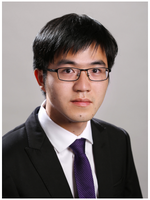
Taming Nonconvexity in Tensor Completion: Fast Convergence & Uncertainty Quantification

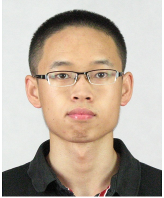
Yuxin Chen

Princeton University

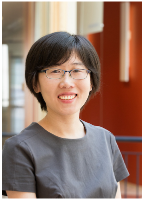
This talk: nonconvex tensor completion



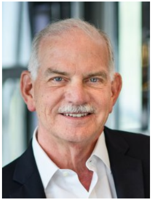
Changxiao Cai
Princeton



Gen Li
Tsinghua

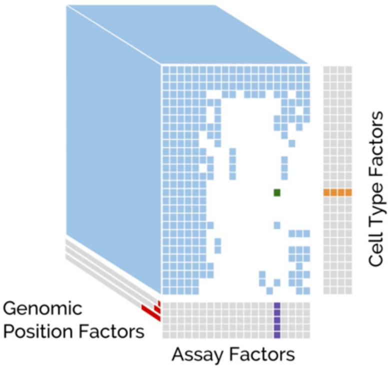


Yuejie Chi
CMU

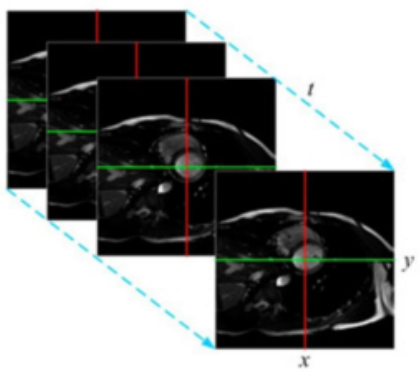


H. Vincent Poor
Princeton

Ubiquity of high-dimensional tensor data

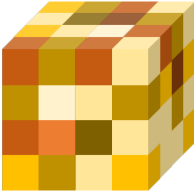


computational genomics
— *fig. credit: Schreiber et al. 19*

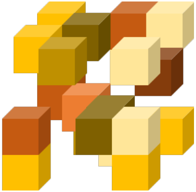


dynamic MRI
— *fig. credit: Liu et al. 17*

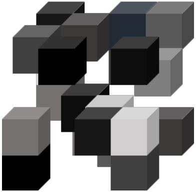
Imperfect data acquisition



a tensor of interest



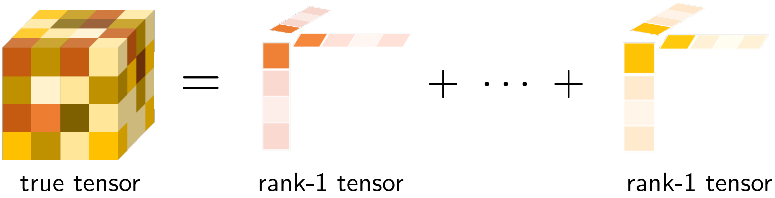
missing data



noise

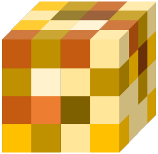
Key to enabling reliable reconstruction from incomplete data

— exploiting **low CP-rank structure**

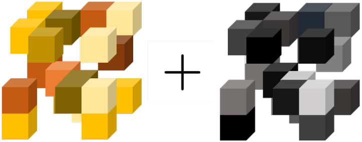


$$\boldsymbol{T}^{\star} = \sum_{i=1}^{r} \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star}$$

Setup



$\boldsymbol{T}^{\star}$

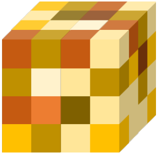


$\boldsymbol{T}$

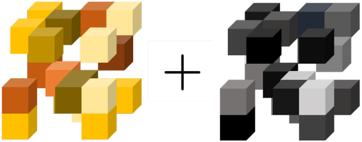
- unknown rank-r tensor $\boldsymbol{T}^{\star}$:

$$\boldsymbol{T}^{\star} = \sum_{i=1}^{r} \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star} \in \mathbb{R}^{d \times d \times d}$$

Setup



$\boldsymbol{T}^\star$



$\boldsymbol{T}$

- unknown rank-r tensor $\boldsymbol{T}^\star$:

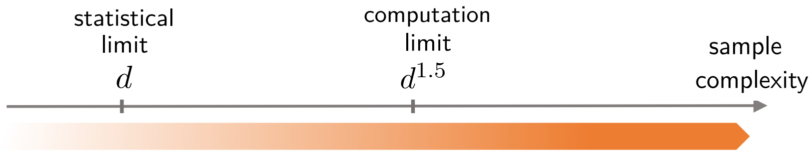
$$\boldsymbol{T}^\star = \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star \in \mathbb{R}^{d\times d\times d}$$

- incomplete & noisy observations over a random sampling set Ω:

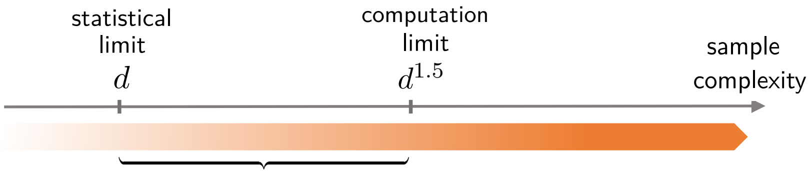
$$T_{i,j,k} = T^\star_{i,j,k} + \text{noise}, \qquad (i,j,k) \in \Omega$$

Goal: recover $\{\boldsymbol{u}_i^\star\}_{i=1}^{r}$ and $\boldsymbol{T}^\star$

Statistical-computational gap ($r = O(1)$)

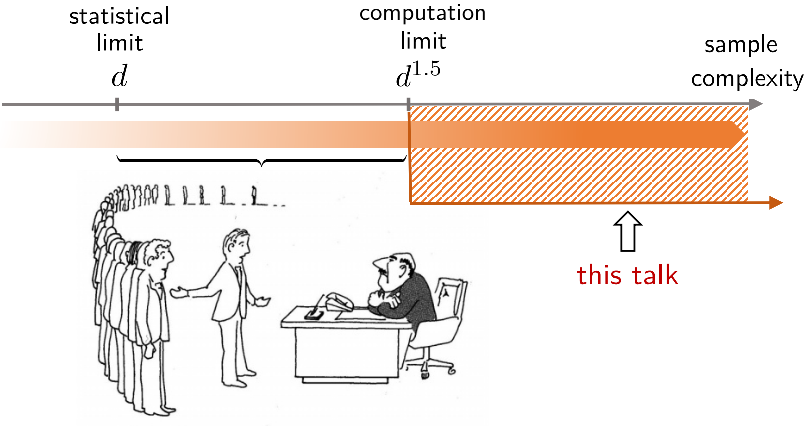


Statistical-computational gap ($r = O(1)$)



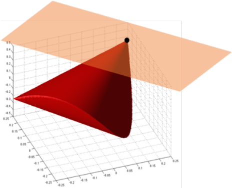
"I can't find an efficient algorithm, but neither can all these people."

Statistical-computational gap ($r = O(1)$)

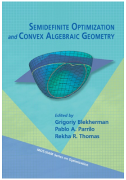


"I can't find an efficient algorithm, but neither can all these people."

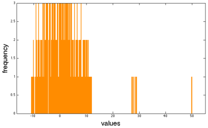
Prior art



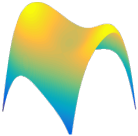
convex relaxation



sum-of-squares hierarchy



spectral methods



nonconvex optimization

- Gandy, Recht, Yamada '11
- Liu, Musialski, Wonka, Ye '12
- Kressner, Steinlechner, Vandereycken '13
- Xu, Hao, Yin, Su '13
- Romera-Paredes, Pontil '13
- Jain, Oh '14
- Huang, Mu, Goldfarb, Wright '15
- Barak, Moitra '16
- Zhang, Aeron '16
- Yuan, Zhang '16
- Montanari, Sun '16
- Kasai, Mishra '16
- Potechin, Steurer '17
- Dong, Yuan, Zhang '17
- Xia, Yuan '19
- Zhang '19
- . . .

Prior art ($r = O(1)$)

	algorithm	sample size	comput. cost	recovery type (noiseless)
Yuan, Zhang '16	tensor nuclear norm	d	NP-hard	exact
Xia, Yuan '17	spectral method + GD on manifold	$d^{3/2}$	slow	exact
Montanari, Sun '18	spectral method	$d^{3/2}$	d^3	inexact
Barak, Moitra '16	sum-of-squares	$d^{3/2}$	slow (d^{15})	exact
Potechin et al. '17	sum-of-squares	$d^{3/2}$	slow (d^{10})	exact

Prior art ($r = O(1)$)







	algorithm	sample size	comput. cost	recovery type (noiseless)
Yuan, Zhang '16	tensor nuclear norm	d	NP-hard	exact
Xia, Yuan '17	spectral method + GD on manifold	$d^{3/2}$	slow	exact
Montanari, Sun '18	spectral method	$d^{3/2}$	d^3	inexact
Barak, Moitra '16	sum-of-squares	$d^{3/2}$	slow (d^{15})	exact
Potechin et al. '17	sum-of-squares	$d^{3/2}$	slow (d^{10})	exact

	algorithm	ℓ_2 error (noisy)	ℓ_∞ error (noisy)
Xia, Yuan, Zhang '17	spectral method + tensor power method	suboptimal	n/a
Barak, Moitra '16	sum-of-squares	suboptimal	n/a

Can we design an algorithm that is simultaneously
<u>sample-efficient</u>, <u>computationally fast</u>, & <u>minimax-optimal</u>?

A nonconvex least squares formulation

$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}} \; f(\boldsymbol{U}) := \underbrace{\sum_{(i,j,k)\in\Omega} \left\{ \left(\sum\nolimits_{s=1}^{r} \boldsymbol{u}_s^{\otimes 3} \right)_{i,j,k} - T_{i,j,k} \right\}^2}_{\text{squared loss over observed entries}}$$

A nonconvex least squares formulation

$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}} \; f(\boldsymbol{U}) := \underbrace{\sum_{(i,j,k)\in\Omega}\left\{\left(\sum\nolimits_{s=1}^{r}\boldsymbol{u}_s^{\otimes 3}\right)_{i,j,k} - T_{i,j,k}\right\}^2}_{\text{squared loss over observed entries}}$$

- **pros:** statistically efficient *if we can find global solutions*

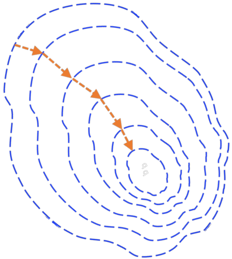
A nonconvex least squares formulation

$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}}\ f(\boldsymbol{U}) := \underbrace{\sum_{(i,j,k)\in\Omega}\Big\{\Big(\sum\nolimits_{s=1}^{r}\boldsymbol{u}_s^{\otimes 3}\Big)_{i,j,k} - T_{i,j,k}\Big\}^2}_{\text{squared loss over observed entries}}$$

- **pros:** statistically efficient *if we can find global solutions*
- **cons:** highly nonconvex $\longrightarrow$ computationally challenging

Gradient descent (GD) with random initialization?

$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}} \; f(\boldsymbol{U}) := \sum_{(i,j,k)\in\Omega} \Big\{ \Big(\sum_{s=1}^{r} \boldsymbol{u}_s^{\otimes 3} \Big)_{i,j,k} - T_{i,j,k} \Big\}^2$$



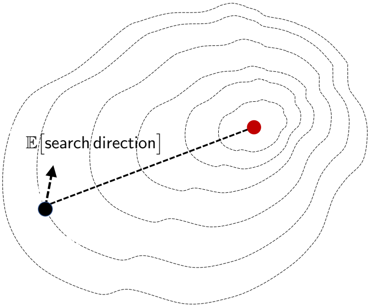
- **initialize** $\boldsymbol{U}^0$ randomly
- **gradient descent:** for $t = 0, 1, \cdots,$

$$\boldsymbol{U}^{t+1} = \boldsymbol{U}^{t} - \eta_t \nabla f(\boldsymbol{U}^{t})$$

— *succeeds for phase retrieval (Chen et al. '18)*

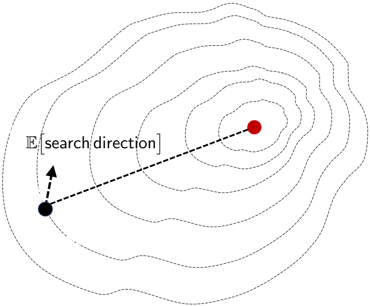
A negative conjecture

Randomly initialized GD does NOT work unless sample size $> \ d^2$



A negative conjecture

Randomly initialized GD does NOT work unless sample size $> \ d^2$

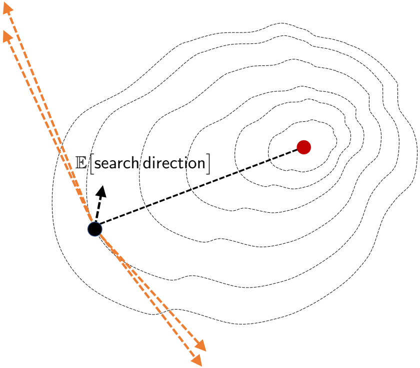


When sample size $\asymp \ d^{1.5}$:

- $\mathbb{E}\big[\text{search direction}\big]$ is desirable

A negative conjecture

Randomly initialized GD does NOT work unless sample size $> \ d^2$

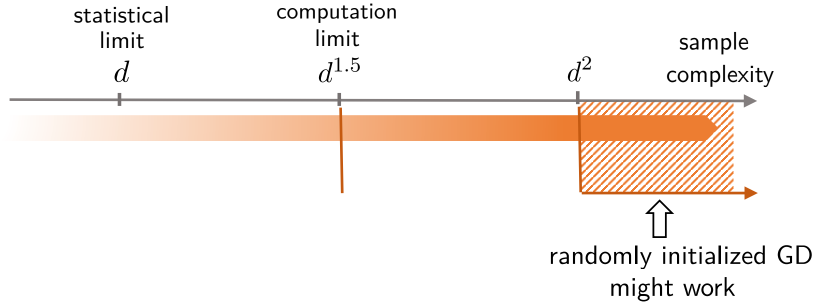


When sample size $\asymp \ d^{1.5}$:

- $\mathbb{E}\big[\text{search direction}\big]$ is desirable
- **issue:** variance $\gtrsim \sqrt{d}\,\text{mean}^2$

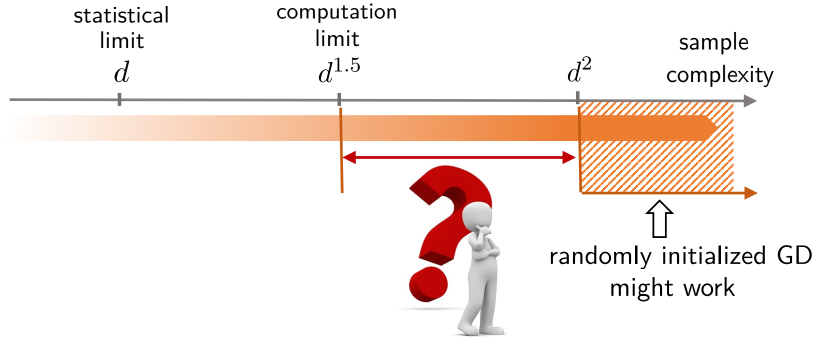
A negative conjecture

Randomly initialized GD does NOT work unless sample size $> \ d^2$



A negative conjecture

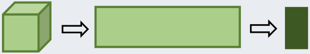
Randomly initialized GD does NOT work unless sample size $> \; d^2$



Our proposal: a two-stage nonconvex algorithm

1. **estimating subspace spanned by low-rank tensor factors**

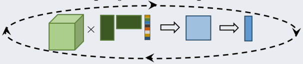
   — unfolding + spectral methods



⇩

2. **successive retrieval of tensor factors from subspace estimates**

   — random projection + spectral methods

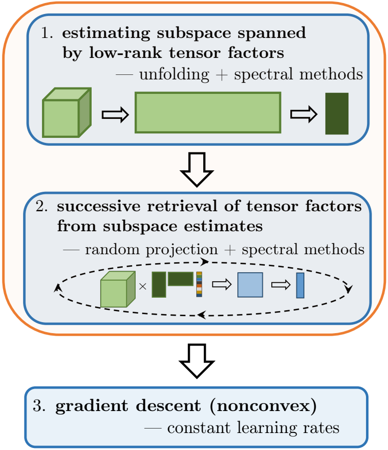


⇩

3. **gradient descent (nonconvex)**

   — constant learning rates

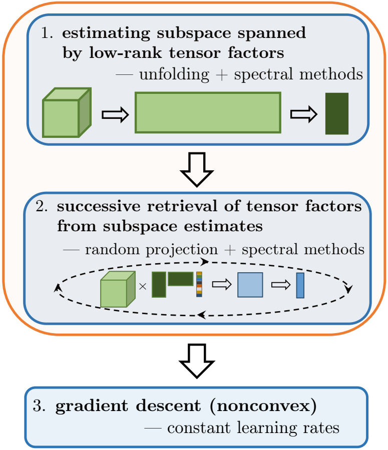
Our proposal: a two-stage nonconvex algorithm



1. initialization: $\boldsymbol{U}^0$

- estimate span$\{\boldsymbol{u}_i^\star\}$ via spectral method
- disentangle individual factors $\{\boldsymbol{u}_i^\star\}$ from subspace estimate

Our proposal: a two-stage nonconvex algorithm



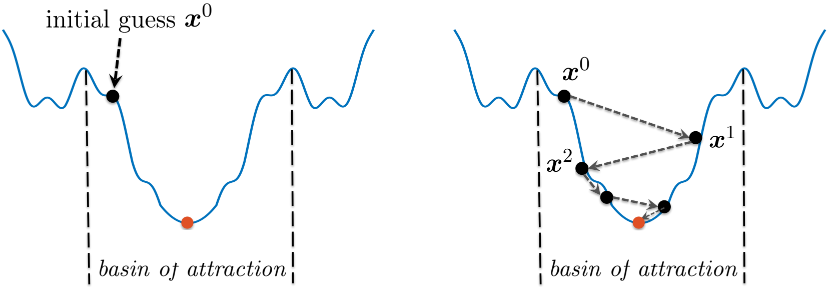
1. initialization: $\boldsymbol{U}^0$

- estimate span$\{\boldsymbol{u}_i^\star\}$ via spectral method
- disentangle individual factors $\{\boldsymbol{u}_i^\star\}$ from subspace estimate

2. gradient descent: for $t = 0, 1, \cdots$

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

Rationale of two-stage approach



1. initialize within $\underbrace{\text{a local basin sufficiently close to global min}}_{\text{(restricted) strongly convex}}$

2. iterative refinement

A bit more details about initialization

Step 1.1: estimate $\text{span}\{\boldsymbol{u}_i^\star\}_{1\le i\le r} \quad \longrightarrow \quad \boldsymbol{U}_{\text{sub}}$

- matricizition: $\boldsymbol{A} = \text{unfold}(\boldsymbol{T})$
- estimate rank-r subspace of $\mathcal{P}_{\text{off-diag}}(\boldsymbol{A}\boldsymbol{A}^\top)$ (diagonal deletion)

A bit more details about initialization

Step 1.2: retrieve tensor factors from subspace estimate

- generate a random vector $\boldsymbol{g}$ from $\boldsymbol{U}_{\mathsf{sub}}$
- compute leading eigenvector of $\boldsymbol{T} \otimes \boldsymbol{g} = \underbrace{\sum_i \langle \boldsymbol{u}_i^\star, \boldsymbol{g} \rangle \boldsymbol{u}_i^\star \boldsymbol{u}_i^{\star\top} + \text{noise}}_{\text{find the } \boldsymbol{u}_i^\star \text{ most aligned with } \boldsymbol{g}}$
- repeat ...

Assumptions

$$\boldsymbol{T}^{\star} = \sum_{i=1}^{r} \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star} \in \mathbb{R}^{d\times d\times d}$$

- **random sampling**: each entry is observed independently with prob. p

Assumptions

$$\boldsymbol{T}^\star = \sum_{i=1}^{r} \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star \otimes \boldsymbol{u}_i^\star \in \mathbb{R}^{d\times d\times d}$$

- **random sampling**: each entry is observed independently with prob. p
- **random noise**: independent zero-mean sub-Gaussian noise with variance $O(\sigma^2)$

Assumptions

$$\boldsymbol{T}^{\star} = \sum_{i=1}^{r} \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star} \otimes \boldsymbol{u}_i^{\star} \in \mathbb{R}^{d \times d \times d}$$

- **random sampling**: each entry is observed independently with prob. p
- **random noise**: independent zero-mean sub-Gaussian noise with variance $O(\sigma^2)$
- **ground truth**: low-rank $\big(r = O(1)\big)$, well-conditioned, incoherent $\big(\{\boldsymbol{u}_i^{\star}\}$ are de-localized and not aligned$\big)$

ℓ_2 and ℓ_∞ theoretical guarantees

Theorem 1 (Cai, Li, Poor, Chen '19)

There exists some constant $\rho < 1$ *and some permutation matrix* $\mathbf{\Pi} \in \mathbb{R}^{r \times r}$ *s.t. with high prob., the* t*-th iterate satisfies*

$$\left\| \boldsymbol{U}^t \mathbf{\Pi} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right) \left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}}$$

$$\left\| \boldsymbol{T}^t - \boldsymbol{T}^\star \right\|_{\mathrm{F}} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right) \left\| \boldsymbol{T}^\star \right\|_{\mathrm{F}}$$

ℓ_2 and ℓ_∞ theoretical guarantees

Theorem 1 (Cai, Li, Poor, Chen '19)

There exists some constant $\rho < 1$ *and some permutation matrix* $\mathbf{\Pi} \in \mathbb{R}^{r\times r}$ *s.t. with high prob., the* t*-th iterate satisfies*

$$\left\|\boldsymbol{U}^t\mathbf{\Pi} - \boldsymbol{U}^\star\right\|_{\mathrm{F}} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right)\left\|\boldsymbol{U}^\star\right\|_{\mathrm{F}}$$

$$\left\|\boldsymbol{T}^t - \boldsymbol{T}^\star\right\|_{\mathrm{F}} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right)\left\|\boldsymbol{T}^\star\right\|_{\mathrm{F}}$$

$$\left\|\boldsymbol{U}^t\mathbf{\Pi} - \boldsymbol{U}^\star\right\|_{\infty} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right)\left\|\boldsymbol{U}^\star\right\|_{\infty}$$

$$\left\|\boldsymbol{T}^t - \boldsymbol{T}^\star\right\|_{\infty} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right)\left\|\boldsymbol{T}^\star\right\|_{\infty}$$

provided that sample size $\gtrsim d^{1.5}\mathrm{poly}\log(d)$

ℓ_2 and ℓ_∞ theoretical guarantees

Theorem 1 (Cai, Li, Poor, Chen '19)

There exists some constant $\rho < 1$ and some permutation matrix $\mathbf{\Pi} \in \mathbb{R}^{r \times r}$ s.t. with high prob., the t-th iterate satisfies

$$\begin{aligned}
\left\| \boldsymbol{U}^t \mathbf{\Pi} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} &\lesssim \left(\rho^t + \sigma \sqrt{d/p}\right) \left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}} \\
\left\| \boldsymbol{T}^t - \boldsymbol{T}^\star \right\|_{\mathrm{F}} &\lesssim \left(\rho^t + \sigma \sqrt{d/p}\right) \left\| \boldsymbol{T}^\star \right\|_{\mathrm{F}} \\
\left\| \boldsymbol{U}^t \mathbf{\Pi} - \boldsymbol{U}^\star \right\|_{\infty} &\lesssim \left(\rho^t + \sigma \sqrt{d/p}\right) \left\| \boldsymbol{U}^\star \right\|_{\infty} \\
\left\| \boldsymbol{T}^t - \boldsymbol{T}^\star \right\|_{\infty} &\lesssim \left(\rho^t + \sigma \sqrt{d/p}\right) \left\| \boldsymbol{T}^\star \right\|_{\infty}
\end{aligned}$$

provided that sample size $\gtrsim d^{1.5} \mathrm{poly}\log(d)$

- linear/geometric convergence $\longrightarrow$ linear-time algorithm

ℓ_2 and ℓ_∞ theoretical guarantees

Theorem 1 (Cai, Li, Poor, Chen '19)

There exists some constant $\rho < 1$ and some permutation matrix $\mathbf{\Pi} \in \mathbb{R}^{r \times r}$ s.t. with high prob., the t-th iterate satisfies

$$\left\| \boldsymbol{U}^t \mathbf{\Pi} - \boldsymbol{U}^\star \right\|_{\mathrm{F}} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right) \left\| \boldsymbol{U}^\star \right\|_{\mathrm{F}}$$

$$\left\| \boldsymbol{T}^t - \boldsymbol{T}^\star \right\|_{\mathrm{F}} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right) \left\| \boldsymbol{T}^\star \right\|_{\mathrm{F}}$$

$$\left\| \boldsymbol{U}^t \mathbf{\Pi} - \boldsymbol{U}^\star \right\|_{\infty} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right) \left\| \boldsymbol{U}^\star \right\|_{\infty}$$

$$\left\| \boldsymbol{T}^t - \boldsymbol{T}^\star \right\|_{\infty} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right) \left\| \boldsymbol{T}^\star \right\|_{\infty}$$

provided that sample size $\gtrsim d^{1.5} \mathrm{poly} \log(d)$

- near-optimal sample complexity (among poly-time algorithms)

ℓ_2 and ℓ_∞ theoretical guarantees

Theorem 1 (Cai, Li, Poor, Chen '19)

There exists some constant $\rho < 1$ *and some permutation matrix* $\mathbf{\Pi} \in \mathbb{R}^{r \times r}$ *s.t. with high prob., the* t*-th iterate satisfies*

$$\left\|\boldsymbol{U}^t\mathbf{\Pi} - \boldsymbol{U}^\star\right\|_{\mathrm{F}} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right)\left\|\boldsymbol{U}^\star\right\|_{\mathrm{F}}$$

$$\left\|\boldsymbol{T}^t - \boldsymbol{T}^\star\right\|_{\mathrm{F}} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right)\left\|\boldsymbol{T}^\star\right\|_{\mathrm{F}}$$

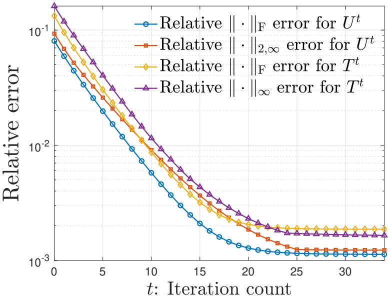
$$\left\|\boldsymbol{U}^t\mathbf{\Pi} - \boldsymbol{U}^\star\right\|_{\infty} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right)\left\|\boldsymbol{U}^\star\right\|_{\infty}$$

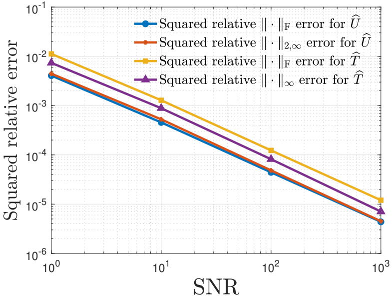
$$\left\|\boldsymbol{T}^t - \boldsymbol{T}^\star\right\|_{\infty} \lesssim \left(\rho^t + \sigma\sqrt{d/p}\right)\left\|\boldsymbol{T}^\star\right\|_{\infty}$$

provided that sample size $\gtrsim d^{1.5}\mathrm{poly}\log(d)$

- near-optimal statistical accuracy (both ℓ_2 and ℓ_∞)

Numerical experiments



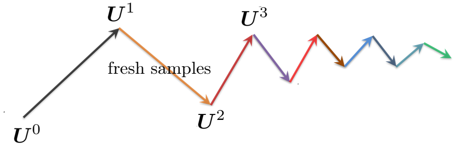


$$d = 100,\ r = 4,\ p = 0.1$$

No need of sample splitting

Sample-splitting (fresh samples at each iteration)?

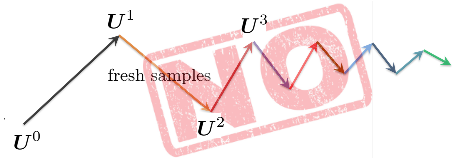
— helps analysis but waste data



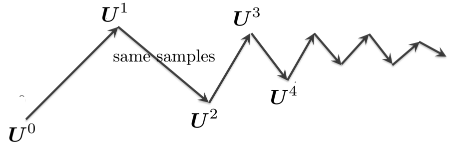
No need of sample splitting

Sample-splitting (fresh samples at each iteration)?

— helps analysis but waste data



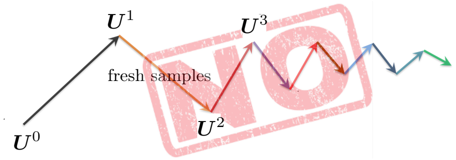
Our results: reusing all samples in all iterations



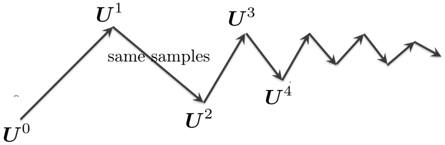
No need of sample splitting

Sample-splitting (fresh samples at each iteration)?

— helps analysis but waste data



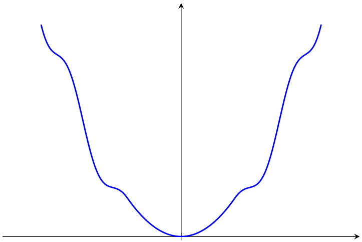
Our results: reusing all samples in all iterations



How to deal with complicated statistical dependency across iterations?

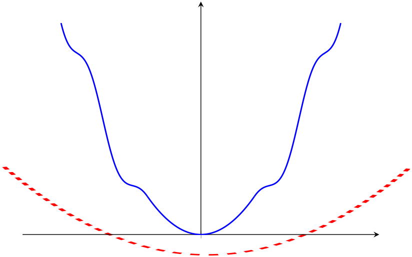
A little analysis

Gradient descent theory revisited



Standard conditions that enable fast convergence of GD

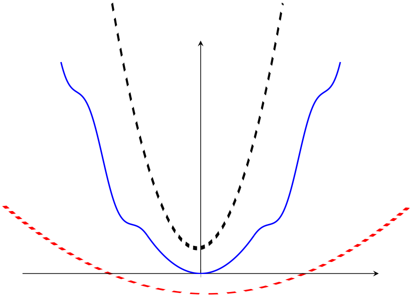
Gradient descent theory revisited



Standard conditions that enable fast convergence of GD

- α-strong convexity within ℓ_2 ball

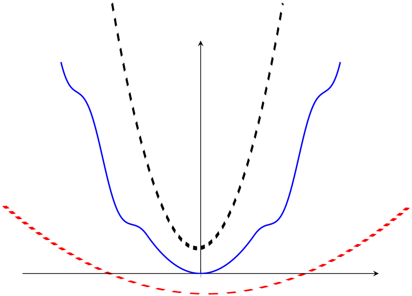
Gradient descent theory revisited



Standard conditions that enable fast convergence of GD

- α-strong convexity within ℓ_2 ball
- β-Lipschitz gradients within ℓ_2 ball

Gradient descent theory revisited

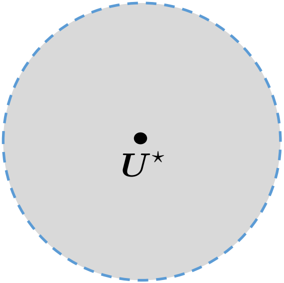


Standard conditions that enable fast convergence of GD

- α-strong convexity within ℓ_2 ball
- β-Lipschitz gradients within ℓ_2 ball

error contraction: $\|\boldsymbol{x}^{t+1} - \boldsymbol{x}^\star\|_2 \le \left(1 - \frac{\alpha}{\beta}\right) \|\boldsymbol{x}^t - \boldsymbol{x}^\star\|_2$

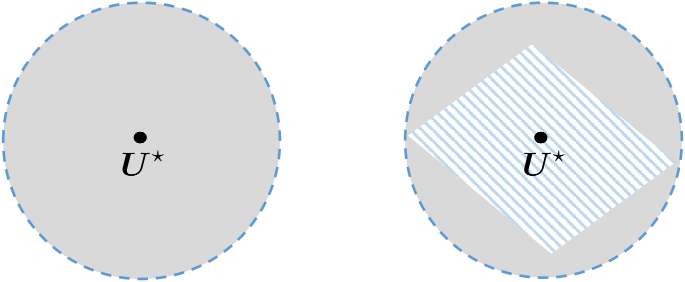
Local optimization landscape



- **Bad news:** f is NOT strongly convex in local ℓ_2 ball (unless the radius is exceedingly small)

Local optimization landscape

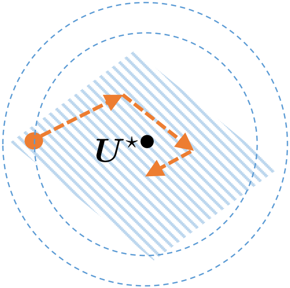
 region of local strong convexity + smoothness



- **Bad news:** f is NOT strongly convex in local ℓ_2 ball (unless the radius is exceedingly small)
- f is strongly convex and well-conditioned in (restricted) ℓ_∞ ball

Our findings: GD controls entrywise error

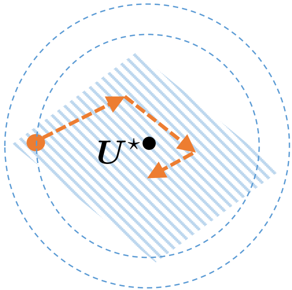
region of local strong convexity + smoothness



Good news: GD implicitly controls ℓ_∞ error

Our findings: GD controls entrywise error

region of local strong convexity + smoothness



Good news: GD implicitly controls ℓ_∞ error

— cannot be derived from generic optimization theory

— requires fine-grained statistical analysis for entire trajectory

Key proof idea: leave-one-out decoupling

Leave out a small amount of randomness and re-run the algorithm

Key proof idea: leave-one-out decoupling

Leave out a small amount of randomness and re-run the algorithm

- El Karoui, Bean, Bickel, Lim, Yu '13
- El Karoui '15
- Javanmard, Montanari '15
- Zhong, Boumal '17
- Lei, Bickel, El Karoui '17
- Sur, Chen, Candès '17
- Abbe, Fan, Wang, Zhong '17
- Chen, Fan, Ma, Wang '17
- Ma, Wang, Chi, Chen '17
- Chen, Chi, Fan, Ma '18
- Ding, Chen '18
- Dong, Shi '18
- Sur, Candès '18
- Chen, Liu, Li '19
- Chen, Fan, Ma, Yan '19
- Pananjady, Wainwright '19
- Ling '20
- Chen, Fan, Ma, Yan '20
- Agarwal, Kakade, Yang '20
- Abbe, Fan, Wang '20
- Li, Wei, Chi, Gu, Chen '20

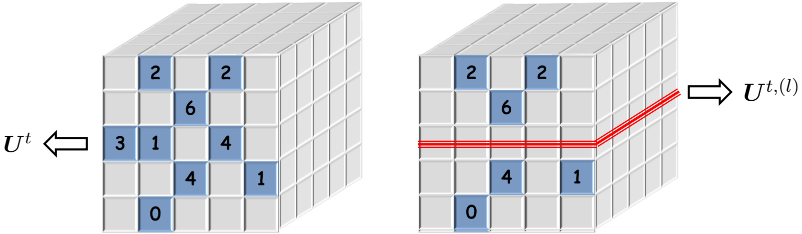
Foundations and Trends® in Machine Learning

Spectral Methods for Data Science: A Statistical Perspective

Suggested Citation: Yuxin Chen, Yuejie Chi, Jianqing Fan and Cong Ma (2020), "Spectral Methods for Data Science: A Statistical Perspective", Foundations and Trends® in

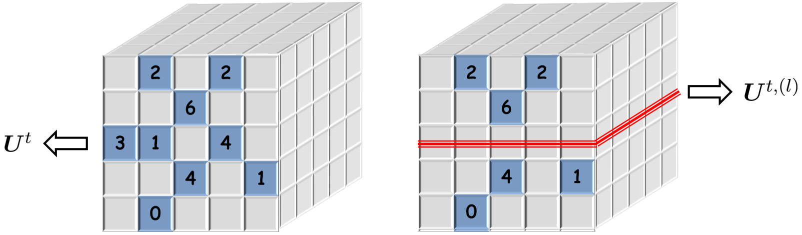
Key proof idea: leave-one-out decoupling

For each $1 \leq l \leq d$, generate leave-one-out auxiliary iterates $\{\boldsymbol{U}^{t,(l)}\}$ by replacing l^{th} slice with true values



Key proof idea: leave-one-out decoupling

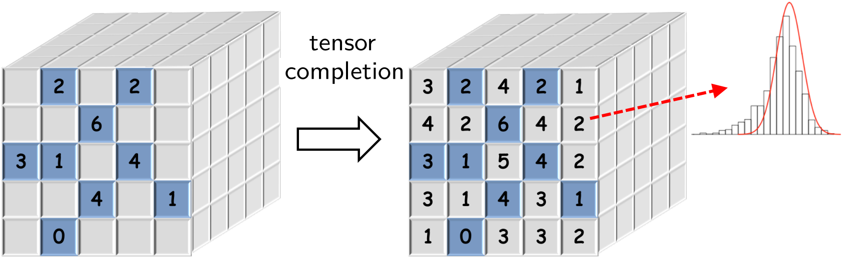
For each $1 \le l \le d$, generate leave-one-out auxiliary iterates $\{\boldsymbol{U}^{t,(l)}\}$ by replacing l^{th} slice with true values



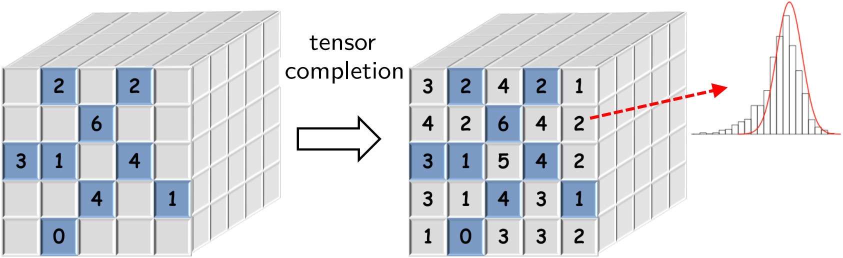
- exploit partial statistical independence
- exploit leave-one-out stability
- enable optimal ℓ_∞ error control

Inference and uncertainty quantification

One step further: uncertainty quantification?



One step further: uncertainty quantification?



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

- noise
- missing data

Challenges

$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}} \; f(\boldsymbol{U}) := \underbrace{\sum_{(i,j,k)\in\Omega} \left\{ \Big(\sum\nolimits_{s=1}^{r} \boldsymbol{u}_s^{\otimes 3} \Big)_{i,j,k} - T_{i,j,k} \right\}^2}_{\text{squared loss}}$$

- How to pin down distributions of nonconvex solutions?

Challenges

$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}} \; f(\boldsymbol{U}) := \underbrace{\sum_{(i,j,k)\in\Omega}\Big\{\Big(\sum\nolimits_{s=1}^{r}\boldsymbol{u}_s^{\otimes 3}\Big)_{i,j,k} - T_{i,j,k}\Big\}^2}_{\text{squared loss}}$$

- How to pin down distributions of nonconvex solutions?
- How to adapt to unknown noise distributions and heteroscedasticity (i.e. location-varying noise variance)?

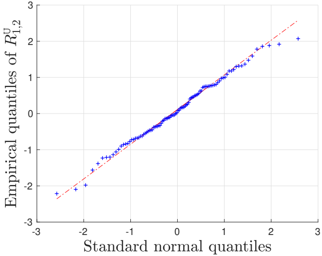
Challenges

$$\underset{\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_r]\in\mathbb{R}^{d\times r}}{\text{minimize}} \ f(\boldsymbol{U}) := \underbrace{\sum_{(i,j,k)\in\Omega} \left\{ \left(\sum\nolimits_{s=1}^{r} \boldsymbol{u}_s^{\otimes 3} \right)_{i,j,k} - T_{i,j,k} \right\}^2}_{\text{squared loss}}$$

- How to pin down distributions of nonconvex solutions?
- How to adapt to unknown noise distributions and heteroscedasticity (i.e. location-varying noise variance)?
- Existing estimation guarantees are highly insufficient
  $\longrightarrow$ Overly wide confidence intervals

Distributional theory

- random sampling
- independent Gaussian noise
- ground truth: low-rank, incoherent, well-conditioned



Theorem 2 (Cai, Poor, Chen '20)

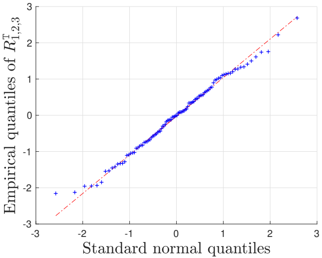
Consider any (i, j, k) s.t. the corresponding "SNR" is not exceedingly small. Then with high prob.,

$$\widehat{T}_{i,j,k} - T^{\star}_{i,j,k} \sim \mathcal{N}(0, \textit{Cramér-Rao}) + \textit{negligible term}$$

— *asymptotically optimal*

Distributional theory

- random sampling
- independent Gaussian noise
- ground truth: low-rank, incoherent, well-conditioned

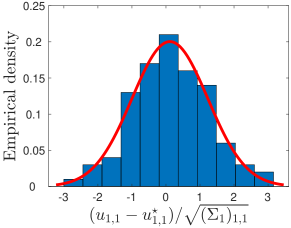


Theorem 2 (Cai, Poor, Chen '20)

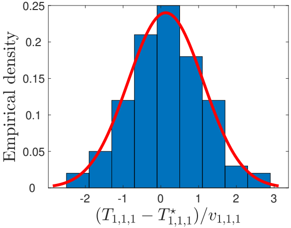
Consider any (i, j, k) s.t. the corresponding "SNR" is not exceedingly small. Then with high prob.,

$$\widehat{T}_{i,j,k} - T^{\star}_{i,j,k} \sim \mathcal{N}(0, \textit{Cramér-Rao}) + \textit{negligible term}$$

— *asymptotically optimal*

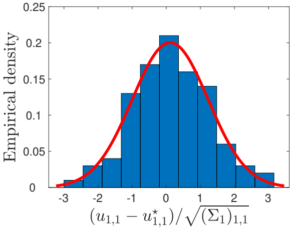


tensor factor entry

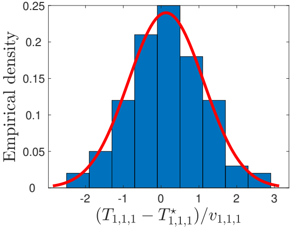


tensor entry

- **approximate Gaussianality:** estimation error of our nonconvex approach is zero-mean Gaussian

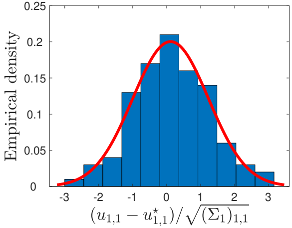


tensor factor entry

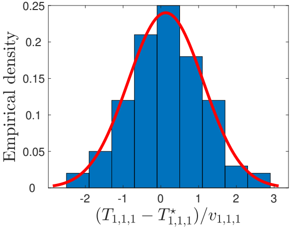


tensor entry

- **approximate Gaussianality:** estimation error of our nonconvex approach is zero-mean Gaussian
- **confidence intervals:** error (co)-variance can be accurately estimated, leading to valid CI construction



tensor factor entry



tensor entry

- **approximate Gaussianality:** estimation error of our nonconvex approach is zero-mean Gaussian
- **confidence intervals:** error (co)-variance can be accurately estimated, leading to valid CI construction
- **adaptivity:** our procedure is data-driven, and adaptive to unknown and heteroscedastic noise levels

Back to estimation: ℓ_2 optimality

Distributional theory in turn allows us to track estimation accuracy

Theorem 3 (Cai, Poor, Chen '20)

Suppose noise is i.i.d. $\mathcal{N}(0, \sigma^2)$. Then one has

$$\|\widehat{\boldsymbol{T}} - \boldsymbol{T}^{\star}\|_{\mathrm{F}}^2 = \underbrace{\frac{(6 + o(1))\sigma^2 r d}{p}}_{\text{Cramér-Rao lower bound}}$$

Back to estimation: ℓ_2 optimality

Distributional theory in turn allows us to track estimation accuracy

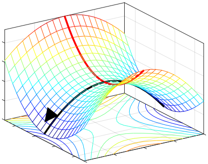
Theorem 3 (Cai, Poor, Chen '20)

Suppose noise is i.i.d. $\mathcal{N}(0, \sigma^2)$. Then one has

$$\|\widehat{\boldsymbol{T}} - \boldsymbol{T}^\star\|_{\mathrm{F}}^2 = \underbrace{\frac{(6 + o(1))\sigma^2 r d}{p}}_{\text{Cramér-Rao lower bound}}$$

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

Summary



nonconvex
optimization

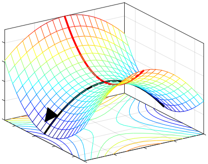
optimal estimation guarantees

linear-time algorithm

minimal sample size

fine-grained uncertainty quantification

Summary



nonconvex
optimization

optimal estimation guarantees

linear-time algorithm

minimal sample size

fine-grained uncertainty quantification

phase retrieval | matrix completion | ranking | blind deconvolution | reinforcement learning

Papers

"Nonconvex low-rank tensor completion from noisy data" C. Cai, G. Li, H. V. Poor, Y. Chen, Operation Research, 2021+

"Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees," C. Cai, G. Li, Y. Chi, H. V. Poor, Y. Chen, Annals of Statistics, 2021+

"Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality," C. Cai, H. V. Poor, Y. Chen, ICML 2020