UNDERSTANDING INCREMENTAL LEARNING OF GRADIENT DESCENT: A FINE-GRAINED ANALYSIS OF MATRIX SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

The implicit bias of optimization algorithms such as gradient descent (GD) is believed to play an important role in generalization of modern machine learning methods such as deep learning. 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. With small initialization, we that GD behaves similarly to the greedy low-rank learning heuristics (Li et al., 2020) and follows an incremental learning procedure (Gissin et al., 2019). That is, 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 is stronger because it characterizes the whole learning process. Moreover, our analysis of the incremental learning procedure applies to the *under-parameterized* regime as well. As a key ingredient of our analysis, we observe that GD always follows an approximately low-rank trajectory and develops novel landscape properties for matrix sensing with low-rank parameterization. Finally, we conduct numerical experiments which 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., 2021; Sun, 2019). It has long been a mystery why simple algorithms such as Gradient Descent (GD) 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 attempts to provide a better understanding of the generalization aspect of GD by studying its *implicit bias* towards certain solutions. Specifically, it has been observed both theoretically (Gissin et al., 2019; Arora et al., 2019; Lyu et al., 2021) and empirically (Arpit et al., 2017; Valle-Perez et al., 2018; Kalimeris et al., 2019; Huh et al., 2021) that GD learns models starting from simple ones and continues learning with increasing complexity, which is commonly referred to as *simplicity bias/incremental learning* in the literature. This simplicity bias could be the reason why models with high memorization capacity, such as deep neural networks, do not overfit to training data, resolving the apparent paradox between their good generalization performance in practice and the worst-case generalization bounds in traditional statistical learning literature (Arpit et al., 2017). Hence, understanding the incremental learning phenomenon gives a more complete picture of the optimization algorithm dynamics and sheds light on why the solution returned by the algorithm generalizes well.

Toward better understanding this simplicity bias, we give a thorough analysis of the incremental learning dynamics of GD in the canonical matrix sensing model. Matrix sensing is a classical model for analyzing the behavior of GD, because it is mathematically simpler to deal with compared with other models such as deep neural networks, but still retains some of the key challenges including non-convexity in particular. Moreover, it arises in a wide range of real-world applications such as 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).

Formally, we consider the matrix sensing problem of recovering the ground-truth matrix $\boldsymbol{Z}^* \in \mathbb{R}^{d\times d}$ from m observations $y_i = \langle \boldsymbol{A}_i, \boldsymbol{Z}^* \rangle, \quad i = 1, 2, \cdots, m$ where $\{\boldsymbol{A}_i\}_{1\le i\le m}$ is a collection of symmetric measurement matrices. We assume that $\boldsymbol{Z}^*$ is positive semi-definite (PSD) and is of low-rank, i.e. $\operatorname{rank}(\boldsymbol{Z}) = r_* \ll d$. Hence, we may write $\boldsymbol{Z}^* = \boldsymbol{X}\boldsymbol{X}^\top$ where $\boldsymbol{X} \in \mathbb{R}^{d\times r_*}$ has orthogonal columns. A common approach for solving this problem is to consider the Burer-Monteiro type decomposition $\boldsymbol{Z}^* = \boldsymbol{U}\boldsymbol{U}^\top$ with $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ and minimize the corresponding square loss. In other words, we consider the following optimization problem:

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

In this paper, we refer to the case $r > r_*$ as the over-parameterized regime and the case $r < r_*$ as the under-parameterized regime.

It is first conjectured in Gunasekar et al. (2017) that for matrix sensing problems, gradient flow (GF) with small initialization finds the solution with minimum nuclear norm. This conjecture is then challenged by Arora et al. (2019), who studies matrix sensing with a single observation and finds that GD with small initialization tends to learn the singular components of ground-truth one by one. Such implicit low-rank bias is also theoretically investigated in different settings (Gissin et al., 2019; Li et al., 2020; Jacot et al., 2021) and observed in practice (Huh et al., 2021).

In spite of these progresses, theoretical understanding of the simplicity bias of GD remains limited. Indeed, a vast majority of existing analysis only shows that GD is initially biased towards learning a rank-1 solution, but their analysis cannot be generalized to higher ranks, unless additional assumptions on the GD dynamics are made (Li et al., 2020, Appendix H). To the best of our knowledge, the only exception is Jiang et al. (2022) for matrix factorization problems.

Consider solving (1) via the gradient descent (GD) algorithm $\boldsymbol{U}_{\alpha,0} = \alpha\boldsymbol{U}$, $\boldsymbol{U}_{\alpha,t+1} = \boldsymbol{U}_{\alpha,t} - \mu\nabla f(\boldsymbol{U}_{\alpha,t})$ where $\|\boldsymbol{U}\|_2 = 1$. $\alpha > 0$ is the initialization scale and $\mu > 0$ is the step size. Although (1) is non-convex, it has been shown that when $r \ge r_*$, GD with small initialization and early stopping can recover the ground-truth. In other words, there exists $\{T_\alpha : \alpha > 0\}$ such that $\lim_{\alpha\to 0}\left\|\boldsymbol{U}_{\alpha,T_\alpha}\boldsymbol{U}_{\alpha,T_\alpha}^\top - \boldsymbol{Z}^*\right\|_F = 0$ (Li et al., 2018; Stöger & Soltanolkotabi, 2021). Although partially revealing the algorithmic regularization of GD, these results do not characterize the learning trajectory of GD, thereby providing limited implications on *how* GD achieves generalization.

1.1 Our Contributions

In this paper, we take a step towards bridging this gap by providing a fine-grained analysis of GD for matrix sensing. Our main result is summarized in the following:

Theorem 1.1 *Suppose that* $\operatorname{rank}(\boldsymbol{X}) = r_*$ *has singular values* $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_{r_*} > 0$ *and the measurements* $\{\boldsymbol{A}_i\}_{i=1}^m$ *satisfy the* (δ, r_*)*-RIP condition (cf. Definition 3.1). For any* $1 \le s \le \min\{r, r_*\}$*, suppose that* $\sigma_s > \sigma_{s+1}$ *(with* $\sigma_{r_*+1} := 0$*) and* $\boldsymbol{U}_{\alpha,0} = \alpha\boldsymbol{U} \in \mathbb{R}^{d\times r}$ *where* $\|\boldsymbol{U}\|_2 = 1$*, then when* $\delta \le \operatorname{poly}(r_*)^{-1}$ *and sufficiently small* μ *there exists* $T_\alpha \in \mathbb{N}$ *for all sufficiently small* α *such that* $\lim_{\alpha\to 0}\left\|\boldsymbol{U}_{\alpha,T_\alpha}\boldsymbol{U}_{\alpha,T_\alpha}^\top - \boldsymbol{Z}_s\right\|_2 = 0$*, where* $\boldsymbol{Z}_s$ *is the (unique) global minimizer of*

$$\min \quad \frac{1}{4m}\sum_{i=1}^{m}\left(y_i - \langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle\right)^2 \quad s.t. \quad \boldsymbol{Z} \succeq \boldsymbol{0} \text{ and } \operatorname{rank}(\boldsymbol{Z}) \le s. \tag{2}$$

A formal version of Theorem 1.1 will be given in Theorem 4.2, and main ideas of the proof are outlined in Section 4.

The result and proof techniques of Theorem 1.1 goes beyond previous results (Li et al., 2018; Stöger & Soltanolkotabi, 2021) in a number of ways. Firstly, in the over-parameterized regime (i.e., $r \ge r_*$), it implies that the trajectory of GD exhibits an *incremental learning* phenomenon, learning solutions with increasing ranks until it finds the ground-truth. Second , Theorem 1.1 states that in the *under-parameterized regime* (i.e., $r < r_*$), GD also exhibits the same implicit bias, but finally it converges to an optimal low-rank solution of the matrix sensing loss. By contrast, to the best of our knowledge, only final convergence to the ground-truth when $r \ge r_*$ is known in existing literature.

Theorem 1.1 can also be considered as a generalization of the matrix factorization results in Jiang et al. (2022). However, there are two major challenges in our setting which make our analysis completely different and more complicated. First, due to the presence of noisy measurements, the decomposition of signal and error terms in Jiang et al. (2022) cannot be applied in our setting. Instead, we adopt a different approach which is motivated by Stöger & Soltanolkotabi (2021); intuitive explanations of our approach is discussed in Appendix B.

Second, it is well-known that the optimal rank-s solution of matrix factorization is $\boldsymbol{X}_s$ (defined in Section 3.1), but little is known for $\boldsymbol{Z}_s$. In Section 5 we provide a comprehensive study of the landscape of (2), establishing the uniqueness of $\boldsymbol{Z}_s$ and local landscape properties under a restricted isometry property (RIP). Indeed we find that near $\boldsymbol{Z}_s$, GD follows an approximate low-rank trajectory, so that it behaves similarly to the under-parameterized regime. With the help of our landscape results, we can finally prove Theorem 1.1.

Organization. The subsequent sections are organized as follows. 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 where we also give a proof sketch; the full proof is deferred to the Appendix. In Section 5, we give a comprehensive landscape analysis of under-parameterized matrix sensing, which is used to prove Theorem 1.1 but presented separately it may be of independent interest. 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 proof of all results in this paper 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 finite 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). A line of works (Fornasier et al., 2011; Ngo & Saad, 2012; Wei et al., 2016; Tong et al., 2021) studies efficient optimization algorithms with convergence guarantees. Low rank matrix recovery has a wide variety of applications (cf. discussions in Section 1) and interested readers can refer to Davenport & Romberg (2016) for an overview of this topic.

Simplicity bias/incremental learning of gradient descent. There is a line of existing works devoted to theoretically understand incremental learning as an inductive bias of optimization algorithms in matrix/tensor factorization (Arora et al., 2019; Gissin et al., 2019; Li et al., 2020; Razin et al., 2021), deep linear networks (Jacot et al., 2021), ReLU networks (Timor et al., 2022; Boursier et al., 2022) among others. For matrix sensing problems, Li et al. (2018); Stöger & Soltanolkotabi (2021) show that GD with small initialization recovers the ground-truth despite over-parameterization. These analyses go beyond the neural tangent kernel (NTK) regime (Oymak & Soltanolkotabi, 2019; Woodworth et al., 2020) in which the model overfits to the training data, but fail to reveal the incremental learning procedure that GD actually follows. In this paper, we bridge this gap by rigorously establishing this property of GD in the same setting.

Landscape analysis of non-convex low-rank problems. The strict saddle property Ge et al. (2016; 2015); Lee et al. (2016) has been established for non-convex low rank problems in a unified framework by Ge et al. (2017). Tu et al. (2016) establishes 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). However, only Zhu et al. (2021) considers the under-parameterized regime, in which they establishes a strict saddle property for matrix factorization. We provide a comprehensive analysis in this regime for matrix factorization as well as matrix sensing that improves over their results.

3 Preliminaries

Define $\mathcal{A} : \mathbb{R}^{d\times d} \mapsto \mathbb{R}^m$ as a linear mapping that maps $\boldsymbol{Z}$ to a vector $\mathcal{A}(\boldsymbol{Z})$ whose i-th entry is $\frac{1}{\sqrt{m}}\left(\langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle\right)$. Then our loss function (1) can 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 following *Restricted Isometry Property (RIP)* is a key condition that ensures the ground truth to be recoverable with partial observation:

Definition 3.1 (Restricted Isometry Property) *A measurement $\mathcal{A}$ satisfies the (δ, r)-RIP condition if $(1-\delta)\|\boldsymbol{Z}\|_F^2 \leq \|\mathcal{A}(\boldsymbol{Z})\|_2^2 \leq (1+\delta)\|\boldsymbol{Z}\|_F^2$ for all matrices $\boldsymbol{Z} \in \mathbb{R}^{d\times d}$ with $\operatorname{rank}(\boldsymbol{Z}) \leq r$.*

An important property 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 3.1 *(Stöger & Soltanolkotabi, 2021, Lemma 7.3) Suppose that $\mathcal{A}$ satisfies (r, δ)-RIP with $r \geq 2$, then for all asymmetric $\boldsymbol{Z}$,*

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

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

We make the following standard assumption in the matrix sensing literature:

Assumption 3.1 *The measurement operator $\mathcal{A}$ satisfies the $(2r_*+1, \delta)$-RIP property.*

Let $\mathcal{A}^*(\boldsymbol{v}) = \frac{1}{\sqrt{m}}\sum_{i=1}^m v_i \boldsymbol{A}_i$ be the adjoint of $\mathcal{A}$ and $\boldsymbol{y} = (y_1, y_2, \cdots, y_m)^\top$, then we have $\nabla f(\boldsymbol{U}) = \mathcal{A}^*\left[\boldsymbol{y} - \mathcal{A}(\boldsymbol{U}_t\boldsymbol{U}_t^\top)\right]\boldsymbol{U}_t = \mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{U}_t$, so that the update rule of GD with step size μ is

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \mu\nabla f(\boldsymbol{U}_t) = \left(\boldsymbol{I} + \mu\left[(\mathcal{A}^*\mathcal{A})\left(\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\right)\right]\right)\boldsymbol{U}_t =: (\boldsymbol{I} + \boldsymbol{M}_t)\boldsymbol{U}_t. \tag{3}$$

In this paper, we consider GD with small initialization. Specifically, let $\boldsymbol{U}_{\alpha,0} = \alpha\boldsymbol{U}$ for some matrix $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ with $\|\boldsymbol{U}\| = 1$, we are interested in the trajectory of GD when $\alpha \to 0$. In our analysis we need the following regularity condition on the initialization:

Assumption 3.2 *We have $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}\right) \geq \rho$ for some positive constant ρ.*

Note that for proving Theorem 1.1, we only require Assumption 3.2 to hold for a fixed $s \in [1, \min\{r, r_*\}]$ rather than every s. The following proposition implies that Assumption 3.2 is satisfied with high probability with a Gaussian initialization.

Proposition 3.2 *(Rudelson & Vershynin, 2009) Suppose that all entries of $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ are independently drawn from $\mathcal{N}\left(0, \frac{1}{\sqrt{r}}\right)$ and $\rho = \epsilon\frac{\sqrt{r}-\sqrt{s-1}}{\sqrt{r}}$, then $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}\right) \geq \rho$ with probability at least $1 - e^{-cr} - (C\epsilon)^{r-s+1}$. Here $c, C > 0$ are universal constants.*

3.1 Notations

Throughout this paper, we use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix, and use $\|\cdot\|$ to denote the spectral norm $\|\cdot\|_2$.

For arbitrary matrix $\boldsymbol{A}$, we denote its singular value decomposition by $\boldsymbol{A} = \boldsymbol{V}_{\boldsymbol{A}}\boldsymbol{\Sigma}_A\boldsymbol{W}_A^\top$. We use $\boldsymbol{V}_{\boldsymbol{A}^\perp}$ to denote the orthogonal complement of $\boldsymbol{V}_{\boldsymbol{A}}$. As mentioned before, we assume that $\boldsymbol{Z}^* = \boldsymbol{X}\boldsymbol{X}^\top$ where $\boldsymbol{X} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{r_*}]$ has orthogonal columns. Define $\boldsymbol{X}_s = [v_1, v_2, \cdots, v_s]$, and we write $\boldsymbol{\Sigma}_s = \operatorname{diag}(\sigma_1, \sigma_2, \cdots, \sigma_s)$ to denote the singular values of $\boldsymbol{X}_s$ and $\boldsymbol{\Sigma}_{s,\perp} = \operatorname{diag}(\sigma_{s+1}, \sigma_{s+2}, \cdots, \sigma_{r_*})$. Without loss of generality, we assume that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{r_*} > \sigma_{r_*+1} := 0$ throughout this paper, and define the eigengap of $\boldsymbol{Z}^*$ as $\tau = \min_{1\leq i\leq r_*}\tau_i$ where $\tau_i = \sigma_i^2 - \sigma_{i+1}^2$. We also define the condition number $\kappa_s = \frac{\|\boldsymbol{X}\|^2}{\tau_s}$.

Let $\mathcal{A} : \mathbb{R}^{d\times d} \mapsto \mathbb{R}^m$ be a linear mapping defined as $\mathcal{A}(\boldsymbol{Z})_i = \frac{1}{\sqrt{m}}\langle \boldsymbol{A}_i, \boldsymbol{Z}\rangle, i = 1, 2, \cdots, m$ and $\mathcal{A}^*$ be its adjoint. We define $\boldsymbol{M} = \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top)$ and $\boldsymbol{M}_t = \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)$, then the update of GD (3) can be written as $\boldsymbol{U}_{t+1} = (I + \mu\boldsymbol{M}_t)\boldsymbol{U}_t$.

4 MAIN RESULTS

In this section, we present our main results in this paper and briefly sketch their proof. The full proof can be found in the appendix. Our first result states that small initialization, GD is able to fit the top-s components of the ground-truth with an error that scales linearly in δ.

Theorem 4.1 *When α is sufficiently small and $\delta \leq 10^{-7}\kappa^{-2}r_*^{-\frac{1}{2}}$, there exists a time $T_\alpha \in \mathbb{Z}_+$ such that*

$$\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_{\alpha,T_\alpha}\boldsymbol{U}_{\alpha,T_\alpha}^\top\right\|_F \leq \kappa^2\sqrt{r_*}\|\boldsymbol{X}\|^2\delta.$$

***Proof sketch*:** The proof roughly follows the three-phase analysis in Stöger & Soltanolkotabi (2021), with some technical modifications. Starting from a small initialization, GD initially behaves similar to power iteration since $\boldsymbol{U}_{t+1} \approx (\boldsymbol{I} + \mu\boldsymbol{M})\boldsymbol{U}_t$. Let $\boldsymbol{M} = \sum_{k=1}^d \hat{\sigma}_k^2\hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i^\top$, then we have

$$\boldsymbol{U}_t \approx (\boldsymbol{I} + \mu\boldsymbol{M})^T\boldsymbol{U}_0 = \sum_{i=1}^d(1+\mu\hat{\sigma}_i^2)^T\hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i^\top\boldsymbol{U}_0 \approx \sum_{i=1}^s(1+\mu\hat{\sigma}_i^2)^T\hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i^\top\boldsymbol{U}_0 \tag{4}$$

where the last step holds since there exists an exponential separation between the magnitude of the top-s and the remaining components.

When (4) no longer holds, we enter the *parallel improvement phase*. In this phase we consider the decomposition $\boldsymbol{U}_t = \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top + \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top$ where $\boldsymbol{W}_t \in \mathbb{R}^{r\times s}$ is defined by the SVD $\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t = \boldsymbol{V}_t\boldsymbol{\Sigma}_t\boldsymbol{W}_t^\top$. Intuitions for this decomposition are given in Appendix B. By the end of the spectral phase we have $\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t) \gg \|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$, and moreover, we can show that the *parallel component* $\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t)$ grows exponentially faster than the *orthogonal component* $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$ until the former reaches a constant scale. Details are provided in Appendix C.2.

After $\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t) = \Theta(1)$, we enter the *refinement phase* and we show in Appendix C.4 that $\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_F$ decreases linearly until it is $\mathcal{O}\left(\kappa^2\sqrt{r_*}\|\boldsymbol{X}\|^2\delta\right)$. The conclusion follows. □

It is not surprising that the bound in Theorem 4.1 is not $o(1)$ when $\alpha \to 0$, because in general $\boldsymbol{Z}_s \neq \boldsymbol{X}_s\boldsymbol{X}_s^\top$ due to measurement error. As a result, according to Theorem 1.1, we should expect that GD visits an $o(1)$ neighbourhood of $\boldsymbol{Z}_s$ rather than $\boldsymbol{X}_s\boldsymbol{X}_s^\top$. Indeed, with a measurement error of δ, intuitively we have that $\left\|\boldsymbol{Z}_s - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F \lesssim \delta$, which is consistent with the bound of Theorem 4.1.

Nonetheless, there is still a gap between Theorem 4.1 and Theorem 1.1. To bridge this gap we need to characterize how gradient descent behaves *after* it nearly fits the top s components, and show it converges to $\boldsymbol{Z}_s$. The following lemma is a key step towards this goal:

Lemma 4.1 *The following statements hold:*

(1). There exists a rank-s matrix $\widetilde{\boldsymbol{U}}_{\alpha,0}$ such that $\lim_{\alpha\to 0}\left\|\widetilde{\boldsymbol{U}}_{\alpha,0} - \boldsymbol{U}_{\alpha,T_\alpha}\right\|_F = 0$.

(2). Let $\bar{\boldsymbol{U}}_{\alpha,0} \in \mathbb{R}^{d\times s}$ be such that $\bar{\boldsymbol{U}}_{\alpha,0}\bar{\boldsymbol{U}}_{\alpha,0}^\top = \widetilde{\boldsymbol{U}}_{\alpha,0}\widetilde{\boldsymbol{U}}_{\alpha,0}^\top$. Consider under-parameterized matrix sensing loss

$$f_s(\boldsymbol{U}) = \frac{1}{4}\left\|\mathcal{A}(\boldsymbol{Z}^* - \boldsymbol{U}\boldsymbol{U}^\top)\right\|_2^2, \quad \boldsymbol{U} \in \mathbb{R}^{d\times s}, \tag{5}$$

then the global minimizer of f_s is unique up to rotation. In other words, for any two global minimizers $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$, there exists an orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{s\times s}$ such that $\boldsymbol{U}_1 = \boldsymbol{U}_2\boldsymbol{R}$. Moreover, there exists a global minimizer $\boldsymbol{U}_s^*$ such that $\left\|\bar{\boldsymbol{U}}_{\alpha,0} - \boldsymbol{U}_s^*\right\|_F \lesssim \kappa^2\sqrt{r_*}\|X\|\delta$.

(3). The function f_s satisfies a local Polyak-Lojasiewicz (PL) condition. Specifically, let $\boldsymbol{U} \in \mathbb{R}^{d\times s}$ and $\boldsymbol{U}_s^*$ be the global minimizer with minimal distance (under $\|\cdot\|_F$) to $\boldsymbol{U}$, then we have

$$\langle\nabla f_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{U}_s^*\rangle \geq 0.1\tau_s\left\|\boldsymbol{U} - \boldsymbol{U}_s^*\right\|_F^2$$

whenever $\|\boldsymbol{U} - \boldsymbol{U}_s^*\| \leq 10^{-2}\kappa^{-1}\|X\|$.

***Proof sketch*:** The statement of *(1)* follows directly by choosing $\widetilde{\boldsymbol{U}}_0 = \boldsymbol{U}_{\alpha,T_\alpha}\boldsymbol{W}_{\alpha,T_\alpha}\boldsymbol{W}_{\alpha,T_\alpha}^\top$ and the observation that the parallel term $\boldsymbol{U}_{\alpha,t}\boldsymbol{W}_{\alpha,t,\perp}$ is $o(1)$ when $t = T_\alpha$. *(2)* and *(3)* follows from our landscape analysis in Section 5. □

To see why Lemma 4.1 is useful, note that (1) implies that $\boldsymbol{U}_{\alpha,T_\alpha}$ is approximately rank-s, so within a time period after T_α, the GD trajectory remains close to another GD initialized at $\bar{\boldsymbol{U}}_{\alpha,0}$ for the rank-s matrix sensing loss (5); that is

$$\bar{\boldsymbol{U}}_{\alpha,t}\bar{\boldsymbol{U}}_{\alpha,t}^\top \approx \boldsymbol{U}_{\alpha,T_\alpha+t}\boldsymbol{U}_{\alpha,T_\alpha+t}^\top \tag{6}$$

when t is below a certain time threshold.

We know from (2) that the initialization $\bar{\boldsymbol{U}}_{\alpha,0}$ is within a small neighbourhood of the global minima of $f_s(\boldsymbol{U})$. Furthermore, (3) implies that $f_s(\boldsymbol{U})$ is locally PL so that GD with good initialization would converge linearly to its global minima (Karimi et al., 2016). We need to choose a time t such that (6) remains true while this linear convergence takes place for sufficiently many steps. Since the initial gap is $o(1)$, we can always choose some t such that both $\left\|\bar{\boldsymbol{U}}_{\alpha,t}\bar{\boldsymbol{U}}_{\alpha,t}^\top - \boldsymbol{U}_{\alpha,T_\alpha+t}\boldsymbol{U}_{\alpha,T_\alpha+t}^\top\right\|_F$ and $\left\|\bar{\boldsymbol{U}}_{\alpha,t} - \boldsymbol{U}_s^*\right\|_F$ are $o(1)$. Hence, we arrive at the conclusion of Theorem 1.1, of which a formal version is stated below.

Theorem 4.2 *Suppose that $\tau_s = \sigma_s^2 - \sigma_{s+1}^2 > 0$, $\delta \le 10^{-7}r_*^{-1}\kappa_s^{-3}$ and $\mu \le 10^{-3}\|X\|^{-2}$, then there exists a constant $C = C(r,\kappa_s)$ such that $\min_{t\ge1}\left\|\boldsymbol{U}_{\alpha,t}\boldsymbol{U}_{\alpha,t}^\top - \boldsymbol{Z}_s\right\|_F \le C\alpha^{\gamma_s}$ for all sufficiently small α, where $\gamma_s \ge 0.1\kappa_s$.*

Remark 4.1 (The implicit regularization effect of GD) *Assuming that $\sigma_i, 1 \le i \le r^*$ are different, then the result of Theorem 4.2 holds for all $1 \le s \le \min\{r, r^*\}$.*

Contrary to what classical statistical learning theory suggests, it is empirically observed that highly expressive and over-parameterized models such as deep neural networks are still able to achieve good generalization performance. It is suggested that this is probably due to the implicit regularization of optimization algorithms. Compared with existing works (Li et al., 2018; Stöger & Soltanolkotabi, 2021), 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 increasing the complexity of search space until it finds the ground-truth.*

The algorithmic regularization of GD can also be viewed from an optimization perspective. Our analysis shows that GD follows a saddle-to-saddle dynamics (Jacot et al., 2021). Theoretical studies of this dynamics are often restricted to the rank-1 case, since extending to higher ranks requires knowledge on how GD approaches the rank-1 saddle (Li et al., 2020). We show that for matrix sensing problems, GD approaches the rank-s saddle through a specific trajectory that lies in the rank-s solution space, which in turn allows GD to efficiently escaping it. The proof idea of Theorem 4.2 is similar to (Li et al., 2020, Theorem 5.6) which shows that GF with infinitesimal initialization behaves like a rank-1 GF, but it is unknown whether the rank-1 GF converges to the minimizer for general PSD-constrained optimization problems.

With the help of the landscape results in Section 5, we can obtain the following corollary which states that when $r \le r_*$, GD with a small initialization can actually find global minima of (5).

Corollary 4.1 *Suppose that $r \le r^*$, then there exists a constant $c = c(r,\kappa_s) > 0$ such that when $\alpha < c$, we have $\lim_{t\to+\infty}\boldsymbol{U}_{\alpha,t}\boldsymbol{U}_{\alpha,t}^\top = \boldsymbol{Z}_s$.*

5 THE LANDSCAPE OF UNDER-PARAMETERIZED MATRIX SENSING

In this section, we analyze the properties enjoyed by the landscape of the matrix sensing loss (1). While the landscape of over-parameterized matrix sensing (i.e. $r \ge r^*$) is well-studied (Tu et al., 2016; Ge et al., 2017), few results are known for the under-parameterized case. A strict saddle property of matrix factorization is established in Zhu et al. (2021), but this is still insufficient for fast convergence of GD to global minima (Du et al., 2017).

We point out that the results in this section serve as technical tools for proving Theorem 4.2. However, we believe that they are of independent interest as well, especially because low-rank matrix

approximation itself is an active research area (Lee et al., 2013; Kishore Kumar & Schneider, 2017; Tropp et al., 2019) and is also of practical importance (Chen et al., 2017; Lu et al., 2020).

Consider the loss function f_s defined in (5); we are interested in the landscape of f_s near the set of global minimizers. Clearly, if $\boldsymbol{U} \in \mathbb{R}^{d\times s}$ is a global minimizer of f_s, then so is $\boldsymbol{U}\boldsymbol{R}$ for arbitrary orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{s\times s}$. As a result, except from the case $s = 1$, f_s cannot be strongly-convex (the case $s = 1$ is separately analyzed in Appendix G). This motivates the following definition:

Definition 5.1 (Procrutes distance) *Let $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{R}^{d\times s}$, we define*

$$\operatorname{dist}(\boldsymbol{U}_1, \boldsymbol{U}_2) = \inf \left\{ \|\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{R}\|_F : \boldsymbol{R} \in \mathbb{R}^{s\times s} \text{ is orthogonal} \right\}.$$

To study the landscape of (5), we first need to characterize the matrix factorization loss

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

since it corresponds to the full-observation case $\delta = 0$. The following lemma summarizes our findings about the landscape properties of F_s:

Theorem 5.1 (Landscape of under-parameterized matrix factorization) *The following statements hold:*

(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 $\{\sigma_i \boldsymbol{v}_i : 1 \le i \le r^*\}$. The set of global minimizers of F_s is $\{\boldsymbol{X}_s\boldsymbol{R} : \boldsymbol{R} \in \mathbb{R}^{s\times s} \text{is orthogonal}\}$.

(2). Given $\boldsymbol{U} \in \mathbb{R}^{d\times s}$ and let $\boldsymbol{R}$ be an orthogonal matrix such that $\|\boldsymbol{U} - \boldsymbol{X}_s\boldsymbol{R}\|_F = \operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}_s)$. Suppose that $\delta \le 0.01\kappa_s^{-2}r^{-1}\|\boldsymbol{X}\|^2$ and $\operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}_s) \le 0.1\|\boldsymbol{X}\|^{-1}\tau_s$, then we have $\langle \nabla F(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{X}_s\boldsymbol{R}\rangle \ge 0.1\tau_s \operatorname{dist}^2(\boldsymbol{U}, \boldsymbol{X}_s)$.

The following lemma formalizes the intuition that all global minimizers of the f_s must be close to the set of minimizers of F_s.

Lemma 5.1 *Under the conditions in Theorem 5.1, we have $\operatorname{dist}^2(\boldsymbol{U}_s^*, \boldsymbol{X}_s) \le 40\delta\kappa_s\|X\|_F$ for any global minimizer $\boldsymbol{U}_s^*$ of f_s.*

We now present the main result in this section, which establishes a local PL condition for f_s.

Theorem 5.2 *Let $\boldsymbol{U} \in \mathbb{R}^{d\times s}$ and $\boldsymbol{U}_s^*$ be a global minimizer with minimal distance (under $\|\cdot\|_F$) to $\boldsymbol{U}$. Suppose that $\|\boldsymbol{U} - \boldsymbol{U}_s^*\| \le 10^{-2}\kappa^{-1}\|\boldsymbol{X}\|$ and $\delta \le 10^{-3}r_*^{-1}\kappa_s^{-2}$, then we have*

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

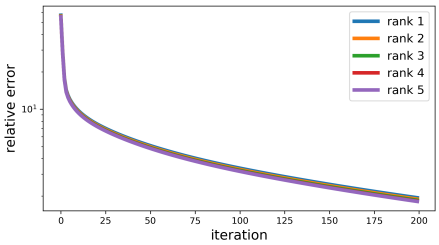
We conclude this section with the following corollary which states that the global minimizer of the under-parameterized loss (5) is unique.

Corollary 5.1 *Under the conditions in Theorem 5.2, the global minimizer of f_s is unique up to rotation i.e. for any two minimizer $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$, there must exists an orthogonal matrix $\boldsymbol{R}$ such that $\boldsymbol{U}_1 = \boldsymbol{U}_2\boldsymbol{R}$.*

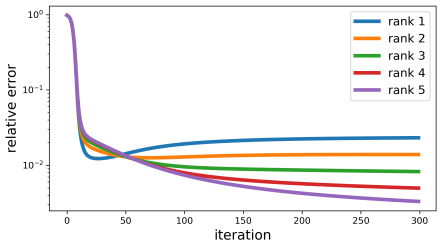
6 Experiments

In this section, we perform some numerical experiments to illustrate our theoretical findings.

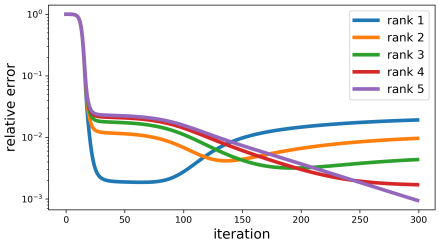
Experimental setup. Consistent with our theory, we consider the matrix sensing problem (1) with $d = 50, r^* = 5$ and $\alpha \in \{1, 0.1, 0.01, 0.001\}$ $m \in \{1000, 2000, 5000\}$. The ground-truth $\boldsymbol{Z}^* = \boldsymbol{X}\boldsymbol{X}^\top$ is generated such that the entries of $\boldsymbol{X}$ are i.i.d. standard Gaussian variables. Throughout our experiments, we use the same ground-truth to compare between different choices of parameters. For $i = 1, 2, \cdots, m$, the measurement $\boldsymbol{A}_i \in \mathbb{R}^{d\times d}$ is chosen as follows: the diagonal entries are drawn i.i.d. from $\mathcal{N}\left(0, \frac{1}{\sqrt{2}}\right)$, while the off-diagonal entries are drawn i.i.d. from the standard Gaussian



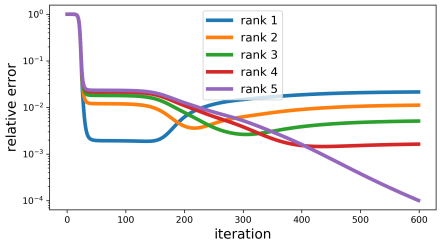
(a) $\alpha = 1$, 1000 measurements.



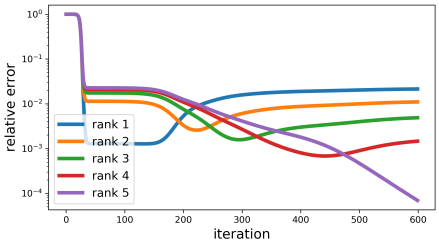
(b) $\alpha = 0.1$, 1000 measurements.



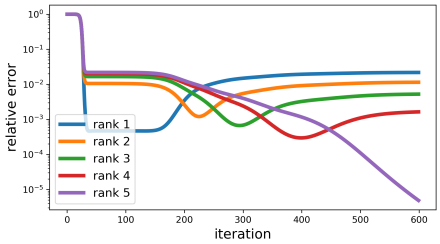
(c) $\alpha = 0.01$, 1000 measurements.



(d) $\alpha = 0.001$, 1000 measurements.



(e) $\alpha = 0.001$, 2000 measurements.



(f) $\alpha = 0.001$, 5000 measurements.

Figure 1: The evolution of relative error w.r.t rank s for $1 \leq s \leq r^* = 5$.

$\mathcal{N}(0,1)$. It is well-known that when $m \gtrsim dr^*\delta^{-2}$, this set of measurements satisfies the RIP with high probability (Recht et al., 2010, Theorem 4.2).

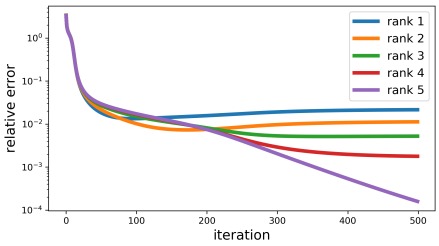
We solve the problem (1) via running GD for $T = 10^4$ iterations starting with small initialization with scale α. Specifically, we choose $\boldsymbol{U}_0 = \alpha \boldsymbol{U}$ where $\alpha = 10^{-6}$ and the entries of $\boldsymbol{U} \in \mathbb{R}^{d \times r}$ are drawn i.i.d. from standard Gaussian distribution. 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

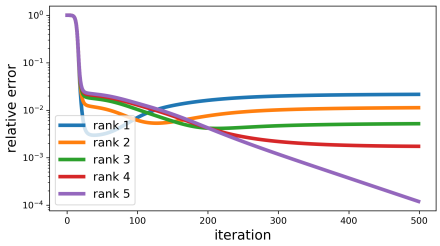
As in Section 3.1, let $\boldsymbol{Z}^* = \sum_{i=1}^{r^*} \sigma_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^\top$ and $\boldsymbol{X}_s \boldsymbol{X}_s^\top = \sum_{i=1}^{s} \sigma_i^2 \boldsymbol{v}_i \boldsymbol{v}_i^\top$. For each $t \in [T]$ and $s \in [r^*]$, we define the relative error $\mathcal{E}(\boldsymbol{U}, s) = \frac{\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{X}_s \boldsymbol{X}_s^\top\|_F^2}{\|\boldsymbol{X}_s \boldsymbol{X}_s^\top\|_F^2}$ 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 neighbourhood 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 matrix. This is in align with the conclusion of Theorem 4.1. By contrast, for large initialization we do not have this implicit bias.

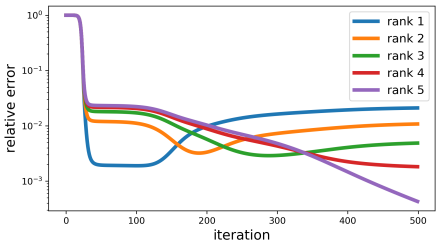
The effect of measurement error. For fixed α, one can observe the the relative error becomes smaller when the number of measurement increases. This is also in align with Theorem 4.1 in which



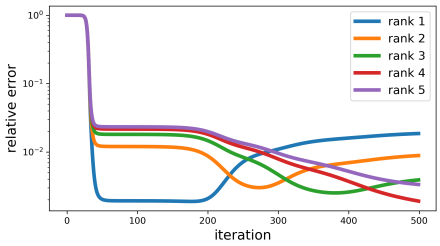
(a) $\alpha = 1$, 1000 measurements.



(b) $\alpha = 0.1$, 1000 measurements.



(c) $\alpha = 0.01$, 1000 measurements.



(d) $\alpha = 0.001$, 1000 measurements.

Figure 2: The evolution of the loss and relative error w.r.t rank s for $1 \leq s \leq r^* = 5$ in the exact-parameterized case $r = 5$.

the bound depends on δ. In particular, for the case $s = r^*$, although GD with fixed initialization does not converge to global minima, but the distance to the set of global minima scales as $\mathrm{poly}(\alpha)$.

6.2 Matrix sensing with exact parameterization

In this subsection, we study the behavior of GD in the exact ($r = r^*$) regime. We fix $m = 1000$ and $r = r^* = 5$ and run GD for $T = 1000$ iterations with different scale of initializations $\alpha \in \{1, 0.1, 0.01, 0.001\}$. We plot the relative error w.r.t. different ranks in Figure 2.

We can see that GD exhibits an implicit low-rank bias when α is small. However, choosing very small α would slow down the convergence speed. This is because GD would get into a $\mathrm{poly}(\alpha)$-neighbourhood of the saddle point $\boldsymbol{Z}_s$ and spend a long time escaping the saddle. Also, convergence to global minimizers is guaranteed as long as α is below a certain threshold (cf. Corollary 4.1).

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 in which 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 partially explaining 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 networks under square loss (Timor et al., 2022). Also, it was shown (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 further studies are necessary for a comprehensive understanding of generalization aspects of gradient-based optimization methods.

REFERENCES

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019.

Devansh Arpit, Stanislaw Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017.

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

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

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

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

Yongyong Chen, Yanwen Guo, Yongli Wang, Dong Wang, Chong Peng, and Guoping He. Denoising of hyperspectral images using nonconvex low rank matrix approximation. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5366–5380, 2017.

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

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f79921bbae40a577928b76d2fc3edc2a-Paper.pdf.

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

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on learning theory*, pp. 797–842. PMLR, 2015.

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

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

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

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

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

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *arXiv preprint arXiv:2103.10427*, 2021.

Arthur Jacot, François Ged, Franck Gabriel, Berfin Şimşek, and Clément Hongler. Deep linear networks dynamics: Low-rank biases induced by initialization scale and l2 regularization. *arXiv preprint arXiv:2106.15933*, 2021.

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

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/b432f34c5a997c8e7c806a895ecc5e25-Paper.pdf.

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

Hamed Karimi, Julie Nutini, and Mark Schmidt. 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.

N Kishore Kumar and Jan Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2017.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pp. 1246–1257, 2016.

Joonseok Lee, Seungyeon Kim, Guy Lebanon, and Yoram Singer. Local low-rank matrix approximation. In *International conference on machine learning*, pp. 82–90. PMLR, 2013.

Xingguo Li, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, Zhaoran Wang, and Tuo Zhao. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514, 2019.

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

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

Xuequan Lu, Scott Schaefer, Jun Luo, Lizhuang Ma, and Ying He. Low rank matrix approximation for 3d geometry filtering. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021.

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

Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pp. 4951–4960. PMLR, 2019.

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

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

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

Mark Rudelson and Roman Vershynin. 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.

Xiaohui Shen and Ying Wu. 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.

Dominik Stöger and Mahdi Soltanolkotabi. 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.

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

Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. *arXiv preprint arXiv:2201.12760*, 2022.

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

Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Streaming low-rank matrix approximation with an application to scientific simulation. *SIAM Journal on Scientific Computing*, 41(4):A2430–A2463, 2019.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pp. 964–973. PMLR, 2016.

Guillermo Valle-Perez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2018.

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

Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. *Advances in neural information processing systems*, 23, 2010.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Bo Zhao, Justin P Haldar, Cornelius Brinegar, and Zhi-Pei Liang. 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.

Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

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

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

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.1 *(Stöger & Soltanolkotabi, 2021, Lemma 7.3) Suppose that $\mathcal{A}$ satisfies (r, δ)-RIP, then for all symmetric matrix Z of rank $\leq r-1$, we have*

$$\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})\boldsymbol{Z}\| \leq \sqrt{r}\delta\|\boldsymbol{Z}\|.$$

A.2 MATRIX ANALYSIS

The following lemma is a direct corollary of Proposition A.1 and will be frequently used in our proof.

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

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

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

Equipped with Lemma A.1, we can have the following characterization of the eigenvalues of M (recall that $M = \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top)$):

Lemma A.4 *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

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

as desired. □

Lemma A.5 *Suppose that $\boldsymbol{M}, \boldsymbol{N}$ are two symmetric matrices where N is positive semi-definite with rank r. If $\|\boldsymbol{M} - \boldsymbol{N}\| \leq \delta \cdot \lambda_r(\boldsymbol{N})$, then it holds that $\|\lambda_1(\boldsymbol{M}) - \lambda_1(\boldsymbol{N})\| \leq \delta\|\boldsymbol{N}\|$, $\lambda_{r+1}(\boldsymbol{M}) \leq \delta\lambda_r(\boldsymbol{N})$, $\lambda_r(\boldsymbol{M}) \geq (1-\delta)\lambda_r(\boldsymbol{N})$ and $\left\|\boldsymbol{V}_{\boldsymbol{N},\perp}^\top \boldsymbol{V}_{\boldsymbol{M}}\right\| \leq 2\delta$.*

A.3 OPTIMIZATION

Lemma A.6 *Suppose that a smooth function $f \in \mathbb{R}^m \mapsto \mathbb{R}$ 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 $f(\boldsymbol{x}) - f^* \geq 2\epsilon$. Moreover, we also have $f(\boldsymbol{x}) - f^* \geq 2\epsilon$ on ∂S.

Then we have

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

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 $f(\boldsymbol{x}^*) - f^* \geq 2\epsilon$, as desired. □

Lemma A.7 *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 $\|\boldsymbol{x}_k\| \leq B$ and $\|\boldsymbol{y}_k\| \leq B$ for all k and f is L-smooth in $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq B\}$, then we have*

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

Proof **:** The update rule implies that

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

which yields the desired inequality. □

B MAIN IDEA FOR THE PROOF OF THEOREM 4.1 AND THEOREM 4.2

Motivated by Stöger & Soltanolkotabi (2021) we decompose the matrix $\boldsymbol{U}_t$ into a parallel component and an orthogonal component. Specifically, let $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t \boldsymbol{V}_t \boldsymbol{\Sigma}_t \boldsymbol{W}_t^\top$ be the SVD (where we recall that $\boldsymbol{Z}_s$ is defined in Theorem 1.1), then we can write

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

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 $\left\|\boldsymbol{U}_t \boldsymbol{U}_t^\top - \boldsymbol{X}_s \boldsymbol{X}_s^\top\right\| \approx 0$.s

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 $V_{X_s}^\top U_t$ 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 X_s.

The dynamics of GD implies that

$$\begin{aligned}
\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_{t+1} &= \boldsymbol{V}_{\boldsymbol{X}_s,\perp} \left[\boldsymbol{I} + \mu \mathcal{A}^* \mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t \boldsymbol{U}_t^\top)\right] U_t \\
&= (\boldsymbol{I} + \mu \boldsymbol{\Sigma}_{s,\perp}^2) \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t - \mu \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{U}_t + \mu \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \left(\mathcal{A}^* \mathcal{A} - \boldsymbol{I}\right)(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t^\top \boldsymbol{U}_t)\boldsymbol{U}_t \\
&= (\boldsymbol{I} + \mu \boldsymbol{\Sigma}_{s,\perp}^2) \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t (\boldsymbol{I} - \mu \boldsymbol{U}_t \boldsymbol{U}_t^\top) + \mu \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \left(\mathcal{A}^* \mathcal{A} - \boldsymbol{I}\right)(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t \boldsymbol{U}_t^\top)\boldsymbol{U}_t + \mathcal{O}(\mu^2).
\end{aligned} \tag{8}$$

Note that $\|\boldsymbol{\Sigma}_{s,\perp}\| = \sigma_{s+1}$, so that

$$\left\|(\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2)\boldsymbol{V}_{X_s^\perp}^\top \boldsymbol{U}_t(\boldsymbol{I} - \mu\boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \leq (1 + \sigma_{s+1}^2)\left\|\boldsymbol{V}_{X_s^\perp}\boldsymbol{U}_t\right\|.$$

The last term $\mathcal{O}(\mu^2)$ is negligible when μ is chosen sufficiently small. However, the main difference with the full-observation case Jiang et al. (2022) is the second term. Since the measurement error $(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(XX^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)$ is small but arbitrary, it is hard to compare this term with $\boldsymbol{V}_{X_s^\perp}\boldsymbol{U}_{t+1}$.

As a result, we cannot directly bound the growth of $\left\|\boldsymbol{V}_{X_s^\perp}\boldsymbol{U}_t\right\|$. However, the aforementioned problem disappear when the second term of (8) is multiplied on the right side by $\boldsymbol{W}_{t,\perp}$, because from its definition we know that $\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = 0$, so that we can write

$$\boldsymbol{V}_{\boldsymbol{X}_s^\perp}(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = \boldsymbol{V}_{\boldsymbol{X}_s^\perp}(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}.$$

The term $\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}$ appears in the above expression, so plugging into (8) gives

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

It thus look promising to instead analyze the growth of $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|$. Note that

$$\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^T\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp} = \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^T\boldsymbol{U}_{t+1}\boldsymbol{W}_t\boldsymbol{W}_t^T\boldsymbol{W}_{t+1,\perp} + \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^T\boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^T\boldsymbol{W}_{t+1,\perp}, \tag{9}$$

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 $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^T\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \ll 1$, which implies that the first term in (9) is negligible.

B.2 Proof outline

We begin with the *spectral alignment phase*, where we can make the following approximation

$$\boldsymbol{U}_{t+1} \approx (\boldsymbol{I} + \mu\boldsymbol{M})\boldsymbol{U}_t \tag{10}$$

since $\boldsymbol{U}_t$ is initially small. At some time $t = T_0$, $\boldsymbol{U}_t$ would become approximately aligned with the first s components, as long as there is a positive gap between the s-th and $(s+1)$-th largest eigenvalues of $\boldsymbol{M}$. The choice of T_0 is subject to a trade-off such that the alignment takes effect while (10) does not induce large error.

We then enter the second phase which we call the *parallel matching* phase, in which the parallel components grow to a constant magnitude and is well-matched with the ground-truth (the orthogonal components remain small). Specifically, for small constants $c_i, 1 \leq i \leq 3$, we show that the followings are true in this phase:

(1). $\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t)$ grows exponentially fast until it reaches $c \cdot \sigma_s$ for some constant c at some time T_s. Specifically, we have

$$\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_{t+1}\boldsymbol{W}_t) \geq \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t)\left(1 + \mu\sigma_s^2 - c_1 - \mu\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t)\right).$$

(2). When $t \leq T_s$, the growth speed of $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$ is slower than $\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t)$:

$$\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \leq \left(1 + \mu\sigma_{s+1}^2 + c_2\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|.$$

(3). $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \leq c_3$ remains true until $t = T_s$.

These statements will be proven by induction. At $t = T_s$, we have $\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t) = \Theta(1)$, and we enter the *refinement phase* in which we show that the quantity $\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top \left(\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right)\right\|_F$ decrease exponentially until it reaches $\mathcal{O}(\delta)$. Note that in general GD would not converge to an $o(1)$-neighbourhood of $\boldsymbol{X}_s\boldsymbol{X}_s^\top$ as the initialization scale $\alpha \to 0$, because $\boldsymbol{X}_s\boldsymbol{X}_s^\top$ is not the rank-s minima of the RIP loss. As a result, in the refinement phase we can only expect to obtain $\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F \le \mathrm{poly}(r)\cdot\delta$.

To conclude the proof of Theorem 1.1, we prove in Section 5 that the landscape of the matrix sensing loss with rank-s parameterization, though non-convex, satisfies a local Polyak-Lojasiewicz (PL) condition within a neighborhood of *constant* radius. As a result, for sufficiently small δ, GD converges linearly to global minima with good initialization. We then show that within a time period after the refinement phase, the GD trajectory is close to the trajectory of well-initialized GD for rank-s parameterized matrix sensing. The length of this period goes to infinity when $\alpha \to 0$, thereby implying that GD finds the rank-s minimizer with $o(1)$ error. The details are given in Appendix F.

C Proof of Theorem 4.1

C.1 The spectral phase

Starting from a small $\boldsymbol{U}_0 = \alpha\boldsymbol{U}, \alpha \ll 1$, 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

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

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.1, we have $\|u_1 - \boldsymbol{v}_1\| = \mathcal{O}(\delta\sqrt{r})$ 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 $\boldsymbol{M} = \mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top)$, $\boldsymbol{Z}_t = (\boldsymbol{I} + \mu\boldsymbol{M})^t$ and $\widetilde{\boldsymbol{U}}_t = \boldsymbol{Z}_t\boldsymbol{U}_0$. 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. We additionally define $\boldsymbol{M}_s = \sum_{i=1}^{\min\{s,\mathrm{rank}(\boldsymbol{M})\}} \hat{\sigma}_i^2 \hat{\boldsymbol{v}}_i\hat{\boldsymbol{v}}_i^\top$. Due to Lemma A.1, we view $\hat{\sigma}_i$ as constants as well. Similarly, let L_t be the span of the top-s left singular vectors of $\boldsymbol{U}_t$. We make the following assumption on the initialization, which holds with high probability when it is i.i.d. Gaussian:

Assumption C.1 *The matrix $\boldsymbol{V}_{\boldsymbol{M}_s}^\top \boldsymbol{U} \in \mathbb{R}^{s\times r}$ has full row-rank i.e. $\rho = \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{M}_s}^\top \boldsymbol{U}\right) > 0$.*

Let

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

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

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*

$$\|E_t\| = \left\|\boldsymbol{U}_t - \widetilde{\boldsymbol{U}}_t\right\| \le 4\hat{\sigma}_1^{-2}\alpha^3 r\left(1+\delta_1\right)\left(1+\mu\hat{\sigma}_1^2\right)^{3t}\|\boldsymbol{U}\|^3. \tag{11}$$

Corollary C.1 *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

$$\|E_t\| \le 4\hat{\sigma}_1^{-2}\alpha^3 r\left(1+\delta_1\right)\left(1+\mu\hat{\sigma}_1^2\right)^{3t}\|\boldsymbol{U}\|^3.$$

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

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

Thus, it follows from $\|E_{t^*}\| \geq \|\widetilde{\boldsymbol{U}}_{t^*}\|$ that

$$\left(1+\mu\hat{\sigma}_1^2\right)^{t^*} \geq \sqrt{\frac{\rho\hat{\sigma}_1^2}{4(1+\delta_1)r\|\boldsymbol{U}\|^3}} \cdot \alpha^{-1} \Rightarrow t^* \geq \frac{\log \alpha^{-1} + \frac{1}{2}\log\frac{\rho\hat{\sigma}_1^2}{4(1+\delta_1)r}}{\log\left(1+\mu\hat{\sigma}_1^2\right)}$$

as desired. □

Lemma C.2 *Suppose that $\alpha \leq \mathrm{poly}(r)^{-1}$, then there exists a time*

$$t = T_0 := \frac{2\log\alpha^{-1} + \log\frac{\hat{\sigma}_1^2}{4r(1+\delta)}}{3\log(1+\mu\hat{\sigma_1^2}) - \log(1+\mu\hat{\sigma}_{s+1}^2)} \leq t^*$$

such that

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

where $\gamma = 1 - \frac{2\log(1+\mu\hat{\sigma}_1^2)}{3\log(1+\mu\hat{\sigma}_1^2) - \log(1+\mu\hat{\sigma}_{s+1}^2)}$.

Proof **:** It's easy to check that $T_0 \leq t^*$ when $\alpha \leq \mathrm{poly}(r)^{-1}$ by applying Corollary C.1.

Consider the following decomposition:

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

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

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

For the second term we have

$$\left\|\widetilde{\boldsymbol{U}}_t - \sum_{i=1}^{s} \alpha(1+\mu\hat{\sigma}_i^2)^t \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top \boldsymbol{U}\right\| \leq \left\|\sum_{i=s+1}^{r} \alpha(1+\mu\hat{\sigma}_i^2)^t \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\top \boldsymbol{U}\right\| \leq \alpha\left(1+\mu\hat{\sigma}_{s+1}^2\right)^t.$$

In particular, the definition of T_0 implies that

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

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.3 *The following inequalities hold for $t = T_0$ when $\alpha \leq \mathrm{poly}(r)^{-1}$ is sufficiently small:*

$$\|\boldsymbol{U}_t\| \leq \|\boldsymbol{X}\| \tag{12a}$$

$$\sigma_{\min}\left(\boldsymbol{U}_{T_0}\boldsymbol{W}_{T_0}\right) \geq \mathrm{poly}(r)^{-1} \cdot \alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_s^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \tag{12b}$$

$$\|\boldsymbol{U}_{T_0}\boldsymbol{W}_{T_0,\perp}\| \leq \mathrm{poly}(r) \cdot \alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_{s+1}^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \tag{12c}$$

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{T_0}\boldsymbol{W}_{T_0}}\right\| \leq 200\delta \tag{12d}$$

Proof : We prove this lemma by applying Corollary D.1 to $t = T_0$ defined in the previous lemma. The inequality (12a) can be directly verified by using Lemma C.2:

$$\|\boldsymbol{U}_t\| \leq \alpha \left(1+\mu\hat{\sigma}_i^2\right)^{T_0} + \alpha^{\gamma} \lesssim \mathrm{poly}(r)\cdot\alpha^{\gamma/3} \leq \|\boldsymbol{X}\|.$$

For the remaining inequalities, we first verify that the assumption in Corollary D.1:

$$\alpha\sigma_s(\boldsymbol{Z}_t) > 10\left(\alpha\sigma_{s+1}(\boldsymbol{Z}_t) + \|\boldsymbol{E}_t\|\right). \tag{13}$$

By definition of $\boldsymbol{Z}_t$, we can see that

$$\begin{aligned}\alpha\sigma_{s+1}(\boldsymbol{Z}_{T_0}) + \|\boldsymbol{E}_{T_0}\| &\leq \alpha\left(1+\mu\hat{\sigma}_{s+1}^2\right)^{T_0} + \|\boldsymbol{E}_{T_0}\| \\ &\lesssim \alpha^{\gamma} \leq \mathrm{poly}(r)^{-1}\cdot 0.1\rho\alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_s^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \\ &\leq 0.1\alpha\sigma_s(\boldsymbol{Z}_{T_0})\end{aligned}$$

when $\alpha \leq \mathrm{poly}(r)^{-1}$, so that (13) holds. As a result, we have

$$\begin{aligned}\sigma_s\left(\boldsymbol{U}_t\boldsymbol{W}_t\right) &\geqslant 0.4\alpha\sigma_s\left(\boldsymbol{Z}_t\right)\sigma_{\min}\left(\boldsymbol{V}_L^T\boldsymbol{U}\right) \\ &\geq 0.4\mathrm{poly}(r)^{-1}\cdot\alpha\rho\left(1+\mu\hat{\sigma}_s^2\right)^{T_0} \\ &\gtrsim \mathrm{poly}(r)^{-1}\cdot\alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_s^2)}{3\log(1+\mu\hat{\sigma}_1)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \\ \|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| &\leqslant 2\left(\alpha\sigma_{s+1}^2\left(\boldsymbol{Z}_t\right)\|\boldsymbol{U}\| + \|\boldsymbol{E}_t\|\right) \\ &\lesssim \mathrm{poly}(r)\cdot\alpha^{1-\frac{2\log(1+\mu\hat{\sigma}_{s+1}^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}} \\ \left\|\boldsymbol{V}_{\boldsymbol{X}^\perp}^T\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| &\leqslant 100\left(\delta + \frac{\alpha\sigma_{s+1}\left(\boldsymbol{Z}_t\right)\|\boldsymbol{U}\| + \|\boldsymbol{E}_t\|}{\alpha\rho\sigma_s\left(\boldsymbol{Z}_t\right)}\right) \\ &\leq 100\left(\delta + \mathrm{poly}(r)\cdot\alpha^{\frac{2\log(1+\mu\hat{\sigma}_s^2)-2\log(1+\mu\hat{\sigma}_{s+1}^2)}{3\log(1+\mu\hat{\sigma}_1^2)-\log(1+\mu\hat{\sigma}_{s+1}^2)}}\right) \quad \leq 200\delta.\end{aligned} \tag{14}$$

The conclusion follows. □

C.2 THE PARALLEL IMPROVEMENT PHASE

C.2.1 THE SIGNAL TERM

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

Lemma C.4 *Suppose that $V_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t \in \mathbb{R}^{s\times r}$ is of full rank, $\mu < 10^{-4}\|\boldsymbol{X}\|^{-2}$, $c_3 < 10^{-3}\kappa_s^{-1}$, $\left\|(\mathcal{A}^*\mathcal{A}-I)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\| < 10^{-3}\kappa_s^{-1}\|\boldsymbol{X}\|^2$ and $\|V_{\boldsymbol{X}_s^\perp}V_{\boldsymbol{U}_t\boldsymbol{W}_t}\| \leq c_3$, then we have*

$$\begin{aligned}&\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_{t+1}) \geq \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_{t+1}\boldsymbol{W}_t) \\ &\geq \left(1+\mu\left(\sigma_s^2 - (5c_3+20\delta)\|\boldsymbol{X}\|^2\right) - 500\mu^2\|\boldsymbol{X}\|^4\right)\left(1-\mu\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t)\right)\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t).\end{aligned}$$

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}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top) + \mu\left[(\mathcal{A}^*\mathcal{A}-\boldsymbol{I})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\right)\boldsymbol{U}_t\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\left[(\mathcal{A}^*\mathcal{A}-\boldsymbol{I})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\right)\boldsymbol{U}_t\boldsymbol{W}_t \qquad (15a) \\ &= (\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\left[(\mathcal{A}^*\mathcal{A}-\boldsymbol{I})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\boldsymbol{U}_t\boldsymbol{W}_t \qquad (15b) \\ &= (\boldsymbol{I}+\mu\boldsymbol{\Sigma}_s^2)\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}\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}}^\top\left[(\mathcal{A}^*\mathcal{A}-\boldsymbol{I})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\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\left[(\mathcal{A}^*\mathcal{A}-\boldsymbol{I})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\boldsymbol{U}_t\boldsymbol{W}_t \\ &\quad - \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 + \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 \qquad (15c)\end{aligned}$$

where (15a) 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 $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{X}_{s,\perp} = 0$; (15b) 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 (15c) 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 (15c) 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 assumed to be invertible, so is $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}, \boldsymbol{\Sigma}_{\boldsymbol{U}_t\boldsymbol{W}_t}$ and $\boldsymbol{W}_{\boldsymbol{U}_t\boldsymbol{W}_t}$, thus we have

$$\begin{aligned}\boldsymbol{U}_t\boldsymbol{W}_t &= \boldsymbol{U}_t\boldsymbol{W}_t(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t \\ &= \boldsymbol{U}_t\boldsymbol{W}_t\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{\Sigma}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{W}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t \\ &= \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t.\end{aligned} \tag{16}$$

Plugging (16) into the second and third terms of (15) and re-arranging, we deduce that

$$\begin{aligned}&\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t \\ &= \left(\boldsymbol{I} + \mu(\boldsymbol{\Sigma}_s^2 + \boldsymbol{P}_1 + \boldsymbol{P}_2)\right)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t(\boldsymbol{I} - \mu\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t) \\ &\quad + \mu^2\left(\boldsymbol{\Sigma}_s^2 + \boldsymbol{P}_1 + \boldsymbol{P}_2\right)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t \\ &= \left[\boldsymbol{I} + \mu\left(\boldsymbol{\Sigma}_s^2 + \boldsymbol{P}_1 + \boldsymbol{P}_2\right) + \mu^2\left(\boldsymbol{\Sigma}_s^2 + \boldsymbol{P}_1 + \boldsymbol{P}_2\right)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\left(\boldsymbol{I} - \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-1}\right]\cdot \\ &\quad \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t(\boldsymbol{I} - \mu\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t)\end{aligned} \tag{17}$$

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), and

$$\begin{aligned}\boldsymbol{P}_1 &= \boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)^{-1} \\ \boldsymbol{P}_2 &= \boldsymbol{V}_{\boldsymbol{X}_s}^\top\left[(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right]\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)^{-1}\end{aligned} \tag{18}$$

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

$$\|\boldsymbol{P}_1\| \leq \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s^\perp}\right\| \cdot \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \cdot \left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)^{-1}\right\| \leq 5c_3\|\boldsymbol{X}\|^2 \leq 5\|\boldsymbol{X}\|^2 \tag{19}$$

and by our assumption we have

$$\|\boldsymbol{P}_2\| \leq \left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)^{-1}\right\| \cdot \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \leq 10^{-2}\kappa^{-1}r^{-\frac{1}{2}}. \tag{20}$$

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

$$\left\|\left(\boldsymbol{\Sigma}_s^2 + \boldsymbol{P}_1 + \boldsymbol{P}_2\right)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\left(\boldsymbol{I} - \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-1}\right\| \leq 500\|\boldsymbol{X}\|^4.$$

The equation (18) implies that

$$\begin{aligned}&\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t) \\ &\geq \sigma_{\min}\left(\boldsymbol{I} + \mu\boldsymbol{\Sigma}_s^2 + \boldsymbol{P}_1 + \boldsymbol{P}_2 + \left(\boldsymbol{\Sigma}_s^2 + \boldsymbol{P}_1 + \boldsymbol{P}_2\right)\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\left(\boldsymbol{I} - \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\right)^{-1}\right)\cdot \\ &\qquad \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t(\boldsymbol{I} - \mu\boldsymbol{W}_t^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t)\right) \\ &\geq \left(1 + \mu\sigma_{\min}^2(\boldsymbol{\Sigma}_s) - \mu\|\boldsymbol{P}_1\| - \mu\|\boldsymbol{P}_2\| - 500\mu^2\|\boldsymbol{X}\|^4\right)\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\left(1 - \mu\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\right) \\ &= \left(1 + \mu\sigma_s^2 - \mu\|\boldsymbol{P}_1\| - \mu\|\boldsymbol{P}_2\| - 500\mu^2\|\boldsymbol{X}\|^4\right)\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\left(1 - \mu\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\right)\end{aligned}$$

Recall that P_1 and P_2 are bounded in (19) and (20) respectively, so we have that

$$\begin{aligned}&\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}) \geq \sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t)\\ &\geq \left(1+\mu\left(\sigma_s^2-(5c_3+20\delta)\|\boldsymbol{X}\|^2\right)-500\mu^2\|\boldsymbol{X}\|^4\right)\left(1-\mu\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t)\right)\sigma_{\min}(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t).\end{aligned}$$

The conclusion follows. □

The corollaries below immediately follow from Lemma C.4.

Corollary C.2 *Under the conditions in Lemma C.4, if* $\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t) < 0.3(\sigma_s^2-\sigma_{s+1}^2) = 0.3\kappa_s^{-1}\|\boldsymbol{X}\|^2$*, then we have*

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

Corollary C.3 *Under the conditions in Lemma C.4, if*

$$\sigma_{\min}^2(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t) \leq \sigma_s^2 - 50(c_3+\delta\sqrt{r})\|\boldsymbol{X}\|^2,$$

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

C.2.2 THE NOISE TERM

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

Lemma C.5 *Suppose that* $\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t \in \mathbb{R}^{s\times r}$ *is of full rank,*

$$\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\| \leq c_3 < 10^{-3}\kappa_s^{-1}$$

and

$$\left\|(\mathcal{A}^*\mathcal{A}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top-\boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \leq 10^{-3}\kappa_s^{-1}c_3\|\boldsymbol{X}\|^2,$$

then we have

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

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

$$\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1,\perp} = -\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\right)^{-1}\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{W}_{t+1,\perp}. \tag{21}$$

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{(22a)} \\
&= -\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{(22b)} \\
&= -\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) + \left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\cdot \\
&\quad \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top \boldsymbol{W}_{t+1,\perp} && \text{(22c)} \\
&= \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 - \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}\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 $\boldsymbol{M}_1 = \boldsymbol{V}_{\boldsymbol{X}_s}^\top\left[\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s^\perp} - \left(\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right)\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\right]$. In (22), (22a) follows from (21), (22b) 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 (22c) we use $\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = 0$. It follows that

$$
\|(a)\| \le \mu\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right\|\cdot\left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right)^{-1}\right\|\|\boldsymbol{M}_1\|\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|. \tag{23}
$$

By Lemma D.4 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

$$
\left\|\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right)^{-1}\right\| = \sigma_{\min}^{-1}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right) = \left(1 - \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_t}\right\|^2\right)^{-\frac{1}{2}} \ge \frac{1}{2}. \tag{24}
$$

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

$$
\begin{aligned}
\|\boldsymbol{M}_1\| &\le \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s^\perp}\right\| + \left\|\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\| \\
&\le \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\right\|\cdot\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\right\| + 10^{-3}\kappa^{-1}c_3\|\boldsymbol{X}\|^2 \\
&\le 10\|\boldsymbol{X}\|^2 c_3.
\end{aligned} \tag{25}
$$

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 (23), (24) and (25) 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

$$
\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp} \\
&= \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\left[\boldsymbol{I} + \mu(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top) + \mu\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} && \text{(26a)} \\
&= \left(\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2 - \mu\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s^\perp} + \mu\underbrace{\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\left[\left(\mathcal{A}^*\mathcal{A} - \boldsymbol{I}\right)\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\boldsymbol{V}_{\boldsymbol{X}_s^\perp}}_{=:\boldsymbol{M}_2}\right)\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp} && \text{(26b)} \\
&= \left(\boldsymbol{I} + \mu\boldsymbol{\Sigma}_{s,\perp}^2 - \mu\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s^\perp} + \mu\boldsymbol{M}_2\right)\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\left(\boldsymbol{I} - \mu\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right) && \text{(26c)} \\
&\quad + \mu^2\left(\boldsymbol{\Sigma}_{s,\perp}^2 - \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s^\perp} + \boldsymbol{M}_2\right)\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} && \text{(26d)}
\end{aligned}
$$

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 (26), (26a) follows from the update rule of GD, (26b) 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 (26d) 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

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

To summarize, we have

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

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.6 *Suppose that $\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\| \le c_3$ and $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \le c_4$, with $\max\left\{c_3, c_4\|\boldsymbol{X}\|^{-1}\right\} \le 10^{-3}\kappa_s^{-1}$ and $\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \le 10^{-3}\kappa_s^{-1}c_3\|\boldsymbol{X}\|^2$ (where κ is the condition number defined in Section 3.1) and $\mu \le 10^{-4}\kappa_s^{-1}\|\boldsymbol{X}\|^{-2}c_3$, 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

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

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 $\operatorname{rank}(\boldsymbol{U}_t\boldsymbol{W}_t) \ge \operatorname{rank}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{W}_t\right) = \operatorname{rank}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\right) = s$, and $\boldsymbol{W}_t^\top\boldsymbol{W}_{t+1}$ is invertible by Lemma D.6 and the assumptions on μ and $\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\|$.

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 $\boldsymbol{Z}$. Following the line of proof of (Stöger & Soltanolkotabi, 2021, Lemma 9.3) (for completeness, we provide details in Lemma D.7), we deduce that

$$\begin{aligned}
&\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1}}\right\| = \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{V}_{\boldsymbol{Z}}\boldsymbol{W}_{\boldsymbol{Z}}^\top\right\|\\
&\le \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\left[\left(\boldsymbol{I} + \boldsymbol{B} - \frac{1}{2}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{B} + \boldsymbol{B}^\top\right)\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} - \boldsymbol{B}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{B} + \boldsymbol{B}^\top\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} + \boldsymbol{D}\right]\right\|\\
&\le \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\left(\boldsymbol{I} + \boldsymbol{B} - \frac{1}{2}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{B} + \boldsymbol{B}^\top\right)\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| + 2\|\boldsymbol{B}\|^2 + \|\boldsymbol{D}\|
\end{aligned} \tag{27}$$

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

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

so that

$$\begin{aligned}&\left\|\boldsymbol{B} - \mu(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t \boldsymbol{U}_t^\top)\right\| \\ &\leq \mu \|\boldsymbol{M}_t - (\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t \boldsymbol{U}_t^\top)\| + \|\boldsymbol{P}\| + \mu \|\boldsymbol{M}_t\| \|\boldsymbol{P}\| \\ &\leq \mu \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t \boldsymbol{U}_t^\top)\right\| + 2 \, \|\boldsymbol{W}_{t,\perp} \boldsymbol{W}_{t+1}\| + 4\mu \|\boldsymbol{X}\|^2 \, \|\boldsymbol{W}_{t,\perp} \boldsymbol{W}_{t+1}\| \\ &\leq \mu \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t \boldsymbol{U}_t^\top)\right\| + 6 \, \|\boldsymbol{W}_{t,\perp} \boldsymbol{W}_{t+1}\| \\ &\leq 18\mu \left(10\mu \|\boldsymbol{X}\|^3 + c_4\right) c_3 \|\boldsymbol{X}\| + 7\mu \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t \boldsymbol{U}_t^\top)\right\| \\ &\leq 18\mu \left(10\mu \|\boldsymbol{X}\|^3 + c_4\right) c_3 \|\boldsymbol{X}\| + 0.01\mu\kappa_s^{-1} c_3 \|\boldsymbol{X}\|^2\end{aligned} \tag{28}$$

where we use Lemma D.6 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

$$\begin{aligned}\boldsymbol{R}_1 &= \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \left(\boldsymbol{I} + \mu \left(\boldsymbol{I} - \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t} \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}^\top\right) \left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t \boldsymbol{U}_t^\top\right)\right) \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t} \\ &= \left(\boldsymbol{I} + \mu \boldsymbol{\Sigma}_{s,\perp}^2\right) \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t} \left(\boldsymbol{I} - \mu \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}\right) \\ &\quad - \mu \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \left(\boldsymbol{I} - \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t} \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}^\top\right) \boldsymbol{U}_t \boldsymbol{W}_{t,\perp} \boldsymbol{W}_{t,\perp}^\top \boldsymbol{U}_t^\top \boldsymbol{V}_{\boldsymbol{X}_s^\perp} \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t} \\ &\quad + \mu^2 \boldsymbol{\Sigma}_{s,\perp}^2 \boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t} \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}^\top \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}.\end{aligned} \tag{29}$$

By Weyl's inequality (cf. Lemma A.2) and our assumption on c_3,

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

So we have

$$\|\boldsymbol{R}_1\| \leq \left(1 - \frac{\mu}{2}(\sigma_s^2 - \sigma_{s+1}^2)\right) \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}\right\| + 3\mu \|\boldsymbol{X}\| c_3 c_4 + \mu^2 \|\boldsymbol{X}\|^4.$$

It thus follows from (27) that

$$\begin{aligned}&\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1} \boldsymbol{W}_{t+1}}\right\| \\ &\leq \|\boldsymbol{R}_1\| + 2\|\boldsymbol{B} - \boldsymbol{B}_1\| + 102\|\boldsymbol{B}\|^2 \\ &\leq \left(1 - \frac{\mu}{2}(\sigma_s^2 - \sigma_{s+1}^2)\right) \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t}\right\| + 40\mu c_3 c_4 \|\boldsymbol{X}\| + 0.02\mu\kappa_s^{-1} c_3 \|\boldsymbol{X}\|^2 + 10^3 \mu^2 \|\boldsymbol{X}\|^4.\end{aligned}$$

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

$$\hat{T}_s = \min\left\{t \geqslant 0 : \sigma_{\min}^2\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{\alpha,t+1}\right) > 0.3\left(\sigma_s^2 - \sigma_{s+1}^2\right) =: \tau_s\right\}.$$

In this section, we show that when $T_0 \leq t < \hat{T}_{\alpha,s}$, 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.3.

Lemma C.7 *Let* $\max\{c_3, c_4\|\boldsymbol{X}\|^{-1}\} \leq 10^{-3}\kappa_s^{-1}$, $\delta \leq 10^{-4}\kappa_s^{-1}r_*^{-\frac{1}{2}}c_3$ *and* $\mu \leq 10^{-4}\kappa_s^{-1}\|\boldsymbol{X}\|^{-2}$. *Then the following holds for all* $T_0 \leq t < \hat{T}_{\alpha,s}$ *as long as* $\alpha \leq \mathrm{poly}(r)^{-1}$*:*

$$\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right) \geq \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\right) \geq \left(1 + 0.5\mu\left(\sigma_s^2 + \sigma_{s+1}^2\right)\right)\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{\alpha,t}\right) \tag{30a}$$

$$\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \leq \min\left\{\left(1 + \mu\left(0.4\sigma_s^2 + 0.6\sigma_{s+1}^2\right)\right)\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|, c_4\right\} \tag{30b}$$

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1}}\right\| \leq c_3. \tag{30c}$$

$$\mathrm{rank}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right) = \mathrm{rank}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_t\right) = s. \tag{30d}$$

Proof **:** The base case $t = T_0$ is already proved in (12). 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 $\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\|$ as follows:

$$\begin{aligned}
&\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \\
&\leq \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top)\right\| + \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\right\| \\
&\leq 10\delta\sqrt{r_*}\|\boldsymbol{X}\|^2 + \delta\left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\right\|_* \\
&\leq 10\delta\sqrt{r_*}\|\boldsymbol{X}\|^2 + \delta\sqrt{d}\left(1 + \mu(0.4\sigma_s^2 + 0.6\sigma_{s+1}^2)\right)^{t-T_0}\|\boldsymbol{U}_{T_0}\boldsymbol{W}_{T_0,\perp}\|
\end{aligned}$$

By induction hypothesis, it's easy to see that

$$\frac{\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\right)}{\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|} \geq \frac{\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{T_0}\right)}{\|\boldsymbol{U}_{T_0}\boldsymbol{W}_{T_0,\perp}\|} \geq \mathrm{poly}(r)\cdot\alpha^{-\gamma_s} \tag{31}$$

where

$$\gamma_s = \frac{2\left(\log\left(1 + \mu\hat{\sigma}_s^2\right) - \log\left(1 + \mu\hat{\sigma}_{s+1}^2\right)\right)}{3\log\left(1 + \mu\hat{\sigma}_1^2\right) - \log\left(1 + \mu\hat{\sigma}_{s+1}^2\right)} > 0.$$

Since we must have $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\right) \leq 0.3\tau_s = \mathcal{O}(1)$ by definition of $\hat{T}_{\alpha,s}$, it follows that $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \leq \mathrm{poly}(r)\alpha^{\gamma_s}$, so for sufficiently small $\alpha \leq \mathrm{poly}(r)^{-1}$, we have $\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \leq 11\delta\sqrt{r}\|\boldsymbol{X}\|^2$.

The above inequality combined with our assumption on δ implies that the conditions on $\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\|$ in Lemmas C.4 to C.6 hold. We now show that (30a) to (30d) hold for $t+1$, which completes the induction step.

First, since $t < \hat{T}_s$, we have $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right) \leq \tau_s$. 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.2 are all satisfied, and we deduce that (30a) holds.

Second, the assumptions on c_3, c_4 and δ, combined with Lemma C.5, 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 (31) we observe that

$$\frac{\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right)}{\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\|} \geq \frac{\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{T_0}\right)}{\|\boldsymbol{U}_{T_0}\boldsymbol{W}_{T_0,\perp}\|} \geq \mathrm{poly}(r)\cdot\alpha^{-\gamma_s}.$$

Since $\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\right) \leq \|\boldsymbol{X}\|$, when $\alpha < \mathrm{poly}(r)^{-1}$ we must have that $\|\boldsymbol{U}_{t+1}\boldsymbol{W}_{t+1,\perp}\| \leq c_4$.

Finally, Lemma C.6 implies that (30c) is true, and (30d) follows from our application of Lemma C.4. This concludes the proof. □

C.4 The refinement phase and concluding the proof of Theorem 4.1

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* $\hat{T}_s$. We begin with the following lemma, which is straightforward from the proof of Lemma C.7.

Lemma C.8 *The following inequality holds when $\alpha \le \mathrm{poly}(r)^{-1}$ is sufficiently small:*

$$\left\| \boldsymbol{U}_{\hat{T}_s} \boldsymbol{W}_{\hat{T}_s,\perp} \right\| \le \mathrm{poly}(r) \cdot \alpha^{\gamma_s}.$$

The following lemma states that in a certain time period after $\hat{T}_s$, the parallel and orthogonal components still behave similarly to the second (parallel improvement) phase.

Lemma C.9 *There exists $\hat{t}_s = \Theta\left(\log \frac{1}{\alpha}\right)$ when $\alpha \to 0$ (here we omit the dependence of $\hat{t}_s$ on α for simplicity) such that when $0 \le t - \hat{T}_s \le \hat{t}_s$, we have*

$$\sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\right) \ge \sigma_{\min}\left(\boldsymbol{U}_t \boldsymbol{W}_t\right) \ge 0.3\tau_s \tag{32a}$$

$$\|\boldsymbol{U}_t \boldsymbol{W}_t\| \le \left(1 + \mu(0.4\sigma_s^2 + 0.6\sigma_{s+1}^2)\right)^{t-\hat{T}_s} \left\| \boldsymbol{U}_{\hat{T}_s} \boldsymbol{W}_{\hat{T}_s} \right\| \tag{32b}$$

$$\left\| \boldsymbol{V}_{\boldsymbol{X}_s^\perp} \boldsymbol{V}_{\boldsymbol{U}_t \boldsymbol{W}_t} \right\| \le c_3. \tag{32c}$$

Proof **:** We choose $\hat{t}_s = \min\left\{t \ge 0 : \|\boldsymbol{U}_{t+1} \boldsymbol{W}_{t+1,\perp}\|^2 \le c_5\right\}$ where

$$c_5 = 10^{-4} d^{-\frac{1}{2}} \kappa_s^{-1} c_3 \|\boldsymbol{X}\|^2 \tag{33}$$

We prove (32) by induction. The proof follows the idea of Lemma C.7, except that we need to bound $\left\|(\boldsymbol{I} - \mathcal{A}^*\mathcal{A})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}\boldsymbol{U}^\top\right)\right\|$ in each induction step. Concretely, suppose that (32) holds at time t, then

$$\begin{aligned}
&\left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right\| \\
\le& \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top)\right\| + \left\|(\mathcal{A}^*\mathcal{A} - \boldsymbol{I})\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\right\| \\
\le& 10\delta\sqrt{r_*}\|\boldsymbol{X}\|^2 + \delta\left\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\boldsymbol{W}_{t,\perp}^\top\boldsymbol{U}_t^\top\right\|_* \\
\le& 10\delta\sqrt{r_*}\|\boldsymbol{X}\|^2 + \delta c_5\sqrt{d} \le 10^{-3}\kappa_s^{-1}c_3\|\boldsymbol{X}\|^2
\end{aligned} \tag{34}$$

where we used the definition of c_5 in the last step. As a result, we can apply the conclusion of Lemmas C.4 to C.6 which implies that (32) holds for $t+1$. Finally, combining Lemma C.8 and (32b) yields $\hat{t}_s = \omega(1)$. □

We now present the main result of this section:

Lemma C.10 *Suppose that $0 \le t - \hat{T}_s \le \hat{t}_s$, $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top \boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \le c_3$ and the conditions on c_3, c_4, δ and μ in Lemma C.7 hold, then we have*

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

where we recall that $\tau_s = \sigma_s^2 - \sigma_{s+1}^2 > 0$.

Proof **:** The update of GD implies that

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

where

$$\begin{aligned}
\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\mathcal{E}_{t,1}\right\|_F &= \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{U}_t\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_F \\
&\le \sqrt{r_*}\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{U}_t\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_2 \\
&\le 10^3\sqrt{r_*}\|\boldsymbol{X}\|^2
\end{aligned}$$

and

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

Note that we would like to bound $\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top\right)\right\|$. 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 \\ &\quad + \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 \qquad (35a) \\ &\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}\|^4 c_3 \qquad (35b) \\ &\le \left(1 - \frac{1}{2}\mu\tau_s\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}\|^4 c_3, \qquad (35c)\end{aligned}$$

where in (35a) we use $\left\|\boldsymbol{I} - \mu\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\| \le 1$, (35b) 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 (35c) 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

$$\begin{aligned}&\left\|\left[(\boldsymbol{I} - \mathcal{A}^*\mathcal{A})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t\right)\right]\boldsymbol{U}_t\boldsymbol{U}_t^\top + \boldsymbol{U}_t\boldsymbol{U}_t^\top\left[(\boldsymbol{I} - \mathcal{A}^*\mathcal{A})\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right]\right\| \\ &\le 0.1\kappa_s^{-1}c_3\|\boldsymbol{X}\|^4\end{aligned} \qquad (36)$$

where we use the estimate in (34). Combining (35) and (36) yields

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

□

To apply the result of Lemma C.10, we need to verify that $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \le c_3$ still holds when $t \ge \hat{T}_s$. In fact, this is true as long as $t - \hat{T}_s \le \mathcal{O}\left(\log\frac{1}{\alpha}\right)$.

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

Theorem C.1 *When $\alpha < \mathrm{poly}(r_*)^{-1}$ and $\delta = 10^{-4}\kappa_s^{-1}c_3$ with $c_3 < 10^{-3}\kappa_s^{-1}r_*^{-\frac{1}{2}}$, there exists a time $T_s \in \mathbb{Z}_+$ such that*

$$\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\| \le 10^3\tau_s^{-2}\|\boldsymbol{X}\|^6\sqrt{r_*}c_3.$$

Proof **:** First, observe that for all $t \geq 0$,

$$\begin{aligned}\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_F &\leq \left\|\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\right\|_F + \left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\right\|_F \\ &\leq \left\|\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{V}_{\boldsymbol{X}_s}\boldsymbol{V}_{\boldsymbol{X}_s}^\top\right\|_F + \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\right\|_F \\ &\leq \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F + \sqrt{r_*}\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_t\right\|^2 + \sqrt{d}\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\right\|^2 \\ &\leq \left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\| + 9\sqrt{r_*}\|\boldsymbol{X}\|^2\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\|^2 + \sqrt{d}\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|^2.\end{aligned} \tag{37}$$

We set $\delta = 10^{-3}\|\boldsymbol{X}\|^{-2}\tau_s c_3$ and

$$T_s = \hat{T}_s - \frac{\log\left(10^{-2}\|\boldsymbol{X}\|^{-2}\tau_s c_3^{-1}\right)}{\log\left(1 - \frac{1}{2}\mu\tau_s\right)},$$

then for small α we have $T_s \leq \hat{T}_s + \hat{t}_s$ (defined in Lemma C.9). Hence for $\hat{T}_s \leq t < T_s$ we always have $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \leq c_3$. By Lemma C.10 and the choice of c_3 and δ, we have for $\hat{T}_s \leq t < T_s$ that

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top\right)\right\|_F \leq \left(1 - \frac{1}{2}\mu\tau_s\right)\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right\|_F + 150\mu\|\boldsymbol{X}\|^4\sqrt{r_*}c_3$$

which implies that

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_{T_s}\boldsymbol{U}_{T_s}^\top\right)\right\|_F \leq 400\tau_s^{-1}\|\boldsymbol{X}\|^4\sqrt{r_*}c_3.$$

Meanwhile, by Lemma C.9 we have $\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \leq c_5$ and $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \leq c_3$ at $t = T_s$. Plugging into (37) yields

$$\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_{T_s}\boldsymbol{U}_{T_s}^\top\right\|_F \leq 400\tau_s^{-1}\|\boldsymbol{X}\|^4\sqrt{r_*}c_3 + 9\|\boldsymbol{X}\|^2c_3^2\sqrt{r_*} + c_5^2\sqrt{d}.$$

By definition of c_3 and c_5 we deduce that $\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \boldsymbol{U}_{T_s}\boldsymbol{U}_{T_s}^\top\right\|_F \leq 10^3\tau_s^{-2}\|\boldsymbol{X}\|^6\sqrt{r_*}c_3$, as desired. □

D AUXILIARY RESULTS FOR PROVING THEOREM 4.1

D.1 THE SPECTRAL PHASE

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

Recall that $\boldsymbol{Z}_t = (\boldsymbol{I} + \mu\boldsymbol{M})^t$ and $\boldsymbol{U}_t = \widetilde{\boldsymbol{U}}_t + \boldsymbol{E}_t = \boldsymbol{Z}_t\boldsymbol{U}_0 + \boldsymbol{E}_t$ and $\boldsymbol{U}_0 = \alpha\boldsymbol{U}$. 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 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 $\rho = \sigma_{\min}\left(\boldsymbol{V}_{\boldsymbol{M}_s}^\top\boldsymbol{U}\right) > 0$, then the following three inequalities hold, given that the denominator of the third is positive.*

$$\sigma_s(\boldsymbol{U}_t) \geq \alpha\left(\rho\sigma_s\left(\boldsymbol{Z}_t\right) - \sigma_{s+1}\left(\boldsymbol{Z}_t\right)\|\boldsymbol{U}\|\right) - \|\boldsymbol{E}_t\|, \tag{38a}$$

$$\sigma_{s+1}(\boldsymbol{U}_t) \leq \alpha\sigma_{s+1}\left(\boldsymbol{Z}_t\right)\|\boldsymbol{U}\| + \|\boldsymbol{E}_t\|, \tag{38b}$$

$$\left\|\boldsymbol{V}_{\boldsymbol{M}_s^\perp}^\top\boldsymbol{V}_{L_t}\right\| \leq \frac{\alpha\sigma_{s+1}\left(\boldsymbol{Z}_t\right)\|\boldsymbol{U}\| + \|\boldsymbol{E}_t\|}{\alpha\rho\sigma_s\left(\boldsymbol{Z}_t\right) - 2\left(\alpha\sigma_{s+1}\left(\boldsymbol{Z}_t\right)\|\boldsymbol{U}\| + \|\boldsymbol{E}_t\|\right)}. \tag{38c}$$

Proof **:** By Weyl's inequality we have

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

Thus (38b) holds. Similarly,

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

Finally, note that we can write

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

so that the subspace spanned by the left singular vectors of $\alpha(1+\mu\boldsymbol{M}_s)^t\boldsymbol{U}$ 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 (38c). □

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* $\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{V}_{L_t}\right\| \leq 0.1$ *for some* $t \geq 1$. *Then it holds that*

$$\sigma_s\left(\boldsymbol{U}_t\boldsymbol{W}_t\right) \geq \frac{1}{2}\sigma_s\left(\boldsymbol{U}_t\right), \tag{39a}$$

$$\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| \leq 10\left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{V}_{L_t}\right\|, \tag{39b}$$

$$\|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| \leq 2\sigma_{s+1}\left(\boldsymbol{U}_t\right). \tag{39c}$$

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

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

$$\begin{aligned}\sigma_s\left(\boldsymbol{U}_t\boldsymbol{W}_t\right) &\geq 0.4\alpha\sigma_{r_\star}\left(\boldsymbol{Z}_t\right)\sigma_{\min}\left(\boldsymbol{V}_L^\top\boldsymbol{U}\right) \\ \|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\| &\leq 2\left(\alpha\sigma_{s+1}\left(\boldsymbol{Z}_t\right)\|\boldsymbol{U}\| + \|\boldsymbol{E}_t\|\right) \\ \left\|\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right\| &\leq 100\left(\delta + \frac{\alpha\sigma_{s+1}\left(\boldsymbol{Z}_t\right)\|\boldsymbol{U}\| + \|\boldsymbol{E}_t\|}{\alpha\rho\sigma_s\left(\boldsymbol{Z}_t\right)}\right)\end{aligned} \tag{40}$$

D.2 The parallel improvement phase

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

Lemma D.3 *(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.4 *Under the assumptions in Lemma C.5, 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.5 *Under the assumptions in Lemma C.6, 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.6 *Under the assumptions in Lemma C.6, 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

$$\begin{aligned}
&\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_{t+1}\boldsymbol{W}_{t,\perp} \\
&= \boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{I} + \mu\mathcal{A}^*\mathcal{A}(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} \\
&= \boldsymbol{V}_{\boldsymbol{X}_s}^\top\left(\boldsymbol{I} + \mu(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\right)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top(\mathcal{A}^*\mathcal{A}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} \\
&= -\mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top(\mathcal{A}^*\mathcal{A}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} && \text{(41a)} \\
&= -\mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} + \mu\boldsymbol{V}_{\boldsymbol{X}_s}^\top(\mathcal{A}^*\mathcal{A}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} && \text{(41b)} \\
&= -\mu\underbrace{\boldsymbol{V}_{\boldsymbol{X}_s}^\top \boldsymbol{U}_t\boldsymbol{W}_t\boldsymbol{W}_t^\top\boldsymbol{U}_t^\top\boldsymbol{V}_{\boldsymbol{X}_s^\perp}\boldsymbol{V}_{\boldsymbol{X}_s^\perp}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}}_{=:\boldsymbol{K}_1} + \mu\underbrace{\boldsymbol{V}_{\boldsymbol{X}_s}^\top(\mathcal{A}^*\mathcal{A}-\boldsymbol{I})(\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{U}_t\boldsymbol{U}_t^\top)\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}}_{:=\boldsymbol{K}_2} && \text{(41c)}
\end{aligned}$$

where (41a) 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 (41b) and (41c) we use $\boldsymbol{V}_{\boldsymbol{X}_s}^\top\boldsymbol{U}_t\boldsymbol{W}_{t,\perp} = 0$.

For $\boldsymbol{K}_1$, note that

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

so that

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

Plugging into (41), we deduce that

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

where in the last step we use Lemma D.5 and the induction hypothesis which implies that $\sigma_{\min}(\boldsymbol{U}_t\boldsymbol{W}_t) \geq \|\boldsymbol{U}_t\boldsymbol{W}_{t,\perp}\|$. □

Lemma D.7 *The matrix $\boldsymbol{Z}$ defined in the proof of Lemma C.6 satisfies the following:*

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

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

Proof **:** By definition of $\boldsymbol{Z}$ we have

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

It follows from (28) and our assumptions on c_3 and c_4 that

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

(note that this step is independent and does not rely on earlier derivations in the proof of Lemma C.6), so by Taylor's formula, we have

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

Hence,

$$\begin{aligned}
&\left\|\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-\frac{1}{2}} - \left(\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} + \boldsymbol{B}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t} - \frac{1}{2}(\boldsymbol{I}+\boldsymbol{B})\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\left(\boldsymbol{B}+\boldsymbol{B}^\top\right)\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)\right\| \\
&= \left\|(\boldsymbol{I}+\boldsymbol{B})\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\left((\boldsymbol{I}+\Delta)^{-\frac{1}{2}} - \boldsymbol{I} + \frac{1}{2}\Delta - \frac{1}{2}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}^\top\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{V}_{\boldsymbol{U}_t\boldsymbol{W}_t}\right)\right\| \\
&\leq (1+\|\boldsymbol{B}\|)\left(3\|\Delta\|^2 + \frac{1}{2}\|\boldsymbol{B}\|^2\right) < 30\|\boldsymbol{B}\|^2
\end{aligned}$$

as desired. □

E Proof of results in Section 5

Theorem 1.1 states that GD approximately learns the rank-s constrained minimizer of the matrix sensing loss. However, Theorem C.1 only implies that GD would get into an $\mathcal{O}(\delta)$ neighborhood of $\boldsymbol{X}_s\boldsymbol{X}_s^\top$. As a result, a more fine-grained analysis is needed. Note that it is not even clear whether the rank-s minimizer of is unique. If it is not, then we may naturally ask which minimizer it converges to.

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

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

and establish local convergence of gradient descent. Most existing results only study the landscape of (1) in the exact- and over-parameterized case. Zhu et al. (2021) studies the landscape of under-parameterized matrix factorization problem, but they only prove a strict saddle property without asymptotic convergence rate of GD.

When the measurement satisfies the RIP condition, we can expect that the landscape of f 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 $s < r$. 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's 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 $\left(\lambda_i \boldsymbol{I}_{m_i} - \boldsymbol{W}_i\right)\boldsymbol{W}_i = 0$. 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 minima, then so is $\boldsymbol{U}\boldsymbol{R}$ for any orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{s\times s}$. Instead, we will establish a local PL property w.r.t the procrustes distance:

Definition E.1 *Let $\boldsymbol{U}_1, \boldsymbol{U}_2 \in \mathbb{R}^{d\times s}$, we define*

$$\operatorname{dist}(\boldsymbol{U}_1, \boldsymbol{U}_2) = \inf\left\{\|\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{R}\|_F : \boldsymbol{R} \in \mathbb{R}^{s\times s} \text{ is orthogonal}\right\}.$$

The following characterization of the optimal $\boldsymbol{R}$ in Definition 5.1 is known in the literature (see e.g. (Tu et al., 2016, Section 5.2.1)) but we provide a proof of it for completeness.

Lemma E.2 *Let $\boldsymbol{R}$ be the orthogonal matrix which minimizes $\|\boldsymbol{U}_1 - \boldsymbol{U}_2\boldsymbol{R}\|_F$, then $\boldsymbol{U}_1^\top \boldsymbol{U}_2 \boldsymbol{R}$ is positive semi-definite.*

Proof **:** Observe that

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

Let $\boldsymbol{U}_2^\top \boldsymbol{U}_1 = A\boldsymbol{\Sigma}B^\top$ be its SVD, then

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

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 $\boldsymbol{B}^\top \boldsymbol{R}^\top \boldsymbol{A} = \boldsymbol{I}$ i.e. $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{B}^\top$. Thus $\boldsymbol{U}_1^\top \boldsymbol{U}_2 \boldsymbol{R} = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top$ is positive semi-definite. □

The following lemma states that the minimizer of matrix sensing loss is also near-optimal for the matrix factorization loss. The main idea for proving this result is to utilize

Lemma E.3 *Let $\boldsymbol{U}_s^*$ be a global minimizer of f, then we have*

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

Proof **:** Since $\boldsymbol{U}_s^*$ is the global minimizer of f, by the RIP property we have

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

as desired. □

The next lemma show that F_s satisfies a local PL property:

Lemma E.4 *Given $\boldsymbol{U} \in \mathbb{R}^{d\times s}$ and let $\boldsymbol{R}$ be an orthogonal matrix such that $\|\boldsymbol{U} - \boldsymbol{X}_s\boldsymbol{R}\|_F = \mathrm{dist}(\boldsymbol{U}, \boldsymbol{X}_s)$. Suppose that $\mathrm{dist}(\boldsymbol{U}, \boldsymbol{X}_s) \leq 0.1\|\boldsymbol{X}\|^{-1}\tau_s$ (where we recall that $\tau_s = \sigma_s^2 - \sigma_{s+1}^2$ is the eigengap of $\boldsymbol{X}\boldsymbol{X}^\top$), then we have*

$$\langle \nabla F_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{X}_s\boldsymbol{R}\rangle \geq 0.1\tau_s \mathrm{dist}^2(\boldsymbol{U}, \boldsymbol{X}_s).$$

Proof **:** Assume WLOG that $\boldsymbol{R} = \boldsymbol{I}$, then $\boldsymbol{U}^\top\boldsymbol{X}_s$ is 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 that

$$\begin{aligned}&\langle \nabla F_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{X}_s\rangle \\ &= \left\langle \left[(\boldsymbol{H}+\boldsymbol{X}_s)(\boldsymbol{H}+\boldsymbol{X}_s)^\top - \boldsymbol{X}\boldsymbol{X}^\top\right](\boldsymbol{H}+\boldsymbol{X}_s), \boldsymbol{H}\right\rangle \\ &= \mathrm{tr}\left(\boldsymbol{H}^\top\left[(\boldsymbol{H}+\boldsymbol{X}_s)(\boldsymbol{H}+\boldsymbol{X}_s)^\top - \boldsymbol{X}\boldsymbol{X}^\top\right]\boldsymbol{H} + \boldsymbol{H}^\top\left(\boldsymbol{H}\boldsymbol{H}^\top + \boldsymbol{H}\boldsymbol{X}_s^\top + \boldsymbol{X}_s\boldsymbol{H}^\top\right)\boldsymbol{X}_s\right) \\ &\geq -\mathrm{tr}\left(\boldsymbol{H}^\top\boldsymbol{X}_{s,\perp}\boldsymbol{X}_{s,\perp}^\top\boldsymbol{H}\right) - 3\|\boldsymbol{X}\|\|\boldsymbol{H}\|_F^3 + \mathrm{tr}\left(\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{X}_s^\top\boldsymbol{X}_s\right) \qquad (42a) \\ &\geq \left(\sigma_s^2 - \sigma_{s+1}^2\right)\|\boldsymbol{H}\|_F^2 - 3\|\boldsymbol{X}\|\|\boldsymbol{H}\|_F^3 \geq 0.1\tau\|\boldsymbol{H}\|_F^2 \qquad (42b)\end{aligned}$$

where in (42a) we use $\mathrm{tr}\left((\boldsymbol{H}^\top\boldsymbol{X}_s)^2\right) \geq 0$ (since $\boldsymbol{H}^\top\boldsymbol{X}_s$ is symmetric as noticed in the beginning of the proof), and (42b) is because of

$$\begin{aligned}\mathrm{tr}\left(\boldsymbol{H}^\top\boldsymbol{H}\boldsymbol{X}_s^\top\boldsymbol{X}_s\right) &= \mathrm{tr}\left(\boldsymbol{H}^\top\boldsymbol{H}W_{\boldsymbol{X}_s}\boldsymbol{\Sigma}_{\boldsymbol{X}_s}^2W_{\boldsymbol{X}_s}^\top\right) \\ &= \mathrm{tr}\left(W_{\boldsymbol{X}_s}^\top\boldsymbol{H}^\top\boldsymbol{H}W_{\boldsymbol{X}_s}\boldsymbol{\Sigma}_{\boldsymbol{X}_s}^2\right) \\ &\geq \sigma_s^2\,\mathrm{tr}\left(W_{\boldsymbol{X}_s}^\top\boldsymbol{H}^\top\boldsymbol{H}W_{\boldsymbol{X}_s}\right) = \sigma_s^2\|\boldsymbol{H}\|_F^2\end{aligned}$$

and

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

□

Corollary E.1 *Under the conditions of Lemma E.4, we have $\|\nabla F_s(\boldsymbol{U})\|_F \geq 0.1\tau_s\mathrm{dist}(\boldsymbol{U}, \boldsymbol{X}_s)$.*

The following lemma shows that the rank-s global minima of matrix sensing must lie in an $\mathcal{O}(\delta)$-neighbourhood of the minima of F_s.

Lemma E.5 *Suppose that the conditions in Lemma E.4 hold. Let $\boldsymbol{U}_s^*$ be a global minimizer of f, then we have*

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

where we recall that $\kappa_s = \tau_s^{-1}\|\boldsymbol{X}\|$ is the condition number of $\boldsymbol{X}\boldsymbol{X}^\top$.

Proof **:** Define

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

First we can show that $\boldsymbol{U}_s^* \in S$. The main idea is to apply Lemma A.6. Indeed, it's easy to see that

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

Let $\boldsymbol{U} \in \partial S$ i.e. $\operatorname{dist}^2(\boldsymbol{U}, \boldsymbol{X}_s) = 0.1\|\boldsymbol{X}\|^{-1}\tau_s$. Assume WLOG that $\operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}_s) = \|\boldsymbol{U} - \boldsymbol{X}_s\|_F$, then by Lemma E.4 we have

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

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

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

On the other hand, we know from Lemma E.3 that

$$F_s(\boldsymbol{U}_s^*) - F_s^* \le 5\delta r\|\boldsymbol{X}\|^4 < 0.5\tau_s^2, \tag{43}$$

so Lemma A.6 implies that $\boldsymbol{U}_s^* \in S$.

Inside S, we can apply the local PL property that we previously derived. Indeed, note that

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

Hence we have that

$$\operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}_s) \le 40\delta\tau_s^{-1}\|\boldsymbol{X}\|^2\|\boldsymbol{X}\|_F.$$

□

Corollary E.2 *Under the conditions in Lemma E.4, we have* $\sigma_{\min}\left(\left(\boldsymbol{U}_s^*\right)^\top \boldsymbol{U}_s^*\right) \ge \sigma_s^2 - 80\delta\kappa\|\boldsymbol{X}\|\|\boldsymbol{X}\|_F$.

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 5.1. By Lemma E.5, we have that

$$\begin{aligned} \left\| \left(\boldsymbol{U}_s^*\right)^\top \boldsymbol{U}_s^* - \boldsymbol{X}_s^\top \boldsymbol{X}_s \right\| &\le \left\| \left(\boldsymbol{U}_s^*\right)^\top \boldsymbol{U}_s^* - \boldsymbol{X}_s^\top \boldsymbol{X}_s \right\| \\ &\le \max\left\{\|\boldsymbol{U}_s^*\|, \|\boldsymbol{X}_s\|\right\} \cdot \|\boldsymbol{U}_s^* - \boldsymbol{X}_s\| \\ &\le 80\delta\kappa_s\|\boldsymbol{X}\|\|\boldsymbol{X}\|_F. \end{aligned}$$

□

Lemma E.6 *Given* $\boldsymbol{U} \in \mathbb{R}^{d\times s}$*, let* $\boldsymbol{U}_s^* \in \mathbb{R}^{d\times s}$ *be a rank-s minimizer of* f_s*, and* $\boldsymbol{U}_s^*\boldsymbol{R}$ *be the rank-s minimizer which is closest to* $\boldsymbol{U}$ *(*$\boldsymbol{R} \in \mathbb{R}^{s\times s}$ *is orthogonal). When* $\|\boldsymbol{U} - \boldsymbol{U}_s^*\boldsymbol{R}\| \le 10^{-2}\kappa_s^{-1}\|\boldsymbol{X}\|$ *and* $\delta \le 10^{-3}r_*^{-1}\kappa_s^{-2}$*, we have*

$$\left\langle \nabla f_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{U}_s^*\boldsymbol{R} \right\rangle \ge 0.1\tau_s \operatorname{dist}(\boldsymbol{U}, \boldsymbol{U}_s^*)^2.$$

Proof **:** We assume WLOG that $\boldsymbol{R} = \boldsymbol{I}$, then $\boldsymbol{U}^\top \boldsymbol{U}_s^*$ is positive semi-definite. Let $\boldsymbol{H} = \boldsymbol{U} - \boldsymbol{U}_s^*$, then

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

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 E.5, we may thus deduce that

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

Hence

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

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

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

When $\delta \le 10^{-3} r_*^{-\frac{1}{2}}\kappa_s^{-2}$ and $\|\boldsymbol{H}\| \le 10^{-2}\tau_s\|\boldsymbol{X}\|^{-1}$, the above implies that $\langle \nabla f_s(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{U}_s^*\rangle \ge 0.5\tau_s\|\boldsymbol{H}\|_F^2$, as desired. □

F Proof of Theorem 4.2

With the landscape results in hand, we are now ready to characterize the saddle-to-saddle dynamics of GD. We first note the following proposition, with is straightforward from Lemma C.9 and Theorem C.1. In the following we use $\boldsymbol{U}_{\alpha,t}$ to denote the t-th iteration of GD when initialized at $\boldsymbol{U}_0 = \alpha\boldsymbol{U}$.

Proposition F.1 *There exists matrices $\bar{\boldsymbol{U}}_{0,\alpha}$ with rank $\le s$ such that*

$$\left\|\bar{\boldsymbol{U}}_{0,\alpha} - \boldsymbol{U}_{\alpha,T_s^\alpha}\right\|_F = o(1) \quad (\alpha \to 0)$$

where T_s^α is defined in Theorem C.1 (we omit the dependence on α there) and moreover

$$\left\|\bar{\boldsymbol{U}}_0\bar{\boldsymbol{U}}_0^\top - \boldsymbol{Z}_s\right\| \le 100\kappa^2\|\boldsymbol{X}\|^2 r_*\delta.$$

where $\boldsymbol{Z}_s = \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top$ is the rank-s minimizer of the matrix sensing loss i.e.

$$\boldsymbol{Z}_s = \arg\min\left\{\frac{1}{2}\left\|\mathcal{A}(\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{X}^\top)\right\|_2^2 : \boldsymbol{Z} \in \mathbb{R}^{d\times d} \text{ is positive semi-definite and } \operatorname{rank}(\boldsymbol{Z}) \le s\right\}. \tag{44}$$

Remark F.1 *We omit the dependence on α for simplicity of notations, when it is clear from context.*

Proof **:** Since $T_s - \hat{T}_s = \mathcal{O}(1)$, it can be seen from the proof of Lemma C.9 that $\|\boldsymbol{U}_{T_s}\boldsymbol{W}_{T_s,\perp}\| = o(1)$. We choose $\bar{\boldsymbol{U}}_0 = \boldsymbol{U}_{T_s}\boldsymbol{W}_{T_s}\boldsymbol{W}_{T_s}^\top$, then $\operatorname{rank}\left(\bar{\boldsymbol{U}}_0\right) \le s$ and moreover by Theorem C.1 we have $\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \bar{\boldsymbol{U}}_0\bar{\boldsymbol{U}}_0^\top\right\|_F \le 2\kappa_s^2\|\boldsymbol{X}\|^2\sqrt{r_*}\delta$. On the other hand, similar to Corollary E.2 we have that $\left\|\boldsymbol{Z}_s - \boldsymbol{X}_s\boldsymbol{X}_s^\top\right\|_F \le 80\delta\kappa_s\|\boldsymbol{X}\|_F^2$. Thus $\left\|\bar{\boldsymbol{U}}_0\bar{\boldsymbol{U}}_0^\top - \boldsymbol{Z}_s\right\| \le 100\kappa_s^2\|\boldsymbol{X}\|^2 r_*\delta$ as desired. □

Let $\widetilde{\boldsymbol{U}}_0 \in \mathbb{R}^{d\times s}$ satisfies $\widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{U}}_0^\top = \bar{\boldsymbol{U}}_0\bar{\boldsymbol{U}}_0^\top$. The following corollary shows that $\widetilde{\boldsymbol{U}}_0$ is close to $\boldsymbol{U}_s^*$ in terms of the procrutes distance.

Corollary F.1 *We have* $\operatorname{dist}(\widetilde{\boldsymbol{U}}_0, \boldsymbol{U}_s^*) \le 80\kappa_s^3 r_*^{\frac{1}{2}}\|\boldsymbol{X}\|\delta$.

Proof **:** We know from Lemma E.5 that $\operatorname{dist}(\boldsymbol{U}_s^*, \boldsymbol{X}_s) \le 40\delta\kappa\|\boldsymbol{X}\|_F$, so it remains to bound $\operatorname{dist}(\widetilde{\boldsymbol{U}}_0, \boldsymbol{X}_s)$.

The proof idea is the same as that of Lemma E.5, so we only provide a proof sketch here. It has been shown in the proof of Proposition F.1 that

$$F_s(\widetilde{\boldsymbol{U}}_0) := \frac{1}{2}\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{U}}_0^\top\right\|_F^2 \le r_*\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{U}}_0^\top\right\|^2 \le 4\kappa_s^4 r_*\|\boldsymbol{X}\|^4\delta^2 \le 0.5\tau_s^2.$$

Note that F_s is the matrix factorization loss with $\boldsymbol{X}_s\boldsymbol{X}_s^\top$ as the ground-truth, so the local PL property (cf. Lemma E.4) still holds here, and by the same reason as (43), we deduce that $\operatorname{dist}(\widetilde{\boldsymbol{U}}_0, \boldsymbol{X}_s) \le 0.1\|\boldsymbol{X}\|^{-1}\tau$ i.e. $\widetilde{\boldsymbol{U}}_0$ is in the local PL region around $\boldsymbol{X}_s$. Finally, it follows from the PL property that

$$\operatorname{dist}(\widetilde{\boldsymbol{U}}_0, \boldsymbol{X}_s) \le 10\tau_s^{-1}\left\|\nabla F_s(\widetilde{\boldsymbol{U}}_0)\right\|_F \le 10\tau_s^{-1}\|\widetilde{\boldsymbol{U}}_0\|\left\|\boldsymbol{X}_s\boldsymbol{X}_s^\top - \widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{U}}_0^\top\right\|_F \le 40\kappa_s^3 r_*^{\frac{1}{2}}\|\boldsymbol{X}\|\delta.$$

The conclusion follows. □

Recall that matrix sensing loss satisfies a local PL property (cf. Lemma E.6). As a result, when δ is small, we can show that GD initialized at $\widetilde{\boldsymbol{U}}_0$ converges linearly to the ground-truth.

Lemma F.1 *Let $\widetilde{\boldsymbol{U}}_t$ be the t-th iteration of GD initialized at $\widetilde{\boldsymbol{U}}_0$. Suppose that $\delta \le 10^{-2}r_*^{-\frac{1}{2}}\kappa_s^{-3}$ and $\mu \le 10^{-3}\|\boldsymbol{X}\|^{-2}$, then we have that*

$$\operatorname{dist}^2(\widetilde{\boldsymbol{U}}_t, \boldsymbol{U}_s^*) \le (1 - 0.05\tau_s\mu)^t \operatorname{dist}^2(\widetilde{\boldsymbol{U}}_0, \boldsymbol{U}_s^*).$$

Proof **:** We know from Corollary F.1 that

$$\left\|\widetilde{\boldsymbol{U}}_0\right\|_F \le \|\boldsymbol{X}\|_F + 40\kappa_s^3 r_*^{\frac{1}{2}}\|\boldsymbol{X}\|\delta.$$

We will prove the following result, which immediately implies Lemma F.1: suppose that $\operatorname{dist}(\widetilde{\boldsymbol{U}}_t, \boldsymbol{U}_s^*) \le \operatorname{dist}(\widetilde{\boldsymbol{U}}_0, \boldsymbol{U}_s^*)$, then

$$\operatorname{dist}^2(\widetilde{\boldsymbol{U}}_{t+1}, \boldsymbol{U}_s^*) \le (1 - 0.05\tau_s\mu)\operatorname{dist}^2(\widetilde{\boldsymbol{U}}_t, \boldsymbol{U}_s^*). \tag{45}$$

Let $\boldsymbol{R}$ be the orthogonal matrix satisfying $\|\boldsymbol{U}_t - \boldsymbol{U}_s^*\boldsymbol{R}\|_F = \operatorname{dist}(\boldsymbol{U}_t, \boldsymbol{U}_s^*)$. Assume WLOG that $\boldsymbol{R} = \boldsymbol{I}$. We first bound the gradient $\nabla f(\widetilde{\boldsymbol{U}}_t)$ as follows:

$$\begin{aligned}
\left\|\nabla f(\widetilde{\boldsymbol{U}}_t)\right\|_F &= \left\|\mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^\top\right)\widetilde{\boldsymbol{U}}_t\right\|_F \\
&\le \left\|\mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}\boldsymbol{X}^\top - \widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^\top\right)\right\|\left\|\widetilde{\boldsymbol{U}}_t - \boldsymbol{U}_s^*\right\|_F + \left\|\left(\widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right)\boldsymbol{U}_s^*\right\|_F \\
&\le 20\|\boldsymbol{X}\|^2\left\|\widetilde{\boldsymbol{U}}_t - \boldsymbol{U}_s^*\right\|_F
\end{aligned} \tag{46}$$

where we use $\left\|\widetilde{\boldsymbol{U}}_t\right\| \leq \|\boldsymbol{X}\| + 40\kappa_s^3 r_*^{\frac{1}{2}}\|\boldsymbol{X}\|\delta \leq 2\|\boldsymbol{X}\|$ and the RIP property. It follows that

$$\begin{aligned}
\operatorname{dist}^2(\widetilde{\boldsymbol{U}}_{t+1}, \boldsymbol{U}_s^*) &\leq \left\|\widetilde{\boldsymbol{U}}_{t+1} - \boldsymbol{U}_s^* R\right\|_F^2 && \text{(47a)}\\
&= \left\|\widetilde{\boldsymbol{U}}_t - \mu\nabla f(\widetilde{\boldsymbol{U}}_t) - \boldsymbol{U}_s^* \boldsymbol{R}\right\|_F^2 \\
&= \left\|\widetilde{\boldsymbol{U}}_t - \boldsymbol{U}_s^* R\right\|_F^2 - \mu\left\langle \nabla f(\widetilde{\boldsymbol{U}}_t), \widetilde{\boldsymbol{U}}_t - \boldsymbol{U}_s^* \boldsymbol{R}\right\rangle + \mu^2\left\|\nabla f(\widetilde{\boldsymbol{U}}_t)\right\|_F^2 \\
&\leq \left(1 - 0.1\tau_s\mu + 400\|\boldsymbol{X}\|^4\mu^2\right)\left\|\widetilde{\boldsymbol{U}}_t - \boldsymbol{U}_s^* \boldsymbol{R}\right\|_F^2 && \text{(47b)}
\end{aligned}$$

where (47a) follows from the definition of dist, and (47b) is due to Lemma E.6 and (46). Finally, (45) follows from the condition $\mu \leq 10^{-3}\kappa_s^{-2}$. □

We are now ready to state and prove our main theorem.

Theorem F.1 *Suppose that $\delta \leq 10^{-7} r_*^{-\frac{1}{2}}\kappa_s^{-3}$ and $\mu \leq 10^{-3}\|\boldsymbol{X}\|^{-2}$, then there exists $\epsilon_\alpha = o(1)$ ($\alpha \to 0$) such that*

$$\min_{t\geq 1}\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{Z}_s\right\|_F \leq \epsilon_\alpha$$

for all sufficiently small α, where $\{\boldsymbol{U}_t\}$ is obtained by GD initialized at $\boldsymbol{U}_0 = \alpha\boldsymbol{U}$ and we recall that $\boldsymbol{Z}_s$ is the rank-s minimizer of the matrix sensing loss defined in (44).

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

By Lemma F.1, we have that $\operatorname{dist}^2(\widetilde{\boldsymbol{U}}_t, \boldsymbol{U}_s^*) \leq (1 - 0.05\tau_s\mu)^t \operatorname{dist}^2(\widetilde{\boldsymbol{U}}_0, \boldsymbol{U}_s^*)$ and, in particular, $\left\|\widetilde{\boldsymbol{U}}_t\right\| \leq 2\|\boldsymbol{X}\|$ for all t. Thus $\left\|\bar{\boldsymbol{U}}_t\right\| \leq 2\|\boldsymbol{X}\|$ as well. Moreover, recall that $\|\boldsymbol{U}_t\| \leq 3\|\boldsymbol{X}\|$ for all t. It's easy to see that that the matrix sensing loss f is L-smooth in $\left\{\boldsymbol{U} \in \mathbb{R}^{d\times r} : \|\boldsymbol{X}\| \leq 3\|\boldsymbol{X}\|\right\}$ for some constant L, so it follows from Lemma A.7 that

$$\left\|\boldsymbol{U}_{T_s+t} - \bar{\boldsymbol{U}}_t\right\|_F \leq (1+\mu L)^t\left\|\boldsymbol{U}_{T_s} - \bar{\boldsymbol{U}}_0\right\|_F.$$

On the other hand, since $\operatorname{dist}^2(\widetilde{\boldsymbol{U}}_t, \boldsymbol{U}_s^*) \leq (1 - 0.05\tau\mu)^t \operatorname{dist}^2(\widetilde{\boldsymbol{U}}_0, \boldsymbol{U}_s^*)$, we can deduce that

$$\begin{aligned}
\left\|\boldsymbol{U}_{T_s+t}\boldsymbol{U}_{T_s+t}^\top - \boldsymbol{Z}_s\right\|_F &\leq \left\|\boldsymbol{U}_{T_s+t}\boldsymbol{U}_{T_s+t}^\top - \bar{\boldsymbol{U}}_t\bar{\boldsymbol{U}}_t^\top\right\|_F + \left\|\bar{\boldsymbol{U}}_t\bar{\boldsymbol{U}}_t^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right\|_F \\
&= \left\|\boldsymbol{U}_{T_s+t}\boldsymbol{U}_{T_s+t}^\top - \bar{\boldsymbol{U}}_t\bar{\boldsymbol{U}}_t^\top\right\|_F + \left\|\widetilde{\boldsymbol{U}}_t\widetilde{\boldsymbol{U}}_t^\top - \boldsymbol{U}_s^*\left(\boldsymbol{U}_s^*\right)^\top\right\|_F \\
&\leq 3\|\boldsymbol{X}\|\left(\left\|\boldsymbol{U}_{T_s+t} - \bar{\boldsymbol{U}}_t\right\|_F + \operatorname{dist}(\widetilde{\boldsymbol{U}}_t, \boldsymbol{U}_s^*)\right) \\
&\leq 3\|\boldsymbol{X}\|\left((1+\mu L)^t\left\|\boldsymbol{U}_{T_s} - \bar{\boldsymbol{U}}_0\right\|_F + (1 - 0.05\tau_s\mu)^{\frac{t}{2}}\operatorname{dist}^2(\widetilde{\boldsymbol{U}}_0, \boldsymbol{U}_s^*)\right)
\end{aligned}$$

Since when $\alpha \to 0$, $\left\|\boldsymbol{U}_{T_s} - \bar{\boldsymbol{U}}_0\right\|_F = o(1)$, it's easy to see that we have $\min_{t\geq 0}\left\|\boldsymbol{U}_{T_s+t}\boldsymbol{U}_{T_s+t}^\top - \boldsymbol{Z}_s\right\|_F = o(1)$ as well, which concludes the proof. □

G THE LANDSCAPE OF MATRIX SENSING WITH RANK-1 PARAMETERIZATION

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 $g(\boldsymbol{u}) - g(\boldsymbol{u}^*) \leq 0.5\tau_1$ where $\tau_1 = \sigma_1^2 - \sigma_2^2$ is the eigengap, then we must have

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

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

$$g_1(\boldsymbol{u}) = \frac{1}{4}\left(\|\boldsymbol{u}\|_2^4 - 2\sum_{i=1}^{s}\sigma_i^2\boldsymbol{u}_i^2 + \|\boldsymbol{X}^T\boldsymbol{X}\|_F^2\right) \tag{48a}$$

$$\geq \frac{1}{4}\left(\|\boldsymbol{u}\|_2^4 - 2\sigma_1^2\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{X}^T\boldsymbol{X}\|_F^2\right) \tag{48b}$$

$$\geq \frac{1}{4}\left(\|\boldsymbol{X}^T\boldsymbol{X}\|_F^2 - \sigma_1^4\right) \tag{48c}$$

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 $g_1(\boldsymbol{u}) - g_1(\boldsymbol{u}^*) \leq 0.5\tau_1$, it follows from (48b) that $\tau_1\sum_{i=2}^d \boldsymbol{u}_i^2 \leq 2(g_1(\boldsymbol{u}) - g_1(\boldsymbol{u}^*))$ which implies that $\sum_{i=2}^d \boldsymbol{u}_i^2 \leq 2\tau_1^{-1}\left(g_1(\boldsymbol{u}) - g_1(\boldsymbol{u}^*)\right)$. Also (48c) yields $\left|\|\boldsymbol{u}\|^2 - \sigma_1^2\right| \leq 4\sqrt{g_1(\boldsymbol{u}) - g_1(\boldsymbol{u}^*)}$. Assume WLOG that $\boldsymbol{u}_1 > 0$, then we have

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

□

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 $\delta \leq 10^{-3}\|\boldsymbol{X}\|^{-2}\tau_1$, 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

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

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

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

Hence we have

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

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 $\|\boldsymbol{u}^*\| \leq 2\|\boldsymbol{X}\|$ (otherwise $f_1(\boldsymbol{u}) > f_1(0)$). We can thus deduce that

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

It follows from Lemma G.1 and our assumption on δ that $\min\left\{\|\boldsymbol{u}^* - \sigma_1 v_1\|^2, \|\boldsymbol{u}^* + \sigma_1 v_1\|^2\right\} \leq \frac{1}{2}a_1^2$. Moreover, by strong convexity, there exists only one global minima in $\mathcal{B}_1$, which concludes the proof. □