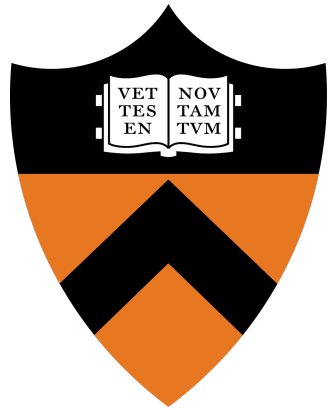
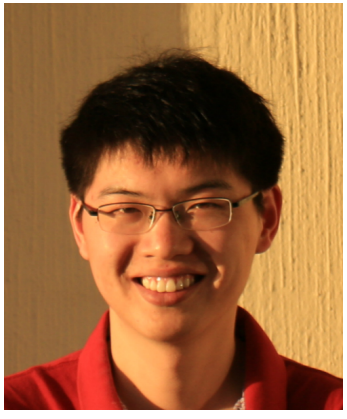
Nonconvex Optimization Meets Statistics:

from random initialization to uncertainty quantification

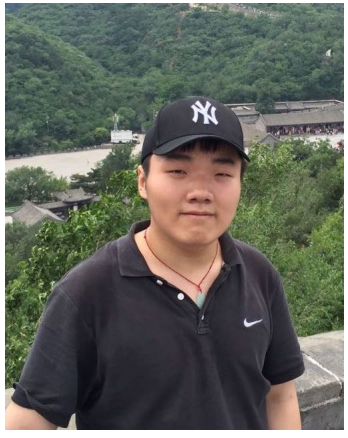


Yuxin Chen

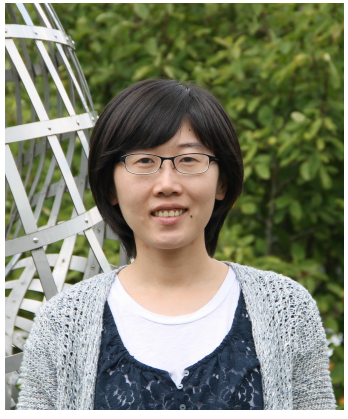
Electrical Engineering, Princeton University



Cong Ma
Princeton ORFE



Yuling Yan
Princeton ORFE



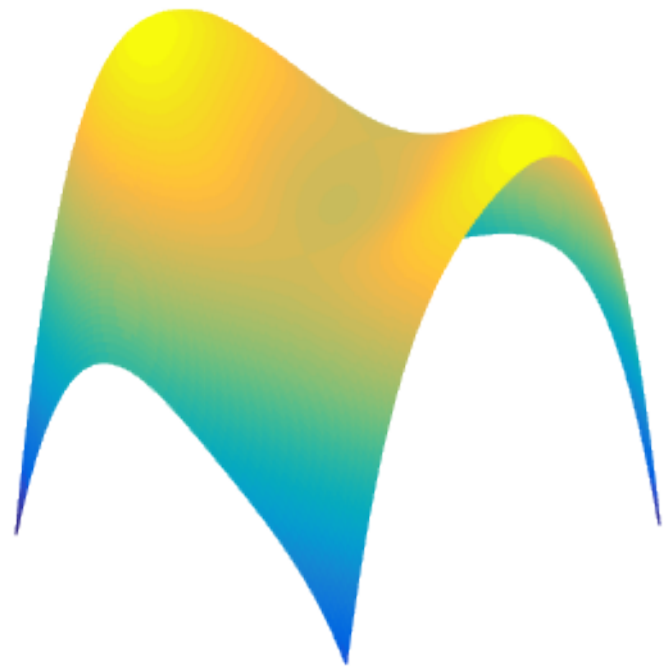
Yuejie Chi
CMU ECE

Jianqing Fan
Princeton ORFE

Nonconvex problems are everywhere

Empirical risk minimization is usually nonconvex

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

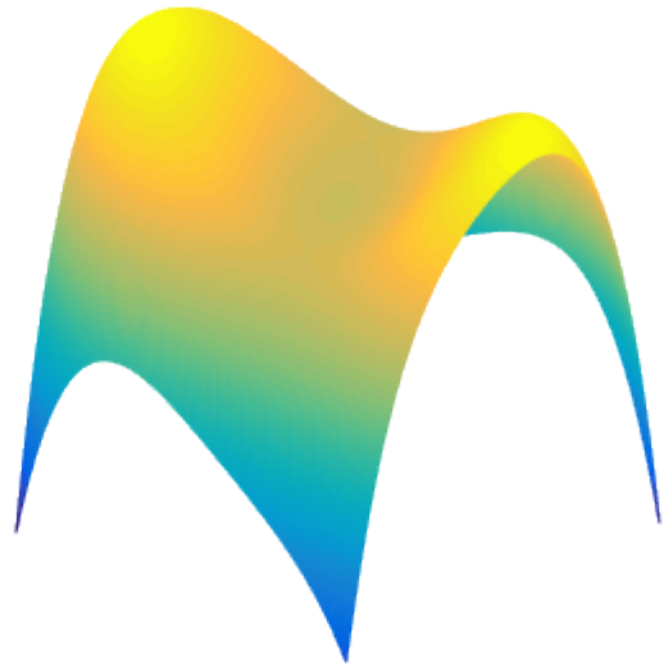


Nonconvex problems are everywhere

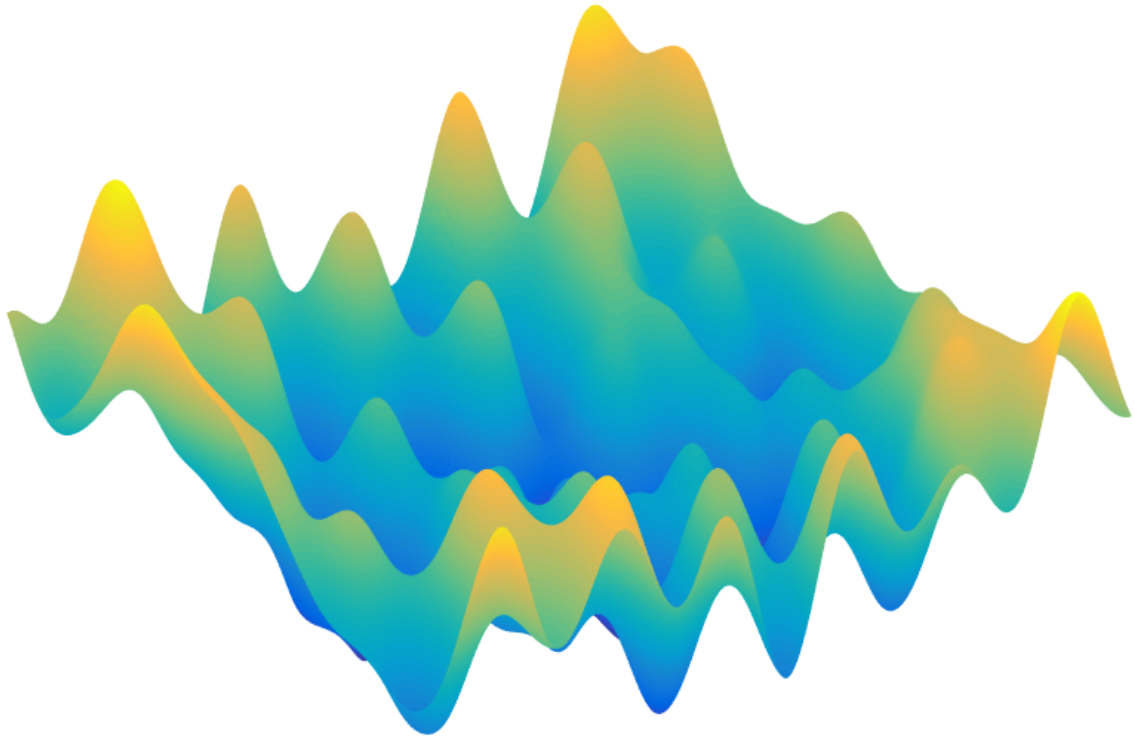
Empirical risk minimization is usually nonconvex

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

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



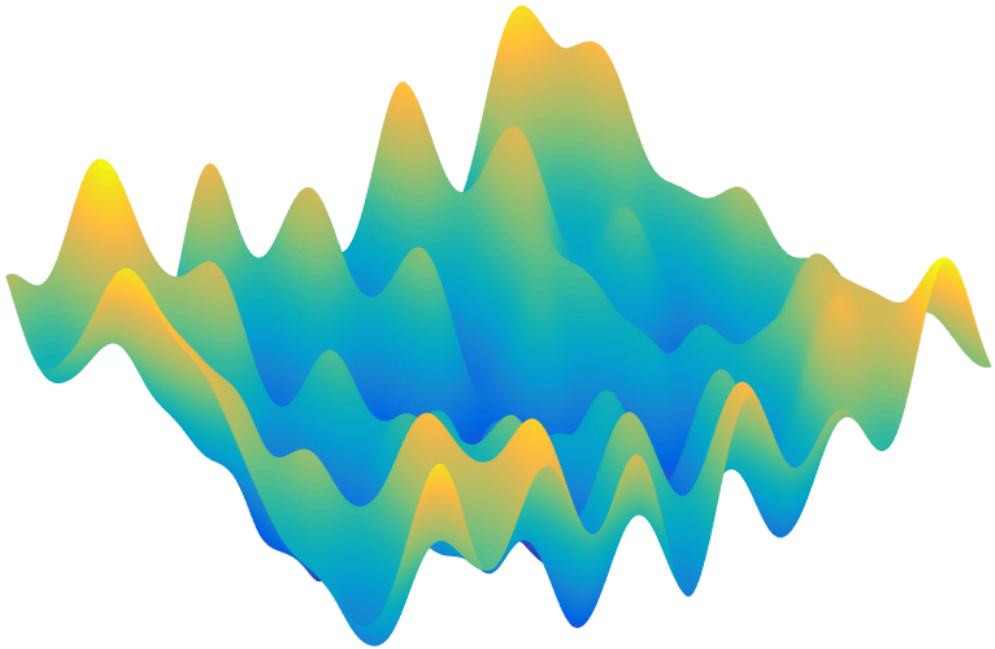
Nonconvex optimization may be super scary

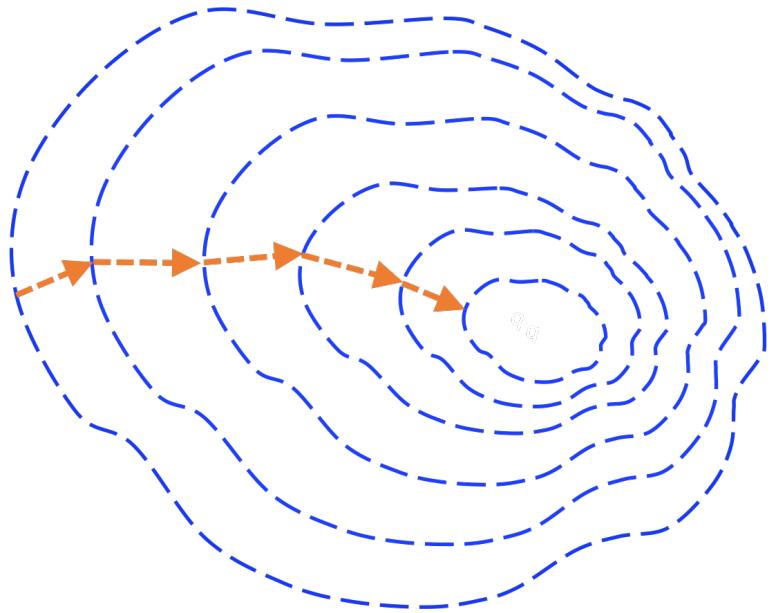


There may be bumps everywhere and exponentially many local optima

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

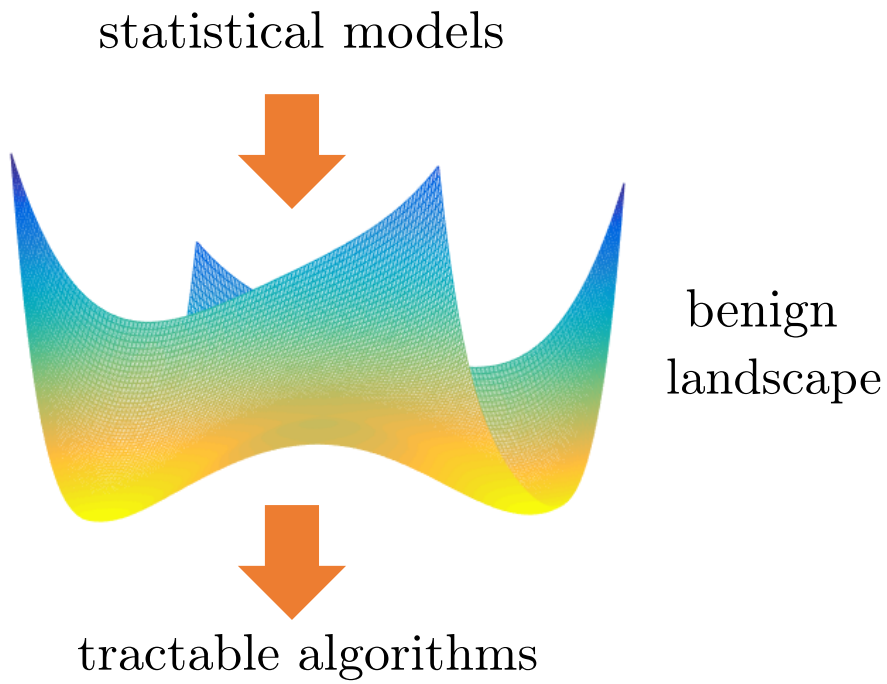
Nonconvex optimization may be super scary





But they are solved on a daily basis via simple algorithms like *(stochastic) gradient descent*

Statistical models come to rescue



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

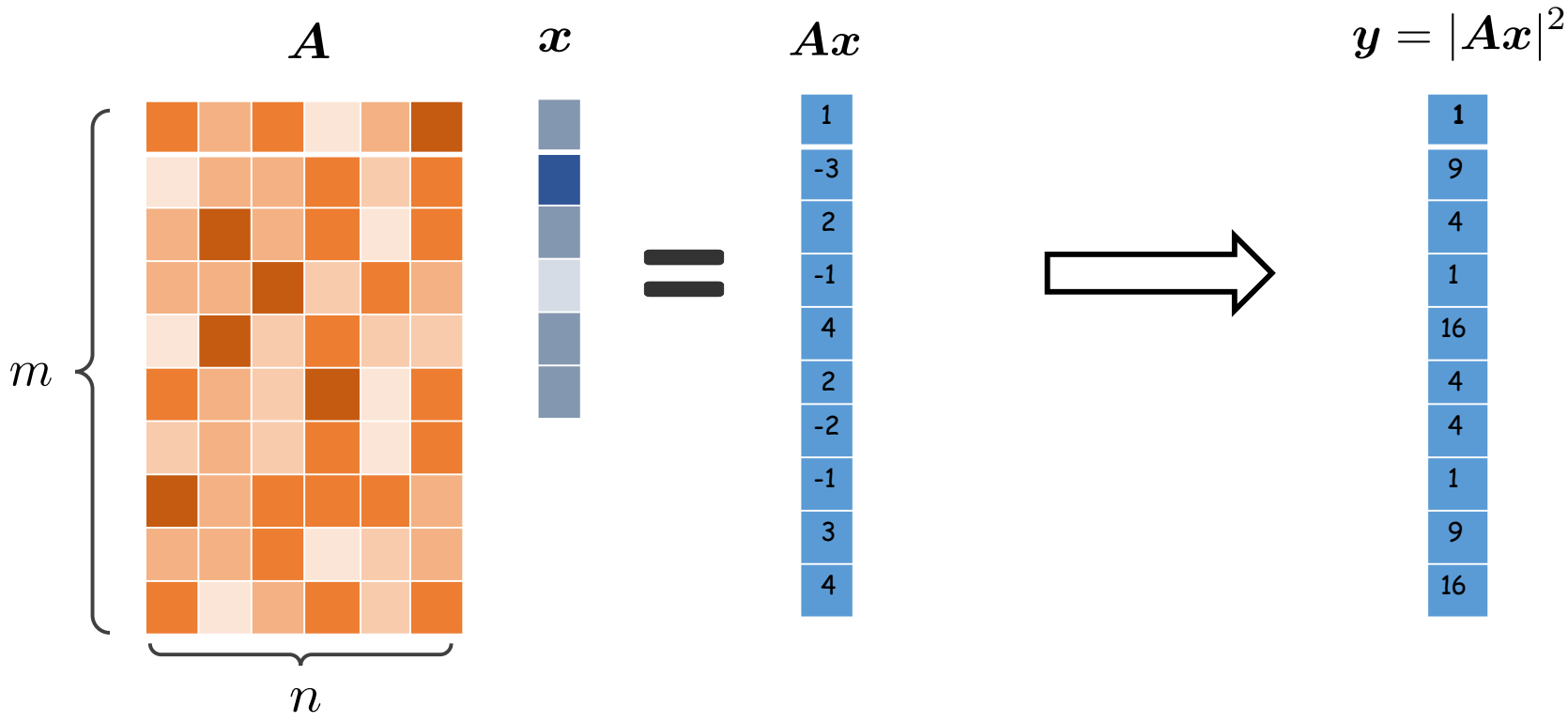
This talk: two recent stories

1. Random initialization when solving random quadratic systems of equations

2. Inference and uncertainty quantification for noisy matrix completion

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

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

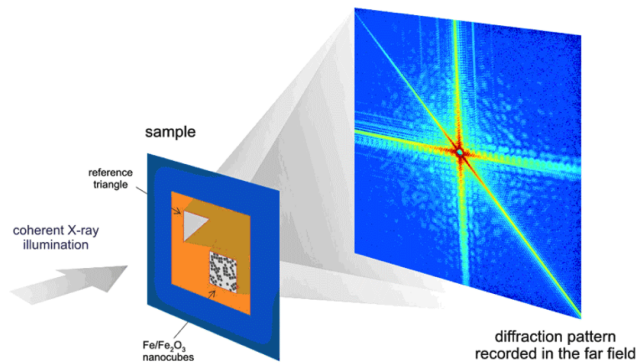
assume w.l.o.g. $\|\boldsymbol{x}^\star\|_2 = 1$

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



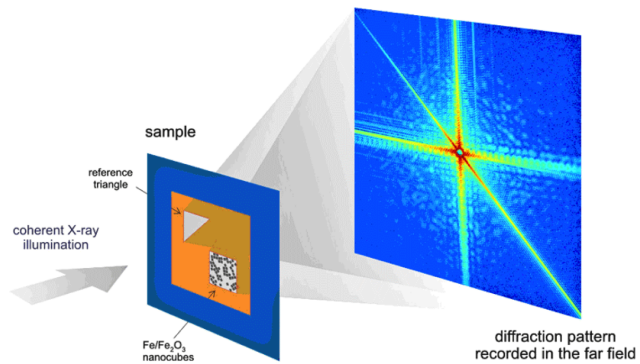
intensity of electrical field: $\left|\widehat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

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

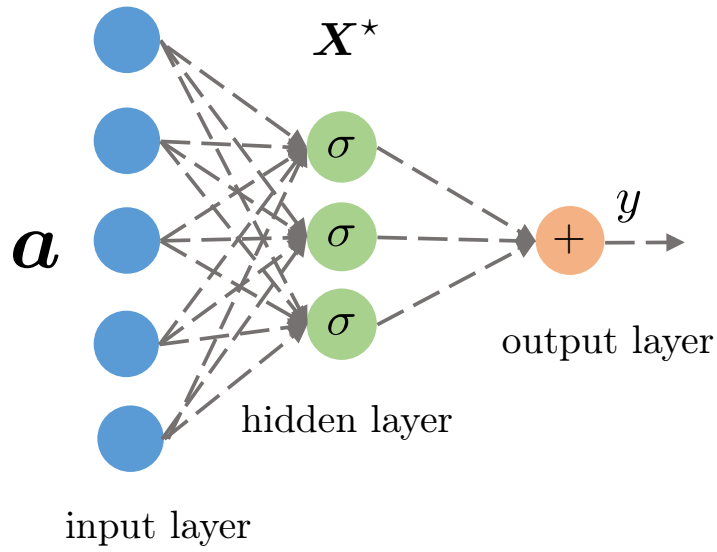


intensity of electrical field: $\left|\widehat{x}(f_1, f_2)\right|^2 = \left|\int x(t_1, t_2) e^{-i2\pi(f_1 t_1 + f_2 t_2)} \mathrm{d}t_1 \mathrm{d}t_2\right|^2$

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*

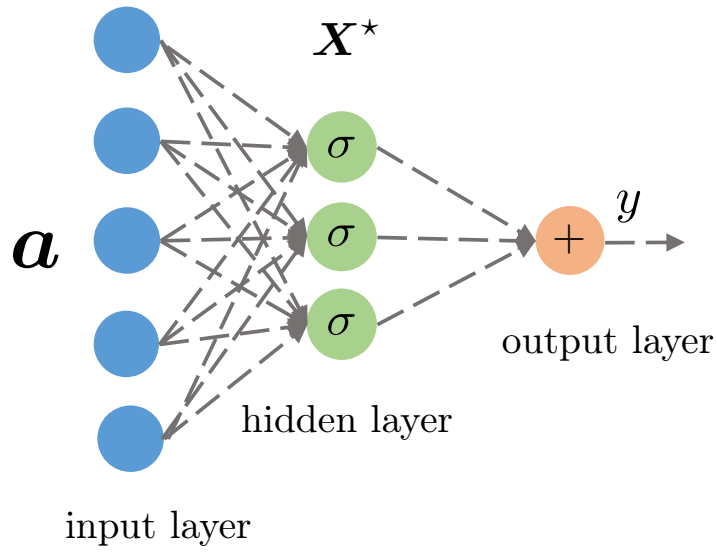


input features: $\boldsymbol{a}$; weights: $\boldsymbol{X}^\star = [\boldsymbol{x}_1^\star, \cdots, \boldsymbol{x}_r^\star]$

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

Motivation: learning neural nets with quadratic activation

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

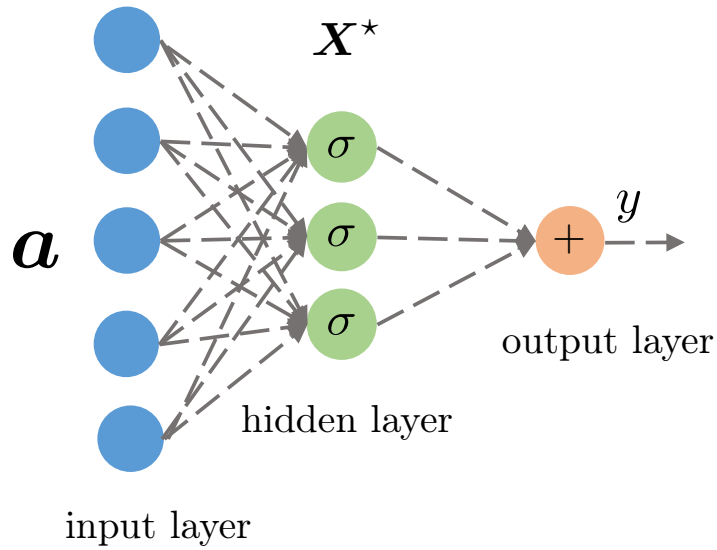


input features: $\boldsymbol{a}$; weights: $\boldsymbol{X}^\star = [\boldsymbol{x}_1^\star, \cdots, \boldsymbol{x}_r^\star]$

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

Motivation: learning neural nets with quadratic activation

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



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

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

We consider simplest model when $r = 1$

A natural least squares formulation

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

A natural least squares formulation

$$\text{minimize}_{\boldsymbol{x}\in\mathbb{R}^n} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\Big[\big(\boldsymbol{a}_k^\top \boldsymbol{x}\big)^2 - y_k\Big]^2$$

- **issue:** $f(\cdot)$ is highly nonconvex

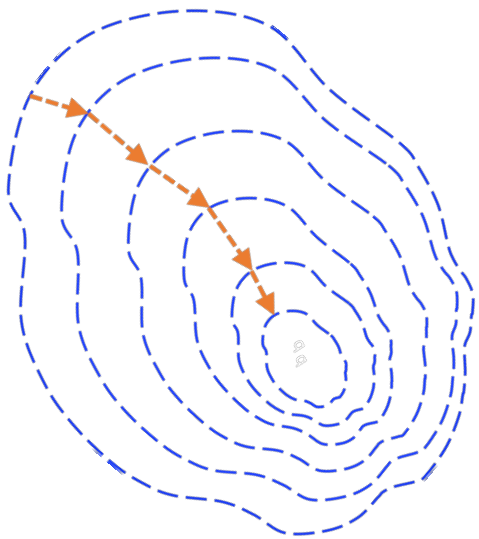
  $\longrightarrow$ *computationally challenging!*

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)

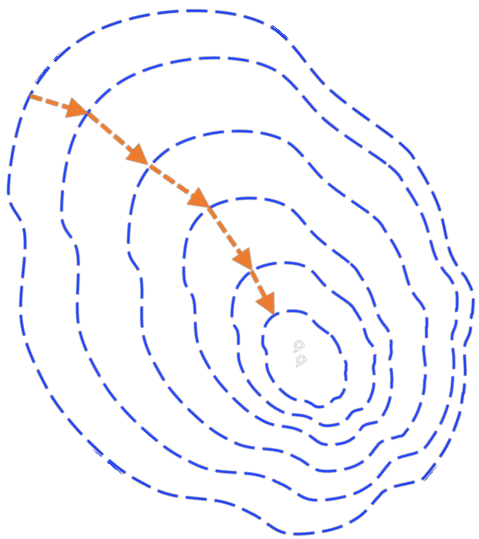
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m}\sum_{k=1}^{m}\left[\left(\boldsymbol{a}_k^\top \boldsymbol{x}\right)^2 - y_k\right]^2$$



- **spectral initialization:** $\boldsymbol{x}^0 \leftarrow$ leading eigenvector of certain data matrix

Wirtinger flow (Candès, Li, Soltanolkotabi '14)

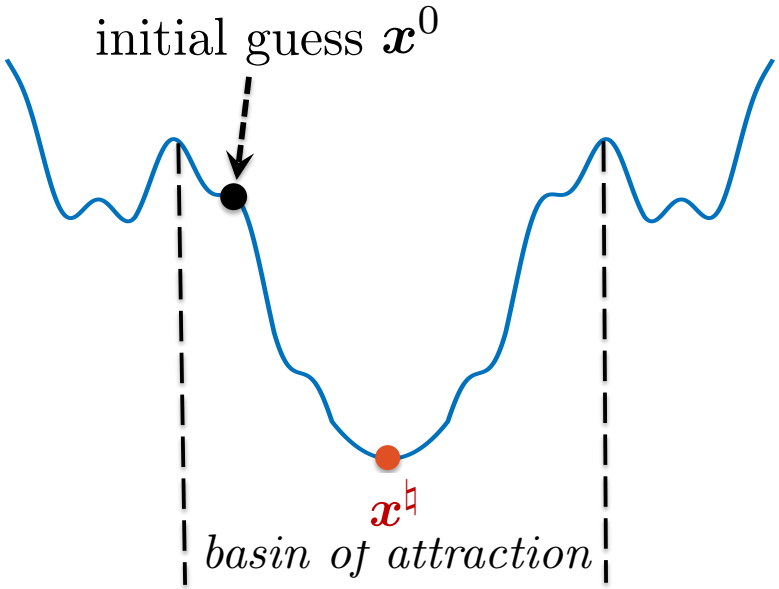
$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) = \frac{1}{4m} \sum_{k=1}^{m} \left[\left(\boldsymbol{a}_k^\top \boldsymbol{x} \right)^2 - y_k \right]^2$$



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

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

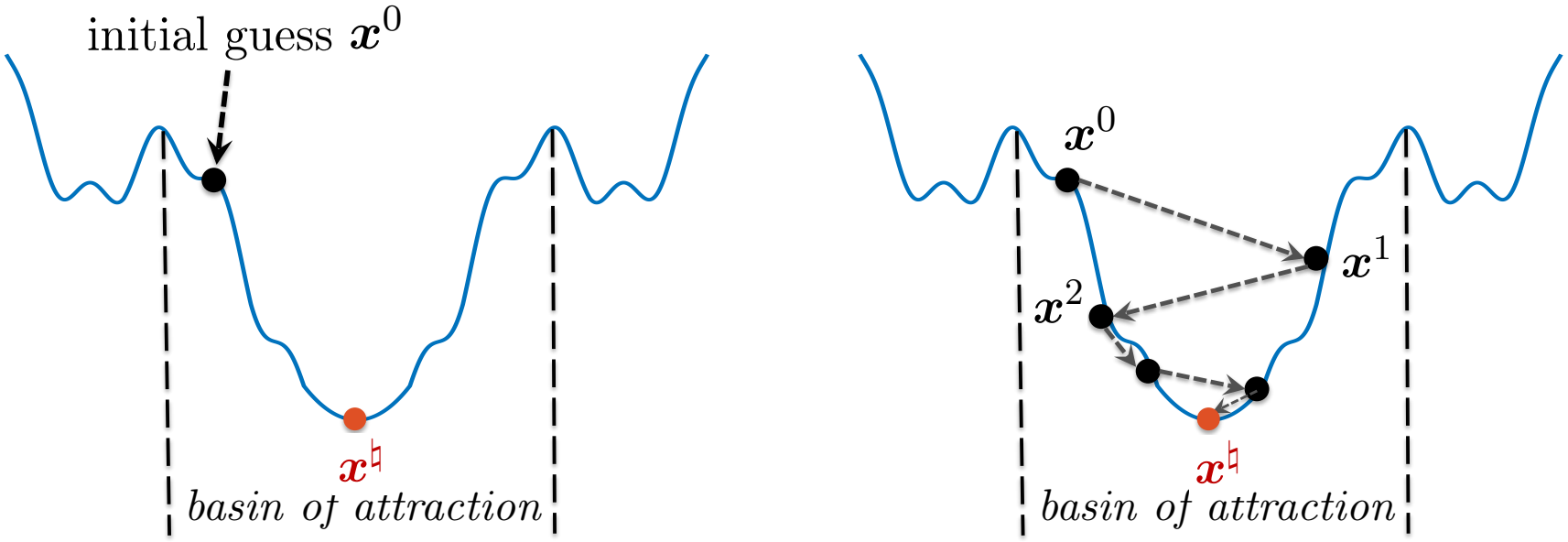
Rationale of two-stage approach



1. initialize within $\underbrace{\text{local basin sufficiently close to } x^\star}$

(restricted) strongly convex; no saddles / spurious local mins

Rationale of two-stage approach



1. initialize within $\underbrace{\text{local basin sufficiently close to } x^\star}$

(restricted) strongly convex; no saddles / spurious local mins

2. iterative refinement

A highly incomplete list of two-stage methods

phase retrieval:

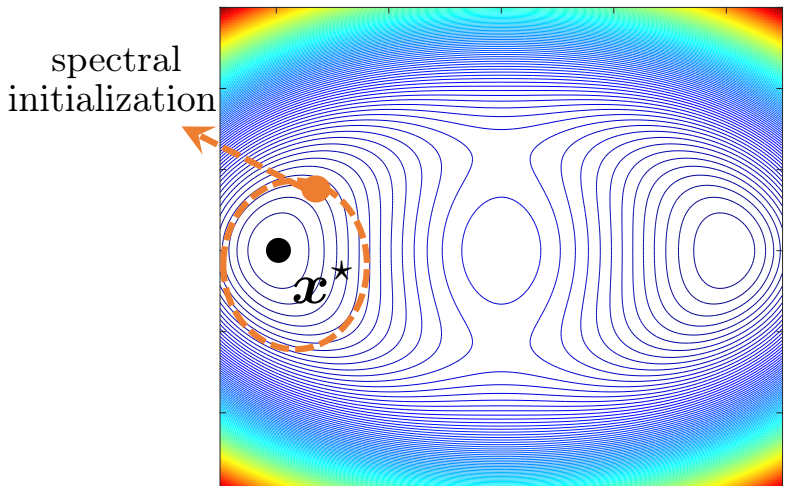
- Netrapalli, Jain, Sanghavi '13
- Candès, Li, Soltanolkotabi '14
- Chen, Candès '15
- Cai, Li, Ma '15
- Wang, Giannakis, Eldar '16
- Zhang, Zhou, Liang, Chi '16
- Kolte, Ozgur '16
- Zhang, Chi, Liang '16
- Soltanolkotabi '17
- Vaswani, Nayer, Eldar '16
- Chi, Lu '16
- Wang, Zhang, Giannakis, Akcakaya, 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
- ...

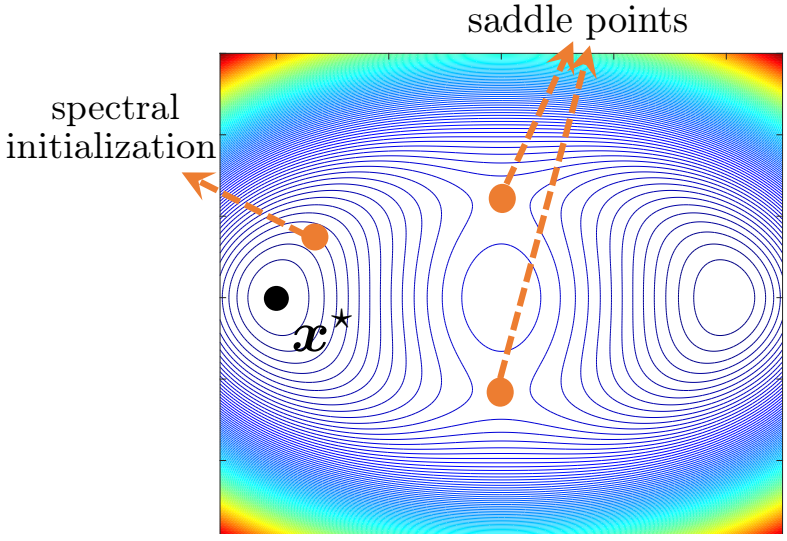
Is carefully-designed initialization necessary for fast convergence?

Initialization



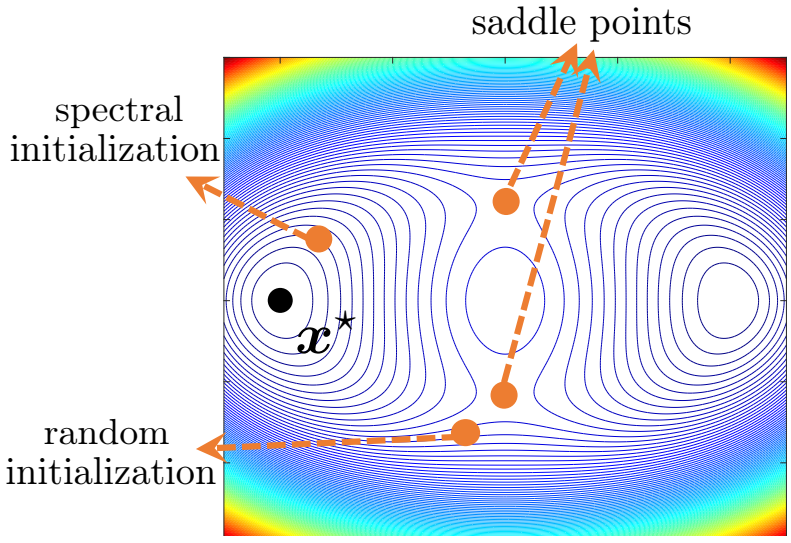
- spectral initialization gets us to (restricted) strongly cvx region

Initialization



- spectral initialization gets us to (restricted) strongly cvx region
- cannot initialize GD anywhere, e.g. might get stuck at saddles

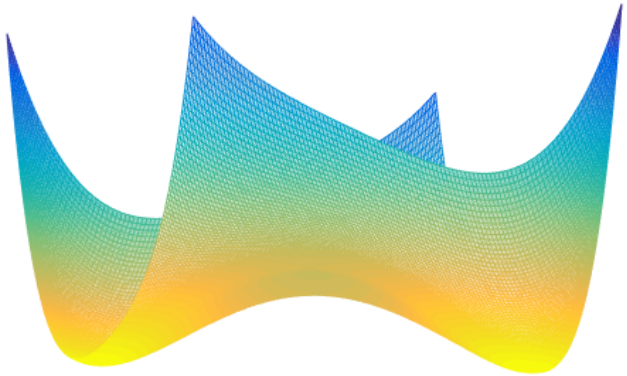
Initialization



- spectral initialization gets us to (restricted) strongly cvx region
- cannot initialize GD anywhere, e.g. might get stuck at saddles

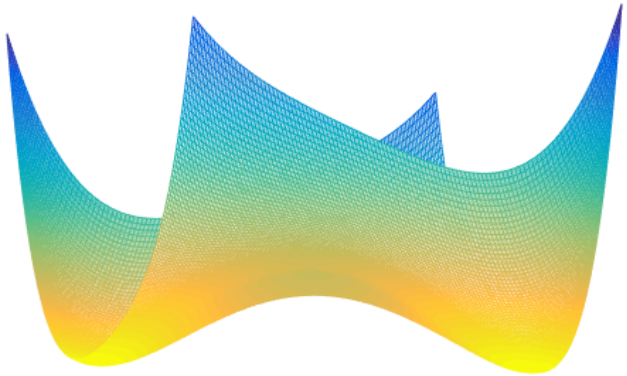
Can we initialize GD randomly, which is simpler and model-agnostic?

What does prior theory say?



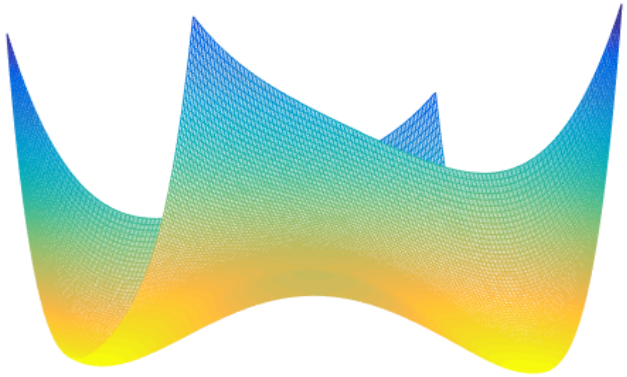
- **landscape:** no spurious local mins (Sun, Qu, Wright '16)

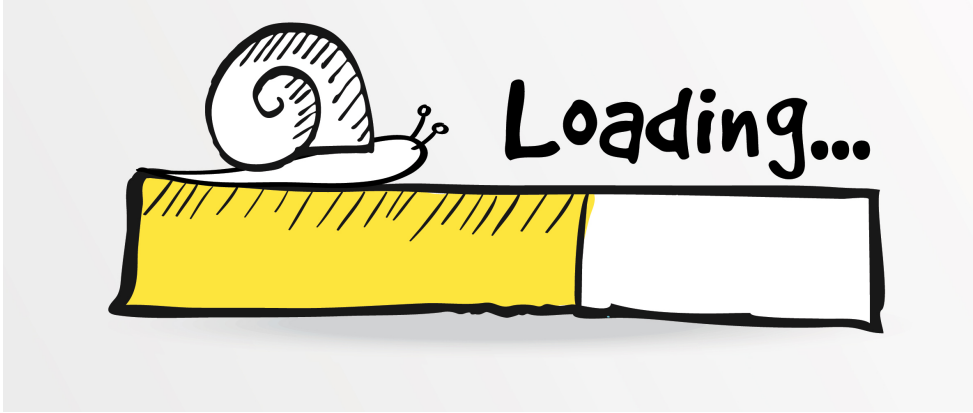
What does prior theory say?



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

What does prior theory say?



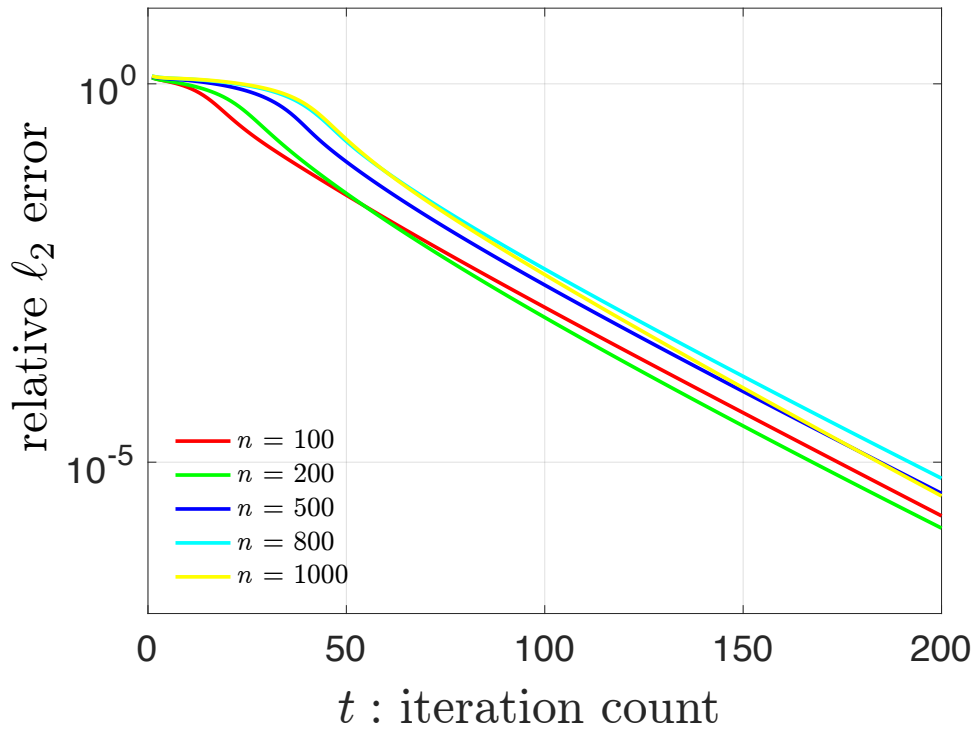


- **landscape:** no spurious local mins (Sun, Qu, Wright '16)
- randomly initialized GD converges almost surely (Lee et al. '16)

"almost surely" might mean "take forever"

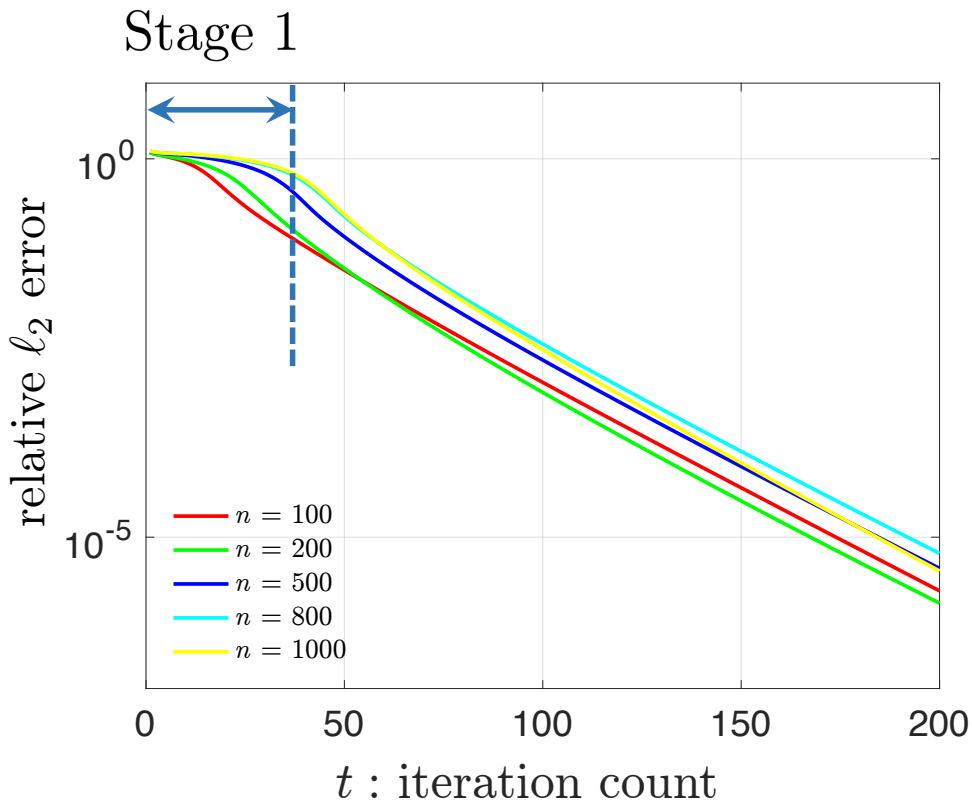
Numerical efficiency of randomly initialized GD

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



Numerical efficiency of randomly initialized GD

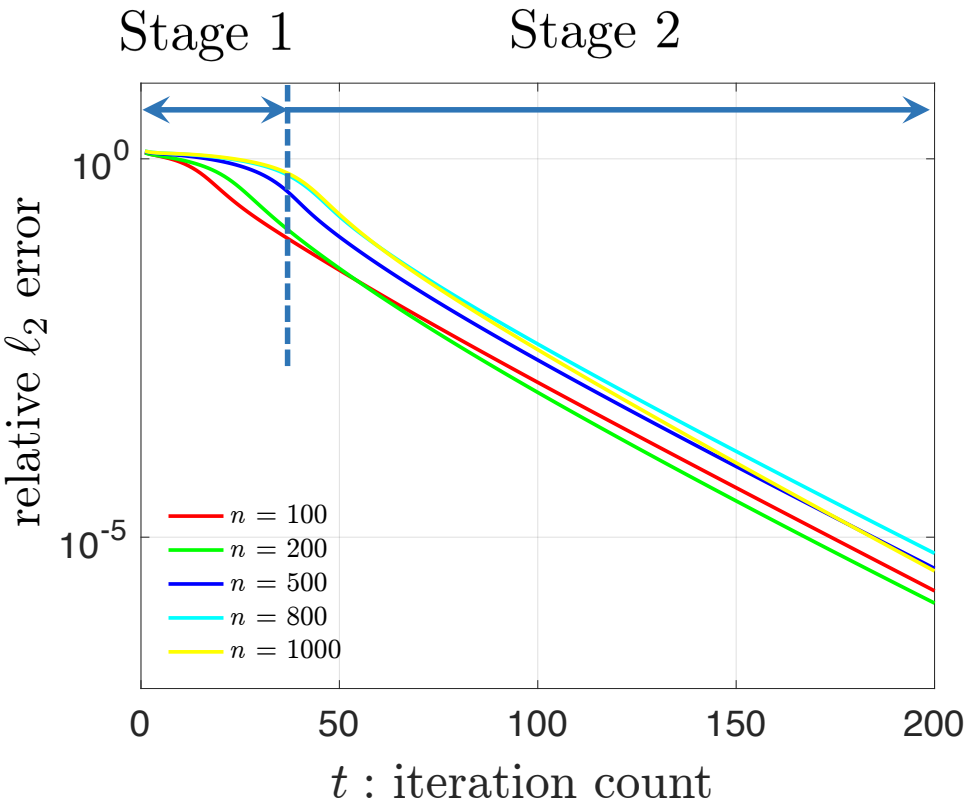
$$\eta = 0.1,\ \boldsymbol{a}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n),\ m = 10n,\ \boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$$



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

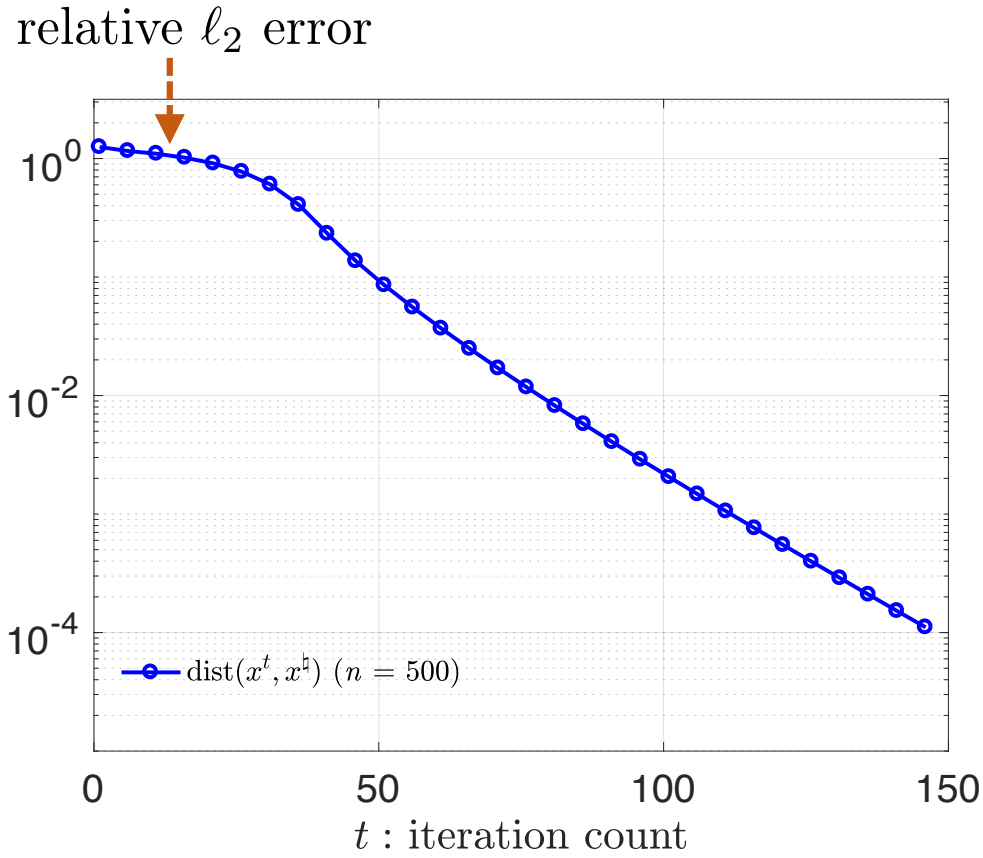
Numerical efficiency of randomly initialized GD

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

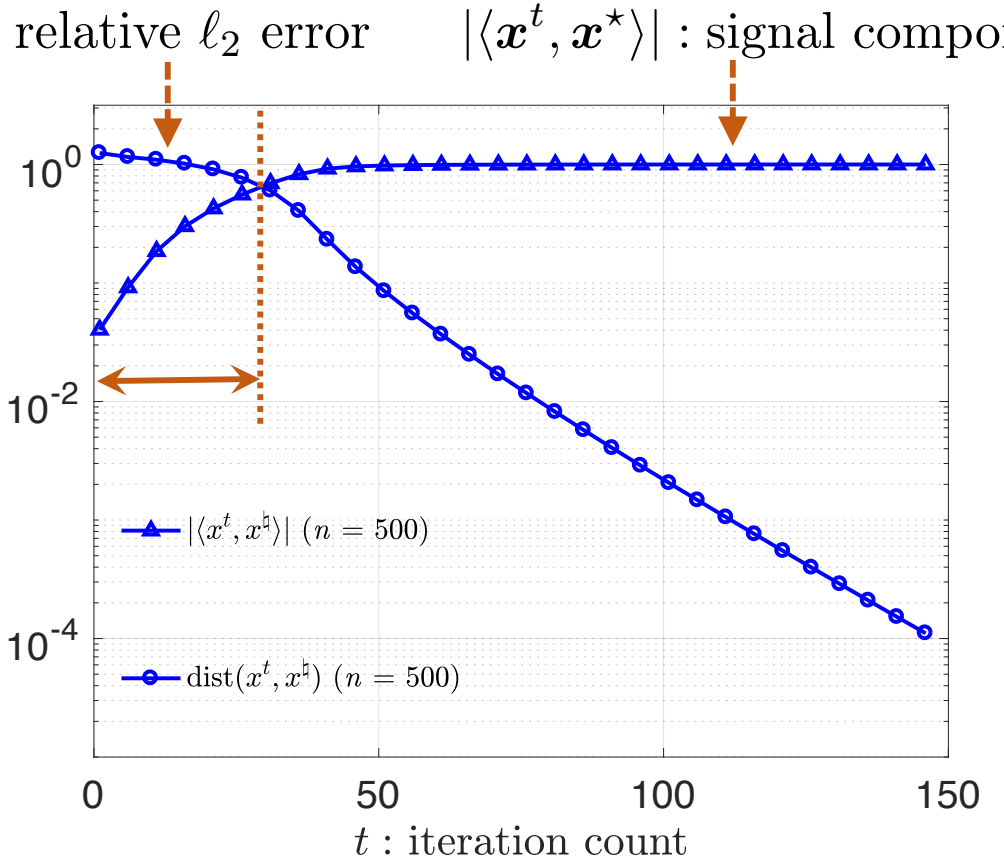


Randomly initialized GD enters local basin within **tens of iterations**

Exponential growth of signal strength in Stage 1



Exponential growth of signal strength in Stage 1



Numerically, a few iterations suffice for entering local region

Our theory: noiseless case

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: noiseless case

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*

Our theory: noiseless case

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)

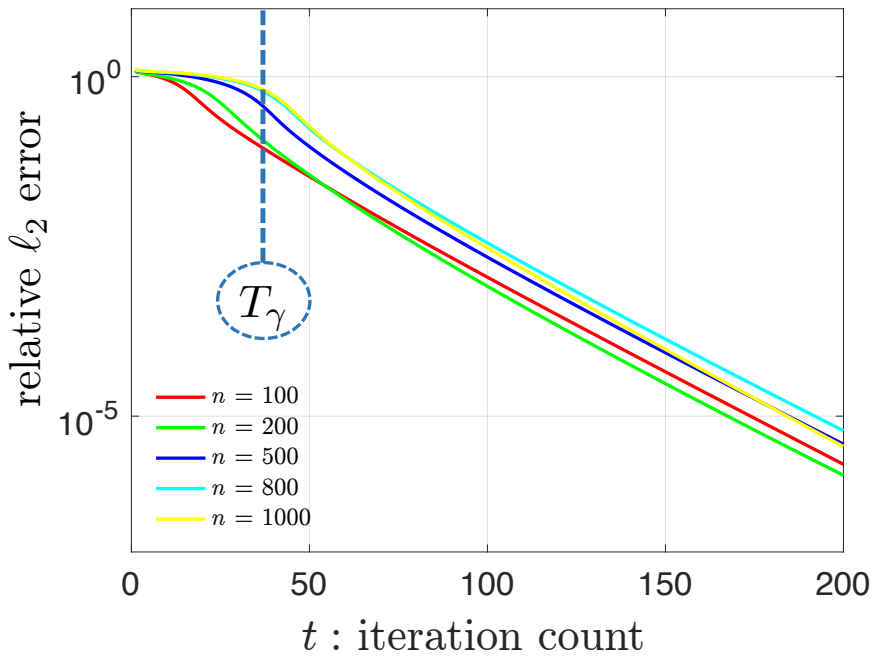
Under i.i.d. Gaussian design, GD with $\boldsymbol{x}^0 \sim \mathcal{N}(\boldsymbol{0}, n^{-1}\boldsymbol{I}_n)$ *achieves*

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

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* $m \gtrsim n\,\text{poly}\log m$

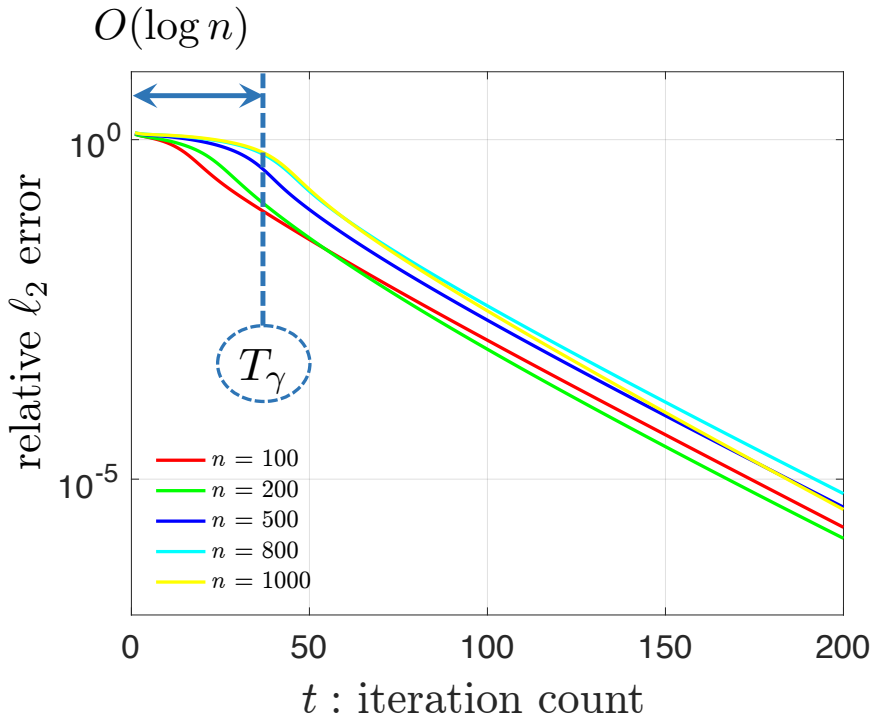
Our theory: noiseless case

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



Our theory: noiseless case

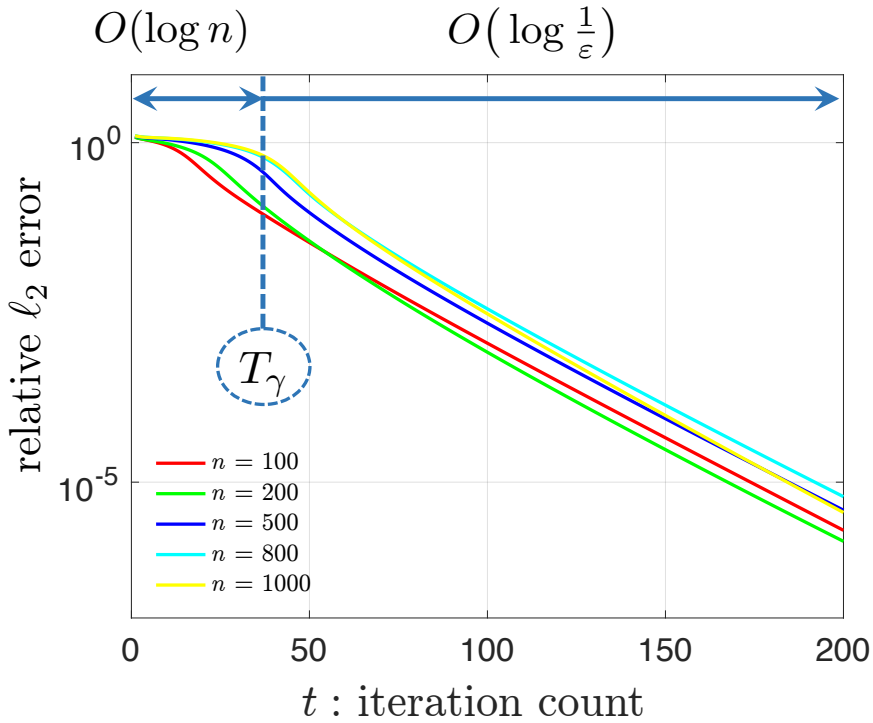
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \leq \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \geq T_\gamma \asymp \log n$$



- *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: noiseless case

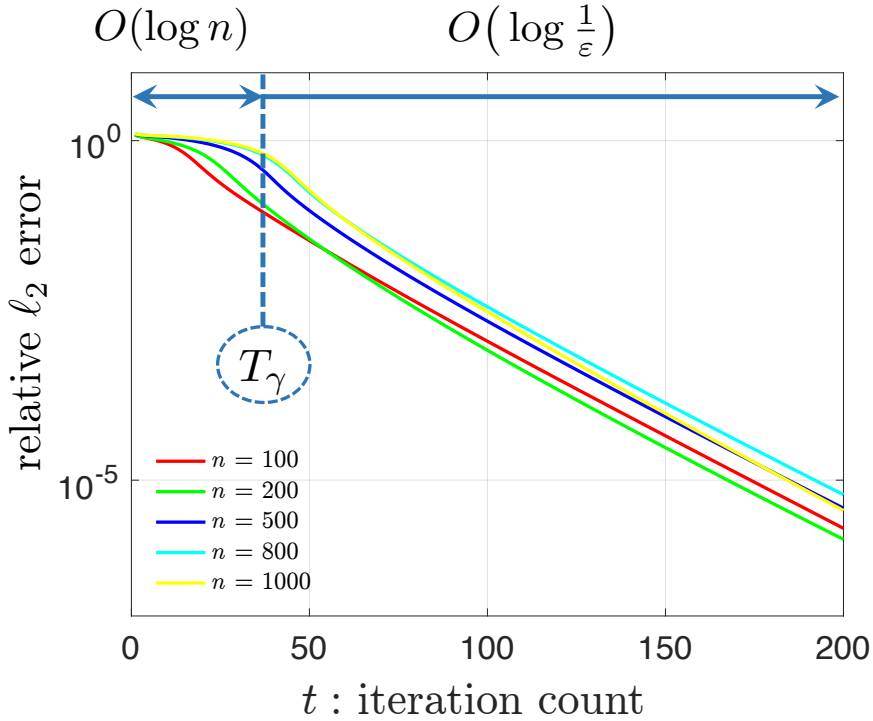
$$\mathrm{dist}(\boldsymbol{x}^t, \boldsymbol{x}^\star) \leq \gamma(1-\rho)^{t-T_\gamma}\|\boldsymbol{x}^\star\|_2, \quad t \geq T_\gamma \asymp \log n$$



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

Our theory: noiseless case

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

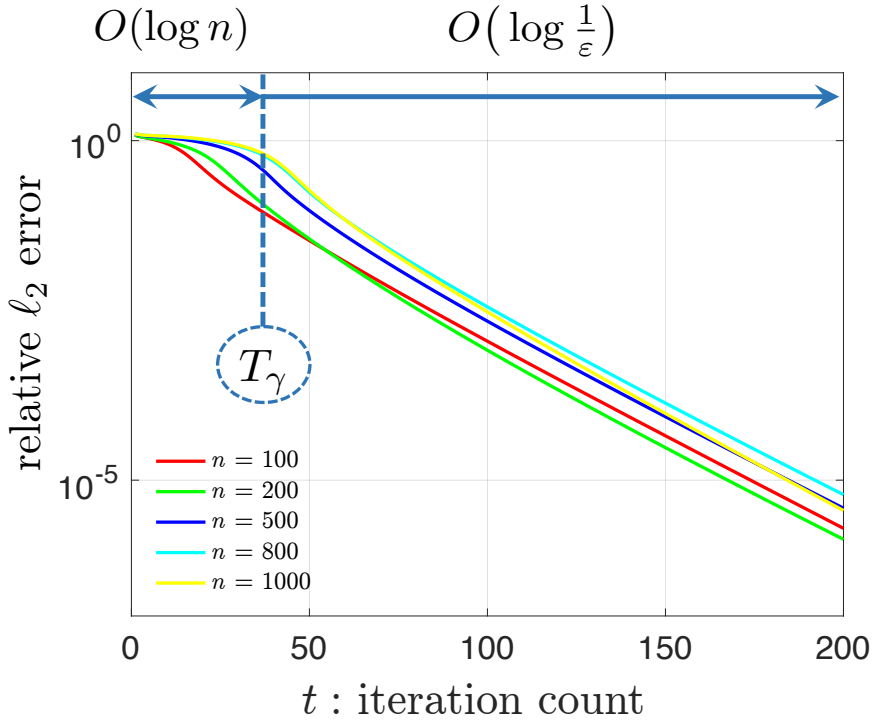


- *near-optimal computational cost*:

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

Our theory: noiseless case

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



- *near-optimal computational cost*:
  — $O\big(\log n + \log \frac{1}{\varepsilon}\big)$ iterations to yield ε accuracy
- *near-optimal sample size*: $m \gtrsim n\mathrm{poly}\log m$

A little analysis

What if we have infinite samples?

Gaussian designs: $\boldsymbol{a}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \quad 1 \le k \le m$

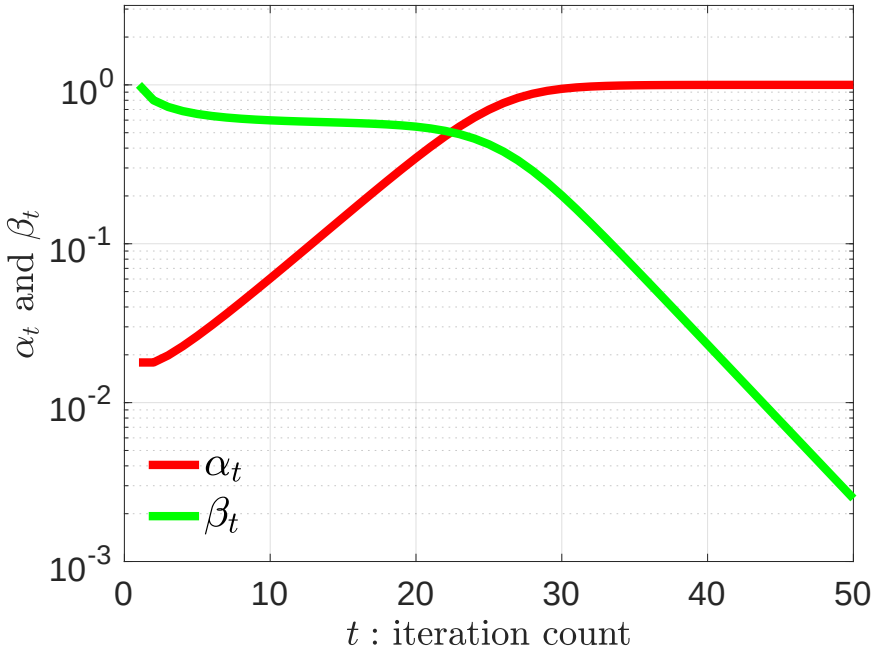
Population level (infinite samples)

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t),$$

where

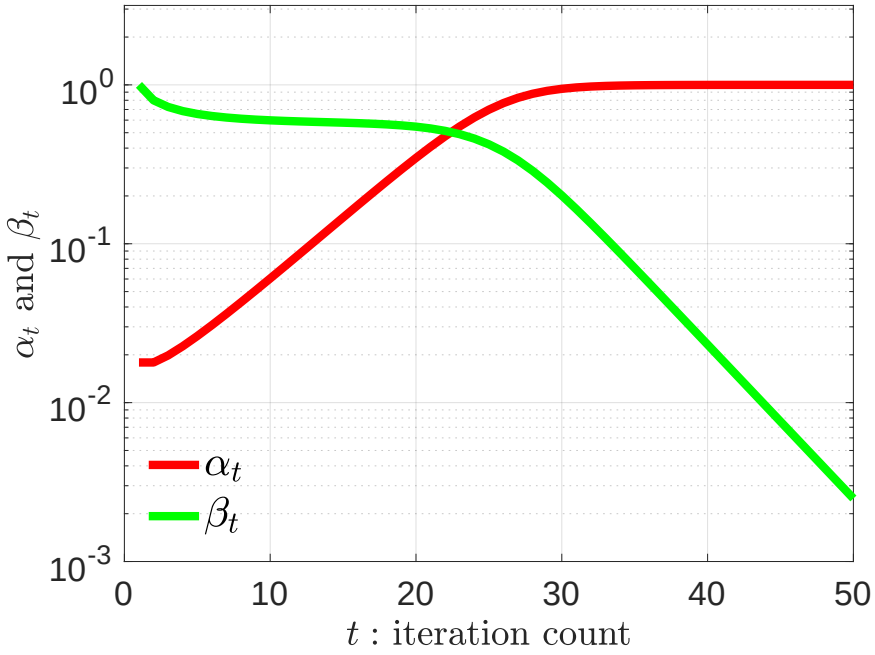
$$\nabla F(\boldsymbol{x}) := \mathbb{E}[\nabla f(\boldsymbol{x})] = (3\|\boldsymbol{x}\|_2^2 - 1)\boldsymbol{x} - 2(\boldsymbol{x}^{\star\top}\boldsymbol{x})\boldsymbol{x}^\star$$

Population-level state evolution



Let $\alpha_t := \underbrace{\left|\langle \boldsymbol{x}^t, \boldsymbol{x}^\star \rangle\right|}_{\text{signal strength}}$ and $\beta_t = \underbrace{\|\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\star \rangle \boldsymbol{x}^\star\|_2}_{\text{size of residual component}}$, then

Population-level state evolution



Let $\alpha_t := \underbrace{|\langle \boldsymbol{x}^t, \boldsymbol{x}^\star \rangle|}_{\text{signal strength}}$ and $\beta_t = \underbrace{\|\boldsymbol{x}^t - \langle \boldsymbol{x}^t, \boldsymbol{x}^\star \rangle \boldsymbol{x}^\star\|_2}_{\text{size of residual component}}$, then

$$\alpha_{t+1} = \{1 + 3\eta[1 - (\alpha_t^2 + \beta_t^2)]\}\alpha_t$$
$$\beta_{t+1} = \{1 + \eta[1 - 3(\alpha_t^2 + \beta_t^2)]\}\beta_t$$

2-parameter dynamics

Back to finite-sample analysis

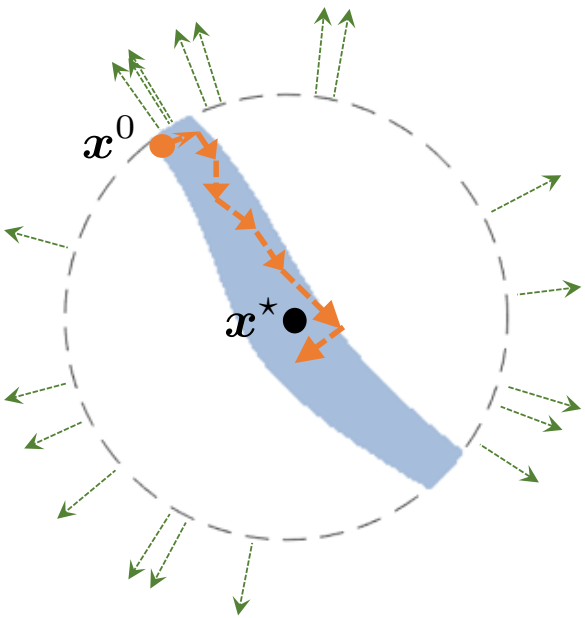
$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$$

Back to finite-sample analysis

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t) = \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t) - \underbrace{\eta\big(\nabla f(\boldsymbol{x}^t) - \nabla F(\boldsymbol{x}^t)\big)}_{\text{residual}}$$

Back to finite-sample analysis

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t) = \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t) - \underbrace{\eta\big(\nabla f(\boldsymbol{x}^t) - \nabla F(\boldsymbol{x}^t)\big)}_{\text{residual}}$$

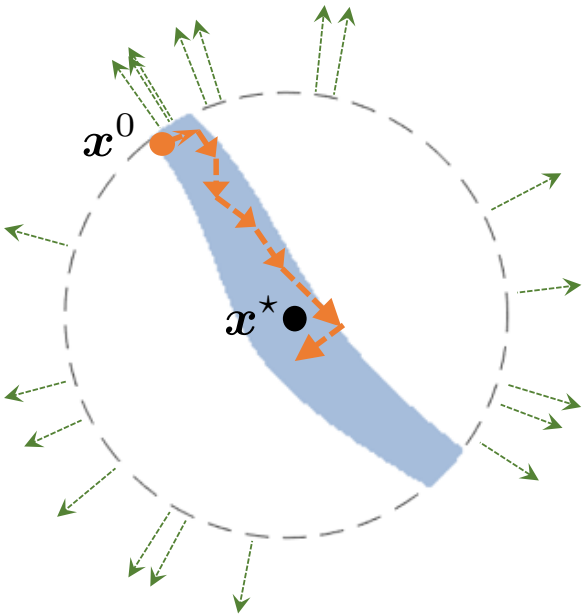


a region with
well-controlled residual

- population-level analysis holds *approximately* if $\boldsymbol{x}^t$ is independent of $\{\boldsymbol{a}_l\}$

Back to finite-sample analysis

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t) = \boldsymbol{x}^t - \eta \nabla F(\boldsymbol{x}^t) - \underbrace{\eta\big(\nabla f(\boldsymbol{x}^t) - \nabla F(\boldsymbol{x}^t)\big)}_{\text{residual}}$$



a region with
well-controlled residual

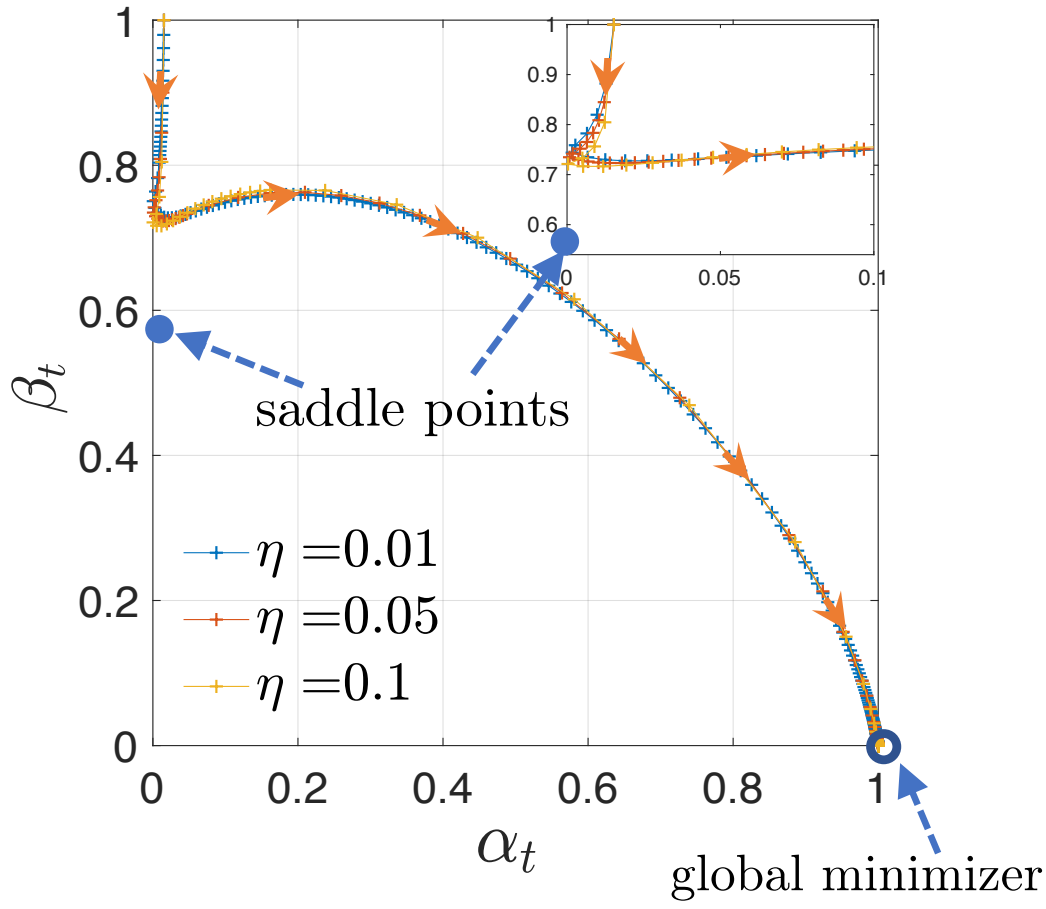
- population-level analysis holds *approximately* if $\boldsymbol{x}^t$ is independent of $\{\boldsymbol{a}_l\}$
- **key analysis ingredient:** show $\boldsymbol{x}^t$ is "nearly-independent" of each $\boldsymbol{a}_l$

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
- Chen, Liu, Li '19
- Chen, Chi, Fan, Ma, Yan '19
- Chen, Chi, Fan, Yan '19
- Cai, Li, Chi, Poor, Chen '19
- Lei '19

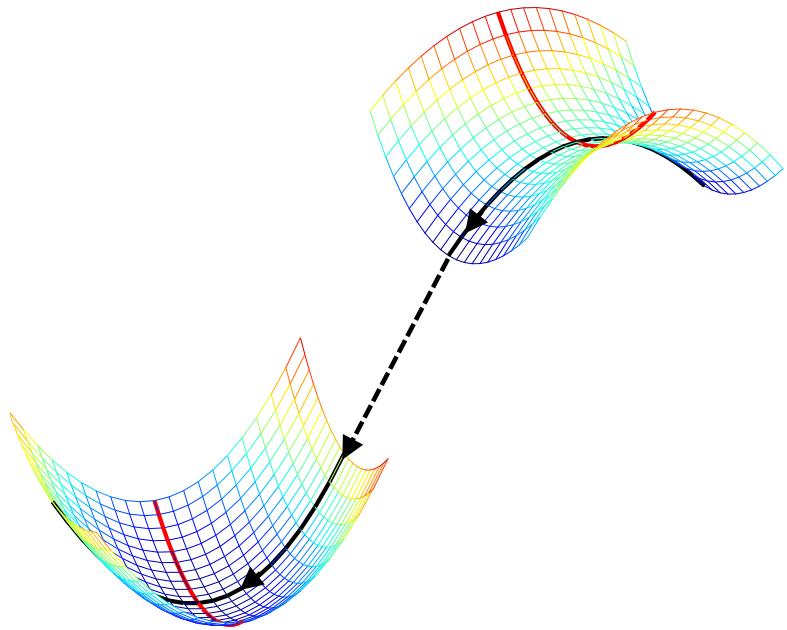
Automatic saddle avoidance



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

Summary for Part 1

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

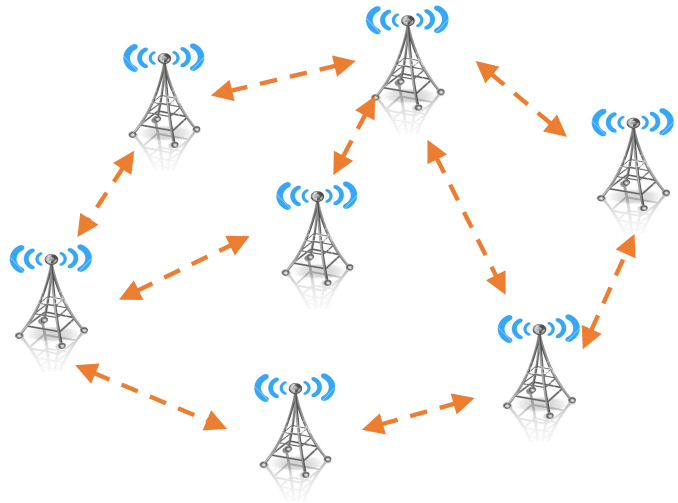
Inference and uncertainty quantification for noisy matrix completion

Low-rank matrix completion

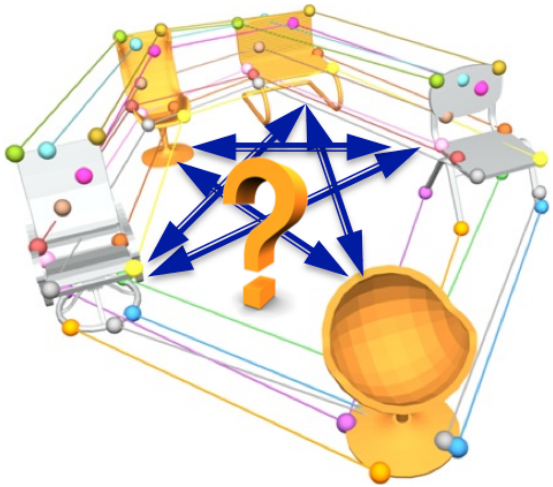
figure credit: E. J. Candès

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

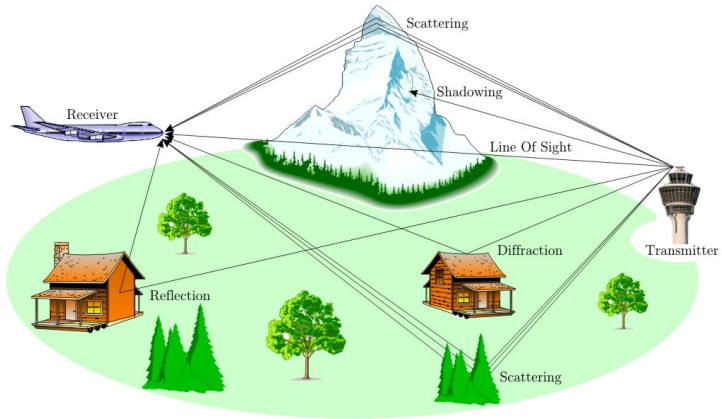
recommendation systems



localization



shape matching

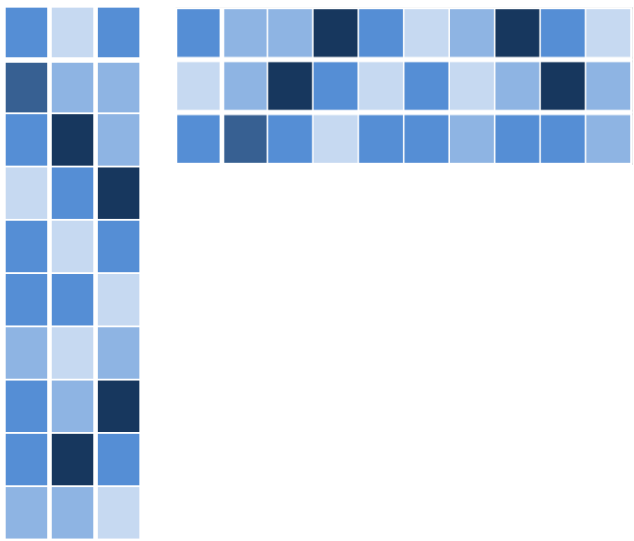


channel estimation

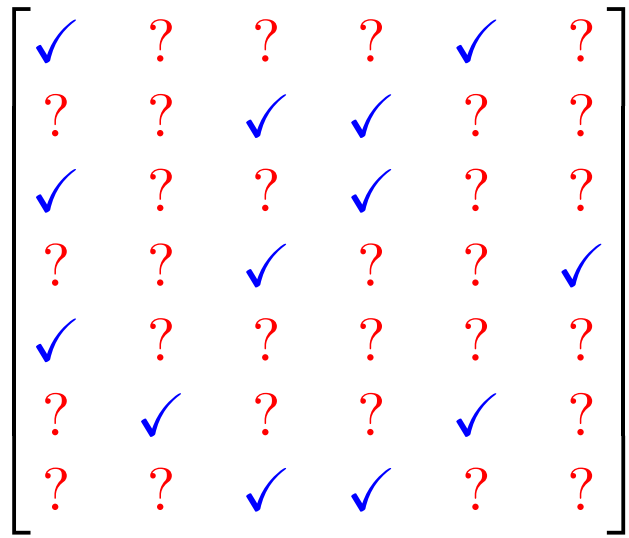
Noisy low-rank matrix completion

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

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



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



sampling set Ω

Noisy low-rank matrix completion

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

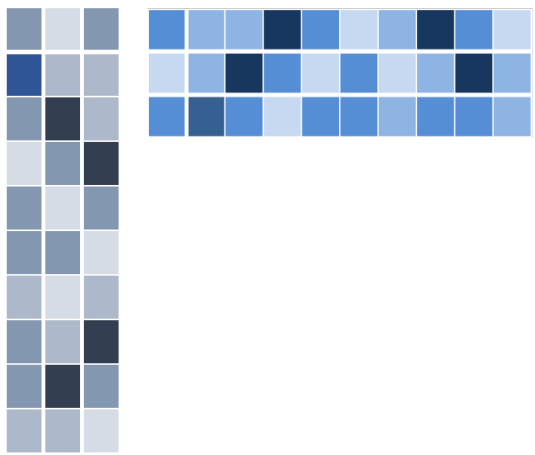
goal: estimate $\boldsymbol{M}^{\star}$

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

Nonconvex matrix completion

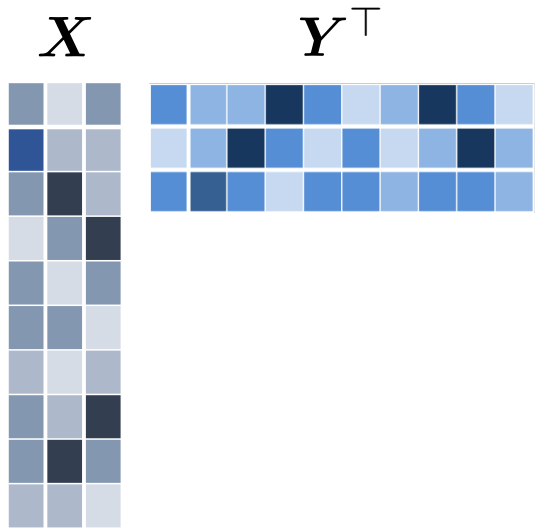
Burer–Monteiro: represent $\boldsymbol{Z}$ by $\boldsymbol{X}\boldsymbol{Y}^\top$ with $\underbrace{\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n \times r}}_{\text{low-rank factors}}$

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



Nonconvex matrix completion

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



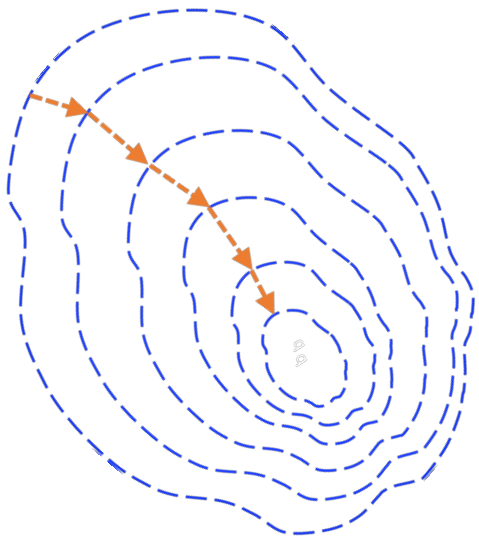
$$\underset{\boldsymbol{X},\boldsymbol{Y} \in \mathbb{R}^{n \times r}}{\text{minimize}} \quad f(\boldsymbol{X}, \boldsymbol{Y}) = \underbrace{\sum_{(i,j) \in \Omega} \left[\left(\boldsymbol{X}\boldsymbol{Y}^\top \right)_{i,j} - M_{i,j} \right]^2}_{\text{squared loss}} + \text{reg}(\boldsymbol{X}, \boldsymbol{Y})$$

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

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

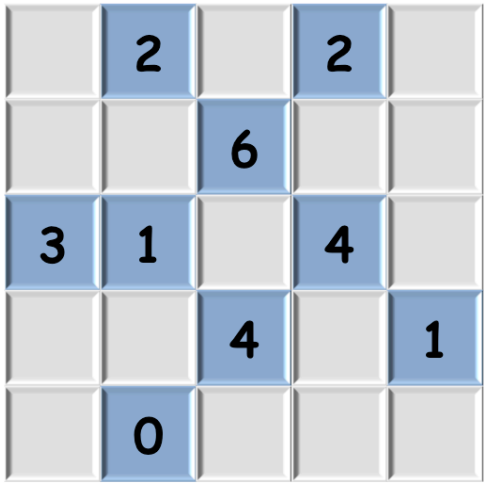


- **suitable initialization:** $(\boldsymbol{X}^0, \boldsymbol{Y}^0)$
- **gradient descent:** for $t = 0, 1, \ldots$

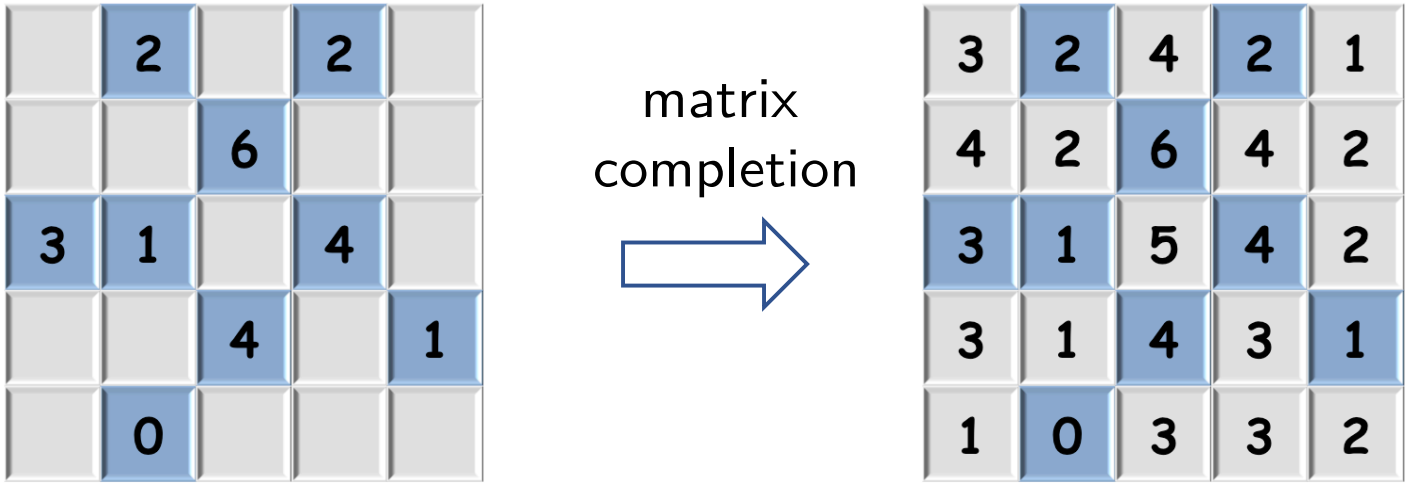
$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla_{\boldsymbol{X}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$
$$\boldsymbol{Y}^{t+1} = \boldsymbol{Y}^t - \eta_t \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}^t, \boldsymbol{Y}^t)$$

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

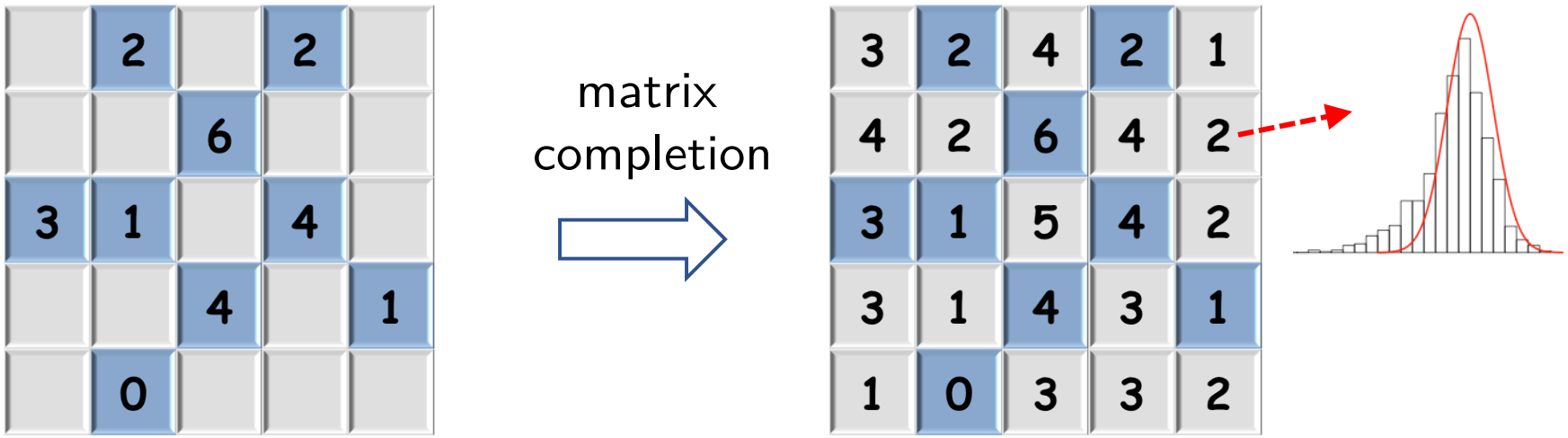
One step further: reasoning about uncertainty?



One step further: reasoning about uncertainty?



One step further: reasoning about uncertainty?

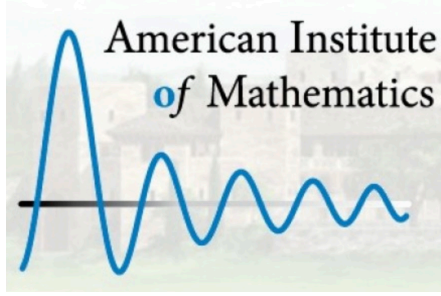


How to assess uncertainty, or "confidence", of obtained estimates due to imperfect data acquisition?

- noise
- incomplete measurements
- ...

INFERENCE IN HIGH DIMENSIONAL REGRESSION

organized by
Peter Buehlmann, Andrea Montanari, and Jonathan Taylor



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

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

Challenges

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

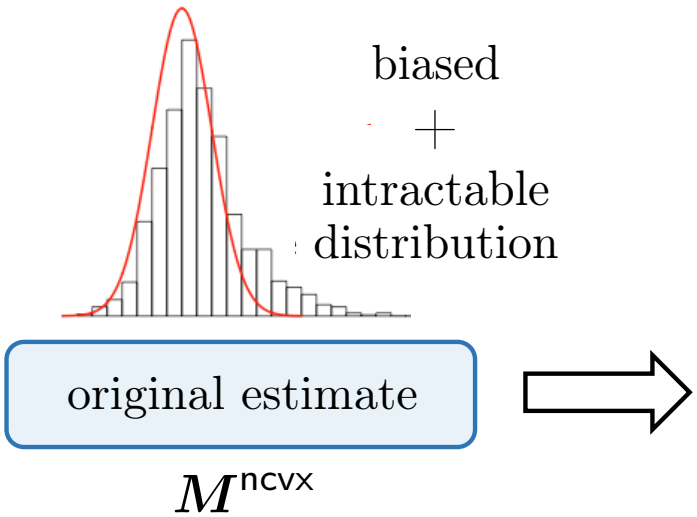
- very challenging to pin down distributions of obtained estimates

Challenges

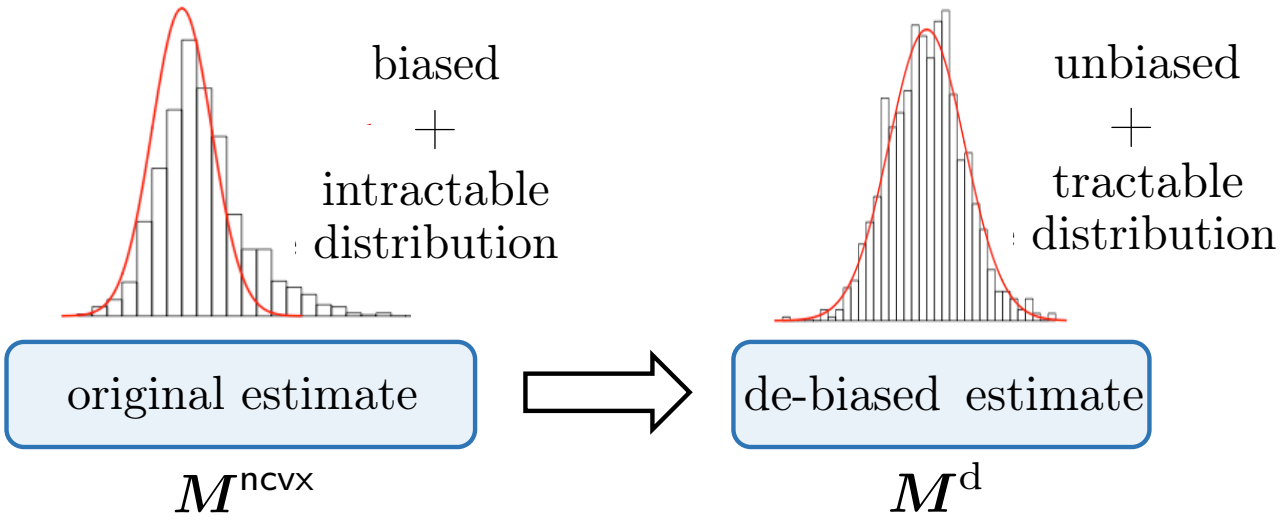
$$\boldsymbol{M}^{\text{ncvx}} \triangleq \arg\min_{\boldsymbol{Z}} \quad \underbrace{f(\boldsymbol{X}, \boldsymbol{Y}; \text{data})}_{\text{empirical loss}} + \text{reg}(\boldsymbol{X}, \boldsymbol{Y})$$

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

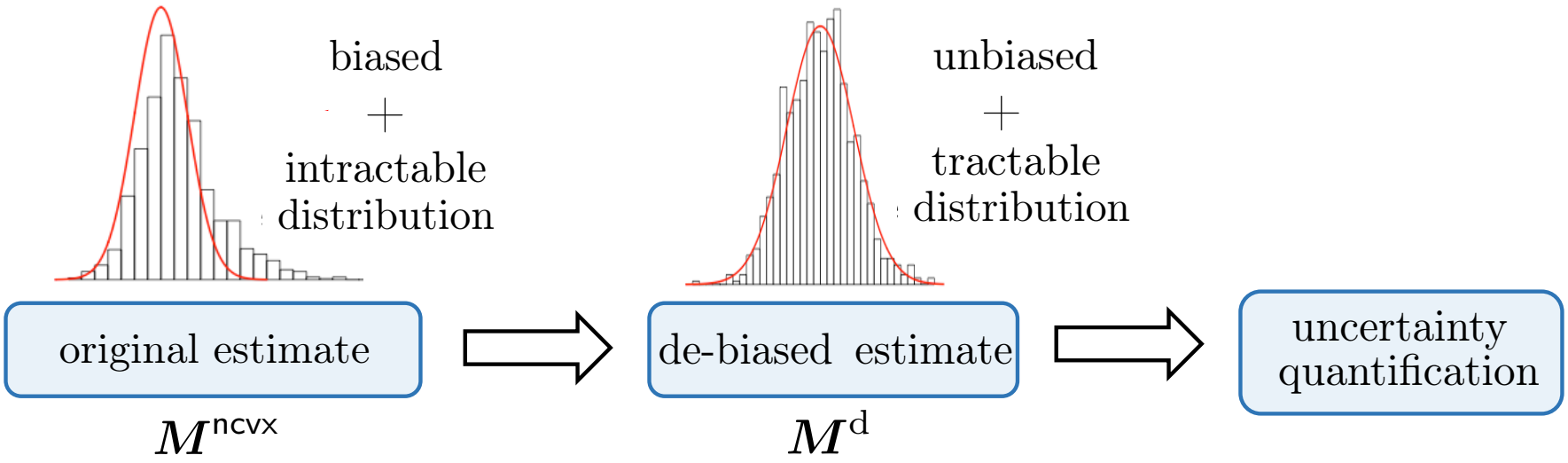
— *inspired by Zhang, Zhang '11, van de Geer et al. '13, Javanmard, Montanari '13*



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



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



De-biasing nonconvex estimate

$$\boldsymbol{M}^{\text{ncvx}} \quad \overset{\text{de-biasing}}{\longrightarrow} \quad \underbrace{\boldsymbol{M}^{\text{ncvx}} + \tfrac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E} - \boldsymbol{M}^{\text{ncvx}})}_{\text{(nearly) unbiased estimate of } \boldsymbol{M}^{\star}}$$

De-biasing nonconvex estimate

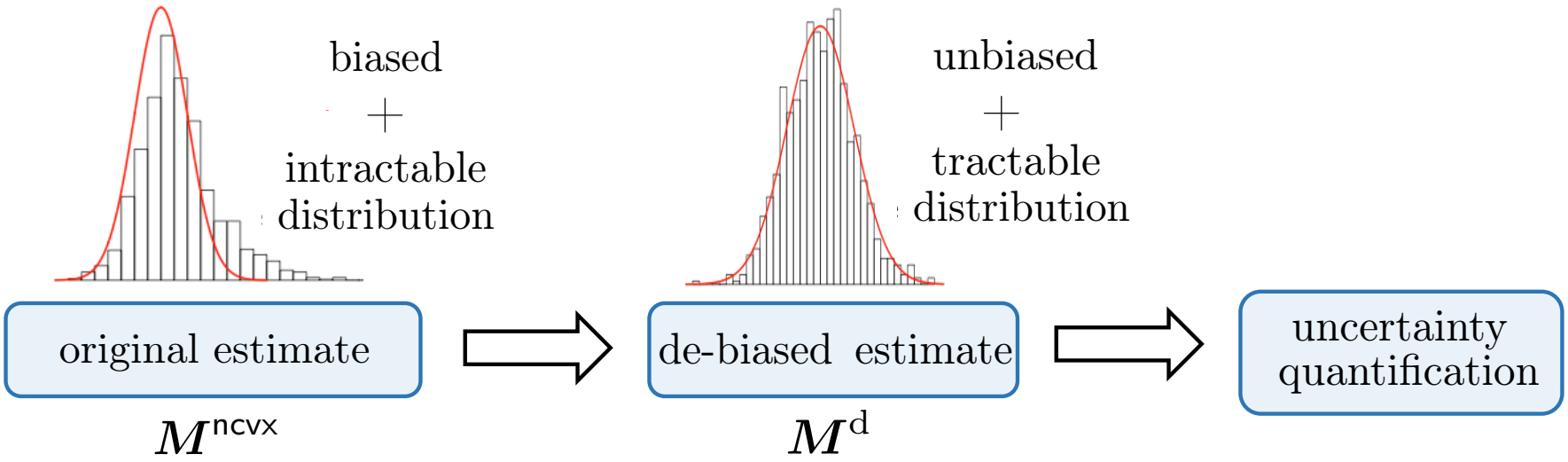
$$\boldsymbol{M}^{\text{ncvx}} \quad \overset{\text{de-biasing}}{\longrightarrow} \quad \underbrace{\boldsymbol{M}^{\text{ncvx}} + \tfrac{1}{p}\mathcal{P}_{\Omega}(\boldsymbol{M}^{\star} + \boldsymbol{E} - \boldsymbol{M}^{\text{ncvx}})}_{\text{(nearly) unbiased estimate of } \boldsymbol{M}^{\star}}$$

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

De-biasing nonconvex estimate

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

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


biased
+
intractable
distribution
unbiased
+
tractable
distribution
original estimate
M^{ncvx}
de-biased estimate
M^{d}
uncertainty
quantification

Distributional guarantees for low-rank factors

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

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

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

Distributional guarantees for low-rank factors

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

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

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

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

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

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

Implications

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

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

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

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

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

Implications

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

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

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

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

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

Implications

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

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

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

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

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

— *asymptotically optimal*

Implications

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

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

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

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

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

Implications

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

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

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

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

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

Implications

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

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

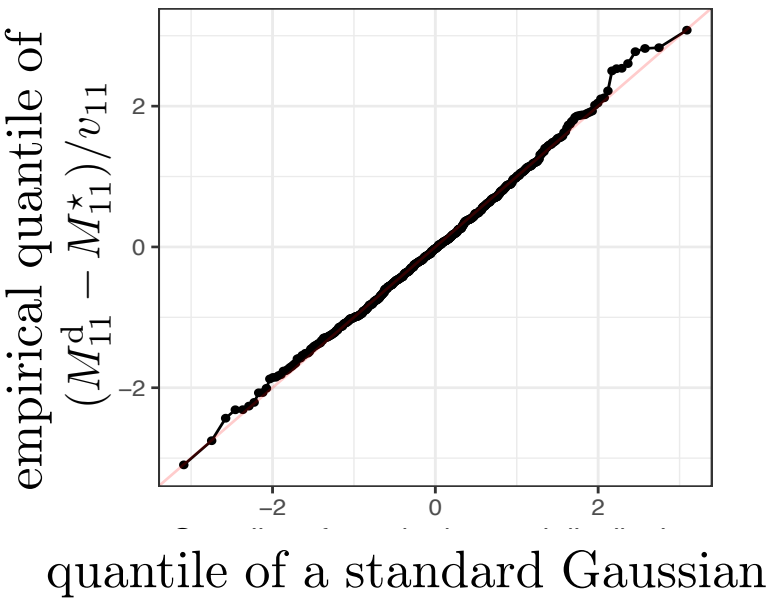
- accurate uncertainty quantification for matrix entries: if $\|\boldsymbol{X}^{\star}_{i,\cdot}\|_2 + \|\boldsymbol{Y}^{\star}_{j,\cdot}\|_2$ is not too small, then

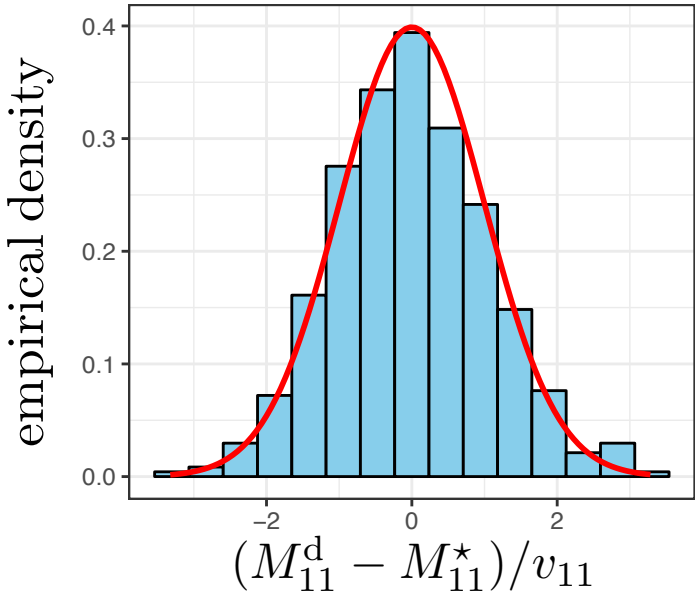
$$M^{\mathsf{d}}_{i,j} - M^{\star}_{i,j} \;\sim\; \mathcal{N}\big(0, \widehat{v}_{i,j}\big) + \text{negligible term}$$

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

— *asymptotically optimal*

Numerical experiments





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

— *Chen, Chi, Fan, Ma, Yan '19*

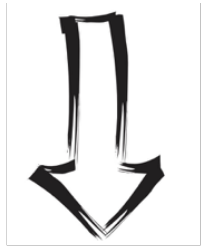
convex

nonconvex

— *Chen, Chi, Fan, Ma, Yan '19*

convex

nonconvex



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

Same inference procedures work for both cvx & noncvx estimates!

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)

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

Back to estimation: de-biased estimator is optimal

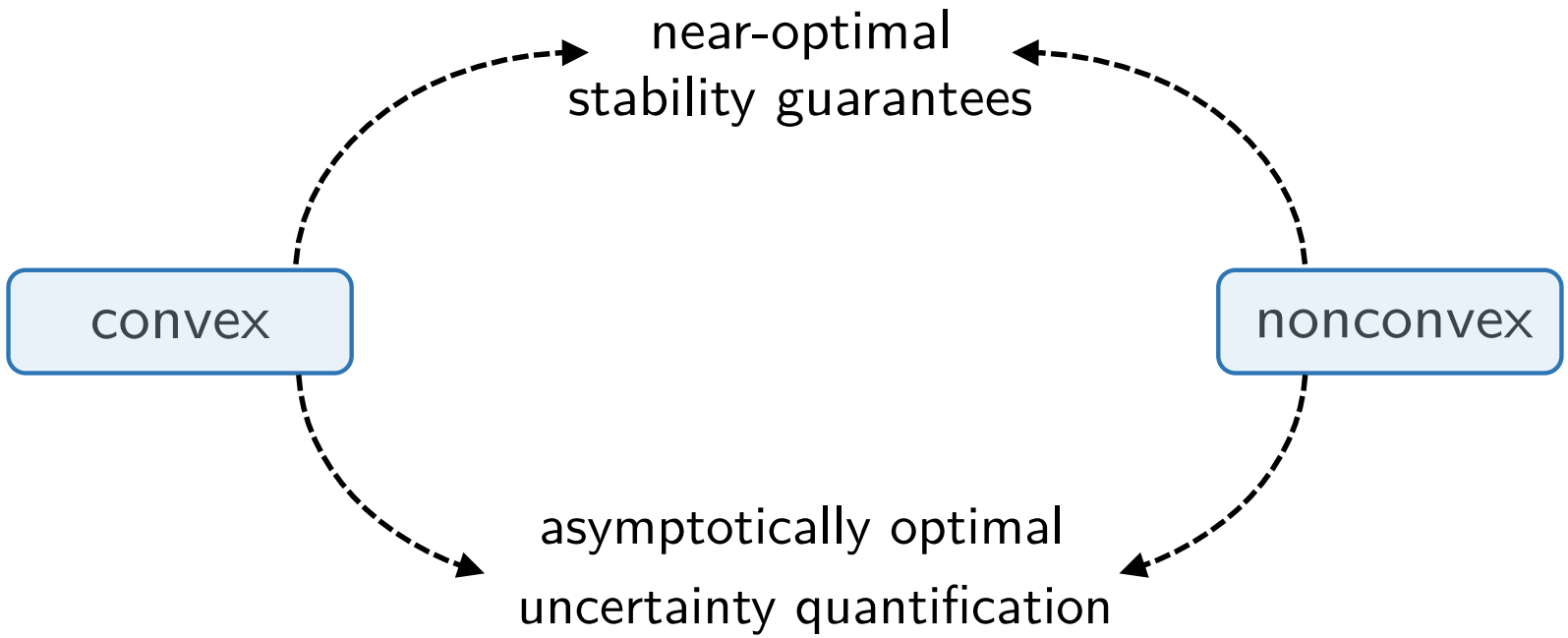
Distributional theory in turn allows us to track estimation accuracy

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

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

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

Summary for Part 2



1. "Inference and uncertainty quantification for noisy matrix completion", 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