Non-convex matrix sensing: Breaking the quadratic rank barrier in the sample complexity

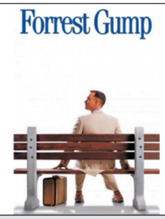
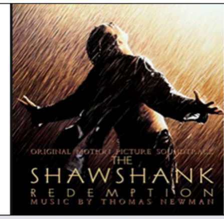
Dominik Stöger, KU Eichstätt-Ingolstadt

Collaborator

Yizhe Zhu (University of Southern California)

Low-rank matrix recovery problems

Matrix Completion:

				
Bob	?	?	1	2
Alice	?	?	3	?
Joe	3	1	?	?
Sam	?	?	?	5

Many other problems: Blind deconvolution, Phase Retrieval, …

Problem setting

- Linear observations $y_i = \langle \mathbf{A}_i, \mathbf{X}_\star \rangle :=$ trace$(\mathbf{A}_i \mathbf{X}_\star)$ for $i = 1, \ldots, m$
- $\mathbf{A}_i \in \mathbb{R}^{d \times d}$ known measurement matrices
- low-rank matrices $\mathbf{X}_\star \in \mathbb{R}^{d \times d}$ of rank r
- **Goal**: estimate $\mathbf{X}_\star$ from samples $y_1, y_2, \ldots, y_m$

Convex approach

Solve optimization problem

$$\min \|\mathbf{Z}\|_* \quad \text{such that } y_i = \langle \mathbf{A}_i, \mathbf{Z} \rangle \text{ for all } i = 1,\ldots,m$$

Here, $\|\cdot\|_*$ denotes the nuclear norm, i.e., sum of singular values

☺ Strong theoretical guarantees: Sample complexity $O\left(rd\right)$ suffices

☹ Computationally expensive! Requires working at least with d^2 variables

Non-convex approach

Objective function

$$f(\mathbf{U}, \mathbf{V}) = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \langle \mathbf{A}_i, \mathbf{U}\mathbf{V}^\top \rangle \right)^2$$

with $\mathbf{U} \in \mathbb{R}^{d \times r}, \mathbf{V} \in \mathbb{R}^{d \times r}$

Solve optimization problem via gradient descent, alternating minimization

☺ Computationally much faster (only $2rd$ optimization variables)

☹ **Theoretical guarantees much weaker**! At least $r^2 d$ samples needed!

The r^2-factor is everywhere!

- **Matrix sensing**: Tu, Boczar, Soltanolkotabi, Recht (2015), Li, Zhu, So, and Vidal (2020); Tong, Ma, and Chi (2021); Charisopoulos, Chen, Davis, Diaz, Ding, and Drusvyatskiy (2021); Zilber and Nadler (2022)…
- **Matrix completion**: Keshavan, Montanari, and Oh (2010); Sun and Luo (2016); Zheng, Lafferty (2016); Ge, Ma, and Lee (2016); Ma, Wang, Chi, Chen (2020); Chen, Liu and Li (2020), …
- **Blind deconvolution and demixing**: Ling and Strohmer (2019), Dong and Shi (2019)
- **Overparameterized models**: Li, Ma, and Zhang (2018); Stöger and Soltanolkotabi (2021); Jin, Li, Lyu, Du, and Li (2023); Xu, Chen, Shi, and Ma (2023); Ma and Fattahi (2023)…
- **Rank-one measurement matrices**: Li, Ma, Chen, and Chi (2020); Bahmani and Lee (2021)

This talk:

Can we get recovery guarantees, where the sample complexity depends linearly on the rank?

Our setup

- Samples $y_i = \langle \mathbf{X}_\star, \mathbf{A}_i \rangle$, $i = 1, \ldots, m$
- $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ symmetric Gaussian matrices (diagonal entries have distribution $\mathcal{N}(0,1)$ and off-diagonals have distribution $\mathcal{N}(0,1/2)$)
- **Symmetric, positive definite** ground truth $\mathbf{X}_\star \in \mathbb{R}^{d\times d}$ with rank r
- Condition number $\kappa := \lambda_1\left(\mathbf{X}_\star\right)/\lambda_r\left(\mathbf{X}_\star\right)$

Two-stage approach
(Keshavan, Montanari, Oh 2010)

Stage 1: Spectral Initialization

- Let $\mathbf{M} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^\top$ be truncated rank-r SVD of $\frac{1}{m}\sum_{i=1}^{m} y_i \mathbf{A}_i = \frac{1}{m}\sum_{i=1}^{m} \langle \mathbf{A}_i, \mathbf{X}_\star \rangle \mathbf{A}_i$
- Set $\mathbf{U}_0 := \mathbf{V}\mathbf{\Sigma_0}^{1/2} \in \mathbb{R}^{d\times r}$

Intuition**:**

For large enough sample size we have w.h.p. $\frac{1}{m}\sum_{i=1}^{m} \langle \mathbf{X}_\star, \mathbf{A}_i \rangle \mathbf{A}_i \approx \mathbf{X}_\star$

Two-stage approach
(Keshavan, Montanari, Oh 2010)

Objective function:

$$f(\mathbf{U}) = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \langle \mathbf{A}_i, \mathbf{U}\mathbf{U}^\top \rangle \right)^2$$

with $\mathbf{U} \in \mathbb{R}^{d \times r}$

Stage 2: Run gradient descent

- $\mathbf{U}_t = \mathbf{U}_{t-1} - \mu \nabla f\left(\mathbf{U}_{t-1}\right)$ for $t = 1,2,\ldots$
- $\mu > 0$ step size

Our result (S., Zhu 2024)

Let $\mathbf{X}_\star = \mathbf{M}_\star \mathbf{M}_\star^\top$ with $\mathbf{M}_\star \in \mathbb{R}^{d \times r}$. Define

$$\operatorname{dist}\left(\mathbf{U}_t, \mathbf{M}_\star\right) := \min_{\mathbf{R}\text{ rotation}} \|\mathbf{U}_t \mathbf{R} - \mathbf{M}_\star\|_F$$

Assume

- sample size $m \gtrsim r d \kappa^2$
- step size $\mu \leq \dfrac{c}{\kappa \|\mathbf{X}_\star\|}$

Let $\mathbf{U}_0, \mathbf{U}_1, \ldots$ be the iterates from the two-stage algorithm. Then w.h.p. it holds that

$$\operatorname{dist}\left(\mathbf{U}_t, \mathbf{M}_\star\right) \lesssim r \left(1 - c\mu\lambda_{\min}(\mathbf{X}_\star)\right)^t \sqrt{\lambda_{\min}(\mathbf{X}_\star)}$$

Open questions

- Improve step size from $\frac{1}{\kappa\|\mathbf{X}_\star\|}$ to $\frac{1}{\|\mathbf{X}_\star\|}$?!
- Asymmetric ground truth matrix $\mathbf{X}_\star$, convergence from random initialization…?!
- Going beyond Gaussian measurement ensembles?!

Proof ideas

Why is the problem difficult?

Typical proof ingredient:

Decompose gradient into population term and *error* term:

$$\nabla f(\mathbf{Z}) = \mathbb{E}_{(\mathbf{A_i})_{\mathbf{i=1}}^{\mathbf{m}}}\left[\nabla f(\mathbf{Z})\right] + \left(\nabla f(\mathbf{Z}) - \mathbb{E}_{(\mathbf{A_i})_{\mathbf{i=1}}^{\mathbf{m}}}\left[\nabla f(\mathbf{Z})\right]\right)$$

Need to show that second term has small spectral norm.

Key quantity: To control the second term, we need an estimate of the form

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \langle \mathbf{A}_i, \mathbf{\Delta}_t \rangle \mathbf{A}_i - \mathbf{\Delta}_t \right\| \leq c \|\mathbf{\Delta}_t\|$$

where $\mathbf{\Delta}_t = \mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top$

Major difficulty: $\mathbf{\Delta}_t$ (stochastically) depends on $\left(\mathbf{A}_i\right)_{i=1}^{n}$ in a complicated, nonlinear way

Why is the problem difficult?

Previous work: Establish uniform bound of the form w.h.p

$$\sup_{\|\mathbf{Z}\|=1,\ \text{rank}\ \mathbf{Z}=2r} \left\| \frac{1}{m} \sum_{i=1}^{m} \langle \mathbf{A}_i, \mathbf{Z} \rangle \mathbf{A}_i - \mathbf{Z} \right\| \lesssim \sqrt{\frac{r^2 d}{m}}$$

Then this bounds applies in particular for all iterates $\mathbf{\Delta}_0, \mathbf{\Delta}_1, \mathbf{\Delta}_2, \ldots$

Proof techniques: Empirical process theory, Restricted Isometry Property, etc.

Can we improve this bound?

$$\sup_{\|\mathbf{Z}\|=1,\ \text{rank}\,(\mathbf{Z})=2r} \left\| \frac{1}{m} \sum_{i=1}^{m} \langle \mathbf{A}_i, \mathbf{Z} \rangle \mathbf{A}_i - \mathbf{Z} \right\|$$

$$= \sup_{\|\mathbf{Z}\|=1,\ \text{rank}\,(\mathbf{Z})=2r, \|\mathbf{v}\|_2=1} \left| \langle \mathbf{v}\mathbf{v}^\top, \frac{1}{m} \sum_{i=1}^{m} \langle \mathbf{A}_i, \mathbf{Z} \rangle \mathbf{A}_i - \mathbf{Z} \rangle \right|$$

$$\geq \sup_{\|\mathbf{Z}\|=1,\ \text{rank}\,(\mathbf{Z})=2r} \left| \langle \mathbf{e}_1\mathbf{e}_1^\top, \frac{1}{m} \sum_{i=1}^{m} \langle \mathbf{A}_i, \mathbf{Z} \rangle \mathbf{A}_i - \mathbf{Z} \rangle \right|$$

$$\geq \sup_{\|\mathbf{Z}\|=1,\ \text{rank}\,(\mathbf{Z})=2r,\ \mathbf{Z}\mathbf{e}_1=\mathbf{0}} \left| \langle \mathbf{e}_1\mathbf{e}_1^\top, \frac{1}{m} \sum_{i=1}^{m} \langle \mathbf{A}_i, \mathbf{Z} \rangle \mathbf{A}_i \rangle \right|$$

Can we improve this bound?

Set $\mathbf{B} := \frac{1}{m}\sum_{i=1}^{m}\langle \mathbf{e}_1\mathbf{e}_1, \mathbf{A}_i\rangle \mathbf{A}_i$.

We have shown that

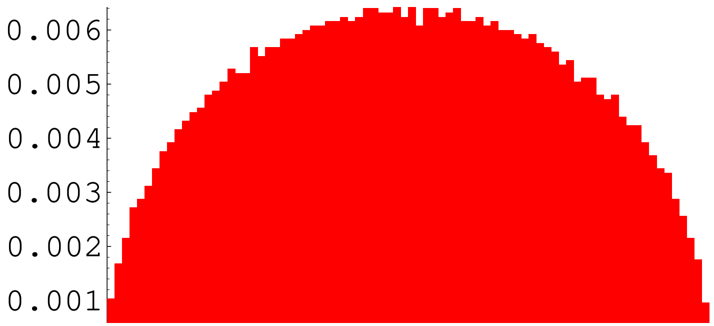
$$\sup_{\|\mathbf{Z}\|=1,\ \text{rank }(\mathbf{Z})=2r} \left\| \frac{1}{m}\sum_{i=1}^{m}\langle \mathbf{A}_i, \mathbf{Z}\rangle \mathbf{A}_i - \mathbf{Z} \right\| \geq \sup_{\|\mathbf{Z}\|=1,\ \text{rank }(\mathbf{Z})=2r, \mathbf{Z}\mathbf{e}_1=\mathbf{0}} \left| \langle \mathbf{Z}, \mathbf{B}\rangle \right|$$

$$= \sum_{i=1}^{2r} \sigma_i\left(\mathbf{B}_{2:d,2:d}\right)$$

Can we improve this bound?

- Conditional on $\left(\langle \mathbf{A}_i, \mathbf{e}_1\mathbf{e}_1^\top\rangle\right)_{i=1}^m$ the matrix $\mathbf{B}_{2:d,2:d}$ has i.i.d. Gaussian entries
- Standard random matrix theory then tells us w.h.p.

$$\sup_{\|\mathbf{Z}\|=1,\ \text{rank } \mathbf{Z}=2r} \left\| \frac{1}{m} \sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{Z} \rangle \mathbf{A}_i - \mathbf{Z} \right\| \gtrsim r\sqrt{\frac{d}{m}} = \sqrt{\frac{r^2 d}{m}}$$

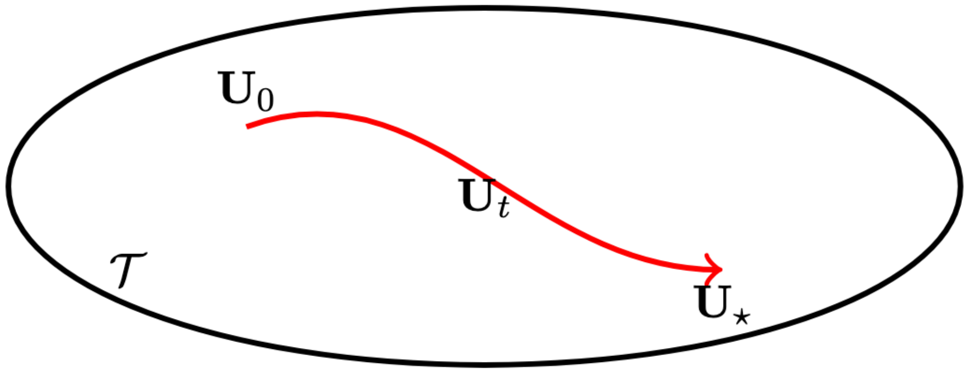


Semicircle law

The previous upper bound is sharp!

All hope is lost?!

- Matrix $\mathbf{Z}$ which we constructed in the proof of lower bounds depends strongly on $\left(\langle \mathbf{A}_i, \mathbf{e}_1\mathbf{e}_1^\top \rangle\right)_{i=1}^m$
- We only need a control **over the trajectory.** Uniform concentration bounds pay the "entropy" cost even for all possible "corners" of the parameter space.
- **Intuition**: The gradient descent iterates $\mathbf{U}_0, \mathbf{U}_1, \mathbf{U}_2, \ldots$ should depend only weakly (in a certain sense) on $\left(\langle \mathbf{A}_i, \mathbf{v}\mathbf{v}^\top \rangle\right)_{i=1}^m$ for all $\mathbf{v}$ with $\|\mathbf{v}\|_2 = 1$



How can we make this intuition rigorous?

Key proof technique: Virtual sequences

Summary

Pure landscape analysis can sometimes lead to overly pessimistic results

$\Longrightarrow$ Gradient descent iterates often enjoy additional randomness which one can exploit via virtual sequences

Related work :

- Leave-one out sequences to analyse GD in phase retrieval (Ma, Wang, Chi, Chen 2020)
- Virtual sequences to establish GD convergence from random initialization (Ma et al.)
- Virtual sequence to establish convergence from random initialization for Alternating Least Squares (Lee, DS 2022)

Main conceptual novelty: Combine virtual sequences with ε-net argument!