Convergence of Gradient Descent with Small Initialization for Unregularized Matrix Completion

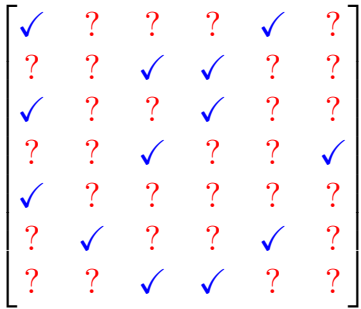
Jianhao Ma
Joint work with Salar Fattahi
University of Michigan

COLT 2024

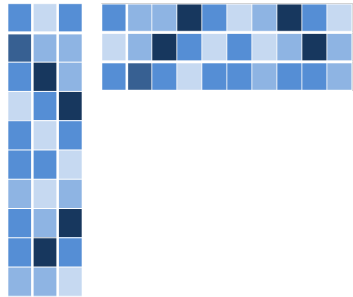
Matrix completion

Goal

Recover a low-rank matrix from a subset of its entries.



(a) matrix completion



(b) matrix factorization[1]

[1] Adapted from https://yuxinchen2020.github.io/slides/NoisyMC_slides.pdf

Mathematical formulation

Unregularized nonconvex optimization

$$\min_{\boldsymbol{U}\in\mathbb{R}^{d\times r'}} f(\boldsymbol{U}) = \frac{1}{4p}\left\|\mathcal{P}_{\Omega}(\boldsymbol{X}^{\star}-\boldsymbol{U}\boldsymbol{U}^{\top})\right\|_{\mathrm{F}}^{2} = \frac{1}{4p}\sum_{(i,j)\in\Omega}\left(X_{i,j}^{\star}-[\boldsymbol{U}\boldsymbol{U}^{\top}]_{i,j}\right)^{2}.$$

Mathematical formulation

Unregularized nonconvex optimization

$$\min_{\boldsymbol{U}\in\mathbb{R}^{d\times r'}} f(\boldsymbol{U}) = \frac{1}{4p}\left\|\mathcal{P}_{\Omega}(\boldsymbol{X}^{\star} - \boldsymbol{U}\boldsymbol{U}^{\top})\right\|_{\mathrm{F}}^{2} = \frac{1}{4p}\sum_{(i,j)\in\Omega}\left(X_{i,j}^{\star} - [\boldsymbol{U}\boldsymbol{U}^{\top}]_{i,j}\right)^{2}.$$

- $\boldsymbol{X}^{\star}$ is a PSD matrix with rank $r \ll d$.

Mathematical formulation

Unregularized nonconvex optimization

$$\min_{\boldsymbol{U}\in\mathbb{R}^{d\times r'}} f(\boldsymbol{U}) = \frac{1}{4p}\left\|\mathcal{P}_{\Omega}(\boldsymbol{X}^{\star}-\boldsymbol{U}\boldsymbol{U}^{\top})\right\|_{\mathrm{F}}^{2} = \frac{1}{4p}\sum_{(i,j)\in\Omega}\left(X_{i,j}^{\star}-[\boldsymbol{U}\boldsymbol{U}^{\top}]_{i,j}\right)^{2}.$$

- $\boldsymbol{X}^{\star}$ is a PSD matrix with rank $r \ll d$.
- **Search rank:** Overparameterization ($r' > r$) v.s. exact-parameterization ($r' = r$).

Mathematical formulation

Unregularized nonconvex optimization

$$\min_{\boldsymbol{U}\in\mathbb{R}^{d\times r'}} f(\boldsymbol{U}) = \frac{1}{4p}\left\|\mathcal{P}_{\Omega}(\boldsymbol{X}^{\star} - \boldsymbol{U}\boldsymbol{U}^{\top})\right\|_{\mathrm{F}}^{2} = \frac{1}{4p}\sum_{(i,j)\in\Omega}\left(X_{i,j}^{\star} - [\boldsymbol{U}\boldsymbol{U}^{\top}]_{i,j}\right)^{2}.$$

- $\boldsymbol{X}^{\star}$ is a PSD matrix with rank $r \ll d$.
- **Search rank:** Overparameterization ($r' > r$) v.s. exact-parameterization ($r' = r$).
- **Bernoulli sampling model:** $\mathbf{1}_{\{(i,j)\in\Omega\}} \overset{i.i.d.}{\sim} \mathrm{Ber}(p)$.

Incoherence

- **Incoherence:** $\boldsymbol{X}^\star = \boldsymbol{V}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$ satisfies $\|\boldsymbol{V}^\star\|_{2,\infty} \leq \sqrt{\frac{\mu r}{d}}$ where $\mu \geq 1$ and $\|\boldsymbol{A}\|_{2,\infty} := \max_i \{\|\boldsymbol{A}_{i,\cdot}\|\}$.

Incoherence

- **Incoherence:** $\boldsymbol{X}^\star = \boldsymbol{V}^\star \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}$ satisfies $\|\boldsymbol{V}^\star\|_{2,\infty} \le \sqrt{\frac{\mu r}{d}}$ where $\mu \ge 1$ and $\|\boldsymbol{A}\|_{2,\infty} := \max_i \{\|\boldsymbol{A}_{i,\cdot}\|\}$.
- Incoherence determines the **difficulty** (sample complexity) of matrix completion.

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}}_{\text{hard } \mu=d} \quad \text{v.s.} \quad \underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}}_{\text{easy } \mu=1}$$

Prior arts

- **Gradient descent (GD):** $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta \nabla f(\boldsymbol{U}_t)$.

Prior arts

- **Gradient descent (GD):** $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta \nabla f(\boldsymbol{U}_t)$.
- GD and its variants converge to the global minimum efficiently for matrix sensing and other similar problems in both exactly and over-parameterized settings [CLC19].

Prior arts

- **Gradient descent (GD):** $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta \nabla f(\boldsymbol{U}_t)$.
- GD and its variants converge to the global minimum efficiently for matrix sensing and other similar problems in both exactly and over-parameterized settings [CLC19].
- Their analysis relies on the restricted isometry property (RIP).

Prior arts

- **Gradient descent (GD):** $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta \nabla f(\boldsymbol{U}_t)$.
- GD and its variants converge to the global minimum efficiently for matrix sensing and other similar problems in both exactly and over-parameterized settings [CLC19].
- Their analysis relies on the restricted isometry property (RIP).
- Matrix completion only satisfies RIP in the incoherent region $\left\{\boldsymbol{U} : \|\boldsymbol{U}\|_{2,\infty} \lesssim \sqrt{\frac{\mu r}{d}}\right\}$.

Prior arts (cont.)

- **Explicit regularization:** add extra regularizer [GJZ17] or projection [ZL16].

Prior arts (cont.)

- **Explicit regularization:** add extra regularizer [GJZ17] or projection [ZL16].
- **Implicit regularization:**
  - **Local result [MWCC18]:** Suppose that ① $r' = r$; ② $\boldsymbol{U}_0$ is close to $\boldsymbol{U}^\star = \boldsymbol{X}^{\star 1/2}$. Then $\boldsymbol{U}_t$ converges to $\boldsymbol{U}^\star$ at a linear rate.
  - **Global result [KC22]:** Suppose that ① $r' = r = 1$; ② $\boldsymbol{U}_0 \approx 0$. Then $\boldsymbol{U}_t$ converges to $\boldsymbol{U}^\star$ at a linear rate.

Prior arts (cont.)

- **Explicit regularization:** add extra regularizer [GJZ17] or projection [ZL16].
- **Implicit regularization:**
  - **Local result [MWCC18]:** Suppose that ① $r' = r$; ② $\boldsymbol{U}_0$ is close to $\boldsymbol{U}^\star = \boldsymbol{X}^{\star 1/2}$. Then $\boldsymbol{U}_t$ converges to $\boldsymbol{U}^\star$ at a linear rate.
  - **Global result [KC22]:** Suppose that ① $r' = r = 1$; ② $\boldsymbol{U}_0 \approx 0$. Then $\boldsymbol{U}_t$ converges to $\boldsymbol{U}^\star$ at a linear rate.

Question: Does GD with small initialization converge to the true solution in the **general rank-r** and **overparameterized** setting?

Overview of our results

Algorithm	Sample complexity	Computational complexity	Global	Exact	Over-param.
[MWCC18]	$dr^3 \log^3(d)$	$\kappa^2 \log\left(\frac{1}{\epsilon}\right)$	✘	✔	✘
[CLL20]	$dr^2 \log(d)$	$\kappa^2 \log\left(\frac{1}{\epsilon}\right)$	✘	✔	✘
[KC22]	$d \log^{22}(d)$	$\log\left(\frac{1}{\epsilon}\right)$	✔	✱	✘
Ours (Thm 1)	$dr^9 \log^8(d)$	$\kappa^4 \log\left(\frac{1}{\epsilon}\right)$	✔	✔	✘
Ours (Thm 2)	$dr^9 \log^2(d) \log^6\left(\frac{1}{\epsilon}\right)$	$\kappa^4 \log^4\left(\frac{1}{\epsilon}\right)$	✔	✔	✔

- ✱ This result only holds for $r' = r = 1$.
- κ is the condition number.

Exactly-parameterized regime

$$\boxed{\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta \nabla f(\boldsymbol{U}_t), \quad \text{where} \quad \boldsymbol{U}_0 = \alpha \boldsymbol{B}} \tag{GD}$$

Theorem 1 (Thorem 3 in [MF24])

Suppose that the sampling rate $p \gtrsim \frac{\kappa^6 \mu^4 r^9 \log^8(d)}{d}$. Consider GD with the step-size $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$ and the initial point $\boldsymbol{U}_0 = \alpha \boldsymbol{B}$, where $\alpha \asymp \frac{\sigma_r^\star}{\kappa^{1.5} d}$. Given any accuracy $\epsilon > 0$, with probability at least $1 - O\left(\frac{1}{d^3}\right)$ and after $T \lesssim \frac{1}{\eta \sigma_r^\star} \log\left(\frac{1}{\epsilon}\right)$ iterations, we have

$$\left\| \boldsymbol{U}_T \boldsymbol{U}_T^\top - \boldsymbol{X}^\star \right\|_{\mathrm{F}} \leq \epsilon.$$

Overparameterized regime

Theorem 2 (Thorem 2 in [MF24])

Suppose that the sampling rate $p \gtrsim \frac{\kappa^6 \mu^4 r^9 \log^6(\frac{1}{\alpha}) \log^2(d)}{d}$. Consider GD with the step-size $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$ and the initial point $\boldsymbol{U}_0 = \alpha \boldsymbol{B}$, where $0 < \alpha \leq \sqrt{\frac{\sigma_1^\star}{d}}$. With probability at least $1 - \frac{2}{d^3}$, after $T \lesssim \frac{1}{\eta \sigma_r^\star} \log\left(\frac{1}{\alpha}\right)$ iterations, we have

$$\left\| \boldsymbol{U}_T \boldsymbol{U}_T^\top - \boldsymbol{X}^\star \right\|_{\mathrm{F}} \lesssim \sqrt{\frac{\sigma_1^\star \kappa^2 \mu r^2}{p}} \alpha.$$

Remark: To achieve ϵ-accuracy, it suffices to set $\alpha \lesssim \sqrt{\frac{p}{\sigma_1^\star \kappa^2 \mu r^2}} \epsilon$.

Proof outline

- **Dynamic signal-residual decomposition (similar to [LMZ18][2]):** Project each iterate $\boldsymbol{U}_t$ onto a dynamic rank-r subspace $\boldsymbol{V}_t$ and its complement:

$$\boldsymbol{U}_t = \underbrace{\mathcal{P}_{\boldsymbol{V}_t}\boldsymbol{U}_t}_{\text{rank-}r} + \underbrace{\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{U}_t}_{\approx 0}.$$

[2] COLT 2018 best paper award.

Proof outline

- **Dynamic signal-residual decomposition (similar to [LMZ18][2]):** Project each iterate $\boldsymbol{U}_t$ onto a dynamic rank-r subspace $\boldsymbol{V}_t$ and its complement:

$$\boldsymbol{U}_t = \underbrace{\mathcal{P}_{\boldsymbol{V}_t}\boldsymbol{U}_t}_{\text{rank-}r} + \underbrace{\mathcal{P}^{\perp}_{\boldsymbol{V}_t}\boldsymbol{U}_t}_{\approx 0}.$$

- **Weakly-coupled leave-one-out (WC-LOO) analysis:** Show that the implicit regularization effect of the low-rank part $\mathcal{P}_{\boldsymbol{V}_t}\boldsymbol{U}_t$:

$$\|\mathcal{P}_{\boldsymbol{V}_t}\boldsymbol{U}_t\|_{2,\infty} \lesssim \sqrt{\frac{\sigma_1^\star \mu r}{d}}, \quad \forall t = 0, \cdots, T.$$

[2] COLT 2018 best paper award.

Classic leave-one-out analysis

- **Idea:** decouple the statistical correlation between algorithm and train set.

Classic leave-one-out analysis

- **Idea:** decouple the statistical correlation between algorithm and train set.
- **Leave-one-out sequences:** For each $1 \leq l \leq d$, $\left\{\boldsymbol{U}_t^{(l)}\right\}$ is generated by

$$\boldsymbol{U}_{t+1}^{(l)} = \left(I - \eta \frac{1}{p} \mathcal{P}_{\Omega^{(l)}} \left(\boldsymbol{U}_t^{(l)} \boldsymbol{U}_t^{(l)\top} - \boldsymbol{X}^\star\right)\right) \boldsymbol{U}_t^{(l)}.$$

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}}_{\text{sampling matrix } \boldsymbol{\Omega}} \quad \Longrightarrow \quad \underbrace{\begin{bmatrix} 1 & p & 0 & 1 & 0 \\ p & p & p & p & p \\ 0 & p & 0 & 1 & 0 \\ 1 & p & 1 & 0 & 0 \\ 0 & p & 0 & 0 & 1 \end{bmatrix}}_{\text{leave-one-out version } \boldsymbol{\Omega}^{(2)}}$$

Classic leave-one-out analysis

- **Idea:** decouple the statistical correlation between algorithm and train set.
- **Leave-one-out sequences:** For each $1 \le l \le d$, $\left\{ \boldsymbol{U}_t^{(l)} \right\}$ is generated by

$$\boldsymbol{U}_{t+1}^{(l)} = \left(I - \eta \frac{1}{p} \mathcal{P}_{\Omega^{(l)}} \left(\boldsymbol{U}_t^{(l)} \boldsymbol{U}_t^{(l)\top} - \boldsymbol{X}^\star \right) \right) \boldsymbol{U}_t^{(l)}.$$

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}}_{\text{sampling matrix } \boldsymbol{\Omega}} \quad \Longrightarrow \quad \underbrace{\begin{bmatrix} 1 & p & 0 & 1 & 0 \\ p & p & p & p & p \\ 0 & p & 0 & 1 & 0 \\ 1 & p & 1 & 0 & 0 \\ 0 & p & 0 & 0 & 1 \end{bmatrix}}_{\text{leave-one-out version } \boldsymbol{\Omega}^{(2)}}$$

Key observation: The l-th row of $\boldsymbol{U}_t^{(l)}$ is *deterministic* for all t!

Control incoherence

$$\|\boldsymbol{U}_t\|_{2,\infty} \le \underbrace{\max_{1\le l\le d}\left\{\left\|\left(\boldsymbol{U}^\star - \boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right)_{l,\cdot}\right\|\right\}}_{\text{leave-one-out error}} + \underbrace{\max_{1\le l\le d}\left\{\left\|\boldsymbol{U}_t - \boldsymbol{U}_t^{(l)}\boldsymbol{H}_t^{(l)}\right\|_{\mathrm{F}}\right\}}_{\text{proximal error}} + \sqrt{\frac{\sigma_1^\star \mu r}{d}}.$$

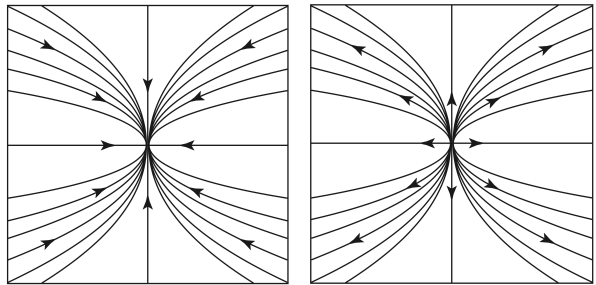
- **Leave-one-out error:** By our construction of $\Omega^{(l)}$, the leave-one-out error is *deterministic*!

Control incoherence

$$\|\boldsymbol{U}_t\|_{2,\infty} \le \underbrace{\max_{1\le l\le d}\left\{\left\|\left(\boldsymbol{U}^\star - \boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right)_{l,\cdot}\right\|\right\}}_{\text{leave-one-out error}} + \underbrace{\max_{1\le l\le d}\left\{\left\|\boldsymbol{U}_t - \boldsymbol{U}_t^{(l)}\boldsymbol{H}_t^{(l)}\right\|_{\mathrm{F}}\right\}}_{\text{proximal error}} + \sqrt{\frac{\sigma_1^\star \mu r}{d}}.$$

- **Leave-one-out error:** By our construction of $\Omega^{(l)}$, the leave-one-out error is *deterministic*!
- **Proximal error:** Suppose that $\boldsymbol{U}_t \approx \boldsymbol{U}^\star$. Then, due to the local restricted *strong convexity* of the loss function, the proximal error decreases geometrically fast.

Local minimum v.s. saddle point

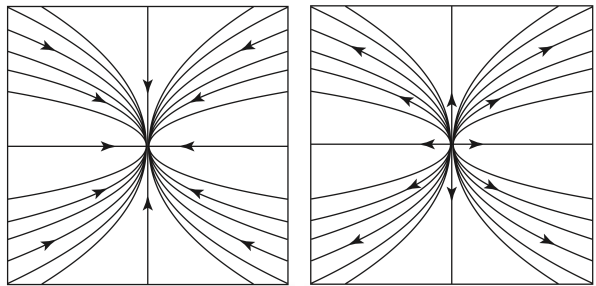


(a) $\boldsymbol{U}^\star$ is stable (b) $\mathbf{0}$ is unstable

Figure: Gradient flow around $\boldsymbol{U}^*$ and $\mathbf{0}$.

Local minimum v.s. saddle point

- $\boldsymbol{U}^\star$ **is a local minimum:** The distance between $\boldsymbol{U}_t$ and $\boldsymbol{U}_t^{(l)}$ decreases linearly.

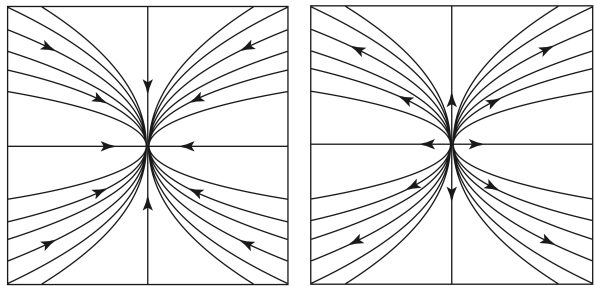


(a) $\boldsymbol{U}^\star$ is stable (b) $\mathbf{0}$ is unstable

Figure: Gradient flow around $\boldsymbol{U}^*$ and $\mathbf{0}$.

Local minimum v.s. saddle point

- $\boldsymbol{U}^\star$ **is a local minimum:** The distance between $\boldsymbol{U}_t$ and $\boldsymbol{U}_t^{(l)}$ decreases linearly.
- $\mathbf{0}$ **is a saddle point:** Under small initialization, the distance between $\boldsymbol{U}_t$ and $\boldsymbol{U}_t^{(l)}$ increases exponentially fast.



(a) $\boldsymbol{U}^\star$ is stable (b) $\mathbf{0}$ is unstable

Figure: Gradient flow around $\boldsymbol{U}^*$ and $\mathbf{0}$.

Weakly-coupled leave-one-out sequences

original sequence

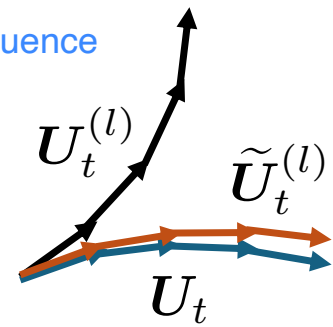
classic LOO sequence

$$\boldsymbol{U}_{t+1} = \left(I - \eta\frac{1}{p}\mathcal{P}_{\Omega}\left(\boldsymbol{V}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{\top} - \boldsymbol{X}^{\star}\right)\right)\boldsymbol{U}_t$$

$$\boldsymbol{U}_{t+1}^{(l)} = \left(I - \eta\frac{1}{p}\mathcal{P}_{\Omega^{(l)}}\left(\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t^{(l)}\boldsymbol{V}_t^{(l)\top} - \boldsymbol{X}^{\star}\right)\right)\boldsymbol{U}_t^{(l)}$$

$$\widetilde{\boldsymbol{U}}_{t+1}^{(l)} = \left(I - \eta\frac{1}{p}\mathcal{P}_{\Omega^{(l)}}\left(\widetilde{\boldsymbol{V}}_t^{(l)}\boldsymbol{\Sigma}_t\widetilde{\boldsymbol{V}}_t^{(l)\top} - \boldsymbol{X}^{\star}\right)\right)\widetilde{\boldsymbol{U}}_t^{(l)}$$

WC-LOO sequence



Key observation: $\widetilde{\boldsymbol{U}}_t^{(l)}$ is much closer to $\boldsymbol{U}_t$ than $\boldsymbol{U}_t^{(l)}$ since we replace $\boldsymbol{\Sigma}_t^{(l)}$ with $\boldsymbol{\Sigma}_t$.

[2] Here we define $\boldsymbol{\Sigma}_t = \boldsymbol{V}_t^{\top}\boldsymbol{U}_t\boldsymbol{U}_t^{\top}\boldsymbol{V}_t \in \mathbb{R}^{r\times r}$.

Control incoherence via WC-LOO analysis

$$\|\boldsymbol{U}_t\|_{2,\infty} \le \underbrace{\max_{1\le l\le d}\left\{\left\|\left(\boldsymbol{U}^\star - \widetilde{\boldsymbol{U}}_t^{(l)}\widetilde{\boldsymbol{R}}_t^{(l)}\right)_{l,\cdot}\right\|\right\}}_{\text{WC-LOO error}} + \underbrace{\max_{1\le l\le d}\left\{\left\|\boldsymbol{U}_t - \widetilde{\boldsymbol{U}}_t^{(l)}\widetilde{\boldsymbol{H}}_t^{(l)}\right\|_{\mathrm{F}}\right\}}_{\text{WC-proximal error}} + \sqrt{\frac{\sigma_1^\star \mu r}{d}}$$

- **Good news ☺:** The WC-proximal error grows mildly for all $0 \le t \le T$.

Control incoherence via WC-LOO analysis

$$\|\boldsymbol{U}_t\|_{2,\infty} \le \underbrace{\max_{1\le l\le d}\left\{\left\|\left(\boldsymbol{U}^\star - \widetilde{\boldsymbol{U}}_t^{(l)}\widetilde{\boldsymbol{R}}_t^{(l)}\right)_{l,\cdot}\right\|\right\}}_{\text{WC-LOO error}} + \underbrace{\max_{1\le l\le d}\left\{\left\|\boldsymbol{U}_t - \widetilde{\boldsymbol{U}}_t^{(l)}\widetilde{\boldsymbol{H}}_t^{(l)}\right\|_{\mathrm{F}}\right\}}_{\text{WC-proximal error}} + \sqrt{\frac{\sigma_1^\star \mu r}{d}}$$

- **Good news ☺:** The WC-proximal error grows mildly for all $0 \le t \le T$.
- **Bad news ☹:** The WC-LOO error is no longer deterministic since we use $\boldsymbol{\Sigma}_t$.

Control incoherence via WC-LOO analysis

$$\|\boldsymbol{U}_t\|_{2,\infty} \le \underbrace{\max_{1\le l\le d}\left\{\left\|\left(\boldsymbol{U}^\star - \widetilde{\boldsymbol{U}}_t^{(l)}\widetilde{\boldsymbol{R}}_t^{(l)}\right)_{l,\cdot}\right\|\right\}}_{\text{WC-LOO error}} + \underbrace{\max_{1\le l\le d}\left\{\left\|\boldsymbol{U}_t - \widetilde{\boldsymbol{U}}_t^{(l)}\widetilde{\boldsymbol{H}}_t^{(l)}\right\|_{\mathrm{F}}\right\}}_{\text{WC-proximal error}} + \sqrt{\frac{\sigma_1^\star \mu r}{d}}$$

- **Good news ☺:** The WC-proximal error grows mildly for all $0 \le t \le T$.
- **Bad news ☹:** The WC-LOO error is no longer deterministic since we use $\boldsymbol{\Sigma}_t$.
- **Good news ☺:** $\boldsymbol{\Sigma}_t$ is an $r \times r$ small matrix. Hence, we can use a dynamic covering trick to bound the deviation uniformly. The price is to increase the sample complexity by a poly(r) factor.

Take-home message

- GD with small initialization converges to the ground truth without any explicit regularization.
- It achieves nearly optimal sample complexity and computational complexity.
- We propose a powerful tool called weakly-coupled leave-one-out analysis, which can decouple the statistical correlation between algorithm and training dataset in the nonconvex regime.

References

Yuejie Chi, Yue M Lu, and Yuxin Chen.
Nonconvex optimization meets low-rank matrix factorization: An overview.
IEEE Transactions on Signal Processing, 67(20):5239–5269, 2019.

Ji Chen, Dekai Liu, and Xiaodong Li.
Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization.
IEEE Transactions on Information Theory, 66(9):5806–5841, 2020.

Rong Ge, Chi Jin, and Yi Zheng.
No spurious local minima in nonconvex low rank problems: A unified geometric analysis.
In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

Daesung Kim and Hye Won Chung.
Rank-1 matrix completion with gradient descent and small random initialization.
arXiv preprint arXiv:2212.09396, 2022.