Understanding Incremental Learning of Gradient Descent: A Fine-grained Analysis of Matrix Sensing

Jikai Jin [1] **Zhiyuan Li** [2] **Kaifeng Lyu** [3] **Simon S. Du** [4] **Jason D. Lee** [5]

Abstract

It is believed that Gradient Descent (GD) induces an implicit bias towards good generalization in training machine learning models. This paper provides a fine-grained analysis of the dynamics of GD for the matrix sensing problem, whose goal is to recover a low-rank ground-truth matrix from near-isotropic linear measurements. It is shown that GD with small initialization behaves similarly to the greedy low-rank learning heuristics (Li et al., 2020) and follows an incremental learning procedure (Gissin et al., 2019): GD sequentially learns solutions with increasing ranks until it recovers the ground truth matrix. Compared to existing works which only analyze the first learning phase for rank-1 solutions, our result provides characterizations for the whole learning process. Moreover, besides the over-parameterized regime that many prior works focused on, our analysis of the incremental learning procedure also applies to the *under-parameterized* regime. Finally, we conduct numerical experiments to confirm our theoretical findings.

1. Introduction

Understanding the optimization and generalization properties of optimization algorithms is one of the central topics in deep learning theory (Zhang et al., 2017; Sun, 2019). It has long been a mystery why simple algorithms such as Gradient Descent (GD) or Stochastic Gradient Descent (SGD) can find global minima even for highly non-convex functions (Du et al., 2019), and why the global minima being found can generalize well (Hardt et al., 2016).

One influential line of works provides theoretical analysis of the *implicit bias* of GD/SGD. These results typically exhibit theoretical settings where the low-loss solutions found by GD/SGD attain certain optimality conditions of a particular generalization metric, *e.g.*, the parameter norm (or the classifier margin) (Soudry et al., 2018; Gunasekar et al., 2018; Nacson et al., 2019; Lyu & Li, 2020; Ji & Telgarsky, 2020), the sharpness of local loss landscape (Blanc et al., 2020; Damian et al., 2021; Li et al., 2022a; Lyu et al., 2022).

Among these works, a line of works seek to characterize the implicit bias even when the training is away from convergence. Kalimeris et al. (2019) empirically observed that SGD learns model from simple ones, such as linear classifiers, to more complex ones. As a result, SGD always tries to fit the training data with minimal model complexity. This behavior, usually referred to as the *simplicity bias* or the *incremental learning* behavior of GD/SGD, can be a hidden mechanism of deep learning that prevents highly over-parameterized models from overfitting. In theory, Hu et al. (2020); Lyu et al. (2021); Frei et al. (2021) established that GD on two-layer nets learns linear classifiers first.

The goal of this paper is to demonstrate this simplicity bias/incremental learning in the *matrix sensing* problem, a non-convex optimization problem that arises in a wide range of real-world applications, *e.g.*, image reconstruction (Zhao et al., 2010; Peng et al., 2014), object detection (Shen & Wu, 2012; Zou et al., 2013) and array processing systems (Kalogerias & Petropulu, 2013). Moreover, this problem can serve as a standard test-bed of the implicit bias of GD/SGD in deep learning theory, since it retains many of the key phenomena in deep learning while being simpler to analyze.

Formally, the matrix sensing problem asks for recovering a ground-truth matrix $\boldsymbol{Z}^* \in \mathbb{R}^{d\times d}$ given m observations $y_1, \ldots, y_m$. Each observation y_i here is resulted from a linear measurement $y_i = \langle \boldsymbol{A}_i, \boldsymbol{Z}^* \rangle$, where $\{\boldsymbol{A}_i\}_{1\le i\le m}$ is a collection of symmetric measurement matrices. In this paper, we focus on the case where $\boldsymbol{Z}^*$ is symmetric, positive semi-definite (PSD) and low-rank, *i.e.*, $\boldsymbol{Z}^* \succeq \boldsymbol{0}$ and $\mathrm{rank}(\boldsymbol{Z}^*) = r_* \ll d$.

An intriguing approach to solve this problem is to use the

[1]School of Mathematical Sciences, Peking University [2]Department of Computer Science, Stanford University [3]Department of Computer Science, Princeton University [4]Paul G. Allen School of Computer Science and Engineering, University of Washington [5]Department of Electrical and Computer Engineering, Princeton University. Correspondence to: Jikai Jin <jinjikai7@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Burer-Monteiro type decomposition $\boldsymbol{Z}^* = \boldsymbol{U}\boldsymbol{U}^\top$ with $\boldsymbol{U} \in \mathbb{R}^{d\times \hat{r}}$, and minimize the squared loss with GD:

$$\min_{\boldsymbol{U}\in\mathbb{R}^{d\times\hat{r}}} \quad f(\boldsymbol{U}) := \frac{1}{4m}\sum_{i=1}^{m}\left(y_i - \left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top\right\rangle\right)^2. \quad (1)$$

In the ideal case, the number of columns of $\boldsymbol{U}$, denoted as $\hat{r}$ above, should be set to r_*. However, r_* may not be known in advance. This leads to two training regimes that are more likely to happen: the *under-parameterized* regime where $\hat{r} \le r_*$, and the *over-parameterized* regime where $\hat{r} > r_*$.

The over-parameterized regime may lead to overfitting at first glance, but surprisingly, with small initialization, GD induces a good implicit bias towards solutions with the exact or approximate recovery of the ground truth. It was first conjectured in Gunasekar et al. (2017) that GD with small initialization finds the matrix with the minimum nuclear norm. Gunasekar et al. (2017) also proved this conjecture for a special case where all measurements are commutable. However, a series of works point out that this nuclear norm minimization view cannot capture the incremental learning behavior of GD, which, in the context of matrix sensing, refers to the phenomenon that GD tends to learn solutions with rank gradually increasing with training steps. Arora et al. (2019) exhibited this phenomenon when there is only one observation ($m = 1$). Gissin et al. (2019); Jiang et al. (2022) studied the full-observation case, where every entry of the ground truth is measured independently $f(\boldsymbol{U}) = \frac{1}{4d^2}\|\boldsymbol{Z}^* - \boldsymbol{U}\boldsymbol{U}^\top\|_{\mathrm{F}}^2$, and GD is shown to sequentially recover singular components of the ground truth from the largest singular value to the smallest one. Li et al. (2020) provided theoretical evidence that the incremental learning behavior generally occurs for matrix sensing. They specifically provided a counterexample for Gunasekar et al. (2017)'s conjecture, where GD converges to a rank-1 solution with a very large nuclear norm. Razin & Cohen (2020) also pointed out a case where GD drives the norm to infinity while keeping the rank to be approximately 1.

Despite these progresses, theoretical understanding of the simplicity bias of GD remains limited. In fact, a vast majority of existing analyses can only show that GD is initially biased towards a rank-1 solution and cannot be generalized to higher ranks, unless additional assumptions on GD dynamics are made (Li et al., 2020, Appendix H), (Belabbas, 2020; Jacot et al., 2021; Razin et al., 2021; 2022). Recently Li et al. (2022b) shows that the implicit bias of Gunasekar et al. (2017) essentially relies on rewriting gradient flow in the space of $\boldsymbol{U}$ as continuous mirror descent in the space of $\boldsymbol{U}\boldsymbol{U}^\top$, which only works a special type of reparametrized model, named "commuting parametrization". However, Li et al. (2022b) also shows that matrix sensing with general (non-commutable) measurements does not fall into this type.

1.1. Our Contributions

In this paper, we take a step towards understanding the generalization of GD with small initialization by firmly demonstrating the simplicity bias/incremental learning behavior in the matrix sensing setting, assuming the Restricted Isometry Property (RIP). Our main result is informally stated below. See Theorem 4.1 for the formal version.

Definition 1.1 (Best Rank-s Solution)**.** We define the *best rank-s solution* as the unique global minimizer $\boldsymbol{Z}_s^*$ of the following constrained optimization problem:

$$\begin{aligned}\min_{\boldsymbol{Z}\in\mathbb{R}^{d\times d}} \quad & \frac{1}{4m}\sum_{i=1}^{m}\left(y_i - \langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle\right)^2 \\ s.t. \quad & \boldsymbol{Z} \succeq \boldsymbol{0}, \quad \mathrm{rank}(\boldsymbol{Z}) \le s.\end{aligned} \quad (2)$$

Theorem 1.2 (Informal version of Theorem 4.1)**.** *Consider the matrix sensing problem (1) with rank-r_* ground-truth matrix $\boldsymbol{Z}^*$ and measurements $\{\boldsymbol{A}_i\}_{i=1}^m$. Assume that the measurements satisfy the RIP condition (Definition 3.2). With small learning rate $\mu > 0$ and small initialization $\boldsymbol{U}_{\alpha,0} = \alpha\boldsymbol{U} \in \mathbb{R}^{d\times\hat{r}}$, the trajectory of $\boldsymbol{U}_{\alpha,t}\boldsymbol{U}_{\alpha,t}^\top$ during GD training enters an $o(1)$-neighborhood of each of the best rank-s solutions in the order of $s = 1, 2, \ldots, \hat{r} \wedge r_*$ when $\alpha \to 0$. Moreover, when $\hat{r} \le r^*$, we have $\lim_{t\to\infty} \boldsymbol{U}_{\alpha,t}\boldsymbol{U}_{\alpha,t}^\top = \boldsymbol{Z}_{\hat{r}}^*$.*

It is shown in Li et al. (2018); Stöger & Soltanolkotabi (2021) that GD exactly recovers the ground truth under the RIP condition, but our theorem goes beyond them in a number of ways. First, in the over-parameterized regime (i.e., $\hat{r} > r_*$), it implies that GD performs *incremental learning*: learning solutions with increasing ranks until it finds the ground truth. Second, this result also shows that in the under-parameterized regime (i.e., $\hat{r} \le r_*$), GD exhibits the same implicit bias, but finally it converges to the best low-rank solution of the matrix sensing loss. By contrast, to the best of our knowledge, only the over-parameterized setting is analyzed in existing literature.

Theorem 1.2 can also be considered as a generalization of previous results in Gissin et al. (2019); Jiang et al. (2022) which show that $\boldsymbol{U}_{\alpha,t}\boldsymbol{U}_{\alpha,t}^\top$ passes by the best low-rank solutions one by one in the full observation case of matrix sensing $f(\boldsymbol{U}) = \frac{1}{4d^2}\|\boldsymbol{Z}^* - \boldsymbol{U}\boldsymbol{U}^\top\|_{\mathrm{F}}^2$. However, our setting has two major challenges which significantly complicate our analysis. First, since our setting only gives partial measurements, the decomposition of signal and error terms in Gissin et al. (2019); Jiang et al. (2022) cannot be applied. Instead, we adopt a different approach which is motivated by Stöger & Soltanolkotabi (2021). Second, it is well-known that the optimal rank-s solution of matrix factorization is $\boldsymbol{X}_s$ (defined in Section 3), but little is known for $\boldsymbol{Z}_s^*$. In Section 4.1 we analyze the landscape of (2), establishing the uniqueness of $\boldsymbol{Z}_s^*$ and local landscape properties under the

RIP condition. We find that when $\boldsymbol{U}_{\alpha,t}\boldsymbol{U}_{\alpha,t}^\top \approx \boldsymbol{Z}_s^*$, GD follows an approximate low-rank trajectory, so that it behaves similarly to GD in the under-parameterized regime. Using our landscape results, we can finally prove Theorem 1.2.

Organization. We review additional related works in Section 2. In Section 3, we provide an overview of necessary background and notations. We then present our main results in Section 4 with proof sketch where we also state some key lemmas that are used in the proof, including Lemma 4.3 and some landscape results. In Section 5 we present a trajectory analysis of GD and prove Lemma 4.3. Experimental results are presented in Section 6 which verify our theoretical findings. Finally, in Section 7, we summarize our main contributions and discuss some promising future directions. Complete proofs of all results are given in the Appendix.

2. Related work

Low-rank matrix recovery. The goal of low-rank matrix recovery is to recover an unknown low-rank matrix from a number of (possibly noisy) measurements. Examples include matrix sensing (Recht et al., 2010), matrix completion (Candès & Recht, 2009; Candes & Plan, 2010) and robust PCA (Xu et al., 2010; Candès et al., 2011). Fornasier et al. (2011); Ngo & Saad (2012); Wei et al. (2016); Tong et al. (2021) study efficient optimization algorithms with convergence guarantees. Interested readers can refer to Davenport & Romberg (2016) for an overview of this topic.

Simplicity bias/incremental learning of GD. Besides the works mentioned in the introduction, there are many other works studying the simplicity bias/incremental learning of GD on tensor factorization (Razin et al., 2021; 2022), deep linear networks (Gidel et al., 2019), two-layer nets with orthogonal inputs (Boursier et al., 2022).

Landscape analysis of non-convex low-rank problems. The strict saddle property (Ge et al., 2016; 2015; Lee et al., 2016) was established for non-convex low-rank problems in a unified framework by Ge et al. (2017). Tu et al. (2016) proved a local PL property for matrix sensing with exact parameterization (i.e. the rank of parameterization and ground-truth matrix are the same). The optimization geometry of general objective function with Burer-Monteiro type factorization is studied in Zhu et al. (2018); Li et al. (2019); Zhu et al. (2021). We provide a comprehensive analysis in this regime for matrix factorization as well as matrix sensing that improves over their results.

3. Preliminaries

In this section, we first list the notations used in this paper, and then provide details of our theoretical setup and necessary preliminary results.

3.1. Notations

We write $\min\{a,b\}$ as $a \wedge b$ for short. For any matrix $\boldsymbol{A}$, we use $\|\boldsymbol{A}\|_{\mathrm{F}}$ to denote the Frobenius norm of $\boldsymbol{A}$, use $\|\boldsymbol{A}\|$ to denote the spectral norm $\|\boldsymbol{A}\|_2$, and use $\sigma_{\min}(\boldsymbol{A})$ to denote the smallest singular value of $\boldsymbol{A}$. We use the following notation for Singular Value Decomposition (SVD):

Definition 3.1 (Singular Value Decomposition)**.** For any matrix $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$ of rank r, we use $\boldsymbol{A} = \boldsymbol{V}_{\boldsymbol{A}}\boldsymbol{\Sigma}_{\boldsymbol{A}}\boldsymbol{W}_{\boldsymbol{A}}^\top$ to denote a Singular Value Decomposition (SVD) of $\boldsymbol{A}$, where $\boldsymbol{V}_{\boldsymbol{A}} \in \mathbb{R}^{d_1\times r}, \boldsymbol{W}_{\boldsymbol{A}} \in \mathbb{R}^{d_2\times r}$ satisfy $\boldsymbol{V}_{\boldsymbol{A}}^\top\boldsymbol{V}_{\boldsymbol{A}} = \boldsymbol{I}, \boldsymbol{W}_{\boldsymbol{A}}^\top\boldsymbol{W}_{\boldsymbol{A}} = \boldsymbol{I}$, and $\boldsymbol{\Sigma}_{\boldsymbol{A}} \in \mathbb{R}^{r\times r}$ is diagonal.

For the matrix sensing problem (1), we write the ground-truth matrix as $\boldsymbol{Z}^* = \boldsymbol{X}\boldsymbol{X}^\top$ for some $\boldsymbol{X} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{r_*}] \in \mathbb{R}^{d\times r_*}$ with orthogonal columns from an orthogonal basis $\{\boldsymbol{v}_i : i \in [d]\}$ of $\mathbb{R}^d$. We denote the singular values of $\boldsymbol{X}$ as $\sigma_1, \sigma_2, \ldots, \sigma_{r_*}$, then the singular values of $\boldsymbol{Z}^*$ are $\sigma_1^2, \sigma_2^2, \ldots, \sigma_{r_*}^2$. We set $\sigma_{r_*+1} := 0$ for convenience. For simplicity, we only consider the case where $\boldsymbol{Z}^*$ has distinct singular values, i.e., $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_{r_*}^2 > 0$. We use $\kappa_* := \frac{\sigma_1^2}{\min_{1\le s\le r_*}\{\sigma_s^2-\sigma_{s+1}^2\}}$ to quantify the degeneracy of the singular values of $\boldsymbol{Z}^*$. We also use the notation $\boldsymbol{X}_s = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_s]$ for the matrix consisting of the first s columns of $\boldsymbol{X}$ and $\boldsymbol{X}_s^\perp = [\boldsymbol{v}_{s+1}, \cdots, \boldsymbol{v}_d]$. Following Definition 3.1, we let $\boldsymbol{V}_{\boldsymbol{X}_s^\perp} = \left[\frac{\boldsymbol{v}_{s+1}}{\|\boldsymbol{v}_{s+1}\|}, \cdots, \frac{\boldsymbol{v}_d}{\|\boldsymbol{v}_d\|}\right]$. Note that the best rank-s solution $\boldsymbol{Z}_s^*$ (Definition 1.1) does not equal $\boldsymbol{X}_s\boldsymbol{X}_s^\top$ in general.

We write the results of the measurements $\{\boldsymbol{A}_i\}_{i=1}^m$ as a linear mapping $\mathcal{A} : \mathbb{R}^{d\times d} \mapsto \mathbb{R}^m$, where $[\mathcal{A}(\boldsymbol{Z})]_i = \frac{1}{\sqrt{m}}\langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle$ for all $1 \le i \le m$. We use $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^{d\times d}, \mathcal{A}^*(\boldsymbol{w}) = \frac{1}{\sqrt{m}}\sum_{i=1}^m w_i\boldsymbol{A}_i$ to denote the adjoint operator of $\mathcal{A}$. Our loss function (1) can then be written as $f(\boldsymbol{U}) = \frac{1}{4}\left\|\mathcal{A}\left(\boldsymbol{Z}^* - \boldsymbol{U}\boldsymbol{U}^\top\right)\right\|_2^2$. The gradient is given by $\nabla f(\boldsymbol{U}) = \mathcal{A}^*\left(\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top)\right)\boldsymbol{U} = \mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top\right)\boldsymbol{U}$.

In this paper, we consider GD with learning rate $\mu > 0$ starting from $\boldsymbol{U}_0$. The update rule is

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \mu\nabla f\left(\boldsymbol{U}_t\right) = (\boldsymbol{I} + \mu\boldsymbol{M}_t)\boldsymbol{U}_t, \qquad (3)$$

where $\boldsymbol{M}_t := \mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\right)$. We specifically focus on GD with *small initialization*: letting $\boldsymbol{U}_0 = \alpha\bar{\boldsymbol{U}}$ for some matrix $\bar{\boldsymbol{U}} \in \mathbb{R}^{d\times\hat{r}}$ with $\|\bar{\boldsymbol{U}}\| = 1$, we are interested in the trajectory of GD when $\alpha \to 0$. Sometimes we write $\boldsymbol{U}_t$ as $\boldsymbol{U}_{\alpha,t}$ to highlight the dependence of the trajectory on α.

3.2. Assumptions

For our theoretical analysis of the matrix sensing problem, we make the following standard assumption in the matrix sensing literature:

Definition 3.2 (Restricted Isometry Property)**.** We say that a measurement operator $\mathcal{A}$ satisfies the (δ, r)-RIP condition if $(1-\delta)\|\boldsymbol{Z}\|_{\mathrm{F}}^2 \le \|\mathcal{A}(\boldsymbol{Z})\|_2^2 \le (1+\delta)\|\boldsymbol{Z}\|_{\mathrm{F}}^2$ for all matrices $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$ with $\mathrm{rank}(\boldsymbol{Z}) \le r$.

Assumption 3.3. The measurement operator $\mathcal{A}$ satisfies the $(2r_*+1, \delta)$-RIP property, where $r_* = \mathrm{rank}(\boldsymbol{Z}^*)$ and $\delta \le 10^{-12}\kappa_*^{-4.5} r_*^{-1}$.

The RIP condition is the key to ensure the ground truth to be recoverable with partial observations. An important consequence of RIP is that it guarantees $\mathcal{A}^*\mathcal{A}(\boldsymbol{Z}) = \frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle \boldsymbol{A}_i \approx \boldsymbol{Z}$ when $\boldsymbol{Z}$ is low-rank. This is made rigorous in the following proposition.

Proposition 3.4. *(Stöger & Soltanolkotabi, 2021, Lemma 7.3) Suppose that $\mathcal{A}$ satisfies (r,δ)-RIP with $r \ge 2$, then for all symmetric $\boldsymbol{Z}$, we have $\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})\boldsymbol{Z}\|_2 \le \delta\|\boldsymbol{Z}\|_*$, where $\|\cdot\|_*$ is the nuclear norm. Moreover, if $\mathrm{rank}(\boldsymbol{Z}) \le r-1$, then $\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})\boldsymbol{Z}\|_2 \le \sqrt{r}\delta\|\boldsymbol{Z}\|$.*

We need the following regularity condition on initialization.

Assumption 3.5. For all $1 \le s \le \hat{r} \wedge r_*$, $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \bar{\boldsymbol{U}}\right) \ge \rho$ for some constant $\rho > 0$, where $\boldsymbol{V}_{\boldsymbol{X}_s}$ is defined as Definition 3.1.

The following proposition implies that Assumption 3.5 is satisfied with high probability with a Gaussian initialization.

Proposition 3.6. *Suppose that all entries of $\bar{\boldsymbol{U}} \in \mathbb{R}^{d\times\hat{r}}$ are independently drawn from $\mathcal{N}\left(0, \frac{1}{\hat{r}}\right)$ and $\rho = \epsilon\frac{\sqrt{\hat{r}}-\sqrt{\hat{r}\wedge r_*-1}}{\sqrt{\hat{r}}} \ge \frac{\epsilon}{2r_*}$, then $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \bar{\boldsymbol{U}}\right) \ge \rho$ holds for all $1 \le s \le \hat{r}\wedge r_*$ with probability at least $1 - \hat{r}\left(C\epsilon + e^{-c\hat{r}}\right)$, where $c, C > 0$ are universal constants.*

Lastly, we make the following assumption on the step size.

Assumption 3.7. The step size $\mu \le 10^{-4}\delta\|\boldsymbol{X}\|^{-2}$.

3.3. Procrustes Distance

Our analysis uses the notion of Procrustes distance defined as in Goodall (1991); Tu et al. (2016).

Definition 3.8 (Procrustes Distance)**.** The Procrustes distance between two matrices $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{R}^{d\times s}$ $(d, s > 0)$ is defined as the optimal value of the classic *orthogonal Procrustes problem*:

$$\mathrm{dist}(\boldsymbol{U}_1, \boldsymbol{U}_2) = \min_{\boldsymbol{R}\in\mathbb{R}^{s\times s}:\boldsymbol{R}^\top\boldsymbol{R}=\boldsymbol{I}} \|\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{R}\|_F. \quad (4)$$

We note that the Procrustes distance is well-defined because the set of $s\times s$ orthogonal matrices is compact and thus the continuous function $\|\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{R}\|_F$ in $\boldsymbol{R}$ can attain its minimum. The Procrustes distance is a pseudometric, *i.e.*, it is symmetric and satisfies the triangle inequality.

The following lemma is borrowed from Tu et al. (2016), which connects the Procrustes distance between $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ with the distance between $\boldsymbol{U}_1\boldsymbol{U}_1^\top$ and $\boldsymbol{U}_2\boldsymbol{U}_2^\top$.

Lemma 3.9 (Tu et al. 2016, Lemma 5.4)**.** *For any two matrices $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{R}^{d\times r}$, we have $\left\|\boldsymbol{U}_1\boldsymbol{U}_1^\top - \boldsymbol{U}_2\boldsymbol{U}_2^\top\right\|_F \ge (2\sqrt{2}-2)^{1/2}\cdot\sigma_r(\boldsymbol{U}_1)\cdot\mathrm{dist}(\boldsymbol{U}_1, \boldsymbol{U}_2)$.*

4. Main results

In this section, we present our main theorems and their proof sketches, following the theoretical setup in Section 3. Full proofs can be found in Appendices C to F.

Theorem 4.1. *Under Assumptions 3.3, 3.5 and 3.7, consider GD (3) with initialization $\boldsymbol{U}_{\alpha,0} = \alpha\bar{\boldsymbol{U}}$ for solving the matrix sensing problem (1). There exist universal constants c, M, constant $C = C(\boldsymbol{X}, \bar{\boldsymbol{U}})$ and a sequence of time points $T_\alpha^1 < T_\alpha^2 < \cdots < T_\alpha^{\hat{r}\wedge r_*}$ such that for all $1 \le s \le \hat{r}\wedge r_*$, the following holds when α is sufficiently small:*

$$\left\|\boldsymbol{U}_{\alpha,T_\alpha^s}\boldsymbol{U}_{\alpha,T_\alpha^s}^\top - \boldsymbol{Z}_s^*\right\|_{\mathrm{F}} \le C\alpha^{\frac{1}{M\kappa_*^2}}, \quad (5)$$

where we recall that $\boldsymbol{Z}_s^*$ is the best rank-s solution defined in Definition 1.1. Moreover, GD follows an incremental learning procedure: we have $\lim_{\alpha\to 0}\max_{1\le t\le T_\alpha^s}\sigma_{s+1}(\boldsymbol{U}_{\alpha,t}) = 0$ for all $1 \le s \le \hat{r}\wedge r_*$, where $\sigma_i(\boldsymbol{A})$ denotes the i-th largest singular value of a matrix $\boldsymbol{A}$.

It is guaranteed that $\boldsymbol{Z}_s^*$ is unique for all $1 \le s \le \hat{r}\wedge r_*$ under our assumptions (see Lemma 4.4). In short, Theorem 4.1 states that GD with small initialization discovers the best rank-s solution ($s = 1, 2, \cdots, \hat{r}\wedge r_*$) sequentially. In particular, when $s = r_*$, the best rank-s solution is exactly the ground truth $\boldsymbol{X}\boldsymbol{X}^\top$. Hence with over-parameterization ($\hat{r} \ge r_*$), GD can discover the ground truth.

At a high level, our result characterizes the complete learning dynamics of GD and reveals an incremental learning mechanism, *i.e.*, GD starts from learning simple solutions and then gradually increases the complexity of search space until it finds the ground truth.

In the under-parameterized setting, we can further establish the following convergence result:

Theorem 4.2 (Convergence in the under-parameterized regime)**.** *Suppose that $\hat{r} \le r^*$, then there exists a constant $\bar{\alpha} > 0$ such that when $\alpha < \bar{\alpha}$, we have $\lim_{t\to+\infty}\boldsymbol{U}_{\alpha,t}\boldsymbol{U}_{\alpha,t}^\top = \boldsymbol{Z}_{\hat{r}}^*$.*

4.1. Key lemmas

In this section, we present some key lemmas for proving our main results. First, we can show that with small initialization, GD can get into a small neighborhood of $\boldsymbol{Z}_s^*$.

Lemma 4.3. *Under Assumptions 3.3 and 3.5, there exists $\hat{T}_\alpha^s > 0$ for all $\alpha > 0$ and $1 \le s \le \hat{r}\wedge r_*$ such*

that $\lim_{\alpha\to 0}\max_{1\le t\le \hat{T}_\alpha^s}\sigma_{s+1}(\boldsymbol{U}_{\alpha,t})=0$. *Furthermore, it holds that* $\left\|\boldsymbol{U}_{\hat{T}_\alpha^s}\boldsymbol{U}_{\hat{T}_\alpha^s}^\top-\boldsymbol{Z}_s^*\right\|_F=\mathcal{O}\left(\kappa_*^3 r_*\delta\|\boldsymbol{X}\|^2\right)$.

The full proof can be found in Appendix C. Motivated by Stöger & Soltanolkotabi (2021), we consider the following decomposition of $\boldsymbol{U}_t$:

$$\boldsymbol{U}_t=\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top+\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top, \tag{6}$$

where $\boldsymbol{W}_t:=\boldsymbol{W}_{\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t}\in\mathbb{R}^{\hat{r}\times s}$ is the matrix consisting of the right singular vectors of $\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t$ (Definition 3.1) and $\boldsymbol{W}_{t,\perp}\in\mathbb{R}^{\hat{r}\times(\hat{r}-s)}$ is any orthogonal complement of $\boldsymbol{W}_t$, *i.e.*, $\boldsymbol{W}_t\boldsymbol{W}_t^\top+\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top=\boldsymbol{I}$. The dependence of $\boldsymbol{W}_t,\boldsymbol{W}_{t,\perp}$ on s is omitted but will be clear from the context.

We will refer to the term $\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top$ as the *parallel component* and $\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top$ as the *orthogonal component*. The idea is to show that the parallel component grows quickly until it gets close to the best rank-s solution at some time $\hat{T}_\alpha^s$ (namely $\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top\approx\boldsymbol{Z}_s^*$ when $t=\hat{T}_\alpha^s$). Meanwhile, the orthogonal term grows exponentially slower and stays $o(1)$ before $\hat{T}_\alpha^s$. See Section 5 for a detailed proof sketch.

Lemma 4.3 shows that $\boldsymbol{U}_t\boldsymbol{U}_t^\top$ would enter a neighborhood of $\boldsymbol{Z}_s^*$ with *constant* radius. However, there is still a gap between Lemma 4.3 and Theorem 4.1, since the latter states that $\boldsymbol{U}_t\boldsymbol{U}_t^\top$ would actually get $o(1)$-close to $\boldsymbol{Z}_s^*$.

To proceed, we define the *under-parameterized* matrix sensing loss f_s for every $1\le s\le r_*$:

$$f_s(\boldsymbol{U})=\frac{1}{4}\left\|\mathcal{A}(\boldsymbol{Z}^*-\boldsymbol{U}\boldsymbol{U}^\top)\right\|_2^2,\quad \boldsymbol{U}\in\mathbb{R}^{d\times s}. \tag{7}$$

While the function we are minimizing is f (defined in (1)) rather than f_s, Lemma 4.3 suggests that for $t\le\hat{T}_\alpha^s$, $\boldsymbol{U}_t$ is always approximately rank-s, so that we use a low-rank approximation for $\boldsymbol{U}_{\hat{T}_\alpha^s}$ and associate the dynamics locally with the GD dynamics of f_s. We will elaborate on how this is done in Section 4.2.

When $\operatorname{dist}(\boldsymbol{U}_1,\boldsymbol{U}_2)=0$, it can be easily shown that $f_s(\boldsymbol{U}_1)=f_s(\boldsymbol{U}_2)$ since f_s is invariant to orthogonal transformations. Moreover, we note that the global minimizer of f_s is unique up to orthogonal transformations.

Lemma 4.4. *Under Assumption 3.3, if* $\boldsymbol{U}_s^*\in\mathbb{R}^{d\times s}$ *is a global minimizer of* f_s*, then the set of global minimizers* $\arg\min f_s$ *is equal to* $\left\{\boldsymbol{U}_s^*\boldsymbol{R}:\boldsymbol{R}\in\mathbb{R}^{s\times s},\boldsymbol{R}^\top\boldsymbol{R}=\boldsymbol{I}\right\}$.

Around the global minimizers, we show that f_s satisfies the Restricted Secant Inequality (RSI) which is useful for optimization analysis.

Definition 4.5. For any $\boldsymbol{U}\in\mathbb{R}^{d\times s}$, we use $\Pi_s(\boldsymbol{U})$ to denote the set of *closest* global minimizers of f_s to $\boldsymbol{U}$, namely $\Pi_s(\boldsymbol{U})=\arg\min\{\|\boldsymbol{U}-\boldsymbol{U}_s^*\|_F:\boldsymbol{U}_s^*\in\arg\min f_s\}$.

Lemma 4.6 (Restricted Secant Inequality)**.** *Under Assumption 3.3, if a matrix* $\boldsymbol{U}\in\mathbb{R}^{d\times s}$ *satisfies* $\|\boldsymbol{U}-\boldsymbol{U}_s^*\|_F\le 10^{-2}\kappa_*^{-1}\|\boldsymbol{X}\|$ *for some* $\boldsymbol{U}_s^*\in\Pi_s(\boldsymbol{U})$*, then we have*

$$\langle\nabla f_s(\boldsymbol{U}),\boldsymbol{U}-\boldsymbol{U}_s^*\rangle\ge 0.1\kappa_*^{-1}\|\boldsymbol{X}\|^2\|\boldsymbol{U}-\boldsymbol{U}_s^*\|_F^2. \tag{8}$$

Remark 4.7. In general, a function $g:\mathbb{R}^n\mapsto\mathbb{R}$ satisfies the RSI condition if for some $\mu>0$, $\langle\nabla g(x),x-\pi(x)\rangle\ge\mu\|x-\pi(x)\|^2$ holds for all x, where $\pi(x)$ is a projection of x onto $\arg\min g$. This condition can be used to prove linear convergence of GD (Zhang & Yin, 2013), but it is weaker than strong convexity and stronger than Polyak-Łojasiewicz(PL) condition (Karimi et al., 2016).

We end this subsection with the following lemma which says that all global minimizers of the f_s must be close to $\boldsymbol{X}_s$ under the procrustes distance, which is used in the proof sketch of Theorem 4.1 in Section 4.

Lemma 4.8. *Under Assumption 3.3, we have* $\operatorname{dist}(\boldsymbol{U}_s^*,\boldsymbol{X}_s)\le 40\delta\kappa_*\|\boldsymbol{X}\|_F$ *for any global minimizer* $\boldsymbol{U}_s^*$ *of* f_s*. Moreover,* $\left\|\boldsymbol{Z}_s^*-\boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F\le 160\delta\kappa_*\sqrt{r_*}\|\boldsymbol{X}\|^2$.

Corollary 4.9. *Under Assumption 3.3, we have* $\sigma_{\min}(\boldsymbol{U}_s^*)\ge\frac{1}{2}\sigma_{\min}(\boldsymbol{X}_s)=\frac{1}{2}\sigma_s\ge\frac{1}{2}\kappa_*^{-\frac{1}{2}}\|\boldsymbol{X}\|$.

The full proofs for Lemmas 4.6 and 4.8 and Corollary 4.9 can be found in Appendix E.

4.2. Proof outline

Based on the key lemmas, here we provide the outlines of the proofs for our main theorems and defer the details to Appendix F. We first prove Theorem 4.2 which can be directly derived by combining the lemmas in Section 4.1.

Proof Sketch of Theorem 4.2. For any global minimizer $\boldsymbol{U}_{\hat{r}}^*$ of (1), we have

$$\begin{aligned}
&\operatorname{dist}(\boldsymbol{U}_{\hat{r}}^*,\boldsymbol{U}_{\alpha,\hat{T}_\alpha^{\hat{r}}})\\
&\le(2\sqrt{2}-2)^{-1/2}\sigma_{\min}^{-1}(\boldsymbol{U}_{\hat{r}}^*)\left\|\boldsymbol{U}_{\alpha,\hat{T}_\alpha^{\hat{r}}}\boldsymbol{U}_{\alpha,\hat{T}_\alpha^{\hat{r}}}^\top-\boldsymbol{Z}_{\hat{r}}^*\right\|_F\\
&\le\mathcal{O}(\kappa_*^{\frac{1}{2}}\|\boldsymbol{X}\|^{-1})\cdot\mathcal{O}(\kappa_*^3 r_*\delta\|\boldsymbol{X}\|^2)\\
&=\mathcal{O}(\kappa_*^{3.5}r_*\delta\|\boldsymbol{X}\|),
\end{aligned}$$

where the first inequality is due to Lemma 3.9 and the second one is due to Corollary 4.9 and Lemma 4.3.

Assumption 3.3 and Lemma 4.6 then imply that $\boldsymbol{U}_{\alpha,\hat{T}_\alpha^{\hat{r}}}$ lies in the small neighborhood of the set of global minimizers of $f=f_{\hat{r}}$, in which the RSI holds. Following a standard non-convex optimization analysis (Karimi et al., 2016), we can show that GD converges linearly to $\arg\min f_{\hat{r}}$ (in the Procrustes distance), which yields the conclusion. □

Now we turn to prove Theorem 4.1. While f is not necessarily local RSI, we use a low-rank approximation for $\boldsymbol{U}_t$

and associate the dynamics in this neighborhood with the GD dynamics of f_s.

Proof sketch of Theorem 4.1. Recall that by Lemma 4.3, $\boldsymbol{U}_{\alpha,\hat{T}_\alpha^s}$ is approximately rank-s. So there must exist a matrix $\bar{\boldsymbol{U}}_{\alpha,0} \in \mathbb{R}^{d\times \hat{r}}$ with $\mathrm{rank}(\bar{\boldsymbol{U}}_{\alpha,0}) \le s$ such that

$$\bar{\boldsymbol{U}}_{\alpha,0}\bar{\boldsymbol{U}}_{\alpha,0}^\top - \boldsymbol{U}_{\alpha,T_\alpha^s}\boldsymbol{U}_{\alpha,T_\alpha^s}^\top = o(1) \quad \text{as } \alpha \to 0. \tag{9}$$

Indeed, we can let $\bar{\boldsymbol{U}}_{\alpha,t}$ be the parallel component of $\boldsymbol{U}_{\alpha,T_\alpha^s}$ because the orthogonal component stays $o(1)$ (see the discussions following (6) and Corollary 5.5 for details).

Let $\{\bar{\boldsymbol{U}}_{\alpha,t}\}_{t\ge 0}$ be the trajectory of GD with step size μ, initialized at $\bar{\boldsymbol{U}}_{\alpha,0}$. Since the gradient of the objective function f is locally Lipschitz, the solution obtained by the two GD trajectories $\{\bar{\boldsymbol{U}}_{\alpha,t}\}_{t\ge 0}$ and $\{\boldsymbol{U}_{\alpha,\hat{T}_\alpha^s+t}\}_{t\ge 0}$ will remain $o(1)$-close for at least constantly many steps. Indeed we can show that they will keep $o(1)$ close for some $\bar{t}_\alpha = \omega(1)$ steps, *i.e.*, for all $t \in [0, \bar{t}_\alpha]$,

$$\bar{\boldsymbol{U}}_{\alpha,t}\bar{\boldsymbol{U}}_{\alpha,t}^\top - \boldsymbol{U}_{\alpha,\hat{T}_\alpha^s+t}\boldsymbol{U}_{\alpha,\hat{T}_\alpha^s+t}^\top = o(1). \tag{10}$$

From (3) it is evident that GD initialized at $\bar{\boldsymbol{U}}_{\alpha,0}$ actually lives in the space of matrices with rank $\le s$. Indeed we can identify its dynamics with another GD on f_s (defined in (7)). Concretely, let $\hat{\boldsymbol{U}}_{\alpha,0} \in \mathbb{R}^{d\times s}$ be a matrix so that $\hat{\boldsymbol{U}}_{\alpha,0}\hat{\boldsymbol{U}}_{\alpha,0}^\top = \bar{\boldsymbol{U}}_{\alpha,0}\bar{\boldsymbol{U}}_{\alpha,0}^\top$, and let $\{\hat{\boldsymbol{U}}_{\alpha,t}\}_{t\ge 0}$ be the trajectory of GD that optimizes f_s with step size μ starting from $\hat{\boldsymbol{U}}_{\alpha,0}$. Then we have $\hat{\boldsymbol{U}}_{\alpha,t}\hat{\boldsymbol{U}}_{\alpha,t}^\top = \bar{\boldsymbol{U}}_{\alpha,t}\bar{\boldsymbol{U}}_{\alpha,t}^\top$ for all $t \ge 0$.

We can now apply our landscape results for f_s to analyze the GD trajectory $\{\hat{\boldsymbol{U}}_{\alpha,t}\}_{t\ge 0}$. By (9) and Lemma 4.3, we have $\left\|\hat{\boldsymbol{U}}_{\alpha,0}\hat{\boldsymbol{U}}_{\alpha,0}^\top - \boldsymbol{Z}_s^*\right\|_F = \mathcal{O}(\kappa_*^3 r_* \delta \|\boldsymbol{X}\|^2)$, so using a similar argument as in the proof sketch of Theorem 4.2, Corollary 4.9 and Lemma 3.9 imply that the initialization $\hat{\boldsymbol{U}}_{\alpha,0}$ is within an $\mathcal{O}\left(\kappa_*^{3.5} r_* \delta \|\boldsymbol{X}\|^2\right)$ neighborhood of the set of global minimizers of $f_s(\boldsymbol{U})$. From Assumption 3.3 and Lemma 4.6 we know that that $f_s(\boldsymbol{U})$ satisfies a local RSI condition in this neighborhood, so following standard non-convex optimization analysis (Karimi et al., 2016), we can show that $\{\hat{\boldsymbol{U}}_{\alpha,t}\}_{t\ge 0}$ converges linearly to its set of global minimizers in the Procrustes distance. We need to choose a time t such that (10) remains true while this linear convergence process takes place for sufficiently many steps. This is possible since $\bar{t}_\alpha = \omega(1)$; indeed we can show that there always exists some $t = t_\alpha^s \le \bar{t}_\alpha$ such that both $\left\|\hat{\boldsymbol{U}}_{\alpha,t}\hat{\boldsymbol{U}}_{\alpha,t}^\top - \boldsymbol{U}_{\alpha,\hat{T}_\alpha^s+t}\boldsymbol{U}_{\alpha,\hat{T}_\alpha^s+t}^\top\right\|_F$ and $\left\|\hat{\boldsymbol{U}}_{\alpha,t} - \boldsymbol{U}_s^*\right\|_F$ are bounded by $\mathcal{O}(\alpha^{\frac{1}{M\kappa_*^2}})$. Hence $\left\|\boldsymbol{U}_{\alpha,t}\boldsymbol{U}_{\alpha,t}^\top - \boldsymbol{Z}_s^*\right\|_F = \mathcal{O}(\alpha^{\frac{1}{M\kappa_*^2}})$ when $t = T_\alpha^s := \hat{T}_\alpha^s + t_\alpha^s$.

For $1 \le s < \hat{r} \wedge r_*$ and $t \le t_\alpha^s$, since (10) holds and $\mathrm{rank}(\hat{\boldsymbol{U}}_{\alpha,t}) \le s$, we have $\max_{1\le t\le T_\alpha^s} \sigma_{s+1}(\boldsymbol{U}_{\alpha,t}) \to 0$ as $\alpha \to 0$. Finally, by Lemma 4.8 and Assumption 3.3 we have $\left\|\boldsymbol{Z}_{s+1}^* - \boldsymbol{X}_{s+1}\boldsymbol{X}_{s+1}^\top\right\| = \mathcal{O}(\delta\kappa_*\sqrt{r_*}) = \mathcal{O}(\kappa_*^{-1}\|\boldsymbol{X}\|^2)$, so $\sigma_{s+1}(\boldsymbol{Z}_{s+1}^*) \gtrsim \sigma_{s+1}^2$. Therefore, $\boldsymbol{U}_{\alpha,t}\boldsymbol{U}_{\alpha,t}^\top$ cannot be close to $\boldsymbol{Z}_{s+1}^*$ when $t \le T_\alpha^s$, so we must have $T_\alpha^{s+1} > T_\alpha^s$. This completes the proof of Theorem 4.1. □

5. Proof sketch of Lemma 4.3

In this section, we outline the proof sketch of Lemma 4.3. We divide the GD dynamics intro three phases and characterize the dynamics separately. Proof details for these three phases can be found in Appendices C.1, C.2 and C.4.

5.1. The spectral phase

Starting from a small initialization, GD initially behaves similarly to power iteration since $\boldsymbol{U}_{t+1} = (\boldsymbol{I} + \mu\boldsymbol{M}_t)\boldsymbol{U}_t \approx (\boldsymbol{I} + \mu\boldsymbol{M})\boldsymbol{U}_t$, where $\boldsymbol{M} := \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top)$ is a symmetric matrix. Let $\boldsymbol{M} = \sum_{k=1}^d \hat{\sigma}_k^2 \hat{\boldsymbol{v}}_k\hat{\boldsymbol{v}}_k^\top$ be the eigendecomposition of $\boldsymbol{M}$, where $\hat{\sigma}_1 \ge \hat{\sigma}_2 \ge \cdots \ge \hat{\sigma}_d \ge 0$. Using our assumption on δ (Assumption 3.3), we can show that $|\sigma_i - \hat{\sigma}_i|, 1 \le i \le s$ are sufficiently small so that $\hat{\sigma}_i$'s are positive and well-separated. Then we have

$$\begin{aligned}\boldsymbol{U}_T \approx (\boldsymbol{I} + \mu\boldsymbol{M})^T\boldsymbol{U}_0 &= \sum_{i=1}^d (1+\mu\hat{\sigma}_i^2)^T \hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i^\top \boldsymbol{U}_0 \\ &\approx \sum_{i=1}^s (1+\mu\hat{\sigma}_i^2)^T \hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i^\top \boldsymbol{U}_0,\end{aligned} \tag{11}$$

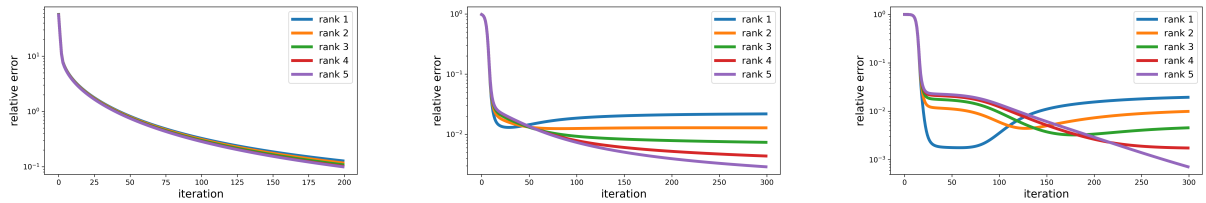
where the last step holds because $(1+\mu\hat{\sigma}_s)^T \gg (1+\mu\hat{\sigma}_{s+1})^T$. In other words, we can expect that there is an exponential separation between the parallel and orthogonal component of $\boldsymbol{U}_T$. Formally, we can prove the following property at the end of the spectral phase:

Lemma 5.1 (Lemma C.4, simplified version)**.** *Suppose that Assumptions 3.3, 3.5 and 3.7 hold. Then there exist positive constants $C_i = C_i(\boldsymbol{X}, \bar{\boldsymbol{U}}), \gamma_i = \gamma_i(\boldsymbol{X}, \bar{\boldsymbol{U}}), i = 2, 3$ independent of α such that $\gamma_2 < \gamma_3$ and the following inequalities hold for $t = T_\alpha^{\mathtt{sp}} = \mathcal{O}\left(\frac{\log \alpha^{-1}}{\log(1+\mu\|\boldsymbol{X}\|^2)}\right)$ when α is sufficiently small:*

$$\|\boldsymbol{U}_t\| \le \|\boldsymbol{X}\|, \quad \sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t) \ge C_2 \cdot \alpha^{\gamma_2},$$
$$\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \le C_3 \cdot \alpha^{\gamma_3}, \text{and} \quad \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \le 200\delta.$$

5.2. The parallel improvement phase

For small α, we have $\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t) \gg \|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$ by the end of the spectral phase. When (11) no longer holds, we enter a new phase which we call the *parallel improvement phase*. In this phase, the ratio $\frac{\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t)}{\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|}$ grows exponentially in t, until the former reaches a constant scale. Formally, let $T_{\alpha,s}^{\mathtt{pi}} =$



(a) $\alpha = 1$, 1000 measurements. (b) $\alpha = 0.1$, 1000 measurements. (c) $\alpha = 0.01$, 1000 measurements.

(d) $\alpha = 0.001$, 1000 measurements. (e) $\alpha = 0.001$, 1500 measurements. (f) $\alpha = 0.001$, 2000 measurements.

Figure 1. The evolution of relative error against the best solution of different ranks over time.

$\min\left\{t \geqslant 0 : \sigma_{\min}^2\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{\alpha,t+1}\right) > 0.3\kappa_*^{-1}\|\boldsymbol{X}\|^2\right\}$, then we can prove the following lemma via induction.

Lemma 5.2. *Suppose that Assumptions 3.3, 3.5 and 3.7 hold and let $c_3 = 10^4\kappa_* r_*^{\frac{1}{2}}\delta$. Then for sufficiently small α, the following inequalities hold when $T_\alpha^{\mathtt{sp}} \leq t < T_{\alpha,s}^{\mathtt{pi}}$:*

$$\begin{aligned}
&\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right) \geq \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\right) \\
&\quad \geq \left(1 + 0.5\mu\left(\sigma_s^2 + \sigma_{s+1}^2\right)\right)\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\right), \quad (13a)\\
&\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \\
&\quad \leq \left(1 + \mu\left(0.4\sigma_s^2 + 0.6\sigma_{s+1}^2\right)\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|, \quad (13b)\\
&\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1}}\right\| \leq c_3, \quad (13c)\\
&\operatorname{rank}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}) = \operatorname{rank}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t) = s. \quad (13d)
\end{aligned}$$

We can immediately deduce from Lemmas 5.1 and 5.2 that the orthogonal term $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$ remains $o(1)$ by the end of the parallel improvement phase:

Corollary 5.3 (Lemma C.11, simplified version)**.** *Under the conditions of Lemma 5.2, when α is sufficiently small we have $\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{pi}}}\boldsymbol{W}_{T_{\alpha,s}^{\mathtt{pi}},\perp}\right\| \leq C_5 \cdot \alpha^{\frac{1}{4\kappa_*}}$ for some constant $C_5 = C_5(\boldsymbol{X}, \bar{\boldsymbol{U}})$.*

5.3. The refinement phase

After $\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t)$ grows to constant scale, we enter the *refinement phase* for which we show that $\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}}$ keeps decreasing until it is $\mathcal{O}\left(\delta\kappa_*^3 r_*\|\boldsymbol{X}\|^2\right)$. Formally, let $\tau = \kappa_*^{-1}\|\boldsymbol{X}\|^2$ and $T_{\alpha,s}^{\mathtt{ft}} = T_{\alpha,s}^{\mathtt{pi}} - \frac{\log\left(10^{-2}\|\boldsymbol{X}\|^{-2}\kappa_*^{-1}c_3^{-1}\right)}{\log\left(1-\frac{1}{2}\mu\tau\right)} > T_{\alpha,s}^{\mathtt{pi}}$ where c_3 is defined in Lemma 5.2, then the following lemma holds.

Lemma 5.4. *Suppose that $T_{\alpha,s}^{\mathtt{pi}} \leq t \leq T_{\alpha,s}^{\mathtt{ft}}$ and all the conditions in Lemma 5.2 hold, then we have*

$$\begin{aligned}
&\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top)\right\|_F \\
&\leq \left(1 - \frac{1}{2}\mu\tau\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F \\
&\quad + 20\mu\|\boldsymbol{X}\|^4(\delta + 5c_3) + 2000\mu^2\sqrt{r_*}\|\boldsymbol{X}\|^6.
\end{aligned}$$

Moreover, we have $\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \leq (1+\sigma_s^2)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$ and $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \leq c_3$.

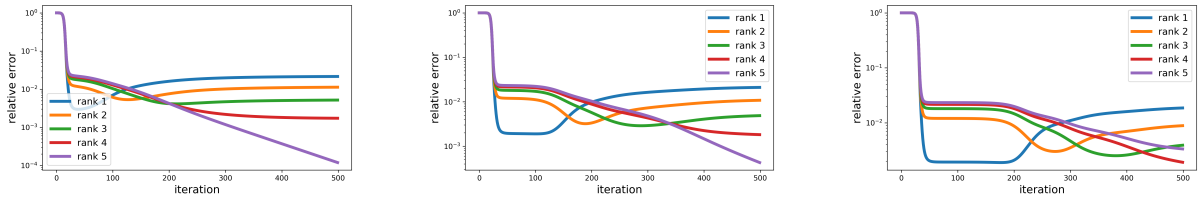
Using Lemma 5.4, we arrive at the following result:

Corollary 5.5. *For sufficiently small α, at $t = T_{\alpha,s}^{\mathtt{ft}}$ we have $\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F \leq 80\delta\kappa_*^3 r_*\|\boldsymbol{X}\|^2$ and $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| = o(1)$ $(\alpha \to 0)$.*

Concluding the proof of Lemma 4.3. At $t = T_{\alpha,s}^{\mathtt{ft}}$, we have

$$\begin{aligned}
&\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_F \\
&\leq \left\|\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\right\|_F \quad (14a)\\
&\quad + \left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\right\|_F \\
&\leq \left\|\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\right\|_F \quad (14b)\\
&\quad + \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\right\|_F \\
&\leq \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F \\
&\quad + \sqrt{r_*}\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_t\right\|^2 + \sqrt{d}\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|^2 \quad (14c)\\
&\leq \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F \\
&\quad + 9\sqrt{r_*}\|\boldsymbol{X}\|^2\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|^2 + \sqrt{d}\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|^2 \quad (14d)\\
&= \mathcal{O}\left(\delta\kappa_*^3 r_*\|\boldsymbol{X}\|^2 + \|\boldsymbol{X}\|^2 c_3^2\sqrt{r_*}\right) + o(1) \quad (14e)\\
&= \mathcal{O}\left(\delta\kappa_*^3 r_*\|\boldsymbol{X}\|^2\right), \quad (14f)
\end{aligned}$$

where (14c) uses $\|\boldsymbol{A}\|_F \leq \sqrt{\operatorname{rank}(\boldsymbol{A})}\|\boldsymbol{A}\|$, (14d) uses $\|\boldsymbol{U}_t\| \leq 3\|\boldsymbol{X}\|$, (14e) follows from Lemma 5.4 and Corollary 5.5 and the last step follows from $c_3 = 10^4\kappa_*\sqrt{r_*}\delta$ and Assumption 3.3.



(a) $\alpha = 0.1$, 1000 measurements. (b) $\alpha = 0.01$, 1000 measurements. (c) $\alpha = 0.001$, 1000 measurements.

Figure 2. The evolution of the loss and relative error against best solution of different ranks in the exact-parameterized case $\hat{r} = r_* = 5$.

By Lemma 4.8, the best rank-s solution is close to the matrix factorization minimizer *i.e.* $\left\|\boldsymbol{Z}_s^* - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F = \mathcal{O}\left(\delta\kappa_*\sqrt{r_*}\|\boldsymbol{X}\|^2\right)$. We thus obtain that $\left\|\boldsymbol{Z}_s^* - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}} = \mathcal{O}\left(\delta\kappa_*^3 r_*\|\boldsymbol{X}\|^2\right)$. Finally, since $\mathrm{rank}(\boldsymbol{U}_t\boldsymbol{W}_t) \le s$ (recall the decomposition (6)), we have $\sigma_{s+1}(\boldsymbol{U}_t) \le \|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| = o(1)$. The conclusion follows.

6. Experiments

In this section, we perform numerical experiments to empirically verify our theoretical findings.

Experimental setup. We consider the matrix sensing problem (1) with $d = 50$, $r_* = 5$, $\alpha \in \{1, 0.1, 0.01, 0.001\}$, $m \in \{1000, 2000, 5000\}$. We will consider different choices for $\hat{r}$ in the experiments. The ground truth $\boldsymbol{Z}^* = \boldsymbol{X}\boldsymbol{X}^\top$ is generated such that the entries of $\boldsymbol{X}$ are i.i.d. standard Gaussian variables. We use the same ground truth throughout our experiments.

For $i = 1, 2, \cdots, m$, all entries of the measurement $\boldsymbol{A}_i \in \mathbb{R}^{d\times d}$ are chosen i.i.d. from the standard Gaussian $\mathcal{N}(0, 1)$. When $m \gtrsim dr_*\delta^{-2}$, this set of measurements satisfies the RIP with high probability (Recht et al., 2010, Theorem 4.2).

We solve the problem (1) via running GD for $T = 10^4$ iterations starting with small initialization with scale α. Specifically, we choose $\boldsymbol{U}_0 = \alpha\bar{\boldsymbol{U}}$ where the entries of $\bar{\boldsymbol{U}} \in \mathbb{R}^{d\times\hat{r}}$ are drawn i.i.d. from $\mathcal{N}(0, 1)$. We consider both the over-parameterized and the exact/under-parameterized regime. The learning rate of GD is set to be $\mu = 0.005$.

6.1. Implicit low-rank bias

In this subsection, we consider the over-parameterized setting with $r = 50$. For each iteration $t \in [T]$ and rank $s \in [r_*]$, we define the relative error $\mathcal{E}_s(t) = \frac{\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F^2}{\|\boldsymbol{X}_s\boldsymbol{X}_s^\top\|_F^2}$ to measure the proximity of the GD iterates to $\boldsymbol{X}_s$. We plot the relative error in Figure 1 for different choices of α and m (which affects the measurement error δ).

Small initialization. The implicit low-rank bias of GD is evident when the initialization scale α is small. Indeed, one can observe that GD first visits a small neighborhood of $\boldsymbol{X}_1$, spends a long period of time near it, and then moves towards $\boldsymbol{X}_2$. It then proceeds to learn $\boldsymbol{X}_3, \boldsymbol{X}_4, \cdots$ in a similar way, until it finally fits the ground truth. This is in align with Theorem 4.1. By contrast, for large initialization we do not have this implicit bias.

The effect of measurement error. For fixed α, one can observe the relative error becomes smaller when the number of measurements increases. This is in align with Lemma 4.3 in which the bound depends on δ. In particular, for the case $s = r_*$, in the end the distance to the set of global minima goes to zero as $\alpha \to 0$.

6.2. Matrix sensing with exact parameterization

Now we study the behavior of GD in the exact parameterization regime ($r = r_*$). We fix $m = 1000$, $r = r_* = 5$ and run GD for $T = 500$ iterations. We plot the relative error in Figure 2. As predicted by Theorem 4.1, we can observe that when α is small, GD exhibits an implicit low-rank bias and takes a longer time to converge. The latter is because GD would get into a $\mathrm{poly}(\alpha)$-small neighborhood of the saddle point $\boldsymbol{Z}_s$ and take a long time to escape the saddle. As guaranteed by Theorem 4.2, we also observe the final convergence to global minimizers for sufficiently small α.

7. Conclusion

In this paper, we study the matrix sensing problem with RIP measurements and show that GD with small initialization follows an incremental learning procedure, where GD finds near-optimal solutions with increasing ranks until it finds the ground-truth. We take a step towards understanding the optimization and generalization aspects of simple optimization methods, thereby providing insights into their success in modern applications such as deep learning (Goodfellow et al., 2016). Also, we provide a detailed landscape analysis in the under-parameterized regime, which to the best of our knowledge is the first analysis of this kind.

Although we focus on matrix sensing in this paper, it has been revealed in a line of works that the implicit regularization effect may vary for different models, including deep matrix factorization (Arora et al., 2019) and nonlinear ReLU/LeakyReLU networks (Lyu et al., 2021; Timor et al., 2022). Also, it is shown in Woodworth et al. (2020) that different initialization scales can lead to distinct inductive

bias and affect the generalization and optimization behaviors. All these results indicate that we need further studies to comprehensively understand gradient-based optimization methods from the generalization aspect.

Acknowledgements

JJ is partially supported by the elite undergraduate training program of School of Mathematical Sciences in Peking University. KL is supported by funding from NSF, ONR, Simons Foundation, DARPA and SRC. JDL acknowledges support of the ARO under MURI Award W911NF-11-1-0304, the Sloan Research Fellowship, NSF CCF 2002272, NSF IIS 2107304, NSF CIF 2212262, ONR Young Investigator Award, and NSF CAREER Award 2144994. SSD acknowledges support of NSF IIS 2110170, NSF DMS 2134106, NSF CCF 2212261, NSF IIS 2143493, NSF CCF 2019844.

References

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Belabbas, M. A. On implicit regularization: Morse functions and applications to matrix factorization. *arXiv preprint arXiv:2001.04264*, 2020.

Blanc, G., Gupta, N., Valiant, G., and Valiant, P. Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 483–513. PMLR, 09–12 Jul 2020.

Boursier, E., Pillaud-Vivien, L., and Flammarion, N. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *arXiv preprint arXiv:2206.00939*, 2022.

Candes, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3): 1–37, 2011.

Damian, A., Ma, T., and Lee, J. D. Label noise SGD provably prefers flat global minimizers. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27449–27461. Curran Associates, Inc., 2021.

Davenport, M. A. and Romberg, J. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4): 608–622, 2016.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Fornasier, M., Rauhut, H., and Ward, R. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21(4):1614–1640, 2011.

Frei, S., Cao, Y., and Gu, Q. Provable generalization of sgd-trained neural networks of any width in the presence of adversarial label noise. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3427–3438. PMLR, 18–24 Jul 2021.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pp. 797–842. PMLR, 2015.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pp. 2973–2981, 2016.

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

Gidel, G., Bach, F., and Lacoste-Julien, S. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Gissin, D., Shalev-Shwartz, S., and Daniely, A. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2019.

Goodall, C. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(2):285–339, 1991. ISSN 00359246.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT Press, 2016.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

Gunasekar, S., Lee, J., Soudry, D., and Srebro, N. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Hu, W., Xiao, L., Adlam, B., and Pennington, J. The surprising simplicity of the early-time learning dynamics of neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17116–17128. Curran Associates, Inc., 2020.

Jacot, A., Ged, F., Gabriel, F., Şimşek, B., and Hongler, C. Deep linear networks dynamics: Low-rank biases induced by initialization scale and l2 regularization. *arXiv preprint arXiv:2106.15933*, 2021.

Ji, Z. and Telgarsky, M. Directional convergence and alignment in deep learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17176–17186. Curran Associates, Inc., 2020.

Jiang, L., Chen, Y., and Ding, L. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *ArXiv*, abs/2203.02839, 2022.

Kalimeris, D., Kaplun, G., Nakkiran, P., Edelman, B., Yang, T., Barak, B., and Zhang, H. Sgd on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Kalogerias, D. S. and Petropulu, A. P. Matrix completion in colocated mimo radar: Recoverability, bounds & theoretical guarantees. *IEEE Transactions on Signal Processing*, 62(2):309–321, 2013.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pp. 1246–1257, 2016.

Li, X., Lu, J., Arora, R., Haupt, J., Liu, H., Wang, Z., and Zhao, T. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514, 2019.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pp. 2–47. PMLR, 2018.

Li, Z., Luo, Y., and Lyu, K. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2020.

Li, Z., Wang, T., and Arora, S. What happens after SGD reaches zero loss? –a mathematical framework. In *International Conference on Learning Representations*, 2022a.

Li, Z., Wang, T., Lee, J. D., and Arora, S. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=k4KHXS6_zOV.

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Lyu, K., Li, Z., Wang, R., and Arora, S. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34: 12978–12991, 2021.

Lyu, K., Li, Z., and Arora, S. Understanding the generalization benefit of normalization layers: Sharpness reduction. *arXiv preprint arXiv:2206.07085*, 2022.

Nacson, M. S., Gunasekar, S., Lee, J., Srebro, N., and Soudry, D. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In *International Conference on Machine Learning*, pp. 4683–4692. PMLR, 2019.

Ngo, T. and Saad, Y. Scaled gradients on grassmann manifolds for matrix completion. *Advances in neural information processing systems*, 25, 2012.

Peng, Y., Suo, J., Dai, Q., and Xu, W. Reweighted low-rank matrix recovery and its application in image restoration. *IEEE transactions on cybernetics*, 44(12):2418–2430, 2014.

Razin, N. and Cohen, N. Implicit regularization in deep learning may not be explainable by norms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H.

(eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21174–21187. Curran Associates, Inc., 2020.

Razin, N., Maman, A., and Cohen, N. Implicit regularization in tensor factorization. In *International Conference on Machine Learning*, pp. 8913–8924. PMLR, 2021.

Razin, N., Maman, A., and Cohen, N. Implicit regularization in hierarchical tensor factorization and deep convolutional neural networks. *arXiv preprint arXiv:2201.11729*, 2022.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Rudelson, M. and Vershynin, R. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009.

Shen, X. and Wu, Y. A unified approach to salient object detection via low rank matrix recovery. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 853–860. IEEE, 2012.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19 (70):1–57, 2018.

Stöger, D. and Soltanolkotabi, M. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34, 2021.

Sun, R. Optimization for deep learning: theory and algorithms. *arXiv preprint arXiv:1912.08957*, 2019.

Timor, N., Vardi, G., and Shamir, O. Implicit regularization towards rank minimization in relu networks. *arXiv preprint arXiv:2201.12760*, 2022.

Tong, T., Ma, C., and Chi, Y. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *J. Mach. Learn. Res.*, 22:150–1, 2021.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. Low-rank solutions of linear matrix equations via Procrustes flow. In *International Conference on Machine Learning*, pp. 964–973. PMLR, 2016.

Wedin, P.-Å. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

Wei, K., Cai, J.-F., Chan, T. F., and Leung, S. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37 (3):1198–1222, 2016.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Xu, H., Caramanis, C., and Sanghavi, S. Robust pca via outlier pursuit. *Advances in neural information processing systems*, 23, 2010.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Zhang, H. and Yin, W. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

Zhao, B., Haldar, J. P., Brinegar, C., and Liang, Z.-P. Low rank matrix recovery for real-time cardiac mri. In *2010 ieee international symposium on biomedical imaging: From nano to macro*, pp. 996–999. IEEE, 2010.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. The global optimization geometry of low-rank matrix optimization. *IEEE Transactions on Information Theory*, 67(2):1308–1331, 2021.

Zou, W., Kpalma, K., Liu, Z., and Ronsin, J. Segmentation driven low-rank matrix recovery for saliency detection. In *24th British machine vision conference (BMVC)*, pp. 1–13, 2013.

The appendix is organized as follows: in Appendix A we present a number of results that will be used for later proof. Appendix B sketches the main idea for proving our main results. Appendix C is devoted to a rigorous proof of Lemma 4.3 ,with some auxiliary lemmas proved in Appendix D. In Appendix E we analyze the landscape of low-rank matrix sensing and prove our landscape results in Section 4.1. These results are then used in Appendix F to prove Theorems 4.1 and 4.2. Finally, Appendix G studies the landscape of rank-1 matrix sensing, which enjoys a strongly convex property, as we mentioned in Section 4.1 without proof.

A. Preliminaries

In this section, we present some useful results that is needed in subsequent analysis.

A.1. The RIP condition and its properties

In this subsection, we collect a few useful properties of the RIP condition, which we recall below:

Definition A.1. We say that the measurement $\mathcal{A}$ satisfies the (δ, r)-RIP condition if for all matrices $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$ with $\operatorname{rank}(\boldsymbol{Z}) \leq r$, we have

$$(1-\delta)\|\boldsymbol{Z}\|_F^2 \leq \|\mathcal{A}(\boldsymbol{Z})\|_2^2 \leq (1-\delta)\|\boldsymbol{Z}\|_F^2.$$

The key intuition behind RIP is that $\mathcal{A}^*\mathcal{A} \approx I$, where $\mathcal{A}^* : \boldsymbol{v} \mapsto \frac{1}{\sqrt{m}}\sum_{i=1}^m v_i \boldsymbol{A}_i$ is the adjoint of $\mathcal{A}$. This intuition is made rigorous by the following proposition:

Proposition A.2. *(Stöger & Soltanolkotabi, 2021, Lemma 7.3) Suppose that $\mathcal{A}$ satisfies (r,δ)-RIP with $r \geq 2$, then for all symmetric $\boldsymbol{Z}$,*

(1). if $\operatorname{rank}(\boldsymbol{Z}) \leq r-1$, we have $\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})\boldsymbol{Z}\|_2 \leq \sqrt{r}\delta\|\boldsymbol{Z}\|$.

(2). $\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})\boldsymbol{Z}\|_2 \leq \delta\|\boldsymbol{Z}\|_*$, where $\|\cdot\|_*$ is the nuclear norm.

A.2. Matrix analysis

The following lemma is a direct corollary of Proposition A.2 and will be frequently used in our proof.

Lemma A.3. *Suppose that the measurement $\mathcal{A}$ satisfies $(\delta, 2r_*+1)$-RIP condition, then for all matrices $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ such that $\operatorname{rank}(\boldsymbol{U}) \leq r_*$, we have*

$$\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top\right)\right\| \leq \delta\sqrt{r_*}\left(\|\boldsymbol{X}\|^2 + \|\boldsymbol{U}\|^2\right).$$

In our proof we will frequently make use of the Weyl's inequality for singular values:

Lemma A.4 (Weyl's inequality)**.** *Let $\boldsymbol{A}, \Delta \in \mathbb{R}^{d\times d}$ be two matrices, then for all $1 \leq k \leq d$, we have*

$$|\sigma_k(\boldsymbol{A}) - \sigma_k(\boldsymbol{A}+\Delta)| \leq \|\Delta\|.$$

We will also need the Wedin's sin theorem for singular value decomposition:

Lemma A.5. *(Wedin, 1972, Section 3) Define $R(\cdot)$ to be the column space of a matrix. Suppose that matrices $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{T}$, $\boldsymbol{A}_1$, $\boldsymbol{B}_1$ are the top-s components in the SVD of $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively, and $\boldsymbol{A}_0 = \boldsymbol{A} - \boldsymbol{A}_1, \boldsymbol{B}_0 = \boldsymbol{B} - \boldsymbol{B}_1$. If $\delta = \sigma_{\min}(\boldsymbol{B}_1) - \sigma_{\max}(\boldsymbol{A}_0) > 0$, then we have*

$$\|\sin\Theta\left(R(\boldsymbol{A}_1), R(\boldsymbol{B}_1)\right)\| \leq \frac{\|\boldsymbol{T}\|}{\delta}$$

where $\Theta(\cdot,\cdot)$ denotes the angle between two subspaces.

Equipped with Lemma A.3, we can have the following characterization of the eigenvalues of $\boldsymbol{M}$ (recall that $\boldsymbol{M} = \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top)$):

Lemma A.6. *Let $\boldsymbol{M} := \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top)$ and $\boldsymbol{M} = \sum_{k=1}^d \hat{\sigma}_k^2 \hat{\boldsymbol{v}}_k\hat{\boldsymbol{v}}_k^\top$ be the eigen-decomposition of $\boldsymbol{M}$. For $1 \leq i \leq d$ we have*

$$\left|\sigma_i^2 - \hat{\sigma}_i^2\right| \leq \delta\|\boldsymbol{X}\|^2.$$

Proof. By Weyl's inequality we have

$$\left|\sigma_i^2 - \hat{\sigma}_i^2\right| \le \left\|\boldsymbol{M} - \boldsymbol{X}\boldsymbol{X}^\top\right\| \le \delta\|\boldsymbol{X}\|^2$$

as desired. □

A.3. Optimization

Lemma A.7. *Suppose that a smooth function $f \in \mathbb{R}^m \mapsto \mathbb{R}$ with minimum value $f^* > -\infty$ satisfies the following conditions with some $\epsilon > 0$:*

(1). $\lim_{\|\boldsymbol{x}\|\to+\infty} f(\boldsymbol{x}) = +\infty$.

(2). There exists an open subset $S \subset \mathbb{R}^m$ such that the set S^* of global minima of f is contained in S, and for all stationary points x of f in $\mathbb{R}^m - S$, we have $f(\boldsymbol{x}) - f^* \ge 2\epsilon$. Moreover, we also have $f(\boldsymbol{x}) - f^* \ge 2\epsilon$ on ∂S.

Then we have

$$\{\boldsymbol{x} \in \mathbb{R}^m : f(\boldsymbol{x}) - f^* \le \epsilon\} \subset S.$$

Proof. Let $\boldsymbol{x}^*$ be the minimizer of f on $\mathbb{R}^m - S$. By condition (1) we can deduce that $\boldsymbol{x}^*$ always exists. Moreover, since any local minimizer of a function defined on a compact set must either be a stationary point or lie on the boundary of its domain, we can see that either $\boldsymbol{x}^* \in \partial S$ or $\nabla f(\boldsymbol{x}^*) = 0$ holds. By condition (2), either cases would imply that $f(\boldsymbol{x}^*) - f^* \ge 2\epsilon$, as desired. □

Lemma A.8. *Let $\{\boldsymbol{x}_k\}, \{\boldsymbol{y}_k\} \subset \mathbb{R}^n$ be two sequences generated by $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \mu\nabla f(\boldsymbol{x}_k)$ and $\boldsymbol{y}_{k+1} = \boldsymbol{y}_k - \mu\nabla f(\boldsymbol{y}_k)$. Suppose that $\|\boldsymbol{x}_k\| \le B$ and $\|\boldsymbol{y}_k\| \le B$ for all k and f is L-smooth in $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \le B\}$, then we have*

$$\|\boldsymbol{x}_k - \boldsymbol{y}_k\| \le (1 + \mu L)^k \|\boldsymbol{x}_0 - \boldsymbol{y}_0\|.$$

Proof. The update rule implies that

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{y}_{k+1}\| &= \|\boldsymbol{x}_k - \boldsymbol{y}_k - \mu\nabla f(\boldsymbol{x}_k) + \mu\nabla f(\boldsymbol{y}_k)\| \\ &\le \|\boldsymbol{x}_k - \boldsymbol{y}_k\| + \mu\|\nabla f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k)\| \\ &\le (1 + \mu L)\|\boldsymbol{x}_k - \boldsymbol{y}_k\|\end{aligned}$$

which yields the desired inequality. □

A.4. Proof for Proposition 3.6

Proposition 3.6. *Suppose that all entries of $\bar{\boldsymbol{U}} \in \mathbb{R}^{d\times\hat{r}}$ are independently drawn from $\mathcal{N}\left(0, \frac{1}{\hat{r}}\right)$ and $\rho = \epsilon\frac{\sqrt{\hat{r}}-\sqrt{\hat{r}\wedge r_*-1}}{\sqrt{\hat{r}}} \ge \frac{\epsilon}{2r_*}$, then $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\bar{\boldsymbol{U}}\right) \ge \rho$ holds for all $1 \le s \le \hat{r}\wedge r_*$ with probability at least $1 - \hat{r}\left(C\epsilon + e^{-c\hat{r}}\right)$, where $c, C > 0$ are universal constants.*

Proposition 3.6 immediately follows from the following result:

Proposition A.9. *(Restatement of Rudelson & Vershynin, 2009, Theorem 1.1) Let A be an $N \times n$ random matrix, $N \ge n$, whose elements are independent copies of a mean zero sub-gaussian random variable with unit variance. Then, for every $\varepsilon > 0$, we have $\mathbb{P}\left(s_n(A) \le \varepsilon(\sqrt{N} - \sqrt{n-1})\right) \le (C\varepsilon)^{N-n+1} + e^{-cN}$ where $C, c > 0$ depend (polynomially) only on the sub-Gaussian moment.*

Now we can complete the proof of Proposition 3.6. Note that the entries of $U \in R^{d\times\hat{r}}$ are independently drawn from $N(0, \frac{1}{\hat{r}})$ and $\boldsymbol{V}_{\boldsymbol{X}_s} \in R^{d\times s}$ is an orthonormal matrix. We write $\boldsymbol{V}_{\boldsymbol{X}_s}^+ \in R^{d\hat{r}\times s\hat{r}}$ as a block diagonal matrix with $\hat{r}$ copies of $\boldsymbol{V}_{\boldsymbol{X}_s}$ on the diagonal, and $\mathrm{vec}(\boldsymbol{U}) \in R^{d\hat{r}}$ be a vector formed by the concatenation of the columns of U. Then $\boldsymbol{V}_{\boldsymbol{X}_s}^+$ is still orthonormal, and $\mathrm{vec}(\boldsymbol{U}) \sim \mathcal{N}(0, \frac{1}{\hat{r}}\boldsymbol{I})$. Since multivariate Gaussian distributions are invariant under orthonormal transformations, we deduce that $(\boldsymbol{V}_{\boldsymbol{X}_s}^+)^\top\mathrm{vec}(\boldsymbol{U}) \sim \mathcal{N}(0, \frac{1}{\hat{r}}\boldsymbol{I})$. Equivalently, the entries of $\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}$ are i.i.d. $\mathcal{N}(0, \frac{1}{\hat{r}})$.

The matrix $\sqrt{\hat{r}}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}$ satisfies all the conditions in Proposition A.9. Thus, with probability at least $1-(C\epsilon)^{\hat{r}-s+1}-e^{-c\hat{r}}$, we have $\sigma_{\min}(\sqrt{\hat{r}}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}) \geq \epsilon(\sqrt{\hat{r}}-\sqrt{s-1})$, or equivalently $\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}) \geq \epsilon\frac{\sqrt{\hat{r}}-\sqrt{s-1}}{\sqrt{\hat{r}}}$. Finally, the conclusion follows from a union bound:

$$\begin{aligned}&\mathbb{P}\left[\exists 1 \leq s \leq \hat{r}\wedge r_* \text{ s.t. } \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}\right) < \frac{\epsilon}{2\hat{r}}\right] \\ \leq &\sum_{s=1}^{\hat{r}\wedge r_*} \mathbb{P}\left[\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}\right) < \epsilon\frac{\sqrt{\hat{r}}-\sqrt{s-1}}{\sqrt{r}}\right] \leq \sum_{s=1}^{\hat{r}\wedge r_*}\left(e^{-c\hat{r}}+(C\epsilon)^{\hat{r}-s+1}\right) \leq r\left(e^{-c\hat{r}}+C\epsilon\right).\end{aligned} \tag{15}$$

A.5. Procrustes Distance

Procrustes distance is introduced in Section 3.3. The following characterization of the optimal $\boldsymbol{R}$ in Definition 3.8 is known in the literature (see e.g. Tu et al., 2016, Section 5.2.1) but we provide a proof for completeness.

Lemma A.10. *Let $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{R}^{d\times r}$ where $r \leq d$. Then for any orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{r\times r}$ that minimizes $\|\boldsymbol{U}_1-\boldsymbol{U}_2\boldsymbol{R}\|_F$ (*i.e.*, any orthogonal $\boldsymbol{R}$ s.t. $\|\boldsymbol{U}_1-\boldsymbol{U}_2\boldsymbol{R}\|_F = \operatorname{dist}(\boldsymbol{U}_1,\boldsymbol{U}_2)$), $\boldsymbol{U}_1^\top\boldsymbol{U}_2\boldsymbol{R}$ is a symmetric positive semi-definite matrix.*

Proof. We only need to consider the case when $\boldsymbol{U}_2^\top\boldsymbol{U}_1 \neq \boldsymbol{0}$. Observe that

$$\begin{aligned}\|\boldsymbol{U}_1-\boldsymbol{U}_2\boldsymbol{R}\|_F^2 &= \|\boldsymbol{U}_1\|_F^2 + \|\boldsymbol{U}_2\boldsymbol{R}\|_F^2 - 2\operatorname{tr}\left(\boldsymbol{R}^\top\boldsymbol{U}_2^\top\boldsymbol{U}_1\right) \\ &= \|\boldsymbol{U}_1\|_F^2 + \|\boldsymbol{U}_2\|_F^2 - 2\operatorname{tr}\left(\boldsymbol{R}^\top\boldsymbol{U}_2^\top\boldsymbol{U}_1\right).\end{aligned}$$

Let $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^\top$ be the SVD of $\boldsymbol{U}_2^\top\boldsymbol{U}_1$, where $\boldsymbol{A}^\top\boldsymbol{A}=\boldsymbol{I}, \boldsymbol{B}^\top\boldsymbol{B}=\boldsymbol{I}$ and $\boldsymbol{\Sigma}\succ\boldsymbol{0}$. Then

$$\operatorname{tr}\left(\boldsymbol{R}^\top\boldsymbol{U}_2^\top\boldsymbol{U}_1\right) = \operatorname{tr}\left(\boldsymbol{B}^\top\boldsymbol{R}^\top\boldsymbol{A}\boldsymbol{\Sigma}\right) \leq \left\|\boldsymbol{B}^\top\boldsymbol{R}^\top\boldsymbol{A}\right\|\operatorname{tr}\left(\boldsymbol{\Sigma}\right) = \operatorname{tr}\left(\boldsymbol{\Sigma}\right),$$

where the final step is due to orthogonality of $\boldsymbol{B}^\top\boldsymbol{R}^\top\boldsymbol{A} \in \mathbb{R}^{s\times s}$, and equality holds if and only if $\boldsymbol{B}^\top\boldsymbol{R}^\top\boldsymbol{A}=\boldsymbol{I}$. Let $\boldsymbol{C}=\boldsymbol{R}^\top\boldsymbol{A}$. Let $\boldsymbol{b}_i, \boldsymbol{c}_i \in \mathbb{R}^d$ be the i-th column of $\boldsymbol{B}$ and $\boldsymbol{C}$ respectively, then $\boldsymbol{B}^\top\boldsymbol{C}=\boldsymbol{I}$ implies that $\boldsymbol{b}_i^\top\boldsymbol{c}_i=1$. Note that $\|\boldsymbol{b}_i\|_2 = \|\boldsymbol{c}_i\|_2 = 1$, so we must have $\boldsymbol{b}_i=\boldsymbol{c}_i$ for all i, *i.e.*, $\boldsymbol{B}=\boldsymbol{C}=\boldsymbol{R}^\top\boldsymbol{A}$. Therefore, $\boldsymbol{U}_1^\top\boldsymbol{U}_2\boldsymbol{R} = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{A}^\top\boldsymbol{R} = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top$, which implies that $\boldsymbol{U}_1^\top\boldsymbol{U}_2\boldsymbol{R}$ is symmetric and positive semi-definite. □

B. Main idea for the proof of Theorem 4.1

In this section, we briefly introduce our main ideas for proving Theorem 4.1. Motivated by Stöger & Soltanolkotabi (2021), we decompose the matrix $\boldsymbol{U}_t$ into a parallel component and an orthogonal component. Specifically, we write

$$\boldsymbol{U}_t = \underbrace{\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top}_{\text{parallel component}} + \underbrace{\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top}_{\text{orthogonal component}}, \tag{16}$$

where $\boldsymbol{W}_t := \boldsymbol{W}_{\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t} \in \mathbb{R}^{\hat{r}\times s}$ is the matrix consisting of the right singular vectors of $\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t$ (Definition 3.1) and $\boldsymbol{W}_{t,\perp} \in \mathbb{R}^{\hat{r}\times(\hat{r}-s)}$ is an orthogonal complement of $\boldsymbol{W}_t$. Our goal is to prove that at some time t, we have $\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right) \approx 0$ and $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \approx 0$. As we will see later, these imply that $\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\| \approx 0$. In the remaining part of this section we give a heuristic explanation for considering (16).

Additional Notations. Let $\boldsymbol{V}_{\boldsymbol{X}_s,\perp} \in \mathbb{R}^{d\times(d-s)}$ be an orthogonal complement of $\boldsymbol{V}_{\boldsymbol{X}_s} \in \mathbb{R}^{d\times s}$. Let $\boldsymbol{\Sigma}_s = \operatorname{diag}(\sigma_1,\ldots,\sigma_s)$ and $\boldsymbol{\Sigma}_{s,\perp} = \operatorname{diag}(\sigma_{s+1},\ldots,\sigma_r,0,\cdots,0) \in \mathbb{R}^{(d-s)\times(d-s)}$. We use $\boldsymbol{\Delta}_t := (\mathcal{A}^*\mathcal{A}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)$ to denote the vector consisting of measurement errors for $\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top$.

B.1. Heuristic explanations of the decomposition

A simple and intuitive approach for showing the implicit low rank bias is to directly analyze the growth of $\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t$ versus $\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{U}_t$. Ideally, the former grows faster than the latter, so that GD only learns the components in $\boldsymbol{X}_s$.

By the update rule of GD (3),

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1} &= \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \left[\boldsymbol{I} + \mu \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right]\boldsymbol{U}_t \\ &= \underbrace{\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \left[\boldsymbol{I} + \mu \boldsymbol{X}\boldsymbol{X}^\top - \mu \boldsymbol{U}_t\boldsymbol{U}_t^\top\right]\boldsymbol{U}_t}_{=:\boldsymbol{G}_{t,1}} + \mu \underbrace{\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{\Delta}_t \boldsymbol{U}_t}_{=:\boldsymbol{G}_{t,2}} \\ &= \boldsymbol{G}_{t,1} + \mu \boldsymbol{G}_{t,2}.\end{aligned}$$

For the first term $\boldsymbol{G}_{t,1}$, we have

$$\begin{aligned}\boldsymbol{G}_{t,1} &= (\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2)\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t - \mu \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{U}_t \\ &= (\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2)\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t(\boldsymbol{I} - \mu\boldsymbol{U}_t\boldsymbol{U}_t^\top) + \mathcal{O}(\mu^2),\end{aligned}$$

where the last term $\mathcal{O}(\mu^2)$ is negligible when μ is sufficiently small. Since $\|\boldsymbol{\Sigma}_{s,\perp}\| = \sigma_{s+1}$, the spectral norm of $\boldsymbol{G}_{t,1}$ can be bounded by

$$\begin{aligned}\|\boldsymbol{G}_{t,1}\| &\le \|\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2\| \cdot \|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\| \cdot \|\boldsymbol{I} - \mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\| + \mathcal{O}(\mu^2) \\ &\le (1 + \mu\sigma_{s+1}^2)\|\boldsymbol{V}_{X_s,\perp}^\top \boldsymbol{U}_t\| + \mathcal{O}(\mu^2).\end{aligned}$$

However, the main difference with the full-observation case (Jiang et al., 2022) is the second term $\boldsymbol{G}_{t,2} := \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{\Delta}_t \boldsymbol{U}_t$. Since the measurement errors $\boldsymbol{\Delta}_t$ are small but arbitrary, it is hard to compare this term with $\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}$. As a result, we cannot directly bound the growth of $\|\boldsymbol{V}_{X_s,\perp}^\top \boldsymbol{U}_t\|$.

However, the aforementioned problem disappears if we turn to bound the growth of $\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\|$. To see this, first we deduce the following by repeatedly using $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = 0$ due to the definition of $\boldsymbol{W}_{t,\perp}$.

$$\begin{aligned}\boldsymbol{G}_{t,1}\boldsymbol{W}_{t,\perp} &= \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \left[\boldsymbol{I} + \mu\boldsymbol{X}\boldsymbol{X}^\top - \mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\right]\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} \\ &= \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top (\boldsymbol{I} + \mu\boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} - \mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} \\ &= (\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2)\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} - \mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t(\boldsymbol{W}_t\boldsymbol{W}_t^\top + \boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top)\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} \\ &= (\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2)\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}(\boldsymbol{I} - \mu\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}) \\ &\quad - \mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \mathcal{O}(\mu^2),\end{aligned}$$

$$\boldsymbol{G}_{t,2}\boldsymbol{W}_{t,\perp} = \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{\Delta}_t\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{\Delta}_t\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp},$$

So we have the following recursion:

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp} &= (\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2 + \mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{\Delta}_t\boldsymbol{V}_{\boldsymbol{X}_s,\perp})\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}(\boldsymbol{I} - \mu\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}) \\ &\quad - \mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \mathcal{O}(\mu^2),\end{aligned}$$

We further note that

$$\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp} = \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1,\perp} + \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp}, \tag{17}$$

which establishes the relationship between $\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}$ and $\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}$. To complete the proof we need to prove the following:

- The minimal eigenvalue of the *parallel component* $\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top$ grows at a linear rate with speed strictly faster than σ_{s+1}.
- The term $\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \ll 1$, which implies that the first term in (17) is negligible.

C. Proof of Lemma 4.3

In this section, we give the full proof of Lemma 4.3, with some additional technical lemmas left to Appendix D.

Lemma 4.3. *Under Assumptions 3.3 and 3.5, there exists $\hat{T}_\alpha^s > 0$ for all $\alpha > 0$ and $1 \le s \le \hat{r} \wedge r_*$ such that $\lim_{\alpha\to 0} \max_{1\le t\le \hat{T}_\alpha^s} \sigma_{s+1}(\boldsymbol{U}_{\alpha,t}) = 0$. Furthermore, it holds that $\left\|\boldsymbol{U}_{\hat{T}_\alpha^s}\boldsymbol{U}_{\hat{T}_\alpha^s}^\top - \boldsymbol{Z}_s^*\right\|_F = \mathcal{O}\left(\kappa_*^3 r_* \delta \|\boldsymbol{X}\|^2\right)$.*

Appendices C.1 and C.2 are devoted to analyzing the spectral phase and parallel improvement phase, respectively. Appendix C.3 uses induction to characterize the low-rank GD trajectory in the parallel improvement phase. In Appendix C.4 we study the refinement phase, which allows us to derive Lemma 4.3.

C.1. The spectral phase

Starting from a small initialization $\boldsymbol{U}_0 = \alpha\bar{\boldsymbol{U}}, \alpha \ll 1$, we first enter the spectral phase where GD behaves similar to power iteration. As in Stöger & Soltanolkotabi (2021), we refer to this phase as the spectral phase. Specifically, we have in the spectral phase that

$$\boldsymbol{U}_{t+1} = \left(\boldsymbol{I} + \mu\left(\mathcal{A}^*\mathcal{A}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right)\boldsymbol{U}_t \approx \left(\boldsymbol{I} + \mu\left(\mathcal{A}^*\mathcal{A}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top\right)\right)\boldsymbol{U}_t.$$

The approximation holds with high accuracy as long as $\|\boldsymbol{U}_t\| \ll 1$. Moreover we have $\boldsymbol{M} := \left(\mathcal{A}^*A\right)\left(\boldsymbol{X}\boldsymbol{X}^\top\right) \approx \boldsymbol{X}\boldsymbol{X}^\top$ by the RIP condition; when δ is sufficiently small, we can still ensure a positive eigen-gap of $\boldsymbol{M}$. As a result, with small initialization $\boldsymbol{U}_t$ would become approximately aligned with the top eigenvector u_1 of $\boldsymbol{M}$. Since $\|\boldsymbol{M} - \boldsymbol{X}\boldsymbol{X}^\top\| = \mathcal{O}(\delta\sqrt{r_*})$ by Proposition A.2, we have $\|\boldsymbol{u}_1 - \boldsymbol{v}_1\| = \mathcal{O}(\delta\sqrt{r_*})$ so that $\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\| = \mathcal{O}(\delta\sqrt{r_*})$. This proves the base case for the induction.

Formally, we define $\boldsymbol{M} = \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top)$, $\boldsymbol{K}_t = (\boldsymbol{I} + \mu\boldsymbol{M})^t$ and $\boldsymbol{U}_t^{\mathtt{sp}} = \boldsymbol{K}_t\boldsymbol{U}_0$. Suppose that $\boldsymbol{M} = \sum_{i=1}^{\mathrm{rank}(\boldsymbol{M})} \hat{\sigma}_i^2 \hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i^\top$ is the spectral decomposition of $\boldsymbol{M}$ where $\{\hat{\sigma}_i\}_{i\ge 1}$ is sorted in non-increasing order. We additionally define $\boldsymbol{M}_s = \sum_{i=1}^{\min\{s,\mathrm{rank}(\boldsymbol{M})\}} \hat{\sigma}_i^2\hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i^\top$. By Lemma A.6 and $\delta\sqrt{r_*} \le 10^{-3}\kappa_*$ by Assumption 3.3, we have $\hat{\sigma}_s^2 \ge \sigma_s^2 - 0.01\tau$ and $\hat{\sigma}_{s+1}^2 \le \sigma_{s+1}^2 + 0.01\tau$, where we recall that $\tau = \min_{s\in[r_*]}\left(\sigma_s^2 - \sigma_{s+1}^2\right) > 0$. Additionally, let $\boldsymbol{L}_t$ be the span of the top-s left singular vectors of $\boldsymbol{U}_t$. Recall that Assumption 3.5 is made on the initialization. Let

$$t^\star := \min\left\{i \in \mathbb{N} : \left\|\boldsymbol{U}_{i-1}^{\mathtt{sp}} - \boldsymbol{U}_{i-1}\right\| > \left\|\boldsymbol{U}_{i-1}^{\mathtt{sp}}\right\|\right\},$$

the following lemma bounds the error of approximating $\boldsymbol{U}_t$ via $\boldsymbol{U}_t^{\mathtt{sp}}$:

Lemma C.1. *(Stöger & Soltanolkotabi, 2021, Lemma 8.1) Suppose that $\mathcal{A}$ satisfies the rank-1 RIP with constant δ_1. For all integers t such that $1 \le t \le t^\star$ it holds that*

$$\|\boldsymbol{E}_t\| = \|\boldsymbol{U}_t - \boldsymbol{U}_t^{\mathtt{sp}}\| \le 4\hat{\sigma}_1^{-2}\alpha^3 r_*\left(1+\delta_1\right)\left(1+\mu\hat{\sigma}_1^2\right)^{3t}. \tag{18}$$

We can derive the following lower bound on t^* from Lemma C.1.

Corollary C.2. *We have*

$$t^* \ge \frac{\log\alpha^{-1} + \frac{1}{2}\log\frac{\rho\hat{\sigma}_1^2}{4(1+\delta_1)r_*}}{\log\left(1+\mu\hat{\sigma}_1^2\right)}.$$

Proof. By Lemma C.1 we have

$$\|\boldsymbol{E}_t\| \le 4\hat{\sigma}_1^{-2}\alpha^3 r_*\left(1+\delta_1\right)\left(1+\mu\hat{\sigma}_1^2\right)^{3t}.$$

for all $t \le t^*$. On the other hand, we have

$$\begin{aligned}\|\boldsymbol{U}_t^{\mathtt{sp}}\| &= \alpha\left\|(\boldsymbol{I}+\mu\boldsymbol{M})^t\bar{\boldsymbol{U}}\right\| \\ &\ge \alpha(1+\mu\hat{\sigma}_1^2)^t\left\|\hat{\boldsymbol{v}}_1\hat{\boldsymbol{v}}_1^\top\bar{\boldsymbol{U}}\right\| \\ &\ge \left(1+\mu\hat{\sigma}_1^2\right)^t\alpha\rho.\end{aligned}$$

Thus, it follows from $\|\boldsymbol{E}_{t^*}\| \ge \|\boldsymbol{U}_{t^*}^{\mathtt{sp}}\|$ that

$$\left(1+\mu\hat{\sigma}_1^2\right)^{t^*} \ge \sqrt{\frac{\rho\hat{\sigma}_1^2}{4(1+\delta_1)r_*}}\cdot\alpha^{-1} \Rightarrow t^* \ge \frac{\log\alpha^{-1} + \frac{1}{2}\log\frac{\rho\hat{\sigma}_1^2}{4(1+\delta_1)r_*}}{\log\left(1+\mu\hat{\sigma}_1^2\right)}$$

as desired. □

Note that a trivial bound for the rank-1 RIP constant is $\delta_1 \le \delta$. We can now show that for small t, GD can be viewed as approximate power iteration.

Lemma C.3. *There exists a time*

$$t = T_\alpha^{\mathtt{sp}} := \frac{2\log\alpha^{-1} + \log\frac{\rho\hat{\sigma}_1^2}{4r_*(1+\delta)}}{3\log(1+\mu\hat{\sigma}_1^2) - \log(1+\mu\hat{\sigma}_{s+1}^2)} \le t^*$$

such that

$$\left\| \boldsymbol{U}_t - \sum_{i=1}^{s} \alpha(1+\mu\hat{\sigma}_i^2)^t \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top \bar{\boldsymbol{U}} \right\| \le C_1 \cdot \alpha^\gamma$$

where $\gamma = 1 - \frac{2\log(1+\mu\hat{\sigma}_{s+1}^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}$ *and* $C_1 = C_1(\boldsymbol{X}, \bar{\boldsymbol{U}})$ *is a constant that only depends on* $\boldsymbol{X}$ *and* $\bar{\boldsymbol{U}}$.

Proof. It's easy to check that $T_\alpha^{\mathtt{sp}} \le t^*$ by applying Corollary C.2.

We consider the following decomposition:

$$\left\| \boldsymbol{U}_t - \sum_{i=1}^{s} \alpha(1+\mu\hat{\sigma}_i^2)^t \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top \bar{\boldsymbol{U}} \right\| \le \|\boldsymbol{U}_t - \boldsymbol{U}_t^{\mathtt{sp}}\| + \left\| \boldsymbol{U}_t^{\mathtt{sp}} - \sum_{i=1}^{s} \alpha(1+\mu\hat{\sigma}_i^2)^t \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top \bar{\boldsymbol{U}} \right\|.$$

When $t \le t^*$, the first term can be bounded as

$$\|\boldsymbol{E}_t\| \le 4\hat{\sigma}_1^{-2}\alpha^3 r_* (1+\delta)\left(1+\mu\hat{\sigma}_1^2\right)^{3t}.$$

For the second term we have

$$\left\| \boldsymbol{U}_t^{\mathtt{sp}} - \sum_{i=1}^{s} \alpha(1+\mu\hat{\sigma}_i^2)^t \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top \boldsymbol{U} \right\| \le \left\| \sum_{i=s+1}^{r_*} \alpha(1+\mu\hat{\sigma}_i^2)^t \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top \boldsymbol{U} \right\| \le \alpha\left(1+\mu\hat{\sigma}_{s+1}^2\right)^t.$$

In particular, the definition of $T_\alpha^{\mathtt{sp}}$ implies that

$$\begin{aligned}
\left\| \boldsymbol{U}_t - \sum_{i=1}^{s} \alpha(1+\mu\hat{\sigma}_i^2)^t \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top \boldsymbol{U} \right\| &\le 2\left(\frac{\rho\hat{\sigma}_1^2}{4r_*(1+\delta)}\right)^{\frac{1-\gamma}{2}} \alpha^\gamma \\
&\le 2\max\left\{1, \frac{\rho\hat{\sigma}_1^2}{4r_*(1+\delta)}\right\}\alpha^\gamma \\
&\le \underbrace{\max\left\{2, \frac{\rho\sigma_1^2}{r_*}\right\}}_{:=C_1}\alpha^\gamma.
\end{aligned}$$

as desired. □

We conclude this section with the following lemma, which states that initially the parallel component $\boldsymbol{U}_t\boldsymbol{W}_t$ would grow much faster than the noise term, and would become well-aligned with $\boldsymbol{X}_s$.

Lemma C.4 (Lemma 5.1, formal version)**.** *There exists positive constants* $C_2 = C_2(\boldsymbol{X}, \bar{\boldsymbol{U}})$ *and* $C_3 = C_3(\boldsymbol{X}, \bar{\boldsymbol{U}})$ *such that the following inequalities hold for* $t = T_\alpha^{\mathtt{sp}}$ *when* $\alpha \in \left(0, \left(\frac{\rho}{10C_1(\boldsymbol{X},\bar{\boldsymbol{U}})}\right)^{10\kappa_*}\right)$*:*

$$\|\boldsymbol{U}_t\| \le \|\boldsymbol{X}\| \tag{19a}$$

$$\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t) \ge C_2 \cdot \alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_s^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \tag{19b}$$

$$\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \le C_3 \cdot \alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_{s+1}^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \tag{19c}$$

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \le 200\delta \tag{19d}$$

Proof. We prove this lemma by applying Corollary D.3 to $t = T_\alpha^{\mathtt{sp}}$ defined in the previous lemma.

The inequality (19a) can be directly verified by using Lemma C.3:

$$\|\boldsymbol{U}_t\| \le \alpha\left(1+\mu\hat{\sigma}_1^2\right)^t + \alpha^\gamma \le \left(1+\left(\frac{C_1(\boldsymbol{X},\bar{\boldsymbol{U}})\hat{\sigma}_1^2}{4r_*(1+\delta)}\right)^{\frac{1}{3}}\right)\cdot\alpha^{\gamma/3} \le \hat{C}_1(\boldsymbol{X},\bar{\boldsymbol{U}})\cdot\alpha^{\gamma/3}\|\boldsymbol{X}\|.$$

where $\hat{C}_1(\boldsymbol{X},\bar{\boldsymbol{U}}) = 1+\left(\frac{C_1(\boldsymbol{X},\bar{\boldsymbol{U}})\|\boldsymbol{X}\|^2}{2r_*(1+\delta)}\right)^{\frac{1}{3}}$ (the constant C_1 is defined in the previous lemma). The last inequality holds when α is sufficiently small. For the remaining inequalities, we first verify that the assumption in Corollary D.3:

$$\alpha\sigma_s(\boldsymbol{K}_t) > 10\left(\alpha\sigma_{s+1}(\boldsymbol{K}_t) + \|\boldsymbol{E}_t\|\right). \tag{20}$$

By definition of $\boldsymbol{K}_t$, we can see that for $\alpha \le \left(\frac{\rho}{10C_1}\right)^{10\kappa_*}$,

$$\begin{aligned}
\alpha\sigma_{s+1}(\boldsymbol{K}_t) + \|\boldsymbol{E}_t\| &\le \alpha\left(1+\mu\hat{\sigma}_{s+1}^2\right)^t + 4\hat{\sigma}_1^{-2}\alpha^3 r_*(1+\delta)\left(1+\mu\hat{\sigma}_1^2\right)^{3t} \\
&\le C_1(\boldsymbol{X},\bar{\boldsymbol{U}})\cdot\alpha^\gamma \le 0.1\rho\alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_s^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \\
&\le 0.1\alpha\sigma_s(\boldsymbol{K}_t)
\end{aligned}$$

where $\|\boldsymbol{E}_{T_\alpha^{\mathtt{sp}}}\|$ is bounded in the previous lemma. Hence (20) holds. Let $\boldsymbol{L}$ be the span of top-s eigenvectors of $\boldsymbol{M}$, then by Corollary D.3, at $t = T_\alpha^{\mathtt{sp}}$ we have

$$\begin{aligned}
\sigma_s\left(\boldsymbol{U}_t\boldsymbol{W}_t\right) &\geqslant 0.4\alpha\sigma_s\left(\boldsymbol{K}_t\right)\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{L}}^\top\bar{\boldsymbol{U}}\right) \\
&\geq 0.1\alpha\rho\left(1+\mu\hat{\sigma}_s^2\right)^t \\
&= 0.1\rho\left(\frac{\rho\hat{\sigma}_1^2}{4r_*(1+\delta)}\right)^{\frac{\log\left(1+\mu\hat{\sigma}_s^2\right)}{3\log\left(1+\mu\hat{\sigma}_1^2\right)-\log\left(1+\mu\hat{\sigma}_{s+1}^2\right)}}\alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_s^2)}{3\log(1+\mu\hat{\sigma}_1)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \\
&\geq \underbrace{0.1\rho\left(\frac{\rho\sigma_1^2}{8r_*}\right)^{\frac{1}{10\kappa_*}}}_{:=C_2(\boldsymbol{X},\bar{\boldsymbol{U}})}\alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_s^2)}{3\log(1+\mu\hat{\sigma}_1)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \\
\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| &\leqslant 2\alpha\sigma_{s+1}^2\left(\boldsymbol{K}_t\right) + \|\boldsymbol{E}_t\| \\
&\le \underbrace{2C_1(\boldsymbol{X},\bar{\boldsymbol{U}})}_{:=C_3(\boldsymbol{X},\bar{\boldsymbol{U}})}\alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_{s+1}^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \\
\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| &\leqslant 100\left(\delta+\frac{\alpha\sigma_{s+1}\left(\boldsymbol{K}_t\right)+\|\boldsymbol{E}_t\|}{\alpha\rho\sigma_s\left(\boldsymbol{K}_t\right)}\right) \\
&\le 100\left(\delta\alpha^{\frac{2\log(1+\mu\hat{\sigma}_s^2)-2\log(1+\mu\hat{\sigma}_{s+1}^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}}\right) \\
&\le 200\delta.
\end{aligned} \tag{21}$$

The conclusion follows. □

C.2. The parallel improvement phase

This subsection is devoted to proving Lemma 5.2 which we recall below.

Lemma 5.2. *Suppose that Assumptions 3.3, 3.5 and 3.7 hold and let $c_3 = 10^4\kappa_* r_*^{\frac{1}{2}}\delta$. Then for sufficiently small α, the following inequalities hold when $T_\alpha^{\mathtt{sp}} \le t < T_{\alpha,s}^{\mathtt{pi}}$:*

$$\begin{aligned}
&\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right) \geq \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\right) \\
&\quad \geq \left(1+0.5\mu\left(\sigma_s^2+\sigma_{s+1}^2\right)\right)\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\right), &\text{(13a)}\\
&\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \\
&\quad \leq \left(1+\mu\left(0.4\sigma_s^2+0.6\sigma_{s+1}^2\right)\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|, &\text{(13b)}\\
&\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1}}\right\| \leq c_3, &\text{(13c)}\\
&\operatorname{rank}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}) = \operatorname{rank}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t) = s. &\text{(13d)}
\end{aligned}$$

C.2.1. THE PARALLEL COMPONENT

In the following we bound $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\right)$. We state our main result of this section in the lemma below.

Lemma C.5. *Suppose that Assumptions 3.3, 3.5 and 3.7 holds, $\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\| \leq c_3 < 10^{-2}\kappa^{-1}$ and $\boldsymbol{\Delta}_t = \left(\mathcal{A}^*\mathcal{A}-\boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)$ satisfies $\|\boldsymbol{\Delta}_t\| \leq 0.2\kappa^{-1}r_*^{-\frac{1}{2}}\|\boldsymbol{X}\|^2$, then we have*

$$\begin{aligned}
&\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}) \geq \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t) \\
&\geq \left[1+\mu\left(\sigma_s^2-5c_3\|\boldsymbol{X}\|^2-2\|\boldsymbol{\Delta}_t\|\right)-20\mu^2\|\boldsymbol{X}\|^4\right]\left(1-\mu\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\right)\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t).
\end{aligned}$$

Proof. The update rule of GD implies that

$$\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t \\
&= \boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{I}+\mu(\boldsymbol{X}_s\boldsymbol{X}_s^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top)+\mu\boldsymbol{\Delta}_t\right)\boldsymbol{U}_t\boldsymbol{W}_t &\text{(22a)}\\
&= (\boldsymbol{I}+\mu\boldsymbol{\Sigma}_s^2)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t - \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{U}_t\boldsymbol{W}_t + \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{\Delta}_t\boldsymbol{U}_t\boldsymbol{W}_t &\text{(22b)}\\
&= (\boldsymbol{I}+\mu\boldsymbol{\Sigma}_s^2)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t - \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t - \mu\boldsymbol{V}_{\boldsymbol{X}}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t \\
&\quad + \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{\Delta}_t\boldsymbol{U}_t\boldsymbol{W}_t \\
&= (\boldsymbol{I}+\mu\boldsymbol{\Sigma}_s^2)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t(\boldsymbol{I}-\mu\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t) + \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{\Delta}_t\boldsymbol{U}_t\boldsymbol{W}_t \\
&\quad - \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t + \mu^2\boldsymbol{\Sigma}_s^2\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t &\text{(22c)}
\end{aligned}$$

where (22a) follows from $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{X}_s\boldsymbol{X}_s^\top + \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{X}_{s,\perp}\boldsymbol{X}_{s,\perp}^\top$ and $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{X}_{s,\perp} = 0$; (22b) follows from $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{X}_s\boldsymbol{X}_s^\top = \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{\Sigma}_s\boldsymbol{V}_{\boldsymbol{X}_s}^\top = \boldsymbol{\Sigma}_s\boldsymbol{V}_{\boldsymbol{X}_s}^\top$, and (22c) follows from $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t = \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top + \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top = \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top$ by definition of $\boldsymbol{W}_t$ and $\boldsymbol{W}_{t,\perp}$.

We now relate the last three terms in (22c) to $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t$. Since $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t$ is invertible by Assumption 3.5, $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}, \boldsymbol{\Sigma}_{\boldsymbol{U}_t\boldsymbol{W}_t}$ and $\boldsymbol{W}_{\boldsymbol{U}_t\boldsymbol{W}_t}$ are also of full rank, thus we have

$$\begin{aligned}
\boldsymbol{U}_t\boldsymbol{W}_t &= \boldsymbol{U}_t\boldsymbol{W}_t(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t \\
&= \boldsymbol{U}_t\boldsymbol{W}_t\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{\Sigma}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{W}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t &\text{(23)}\\
&= \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t.
\end{aligned}$$

Plugging (23) into the second and third terms of (22) and re-arranging, we deduce that

$$\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t \\
&= \left(\boldsymbol{I}+\mu(\boldsymbol{\Sigma}_s^2+\boldsymbol{P}_1+\boldsymbol{P}_2)\right)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t(\boldsymbol{I}-\mu\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t) \\
&\quad + \mu^2\left(\boldsymbol{\Sigma}_s^2+\boldsymbol{P}_1+\boldsymbol{P}_2\right)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t &\text{(24)}\\
&= \left[\boldsymbol{I}+\mu\left(\boldsymbol{\Sigma}_s^2+\boldsymbol{P}_1+\boldsymbol{P}_2\right)+\mu^2\left(\boldsymbol{\Sigma}_s^2+\boldsymbol{P}_1+\boldsymbol{P}_2\right)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\left(\boldsymbol{I}-\mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-1}\right]\cdot \\
&\quad \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t(\boldsymbol{I}-\mu\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t)
\end{aligned}$$

where we use the equation $\boldsymbol{A} = (\boldsymbol{I} - \mu \boldsymbol{A}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}(\boldsymbol{I} - \mu \boldsymbol{A}^\top \boldsymbol{A})$ with $\boldsymbol{A} = \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{W}_t$ (when $\mu < \frac{1}{9\|\boldsymbol{X}\|^2}$, $\boldsymbol{I} - \mu \boldsymbol{A}\boldsymbol{A}^\top$ is invertible by Lemma D.4), and

$$\begin{aligned} \boldsymbol{P}_1 &= \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp} \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t} \left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}\right)^{-1} \\ \boldsymbol{P}_2 &= \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{\Delta}_t \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t} \left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}\right)^{-1} \end{aligned} \tag{25}$$

By assumption we have

$$\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}\right) \geq \sqrt{1 - \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}\right\|^2} \geq \frac{1}{2},$$

so that

$$\|\boldsymbol{P}_1\| \leq \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\right\| \cdot \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}\right\| \cdot \left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}\right)^{-1}\right\| \leq 5c_3\|\boldsymbol{X}\|^2 \leq 0.1\|\boldsymbol{X}\|^2 \tag{26}$$

and by our assumption we have

$$\|\boldsymbol{P}_2\| \leq \left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}\right)^{-1}\right\| \cdot \|\boldsymbol{\Delta}_t\| \leq 2\|\boldsymbol{\Delta}_t\| \leq 0.2\kappa^{-1} r_*^{-\frac{1}{2}} \|\boldsymbol{X}\|^2. \tag{27}$$

Moreover, note that $\|\boldsymbol{\Sigma}_s\|^2 = \sigma_1^2 = \|\boldsymbol{X}\|^2$, and since $\mu < 10^{-4}\|\boldsymbol{X}\|^{-2}$ by Assumption 3.7, we have $\left\|\left(\boldsymbol{I} - \mu \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{W}_t \boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-1}\right\| < 1.1$. Thus

$$\left\|\left(\boldsymbol{\Sigma}_s^2 + \boldsymbol{P}_1 + \boldsymbol{P}_2\right) \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{W}_t \boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s} \left(\boldsymbol{I} - \mu \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{W}_t \boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-1}\right\| \leq 20\|\boldsymbol{X}\|^4.$$

The equation (25) implies that

$$\begin{aligned} &\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1} \boldsymbol{W}_t) \\ &\geq \sigma_{\min}\left(\boldsymbol{I} + \mu\left(\boldsymbol{\Sigma}_s^2 + \boldsymbol{P}_1 + \boldsymbol{P}_2\right) + \left(\boldsymbol{\Sigma}_s^2 + \boldsymbol{P}_1 + \boldsymbol{P}_2\right) \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{W}_t \boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s} \left(\boldsymbol{I} - \mu \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{W}_t \boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-1}\right) \cdot \\ &\quad \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{W}_t (\boldsymbol{I} - \mu \boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s} \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{W}_t)\right) \\ &\geq \left(1 + \mu\sigma_{\min}^2(\boldsymbol{\Sigma}_s) - \mu\|\boldsymbol{P}_1\| - \mu\|\boldsymbol{P}_2\| - 20\mu^2\|\boldsymbol{X}\|^4\right) \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t) \left(1 - \mu\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\right) \\ &= \left(1 + \mu\sigma_s^2 - \mu\|\boldsymbol{P}_1\| - \mu\|\boldsymbol{P}_2\| - 20\mu^2\|\boldsymbol{X}\|^4\right) \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t) \left(1 - \mu\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\right) \end{aligned}$$

Recall that $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are bounded in (26) and (27) respectively, so we have that

$$\begin{aligned} &\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}) \\ &\geq \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1} \boldsymbol{W}_t) \\ &\geq \left[1 + \mu\left(\sigma_s^2 - 5c_3\|\boldsymbol{X}\|^2 - 2\|\boldsymbol{\Delta}_t\|^2\right) - 20\mu^2\|\boldsymbol{X}\|^4\right] \left(1 - \mu\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\right) \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t). \end{aligned}$$

The conclusion follows. □

The corollaries below immediately follow from Lemma C.5.

Corollary C.6. *Under the conditions in Lemma C.5, if $\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t) < 0.3\kappa^{-1}\|\boldsymbol{X}\|^2$, then we have*

$$\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}) \geq \left(1 + 0.5\mu(\sigma_s^2 + \sigma_{s+1}^2)\right) \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t).$$

Proof. By Lemma C.5 it remains to check that

$$\left[1 + \mu\left(\sigma_s^2 - 5c_3\|\boldsymbol{X}\|^2 - 2\|\boldsymbol{\Delta}_t\|^2\right) - 20\mu^2\|\boldsymbol{X}\|^4\right] \left(1 - 0.3\mu\kappa^{-1}\|\boldsymbol{X}\|^2\right) \geq 1 + 0.5\mu(\sigma_s^2 + \sigma_{s+1}^2).$$

Indeed, recall from the conditions of Lemma C.5 that $5c_3\|\boldsymbol{X}\|^2 \leq 0.01\kappa^{-1}\|\boldsymbol{X}\|^2 \leq 0.01(\sigma_s^2 - \sigma_{s+1}^2)$ and similarly $\|\boldsymbol{\Delta}_t\| \leq 0.005(\sigma_s^2 - \sigma_{s+1}^2)$ and $\mu^2\|\boldsymbol{X}\|^4 \leq 10^{-4}\kappa^{-1}\mu\|\boldsymbol{X}\|^2 \leq 10^{-4}(\sigma_s^2 - \sigma_{s+1}^2)$, so that

$$\begin{aligned} &\left[1 + \mu\left(\sigma_s^2 - 5c_3\|\boldsymbol{X}\|^2 - 2\|\boldsymbol{\Delta}_t\|^2\right) - 20\mu^2\|\boldsymbol{X}\|^4\right] \left(1 - 0.3\kappa^{-1}\|\boldsymbol{X}\|^2\right) \\ &\geq \left(1 + \mu(0.9\sigma_s^2 + 0.1\sigma_{s+1}^2)\right) \left(1 - 0.3\mu(\sigma_s^2 - \sigma_{s+1}^2)\right) \\ &= 1 + \mu(0.6\sigma_s^2 + 0.4\sigma_{s+1}^2) - \mu^2\|\boldsymbol{X}\|^2(\sigma_s^2 - \sigma_{s+1}^2) \\ &\geq 1 + 0.5\mu(\sigma_s^2 + \sigma_{s+1}^2) \end{aligned}$$

as desired. □

Corollary C.7. *Under the conditions in Lemma C.5, if*

$$\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t) \le \sigma_s^2 - \mu\sigma_s^4 - 5c_3\|\boldsymbol{X}\|^2 - 2\|\boldsymbol{\Delta}_t\|^2 - 20\mu^2\|\boldsymbol{X}\|^4, \tag{28}$$

then we have that $\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}) \ge \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)$.

Proof. A sufficient condition for $\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}) \ge \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)$ to hold is that

$$\begin{aligned}
&\left[1 + \mu\left(\sigma_s^2 - 5c_3\|\boldsymbol{X}\|^2 - 2\|\boldsymbol{\Delta}_t\|^2\right) - 20\mu^2\|\boldsymbol{X}\|^4\right]\left(1 - \mu\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\right)\\
\Leftarrow\ &\sigma_s^2 - 5c_3\|\boldsymbol{X}\|^2 - 2\|\boldsymbol{\Delta}_t\|^2 - 20\mu\|\boldsymbol{X}\|^2 - \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t) - \mu\sigma_s^2\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t) \ge 0.
\end{aligned}$$

When (28) holds, we have

$$\begin{aligned}
&\sigma_s^2 - 5c_3\|\boldsymbol{X}\|^2 - 2\|\boldsymbol{\Delta}_t\|^2 - 20\mu\|\boldsymbol{X}\|^2 - \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\\
\ge\ &\mu\sigma_s^4 \ge \mu\sigma_s^2\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)
\end{aligned}$$

as desired. □

C.2.2. The orthogonal component

In this section we turn to analyze the noise term.The main result of this section is presented in the following:

Lemma C.8. *Suppose that Assumptions 3.3, 3.5 and 3.7 hold,* $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t \in \mathbb{R}^{s\times r}$ *is of full rank,* $\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\| \le c_3 < 10^{-2}\kappa^{-1}$ *and* $\|\boldsymbol{\Delta}_t\| \le c_3\|\boldsymbol{X}\|^2$*, then we have*

$$\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \le \left(1 + \mu\sigma_{s+1}^2 + 30\mu\|\boldsymbol{X}\|^2 c_3 + 0.1\mu^2\|\boldsymbol{X}\|^4\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|.$$

Proof. By the definition of $\boldsymbol{W}_{t,\perp}$, we have $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = 0$, thus $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| = \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|$. The latter can be decomposed as follows:

$$\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp} = \underbrace{\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1,\perp}}_{=(a)} + \underbrace{\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp}}_{=(b)}.$$

In the following, we are going to show that the term (a) is bounded by $c\cdot\mu$ where c is a small constant, while (b) grows linearly with a slow speed.

Bounding summand (a). Since

$$0 = \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp} = \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1,\perp} + \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp}$$

by definition, we have

$$\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1,\perp} = -\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp}. \tag{29}$$

Thus the summand (a) can be rewritten as follows:

$$\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1,\perp}\\
=\ &-\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp} &\text{(30a)}\\
=\ &-\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\boldsymbol{\Sigma}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\boldsymbol{W}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp}\\
=\ &-\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp} &\text{(30b)}\\
=\ &-\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{I} + \mu\mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp}\\
=\ &-\mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left[\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right) + \boldsymbol{\Delta}_t\right]\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp} &\text{(30c)}\\
=\ &\mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left[\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{\Delta}_t\right]\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp}\\
=\ &\mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right)^{-1}\boldsymbol{M}_1\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp},
\end{aligned}$$

where $\boldsymbol{M}_1 = \boldsymbol{V}_{\boldsymbol{X}_s}^\top \left[\boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp} - \boldsymbol{\Delta}_t \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\right]$. In (30), (30a) follows from (29), (30b) holds since $\boldsymbol{\Sigma}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t} \boldsymbol{W}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}^\top \in \mathbb{R}^{s\times s}$ is invertible, and in (30c) we use $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{W}_{t,\perp} = 0$. It follows that

$$\|(a)\| \le \mu \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right\| \cdot \left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right)^{-1}\right\| \|\boldsymbol{M}_1\| \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t \boldsymbol{W}_{t,\perp}\right\|. \tag{31}$$

By Lemma D.5 we have $\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right\| \le 0.01$, which implies that

$$\left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right)^{-1}\right\| = \sigma_{\min}^{-1}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right) = \left(1 - \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right\|^2\right)^{-\frac{1}{2}} \le 1.1. \tag{32}$$

Lastly, we bound $\boldsymbol{M}_1$ as follows:

$$\begin{aligned}
\|\boldsymbol{M}_1\| &\le \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\right\| + \left\|\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\| \\
&\le \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{W}_t\right\| \cdot \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t \boldsymbol{W}_t\right\| + 10^{-3}\kappa^{-1} c_3 \|\boldsymbol{X}\|^2 \\
&\le 10\|\boldsymbol{X}\|^2 c_3.
\end{aligned} \tag{33}$$

where the second inequality follows from our assumption on $\left\|\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|$. Combining (31), (32) and (33) yields

$$\|(a)\| \le 20\mu\|\boldsymbol{X}\|^2 c_3 \|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|.$$

Bounding summand (b). This is the main component in the error term. We'll see that although this term can grow exponentially fast, the growth speed is slower than the minimal eigenvalue of the parallel component.

We have

$$\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp} \\
&= \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \left[\boldsymbol{I} + \mu(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top) + \mu\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} && \text{(34a)} \\
&= \left(\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2 - \mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp} + \mu\underbrace{\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{\Delta}_t \boldsymbol{V}_{\boldsymbol{X}_s,\perp}}_{=:\boldsymbol{M}_2}\right)\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} && \text{(34b)} \\
&= \left(\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2 - \mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp} + \mu\boldsymbol{M}_2\right)\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\left(\boldsymbol{I} - \mu\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right) && \text{(34c)} \\
&\quad + \mu^2\left(\boldsymbol{\Sigma}_{s,\perp}^2 - \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp} + \boldsymbol{M}_2\right)\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} && \text{(34d)}
\end{aligned}$$

where we recall that $\boldsymbol{\Sigma}_{s,\perp}^2 = \operatorname{diag}\left(\sigma_{s+1}^2, \cdots, \sigma_r^2, 0, \cdots, 0\right) \in \mathbb{R}^{(d-s)\times(d-s)}$. In (34), (34a) follows from the update rule of GD, (34b) is obtained from $\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{\Sigma}_{s,\perp}^2 \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top$ and $\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}$, and lastly in (34d) we use

$$\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} \\
&= \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} \\
&= \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\right\| \\
&\le \left(\left\|\boldsymbol{I} - \mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s,\perp}\right\| + \mu\|\boldsymbol{\Sigma}_{s,\perp}\|^2 + \mu\|\boldsymbol{M}_2\|\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|\left(\boldsymbol{I} - \mu\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|^2\right) \\
&\quad + \mu^2\left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|^3\left(\sigma_{s+1}^2 + \|\boldsymbol{U}_t\|^2 + 10^{-3}\kappa^{-1}c_3\|\boldsymbol{X}\|^2\right) \\
&\le \left(1 + \mu\sigma_{s+1}^2 + \mu\|\boldsymbol{\Delta}_t\|\right)\left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|\left(1 - \mu\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|^2\right) + 0.1\mu^2\|\boldsymbol{X}\|^4\left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\| \\
&\le \left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|\left(1 + \mu\sigma_{s+1}^2 + \mu c_3\|\boldsymbol{X}\|^2 + 0.1\mu^2\|\boldsymbol{X}\|^4\right)
\end{aligned}$$

To summarize, we have

$$\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \le \left(1+\mu\sigma_{s+1}^2+30\mu\|\boldsymbol{X}\|^2 c_3+0.1\mu^2\|\boldsymbol{X}\|^4\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$$

as desired. □

To bound the growth speed of the orthogonal component, we need to show that the quantity $\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|$ remains small. The following lemma serves to complete an induction step from t to $t+1$:

Lemma C.9. *Suppose $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t$ is of full rank, $\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\| \le c_3$ and $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \le \min\{\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t), c_4\}$ with $\max\left\{c_3, c_4\|\boldsymbol{X}\|^{-1}\right\} \le 10^{-2}\kappa^{-1}$, and $\boldsymbol{\Delta}_t = (\mathcal{A}^*\mathcal{A}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)$ satisfies $\|\boldsymbol{\Delta}_t\| \le 10^{-3}\kappa^{-1}c_3\|\boldsymbol{X}\|^2$ and $\mu \le 10^{-4}\kappa^{-1}\|\boldsymbol{X}\|^{-2}c_3$, then we have $\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1}}\| \le c_3$.*

Proof. Let $\boldsymbol{M}_t = \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)$, so the update rule of GD implies that

$$\begin{aligned}
\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1} &= (\boldsymbol{I}+\mu\boldsymbol{M}_t)\boldsymbol{U}_t\boldsymbol{W}_{t+1} \\
&= (\boldsymbol{I}+\mu\boldsymbol{M}_t)\left(\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1} + \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1}\right) \\
&= (\boldsymbol{I}+\mu\boldsymbol{M}_t)\left(\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}}^\top\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1} + \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1}\right) \\
&= \underbrace{(\boldsymbol{I}+\mu\boldsymbol{M}_t)(\boldsymbol{I}+\boldsymbol{P})\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}}_{:=\boldsymbol{H}}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1},
\end{aligned}$$

where

$$\boldsymbol{P} = \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1}\left(\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1}\right)^{-1}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top$$

and $\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1}$ is invertible since $\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{U}_t\boldsymbol{W}_t$ is invertible by our assumption that $\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t$ is of full rank and $\operatorname{rank}(\boldsymbol{U}_t\boldsymbol{W}_t) \ge \operatorname{rank}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{W}_t) = \operatorname{rank}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t) = s$, and $\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1}$ is invertible by Lemma D.7. Indeed, Lemma D.7 implies that $\sigma_{\min}\left(\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1}\right) \ge \frac{1}{2}$ by our condition on μ.

The key observation here is that because the (square) matrix $\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1}$ is invertible, so that the column space of $\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1}$ is the same as that of $\boldsymbol{H}$. Following the line of proof of Stöger & Soltanolkotabi, 2021, Lemma 9.3 (for completeness, we provide details in Lemma D.8), we deduce that

$$\begin{aligned}
&\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1}}\right\| = \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{H}}\boldsymbol{W}_{\boldsymbol{H}}^\top\right\| \\
&\le \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\left[\left(\boldsymbol{I}+\boldsymbol{B}-\frac{1}{2}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{B}+\boldsymbol{B}^\top\right)\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} - \boldsymbol{B}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{B}+\boldsymbol{B}^\top\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} + \boldsymbol{D}\right]\right\| \\
&\le \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\left(\boldsymbol{I}+\boldsymbol{B}-\frac{1}{2}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{B}+\boldsymbol{B}^\top\right)\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| + 2\|\boldsymbol{B}\|^2 + \|\boldsymbol{D}\|
\end{aligned} \tag{35}$$

where $\boldsymbol{B} = (\boldsymbol{I}+\mu\boldsymbol{M}_t)(\boldsymbol{I}+\boldsymbol{P}) - \boldsymbol{I}$ and $\|\boldsymbol{D}\| \le 100\|\boldsymbol{B}\|^2$. By assumption we have

$$\begin{aligned}
\|\boldsymbol{P}\| &\le \frac{\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|\,\|\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t+1}\|}{\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t)\sigma_{\min}(\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1})} \\
&\le 2\,\|\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t+1}\|,
\end{aligned}$$

so that

$$\begin{aligned}
&\left\|\boldsymbol{B}-\mu(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \\
&\le \mu\|\boldsymbol{M}_t-(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top)\| + \|\boldsymbol{P}\| + \mu\|\boldsymbol{M}_t\|\|\boldsymbol{P}\| \\
&\le \mu\,\|\boldsymbol{\Delta}_t\| + 2\,\|\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t+1}\| + 4\mu\|\boldsymbol{X}\|^2\,\|\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t+1}\| \\
&\le \mu\,\|\boldsymbol{\Delta}_t\| + 6\,\|\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t+1}\| \\
&\le 18\mu\left(10\mu\|\boldsymbol{X}\|^3+c_4\right)c_3\|\boldsymbol{X}\| + 7\mu\,\|\boldsymbol{\Delta}_t\| \\
&\le 18\mu\left(10\mu\|\boldsymbol{X}\|^3+c_4\right)c_3\|\boldsymbol{X}\| + 0.01\mu\kappa^{-1}c_3\|\boldsymbol{X}\|^2
\end{aligned} \tag{36}$$

where we use Lemma D.7 to bound $\left\|\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1}\right\|$. Let $\boldsymbol{B}_1 = \mu(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)$ and $\boldsymbol{R}_1 = \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\left(\boldsymbol{I}+\boldsymbol{B}_1 - \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{B}_1\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}$, then we have

$$\begin{aligned}\boldsymbol{R}_1 &= \boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \left(\boldsymbol{I} + \mu\left(\boldsymbol{I} - \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} \\ &= \left(\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2\right)\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\left(\boldsymbol{I} - \mu\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right) \\ &\quad - \mu\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\left(\boldsymbol{I} - \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\right)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} \\ &\quad + \mu^2\boldsymbol{\Sigma}_{s,\perp}^2\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}.\end{aligned} \tag{37}$$

By Weyl's inequality (cf. Lemma A.4) and our assumption on c_3,

$$\begin{aligned}\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right) &\geq \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{X}_s\boldsymbol{X}_s^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right) - \left\|\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{X}_{s,\perp}\boldsymbol{X}_{s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|^2 \\ &\geq \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{X}_s\boldsymbol{X}_s^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right) - \sigma_{s+1}^2\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|^2 \\ &\geq \sigma_s^2\left\|\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{V}_{\boldsymbol{X}_s}\right\|^2 - \sigma_{s+1}^2c_3^2 \\ &= \sigma_s^2 - (\sigma_s^2+\sigma_{s+1}^2)c_3^2 > \frac{1}{2}\left(\sigma_s^2+\sigma_{s+1}^2\right).\end{aligned}$$

So we have

$$\|\boldsymbol{R}_1\| \leq \left(1 - \frac{\mu}{2}(\sigma_s^2 - \sigma_{s+1}^2)\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| + \mu c_3c_4^2 + \mu^2\|\boldsymbol{X}\|^4.$$

It thus follows from (35) that

$$\begin{aligned}&\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1}}\right\| \\ &\leq \|\boldsymbol{R}_1\| + 2\|\boldsymbol{B}-\boldsymbol{B}_1\| + 102\|\boldsymbol{B}\|^2 \\ &\leq \left(1 - \frac{\mu}{2}(\sigma_s^2 - \sigma_{s+1}^2)\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| + 40\mu c_3c_4\|\boldsymbol{X}\| + 0.02\mu\kappa^{-1}c_3\|\boldsymbol{X}\|^2 + 10^3\mu^2\|\boldsymbol{X}\|^4.\end{aligned}$$

Since $\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \leq c_3$, it follows from our assumption on c_3, c_4 and μ that $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1}}\right\| \leq c_3$ as well, which concludes the proof. □

C.3. Induction

Let

$$T_{\alpha,s}^{\mathtt{pi}} = \min\left\{t \geqslant 0 : \sigma_{\min}^2\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_{\alpha,t+1}\right) > 0.3\kappa^{-1}\|\boldsymbol{X}\|^2\right\}.$$

where pi stands for the parallel improvement phase. In this section, we show that when $T_\alpha^{\mathtt{sp}} \leq t < T_{\alpha,s}^{\mathtt{pi}}$, the parallel component grows exponentially faster than the orthogonal component. We prove this via induction and the base case is already shown in Lemma C.4.

Lemma C.10 (Lemma 5.2, detailed version)**.** *Suppose that Assumptions 3.3, 3.5 and 3.7 hold and let $c_3 = 10^4\kappa\sqrt{r_*}\delta$, $c_4 \leq 10^{-3}\kappa^{-1}\|\boldsymbol{X}\|$. Then the following holds for all $T_\alpha^{\mathtt{sp}} \leq t < T_{\alpha,s}^{\mathtt{pi}}$ as long as $\alpha \leq C_4(\boldsymbol{X},\bar{\boldsymbol{U}}) = \left(\kappa\frac{C_2(\boldsymbol{X},\bar{\boldsymbol{U}})^2}{C_3(\boldsymbol{X},\bar{\boldsymbol{U}})^2}\right)^{-2\kappa}$ is sufficiently small:*

$$\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_{t+1}\right) \geq \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_{t+1}\boldsymbol{W}_t\right) \geq \left(1 + 0.5\mu\left(\sigma_s^2+\sigma_{s+1}^2\right)\right)\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_{\alpha,t}\right) \tag{38a}$$

$$\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \leq \min\left\{\left(1 + \mu\left(0.4\sigma_s^2 + 0.6\sigma_{s+1}^2\right)\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|, c_4\right\} \tag{38b}$$

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1}}\right\| \leq c_3. \tag{38c}$$

$$\operatorname{rank}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_{t+1}) = \operatorname{rank}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_{t+1}\boldsymbol{W}_t) = s. \tag{38d}$$

Proof. The base case $t = T_{\alpha,s}^{\mathtt{pi}}$ is already proved in (19). Now suppose that the lemma holds for t, we now show that it holds for $t+1$ as well.

To begin with, we bound the term $\|\boldsymbol{\Delta}_t\|$ as follows:

$$\begin{aligned}\|\boldsymbol{\Delta}_t\| &= \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \\ &\leq \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top)\right\| + \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\right\| \\ &\leq 10\delta\sqrt{r_*}\|\boldsymbol{X}\|^2 + \delta\left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\right\|_* \\ &\leq 10\delta\sqrt{r_*}\|\boldsymbol{X}\|^2 + \delta d\left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|^2\end{aligned} \tag{39}$$

where in the second inequality we use Proposition A.2 and Lemma A.3 and in the third inequality we use $\|\boldsymbol{A}\|_* \leq \sqrt{d}\|\boldsymbol{A}\|, \forall \boldsymbol{A} \in \mathbb{R}^{d\times d}$ and the induction hypotheses.

By induction hypothesis, there exists a constant $\hat{C}_4(\boldsymbol{X}, \bar{\boldsymbol{U}}) = \frac{C_2(\boldsymbol{X},\bar{\boldsymbol{U}})}{C_3(\boldsymbol{X},\bar{\boldsymbol{U}})}$ (see Lemma C.4) such that

$$\frac{\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\right)}{\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|} \geq \frac{\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{T_\alpha^{\mathtt{sp}}}\right)}{\left\|\boldsymbol{U}_{T_\alpha^{\mathtt{sp}}}\boldsymbol{W}_{T_\alpha^{\mathtt{sp}},\perp}\right\|} \geq \hat{C}_4 \cdot \alpha^{-\gamma_s} \tag{40}$$

where

$$\gamma_s = \frac{2\left(\log\left(1+\mu\hat{\sigma}_s^2\right) - \log\left(1+\mu\hat{\sigma}_{s+1}^2\right)\right)}{3\log\left(1+\mu\hat{\sigma}_1^2\right) - \log\left(1+\mu\hat{\sigma}_{s+1}^2\right)} \geq \frac{1}{4\kappa}.$$

Since we must have $\sigma_{\min}^2\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\right) \leq 0.3\kappa^{-1}\|\boldsymbol{X}\|^2$ by definition of $T_{\alpha,s}^{\mathtt{pi}}$, it follows that $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|^2 \leq 10\kappa\|\boldsymbol{X}\|^2\hat{C}_4^2\alpha^{\frac{1}{2\kappa}}$, so for $\alpha \leq (\hat{C}_4^{-2}\kappa)^{-2\kappa}$, $\|\boldsymbol{\Delta}_t\| \leq 11\delta\sqrt{r_*}\|\boldsymbol{X}\|^2$ holds.

The above inequality combined with our assumption on δ implies that the conditions on $\|\boldsymbol{\Delta}_t\|$ in Lemmas C.5, C.8 and C.9 hold. We now show that (38a) to (38d) hold for $t+1$, which completes the induction step.

First, since $t < T_{\alpha,s}^{\mathtt{pi}}$, we have $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right) \leq \kappa^{-1}\|\boldsymbol{X}\|^2$. Moreover, the induction hypothesis implies that $\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t-1}\boldsymbol{W}_{t-1}}\right\| \leq c_3$ and that $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{\alpha,t}$ is of full rank. Thus the conditions of Corollary C.6 are all satisfied, and we deduce that (38a) holds.

Second, the assumptions on c_3, c_4 and δ, combined with Lemma C.8, immediately implies

$$\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \leq \left(1+\mu\left(0.4\sigma_s^2 + 0.6\sigma_{s+1}^2\right)\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|.$$

As a result, similar to (40) we observe that

$$\frac{\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right)}{\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\|} \geq \frac{\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{T_\alpha^{\mathtt{sp}}}\right)}{\left\|\boldsymbol{U}_{T_\alpha^{\mathtt{sp}}}\boldsymbol{W}_{T_\alpha^{\mathtt{sp}},\perp}\right\|} \geq \hat{C}_4 \cdot \alpha^{-\frac{1}{4\kappa}}.$$

Since $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right) \leq \|\boldsymbol{X}\|$, when α is sufficiently small we must have that $\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \leq c_4$.

Finally, Lemma C.9 implies that (38c) is true, and (38d) follows from our application of Lemma C.5. This concludes the proof. □

C.4. The refinement phase and concluding the proof of Lemma 4.3

We have shown that the parallel component $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right)$ grows exponentially faster than the orthogonal component $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$. In this section, we characterize the GD dynamics *after* $T_{\alpha,s}^{\mathtt{pi}}$. We begin with the following lemma, which is straightforward from the proof of Lemma C.10.

Lemma C.11 (Corollary 5.3, formal version)**.** *Under the conditions of Lemma 5.2, the following inequality holds when $\alpha \leq C_4(\boldsymbol{X}, \bar{\boldsymbol{U}})$:*

$$\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{pi}}}\boldsymbol{W}_{T_{\alpha,s}^{\mathtt{pi}},\perp}\right\| \leq C_5(\boldsymbol{X}, \bar{\boldsymbol{U}}) \cdot \alpha^{\frac{1}{4\kappa}}$$

where $C_5 = \sqrt{10\kappa}\|\boldsymbol{X}\|\frac{C_2(\boldsymbol{X},\bar{\boldsymbol{U}})}{C_3(\boldsymbol{X},\bar{\boldsymbol{U}})}$.

The following lemma states that in a certain time period after $T_{\alpha,s}^{\mathtt{pi}}$, the parallel and orthogonal components still behave similarly to the second (parallel improvement) phase.

Lemma C.12. *Under the conditions in Lemma C.10, there exists $\widetilde{t}_{\alpha,s} \geq \frac{1}{\log(1+\mu\sigma_s^2)}\log\left(\frac{10^{-4}c_3\|\boldsymbol{X}\|^2}{\sqrt{d}\kappa C_5}\alpha^{-\frac{1}{4\kappa}}\right) = \Theta\left(\log\alpha^{-1}\right)$ when $\alpha \to 0$ such that when $0 \leq t - T_{\alpha,s}^{\mathtt{pi}} \leq \widetilde{t}_{\alpha,s}$, we have*

$$\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\right) \geq \sigma_{\min}\left(\boldsymbol{U}_t\boldsymbol{W}_t\right) \geq 0.3\kappa^{-1}\|\boldsymbol{X}\|^2, \tag{41a}$$

$$\|\boldsymbol{U}_t\boldsymbol{W}_t\| \leq \left(1+\mu(0.4\sigma_s^2+0.6\sigma_{s+1}^2)\right)^{t-T_{\alpha,s}^{\mathtt{pi}}}\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{pi}}}\boldsymbol{W}_{T_{\alpha,s}^{\mathtt{pi}}}\right\|, \tag{41b}$$

$$\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\| \leq c_3. \tag{41c}$$

Proof. We choose

$$\widetilde{t}_{\alpha,s} = \min\left\{t \geq 0 : \|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\|^2 \leq c_5\right\} \tag{42}$$

where

$$c_5 = 10^{-4}d^{-\frac{1}{2}}\kappa^{-1}c_3\|\boldsymbol{X}\|^2 \tag{43}$$

We prove (41) by induction. The proof follows the idea of Lemma C.10, except that we need to bound $\|\boldsymbol{\Delta}_t\|$ in each induction step. Concretely, suppose that (41) holds at time t, then

$$\begin{aligned}
\|\boldsymbol{\Delta}_t\| &= \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \\
&\leq \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top)\right\| + \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\right\| \\
&\leq 10\delta\sqrt{r_*}\|\boldsymbol{X}\|^2 + \delta\left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\right\|_* \\
&\leq 10\delta\sqrt{r_*}\|\boldsymbol{X}\|^2 + \delta c_5\sqrt{d} \leq 0.02\kappa^{-1}c_3\|\boldsymbol{X}\|^2
\end{aligned} \tag{44}$$

where we used the definition of c_5 in the last step. As a result, we can apply the conclusion of Lemmas C.5, C.8 and C.9 which implies that (41) holds for $t+1$. Finally, combining Lemma C.11 and (41b) yields $\widetilde{t}_{\alpha,s} = \Theta\left(\log\frac{1}{\alpha}\right)$. □

We now present the main result of this section:

Lemma C.13. *Suppose that* $0 \leq t - T^\pi_{\alpha,s} \leq \widetilde{t}_{\alpha,s}$, $\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \leq c_3$ *and the conditions in Lemma C.10 hold, then we have*

$$\begin{aligned}
&\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top)\right\|_F \\
\leq &\left(1 - \frac{1}{2}\mu\tau\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F + 20\mu\|\boldsymbol{X}\|^4\left(\delta + 5c_3\right) + 2000\mu^2\sqrt{r_*}\|\boldsymbol{X}\|^6.
\end{aligned}$$

where we recall that $\tau = \min_{1\leq s\leq\hat{r}\wedge r_*}(\sigma_s^2 - \sigma_{s+1}^2) > 0$.

Proof. Recall that $\boldsymbol{M}_t = \mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)$. The update of GD implies that

$$\begin{aligned}
&\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top \\
&= \boldsymbol{X}\boldsymbol{X}^\top - (\boldsymbol{I} + \mu\boldsymbol{M}_t)\boldsymbol{U}_t\boldsymbol{U}_t^\top(\boldsymbol{I} + \mu\boldsymbol{M}_t) \\
&= \underbrace{\left(\boldsymbol{I} - \mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\left(\boldsymbol{I} - \mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)}_{=(i)} + \mu\underbrace{\boldsymbol{\Delta}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top}_{=(ii)} \\
&\quad + \underbrace{\mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{\Delta}_t}_{=(iii)} + \mu^2\left(\mathcal{E}_{t,1} + \mathcal{E}_{t,2}\right),
\end{aligned}$$

where $\mathcal{E}_{t,1} = -\boldsymbol{U}_t\boldsymbol{U}_t^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{U}_t\boldsymbol{U}_t^\top$ and $\mathcal{E}_{t,2} = -\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t$. Since $\|\boldsymbol{U}_t\| \leq 3\|\boldsymbol{X}\|$ by Lemma D.4, and $\left\|\boldsymbol{M}_t - (\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| = \|\boldsymbol{\Delta}_t\| \leq \|\boldsymbol{X}\|^2$ which is shown in (44), we have

$$\begin{aligned}
\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\mathcal{E}_{t,1}\right\|_F &= \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{U}_t^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_F \\
&\leq \sqrt{r_*}\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{U}_t\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_2 \\
&\leq 10^3\sqrt{r_*}\|\boldsymbol{X}\|^6
\end{aligned}$$

and

$$\begin{aligned}
\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\mathcal{E}_{t,2}\right\|_F &= \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left[(\mathcal{A}^*\mathcal{A})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\boldsymbol{U}_t\boldsymbol{U}_t^\top\left[(\mathcal{A}^*\mathcal{A})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\right\|_F \\
&\leq \sqrt{r_*}\left\|\left[(\mathcal{A}^*\mathcal{A})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\boldsymbol{U}_t\boldsymbol{U}_t^\top\left[(\mathcal{A}^*\mathcal{A})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\right\| \\
&\leq 10^3\sqrt{r_*}\|\boldsymbol{X}\|^6.
\end{aligned}$$

Note that we would like to bound $\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top\right)\right\|_F$. We deal with the above three terms separately. For the first term, we have

$$\begin{aligned}
&\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{I}-\mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\left(\boldsymbol{I}-\mu\boldsymbol{U}_t\boldsymbol{U}_t\right)\right\|_F\\
=&\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{I}-\mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\left(\boldsymbol{I}-\mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F\\
&+\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{I}-\mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\left(\boldsymbol{I}-\mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F\\
\le&\left\|\boldsymbol{I}-\mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s}\right\|\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|+\mu\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F &\text{(45a)}\\
\le&\left(1-\mu\sigma_{\min}^2(\boldsymbol{U}_t\boldsymbol{W}_t)\sigma_{\min}^2\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F+100\mu\|\boldsymbol{X}\|^4c_3 &\text{(45b)}\\
\le&\left(1-\frac{1}{2}\mu\tau\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F+100\mu\|\boldsymbol{X}\|^4c_3, &\text{(45c)}
\end{aligned}$$

where in (45a) we use $\left\|\boldsymbol{I}-\mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|\le 1$, (45b) follows from

$$\begin{aligned}
\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s}\right)&=\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s}\right)\ge\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{W}_t\right)^2\\
&\ge\sigma_{\min}^2(\boldsymbol{U}_t\boldsymbol{W}_t)\sigma_{\min}^2\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)
\end{aligned}$$

and

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\right\|=\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\right\|\le\|\boldsymbol{U}_t\|^2\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_t\right\|\le c_3\|\boldsymbol{U}_t\|^2,$$

and lastly (45c) is obtained from

$$\sigma_{\min}^2\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)\ge 1-\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|^2\ge 1-c_3^2.$$

For the second and the third terms, we have

$$\left\|\boldsymbol{\Delta}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top+\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{\Delta}_t\right\|\le 0.1\kappa c_3\|\boldsymbol{X}\|^4 \tag{46}$$

where we use the estimate in (44). Combining (45) and (46) yields

$$\begin{aligned}
&\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top)\right\|_F\\
\le&\left(1-\frac{1}{2}\mu\tau\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F+200\mu\|\boldsymbol{X}\|^4c_3+110\mu^2\sqrt{r_*}\|\boldsymbol{X}\|^6.
\end{aligned}$$

□

To apply the result of Lemma 5.4, we need to verify that $\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\|\le c_3$ still holds when $t\ge T_{\alpha,s}^{\mathtt{pi}}$. In fact, this is true as long as $t-T_{\alpha,s}^{\mathtt{pi}}\le\mathcal{O}\left(\log\frac{1}{\alpha}\right)$.

Lemma C.14. *Under the conditions in Lemma C.10, if*

$$T_{\alpha,s}^{\mathtt{pi}}\le t\le T_{\alpha,s}^{\mathtt{pi}}+\frac{\gamma_s\log\frac{c_4}{C_5(\boldsymbol{X},\bar{\boldsymbol{U}})}\cdot\log\frac{1}{\alpha}}{\log\left(1+\mu\sigma_s^2\right)}=:T_{\alpha,s}^{\mathtt{re}},$$

then $\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\|\le\left(1+\mu\sigma_s^2\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$ *and* $\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|\le c_3$. *As a consequence, we have* $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|\le\left(1+\mu\sigma_s^2\right)^{t-T_{\alpha,s}^{\mathtt{pi}}}C_5(\boldsymbol{X},\bar{\boldsymbol{U}})\cdot\alpha^{\gamma_s}\le c_4$.

Proof. The proof is basically the same as that of Lemma C.10 and we only provide a sketch here.

We induct on t. The base case $t=T_{\alpha,s}^{\mathtt{pi}}$ is already proved in Lemma C.10. Suppose that the lemma holds for $t-1$ with $t<T_{\alpha,s}^{\mathtt{re}}$, then the choice of $T_{\alpha,s}^{\mathtt{re}}$ combined with Lemma C.8 imply that

$$\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|\le\left(1+\mu\sigma_s^2\right)\|\boldsymbol{U}_{t-1}\boldsymbol{W}_{t-1,\perp}\|\le\left(1+\mu\sigma_s^2\right)^{t-T_{\alpha,s}^{\mathtt{pi}}}\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{pi}}}\boldsymbol{W}_{T_{\alpha,s}^{\mathtt{pi}},\perp}\right\|.$$

Since we have $\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{pi}}}\boldsymbol{W}_{T_{\alpha,s}^{\mathtt{pi}},\perp}\right\|\le C_5(\boldsymbol{X},\bar{\boldsymbol{U}})\cdot\alpha^{\gamma_s}$ by Lemma C.11, the choice of $T_{\alpha,s}^{\mathtt{re}}$ implies that $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|\le c_4$. The bound $\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|\le c_3$ then follows from Lemma C.9. □

We will only use a weaker version of this lemma, namely that the bounds holds for all $T^{\mathtt{pi}}_{\alpha,s} \le t \le T^{\mathtt{ft}}_{\alpha,s}$. When α is sufficiently small, this can be directly derived from Lemmas C.13 and C.14 since $\tilde{t}_{\alpha,s}, T^{\mathtt{re}}_{\alpha,s} - T^{\mathtt{pi}}_{\alpha,s} = \Theta\left(\log \frac{1}{\alpha}\right)$. Specifically, we have proven Lemma 5.4 in the main text:

Lemma 5.4. *Suppose that $T^{\mathtt{pi}}_{\alpha,s} \le t \le T^{\mathtt{ft}}_{\alpha,s}$ and all the conditions in Lemma 5.2 hold, then we have*

$$\begin{aligned}
&\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top)\right\|_F \\
\le &\left(1 - \frac{1}{2}\mu\tau\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F \\
&+ 20\mu\|\boldsymbol{X}\|^4(\delta + 5c_3) + 2000\mu^2\sqrt{r_*}\|\boldsymbol{X}\|^6.
\end{aligned}$$

Moreover, we have $\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \le (1+\sigma_s^2)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$ and $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \le c_3$.

We are now ready to present our first main result, which states that with small initialization, GD would visit the $\mathcal{O}(\delta)$-neighborhood of the rank-s minimizer of the full observation loss i.e. $\boldsymbol{X}_s\boldsymbol{X}_s^\top$.

Theorem C.15 (Restatement of Lemma 4.3)**.** *Under Assumptions 3.3, 3.5 and 3.7, if the initialization scale α is sufficiently small, then for all $1 \le s \le \hat{r} \wedge r_*$ there exists a time $T^{\mathtt{ft}}_{\alpha,s} \in \mathbb{Z}_+$ (where $\mathtt{ft}$ stands for fitting the ground-truth) such that*

$$\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_{T^{\mathtt{pi}}_{\alpha,s}}\boldsymbol{U}_{T^{\mathtt{pi}}_{\alpha,s}}^\top\right\|_F \le 10^7\kappa^3 r_*\|\boldsymbol{X}\|^2\delta.$$

Proof. First, observe that for all $t \ge 0$,

$$\begin{aligned}
\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_F &\le \left\|\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\right\|_F + \left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\right\|_F \\
&\le \left\|\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\right\|_F + \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\right\|_F \\
&\le \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F + \sqrt{r_*}\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_t\right\|^2 + \sqrt{d}\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|^2 \\
&\le \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F + 9\sqrt{r_*}\|\boldsymbol{X}\|^2\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|^2 + \sqrt{d}\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|^2.
\end{aligned} \tag{47}$$

where the last step uses Lemma D.4 to bound $\|\boldsymbol{U}_t\|$. We set $c_3 = 10^3\kappa\sqrt{r_*}\delta$ and

$$T^{\mathtt{ft}}_{\alpha,s} = T^{\mathtt{pi}}_{\alpha,s} - \frac{\log\left(10^{-2}\|\boldsymbol{X}\|^{-2}\tau c_3^{-1}\right)}{\log\left(1 - \frac{1}{2}\mu\tau\right)}, \tag{48}$$

where we recall that $\tau = \kappa^{-1}\|\boldsymbol{X}\|^2$, then for small α we have $T^{\mathtt{ft}}_{\alpha,s} \le T^{\mathtt{pi}}_{\alpha,s} + \widetilde{t}_{\alpha,s}$ (defined in Lemma C.12). Hence for $T^{\mathtt{pi}}_{\alpha,s} \le t < T^{\mathtt{ft}}_{\alpha,s}$ we always have $\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\| \le c_3$. By Lemma C.13 and the choice of c_3 and δ, we have for $T^{\mathtt{pi}}_{\alpha,s} \le t < T^{\mathtt{ft}}_{\alpha,s}$ that

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top\right)\right\|_F \le \left(1 - \frac{1}{2}\mu\tau\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F + 30\mu\|\boldsymbol{X}\|^4\sqrt{r_*}c_3$$

which implies that for $t = T^{\mathtt{ft}}_{\alpha,s}$,

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_{T_s}\boldsymbol{U}_{T_s}^\top\right)\right\|_F \le 80\kappa\|\boldsymbol{X}\|^2\sqrt{r_*}c_3.$$

Meanwhile, by Lemma C.12 we have $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \le c_5$ (c_5 is defined in (43)) and $\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \le c_3$ at $t = T^{\mathtt{ft}}_{\alpha,s}$. Plugging into (47) yields

$$\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_{T^{\mathtt{ft}}_{\alpha,s}}\boldsymbol{U}_{T^{\mathtt{ft}}_{\alpha,s}}^\top\right\|_F \le 80\kappa\|\boldsymbol{X}\|^2\sqrt{r_*}c_3 + 9\|\boldsymbol{X}\|^2c_3^2\sqrt{r_*} + c_5^2\sqrt{d}.$$

By definition of c_3 and c_5 we deduce that $\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_{T^{\mathtt{ft}}_{\alpha,s}}\boldsymbol{U}_{T^{\mathtt{ft}}_{\alpha,s}}^\top\right\|_F \le 10^2\tau^{-2}\|\boldsymbol{X}\|^6\sqrt{r_*}c_3 \le 10^5\kappa^3 r_*\|\boldsymbol{X}\|^2\delta$, as desired. □

Corollary C.16. *There exists a constant*

$$C_6(\boldsymbol{X}, \bar{\boldsymbol{U}}) = C_5(\boldsymbol{X}, \bar{\boldsymbol{U}}) \cdot (1+\mu\sigma_s^2)^{T_{\alpha,s}^{\mathtt{ft}} - T_{\alpha,s}^{\mathtt{pi}}} \tag{49}$$

such that

$$\max_{0 \le t \le T_{\alpha,s}^{\mathtt{ft}}} \|\boldsymbol{U}_t \boldsymbol{W}_{t,\perp}\| \le C_1 \cdot \alpha^{\frac{1}{4\kappa}}.$$

Proof. The case of $t \le T_{\alpha,s}^{\mathtt{pi}}$ directly follows from Lemma C.11. For $t > T_{\alpha,s}^{\mathtt{pi}}$, we know from Lemma C.12 that

$$\|\boldsymbol{U}_t \boldsymbol{W}_{t,\perp}\| \le \left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{pi}}} \boldsymbol{W}_{T_{\alpha,s}^{\mathtt{pi}},\perp}\right\| \cdot \left(1+\mu\sigma_s^2\right)^{T_{\alpha,s}^{\mathtt{ft}} - T_{\alpha,s}^{\mathtt{pi}}}.$$

By (48), the second term is a constant independent of α, so the conclusion follows. □

D. Auxiliary results for proving Lemma 4.3

This section contains a collection of auxiliary results that are used in the previous section.

D.1. The spectral phase

In the section, we provide auxiliary results for the analysis in the spectral phase.

Recall that $\boldsymbol{K}_t = (\boldsymbol{I}+\mu\boldsymbol{M})^t$ and $\boldsymbol{U}_t = \boldsymbol{U}_t^{\mathtt{sp}} + \boldsymbol{E}_t = \boldsymbol{K}_t\boldsymbol{U}_0 + \boldsymbol{E}_t$ and $\boldsymbol{U}_0 = \alpha\bar{\boldsymbol{U}}$ with $\|\bar{\boldsymbol{U}}\| = 1$. Also recall that $\boldsymbol{M} = \sum_{i=1}^{\mathrm{rank}(\boldsymbol{M})} \hat{\lambda}_i \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top$; we additionally define $\boldsymbol{M}_s = \sum_{i=1}^{\min\{s,\mathrm{rank}(\boldsymbol{M})\}} \hat{\lambda}_i \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top$. Similarly, let $\boldsymbol{L}_t$ be the span of the top-s left singular vectors of $\boldsymbol{U}_t$. The following lemma shows that power iteration would result in large eigengap of $\boldsymbol{U}_t$.

Lemma D.1. *Let $\hat{\rho} = \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{M}_s}^\top \bar{\boldsymbol{U}}\right) > 0$, then the following three inequalities hold, given that the denominator of the third is positive.*

$$\sigma_s(\boldsymbol{U}_t) \ge \alpha\left(\hat{\rho}\sigma_s\left(\hat{\boldsymbol{Z}}_t\right) - \sigma_{s+1}\left(\hat{\boldsymbol{Z}}_t\right)\right) - \|\boldsymbol{E}_t\|, \tag{50a}$$

$$\sigma_{s+1}(\boldsymbol{U}_t) \le \alpha\sigma_{s+1}\left(\hat{\boldsymbol{Z}}_t\right) + \|\boldsymbol{E}_t\|, \tag{50b}$$

$$\left\|\boldsymbol{V}_{\boldsymbol{M}_s^\perp}^\top \boldsymbol{V}_{L_t}\right\| \le \frac{\alpha\sigma_{s+1}\left(\hat{\boldsymbol{Z}}_t\right) + \|\boldsymbol{E}_t\|}{\alpha\hat{\rho}\sigma_s\left(\hat{\boldsymbol{Z}}_t\right) - 2\left(\alpha\sigma_{s+1}\left(\hat{\boldsymbol{Z}}_t\right) + \|\boldsymbol{E}_t\|\right)}. \tag{50c}$$

Proof. By Weyl's inequality we have

$$\begin{aligned}
\sigma_{s+1}(\boldsymbol{U}_t) &= \sigma_{s+1}\left((1+\mu\boldsymbol{M})^t\boldsymbol{U}_0\right) + \|\boldsymbol{E}_t\| \\
&= \alpha\sigma_{s+1}\left((1+\mu\boldsymbol{M})^t\bar{\boldsymbol{U}}\right) + \|\boldsymbol{E}_t\| \\
&\le \alpha\sigma_{s+1}\left((1+\mu\boldsymbol{M}_s)^t\bar{\boldsymbol{U}}\right) + \alpha\left\|\left[(1+\mu\boldsymbol{M})^t - (1+\mu\boldsymbol{M}_s)^t\right]\bar{\boldsymbol{U}}\right\| + \|\boldsymbol{E}_t\| \\
&\le \alpha(1+\mu\hat{\lambda}_{s+1})^t + \|\boldsymbol{E}_t\|.
\end{aligned}$$

Thus (50b) holds. Similarly,

$$\begin{aligned}
\sigma_s(\boldsymbol{U}_t) &\ge \alpha\sigma_s\left(\boldsymbol{N}_t\boldsymbol{V}_{\boldsymbol{M}_s}\boldsymbol{V}_{\boldsymbol{M}_s}^\top\bar{\boldsymbol{U}}\right) - \alpha(1+\mu\hat{\lambda}_{s+1})^t - \|\boldsymbol{E}_t\| \\
&\ge \alpha\sigma_s\left(\boldsymbol{N}_t\boldsymbol{V}_{\boldsymbol{M}_s}\right)\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{M}_s}^\top\bar{\boldsymbol{U}}\right) - \alpha(1+\mu\hat{\lambda}_{s+1})^t - \|\boldsymbol{E}_t\| \\
&\ge \alpha\hat{\rho}(1+\mu\hat{\lambda}_s)^t - \alpha(1+\mu\hat{\lambda}_{s+1})^t - \|\boldsymbol{E}_t\|.
\end{aligned}$$

Finally, note that we can write

$$\alpha(1+\mu\boldsymbol{M}_s)^t\bar{\boldsymbol{U}} = \boldsymbol{V}_{\boldsymbol{M}_s}\underbrace{(1+\mu\boldsymbol{\Sigma}_{\boldsymbol{M}_s})^t\boldsymbol{V}_{\boldsymbol{M}_s}^\top\bar{\boldsymbol{U}}}_{\text{invertible}},$$

so that the subspace spanned by the left singular vectors of $\alpha(1+\mu\boldsymbol{M}_s)^t\bar{\boldsymbol{U}}$ coincides with the column span of $\boldsymbol{V}_{\boldsymbol{M}_s}$. Since L_t is the span of top-s left singular vectors of $\boldsymbol{U}_t$, we apply Wedin's sin theorem (Wedin, 1972) and deduce (50c). □

The next lemma relates the quantities studied in Lemma D.1 with those that are needed in the induction. The proof is the same as Stöger & Soltanolkotabi, 2021, Lemma 8.4, so we omit it here.

Lemma D.2. *Suppose that* $\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{L}_t}\right\| \leq 0.1$ *for some* $t \geq 1$. *Then it holds that*

$$\sigma_s\left(\boldsymbol{U}_t\boldsymbol{W}_t\right) \geq \frac{1}{2}\sigma_s\left(\boldsymbol{U}_t\right), \tag{51a}$$

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \leq 10\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{L}_t}\right\|, \tag{51b}$$

$$\left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\| \leq 2\sigma_{s+1}\left(\boldsymbol{U}_t\right). \tag{51c}$$

Combining the above two lemmas, we directly obtain the following corollary:

Corollary D.3. *Suppose that* $\alpha\sigma_s(\boldsymbol{K}_t) > 10\left(\alpha\sigma_{s+1}(\boldsymbol{K}_t) + \|\boldsymbol{E}_t\|\right)$, *then we have that*

$$\begin{aligned}
\sigma_s\left(\boldsymbol{U}_t\boldsymbol{W}_t\right) &\geq 0.4\alpha\sigma_{r_\star}\left(\boldsymbol{K}_t\right)\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{L}}^\top\bar{\boldsymbol{U}}\right)\\
\left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\| &\leq 2\left(\alpha\sigma_{s+1}\left(\boldsymbol{K}_t\right) + \|\boldsymbol{E}_t\|\right)\\
\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| &\leq 100\left(\delta + \frac{\alpha\sigma_{s+1}\left(\boldsymbol{K}_t\right) + \|\boldsymbol{E}_t\|}{\alpha\hat{\rho}\sigma_s\left(\boldsymbol{K}_t\right)}\right)
\end{aligned} \tag{52}$$

D.2. The parallel improvement phase

In the section, we provide auxiliary results for the analysis in the parallel improvement phase.

Lemma D.4. *(Stöger & Soltanolkotabi, 2021, Lemma 9.4) For sufficiently small* μ *and* δ, *suppose that* $\|\boldsymbol{U}_t\| \leq 3\|\boldsymbol{X}\|$, *then we also have* $\|\boldsymbol{U}_{t+1}\| \leq 3\|\boldsymbol{X}\|$.

Lemma D.5. *Under the assumptions in Lemma C.8, we have*

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right\| \leq 2\left(c_3 + 10\mu\|\boldsymbol{X}\|^2\right) \leq 0.01.$$

Proof. The proof of this lemma is essentially the same as Stöger & Soltanolkotabi, 2021, Lemma B.1, and we omit it here. □

Lemma D.6. *Under the assumptions in Lemma C.9, we have*

$$\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right) \geq \frac{1}{2}\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t).$$

Proof. We have

$$\begin{aligned}
&\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right) \geq \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\right)\\
&= \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{I} + \mu\boldsymbol{M}_t\right)\boldsymbol{U}_t\boldsymbol{W}_t\right)\\
&\geq \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{I} + \mu\boldsymbol{M}_t\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)\cdot\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{U}_t\boldsymbol{W}_t\right)\\
&\geq \left[\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right) - \mu\|\boldsymbol{M}_t\|\right]\cdot\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t)\\
&\geq \left(\sqrt{1-c_3^2} - 10\mu\|\boldsymbol{X}\|^2\right)\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t) \geq \frac{1}{2}\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t)
\end{aligned}$$

where the last step follows from

$$\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)^2 \geq 1 - \left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|^2 \geq 1 - c_3^2.$$

The conclusion follows. □

Lemma D.7. *Under the assumptions in Lemma C.9, we have*

$$\left\|\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1}\right\| \leq 3\mu\left(10\mu\|\boldsymbol{X}\|^2 + c_4\right)c_3\|\boldsymbol{X}\| + \mu\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\|.$$

Proof. The proof roughly follows [Stöger & Soltanolkotabi, 2021, Lemma B.3], but we include it here for completeness. Since $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1} = \boldsymbol{V}_{t+1}\boldsymbol{\Sigma}_{t+1}\boldsymbol{W}_{t+1}$ and $\boldsymbol{V}_{t+1}\boldsymbol{\Sigma}_{t+1} \in \mathbb{R}^{s\times s}$ is invertible, we have

$$\left\|\boldsymbol{W}_{t,\perp}^\top \boldsymbol{W}_{t+1}\right\| = \left\|\boldsymbol{W}_{t,\perp}^\top \boldsymbol{U}_{t+1}^\top \boldsymbol{V}_{\boldsymbol{X}_s} \left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-\frac{1}{2}}\right\|.$$

Since

$$\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp} \\
&= \boldsymbol{V}_{\boldsymbol{X}_s}^\top \left(\boldsymbol{I} + \mu\mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} \\
&= \boldsymbol{V}_{\boldsymbol{X}_s}^\top \left(\boldsymbol{I} + \mu(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{\Delta}_t\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} \\
&= -\mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{\Delta}_t\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} && \text{(53a)} \\
&= -\mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{\Delta}_t\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} && \text{(53b)} \\
&= -\mu\underbrace{\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}}_{=:\boldsymbol{K}_1} + \mu\underbrace{\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{\Delta}_t\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}}_{:=\boldsymbol{K}_2} && \text{(53c)}
\end{aligned}$$

where (53a) follows from $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = \boldsymbol{\Sigma}_s\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = 0$, and in (53b) and (53c) we use $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = 0$. For $\boldsymbol{K}_1$, note that

$$\begin{aligned}
&\left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-\frac{1}{2}} \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\right\| \\
&\le \left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-\frac{1}{2}} \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right\| + \mu\left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-\frac{1}{2}} \boldsymbol{V}_{\boldsymbol{X}_s}^\top \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\boldsymbol{U}_t\right\| \\
&\le 1 + 10\mu\|\boldsymbol{X}\|^3\sigma_{\min}^{-1}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right)
\end{aligned}$$

so that

$$\left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-\frac{1}{2}} K_1\right\| \le \left[1 + 10\mu\|\boldsymbol{X}\|^3\sigma_{\min}^{-1}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right)\right]\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\right\|\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|.$$

Plugging into (53), we deduce that

$$\begin{aligned}
&\left\|\boldsymbol{W}_{t,\perp}^\top \boldsymbol{W}_{t+1}\right\| \\
&\le 3\mu\left(1 + 10\mu\|\boldsymbol{X}\|^3\sigma_{\min}^{-1}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right)\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|\|\boldsymbol{X}\|\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \\
&\quad + \mu\sigma_{\min}^{-1}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \\
&\le 3\mu\left(\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| + 10\mu\|\boldsymbol{X}\|^3\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|\|\boldsymbol{X}\| \\
&\quad + \mu\sigma_{\min}^{-1}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \\
&\le 3\mu\left(10\mu\|\boldsymbol{X}\|^2 + c_4\right)c_3\|\boldsymbol{X}\| + \mu\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\|.
\end{aligned}$$

where in the last step we use Lemma D.6 and the induction hypothesis which implies that $\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t) \ge \|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$. □

Lemma D.8. *The matrix $\boldsymbol{H}$ defined in the proof of Lemma C.9 satisfies the following:*

$$\boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{H})^{-\frac{1}{2}} = \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} + \boldsymbol{B}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} - \frac{1}{2}(\boldsymbol{I} + \boldsymbol{B})\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{B} + \boldsymbol{B}^\top\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} - \boldsymbol{D},$$

where $\|\boldsymbol{D}\| \le 30\|\boldsymbol{B}\|^2$.

Proof. By definition of $\boldsymbol{H}$ we have

$$\begin{aligned}
&\boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{H})^{-\frac{1}{2}} \\
&= (\boldsymbol{I} + \mu\boldsymbol{M})(\boldsymbol{I} + \boldsymbol{P})\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\left(\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{I} + \boldsymbol{P}^\top\right)(\boldsymbol{I} + \mu\boldsymbol{M})^2(\boldsymbol{I} + \boldsymbol{P})\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)^{-\frac{1}{2}} \\
&= (\boldsymbol{I} + \boldsymbol{B})\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\left[\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{I} + \boldsymbol{B}^\top + \boldsymbol{B} + \boldsymbol{B}^\top\boldsymbol{B}\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right]^{-\frac{1}{2}} \\
&= (\boldsymbol{I} + \boldsymbol{B})\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\left[\boldsymbol{I} + \underbrace{\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{B}^\top + \boldsymbol{B} + \boldsymbol{B}^\top\boldsymbol{B}\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}}_{=:\boldsymbol{\Theta}}\right]^{-\frac{1}{2}}.
\end{aligned}$$

It follows from (36) and our assumptions on c_3 and c_4 that

$$\begin{aligned}\|\boldsymbol{B}\| &\le \mu\left\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\| + 6\mu\left(c_3c_4\|\boldsymbol{X}\| + 50\|\boldsymbol{X}\|^2\delta\right)\\ &\le 10\mu\|\boldsymbol{X}\|^2 + 6\mu c_3\left(c_4+1\right)\|\boldsymbol{X}\| < 0.1\end{aligned}$$

(note that this step is independent and does not rely on earlier derivations in the proof of Lemma C.9), so by Taylor's formula, we have

$$\left\|(\boldsymbol{I}+\boldsymbol{\Theta})^{-\frac{1}{2}} - \boldsymbol{I} + \frac{1}{2}\boldsymbol{\Theta}\right\| \le 3\|\boldsymbol{\Theta}\|^2.$$

Hence,

$$\begin{aligned}&\left\|\boldsymbol{H}(\boldsymbol{H}^\top\boldsymbol{H})^{-\frac{1}{2}} - \left(\boldsymbol{V}_{U_t\boldsymbol{W}_t} + \boldsymbol{B}\boldsymbol{V}_{U_t\boldsymbol{W}_t} - \frac{1}{2}(\boldsymbol{I}+\boldsymbol{B})\boldsymbol{V}_{U_t\boldsymbol{W}_t}\boldsymbol{V}_{U_t\boldsymbol{W}_t}^\top\left(\boldsymbol{B}+\boldsymbol{B}^\top\right)\boldsymbol{V}_{U_t\boldsymbol{W}_t}\right)\right\|\\ =&\left\|(\boldsymbol{I}+\boldsymbol{B})\boldsymbol{V}_{U_t\boldsymbol{W}_t}\left((\boldsymbol{I}+\boldsymbol{\Theta})^{-\frac{1}{2}} - \boldsymbol{I} + \frac{1}{2}\boldsymbol{\Theta} - \frac{1}{2}\boldsymbol{V}_{U_t\boldsymbol{W}_t}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{V}_{U_t\boldsymbol{W}_t}\right)\right\|\\ \le&\left(1+\|\boldsymbol{B}\|\right)\left(3\|\boldsymbol{\Theta}\|^2 + \frac{1}{2}\|\boldsymbol{B}\|^2\right) < 30\|\boldsymbol{B}\|^2\end{aligned}$$

as desired. □

E. Proofs for the Landscape Results in Section 4.1

In this section, we study the landscape of under-parameterized matrix sensing problem

$$f_s(\boldsymbol{U}) = \frac{1}{2}\left\|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}\boldsymbol{X}^\top)\right\|_2^2, \quad \boldsymbol{U}\in\mathbb{R}^{d\times s}$$

Our key result in this section is Lemma 4.6, which states a local RSI condition for the matrix sensing loss. Most existing results only study the landscape of (1) in the exact- and over-parameterized case. Zhu et al. (2021) have studied the landscape of under-parameterized matrix factorization problem, but their main focus is the strict-saddle property of the loss.

E.1. Analysis of the matrix factorization loss

When the measurement satisfies the RIP condition, we can expect that the landscape of f_s looks similar to that of the (under-parameterized) matrix factorization loss:

$$F_s(\boldsymbol{U}) = \frac{1}{2}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}\boldsymbol{X}^\top\right\|_F^2, \quad \boldsymbol{U}\in\mathbb{R}^{d\times s}$$

for some $s < \hat{r}$.For this reason, we first look into the landscape of F_s before analyzing f_s.

Recall that $\boldsymbol{X}\boldsymbol{X}^\top = \sum_{i=1}^{r_*}\sigma_i^2\boldsymbol{v}_i\boldsymbol{v}_i^\top$. The critical points of $F_s(\boldsymbol{U})$ is characterized by the following lemma:

Lemma E.1. *$\boldsymbol{U}\in\mathbb{R}^{d\times s}$ is a critical point of $F_s(\boldsymbol{U})$ if and only if there exists an orthogonal matrix $\boldsymbol{R}\in\mathbb{R}^{s\times s}$, such that all columns of $\boldsymbol{U}\boldsymbol{R}$ are in $\{\sigma_i v_i : 1\le i\le r_*\}$.*

Proof. Assume WLOG that $\boldsymbol{X}\boldsymbol{X}^\top = \operatorname{diag}(\sigma_1^2,\sigma_2^2,\cdots,\sigma_r^2,0,\cdots,0) =: \boldsymbol{\Sigma}$. Let $\boldsymbol{U}$ be a critical point of F_s, then we have that $\left(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}\boldsymbol{X}^\top\right)\boldsymbol{U} = 0$. Let $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{U}^\top$, then $(\boldsymbol{\Sigma}-\boldsymbol{W})\boldsymbol{W} = 0$.

Since $\boldsymbol{W}$ is symmetric, so is $\boldsymbol{W}^2$, and we obtain that $\boldsymbol{\Sigma}\boldsymbol{W}$ is also symmetric. It is then easy to see that that if $\boldsymbol{\Sigma} = \operatorname{diag}\left(\lambda_1\boldsymbol{I}_{m_1},\cdots,\lambda_t\boldsymbol{I}_{m_t}\right)$ with $\lambda_1 > \lambda_2 > \cdots > \lambda_t \ge 0$, then $\boldsymbol{W}$ is also in block-diagonal form: $\boldsymbol{W} = \operatorname{diag}\left(\boldsymbol{W}_1,\boldsymbol{W}_2,\cdots,\boldsymbol{W}_t\right)$ where $\boldsymbol{W}_i\in\mathbb{R}^{m_i\times m_i}$. For each $1\le i\le t$, we then have the equation $(\lambda_i\boldsymbol{I}_{m_i} - \boldsymbol{W}_i)\boldsymbol{W}_i = 0$. Hence, there exists an orthogonal matrix $\boldsymbol{R}_i$ such that $\boldsymbol{R}_i^\top\boldsymbol{W}_i\boldsymbol{R}_i$ is a diagonal matrix where the diagonal entries are either 0 or $\sqrt{\lambda_i} = \sigma_i$. Let $\boldsymbol{R} = \operatorname{diag}\left(\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_t\right)$, then $\boldsymbol{R}^\top\boldsymbol{W}\boldsymbol{R}$ is diagonal and its nonzero diagonal entries form an s-subset of the multi-set $\{\sigma_i : 1\le i\le r_*\}$. The conclusion follows. □

In the case of $s=1$, the global minimizers of F_s are $\pm\sigma_1\boldsymbol{v}_1$, and we can show that F_s is locally strongly convex around these minimizers. Therefore, we can deduce that f is locally strongly-convex as well. Since our main focus is on $s>1$, we put these details in Appendix G. When $s>1$, $F_s(\boldsymbol{U})$ is not locally strongly-convex due to rotational invariance: if $\boldsymbol{U}$ is a global minimizer, then so is $\boldsymbol{U}\boldsymbol{R}$ for any orthogonal matrix $\boldsymbol{R}\in\mathbb{R}^{s\times s}$. Instead, we establish a Restricted Secant Inequality for F_s, as shown below.

Lemma E.2. *For $\boldsymbol{U} \in \mathbb{R}^{d\times s}$, let $\boldsymbol{R}$ be an orthogonal matrix that minimizes $\|\boldsymbol{U} - \boldsymbol{X}_s\boldsymbol{R}\|_F$. Suppose that $\operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}_s) \le 0.1\|\boldsymbol{X}\|^{-1}\tau$ (where we recall that $\tau = \min_{s\in[r_*]}\left(\sigma_s^2 - \sigma_{s+1}^2\right)$ is the eigengap of $\boldsymbol{X}\boldsymbol{X}^\top$), then we have*

$$\langle \nabla F_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{X}_s\boldsymbol{R}\rangle \ge 0.1\tau \cdot \operatorname{dist}^2(\boldsymbol{U}, \boldsymbol{X}_s).$$

Proof. Assume WLOG that $\boldsymbol{R} = \boldsymbol{I}$. Then by Lemma A.10, $\boldsymbol{U}^\top\boldsymbol{X}_s$ is symmetric and positive semi-definite. Let $\boldsymbol{H} = \boldsymbol{U} - \boldsymbol{X}_s$, then

$$\begin{aligned}\nabla F_s(\boldsymbol{U}) &= (\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{U}\\ &= \left[(\boldsymbol{H}+\boldsymbol{X}_s)(\boldsymbol{H}+\boldsymbol{X}_s)^\top - \boldsymbol{X}\boldsymbol{X}^\top\right](\boldsymbol{H}+\boldsymbol{X}_s).\end{aligned}$$

So we have

$$\begin{aligned}\langle \nabla F_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{X}_s\rangle &= \left\langle \left[(\boldsymbol{H}+\boldsymbol{X}_s)(\boldsymbol{H}+\boldsymbol{X}_s)^\top - \boldsymbol{X}\boldsymbol{X}^\top\right](\boldsymbol{H}+\boldsymbol{X}_s), \boldsymbol{H}\right\rangle\\ &= \operatorname{tr}\left(\boldsymbol{H}^\top\left[(\boldsymbol{H}+\boldsymbol{X}_s)(\boldsymbol{H}+\boldsymbol{X}_s)^\top - \boldsymbol{X}\boldsymbol{X}^\top\right]\boldsymbol{H} + \boldsymbol{H}^\top\left(\boldsymbol{H}\boldsymbol{H}^\top + \boldsymbol{H}\boldsymbol{X}_s^\top + \boldsymbol{X}_s\boldsymbol{H}^\top\right)\boldsymbol{X}_s\right)\\ &\ge -\operatorname{tr}\left(\boldsymbol{H}^\top\boldsymbol{X}_{s,\perp}\boldsymbol{X}_{s,\perp}^\top\boldsymbol{H}\right) - 3\|\boldsymbol{X}\|\|\boldsymbol{H}\|_F^3 + \operatorname{tr}\left(\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{X}_s^\top\boldsymbol{X}_s\right) \qquad (54a)\\ &\ge \left(\sigma_s^2 - \sigma_{s+1}^2\right)\|\boldsymbol{H}\|_F^2 - 3\|\boldsymbol{X}\|\|\boldsymbol{H}\|_F^3 \qquad (54b)\\ &\ge 0.1\tau\|\boldsymbol{H}\|_F^2\end{aligned}$$

where in (54a) we use $\operatorname{tr}\left((\boldsymbol{H}^\top\boldsymbol{X}_s)^2\right) \ge 0$ (since $\boldsymbol{H}^\top\boldsymbol{X}_s$ is symmetric as noticed in the beginning of the proof), and (54b) is because of

$$\operatorname{tr}\left(\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{X}_s^\top\boldsymbol{X}_s\right) \ge \sigma_{\min}\left(\boldsymbol{X}_s^\top\boldsymbol{X}_s\right)\cdot\operatorname{tr}\left(\boldsymbol{H}^\top\boldsymbol{H}\right) = \sigma_s^2\|\boldsymbol{H}\|_F^2$$

and

$$\begin{aligned}\operatorname{tr}\left(\boldsymbol{H}^\top\boldsymbol{X}_{s,\perp}\boldsymbol{X}_{s,\perp}^\top\boldsymbol{H}\right) &= \operatorname{tr}\left(\boldsymbol{H}^\top\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\boldsymbol{\Sigma}_{s,\perp}\boldsymbol{V}_{\boldsymbol{X}_s,\perp}^\top\boldsymbol{H}\right)\\ &\le \|\boldsymbol{\Sigma}_{s,\perp}\|\cdot\left\|\boldsymbol{H}^\top\boldsymbol{V}_{\boldsymbol{X}_s,\perp}\right\|_F^2 \le \sigma_{s+1}^2\|\boldsymbol{H}\|_F^2.\end{aligned}$$

□

Corollary E.3. *Under the conditions of Lemma E.2, we have $\|\nabla F_s(\boldsymbol{U})\|_F \ge 0.1\tau\operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}_s)$.*

E.2. Analysis of the matrix sensing loss

The following lemma states that the minimizer of matrix sensing loss is also near-optimal for the matrix factorization loss.

Lemma E.4. *Let $\boldsymbol{Z}_s^*$ be a best rank-s solution as defined in Definition 1.1, then we have*

$$\left\|\boldsymbol{Z}_s^* - \boldsymbol{X}\boldsymbol{X}^\top\right\|_F^2 \le \left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{X}\boldsymbol{X}^\top\right\|_F^2 + 10\delta\left\|\boldsymbol{X}\boldsymbol{X}^\top\right\|_F^2.$$

Proof. By the RIP property Definition 3.2 we have

$$\begin{aligned}\left\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{Z}_s^*\right\|_F^2 &\le (1-\delta)^{-1}\left\|\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{Z}_s^*\right)\right\|_2^2\\ &\le (1-\delta)^{-1}\left\|\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right)\right\|_2^2\\ &\le \frac{1+\delta}{1-\delta}\left\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F^2\\ &\le \left\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F^2 + 10\delta\|\boldsymbol{X}\boldsymbol{X}^\top\|_F^2,\end{aligned}$$

where the second inequality holds due to Definition 1.1. □

We now recall Lemma 4.8.

Lemma 4.8. *Under Assumption 3.3, we have $\operatorname{dist}(\boldsymbol{U}_s^*, \boldsymbol{X}_s) \le 40\delta\kappa_*\|\boldsymbol{X}\|_F$ for any global minimizer $\boldsymbol{U}_s^*$ of f_s. Moreover, $\left\|\boldsymbol{Z}_s^* - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F \le 160\delta\kappa_*\sqrt{r_*}\|\boldsymbol{X}\|^2$.*

We prove the statements in this lemma separately in Lemma E.5 and Corollary E.6 below.

Lemma E.5. *Suppose that Assumption 3.3 holds. Let $\boldsymbol{U}_s^*$ be a global minimizer of f_s, then we have*

$$\operatorname{dist}(\boldsymbol{U}_s^*, \boldsymbol{X}_s) \leq 40\delta\kappa\|\boldsymbol{X}\|_F.$$

Proof. Define

$$S = \left\{\boldsymbol{U} \in \mathbb{R}^{d\times s} : \operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}_s) < 0.1\kappa^{-1}\|\boldsymbol{X}\|\right\}.$$

First we can show that $\boldsymbol{U}_s^* \in S$. The main idea is to apply Lemma A.7. Indeed, it is easy to see that

$$\lim_{\|\boldsymbol{U}\|_F\to+\infty} F_s(\boldsymbol{U}) = +\infty,$$

so the condition *(1)* in Lemma A.7 holds. To check condition *(2)*, we separately analyze the two cases $\boldsymbol{U} \in \partial S$ and $\boldsymbol{U} \notin S$.

Firstly, let $\boldsymbol{U} \in \partial S$, i.e., $\operatorname{dist}^2(\boldsymbol{U}, \boldsymbol{X}_s) = 0.1\|\boldsymbol{X}\|^{-1}\tau$. Assume WLOG that $\operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}_s) = \|\boldsymbol{U} - \boldsymbol{X}_s\|_F$, then by Lemma E.2 we have

$$\begin{aligned} F_s(\boldsymbol{U}) - F_s(\boldsymbol{X}_s) &= \int_0^1 t\left\langle \nabla F_s(t\boldsymbol{U} + (1-t)\boldsymbol{X}_s), \boldsymbol{U} - \boldsymbol{X}_s\right\rangle \mathrm{d}t \\ &\geq \int_0^1 0.1\tau t^2 \left\|\boldsymbol{U} - \boldsymbol{X}_s\right\|_F^2 \mathrm{d}t \\ &\geq 10^{-3}\|\boldsymbol{X}\|^{-2}\tau^3 = 10^{-3}\kappa^{-3}\|\boldsymbol{X}\|^2. \end{aligned}$$

Secondly, let $\boldsymbol{U} \notin S$ be a stationary point of f_s. Recall that all the stationary points of F_s are characterized in Lemma E.1, so that for all $\boldsymbol{U} \notin S$ with $\nabla F_s(\boldsymbol{U}) = 0$, we have

$$F_s(\boldsymbol{U}) - F_s^* \geq 0.5\left(\sigma_s^4 - \sigma_{s+1}^4\right) \geq 0.5\tau^2.$$

On the other hand, we know from Lemma E.4 that

$$F_s(\boldsymbol{U}_s^*) - F_s^* \leq 5\delta r_*\|\boldsymbol{X}\|^4. \tag{55}$$

By Assumption 3.3, we have $5\delta r_*\|\boldsymbol{X}\|^4 < 10^{-3}\kappa^{-3}\|\boldsymbol{X}\|^2 < 0.5\tau^2$, so Lemma A.7 implies that $\boldsymbol{U}_s^* \in S$.

Since $\nabla f_s(\boldsymbol{U}_s^*) = 0$, we have $\mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{U}_s^* = 0$, so that

$$\begin{aligned} \|\nabla F_s(\boldsymbol{U}_s^*)\|_F &= \left\|\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{U}_s^*\right\|_F \\ &= \left\|\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{U}_s^*\right\|_F \\ &\leq \delta\left\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right\|_F\|\boldsymbol{U}_s^*\| \\ &\leq 4\delta\|\boldsymbol{X}\|\cdot\left\|\boldsymbol{X}\boldsymbol{X}^\top\right\|_F. \end{aligned}$$

From $\boldsymbol{U}_s^* \in S$ and Corollary E.3 we can deduce that

$$\operatorname{dist}(\boldsymbol{U}_s^*, \boldsymbol{X}_s) \leq 40\delta\tau^{-1}\|\boldsymbol{X}\|^2\|\boldsymbol{X}\|_F = 40\delta\kappa\|\boldsymbol{X}\|_F.$$

□

Corollary E.6. *Suppose that Assumption 3.3 holds, then we have $\left\|\boldsymbol{Z}_s^* - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F \leq 80\delta\kappa\sqrt{r_*}\|\boldsymbol{X}\|^2$ and $\sigma_{\min}\left(\left(\boldsymbol{U}_s^*\right)^\top\boldsymbol{U}_s^*\right) \geq \sigma_s^2 - 80\delta\kappa\sqrt{r_*}\|\boldsymbol{X}\|^2$.*

Proof. We assume WLOG that $\|\boldsymbol{U}_s^* - \boldsymbol{X}_s\|_F = \operatorname{dist}(\boldsymbol{U}_s^*, \boldsymbol{X}_s)$ i.e. $\boldsymbol{R} = \boldsymbol{I}$ in Definition 3.8. By Lemma 4.8, we have that

$$\begin{aligned} \left\|\boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F &\leq 2\max\left\{\|\boldsymbol{U}_s^*\|, \|\boldsymbol{X}_s\|\right\}\cdot\|\boldsymbol{U}_s^* - \boldsymbol{X}_s\|_F \\ &\leq 160\delta\kappa\|\boldsymbol{X}\|\|\boldsymbol{X}\|_F \leq 160\delta\kappa\sqrt{r_*}\|\boldsymbol{X}\|^2. \end{aligned}$$

which proves the first inequality. Similarly, we have

$$\left\|\left(\boldsymbol{U}_s^*\right)^\top \boldsymbol{U}_s^* - \boldsymbol{X}_s^\top \boldsymbol{X}_s\right\|_F \le 160\delta\kappa\sqrt{r_*}\|\boldsymbol{X}\|^2.$$

Hence $\sigma_s^2 - \sigma_{\min}\left(\left(\boldsymbol{U}_s^*\right)^\top \boldsymbol{U}_s^*\right) \le \left\|\left(\boldsymbol{U}_s^*\right)^\top \boldsymbol{U}_s^* - \boldsymbol{X}_s^\top \boldsymbol{X}_s\right\| \le 160\delta\kappa\sqrt{r_*}\|\boldsymbol{X}\|^2$, as desired. □

Corollary 4.9. *Under Assumption 3.3, we have* $\sigma_{\min}(\boldsymbol{U}_s^*) \ge \frac{1}{2}\sigma_{\min}(\boldsymbol{X}_s) = \frac{1}{2}\sigma_s \ge \frac{1}{2}\kappa_*^{-\frac{1}{2}}\|\boldsymbol{X}\|$.

Proof. Assumption 3.3 implies that $160\delta\kappa\sqrt{r_*}\|\boldsymbol{X}\|^2 \le 0.1\kappa^{-1}\|\boldsymbol{X}\|^2 \le 0.1\sigma_s^2$, so that the conclusion immediately follows from Corollary E.6. □

Lemma E.7. *Under Assumption 3.3, suppose that* $\boldsymbol{U}, \boldsymbol{U}_s^* \in \mathbb{R}^{d\times s}$ *such that* $\boldsymbol{U}_s^*$ *is a global minimizer of* f_s *and* $\|\boldsymbol{U} - \boldsymbol{U}_s^*\|_F = \mathrm{dist}(\boldsymbol{U}, \boldsymbol{U}_s^*) \le 10^{-2}\kappa^{-1}\|\boldsymbol{X}\|$ *(recall that* dist *is defined in Definition 3.8), then we have*

$$\langle \nabla f_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{U}_s^*\rangle \ge 0.1\kappa^{-1}\|\boldsymbol{X}\|^2\|\boldsymbol{U} - \boldsymbol{U}_s^*\|_F^2.$$

Proof. By Lemma A.10, $\boldsymbol{U}^\top\boldsymbol{U}_s^*$ is symmetric and positive semi-definite. Let $\boldsymbol{H} = \boldsymbol{U} - \boldsymbol{U}_s^*$, then

$$\begin{aligned}
\nabla f_s(\boldsymbol{U}) &= \left(\mathcal{A}^*\mathcal{A}\right)\left(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}\boldsymbol{X}^\top\right)\boldsymbol{U} \\
&= \left(\mathcal{A}^*\mathcal{A}\right)\left[(\boldsymbol{H}+\boldsymbol{U}_s^*)(\boldsymbol{H}+\boldsymbol{U}_s^*)^\top - \boldsymbol{X}\boldsymbol{X}^\top\right](\boldsymbol{H}+\boldsymbol{U}_s^*) \\
&= \left[\left(\mathcal{A}^*\mathcal{A}\right)\left(\boldsymbol{H}\boldsymbol{H}^\top + \boldsymbol{U}_s^*\boldsymbol{H}^\top + \boldsymbol{H}\left(\boldsymbol{U}_s^*\right)^\top\right)\right](\boldsymbol{H}+\boldsymbol{U}_s^*) - \mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{H}
\end{aligned}$$

where we use the first-order optimality condition

$$\mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{U}_s^* = 0.$$

Since $\|\boldsymbol{U}_s^*\| \le 2\|\boldsymbol{X}\|$ by Lemma 4.8, we may thus deduce that

$$\begin{aligned}
&\left\|\nabla f_s(\boldsymbol{U}) - \left[\left(\boldsymbol{H}\boldsymbol{H}^\top + \boldsymbol{U}_s^*\boldsymbol{H}^\top + \boldsymbol{H}\left(\boldsymbol{U}_s^*\right)^\top\right)(\boldsymbol{H}+\boldsymbol{U}_s^*) - \left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{H}\right]\right\|_F \\
\le& \left\|\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{H}\boldsymbol{H}^\top + \boldsymbol{U}_s^*\boldsymbol{H}^\top + \boldsymbol{H}\left(\boldsymbol{U}_s^*\right)^\top\right)(\boldsymbol{H}+\boldsymbol{U}_s^*)\right\|_F + \left\|\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{H}\right\| \\
\le& 50\delta\|\boldsymbol{X}\|^2\|\boldsymbol{H}\|_F
\end{aligned}$$

Hence

$$\begin{aligned}
&\langle \nabla f_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{U}_s^*\rangle \\
\ge& \left\langle \left(\boldsymbol{H}\boldsymbol{H}^\top + \boldsymbol{U}_s^*\boldsymbol{H}^\top + \boldsymbol{H}\left(\boldsymbol{U}_s^*\right)^\top\right)(\boldsymbol{H}+\boldsymbol{U}_s^*) - \left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{H}, \boldsymbol{H}\right\rangle - 50\delta\|\boldsymbol{X}\|^2\|\boldsymbol{H}\|_F^2 \\
\ge& \operatorname{tr}\left(\boldsymbol{H}(\boldsymbol{H}+\boldsymbol{U}_s^*)^\top(\boldsymbol{H}+\boldsymbol{U}_s^*)\boldsymbol{H}^\top + \boldsymbol{H}^\top\boldsymbol{U}_s^*\boldsymbol{H}^\top\boldsymbol{H} + \left(\left(\boldsymbol{U}_s^*\right)^\top\boldsymbol{H}\right)^2\right. \\
&\left. -\boldsymbol{H}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{H}\right) - 50\delta\|\boldsymbol{X}\|^2\|\boldsymbol{H}\|_F^2 \\
\ge& \left[\sigma_{\min}\left(\left(\boldsymbol{U}_s^*\right)^\top\boldsymbol{U}_s^*\right) - \left\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right\| - 50\delta\|\boldsymbol{X}\|^2 - 3\|\boldsymbol{U}_s^*\|\|\boldsymbol{H}\| - \|\boldsymbol{H}\|^2\right]\|\boldsymbol{H}\|_F^2.
\end{aligned}$$

By Corollary E.6 we have $\sigma_{\min}\left(\left(\boldsymbol{U}_s^*\right)^\top\boldsymbol{U}_s^*\right) \ge \sigma_s^2 - 80\delta\kappa\|\boldsymbol{X}\|\|\boldsymbol{X}\|_F$ and $\left\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right\| \le \sigma_{s+1}^2 + 80\delta\kappa\|\boldsymbol{X}\|_F^2$, so that

$$\langle \nabla f_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{U}_s^*\rangle \ge \left(\sigma_s^2 - \sigma_{s+1}^2 - 160\delta\kappa\|\boldsymbol{X}\|\|\boldsymbol{X}\|_F - 50\delta\|\boldsymbol{X}\|^2 - 3\|\boldsymbol{U}_s^*\|\|\boldsymbol{H}\| - \|\boldsymbol{H}\|^2\right)\|\boldsymbol{H}\|_F^2.$$

When Assumption 3.3 on δ is satisfied and $\|\boldsymbol{H}\| \le 10^{-2}\tau\|\boldsymbol{X}\|^{-1}$, the above implies that $\langle \nabla f_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{U}_s^*\rangle \ge 0.5\tau\|\boldsymbol{H}\|_F^2$, as desired. □

We now prove Lemmas 4.4 and 4.6 to conclude this section. Both results follow immediately from Lemma E.7.

Lemma 4.4. *Under Assumption 3.3, if $\boldsymbol{U}_s^* \in \mathbb{R}^{d\times s}$ is a global minimizer of f_s, then the set of global minimizers $\arg\min f_s$ is equal to $\left\{\boldsymbol{U}_s^*\boldsymbol{R} : \boldsymbol{R} \in \mathbb{R}^{s\times s}, \boldsymbol{R}^\top\boldsymbol{R} = \boldsymbol{I}\right\}$.*

Proof. By Lemma E.5 we have that $\operatorname{dist}(\boldsymbol{U}_s^*, \boldsymbol{X}_s^*) \le 40\delta\kappa\|\boldsymbol{X}\|_F$ holds for *any* $\boldsymbol{U}_s^* \in \arg\min f_s$. Suppose now that $\boldsymbol{U}_s^*, \hat{\boldsymbol{U}}_s^* \in \arg\min f_s$ such that $\operatorname{dist}(\boldsymbol{U}_s^*, \hat{\boldsymbol{U}}_s^*) > 0$. Then we also have that $\operatorname{dist}(\boldsymbol{U}_s^*, \hat{\boldsymbol{U}}_s^*) \le \operatorname{dist}(\boldsymbol{U}_s^*, \boldsymbol{X}_s) + \operatorname{dist}(\boldsymbol{X}_s, \hat{\boldsymbol{U}}_s^*) \le 80\delta\kappa\|\boldsymbol{X}\|_F < 10^{-2}\kappa^{-1}\|\boldsymbol{X}\|$, where the last inequality follows from Assumption 3.3. Without loss of generality, we can assume that $\|\boldsymbol{U}_s^* - \hat{\boldsymbol{U}}_s^*\|_F = \operatorname{dist}(\boldsymbol{U}_s^*, \hat{\boldsymbol{U}}_s^*)$, so that we can apply Lemma E.7 to obtain

$$\left\langle \nabla f_s(\hat{\boldsymbol{U}}_s^*), \hat{\boldsymbol{U}}_s^* - \boldsymbol{U}_s^* \right\rangle \ge 0.1\kappa^{-1}\|\boldsymbol{X}\|^2\|\hat{\boldsymbol{U}}_s^* - \boldsymbol{U}_s^*\|_F^2.$$

However, since $\hat{\boldsymbol{U}}_s^*$ is a global minimizer of f_s, we have $\nabla f_s(\hat{\boldsymbol{U}}_s^*)$ which is a contradiction. Thus the global minimizer must be unique under the procrustes distance. □

We recall Definition 4.5 which is now guaranteed to be well-defined by Lemma 4.4.

Definition 4.5. For any $\boldsymbol{U} \in \mathbb{R}^{d\times s}$, we use $\Pi_s(\boldsymbol{U})$ to denote the set of *closest* global minimizers of f_s to $\boldsymbol{U}$, namely $\Pi_s(\boldsymbol{U}) = \arg\min\{\|\boldsymbol{U} - \boldsymbol{U}_s^*\|_F : \boldsymbol{U}_s^* \in \arg\min f_s\}$.

Equipped with this definition, Lemma E.7 directly translates into Lemma 4.6:

Lemma 4.6 (Restricted Secant Inequality)**.** *Under Assumption 3.3, if a matrix $\boldsymbol{U} \in \mathbb{R}^{d\times s}$ satisfies $\|\boldsymbol{U} - \boldsymbol{U}_s^*\|_F \le 10^{-2}\kappa_*^{-1}\|\boldsymbol{X}\|$ for some $\boldsymbol{U}_s^* \in \Pi_s(\boldsymbol{U})$, then we have*

$$\langle \nabla f_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{U}_s^* \rangle \ge 0.1\kappa_*^{-1}\|\boldsymbol{X}\|^2\|\boldsymbol{U} - \boldsymbol{U}_s^*\|_F^2. \tag{8}$$

F. Proofs for Theorems 4.1 and 4.2

In this section, we prove the main theorems based on our key lemmas introduced in Section 4.1.

Based on Lemma 4.6, we first prove the following lemma, which shows that GD initialized near global minimizers converges linearly.

Lemma F.1. *Suppose that Assumptions 3.3 and 3.7 hold. Let $\{\hat{\boldsymbol{U}}_t\}_{t\ge 0}$ be a trajectory of GD that optimizes f_s with step size μ, starting with $\hat{\boldsymbol{U}}_0$. Also let $\boldsymbol{U}_s^*$ be a global minimizer of f_s. If $\operatorname{dist}(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_s^*) \le 10^{-2}\kappa^{-1}\|\boldsymbol{X}\|$, then for all $t \ge 0$,*

$$\operatorname{dist}^2(\hat{\boldsymbol{U}}_t, \boldsymbol{U}_s^*) \le (1 - 0.05\tau\mu)^t \operatorname{dist}^2(\hat{\boldsymbol{U}}_{\alpha,0}, \boldsymbol{U}_s^*). \tag{56}$$

Proof. We prove (56) by induction. It is easy to check that (56) holds for $t = 0$. Now we show that (56) holds for $t+1$ assuming it holds for t.

Since $\operatorname{dist}(\hat{\boldsymbol{U}}_0, \boldsymbol{U}_s^*) \le 10^{-2}\kappa^{-1}\|\boldsymbol{X}\|$, we have

$$\left\|\hat{\boldsymbol{U}}_0\right\| \le \|\boldsymbol{U}_s^*\| + 10^{-2}\kappa^{-1}\|\boldsymbol{X}\| \le 2\|\boldsymbol{X}\|. \tag{57}$$

Let $\boldsymbol{R}$ be the orthogonal matrix such that $\boldsymbol{U}_s^*\boldsymbol{R} \in \Pi(\boldsymbol{U}_t)$, then $\|\boldsymbol{U} - \boldsymbol{U}_s^*\boldsymbol{R}\|_F = \operatorname{dist}(\boldsymbol{U}_t, \boldsymbol{U}_s^*)$. We first bound the gradient $\nabla f(\hat{\boldsymbol{U}}_{\alpha,t})$ as follows:

$$\begin{aligned}
\left\|\nabla f(\hat{\boldsymbol{U}}_{\alpha,t})\right\|_F &= \left\|\mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \hat{\boldsymbol{U}}_{\alpha,t}\hat{\boldsymbol{U}}_{\alpha,t}^\top\right)\hat{\boldsymbol{U}}_{\alpha,t}\right\|_F \\
&\le \left\|\mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \hat{\boldsymbol{U}}_{\alpha,t}\hat{\boldsymbol{U}}_{\alpha,t}^\top\right)\right\| \left\|\hat{\boldsymbol{U}}_{\alpha,t} - \boldsymbol{U}_s^*\right\|_F + \left\|\left(\hat{\boldsymbol{U}}_{\alpha,t}\hat{\boldsymbol{U}}_{\alpha,t}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{U}_s^*\right\|_F \\
&\le 20\|\boldsymbol{X}\|^2\left\|\hat{\boldsymbol{U}}_{\alpha,t} - \boldsymbol{U}_s^*\right\|_F
\end{aligned} \tag{58}$$

where we use (57) and the RIP property (Assumption 3.3). It follows that

$$\begin{aligned}
\operatorname{dist}^2(\hat{\boldsymbol{U}}_{t+1}, \boldsymbol{U}_s^*) &\le \left\|\hat{\boldsymbol{U}}_{t+1} - \boldsymbol{U}_s^*\boldsymbol{R}\right\|_F^2 && \text{(59a)}\\
&= \left\|\hat{\boldsymbol{U}}_t - \mu\nabla f(\hat{\boldsymbol{U}}_{\alpha,t}) - \boldsymbol{U}_s^*\boldsymbol{R}\right\|_F^2 \\
&= \left\|\hat{\boldsymbol{U}}_{\alpha,t} - \boldsymbol{U}_s^*\boldsymbol{R}\right\|_F^2 - \mu\left\langle\nabla f(\hat{\boldsymbol{U}}_{\alpha,t}), \hat{\boldsymbol{U}}_t - \boldsymbol{U}_s^*\boldsymbol{R}\right\rangle + \mu^2\left\|\nabla f(\hat{\boldsymbol{U}}_{\alpha,t})\right\|_F^2 \\
&\le \left(1 - 0.1\tau\mu + 400\|\boldsymbol{X}\|^4\mu^2\right)\left\|\hat{\boldsymbol{U}}_{\alpha,t} - \boldsymbol{U}_s^*\boldsymbol{R}\right\|_F^2 && \text{(59b)}
\end{aligned}$$

where (59a) follows from the definition of dist, and (59b) is due to Lemma 4.6 and (58). Finally, (56) follows from Assumption 3.7. □

We then note the following proposition, which is straightforward from Lemma C.12 and Theorem C.15. In the following we use $\boldsymbol{U}_{\alpha,t}$ to denote the t-th iteration of GD when initialized at $\boldsymbol{U}_0 = \alpha\bar{\boldsymbol{U}}$.

Proposition F.2. *Suppose that Assumptions 3.3, 3.5 and 3.7 hold and α is sufficiently small. Then there exist matrices $\boldsymbol{U}_{\alpha,t}^{\mathtt{lr}}$ for $t = -T_{\alpha,s}^{\mathtt{ft}}, -T_{\alpha,s}^{\mathtt{ft}}+1, \cdots, 0$ with rank $\le s$ (where $\mathtt{lr}$ stands for low rank) and a constant $C_6 = C_6(\boldsymbol{X}, \bar{\boldsymbol{U}})$ (defined in (49)) such that*

$$\max_{T_{\alpha,s}^{\mathtt{ft}} \le t \le 0}\left\|\boldsymbol{U}_{\alpha,t}^{\mathtt{lr}} - \boldsymbol{U}_{\alpha,T_{\alpha,s}^{\mathtt{ft}}+t}\right\|_F = C_6 \cdot \alpha^{\frac{1}{4\kappa}}$$

where $T_{\alpha,s}^{\mathtt{ft}}$ is defined in Theorem C.15 and moreover

$$\left\|\boldsymbol{U}_{\alpha,0}^{\mathtt{lr}}\left(\boldsymbol{U}_{\alpha,0}^{\mathtt{lr}}\right)^\top - \boldsymbol{Z}_s^*\right\|_F \le 2\times 10^5\kappa^3\|\boldsymbol{X}\|^2 r_*\delta.$$

where $\boldsymbol{Z}_s^* = \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top$ is the best rank-s solution as defined in Definition 1.1.

Proof. It follows from Corollary C.16 that $\max_{1\le t\le T_{\alpha,s}^{\mathtt{ft}}}\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \le C_6(\boldsymbol{X},\bar{\boldsymbol{U}})\cdot\alpha^{\frac{1}{4\kappa}}$ (recall that C_5 is defined in Lemma C.11 and $T_{\alpha,s}^{\mathtt{ft}}$ is defined in (48)). We choose $\boldsymbol{U}_{\alpha,t}^{\mathtt{lr}} = \boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t}\boldsymbol{W}_{T_{\alpha,s}^{\mathtt{ft}}+t}\boldsymbol{W}_{T_{\alpha,s}^{\mathtt{ft}}+t}^\top$, then $\operatorname{rank}(\bar{\boldsymbol{U}}_t) \le s$ and moreover by Theorem C.15 we have $\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_{\alpha,0}^{\mathtt{lr}}\left(\boldsymbol{U}_{\alpha,0}^{\mathtt{lr}}\right)^\top\right\|_F \le 10^5\kappa^3\|\boldsymbol{X}\|^2 r_*\delta$. On the other hand, by Lemma 4.8 we have that $\left\|\boldsymbol{Z}_s^* - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F \le 80\delta\kappa\sqrt{r_*}\|\boldsymbol{X}\|^2$. Thus $\left\|\boldsymbol{U}_{\alpha,0}^{\mathtt{lr}}\left(\boldsymbol{U}_{\alpha,0}^{\mathtt{lr}}\right)^\top - \boldsymbol{Z}_s^*\right\|_F \le 2\times 10^5\kappa^3\|\boldsymbol{X}\|^2 r_*\delta$ as desired. □

Let $\hat{\boldsymbol{U}}_{\alpha,0} = \boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}}\boldsymbol{W}_{T_{\alpha,s}^{\mathtt{ft}}} \in \mathbb{R}^{d\times s}$, then it satisfies $\hat{\boldsymbol{U}}_{\alpha,0}\hat{\boldsymbol{U}}_{\alpha,0}^\top = \boldsymbol{U}_{\alpha,0}^{\mathtt{lr}}\left(\boldsymbol{U}_{\alpha,0}^{\mathtt{lr}}\right)^\top$. The following corollary shows that $\hat{\boldsymbol{U}}_{\alpha,0}$ is close to $\boldsymbol{U}_s^*$ in terms of the procrustes distance.

Corollary F.3. *We have* $\operatorname{dist}(\hat{\boldsymbol{U}}_{\alpha,0}, \boldsymbol{U}_s^*) \le 3\times 10^6\kappa^4 r_*\|\boldsymbol{X}\|\delta$.

Proof. We know from Lemma 4.8 that $\operatorname{dist}(\boldsymbol{U}_s^*, \boldsymbol{X}_s) \le 40\delta\kappa\|\boldsymbol{X}\|_F$, so it remains to bound $\operatorname{dist}(\hat{\boldsymbol{U}}_{\alpha,0}, \boldsymbol{X}_s)$.

The proof idea is the same as that of Lemma 4.8, so we only provide a proof sketch here. It has been shown in the proof of Proposition F.2 that

$$F_s(\hat{\boldsymbol{U}}_{\alpha,0}) := \frac{1}{2}\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \hat{\boldsymbol{U}}_{\alpha,0}\hat{\boldsymbol{U}}_{\alpha,0}^\top\right\|_F^2 \le r_*\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \hat{\boldsymbol{U}}_{\alpha,0}\hat{\boldsymbol{U}}_{\alpha,0}^\top\right\|^2 \le 4\times 10^{10}\kappa^6 r_*^2\|\boldsymbol{X}\|^4\delta^2 \le 0.5\tau^2.$$

Note that F_s is the matrix factorization loss with $\boldsymbol{X}_s\boldsymbol{X}_s^\top$ being the ground-truth, so the local RSI condition (Lemma E.2) still holds. By the same reason as (55), we deduce that $\operatorname{dist}(\hat{\boldsymbol{U}}_{\alpha,0}, \boldsymbol{X}_s) \le 0.1\|\boldsymbol{X}\|^{-1}\tau$, i.e., $\hat{\boldsymbol{U}}_{\alpha,0}$ is in the local region around $\boldsymbol{X}_s$ in which the RSI condition holds. Finally, it follows from the local RSI condition that

$$\operatorname{dist}(\hat{\boldsymbol{U}}_{\alpha,0}, \boldsymbol{X}_s) \le 10\tau^{-1}\left\|\nabla F_s(\hat{\boldsymbol{U}}_{\alpha,0})\right\|_F \le 10\tau^{-1}\|\hat{\boldsymbol{U}}_{\alpha,0}\|\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \hat{\boldsymbol{U}}_{\alpha,0}\hat{\boldsymbol{U}}_{\alpha,0}^\top\right\|_F \le 3\times 10^6\kappa^4 r_*\|\boldsymbol{X}\|\delta.$$

The conclusion follows. □

We are now ready to complete the proof of Theorems 4.1 and 4.2.

Theorem 4.2 (Convergence in the under-parameterized regime)**.** *Suppose that $\hat{r} \le r^*$, then there exists a constant $\bar{\alpha} > 0$ such that when $\alpha < \bar{\alpha}$, we have $\lim_{t\to+\infty} \boldsymbol{U}_{\alpha,t}\boldsymbol{U}_{\alpha,t}^\top = \boldsymbol{Z}_{\hat{r}}^*$.*

Proof. When $\hat{r} \le r_*$, the parameterization itself ensures that $\boldsymbol{U}_{\alpha,t}$ is low-rank, so that we can choose $\boldsymbol{U}_{\alpha,t}^{\mathtt{lr}} = \boldsymbol{U}_{\alpha,T_{\alpha,\hat{r}}^{\mathtt{ft}}+t}$ and $\hat{\boldsymbol{U}}_{\alpha,0} = \boldsymbol{U}_{\alpha,T_{\alpha,s}^{\mathtt{ft}}}$ in Proposition F.2 and Corollary F.3 (for $s = \hat{r}$). The proof that these choices satisfy all required conditions are identical to our proofs for these two lemmas in the general setting, and we omit them here.

Applying Lemma F.1, we can thus deduce that $\lim_{t\to+\infty} \operatorname{dist}(\hat{\boldsymbol{U}}_{\alpha,t}, \boldsymbol{U}_{\hat{r}}^*) = 0$. This means that $\lim_{t\to+\infty} \operatorname{dist}(\boldsymbol{U}_{\alpha,t}, \boldsymbol{U}_{\hat{r}}^*) = 0$. Recall that $\boldsymbol{Z}_{\hat{r}}^* = \boldsymbol{U}_{\hat{r}}^*\boldsymbol{U}_{\hat{r}}^{*\top}$, so the conclusion immediately follows. □

Theorem 4.1. *Under Assumptions 3.3, 3.5 and 3.7, consider GD (3) with initialization $\boldsymbol{U}_{\alpha,0} = \alpha\bar{\boldsymbol{U}}$ for solving the matrix sensing problem (1). There exist universal constants c, M, constant $C = C(\boldsymbol{X}, \bar{\boldsymbol{U}})$ and a sequence of time points $T_\alpha^1 < T_\alpha^2 < \cdots < T_\alpha^{\hat{r}\wedge r_*}$ such that for all $1 \le s \le \hat{r} \wedge r_*$, the following holds when α is sufficiently small:*

$$\left\|\boldsymbol{U}_{\alpha,T_\alpha^s}\boldsymbol{U}_{\alpha,T_\alpha^s}^\top - \boldsymbol{Z}_s^*\right\|_{\mathrm{F}} \le C\alpha^{\frac{1}{M\kappa_*^2}}, \tag{5}$$

where we recall that $\boldsymbol{Z}_s^*$ is the best rank-s solution defined in Definition 1.1. Moreover, GD follows an incremental learning procedure: we have $\lim_{\alpha\to 0} \max_{1\le t\le T_\alpha^s} \sigma_{s+1}(\boldsymbol{U}_{\alpha,t}) = 0$ for all $1 \le s \le \hat{r} \wedge r_*$, where $\sigma_i(\boldsymbol{A})$ denotes the i-th largest singular value of a matrix $\boldsymbol{A}$.

Proof. Recall that $\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}} - \bar{\boldsymbol{U}}_0\right\|_F = o(1)$ $(\alpha \to 0)$ where $T_{\alpha,s}^{\mathtt{ft}}$ is defined in Proposition F.2; we omit the dependence on α to simplify notations. We also note that by the update of GD, we have $\bar{\boldsymbol{U}}_t\bar{\boldsymbol{U}}_t^\top = \hat{\boldsymbol{U}}_{\alpha,t}\hat{\boldsymbol{U}}_{\alpha,t}^\top$ for all $t \ge 0$.

By Lemma F.1, we have that $\operatorname{dist}^2(\hat{\boldsymbol{U}}_{\alpha,t}, \boldsymbol{U}_s^*) \le (1 - 0.05\tau\mu)^t \operatorname{dist}^2(\hat{\boldsymbol{U}}_{\alpha,0}, \boldsymbol{U}_s^*)$ and, in particular, $\left\|\hat{\boldsymbol{U}}_{\alpha,t}\right\| \le 2\|\boldsymbol{X}\|$ for all t. Thus $\left\|\bar{\boldsymbol{U}}_t\right\| \le 2\|\boldsymbol{X}\|$ as well. Moreover, recall that $\|\boldsymbol{U}_t\| \le 3\|\boldsymbol{X}\|$ for all t. It's easy to see that that the matrix sensing loss f is L-smooth in $\left\{\boldsymbol{U} \in \mathbb{R}^{d\times r} : \|\boldsymbol{U}\| \le 3\|\boldsymbol{X}\|\right\}$ for some constant $L = \mathcal{O}(\|\boldsymbol{X}\|^2)$, so it follows from Lemma A.8 that

$$\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t} - \bar{\boldsymbol{U}}_t\right\|_F \le (1+\mu L)^t \left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}} - \bar{\boldsymbol{U}}_0\right\|_F.$$

On the other hand, since $\operatorname{dist}^2(\hat{\boldsymbol{U}}_{\alpha,t}, \boldsymbol{U}_s^*) \le (1 - 0.05\tau\mu)^t \operatorname{dist}^2(\hat{\boldsymbol{U}}_{\alpha,0}, \boldsymbol{U}_s^*)$, we can deduce that

$$\begin{aligned}
\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t}\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t}^\top - \boldsymbol{Z}_s\right\|_F &\le \left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t}\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t}^\top - \bar{\boldsymbol{U}}_t\bar{\boldsymbol{U}}_t^\top\right\|_F + \left\|\bar{\boldsymbol{U}}_t\bar{\boldsymbol{U}}_t^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right\|_F \\
&= \left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t}\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t}^\top - \bar{\boldsymbol{U}}_t\bar{\boldsymbol{U}}_t^\top\right\|_F + \left\|\hat{\boldsymbol{U}}_{\alpha,t}\hat{\boldsymbol{U}}_{\alpha,t}^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right\|_F \\
&\le 3\|\boldsymbol{X}\|\left(\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t} - \bar{\boldsymbol{U}}_t\right\|_F + \operatorname{dist}(\hat{\boldsymbol{U}}_{\alpha,t}, \boldsymbol{U}_s^*)\right) \\
&\le 3\|\boldsymbol{X}\|\left((1+\mu L)^t\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}} - \bar{\boldsymbol{U}}_0\right\|_F + (1-0.05\tau\mu)^{\frac{t}{2}}\operatorname{dist}^2(\hat{\boldsymbol{U}}_{\alpha,0}, \boldsymbol{U}_s^*)\right)
\end{aligned}$$

Since when $\alpha \to 0$, $\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}} - \bar{\boldsymbol{U}}_0\right\|_F = \mathcal{O}(\alpha^{\frac{1}{4\kappa}})$, it's easy to see that there exists a time $t = t_\alpha^s$ so that we have $\max_{-T_{\alpha,s}^{\mathtt{ft}} \le t \le t_\alpha^s} \left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t} - \bar{\boldsymbol{U}}_t\right\|_F = \mathcal{O}\left(\alpha^{\frac{1}{M_1\kappa^2}}\right)$ and $\left\|\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t}\boldsymbol{U}_{T_{\alpha,s}^{\mathtt{ft}}+t}^\top - \boldsymbol{Z}_s\right\|_F = \mathcal{O}\left(\alpha^{\frac{1}{M_1\kappa^2}}\right)$ as well, where c_1 is a universal constant. Let $T_\alpha^s = T_{\alpha,s}^{\mathtt{ft}} + t_\alpha^s$, then $\left\|\boldsymbol{U}_{T_\alpha^s}\boldsymbol{U}_{T_\alpha^s}^\top - \boldsymbol{Z}_s\right\|_F = o(1)$ holds. Recall that $\operatorname{rank}(\boldsymbol{U}_t) \le s$, so that $\max_{0\le t\le T_\alpha^s} \sigma_{s+1}(\boldsymbol{U}_t) = o(1)$. Finally, for all $0 \le s < \hat{r} \wedge r_*$, we need to show that $T_\alpha^s < T_\alpha^{s+1}$. Indeed, by Corollary E.6 and the Assumption 3.3 we have $\sigma_{s+1}^2\left(\boldsymbol{U}_{T_\alpha^s}\right) \ge \sigma_{s+1}\left(\boldsymbol{Z}_{s+1}\right) - o(1) \ge 0.5\sigma_{s+1}^2$, so that $T_\alpha^{s+1} > T_\alpha^s$, as desired. □

G. The landscape of matrix sensing with rank-1 parameterization

In this section, we establish a local strong-convexity result Lemma G.2 for rank-1 parameterized matrix sensing. This result is stronger than the RSI condition we established for general ranks, though the latter is sufficient for our analysis.

Lemma G.1. *Define the full-observation loss with rank-1 parameterization*

$$g_1(\boldsymbol{u}) = \frac{1}{4}\left\|\boldsymbol{u}\boldsymbol{u}^T - \boldsymbol{X}\boldsymbol{X}^T\right\|_F^2.$$

Then the global minima of g_1 are $\boldsymbol{u}^* = \sigma_1 v_1$ and $-\boldsymbol{u}^*$. Moreover, suppose that $g(\boldsymbol{u}) - g(\boldsymbol{u}^*) \le 0.5\tau_1$ where $\tau_1 = \sigma_1^2 - \sigma_2^2$ is the eigengap, then we must have

$$\|\boldsymbol{u} - \boldsymbol{u}^*\|^2 \le 20\tau_1^{-1}\left(g_1(\boldsymbol{u}) - g_1(\boldsymbol{u}^*)\right).$$

Proof. We can assume WLOG that $\boldsymbol{X}\boldsymbol{X}^T = \operatorname{diag}\big(\sigma_1^2, \cdots, \sigma_{r_*}^2, 0, \cdots, 0\big)$. Then

$$g_1(\boldsymbol{u}) = \frac{1}{4}\left(\|\boldsymbol{u}\|_2^4 - 2\sum_{i=1}^{s}\sigma_i^2\boldsymbol{u}_i^2 + \|\boldsymbol{X}^T\boldsymbol{X}\|_F^2\right) \tag{60a}$$

$$\ge \frac{1}{4}\left(\|\boldsymbol{u}\|_2^4 - 2\sigma_1^2\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{X}^T\boldsymbol{X}\|_F^2\right) \tag{60b}$$

$$\ge \frac{1}{4}\left(\|\boldsymbol{X}^T\boldsymbol{X}\|_F^2 - \sigma_1^4\right) \tag{60c}$$

where equality holds if and only if $\boldsymbol{u}_2 = \cdots = \boldsymbol{u}_d = 0$ and $\|\boldsymbol{u}\|^2 = \sigma_1^2$ i.e. $\boldsymbol{u} = \pm\sigma_1\boldsymbol{e}_1$. Moreover, suppose that $g_1(\boldsymbol{u}) - g_1(\boldsymbol{u}^*) \le 0.5\tau_1$, it follows from (60b) that $\tau_1\sum_{i=2}^d \boldsymbol{u}_i^2 \le 2(g_1(\boldsymbol{u}) - g_1(\boldsymbol{u}^*))$ which implies that $\sum_{i=2}^d \boldsymbol{u}_i^2 \le 2\tau_1^{-1}\left(g_1(\boldsymbol{u}) - g_1(\boldsymbol{u}^*)\right)$. Also (60c) yields $\left|\|\boldsymbol{u}\|^2 - \sigma_1^2\right| \le 4\sqrt{g_1(\boldsymbol{u}) - g_1(\boldsymbol{u}^*)}$. Assume WLOG that $\boldsymbol{u}_1 > 0$, then we have

$$\begin{aligned}\|\boldsymbol{u} - \sigma_1\boldsymbol{e}_1\|^2 &\le \sigma_1^{-2}\left(\boldsymbol{u}_1^2 - \sigma_1^2\right)^2 + \sum_{i=2}^d \boldsymbol{u}_i^2 \\ &\le 20\tau_1^{-1}\left(g_1(\boldsymbol{u}) - g_1(\boldsymbol{u}^*)\right).\end{aligned}$$

□

Lemma G.2. *Let*

$$f_1(\boldsymbol{u}) = \frac{1}{4}\left\|\mathcal{A}\left(\boldsymbol{u}\boldsymbol{u}^T - \boldsymbol{X}\boldsymbol{X}^T\right)\right\|_2^2, \quad \boldsymbol{u} \in \mathbb{R}^d.$$

Suppose that $\delta \le 10^{-3}\|\boldsymbol{X}\|^{-2}\tau_1$, then there exists constants a_1 and ι, such that f_1 is locally ι-strongly convex in $\mathcal{B}_1 = \mathcal{B}(\sigma_1 v_1, a_1) \subset \mathbb{R}^d$. Furthermore, there is a unique global minima of f_1 inside $\mathcal{B}_1$.

Proof. Recall that we defined the full observation loss $g_1(\boldsymbol{u}) = \frac{1}{4}\left\|\boldsymbol{u}\boldsymbol{u}^T - \boldsymbol{X}\boldsymbol{X}^T\right\|_F^2$. Let $h_1 = f_1 - g_1$, then

$$\begin{aligned}\left\|\nabla^2 h_1(\boldsymbol{u})\right\| &= \frac{1}{2}\left\|\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{u}\boldsymbol{u}^T - \boldsymbol{X}\boldsymbol{X}^T\right) + 2\left(\mathcal{A}^*\mathcal{A} - I\right)\boldsymbol{u}\boldsymbol{u}^T\right\| \\ &\le \delta\left(2\|\boldsymbol{u}\|^2 + \|\boldsymbol{X}\|^2\right).\end{aligned}$$

When $\|\boldsymbol{u} - \sigma_1 v_1\|^2 \le 0.1\min\left\{\sigma_1^2, \tau_1\right\}$ (recall $\tau_1 = \sigma_1^2 - \sigma_2^2$),

$$\sigma_{\min}\left(\nabla^2 g_1(\boldsymbol{u})\right) = \frac{1}{2}\sigma_{\min}\left(\|\boldsymbol{u}\|^2 I + 2\boldsymbol{u}\boldsymbol{u}^T - \boldsymbol{X}\boldsymbol{X}^T\right) \ge 0.4\tau_1.$$

Hence we have

$$\sigma_{\min}\left(\nabla^2 f_1(\boldsymbol{u})\right) \ge \left(\nabla^2 g_1(\boldsymbol{u})\right) - \|\nabla^2 h_1(\boldsymbol{u})\| \ge 0.4\tau_1 - 4\|\boldsymbol{X}\|^2\delta \ge 0.2\tau_1,$$

i.e. strong-convexity holds for $a_1^2 = 0.1\min\left\{\sigma_1^2, \tau_1\right\}$ and $\iota = 0.2\tau_1$.

Let $\boldsymbol{u}^*$ be a global minima of f_1, then we must have $\|\boldsymbol{u}^*\| \le 2\|\boldsymbol{X}\|$ (otherwise $f_1(\boldsymbol{u}) > f_1(0)$). We can thus deduce that

$$\begin{aligned}g_1(\boldsymbol{u}^*) &\le f_1(\boldsymbol{u}^*) + \frac{1}{4}\left|\left\langle\boldsymbol{u}\boldsymbol{u}^T - \boldsymbol{X}\boldsymbol{X}^T, (\mathcal{A}^*\mathcal{A} - I)(\boldsymbol{u}\boldsymbol{u}^T - \boldsymbol{X}\boldsymbol{X}^T)\right\rangle\right| \\ &\le f_1(\boldsymbol{u}) + 10\delta\|\boldsymbol{X}\|^2 \le g_1(\boldsymbol{u}) + 20\delta\|\boldsymbol{X}\|^2.\end{aligned}$$

It follows from Lemma G.1 and our assumption on δ that $\min\left\{\|\boldsymbol{u}^* - \sigma_1\boldsymbol{v}_1\|^2, \|\boldsymbol{u}^* + \sigma_1\boldsymbol{v}_1\|^2\right\} \le \frac{1}{2}a_1^2$. Moreover, by strong convexity, there exists only one global minima in $\mathcal{B}_1$, which concludes the proof. □