Alternating Least Squares for Matrix Sensing

Convergence guarantees from random initialization

KATHOLISCHE UNIVERSITÄT EICHSTÄTT-INGOLSTADT

MIDS

Dominik Stöger

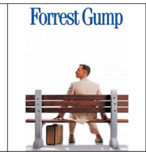
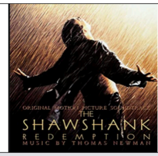
May 15, 2024

Collaborator

Kiryung Lee
(The Ohio State University)

Low-rank matrix recovery problems

Matrix completion:

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

Many other problems can be formulated in this framework:

Blind deconvolution, Phase Retrieval

Problem setting

- Linear observations $y_i = \langle \mathbf{A}_i, \mathbf{X}_\star \rangle :=$ trace $\left(\mathbf{A}_i \mathbf{X}_\star\right)$ for $i = 1,2,\ldots,m$
- $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ known measurement matrices
- low-rank ground truth matrix $\mathbf{X}_\star \in \mathbb{R}^{d\times d}$ with rank $r \ll d$
- **Goal**: estimate $\mathbf{X}_\star$ from samples $y_1, y_2, \ldots, y_m$

Alternating Least Squares (ALS)

- Minimize **non-convex** 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 $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d \times r}$

- **Solution approach**: Alternating least squares

$$\mathbf{U}_{t+1} = \operatorname*{argmin}_{\mathbf{U}} f\left(\mathbf{U}, \mathbf{V}_t\right)$$

$$\mathbf{V}_{t+1} = \operatorname*{argmin}_{\mathbf{V}} f\left(\mathbf{U}_{t+1}, \mathbf{V}\right)$$

Alternating Least Squares (ALS)

☺ Easy to implement

☺ Model-agnostic

☺ Low computational cost (rd optimization variables)

☹ Non-convex objective: Convergence properties unclear and hard to analyze!

When can we establish convergence and recovery guarantees for this non-convex objective function?

Prior work
(e.g., Jain et al. 2013)

Existing convergence and recovery theory requires a good initialization $\mathbf{U}_0$, i.e.,

$$\min_{\mathbf{V}} \|\mathbf{U}_0\mathbf{V}^\top - \mathbf{X}_\star\|_F \ll \sigma_{\min}\left(\mathbf{X}_\star\right)$$

<u>Standard approach for constructing the initialization</u>: Compute top singular vectors of matrix $\frac{1}{m}\sum_{i=1}^{m} y_i \mathbf{A}_i$

<u>Disadvantage</u>: Approach not used by practitioners since it is not model-agnostic!

$\boxed{\text{Practioners often prefer random initialization!}}$

This talk

Can we understand convergence properties of ALS with random initialization?

Major challenge: Many saddle points and local minima exist!

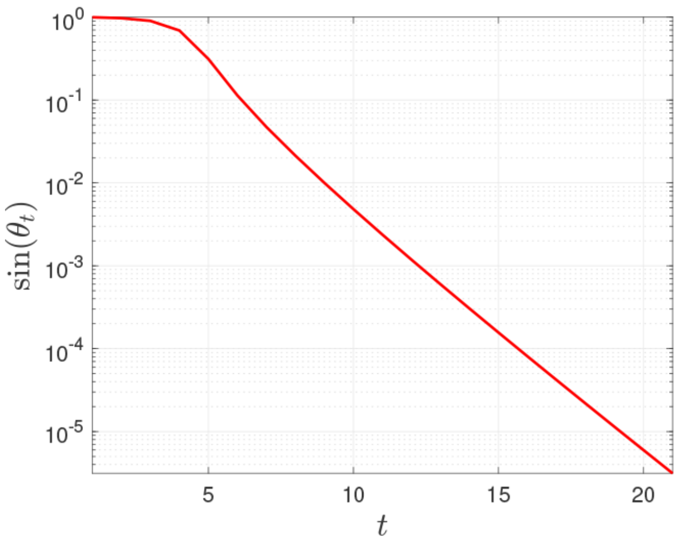
How can we guarantee that ALS avoids those?

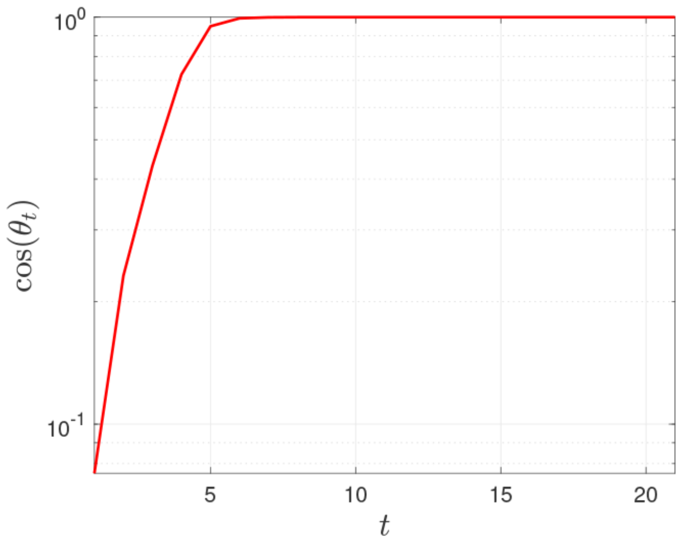
Our setting

- Samples $y_i = \langle \mathbf{X}_\star, \mathbf{A}_i \rangle, i = 1,2,\ldots,m$
- $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ are Gaussian matrices (i.i.d. entries with distribution $\mathcal{N}(0,1)$)
- **rank-one** ground truth matrix $\mathbf{X}_\star = \mathbf{u}_\star \mathbf{v}_\star^\top$
- Without loss of generality $\|\mathbf{u}_\star\|_2 = \|\mathbf{v}_\star\|_2 = 1$
- Initialize $\mathbf{u}_0 \in \mathbb{R}^d$ as a random Gaussian vector

Simulations

$d = 256,\ m = 6d$





- t: number of iterations
- θ_t: angle between $\mathbf{v}_t$ and $\mathbf{v}_\star$

Our result (Lee, S)

(SIAM Journal on Mathematics of Data Science 2023)

Assume for the number of samples that

$$m \gtrsim d \log^4 d.$$

Then it holds with high probability that for every $\varepsilon > 0$ after

$$t \gtrsim \frac{\log d}{\log \log d} + \frac{\log(1/\varepsilon)}{\log \log d}$$

iterations, we have that

$$\max \left\{ \sin \left(\angle(\mathbf{u}_t, \mathbf{u}_\star \right), \sin \left(\angle(\mathbf{v}_t, \mathbf{v}_\star \right) \right\} \leq \varepsilon$$

Insights from our analysis

- Evolution of the ALS iterates can be separated into two phases

$$t \gtrsim \underbrace{\frac{\log d}{\log\log d}}_{\text{Phase 1}} + \underbrace{\frac{\log(1/\varepsilon)}{\log\log d}}_{\text{Phase 2}}$$

- **Phase 1 (Alignment Phase):** $\cos(\theta_t) = \dfrac{|\langle \mathbf{v}_t, \mathbf{v}_\star \rangle|}{\|\mathbf{v}_t\|_2 \|\mathbf{v}_\star\|_2}$ grows geometrically!
- **Phase 2 (Convergence Phase):** $\sin\left(\theta_t\right) = \sqrt{1 - \dfrac{|\langle \mathbf{v}_t, \mathbf{v}_\star \rangle|^2}{\|\mathbf{v}_t\|_2^2 \|\mathbf{v}_\star\|_2^2}}$ converges linearly to 0!

A glimpse of our analysis

A glimpse of the analysis

- Analysis of Phase 2: Established in previous work via RIP (Restricted Isometry Property)
- Major hurdle in our proof: Analysis of Phase 1
- We know that $\mathbf{u}_{t+1}$ satisfies $\nabla_{\mathbf{u}} f\left(\mathbf{u}_{t+1}, \mathbf{v}_t\right) = \mathbf{0}$
- This expression can be rearranged (if $\|\mathbf{v}_t\|_2 = 1$) as follows…

A glimpse of the analysis

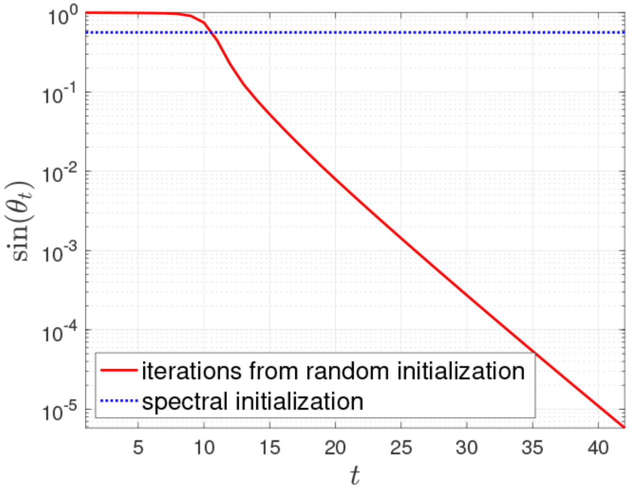
- $$\mathbf{u}_{t+1} = \langle \mathbf{v}_t, \mathbf{v}_\star \rangle \mathbf{u}_\star + \underbrace{\left(\mathbf{M}_t - \frac{1}{m} \sum_{i=1}^{m} \mathbf{A}_i \langle \mathbf{A}_i, \mathbf{M}_t \rangle \right) \mathbf{v}_t}_{=:\mathbf{e}}$$

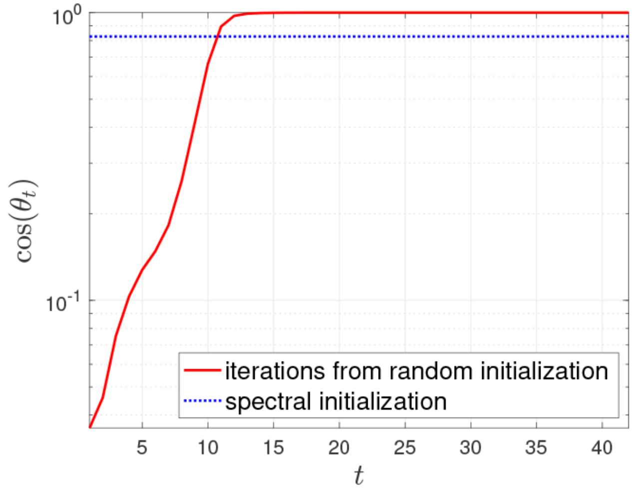
  where $\mathbf{M}_t := \mathbf{u}_{t+1} \mathbf{v}_t^\top - \mathbf{u}_\star \mathbf{v}_\star^\top$

- Second term $\mathbf{e}$ can be interpreted as a perturbation (goes to zero as $m \to \infty$)
- Major difficulty: If $|\langle \mathbf{v}_t, \mathbf{v}_\star \rangle| \ll \|\mathbf{v}_t\|_2$, we will also have $|\langle \mathbf{v}_t, \mathbf{v}_\star \rangle| \ll \|\mathbf{e}\|_2$
- We need to split into $\mathbf{e}$ into part parallel to $\mathbf{u}_\star$ and part perpendicular to $\mathbf{u}_\star$
- Both terms need to be analyzed carefully separately (Key tool: **virtual sequences**)

Open problem: Extension to higher rank case

- $d = 256,\ r = 5,\ m = 2r(2d - r)$





- t number of iterations

- θ_t angle between the subspaces spanned by the columns of $\mathbf{V}_t$ and the left-singular vectors of $\mathbf{X}_\star$

- We again observe that convergence can be separated into two phases!

Outlook

- How to extend our analysis to matrices with rank larger than 1?! (This is open even in a scenario where you take fresh samples in each iteration!)
- How to extend our analysis beyond Gaussian designs?
- What about noisy observations?
- Can we precisely characterize the evolution of $\sin(\theta_t)$ and $\cos(\theta_t)$ depending on the dimension, the number of samples, and noise level? (Lower bounds!)

Our understanding of these non-convex statistical estimation tasks is only in its infancy!

Thank you for your attention!