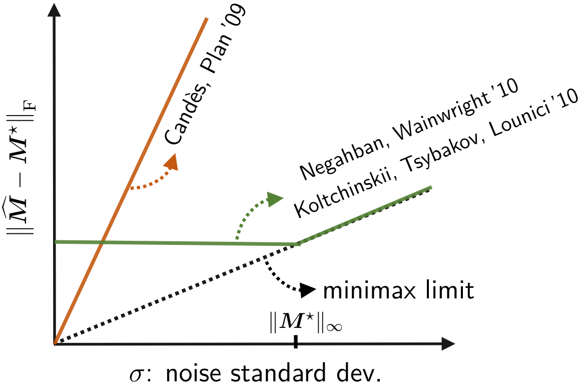
minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$



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

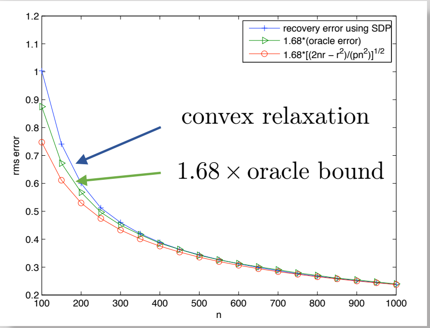


minimax limit	$\sigma\sqrt{n/p}$
Candès, Plan '09	$\sigma n^{1.5}$
Negahban, Wainwright '10	$\max\{\sigma, \|\boldsymbol{M}^{\star}\|_{\infty}\}\sqrt{n/p}$
Koltchinskii, Tsybakov, Lounici '10	$\max\{\sigma, \|\boldsymbol{M}^{\star}\|_{\infty}\}\sqrt{n/p}$



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

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 motivating experiment

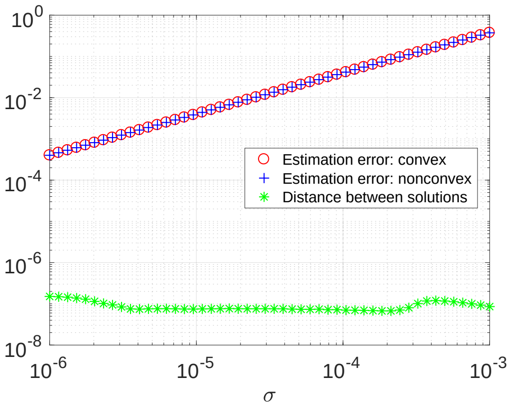
convex: $$\underset{\boldsymbol{Z}\in\mathbb{R}^{n\times n}}{\text{minimize}} \sum_{(i,j)\in\Omega} \left(Z_{i,j} - M_{i,j}\right)^2 + \lambda\|\boldsymbol{Z}\|_*$$

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

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

A motivating experiment

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



Convex and nonconvex solutions are exceedingly close!

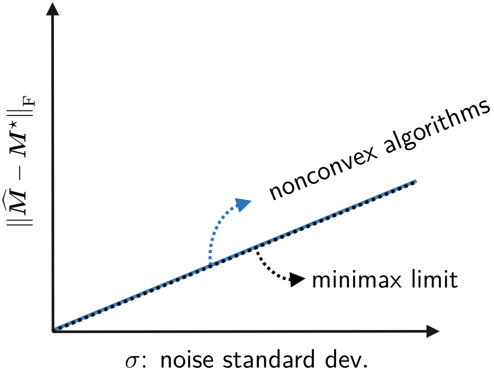
convex

nonconvex

$$\Downarrow$$

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

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



— *Ma, Wang, Chi, Chen '17*

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

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

Main results: $r = O(1)$

- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian noise with variance σ^2
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

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

Theorem 4 (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*

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

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

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

Theorem 4 (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. $$\big\|\widehat{\boldsymbol{M}}_{\mathsf{cvx}} - \boldsymbol{M}^\star\big\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}$$

Main results: $r = O(1)$

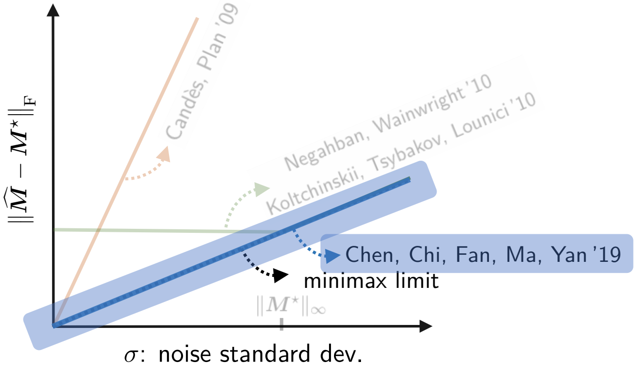
- **random sampling:** each $(i,j) \in \Omega$ with prob. $p \gtrsim \frac{\log^3 n}{n}$
- **random noise:** i.i.d. sub-Gaussian noise with variance σ^2
- true matrix $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$: $r = O(1)$, incoherent, well-conditioned

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

Theorem 4 (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. $$\|\widehat{\boldsymbol{M}}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{n}{p}}$$
$$\|\widehat{\boldsymbol{M}}_{\mathsf{cvx}} - \boldsymbol{M}^\star\|_{\infty} \lesssim \sigma\sqrt{\frac{n\log n}{p}} \cdot \frac{1}{n}$$



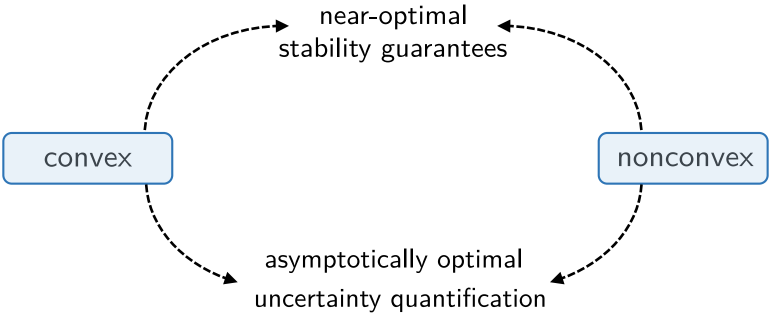
- minimax optimal when $r = O(1)$
- estimation errors are spread out across all entries

convex  nonconvex

$\Downarrow$

inference(convex) $\approx$ inference(nonconvex)

Same inference procedures work for both cvx & noncvx estimates!



1. "Inference and uncertainty quantification for noisy matrix completion", accepted to *Proceedings of the National Academy of Sciences (PNAS)*, Y. Chen, J. Fan, C. Ma, Y. Yan, 2019

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