Convergence of Gradient Descent with Small Initialization for Unregularized Matrix Completion

Jianhao Ma JIANHAO@UMICH.EDU and **Salar Fattahi** FATTAHI@UMICH.EDU
University of Michigan, Ann Arbor

Editors: Shipra Agrawal and Aaron Roth

Abstract

We study the problem of symmetric matrix completion, where the goal is to reconstruct a positive semidefinite matrix $\boldsymbol{X}^\star \in \mathbb{R}^{d\times d}$ of rank-r, parameterized by $\boldsymbol{U}\boldsymbol{U}^\top$, from only a subset of its observed entries. We show that the vanilla gradient descent (GD) with small initialization provably converges to the ground truth $\boldsymbol{X}^\star$ without requiring any explicit regularization. This convergence result holds true even in the over-parameterized scenario, where the true rank r is unknown and conservatively over-estimated by a search rank $r' \gg r$. The existing results for this problem either require explicit regularization, a sufficiently accurate initial point, or exact knowledge of the true rank r.

In the over-parameterized regime where $r' \geq r$, we show that, with $\widetilde{\Omega}(dr^9)$ observations, GD with an initial point $\|\boldsymbol{U}_0\| \leq O(\epsilon)$ converges near-linearly to an ϵ-neighborhood of $\boldsymbol{X}^\star$. Consequently, smaller initial points result in increasingly accurate solutions. Surprisingly, neither the convergence rate nor the final accuracy depends on the over-parameterized search rank r', and they are only governed by the true rank r. In the exactly-parameterized regime where $r' = r$, we further enhance this result by proving that GD converges at a faster rate to achieve an arbitrarily small accuracy $\epsilon > 0$, provided the initial point satisfies $\|\boldsymbol{U}_0\| = O(1/d)$. At the crux of our method lies a novel *weakly-coupled leave-one-out analysis*, which allows us to establish the global convergence of GD, extending beyond what was previously possible using the classical leave-one-out analysis.

Keywords: Matrix completion, implicit regularization, leave-one-out analysis

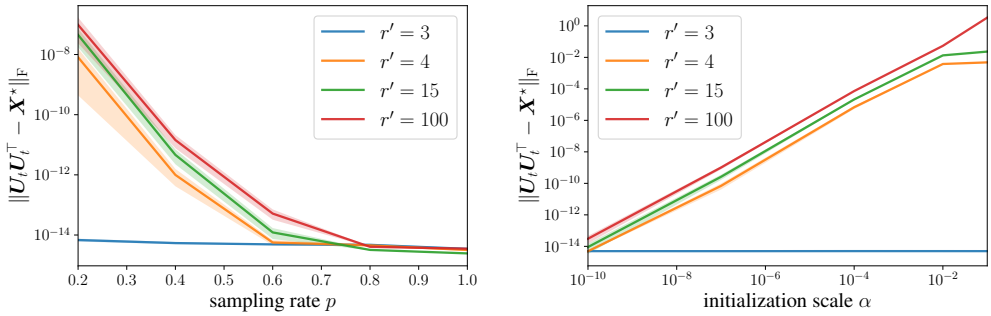
1. Introduction

Matrix completion is a fundamental problem in the field of machine learning, where the objective is to reconstruct a positive semidefinite (PSD) matrix of rank-r, denoted as $\boldsymbol{X}^\star \in \mathbb{R}^{d\times d}$, from only a subset of its observed entries. The most natural approach to solve this problem involves minimizing the following mean squared error:

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

Here, p represents the probability of observing each entry in $\boldsymbol{X}^\star$, Ω denotes the set of observed entries, and $\mathcal{P}_\Omega$ shows the projection operation onto the set of matrices supported by Ω. When the true rank r is unknown, it is often over-estimated by the search rank $r' \geq r$, leading to what is referred to as *over-parameterized* matrix completion.

A prominent application of matrix completion is in collaborative filtering (Gleich and Lim, 2011). Additionally, it has applications in other areas, including image reconstruction (Hu et al., 2018), fast kernel matrix approximation (Graepel, 2002; Paisley and Carin, 2010), and more recently, in teaching arithmetic to transformers (Lee et al., 2023).



(a) (b)

Figure 1: The performance of vanilla GD with small initialization and constant step-size on MC without any explicit regularization. In all experiments, we set the dimension $d = 100$ and the true rank $r = 3$. (a): Higher sampling rates improve the final accuracy of GD for the over-parameterized MC, but have no impact on the exactly-parameterized MC. (b): Increasing the initialization scale hampers the final accuracy of GD for the over-parameterized MC, but has no impact on the exactly-parameterized MC.

Perhaps one of the most natural approaches for solving the above optimization problem is (vanilla) gradient descent (GD): given an initial point $\boldsymbol{U}_0$ and a fixed step-size $\eta > 0$, generate a sequence of iterates $\{\boldsymbol{U}_t\}_{t=1}^T$ according to $\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta \nabla f(\boldsymbol{U}_t)$. Despite its simplicity and desirable practical performance (see Figure 1 and the experiments in (Zheng and Lafferty, 2016)), the conditions under which the GD converges globally to the ground truth $\boldsymbol{X}^\star$ have remained a long-standing mystery.

A line of research has been devoted to studying gradient-based algorithms with *explicit regularization* (Sun and Luo, 2016; Zheng and Lafferty, 2016; Jain et al., 2013; Ge et al., 2016). These methods typically incorporate either an $\ell_{2,\infty}$-norm regularizer or a projection step to constrain the iterates within a set with $\ell_{2,\infty}$-norm bounds to promote *incoherence* (see Definition 2). However, the use of $\ell_{2,\infty}$-norm regularization or projection techniques often introduce more tuning parameters, and has been found to be unnecessary in most cases (Zheng and Lafferty, 2016; Ma et al., 2018)[1].

On the other hand, the convergence of GD without explicit regularization was initially tackled by Ma et al. (2018) in the context of symmetric matrix completion, and subsequently extended by Chen et al. (2020) to asymmetric settings. However, these studies consider a very special case of matrix completion where GD is initialized sufficiently close to the ground truth, and the rank of the ground truth r is known. In practice, however, GD converges even if it is initialized far from the ground truth and the rank is over-parameterized $r' \gg r$ (see Figure 1).

Therefore, the following question still remains open:

> *Why does GD with a small initialization efficiently converge to the ground truth of MC in the absence of explicit regularization, even in the general rank-r case where $r' \geq r \geq 1$?*

Recently, Kim and Chung (2022) answered the above question for the special case of rank-1 symmetric matrix completion with $r' = r = 1$, but their proposed approach does not extend to the general rank-r case. It is also worth noting that the above question has been addressed for another class of matrix factorization problems satisfying a norm-preserving property called *restricted isometry property* (RIP). Problems that satisfy this property include matrix sensing (Li et al., 2018;

1. It is worth noting that explicit regularization can still offer significant advantages in scenarios with extremely limited sample sizes (Sun, 2015).

Stöger and Soltanolkotabi, 2021; Ma and Fattahi, 2023a) and sparse recovery (Vaskevicius et al., 2019). However, a significant challenge arises with matrix completion, as it *does not* satisfy the restricted isometry property. Consequently, the existing methodologies built upon RIP are not directly applicable to matrix completion.

1.1. Summary of Contributions

In this work, we provide a complete answer to the aforementioned question. A comparison of our results with other studies on unregularized matrix completion can be found in Table 1. The key contributions of our work are as follows:

- **Convergence of GD with small initialization in over-parameterized regime:** When the rank of the ground truth r is unknown and over-parameterized by $r' \geq r \geq 1$, we prove that GD with small initialization converges to the ground truth at a near-linear rate. Surprisingly, neither the convergence rate nor the final accuracy depends on the over-parameterized search rank r', and they are only governed by the true rank r. In particular, given an initial point that satisfies $\|\boldsymbol{U}_0\| \leq O(\epsilon)$ for some $\epsilon > 0$ and a sampling rate of $p = \widetilde{\Omega}(r^9 \log^6(1/\epsilon)/d)$, GD converges to ϵ-neighborhood of $\boldsymbol{X}^\star$ in $O(\log^4(1/\epsilon))$ iterations. Therefore, a smaller initial point or a larger sampling rate can improve the final error of GD. The empirical observation presented in Figure 1 provides further support for this result.

- **Improved results in the exactly-parameterized regime:** We show that GD enjoys an improved convergence when the rank of the ground truth r is known and $r' = r \geq 1$. In particular, given an initial point that satisfies $\|\boldsymbol{U}_0\| \leq O(1/d)$ and a sampling rate of $p = \widetilde{\Omega}(r^9/d)$, GD converges to ϵ-neighborhood of $\boldsymbol{X}^\star$ in $O(\log(1/\epsilon))$ iterations for any arbitrarily small $\epsilon > 0$. A key distinction from the over-parameterized setting is that the final error of GD remains unaffected by the initialization scale or the sample size, provided that they meet certain thresholds. This is also evident in Figure 1. When $r = O(1)$, the resulting sample complexity is information-theoretically optimal (modulo logarithmic factors).

- **Weakly-coupled leave-one-out analysis:** In order to establish the implicit regularization of GD for MC, a pivotal technique is a decoupling mechanism known as leave-one-out analysis, a trick rooted in probability and random matrix theory. However, the current theory based on this technique is only effective when the iterates are sufficiently close to the ground truth. At the crux of our technical analysis lies an extension of the classical leave-one-out analysis to the global setting, which we term *weakly-coupled leave-one-out analysis*. In essence, our proposed method relaxes the requirement for the initial iterates to be sufficiently close to the ground truth, making it particularly suitable for the global convergence analysis of GD.

Notations. We use bold uppercase letters $\boldsymbol{X}, \boldsymbol{Y}$ to denote matrices and bold lowercase letters $\boldsymbol{x}, \boldsymbol{y}$ to denote vectors. For vectors, we use $\|\cdot\|$ to denote ℓ_2-norm, and for matrices we use $\|\cdot\|$ and $\|\cdot\|_{\mathrm{F}}$ to denote operator norm and Frobenius norm, respectively. For matrix $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$, we denote by $X_{i,j}$ the (i,j)-th element of $\boldsymbol{X}$, $\boldsymbol{X}_{i,\cdot}$ the i-th row, and $\boldsymbol{X}_{\cdot,j}$ the j-th column. The $\ell_{2,\infty}$-norm of $\boldsymbol{X}$, denoted as $\|\boldsymbol{X}\|_{2,\infty}$, is defined as $\max_i \|\boldsymbol{X}_{i,\cdot}\|$. Additionally, we define the singular values of $\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}$ as $\sigma_1(\boldsymbol{X}) \geq \sigma_2(\boldsymbol{X}) \geq \cdots \geq \sigma_{\min\{d_1,d_2\}}(\boldsymbol{X}) \geq 0$. The set of all orthogonal matrices is denoted by $\mathcal{O}_{d_1 \times d_2} := \{\boldsymbol{O} \in \mathbb{R}^{d_1 \times d_2} : \boldsymbol{O}^\top \boldsymbol{O} = \boldsymbol{I}\}$. For two matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d_1 \times d_2}$, we define

Algorithm	Sample complexity	Computational complexity	Global	Exact	Over-param.
(Ma et al., 2018)	$dr^3\log^3(d)$	$\kappa^2\log\left(\frac{1}{\epsilon}\right)$	✘	✔	✘
(Chen et al., 2020)	$dr^2\log(d)$	$\kappa^2\log\left(\frac{1}{\epsilon}\right)$	✘	✔	✘
(Kim and Chung, 2022)	$d\log^{22}(d)$	$\log\left(\frac{1}{\epsilon}\right)$	✔	✱	✘
Ours (Theorem 3)	$dr^9\log^8(d)$	$\kappa^4\log\left(\frac{1}{\epsilon}\right)$	✔	✔	✘
Ours (Theorem 2)	$dr^9\log^2(d)\log^6\left(\frac{1}{\epsilon}\right)$	$\kappa^4\log^4\left(\frac{1}{\epsilon}\right)$	✔	✔	✔

Table 1: Comparisons between different algorithms for matrix completion without explicit regularization. ✱ The result only holds for $r' = r = 1$.

their Procrustes distance as $\mathrm{dist}(\boldsymbol{X},\boldsymbol{Y}) = \min_{\boldsymbol{O}\in\mathcal{O}_{d_2\times d_2}} \|\boldsymbol{X}-\boldsymbol{Y}\boldsymbol{O}\|_{\mathrm{F}}$. The projection matrix onto the column space of an orthogonal matrix $\boldsymbol{V}\in\mathcal{O}_{d_1\times d_2}$ is defined as $\mathcal{P}_{\boldsymbol{V}} := \boldsymbol{V}\boldsymbol{V}^\top$.

We use the notation $f(n)\lesssim g(n)$ or $f(n)=O(g(n))$ when a constant $C>0$ exists such that $f(n)\le Cg(n)$ for sufficiently large n. Conversely, $f(n)\gtrsim g(n)$ or $f(n)=\Omega(g(n))$ implies the existence of a constant $C>0$ such that $f(n)\ge Cg(n)$ for sufficiently large n. Moreover, we use the notations $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to hide logarithmic dependencies on the dimension or other parameters of the problem. Additionally, we use $f(n)\asymp g(n)$ or $f(n)=\Theta(g(n))$ when $f(n)\lesssim g(n)$ and $f(n)\gtrsim g(n)$.

2. Problem Setup and Main Results

Suppose that the singular value decomposition (SVD) of $\boldsymbol{X}^\star$ is given by $\boldsymbol{X}^\star = \boldsymbol{V}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$, where $\boldsymbol{V}^\star\in\mathcal{O}_{d\times r}$ and $\boldsymbol{\Sigma}^\star$ is an $r\times r$ diagonal matrix with diagonal elements in descending order $\sigma_1^\star\ge\cdots\ge\sigma_r^\star>0$. We denote the condition number of $\boldsymbol{X}^\star$ as $\kappa=\sigma_1^\star/\sigma_r^\star$. Upon defining the symmetrized operator $\mathcal{R}_\Omega = \frac{1}{2p}(\mathcal{P}_\Omega+\mathcal{P}_\Omega^\top)$, the update rule for GD can be written as

$$\boldsymbol{U}_{t+1} = \boldsymbol{U}_t - \eta\nabla f(\boldsymbol{U}_t) = \boldsymbol{U}_t - \eta\mathcal{R}_\Omega(\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star)\boldsymbol{U}_t, \qquad \boldsymbol{U}_0 = \alpha\boldsymbol{Z} \tag{GD}$$

where $\alpha>0$ is the initialization scale and $\boldsymbol{Z}$ is the initialization matrix satisfying $\|\boldsymbol{Z}\|=1$. We assume that $\boldsymbol{Z}$ satisfies the following *alignment condition*.

Condition 1 (Alignment) *We say the matrix $\boldsymbol{Z}\in\mathbb{R}^{d\times r'}$ with $\|\boldsymbol{Z}\|=1$ satisfies the alignment condition if there exists a universal constant $c_0>0$ such that*

$$\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{Z}) \ge c_0. \tag{alignment condition}$$

Intuitively, this condition necessitates that the initialization matrix should have a non-negligible alignment with the column space of the ground truth. The following lemma reveals that this alignment condition is satisfied for common initialization strategies with overwhelming probability.

Lemma 1 (Sufficient condition for alignment) *The following statements are satisfied:*

- ***Gaussian initialization.*** *Given $0.5d\le r'\le d$ and $\boldsymbol{Z}=\boldsymbol{G}/\|\boldsymbol{G}\|$, where $\boldsymbol{G}$ is a standard Gaussian matrix, alignment condition holds with $c_0=0.1$ with probability at least $1-\exp\{-\Omega(d)\}$.*

- ***Orthogonal initialization.*** *Given $r' = d$ and $\boldsymbol{Z} = \boldsymbol{O}$ for any $\boldsymbol{O} \in \mathcal{O}_{d\times d}$, alignment condition is satisfied with $c_0 = 1$.*
- ***Spectral initialization.*** *Let $\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the eigendecomposition of the best rank-r' approximation of $\mathcal{R}_\Omega(\boldsymbol{X}^\star)$ measured in Frobenius norm. Given $r \le r' \le d$ and $\boldsymbol{Z} = \boldsymbol{U}/\|\boldsymbol{U}\|$, where $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$, alignment condition holds with $c_0 = \frac{1}{2\kappa}$ with probability at least $1 - \frac{1}{d^3}$.*

Before proceeding to the main theorem, we introduce two crucial conditions on the ground truth $\boldsymbol{X}^\star$ and the random observation set Ω.

Condition 2 (Incoherence) *The rank-r PSD matrix $\boldsymbol{X}^\star \in \mathbb{R}^{d\times d}$ with SVD $\boldsymbol{X}^\star = \boldsymbol{V}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}$ is μ-incoherent for some $\mu \ge 1$ if $\|\boldsymbol{V}^\star\|_{2,\infty} = \sqrt{\frac{\mu}{d}}\|\boldsymbol{V}^\star\|_{\mathrm{F}} = \sqrt{\frac{\mu r}{d}}$.*

Condition 3 (Random sampling model) *Each entry of $\boldsymbol{X}^\star$ is observed independently with probability p. In other words, $\mathbb{P}((i,j) \in \Omega) = p$ independently for all $1 \le i, j \le d$.*

Chen (2015) shows that the incoherence condition is necessary for the recovery of the ground truth. Intuitively, the incoherence condition with $\mu = O(1)$ entails that none of the columns of $\boldsymbol{V}^\star$ have significant alignment with the standard basis vectors. Such a condition ensures that the ground truth is far from being sparse. We note that successful recovery of a sparse $\boldsymbol{X}^\star$ is only achievable in a near-ideal scenario where $p \approx 1$ and nearly the entirety of $\boldsymbol{X}^\star$ is observed. On the other hand, when $\mu = O(1)$, the recovery is possible even when p scales with $\widetilde{\Omega}(\mathrm{poly}(r)/d)$ (Chen, 2015).

With the aforementioned conditions in place, we can now present our main result, which establishes the global convergence of GD on MC for the over-parameterized setting.

Theorem 2 (Convergence of GD for over-parameterized MC) *Let $\boldsymbol{X}^\star$ be rank-r, and Conditions 2 and 3 are satisfied with a sampling rate of $p \gtrsim \frac{\kappa^6\mu^4 r^9 \log^6(\frac{1}{\alpha})\log^2(d)}{d}$. Consider MC with search rank $r \le r' \le d$. Consider the iterates of GD with the step-size $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$ and the initial point $\boldsymbol{U}_0 = \alpha\boldsymbol{Z}$, where $0 < \alpha \le \sqrt{\frac{\sigma_1^\star}{d}}$ and $\boldsymbol{Z}$ satisfies Condition 1. With probability at least $1 - \frac{2}{d^3}$, after $T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$ iterations, we have*

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

A few observations are in order based on the above theorem.

Computational complexity. The initialization scale α governs the final accuracy of GD. Therefore, to ensure that $\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \epsilon$, it suffices to set the initialization scale to $\alpha \lesssim \sqrt{\frac{p}{\sigma_1^\star\kappa^2\mu r^2}}\epsilon$. Moreover, assuming that $\max\{\kappa, \mu, r\} = O(1)$ and $\epsilon \le 1/d$, this accuracy is achieved within $\widetilde{O}\left(\log^4\left(\frac{1}{\epsilon}\right)\right)$ iterations, which scales only poly-logarithmically with $1/\epsilon$.

Effect of over-parameterization. The level of over-parameterization in the search rank r' does not have any impact on either the sample complexity or the convergence of GD. As a result, our results hold even if $d^2p \ll dr'$. In such cases, MC has many global minima, some of which may not satisfy $\boldsymbol{U}\boldsymbol{U}^\top \approx \boldsymbol{X}^\star$. This sheds light on the implicit regularization of the vanilla GD with small initialization toward low-rank solutions when applied to MC.

Sample complexity. The required sample complexity is given by $d^2p \gtrsim dr^9\kappa^6\mu^4 \log^6\left(\frac{1}{\alpha}\right)\log^2(d)$, which is optimal with respect to the dimension d up to a logarithmic factor. This contrasts with the direct extension of the approach by Kim and Chung (2022), which necessitates a sample size on the order of $d^{1+\Theta(\kappa-1)}$. Moreover, Theorem 2 highlights that the sampling rate p must scale with $\log^6\left(\frac{1}{\epsilon}\right)$ to attain an accuracy level of ϵ, which in turn leads to a mild dependency of the final error on the sampling rate. In other words, given a fixed sampling rate p, GD achieves an accuracy in the order of $\exp(-\Omega(pd))$. We suspect that the observed dependency might be an artifact of our proof technique and could potentially be relaxed with a more detailed analysis, similar to that presented in (Stöger and Soltanolkotabi, 2021). In our next theorem, we demonstrate that in the exact-parameterization regime, it is possible to relax this mild dependency, achieving a sample complexity that does not depend on the desired accuracy level or the initialization scale.

Theorem 3 (Convergence of GD for exactly-parameterized MC) *Let $\boldsymbol{X}^\star$ be rank-r, and Conditions 2 and 3 are satisfied with a sampling rate of $p \gtrsim \frac{\kappa^6\mu^4 r^9 \log^8(d)}{d}$. Consider MC with search rank $r' = r$. Consider the iterates of GD with the step-size $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$ and the initial point $\boldsymbol{U}_0 = \alpha\boldsymbol{Z}$, where $\alpha \asymp \frac{\sigma_r^\star}{\kappa^{1.5}d}$ and $\boldsymbol{Z}$ satisfies Condition 1. Given any accuracy $\epsilon > 0$, with probability at least $1 - O\left(\frac{1}{d^3}\right)$ and after $T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\epsilon}\right)$ iterations, we have*

$$\left\|\boldsymbol{U}_T\boldsymbol{U}_T^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq \epsilon. \tag{1}$$

We next outline the key distinctions between the two aforementioned theorems. In the exactly-parameterized regime, neither the sampling rate p nor the initialization scale α affect the final error ϵ, as long as they meet certain thresholds. In contrast, a smaller initialization scale or a larger sampling rate improves the final error in the over-parameterized regime. Moreover, the convergence rate of GD improves from $O\left(\log^4\left(\frac{1}{\epsilon}\right)\right)$ to $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$. This is because the required sampling rate is smaller in the exactly-parameterized regime, allowing the algorithm to adopt a more aggressive step-size.

3. Proof Outline

In this section, we present the key ideas underpinning our proof techniques. We begin in Section 3.1 with a dynamic signal-residual decomposition. Next, in Section 3.2, we introduce the weakly-coupled leave-one-out analysis which, together with our dynamic signal-residual decomposition, completes the proof for Theorem 2. Section 3.3 explains how these techniques can be further refined to yield improved results for the exactly-parameterized regime. Throughout this section, we occasionally omit the consideration of higher-order terms involving the step-size η. We highlight that while this omission serves to streamline the presentation, our rigorous proofs in the appendix carefully account for these higher-order terms.

3.1. Dynamic Signal-residual Decomposition

We employ a dynamic projection scheme akin to that described by Li et al. (2018), which decomposes the iterates $\boldsymbol{U}_t$ into two distinct components: a low-rank signal part, $\boldsymbol{S}_t$, and a residual part, $\boldsymbol{E}_t$. This decomposition is represented as follows:

$$\boldsymbol{U}_t = \boldsymbol{S}_t + \boldsymbol{E}_t, \quad \text{where} \quad \boldsymbol{S}_t = \mathcal{P}_{\boldsymbol{V}_t}\boldsymbol{U}_t, \text{ and } \boldsymbol{E}_t = \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{U}_t.$$

Upon defining $\boldsymbol{M}_t = \mathcal{R}_\Omega(\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top)$, the dynamic orthonormal matrix $\boldsymbol{V}_t$ is recursively defined as:

$$\boldsymbol{V}_{t+1} = \boldsymbol{Z}_{t+1}\left(\boldsymbol{Z}_{t+1}^\top \boldsymbol{Z}_{t+1}\right)^{-1/2} \quad \text{where} \quad \boldsymbol{Z}_{t+1} = (\boldsymbol{I} + \eta \boldsymbol{M}_t)\,\boldsymbol{V}_t \text{ and } \boldsymbol{V}_0 = \boldsymbol{V}^\star.$$

Upon defining the error matrix as $\boldsymbol{\Delta}_t := \boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top$, our goal is to show that $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$ decreases efficiently to $O(\alpha)$. To show the benefit of the proposed dynamic signal-residual decomposition in achieving this goal, we start by stating the one-step dynamic of the error matrix:

$$\|\boldsymbol{\Delta}_{t+1}\|_{\mathrm{F}}^2 = \|\boldsymbol{\Delta}_{t+1}\|_{\mathrm{F}}^2 - 4\eta\left\langle \boldsymbol{\Delta}_t, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top \right\rangle + O(\eta^2). \tag{2}$$

Therefore, to establish the convergence of $\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$, it suffices to provide a reasonable lower-bound for $\left\langle \boldsymbol{\Delta}_t, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top \right\rangle$. This can be achieved via the following descent lemma:

Lemma 4 (Descent lemma, informal) *Suppose that* $\frac{\sqrt{\sigma_r^\star}}{2} \le \sigma_r(\boldsymbol{S}_t)$, $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$, *and* $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\| \le 0.1$. *Then, we have*

$$\left\langle \boldsymbol{\Delta}_t, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top \right\rangle \ge \frac{\sigma_r^\star}{15}\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}^2 - O\left(\sqrt{\frac{\sigma_r^{\star 3}\mu r^2}{p}}\|\boldsymbol{E}_t\| + \sqrt{r}\sigma_1^\star\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right)\|\boldsymbol{\Delta}_t\|_{\mathrm{F}}.$$

By combining the descent lemma with Equation (2), we arrive at the following expression:

$$\|\boldsymbol{\Delta}_{t+1}\|_{\mathrm{F}} \le \left(1 - \frac{\eta\sigma_r^\star}{10}\right)\|\boldsymbol{\Delta}_t\|_{\mathrm{F}} + O(\eta)\left(\sqrt{\frac{\sigma_r^{\star 3}\mu r^2}{p}}\|\boldsymbol{E}_t\| + \sqrt{r}\sigma_1^\star\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right) + O(\eta^2). \tag{3}$$

The above inequality holds once the conditions $\frac{\sqrt{\sigma_r^\star}}{2} \le \sigma_r(\boldsymbol{S}_t) \le \|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\| \le 0.1$ are met. These conditions entail that during the initial phase of the algorithm, $\sigma_r(\boldsymbol{S}_t)$ must undergo a fast growth, whereas $\boldsymbol{V}_t$ must remain close to $\boldsymbol{V}^\star$. Under these conditions, GD enters a fast linear convergence phase, provided that $\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \ll \|\boldsymbol{\Delta}_t\|_{\mathrm{F}}$. In fact, we can readily establish that $\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \le c\|\boldsymbol{\Delta}_t\|$ for some $c > 0$. However, the challenge lies in ensuring that c remains sufficiently small so as not to negate the effect of a constant factor improvement $1 - 0.1\eta\sigma_r^\star$ in Equation (3). This phase continues until GD reaches an error level controlled by $\|\boldsymbol{E}_t\|$. Therefore, to prove the convergence of GD, we need to establish the following properties:

- **Fast growth of $\boldsymbol{S}_t$:** Recall that $\max\{\|\boldsymbol{S}_0\|, \|\boldsymbol{E}_0\|\} = O(\alpha)$. We need to ensure efficient growth of $\sigma_r(\boldsymbol{S}_t)$ from $O(\alpha)$ to $\frac{\sqrt{\sigma_r^\star}}{2}$, while keeping $\|\boldsymbol{S}_t\|$ below $2\sqrt{\sigma_1^\star}$.
- **Slow growth of $\boldsymbol{E}_t$:** We need to show that while the signal term grows rapidly, the residual term $\boldsymbol{E}_t$ grows at a much slower rate. Specifically, we will demonstrate that $T = O\left(\frac{1}{\eta\sigma_r^\star}\log(\frac{1}{\alpha})\right)$ suffices to ensure $\frac{\sqrt{\sigma_r^\star}}{2} \le \sigma_r(\boldsymbol{S}_t) \le \|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$ while keeping $\|\boldsymbol{E}_t\| = O(\alpha)$.
- **Small values of $\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$ and $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|$:** Equally important is maintaining control over $\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$ and $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|$. While $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\| \le 0.1$ is needed as a crucial condition for Equation (3), the value of $\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$ directly controls the convergence rate of GD.

To establish the above properties, we provide the one-step dynamics of $\boldsymbol{S}_t$, $\boldsymbol{E}_t$, and $(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)$.

Lemma 5 (One-step dynamics, informal) *Under the conditions of Theorem 2, the following inequalities hold with an overwhelming probability:*

$$\|\boldsymbol{S}_{t+1}\| \le \left(1+\eta\left(\sigma_1^\star - \|\boldsymbol{S}_t\|^2 + O\left(\sigma_1^\star \|\boldsymbol{V}_t-\boldsymbol{V}^\star\| + \|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right)\right)\right)\|\boldsymbol{S}_t\|, \tag{4}$$

$$\begin{aligned}\sigma_r(\boldsymbol{S}_{t+1}) \ge &\left(1+\eta\left(\sigma_r^\star - \sigma_r^2(\boldsymbol{S}_t) - O\left(\sigma_1^\star\|\boldsymbol{V}_t-\boldsymbol{V}^\star\| + \|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right)\right)\right)\sigma_r(\boldsymbol{S}_t)\\ &+ O(\eta)\cdot\left(\sigma_1^\star\|\boldsymbol{V}_t-\boldsymbol{V}^\star\| + \|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right)\|\boldsymbol{E}_t\|,\end{aligned} \tag{5}$$

$$\|\boldsymbol{E}_{t+1}\| \le \left(1+O(\eta)\cdot\left(\sigma_1^\star\|\boldsymbol{V}_t-\boldsymbol{V}^\star\| + \|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|\right)\right)\|\boldsymbol{E}_t\|, \tag{6}$$

$$\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \le O\left(\sqrt{\frac{d}{p}}\|\boldsymbol{\Delta}_t\|\left(\|\boldsymbol{V}^\star\|_{2,\infty}^2 + \|\boldsymbol{V}_t\|_{2,\infty}^2\right) + \sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\|\right). \tag{7}$$

Next, we provide a high-level overview of how Lemma 5 can be used to establish the aforementioned properties. To this goal, we only focus on the initial phase of the algorithm, where both $\boldsymbol{S}_t$ and $\boldsymbol{E}_t$ are small. A more formal analysis for the entire trajectory is provided in Appendix C.

To use Lemma 5, it suffices to control two key quantities: $\|\boldsymbol{V}_t\|_{2,\infty}$ and $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$. To illustrate this, let us assume that $\boldsymbol{V}_t$ inherits the incoherence of $\boldsymbol{V}^\star$, that is, $\|\boldsymbol{V}_t\|_{2,\infty} \le O(\sqrt{\mu r/d}) \ll 1$ for all $1 \le t \le T$. Then, Equation (7) suggests that $\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \le c\|\boldsymbol{\Delta}_t\| + O(\alpha)$, where $c = O(\mu r/\sqrt{pd}) \ll 1$, thereby ensuring the necessary control over $\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$.

On the other hand, the small values of $\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$ and $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$ play crucial roles in controlling the behavior of the signal and residual terms. To illustrate this, let us assume that at a certain point, $\|\boldsymbol{S}_t\| \ge 1.5\sqrt{\sigma_1^\star}$. Given that we have considered $\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$ and $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$ to be small, Equation (4) simplifies to $\|\boldsymbol{S}_{t+1}\| \le (1-\eta\Omega(\sigma_1^\star))\|\boldsymbol{S}_t\| + O(\eta\sigma_1^\star\|\boldsymbol{E}_t\|)$, effectively preventing $\|\boldsymbol{S}_{t+1}\|$ from further growth. With a similar reasoning, Equation (5) simplifies to $\sigma_r(\boldsymbol{S}_{t+1}) \ge (1+\eta\Omega(\sigma_r^\star))\sigma_r(\boldsymbol{S}_t)$. Here, we have leveraged the assumption that $\sigma_r^2(\boldsymbol{S}_t) \ll \sigma_r^\star$ during the initial phase. This implies that $\sigma_r(\boldsymbol{S}_t)$ grows at a rate of $1+\eta\Omega(\sigma_r^\star)$. In contrast, Equation (6) implies that $\|\boldsymbol{E}_t\|$ grows at a rate of $1+\eta O(\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| + \sigma_1^\star\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|)$, which is significantly slower than the growth rate of $\sigma_r(\boldsymbol{S}_t)$ because $\max\{\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|, \sigma_1^\star\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|\} \ll \sigma_r^\star$. It is due to this discrepancy in the growth rates of $\|\boldsymbol{S}_t\|$ and $\|\boldsymbol{E}_t\|$ that GD enters the local linear convergence rate and achieves a final error of $O(\alpha)$.

3.2. Refined Leave-one-out Analysis with Weak Coupling

Indeed, it is not immediately evident why both $\|\boldsymbol{V}_t\|_{2,\infty}$ and $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$ would remain small. In fact, our initial intuition might suggest the opposite: recall that GD takes an aggressively large step size. Consequently, even a single GD update has the potential to disrupt the incoherence of $\boldsymbol{V}_t$. Our key contribution is to establish that such disruption does not occur, even when the iterates are arbitrarily far from the ground truth. In essence, we show that despite the gradient update $\|\eta\mathcal{R}_\Omega(\boldsymbol{\Delta}_t)\boldsymbol{U}_t\|$ potentially having a magnitude of $\widetilde{\Omega}(1)$, its impact on $\boldsymbol{V}_t$ is distributed fairly evenly across its elements. As a result, it has minimal influence on $\|\boldsymbol{V}_t\|_{2,\infty}$ and $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$.

We start by showing that a small $\|\boldsymbol{V}_t\|_{2,\infty}$ implies a small $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$.

Lemma 6 (Small $\|\boldsymbol{V}_t\|_{2,\infty}$ implies small $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|_{\mathrm{F}}$, informal) *Suppose that $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\| \le \frac{1}{2\kappa}$. With an overwhelming probability, we have*

$$\|\boldsymbol{V}_{t+1}-\boldsymbol{V}^\star\|_{\mathrm{F}} \le \|\boldsymbol{V}_t-\boldsymbol{V}^\star\|_{\mathrm{F}} + O(\eta)\cdot\sigma_1^\star\sqrt{\frac{dr}{p}}\left(\|\boldsymbol{V}^\star\|_{2,\infty}^2+\|\boldsymbol{V}_t\|_{2,\infty}^2\right)+O(\eta^2).$$

Recall that, due to the incoherence of $\boldsymbol{V}^\star$, we have $\|\boldsymbol{V}^\star\|_{2,\infty}=\sqrt{\mu r/d}$. Now, suppose we can further establish that $\boldsymbol{V}_t$ enjoys a similar incoherence property. In such a case, the aforementioned lemma leads to $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\| \le \|\boldsymbol{V}_t-\boldsymbol{V}^\star\|_{\mathrm{F}} \le O\left(T\cdot\eta\sigma_1^\star\sqrt{\frac{\mu^2 r^3}{pd}}\right)+O(T\cdot\eta^2)$ for every $1\le t\le T$. Given the provided bounds on T and p, this automatically establishes that $\|\boldsymbol{V}_t-\boldsymbol{V}^\star\|$ remains small throughout the iterations. Therefore, it suffices to control the incoherence of $\|\boldsymbol{V}_t\|_{2,\infty}$.

Controlling $\|\boldsymbol{V}_t\|_{2,\infty}$, which necessitates estimating the ℓ_2-norm of each row, requires a more fine-grained analysis than what is needed for the Frobenius norm. The primary challenge lies in the intricate correlations between the orthogonal matrix $\boldsymbol{V}_t$ and the random observation set Ω, which preclude the straightforward application of classical concentration inequalities. To effectively decouple these correlations, we propose a technique called *weakly-coupled leave-one-out analysis*. Before introducing our proposed methodology, it is essential to grasp the core principles of the classical leave-one-out analysis.

Local leave-one-out analysis. When the search rank is exactly parameterized ($r=r'$) and the initial point is sufficiently close to the ground truth $\boldsymbol{U}_0\boldsymbol{U}_0^\top \approx \boldsymbol{X}^\star$, Ma et al. (2018) established the incoherence of the iterates via the following leave-one-out sequences $\left\{\boldsymbol{U}_t^{(l)}\right\}_{t=0}^T$ for each $1\le l\le d$:

$$\operatorname{dist}\left(\boldsymbol{U}_0,\boldsymbol{U}_0^{(l)}\right)\approx 0, \quad \text{and} \quad \boldsymbol{U}_{t+1}^{(l)}=\left(\boldsymbol{I}-\eta\mathcal{R}_{\Omega^{(l)}}\left(\boldsymbol{U}_t^{(l)}\boldsymbol{U}_t^{(l)\top}-\boldsymbol{X}^\star\right)\right)\boldsymbol{U}_t^{(l)}, \tag{8}$$

where $\mathcal{R}_{\Omega^{(l)}}$ is the leave-one-out projection operator defined by

$$\mathcal{R}_{\Omega^{(l)}}=\frac{1}{2p}\left(\mathcal{P}_{\Omega^{(l)}}+\mathcal{P}_{\Omega^{(l)}}^\top\right), \quad \text{and} \quad [\mathcal{P}_{\Omega^{(l)}}(\boldsymbol{X})]_{i,j}=\begin{cases} pX_{i,j} & \text{if } i=l \text{ or } j=l,\\ X_{i,j} & \text{if } (i,j)\in\Omega,\ i\ne l, \text{ and } j\ne l,\\ 0 & \text{otherwise.}\end{cases}$$

The sole distinction between the projection operators $\mathcal{R}_{\Omega^{(l)}}$ and $\mathcal{R}_\Omega$ is in their l-th row and l-th column: in contrast to $\mathcal{R}_\Omega$, the l-th row and l-th column of $\mathcal{R}_{\Omega^{(l)}}(\boldsymbol{X})$ are *deterministically* set to match the corresponding values of $\boldsymbol{X}$. This seemingly minor adjustment yields two important consequences: first, it ensures that $\boldsymbol{U}_t\approx\boldsymbol{U}_t^{(l)}$, and second, it guarantees that the behavior of $\boldsymbol{U}_t^{(l)}$ remains independent of the random measurements in the l-th row and l-th column. This decoupling technique is the key to controlling the deviation of $\|\boldsymbol{U}_t\|_{2,\infty}$. To formalize this intuition, let us define $\boldsymbol{U}^\star=\boldsymbol{V}^\star\boldsymbol{\Sigma}^{\star 1/2}$. One can write

$$\begin{aligned}
\|\boldsymbol{U}_t\|_{2,\infty} &= \max_{1\le l\le d}\left\{\left\|\left[\boldsymbol{U}_t\boldsymbol{H}_t^{(l)}\right]_{l,\cdot}\right\|\right\}\\
&\le \max_{1\le l\le d}\left\{\left\|\left[\boldsymbol{U}^\star-\boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right]_{l,\cdot}\right\|+\left\|\left[\boldsymbol{U}_t\boldsymbol{H}_t^{(l)}-\boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right]_{l,\cdot}\right\|+\left\|[\boldsymbol{U}^\star]_{l,\cdot}\right\|\right\}\\
&\le \underbrace{\max_{1\le l\le d}\left\{\left\|\left(\boldsymbol{U}^\star-\boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right)_{l,\cdot}\right\|\right\}}_{\text{leave-one-out error}}+\underbrace{\max_{1\le l\le d}\left\{\left\|\boldsymbol{U}_t\boldsymbol{H}_t^{(l)}-\boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathrm{F}}\right\}}_{\text{proximal error}}+\sqrt{\frac{\sigma_1^\star\mu r}{d}}.
\end{aligned} \tag{9}$$

Here $\boldsymbol{R}_t^{(l)}$ and $\boldsymbol{H}_t^{(l)}$ are orthogonal matrices defined as $\boldsymbol{R}_t^{(l)} = \arg\min_{\boldsymbol{O}\in\mathcal{O}_{r\times r}} \left\|\boldsymbol{U}_t^{(l)}\boldsymbol{O} - \boldsymbol{U}^\star\right\|_{\mathrm{F}}$ and $\boldsymbol{H}_t^{(l)} = \arg\min_{\boldsymbol{O}\in\mathcal{O}_{r\times r}} \left\|\boldsymbol{U}_t^{(l)}\boldsymbol{O} - \boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}\right\|_{\mathrm{F}}$. Although it may not be immediately obvious, it can be shown that the l-th row of the matrix $\boldsymbol{U}^\star - \boldsymbol{U}_t^{(l)}\boldsymbol{R}_t^{(l)}$ is purely *deterministic*. Consequently, it becomes possible to effectively control the leave-one-out error. To tackle the proximal error, recall that the initial point $\boldsymbol{U}_0\boldsymbol{U}_0^\top$ is assumed to be close to $\boldsymbol{X}^\star$. Within this region, the local landscape exhibits restricted strong convexity. This ensures that the true iterates $\boldsymbol{U}_t$ and the leave-one-out versions $\boldsymbol{U}_t^{(l)}$ become increasingly close, leading to a small proximal error. By combining these two arguments, we can guarantee the incoherence of the true iterates. Furthermore, the incoherence of $\boldsymbol{U}_t$ automatically implies the incoherence of $\boldsymbol{V}_t$, given that $\sqrt{\sigma_r^\star}\left\|\boldsymbol{V}_t\right\|_{2,\infty} \lesssim \left\|\boldsymbol{U}_t\right\|_{2,\infty}$ when $\boldsymbol{U}_t\boldsymbol{U}_t^\top \approx \boldsymbol{X}^\star$. For more details, we refer interested readers to the discussions in (Ma et al., 2018).

While the classical leave-one-out analysis provides precise *local* guarantees within the *exactly-parameterized* regimes, we shed light on its limitations when applied *globally* in the *over-parameterized* settings. A significant challenge arises from the discrepancy of the singular values of $\boldsymbol{U}_t$ and $\boldsymbol{U}_t^{(l)}$: although they may remain close to the singular values of $\boldsymbol{U}^\star$ in the local regime, they can undergo substantial changes when positioned far from the true solution. Consequently, the original measure of proximal error based on $\mathrm{dist}(\boldsymbol{U}_t, \boldsymbol{U}_t^{(l)})$ loses its effectiveness as a reliable metric.

Instead, recall that we only require controlling $\boldsymbol{V}_t$, which unlike $\boldsymbol{U}_t$, has *unit* singular values. This motivates us to switch to a more stable metric—the divergence between the left column spaces of $\boldsymbol{U}_t$ and $\boldsymbol{U}_t^{(l)}$. However, an additional complication is that these left column spaces may also not align perfectly due to over-parameterization. Fortunately, by resorting to our proposed dynamic signal-residual decomposition, we can show that the iterates $\boldsymbol{U}_t$ are well-approximated by the low-rank signal $\boldsymbol{U}_t \approx \boldsymbol{S}_t$. Therefore, it suffices to focus on controlling the discrepancy in the column spaces of $\boldsymbol{S}_t$ and $\boldsymbol{S}_t^{(l)}$, i.e., $\mathrm{dist}(\boldsymbol{V}_t, \boldsymbol{V}_t^{(l)})$. However, the new proximal error $\mathrm{dist}(\boldsymbol{V}_t, \boldsymbol{V}_t^{(l)})$ can still grow exponentially. To explain the root cause of this exponential growth, we employ matrix Taylor expansion to derive the first-order approximations for $\boldsymbol{V}_{t+1}$ and $\boldsymbol{V}_{t+1}^{(l)}$:

$$\boldsymbol{V}_{t+1} = \boldsymbol{V}_t + \eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\boldsymbol{V}_t + O(\eta^2) \quad \text{and} \quad \boldsymbol{V}_{t+1}^{(l)} = \boldsymbol{V}_t^{(l)} + \eta\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)} + O(\eta^2), \tag{10}$$

where we define $\boldsymbol{M}_t^{(l)} = \mathcal{R}_{\Omega^{(l)}}\left(\boldsymbol{X}^\star - \boldsymbol{U}_t^{(l)}\boldsymbol{U}_t^{(l)\top}\right)$. To effectively control the proximal error, it is crucial to establish an upper bound for $\left\|\boldsymbol{M}_t - \boldsymbol{M}_t^{(l)}\right\|$. This distance tends to concentrate around $\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{U}_t^{(l)}\boldsymbol{U}_t^{(l)\top}\right\| \approx \left\|\boldsymbol{S}_t\boldsymbol{S}_t^\top - \boldsymbol{S}_t^{(l)}\boldsymbol{S}_t^{(l)\top}\right\|$ when the sampling rate p is sufficiently large. However, as previously noted, the singular values of $\boldsymbol{S}_t$ and $\boldsymbol{S}_t^{(l)}$ may diverge. This misalignment can lead to $\left\|\boldsymbol{M}_t - \boldsymbol{M}_t^{(l)}\right\| = \Omega(\sigma_1^\star)$ in the worst case. Hence, the proximal error can grow exponentially.

Weakly-coupled leave-one-out analysis. To remedy the alignment challenges identified earlier, we propose the following refined leave-one-out sequences $\left\{\widetilde{\boldsymbol{V}}_t^{(l)}\right\}_{t=0}^{T}$:

$$\widetilde{\boldsymbol{V}}_{t+1}^{(l)} = \widetilde{\boldsymbol{Z}}_{t+1}^{(l)}\left(\widetilde{\boldsymbol{Z}}_{t+1}^{(l)\top}\widetilde{\boldsymbol{Z}}_{t+1}^{(l)}\right)^{-1/2} \quad \text{where} \quad \widetilde{\boldsymbol{Z}}_{t+1}^{(l)} = \left(\boldsymbol{I} + \eta\widetilde{\boldsymbol{M}}_t^{(l)}\right)\widetilde{\boldsymbol{V}}_t^{(l)} \text{ and } \widetilde{\boldsymbol{V}}_0^{(l)} = \boldsymbol{V}^\star. \tag{11}$$

In this context, $\widetilde{\boldsymbol{M}}_t^{(l)}$ is defined as:

$$\widetilde{\boldsymbol{M}}_t^{(l)} = \mathcal{R}_{\Omega^{(l)}}\left(\boldsymbol{X}^\star - \widetilde{\boldsymbol{V}}_t^{(l)}\boldsymbol{\Sigma}_t\widetilde{\boldsymbol{V}}_t^{(l)\top}\right) \quad \text{where } \boldsymbol{\Sigma}_t = \boldsymbol{V}_t^\top\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_t \in \mathbb{R}^{r\times r}. \tag{12}$$

Compared to the original $\boldsymbol{M}_t^{(l)}$, we replace $\boldsymbol{\Sigma}_t^{(l)} = \boldsymbol{V}_t^{(l)\top}\boldsymbol{U}_t^{(l)}\boldsymbol{U}_t^{(l)\top}\boldsymbol{V}_t^{(l)}$ by $\boldsymbol{\Sigma}_t = \boldsymbol{V}_t^\top\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_t$ in the definition of $\widetilde{\boldsymbol{M}}_t^{(l)}$. Our analysis indicates that this adjustment significantly improves the control over the distance $\left\|\boldsymbol{M}_t - \widetilde{\boldsymbol{M}}_t^{(l)}\right\| = o(1)$ when the sampling rate p is sufficiently large. Hence, the proximal error grows at a much slower rate.

Despite their promise, these refined leave-one-out sequences do introduce a trade-off: the statistical independence inherent in the original leave-one-out sequences is compromised due to the inclusion of $\boldsymbol{\Sigma}_t$. In other words, the l-th leave-one-out sequence $\left\{\widetilde{\boldsymbol{V}}_t^{(l)}\right\}_{t=0}^T$ is no longer independent of the random measurements in the l-th row and l-th column. Nonetheless, we demonstrate that the resulting correlation is relatively weak, primarily because $\boldsymbol{\Sigma}_t$ is a comparatively small $r \times r$ matrix. To control this statistical coupling, we employ a novel *adaptive covering argument*, which can be of independent interest. This approach effectively mitigates the statistical coupling while incurring a mild increase in the required sample complexity, which remains only polynomial in r.

To formalize our arguments, we can decompose the refined leave-one-out sequences $\|\boldsymbol{V}_t\|_{2,\infty}$ as:

$$\begin{aligned}\|\boldsymbol{V}_t\|_{2,\infty} &\le \|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{2,\infty} + \|\boldsymbol{V}^\star\|_{2,\infty} \\ &= \max_{1\le l\le d}\left\{\left\|(\boldsymbol{V}^\star - \boldsymbol{V}_t)_{l,\cdot}\right\|\right\} + \|\boldsymbol{V}^\star\|_{2,\infty} \\ &\le \underbrace{\max_{1\le l\le d}\left\{\left\|\left(\boldsymbol{V}^\star - \widetilde{\boldsymbol{V}}_t^{(l)}\right)_{l,\cdot}\right\|\right\}}_{\text{refined leave-one-out error (Proposition 7)}} + \underbrace{\max_{1\le l\le d}\left\{\left\|\boldsymbol{V}_t - \widetilde{\boldsymbol{V}}_t^{(l)}\right\|_{\mathrm{F}}\right\}}_{\text{refined proximal error (Proposition 8)}} + \sqrt{\frac{\mu r}{d}}.\end{aligned} \tag{13}$$

Next, we characterize the dynamic of the refined leave-one-out error.

Proposition 7 (Refined leave-one-out error) *Suppose that $p \gtrsim \frac{\log(d)}{d}$ and $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\| \le \frac{1}{2\kappa}$. With probability at least $1 - \frac{1}{d^3}$, for any $1 \le t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$ and $1 \le l \le d$, we have*

$$\left\|\left(\boldsymbol{V}^\star - \widetilde{\boldsymbol{V}}_{t+1}^{(l)}\right)_{l,\cdot}\right\| \le (1 - 0.5\eta\sigma_r^\star)\left\|\left(\boldsymbol{V}^\star - \widetilde{\boldsymbol{V}}_t^{(l)}\right)_{l,\cdot}\right\| + O(\eta)\cdot\sigma_1^\star\frac{\kappa\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}}.$$

A simple inductive argument based on Proposition 7 reveals that the following inequality holds with an overwhelming probability for all $1 \le t \le T$:

$$\left\|\left(\boldsymbol{V}^\star - \widetilde{\boldsymbol{V}}_t^{(l)}\right)_{l,\cdot}\right\| \lesssim \frac{\kappa^2\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}} \le \sqrt{\frac{\mu r}{4d}}, \quad \text{assuming} \quad p \gtrsim \frac{\kappa^4\mu^2 r^3\log^2\left(\frac{1}{\alpha}\right)\log(d)}{d}.$$

Next, we characterize the dynamic of the refined proximal error.

Proposition 8 (Refined proximal error) *Suppose that $p \gtrsim \frac{\kappa^6\mu^4 r^9\log^6(\frac{1}{\alpha})\log^2(d)}{d}$, $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\| \le \frac{1}{2\kappa}$, and $\left\|\boldsymbol{V}_t - \widetilde{\boldsymbol{V}}_t^{(l)}\right\|_{\mathrm{F}} \le \sqrt{\frac{\mu r}{4d}}$. With probability at least $1 - \frac{1}{d^3}$, for any $1 \le t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$ and $1 \le l \le d$, we have*

$$\left\|\boldsymbol{V}_{t+1} - \widetilde{\boldsymbol{V}}_{t+1}^{(l)}\right\|_{\mathrm{F}} \le \left\|\boldsymbol{V}_t - \widetilde{\boldsymbol{V}}_t^{(l)}\right\|_{\mathrm{F}} + O(\eta)\cdot\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}.$$

The above proposition implies that

$$\left\|\boldsymbol{V}_t - \widetilde{\boldsymbol{V}}_t^{(l)}\right\|_{\mathrm{F}} \lesssim \eta\sigma_1^\star \sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(d\right)}{\sqrt{pd}\cdot d}}\cdot T \lesssim \sqrt{\frac{\kappa^3\mu^3 r^{5.5}\log^3\left(\frac{1}{\alpha}\right)\log\left(d\right)}{\sqrt{pd}\cdot d}} \leq \sqrt{\frac{\mu r}{4d}}.$$

Combining the above inequalities with the proposed decomposition in Equation (13) leads to:

$$\|\boldsymbol{V}_t\|_{2,\infty} \leq \sqrt{\frac{\mu r}{4d}} + \sqrt{\frac{\mu r}{4d}} + \sqrt{\frac{\mu r}{d}} \leq \sqrt{\frac{4\mu r}{d}}, \quad \text{with probably at least } 1 - \frac{2}{d^3}.$$

This establishes the incoherence of $\boldsymbol{V}_t$ for all $1 \leq t \leq T$.

3.3. Improved Results for Exact Parameterization

Finally, we show that our analysis in the over-parameterized regime, combined with the following local convergence result for the exact parameterization regime by Ma et al. (2018), readily establishes the proof of Theorem 3.

Theorem 9 (Local convergence of GD (Ma et al., 2018, Theorem 2)) *Consider MC with search rank $r' = r$. Suppose that the sampling rate satisfies $p \gtrsim \frac{\mu^3 r^3 \log^3(d)}{d}$. Consider the iterates of GD with step-size $\eta \leq \frac{2}{25\kappa\sigma_1^\star}$. Suppose that there exists $t_0 \geq 0$ such that $\boldsymbol{U}_{t_0}$ and the leave-one-out sequences $\left\{\boldsymbol{U}_{t_0}^{(l)}\right\}_{l=0}^{d}$ defined in Equation (8) satisfy:*

$$\operatorname{dist}\left(\boldsymbol{U}_{t_0}, \boldsymbol{U}^\star\right) \leq O\left(\sqrt{\frac{\sigma_r^\star \mu^3 r^3 \log(d)}{pd^2}}\right), \tag{14}$$

$$\max\left\{\operatorname{dist}\left(\boldsymbol{U}_{t_0}, \boldsymbol{U}_{t_0}^{(l)}\right), \operatorname{dist}\left(\boldsymbol{U}_{t_0}^{(l)}, \boldsymbol{U}^\star\right)\right\} \leq O\left(\sqrt{\frac{\sigma_r^\star \mu^3 r^3 \log(d)}{pd^2}}\right), \quad \textit{for all } 1 \leq l \leq d. \tag{15}$$

With probability at least $1 - O\left(\frac{1}{d^3}\right)$, for all $t_0 \leq t \leq t_0 + O(d^5)$, we have

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \leq \left\|\boldsymbol{U}_{t_0}\boldsymbol{U}_{t_0}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} (1 - 0.2\eta\sigma_r^\star)^{t-t_0}.$$

To prove Theorem 3, it suffices to show that the conditions of the above theorem are met at a certain iteration $0 \leq t_0 \leq T$. This can be achieved by leveraging our result for the over-parameterized regime. In particular, upon choosing $\alpha = c \cdot \frac{\sigma_r^\star}{\kappa^{1.5} d}$ for sufficiently small $c > 0$ in Theorem 2, one can show that both Conditions (14) and (15) are satisfied with an overwhelming probability after $t_0 = \widetilde{O}\left(\frac{1}{\eta\sigma_r^\star}\right)$ iterations. From this iteration onward, Theorem 9 shows that the iterations of GD enter a local linear convergence regime, which readily establishes the final result of Theorem 3.

4. Conclusion and Future Directions

In this paper, we prove the convergence of vanilla gradient descent (GD) with small initialization for symmetric matrix completion. Existing convergence results for this problem typically require

explicit regularization or precise initializations. However, our work proves that neither condition is necessary for GD to converge. Moreover, our results also apply to the over-parameterized regime, where the rank of the true solution is unknown and over-estimated instead.

Although our required sample complexity $\widetilde{O}(dr^9)$ is optimal with respect to the dimension d, it remains sub-optimal with respect to the rank r. Specifically, it exceeds the sample complexity of regularized GD, which stands at $\widetilde{O}(dr^2)$ (Chen and Wainwright, 2015). We expect our analysis can be sharpened to achieve a similar sample complexity.

We anticipate that our findings will pave the way for broader results extending beyond symmetric matrix completion. In particular, our proposed weakly-coupled leave-one-out analysis relaxes several stringent conditions of classical leave-one-out analysis, making it highly applicable for the global analysis of GD. We believe that this approach, along with potential variations, holds promise for explaining the favorable performance of GD or its variants in various statistical learning problems.

Acknowledgment

We thank Hongyang R. Zhang for pointing him to the updated version of (Li et al., 2018). We thank Richard Y. Zhang and Cédric Josz for their helpful discussions. We are grateful to the anonymous reviewers for pointing out the limitations of our sample complexity result in the overparameterized regime. This work is supported, in part, by NSF CAREER Award CCF-2337776, NSF Award DMS-2152776, and ONR Award N00014-22-1-2127.

References

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452, 2020.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

Emmanuel Candes and Benjamin Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Joshua Cape, Minh Tang, and Carey E Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. 2019.

Ji Chen and Xiaodong Li. Memory-efficient kernel pca via partial matrix sampling and nonconvex optimization: a model-free analysis of local minima. *arXiv preprint arXiv:1711.01742*, 2017.

Ji Chen and Xiaodong Li. Model-free nonconvex matrix completion: Local minima analysis and applications in memory-efficient kernel pca. *J. Mach. Learn. Res.*, 20(142):1–39, 2019.

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

Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176:5–37, 2019.

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

Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 66(11):7274–7301, 2020.

Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36): 14557–14562, 2013.

Salar Fattahi and Somayeh Sojoudi. Exact guarantees on the absence of spurious local minima for non-negative rank-1 robust principal component analysis. *Journal of machine learning research*, 2020.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 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*, pages 1233–1242. PMLR, 2017.

Rong Ge, Yunwei Ren, Xiang Wang, and Mo Zhou. Understanding deflation process in over-parametrized tensor decomposition. *Advances in Neural Information Processing Systems*, 34: 1299–1311, 2021.

David F Gleich and Lek-heng Lim. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 60–68, 2011.

Thore Graepel. Kernel matrix completion by semidefinite programming. In *International Conference on Artificial Neural Networks*, pages 694–699. Springer, 2002.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

Yue Hu, Xiaohan Liu, and Mathews Jacob. A generalized structured low-rank matrix completion algorithm for mr image recovery. *IEEE transactions on medical imaging*, 38(8):1841–1851, 2018.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

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

Nayoung Lee, Kartik Sreenivasan, Jason D Lee, Kangwook Lee, and Dimitris Papailiopoulos. Teaching arithmetic to small transformers. *arXiv preprint arXiv:2307.03381*, 2023.

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*, pages 2–47. PMLR, 2018.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *arXiv preprint arXiv:1711.10467*, 2017.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

Jianhao Ma and Salar Fattahi. Blessing of depth in linear regression: Deeper models have flatter landscape around the true solution. *Advances in Neural Information Processing Systems*, 35: 34334–34346, 2022.

Jianhao Ma and Salar Fattahi. Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *Journal of Machine Learning Research*, 24(96):1–84, 2023a.

Jianhao Ma and Salar Fattahi. On the optimization landscape of burer-monteiro factorization: When do global solutions correspond to ground truth? *arXiv preprint arXiv:2302.10963*, 2023b.

Jianhao Ma, Lingjun Guo, and Salar Fattahi. Behind the scenes of gradient descent: A trajectory analysis via basis function decomposition. *arXiv preprint arXiv:2210.00346*, 2022.

John Paisley and Lawrence Carin. A nonparametric bayesian model for kernel matrix completion. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2090–2093. IEEE, 2010.

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. *Matrix completion via nonconvex factorization: Algorithms and theory*. PhD thesis, University of Minnesota, 2015.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

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*, pages 964–973. PMLR, 2016.

Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Van Vu. A simple svd algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Xiang Wang, Chenwei Wu, Jason D Lee, Tengyu Ma, and Rong Ge. Beyond lazy training for over-parameterized tensor decomposition. *Advances in Neural Information Processing Systems*, 33:21934–21944, 2020.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

Contents

Appendix A. Related Work

Nonconvex matrix completion. To solve the matrix completion problem, several algorithms based on convex optimization have been developed (Candes and Recht, 2012; Candès and Tao, 2010; Gross, 2011), offering excellent theoretical guarantees. However, in high-dimensional scenarios, convex optimization techniques require significant memory and computational resources due to the iterative singular value decompositions. To overcome these limitations, researchers have shifted towards nonconvex optimization techniques using first-order methods such as GD (Sun and Luo, 2016), projected GD (Zheng and Lafferty, 2016), and alternating minimization (Jain et al., 2013). Specifically, Sun and Luo (2016) demonstrate that GD can achieve local linear convergence provided that the initialization is close to the ground truth. Subsequent studies provide the global convergence guarantees for the first-order methods by showing the benign landscape of these nonconvex optimization formulations. Specifically, they reveal that the loss landscape has no spurious local minima and all the saddle points are strict (Ge et al., 2016, 2017; Chen and Li, 2017; Fattahi and Sojoudi, 2020). Nonetheless, these advancements necessitate either an explicit $\ell_{2,\infty}$-norm regularization or a projection step to maintain the incoherence of the iterates. Moreover, these works are only applicable in the exactly-parameterized setting Ma and Fattahi (2023b). For a more detailed exploration of matrix completion and its variants, we refer the readers to the comprehensive survey by Chi et al. (2019).

Leave-one-out analysis. Leave-one-out analysis is a powerful statistical technique designed to decouple correlations among individual entries of a stochastic process. Initially employed by El Karoui et al. (2013) to establish asymptotic sampling distributions for robust estimators in high or moderate dimensional regression, this technique has been proven invaluable across a broad spectrum of applications. For instance, Abbe et al. (2020) utilized it to control ℓ_∞ estimation errors for eigenvectors in stochastic spectral problems, enabling precise spectral clustering in community detection without the need for data cleaning or regularization. More relevantly, Ma et al. (2017) applied leave-one-out analysis to demonstrate the local linear convergence of GD for the unregularized and symmetric matrix completion. Their approach not only elucidated the convergence properties of GD in matrix completion but also paved the way for similar analyses in other low-rank recovery challenges, such as phase retrieval and blind deconvolution. Extending these insights, Chen et al. (2020) and Kim and Chung (2022) broadened the scope of this analysis to include asymmetric matrix completion and global convergence in rank-1 scenarios, respectively. Furthermore, leave-one-out analysis has facilitated advancements in Singular Value Projection (SVP) for matrix completion, as demonstrated by Ding and Chen (2020) and has been instrumental in analyzing gradient descent with random initialization for phase retrieval, as shown by Chen et al. (2019).

Implicit regularization of GD in other applications. Indeed, the conventional wisdom in statistics suggests that increasing the number of parameters beyond the true dimension without proper regularization would lead to inferior solutions due to *overfitting*. However, a growing body of works show that, for a large class of learning problems, GD leads to surprisingly good solutions, due to its *implicit regularization* property. For instance, it is known that GD recovers the true low-dimensional solutions in matrix factorization and sensing (Gunasekar et al., 2018; Li et al., 2018; Stöger and Soltanolkotabi, 2021), tensor decomposition (Wang et al., 2020; Ge et al., 2021), deep linear neural networks (Arora et al., 2018; Ma and Fattahi, 2022), and beyond (Ma et al., 2022). However, the current theory behind the success of GD in these classes of problems hinges heavily upon a norm-preserving property

of the measurements, known as the restricted isometry property (RIP), limiting its applicability in settings where RIP is not satisfied.

Appendix B. Preliminaries

B.1. Otuline of the Appendix

The structure of the appendix is as follows. In the remainder of this section, we introduce additional notation. Following this, we present key intermediate lemmas (Lemmas 10 to 13) crucial for our main proofs. Section C delves into a detailed proof of the signal and residual dynamics, starting with a refined version of Lemma 5 (Proposition 14) that takes into account the incoherence of $\boldsymbol{V}_t$. Additionally, the proof of Lemma 4 is provided in this section. Moving on to Section D, we present the proofs of our main theorems. Section E presents the key novelty of our paper, focusing on establishing the incoherence of $\boldsymbol{V}_t$ via weakly-coupled leave-one-out analysis. The validation of different initialization schemes, as presented in Lemma 1, is addressed in Section F. In Section G, we compile several known results on matrix completion crucial to our arguments. Lastly, Section H collects several basic lemmas, which we include for completeness.

B.2. Additional Notations

We introduce some additional notations that will be used throughout the appendix. The max-norm of $\boldsymbol{X}$, denoted as $\|\boldsymbol{X}\|_{\max}$, is defined as $\max_{i,j}|X_{i,j}|$. We define the operator and Frobenius norm ball as $\mathcal{B}_{\mathrm{op}}^{d_1\times d_2}(r) := \{\boldsymbol{X}\in\mathbb{R}^{d_1\times d_2} : \|\boldsymbol{X}\|\le r\}$ and $\mathcal{B}_{\mathrm{F}}^{d_1\times d_2}(r) := \{\boldsymbol{X}\in\mathbb{R}^{d_1\times d_2} : \|\boldsymbol{X}\|_{\mathrm{F}}\le r\}$, respectively. For any matrix $\boldsymbol{X}$, we denote its SVD as $\boldsymbol{X}=\boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{R}_{\boldsymbol{X}}^\top$. We denote $\mathcal{S}_{d\times d}$ as the set of all the symmetric matrices $\boldsymbol{X}\in\mathbb{R}^{d\times d}$. In the appendix, $\Gamma,\Gamma_1,\Gamma_2,\ldots$ denote fixed universal constants, while $C,C_1,C_2,c_1,c_2,\ldots$ represent universal constants whose specific values may vary depending on the context.

Throughout the appendix, our arguments are conditioned on the following good event without further explanation. We define the random observation matrix $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega}_{i,j}=\begin{cases}1 & \text{if } (i,j)\in\Omega,\\ 0 & \text{otherwise.}\end{cases} \tag{16}$$

Then, the good event can be defined as

$$E_{\mathrm{good}}=\left\{\left\|\frac{\boldsymbol{\Omega}+\boldsymbol{\Omega}^\top}{2p}-\boldsymbol{J}\right\|\le\Gamma\sqrt{\frac{d}{p}}\right\}. \tag{17}$$

Here $\boldsymbol{J}$ is the all-one matrix. According to Lemma 25, we have $\mathbb{P}(E_{\mathrm{good}})\ge 1-\frac{1}{d^3}$.

B.3. Important Intermediate Lemmas

Next, we collect some useful intermediate results that will be directly used throughout our proofs. We also note that some of these intermediate results rely on concentration inequalities for matrix completion, which are thoroughly discussed in Appendix G. Before proceeding, we define the following notations

$$\boldsymbol{\Delta}_t=\boldsymbol{X}^\star-\boldsymbol{U}_t\boldsymbol{U}_t^\top,\quad \boldsymbol{M}_t=\mathcal{R}_\Omega\left(\boldsymbol{X}^\star-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right),\quad\text{and}\quad \boldsymbol{\Lambda}_t=\boldsymbol{S}_t\boldsymbol{E}_t^\top+\boldsymbol{E}_t\boldsymbol{S}_t^\top+\boldsymbol{E}_t\boldsymbol{E}_t^\top. \tag{18}$$

Moreover, we introduce the following term

$$\boldsymbol{A}_t = -\eta^2 \boldsymbol{M}_t \boldsymbol{V}_t \boldsymbol{V}_t^\top \boldsymbol{M}_t \boldsymbol{V}_t - 0.5\eta^2 \boldsymbol{V}_t \boldsymbol{V}_t^\top \boldsymbol{M}_t^2 \boldsymbol{V}_t - 0.5\eta^3 \boldsymbol{M}_t \boldsymbol{V}_t \boldsymbol{V}^{\star\top} \boldsymbol{M}_t^2 \boldsymbol{V}_t + (\boldsymbol{I} + \eta \boldsymbol{M}_t)\, \boldsymbol{V}_t \boldsymbol{R}(\boldsymbol{Y}_t) \tag{19}$$

where $\boldsymbol{Y}_t = \boldsymbol{V}_t^\top \left(2\eta \boldsymbol{M}_t + \eta^2 \boldsymbol{M}_t^2\right) \boldsymbol{V}_t$ and $\boldsymbol{R}(\boldsymbol{X}) = \sum_{k=2}^{\infty} \frac{(-1)^k (2k)!}{4^k (k!)^2} \boldsymbol{X}^k$. This notion of $\boldsymbol{A}_t$ will be used when controlling the higher-order terms with respect to the step-size η. We are now ready to statement our helper lemmas.

Lemma 10 (Helper lemma for $\boldsymbol{U}_t$) *Suppose that $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \leq \sqrt{\frac{\sigma_1^\star \mu r}{d}}$. Then, we have*

$$\left\| \boldsymbol{U}_t \boldsymbol{U}_t^\top \right\|_{\mathrm{F}} \leq 8\sqrt{r}\sigma_1^\star. \tag{20}$$

Proof Applying triangle inequality, we have

$$\begin{aligned} \left\| \boldsymbol{U}_t \boldsymbol{U}_t^\top \right\|_{\mathrm{F}} &\leq \left\| \boldsymbol{S}_t \boldsymbol{S}_t^\top \right\|_{\mathrm{F}} + 2 \left\| \boldsymbol{S}_t \boldsymbol{E}_t^\top \right\|_{\mathrm{F}} + \left\| \boldsymbol{E}_t \boldsymbol{E}_t^\top \right\|_{\mathrm{F}} \\ &\leq \sqrt{r} \left\| \boldsymbol{S}_t \right\|^2 + 2\sqrt{r} \left\| \boldsymbol{S}_t \right\| \left\| \boldsymbol{E}_t \right\| + \sqrt{d} \left\| \boldsymbol{E}_t \right\|^2 \\ &\leq 8\sqrt{r}\sigma_1^\star. \end{aligned} \tag{21}$$

Here in the last inequality, we use the assumptions $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \leq \sqrt{\frac{\sigma_1^\star \mu r}{d}}$ and the fact that $d \gg \mu r$. ■

Lemma 11 (Helper lemma for $\boldsymbol{\Lambda}_t$) *Suppose that $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$, $\|\boldsymbol{V}_t\|_{2,\infty} \leq 2\sqrt{\frac{\mu r}{d}}$, and $\|\boldsymbol{E}_t\| \leq \sqrt{\frac{\sigma_1^\star \mu r}{81 d}}$. Then, conditioned on E_{good}, we have*

$$\begin{aligned} \left\| \boldsymbol{\Lambda}_t \right\| &\leq 5\sqrt{\sigma_1^\star} \left\| \boldsymbol{E}_t \right\|, \\ \left\| \mathcal{R}_\Omega(\boldsymbol{\Lambda}_t) \right\| &\leq 10\Gamma \sqrt{\frac{\sigma_1^\star \mu r}{p}} \left\| \boldsymbol{E}_t \right\|, \\ \left\| \left(\mathcal{I} - \mathcal{R}_\Omega\right)(\boldsymbol{\Lambda}_t) \right\| &\leq 9\Gamma \sqrt{\frac{\sigma_1^\star \mu r}{p}} \left\| \boldsymbol{E}_t \right\|. \end{aligned} \tag{22}$$

Proof First, we can bound $\|\boldsymbol{\Lambda}_t\|$ as follows

$$\left\| \boldsymbol{\Lambda}_t \right\| \leq \left\| \boldsymbol{S}_t \boldsymbol{E}_t^\top \right\| + \left\| \boldsymbol{E}_t \boldsymbol{S}_t^\top \right\| + \left\| \boldsymbol{E}_t \boldsymbol{E}_t^\top \right\| \leq 5\sqrt{\sigma_1^\star} \left\| \boldsymbol{E}_t \right\| \tag{23}$$

where we use the assumptions $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \le \sqrt{\frac{\sigma_1^\star \mu r}{d}} \le \sqrt{\sigma_1^\star}$. Next, we control $\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Lambda}_t)\|$. To this end, we first apply triangle inequality to obtain

$$\begin{aligned}
\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Lambda}_t)\| &\le 2\left\|(\mathcal{I}-\mathcal{R}_\Omega)\left(\boldsymbol{S}_t\boldsymbol{E}_t^\top\right)\right\| + \left\|(\mathcal{I}-\mathcal{R}_\Omega)\left(\boldsymbol{E}_t\boldsymbol{E}_t^\top\right)\right\| \\
&\overset{(a)}{\le} \Gamma\sqrt{\frac{d}{p}}\|\boldsymbol{E}_t\|_{2,\infty}\left(\|\boldsymbol{E}_t\|_{2,\infty} + 2\|\boldsymbol{S}_t\|_{2,\infty}\right) \\
&\overset{(b)}{\le} \Gamma\sqrt{\frac{d}{p}}\|\boldsymbol{E}_t\|\left(\|\boldsymbol{E}_t\| + 8\sqrt{\frac{\sigma_1^\star\mu r}{d}}\right) \\
&\le 9\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\|.
\end{aligned} \tag{24}$$

Here in (a), we apply Lemma 30. In (b), we use the facts that $\|\boldsymbol{E}_t\|_{2,\infty} \le \|\boldsymbol{E}_t\| \le \sqrt{\frac{\sigma_1^\star\mu r}{d}}$ and

$$\|\boldsymbol{S}_t\|_{2,\infty} = \left\|\boldsymbol{V}_t\boldsymbol{V}_t^\top\boldsymbol{U}_t\right\|_{2,\infty} \overset{\text{Lemma 30}}{\le} \|\boldsymbol{V}_t\|_{2,\infty}\|\boldsymbol{S}_t\| \le 2\sqrt{\frac{\mu r}{d}}\cdot 2\sqrt{\sigma_1^\star} = 4\sqrt{\frac{\sigma_1^\star\mu r}{d}}. \tag{25}$$

In the last inequality, we use the fact that $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{V}_t\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$. Lastly, applying triangle inequality leads to

$$\begin{aligned}
\|\mathcal{R}_\Omega(\boldsymbol{\Lambda}_t)\| &\le \|\boldsymbol{\Lambda}_t\| + \|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Lambda}_t)\| \\
&\le 5\sqrt{\sigma_1^\star}\|\boldsymbol{E}_t\| + 9\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\| \\
&\le 10\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\|,
\end{aligned} \tag{26}$$

where the last inequality is due to $p \le \frac{1}{25}\Gamma^2\mu r$. ■

Lemma 12 (Helper lemma for $\boldsymbol{\Delta}_t$) *Under the same conditions as Lemma 11 with the additional assumption that $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\| \le \Gamma_1\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}$, we have*

$$\begin{aligned}
\|\boldsymbol{\Delta}_t\| &\le 5\sigma_1^\star, \\
\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| &\le 5\Gamma\sqrt{\frac{\mu^2r^2}{pd}}\|\boldsymbol{\Delta}_t\| + 10\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\|, \\
\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| &\le 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2r^2}{pd}}, \\
\|\boldsymbol{M}_t\| &\le \left(1 + 5\Gamma\sqrt{\frac{\mu^2r^2}{pd}}\right)\|\boldsymbol{\Delta}_t\| + 10\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\|\boldsymbol{E}_t\|, \\
\|\boldsymbol{M}_t\| &\le 6\sigma_1^\star, \\
\left\|\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^\perp\right\| &\le 2\Gamma_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.
\end{aligned} \tag{27}$$

Proof We first control $\|\boldsymbol{\Delta}_t\|$ as follows

$$\begin{aligned}\|\boldsymbol{\Delta}_t\| &\leq \left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + \|\boldsymbol{\Lambda}_t\| \\ &\leq \max\{\|\boldsymbol{X}^\star\|, \|\boldsymbol{S}_t\|^2\} + 5\sqrt{\sigma_1^\star}\,\|\boldsymbol{E}_t\| \\ &\leq 4\sigma_1^\star + 5\sqrt{\sigma_1^\star}\,\|\boldsymbol{E}_t\| \\ &\leq 5\sigma_1^\star.\end{aligned} \tag{28}$$

Next, we control $\|(\mathcal{I} - \mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\|$. To this end, applying triangle inequality leads to

$$\|(\mathcal{I} - \mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \leq \underbrace{\left\|(\mathcal{I} - \mathcal{R}_\Omega)\left(\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right\|}_{:=\text{(I)}} + \underbrace{\|(\mathcal{I} - \mathcal{R}_\Omega)(\boldsymbol{\Lambda}_t)\|}_{:=\text{(II)}}. \tag{29}$$

For (I), applying Lemma 31, we have

$$\begin{aligned}\text{(I)} &\leq \Gamma\sqrt{\frac{d}{p}}\left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|\left(\|\boldsymbol{V}^\star\|_{2,\infty}^2 + \|\boldsymbol{V}_t\|_{2,\infty}^2\right) \\ &\overset{(a)}{\leq} 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \\ &\leq 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\left(\|\boldsymbol{\Delta}_t\| + \|\boldsymbol{\Lambda}_t\|\right) \\ &\overset{(b)}{\leq} 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\,\|\boldsymbol{\Delta}_t\| + 25\Gamma\sqrt{\frac{\sigma_1^\star\mu^2 r^2}{pd}}\,\|\boldsymbol{E}_t\|.\end{aligned} \tag{30}$$

Here in (a), we use the assumption that $\|\boldsymbol{V}_t\|_{2,\infty} \leq 2\sqrt{\frac{\mu r}{d}}$. In (b), we use the result from Lemma 11 that $\|\boldsymbol{\Lambda}_t\| \leq 5\sqrt{\sigma_1^\star}\,\|\boldsymbol{E}_t\|$. On the other hand, we know that $\text{(II)} \leq 9\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\,\|\boldsymbol{E}_t\|$ due to Lemma 11. Overall, we conclude that

$$\begin{aligned}\|(\mathcal{I} - \mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| &\leq 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\,\|\boldsymbol{\Delta}_t\| + 25\Gamma\sqrt{\frac{\sigma_1^\star\mu^2 r^2}{pd}}\,\|\boldsymbol{E}_t\| + 9\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\,\|\boldsymbol{E}_t\| \\ &\leq 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\,\|\boldsymbol{\Delta}_t\| + 10\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\,\|\boldsymbol{E}_t\|.\end{aligned} \tag{31}$$

In the final inequality, we make the assumption $d \geq 9\mu r$ without loss of generality. This assumption simplifies the presentation of the proof but does not impact the final result of the paper.

Furthermore, upon noticing that $\left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| \leq \max\{\|\boldsymbol{X}^\star\|, \|\boldsymbol{S}_t\|^2\} \leq 4\sigma_1^\star$, we have

$$\begin{aligned}\|(\mathcal{I} - \mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| &\leq 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\left\|\boldsymbol{X}^\star - \boldsymbol{S}_t\boldsymbol{S}_t^\top\right\| + \|(\mathcal{I} - \mathcal{R}_\Omega)(\boldsymbol{\Lambda}_t)\| \\ &\leq 20\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}} + 9\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\,\|\boldsymbol{E}_t\| \\ &\leq 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}.\end{aligned} \tag{32}$$

Next, combining the above two inequalities, we can control $\|\boldsymbol{M}_t\|$ as follows:

$$\begin{aligned}\|\boldsymbol{M}_t\| &\le \|\boldsymbol{\Delta}_t\| + \|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \\ &\le \|\boldsymbol{\Delta}_t\| + 5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\|\boldsymbol{\Delta}_t\| + 10\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\|\boldsymbol{E}_t\| \\ &\le \left(1+5\Gamma\sqrt{\frac{\mu^2 r^2}{pd}}\right)\|\boldsymbol{\Delta}_t\| + 10\Gamma\sqrt{\frac{\sigma_1^\star \mu r}{p}}\|\boldsymbol{E}_t\|.\end{aligned} \tag{33}$$

Furthermore, we can also bound $\|\boldsymbol{M}_t\|$ as

$$\begin{aligned}\|\boldsymbol{M}_t\| &\le \|\boldsymbol{\Delta}_t\| + \|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| \\ &\le 5\sigma_1^\star + 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}} \\ &\le 6\sigma_1^\star.\end{aligned} \tag{34}$$

Lastly, for $\left\|\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^\perp\right\|$, we have the following decomposition

$$\begin{aligned}\left\|\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^\perp\right\| &\le \left\|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\mathcal{P}_{\boldsymbol{V}_t}^\perp\right\| + \left\|\left(\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\mathcal{P}_{\boldsymbol{V}_t}^\perp\right\| \\ &\le \|(\mathcal{I}-\mathcal{R}_\Omega)(\boldsymbol{\Delta}_t)\| + \left\|\boldsymbol{V}^\star\boldsymbol{\Sigma}(\boldsymbol{V}^\star-\boldsymbol{V}_t)^\top\mathcal{P}_{\boldsymbol{V}_t}^\perp\right\| + \left\|\boldsymbol{U}_t\boldsymbol{E}_t^\top\mathcal{P}_{\boldsymbol{V}_t}^\perp\right\| \\ &\le 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}} + \sigma_1^\star\|\boldsymbol{V}^\star-\boldsymbol{V}_t\| + \|\boldsymbol{E}_t\|(\|\boldsymbol{S}_t\|+\|\boldsymbol{E}_t\|) \\ &\le 22\Gamma\frac{\sigma_1^\star\mu r}{\sqrt{pd}} + \sigma_1^\star\|\boldsymbol{V}^\star-\boldsymbol{V}_t\| \\ &\overset{(a)}{\le} 22\Gamma\frac{\sigma_1^\star\mu r}{\sqrt{pd}} + \Gamma_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \\ &\le 2\Gamma_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.\end{aligned} \tag{35}$$

Here in (a), we apply Lemma 6. ■

Lemma 13 (Helper lemma for $\boldsymbol{A}_t$) *Under the same conditions as Lemma 11, we have*

$$\|\boldsymbol{A}_t\| \le 300\eta^2\sigma_1^{\star 2}. \tag{36}$$

Proof We first use triangle inequality to bound $\|\boldsymbol{A}_t\|$ as follows

$$\|\boldsymbol{A}_t\| \le 1.5\eta^2\|\boldsymbol{M}_t\|^2 + 0.5\eta^3\|\boldsymbol{M}_t\|^3 + (1+\eta\|\boldsymbol{M}_t\|)\|\boldsymbol{R}(\boldsymbol{Y}_t)\|. \tag{37}$$

Next, Lemma 12 tells us that $\|\boldsymbol{M}_t\| \le 6\sigma_1^\star$ conditioned on E_{good}. For $\|\boldsymbol{R}(\boldsymbol{Y}_t)\|$, we first have

$$\|\boldsymbol{R}(\boldsymbol{Y}_t)\| \le \sum_{k=2}^{\infty}\frac{(2k)!}{4^k(k!)^2}\|\boldsymbol{Y}_t\|^k = \frac{2+\sqrt{1-\|\boldsymbol{Y}_t\|}}{2\sqrt{1-\|\boldsymbol{Y}_t\|}\left(1+\sqrt{1-\|\boldsymbol{Y}_t\|}\right)^2}\|\boldsymbol{Y}_t\|^2. \tag{38}$$

Note that $\|\boldsymbol{Y}_t\| \le \left\|2\eta\boldsymbol{M}_t + \eta^2\boldsymbol{M}_t^2\right\| \le 20\eta\sigma_1^\star$ and the right-hand side is an increasing function of $\|\boldsymbol{Y}_t\|$. Therefore, we derive that

$$\|\boldsymbol{R}(\boldsymbol{Y}_t)\| \le \frac{1}{2}\|\boldsymbol{Y}_t\|^2 \le 200\eta^2\sigma_1^{\star 2}. \tag{39}$$

This implies that

$$\|\boldsymbol{A}_t\| \le 300\eta^2\sigma_1^{\star 2}. \tag{40}$$

■

Appendix C. Proofs for Dynamic Signal-residual Decomposition

We first present a more precise version of the one-step dynamics of the signal and residual terms.

Proposition 14 *Suppose that* $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$, $\|\boldsymbol{E}_t\| \le \sqrt{\frac{\sigma_1^\star}{d}}$, $\|\boldsymbol{V}_t\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$ *and* $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}} \le \Gamma_1\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}$. *Then, the following dynamics hold conditioned on* E_{good}*:*

$$\sigma_r(\boldsymbol{S}_{t+1}) \ge \left(1 + 0.8\eta\sigma_r^\star - \eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t) - 6\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\|\boldsymbol{E}_t\|,$$

(minimal signal dynamic)

$$\|\boldsymbol{S}_{t+1}\| \le 2\sqrt{\sigma_1^\star}, \tag{maximal signal dynamic}$$

$$\|\boldsymbol{E}_{t+1}\| \le \left(1 + 2\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)\|\boldsymbol{E}_t\|. \tag{residual dynamic}$$

Additionally, if $\sigma_r(\boldsymbol{S}_t) \ge \frac{\sqrt{\sigma_r^\star}}{2}$, *then*

$$\left\|\boldsymbol{X}^\star - \boldsymbol{U}_{t+1}\boldsymbol{U}_{t+1}^\top\right\|_{\mathrm{F}} \le \left(1 - \frac{1}{10}\eta\sigma_r^\star\right)\left\|\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}} + \Gamma_6\eta\sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}}\|\boldsymbol{E}_t\|.$$

(error dynamic)

The key distinction between the above proposition and Lemma 5 lies in a finer control over the one-step dynamics, accompanied by additional assumptions on $\|\boldsymbol{E}_t\|$, $\|\boldsymbol{V}_t\|_{2,\infty}$, and $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_{\mathrm{F}}$. We emphasize that the one-step dynamics in Lemma 5 are derived from the proof of this proposition. Moreover, the proof of the error dynamics (last inequality in Proposition 14) will incorporate the proof of the descent lemma (Lemma 4).

C.1. Proof of Signal Dynamic

We notice that

$$\boldsymbol{S}_{t+1} = \mathcal{P}_{\boldsymbol{V}_{t+1}}\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\left(\boldsymbol{S}_t + \boldsymbol{E}_t\right) = \left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{S}_t + \mathcal{P}_{\boldsymbol{V}_{t+1}}\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{E}_t. \tag{41}$$

Here the second equality follows from the definition of $\mathcal{P}_{\boldsymbol{V}_{t+1}}$.

Maximal signal dynamic. We first provide an upper-bound for $\|\boldsymbol{S}_{t+1}\|$ by

$$\|\boldsymbol{S}_{t+1}\| \le \underbrace{\|(\boldsymbol{I}+\eta\boldsymbol{M}_t)\,\boldsymbol{S}_t\|}_{:=(\mathrm{I})} + \underbrace{\left\|\mathcal{P}_{\boldsymbol{V}_{t+1}}\,(\boldsymbol{I}+\eta\boldsymbol{M}_t)\,\boldsymbol{E}_t\right\|}_{:=(\mathrm{II})}. \tag{42}$$

Next, we further control (I) by

$$\begin{aligned}
(\mathrm{I}) &\le \|(\boldsymbol{I}+\eta\boldsymbol{\Delta}_t)\,\boldsymbol{S}_t\| + \eta\,\|(\mathcal{I}-\mathcal{R}_\Omega)\,(\boldsymbol{\Delta}_t)\|\,\|\boldsymbol{S}_t\| \\
&\overset{(a)}{\le} \|(\boldsymbol{I}+\eta\boldsymbol{\Delta}_t)\,\boldsymbol{S}_t\| + \eta\cdot 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}\cdot 2\sqrt{\sigma_1^\star} \\
&\le \left\|\left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\right\| + \eta\,\|\boldsymbol{\Lambda}_t\|\,\|\boldsymbol{S}_t\| + 42\Gamma\eta\frac{\sigma_1^{\star 1.5}\mu r}{\sqrt{pd}} \\
&\overset{(b)}{\le} \left\|\left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\right\| + 10\eta\sigma_1^\star\,\|\boldsymbol{E}_t\| + 42\Gamma\eta\frac{\sigma_1^{\star 1.5}\mu r}{\sqrt{pd}} \\
&\le \underbrace{\left\|\left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\right\|}_{:=(\mathrm{I}_1)} + 43\Gamma\eta\frac{\sigma_1^{\star 1.5}\mu r}{\sqrt{pd}}.
\end{aligned} \tag{43}$$

Here we apply Lemma 12 in (a) and (b). In the last inequality, we use the assumption that $\|\boldsymbol{E}_t\| \le \frac{\Gamma\sqrt{\sigma_1^\star}\mu r}{\sqrt{pd}}$. For (I_1), we further decompose it via triangle inequality as follows

$$\begin{aligned}
(\mathrm{I}_1) &\le \left\|\mathcal{P}_{\boldsymbol{V}_t}\left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\right\| + \left\|\mathcal{P}_{\boldsymbol{V}_t^\perp}\left(\boldsymbol{I}+\eta\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\right)\boldsymbol{S}_t\right\| \\
&= \underbrace{\left\|\left(\boldsymbol{I}+\eta\left(\boldsymbol{V}_t^\top\boldsymbol{V}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t-\bar{\boldsymbol{S}}_t\bar{\boldsymbol{S}}_t^\top\right)\right)\bar{\boldsymbol{S}}_t\right\|}_{:=(\mathrm{I}_{1,1})} + \eta\underbrace{\left\|\mathcal{P}_{\boldsymbol{V}_t^\perp}\,(\boldsymbol{V}^\star-\boldsymbol{V}_t)\,\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{S}_t\right\|}_{:=(\mathrm{I}_{1,2})}.
\end{aligned} \tag{44}$$

Here we define $\bar{\boldsymbol{S}}_t = \boldsymbol{V}_t^\top\boldsymbol{U}_t \in \mathbb{R}^{r\times d}$. Note that $\left\|\bar{\boldsymbol{S}}_t\right\| = \|\boldsymbol{S}_t\|$. For $(\mathrm{I}_{1,1})$, we have

$$\begin{aligned}
(\mathrm{I}_{1,1}) &\le \left\|\left(\boldsymbol{I}-\eta\bar{\boldsymbol{S}}_t\bar{\boldsymbol{S}}_t^\top\right)\bar{\boldsymbol{S}}_t\right\| + \eta\left\|\boldsymbol{V}_t^\top\boldsymbol{V}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t\bar{\boldsymbol{S}}_t\right\| \\
&\overset{(a)}{=} \left\|\bar{\boldsymbol{S}}_t\right\|\left(1-\eta\left\|\bar{\boldsymbol{S}}_t\right\|^2\right) + \eta\sigma_1^\star\left\|\bar{\boldsymbol{S}}_t\right\| \\
&= \|\boldsymbol{S}_t\|\left(1+\eta\sigma_1^\star-\eta\,\|\boldsymbol{S}_t\|^2\right).
\end{aligned} \tag{45}$$

where (a) follows from the fact that $\bar{\boldsymbol{S}}_t$ and $\bar{\boldsymbol{S}}_t\bar{\boldsymbol{S}}_t^\top$ share the same eigenvectors, and the assumption $\eta \lesssim 1/\sigma_1$. Next, we control $(\mathrm{I}_{1,2})$:

$$(\mathrm{I}_{1,2}) \le \sigma_1^\star\,\|\boldsymbol{V}^\star-\boldsymbol{V}_t\|\,\|\boldsymbol{S}_t\| \le 2\Gamma_1\frac{\kappa\mu(\sigma_1^\star r)^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}. \tag{46}$$

Here we use Lemma 6. Therefore, we can bound (I_1) by

$$(\mathrm{I}_1) \le \|\boldsymbol{S}_t\|\left(1+\eta\sigma_1^\star-\eta\,\|\boldsymbol{S}_t\|^2\right) + 2\Gamma_1\eta\frac{\kappa\mu(\sigma_1^\star r)^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}. \tag{47}$$

This leads to

$$\begin{aligned}\text{(I)} &\le \|\boldsymbol{S}_t\| \left(1+\eta\sigma_1^\star - \eta\|\boldsymbol{S}_t\|^2\right) + 2\Gamma_1\eta\frac{\kappa\mu(\sigma_1^\star r)^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} + 43\Gamma\eta\frac{\sigma_1^{\star 1.5}\mu r}{\sqrt{pd}} \\ &\le \|\boldsymbol{S}_t\| \left(1+\eta\sigma_1^\star - \eta\|\boldsymbol{S}_t\|^2\right) + 3\Gamma_1\eta\frac{\kappa\mu(\sigma_1^\star r)^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.\end{aligned} \tag{48}$$

Next, we control (II). To this end, we first notice that

$$\mathcal{P}_{\boldsymbol{V}_{t+1}}\left(\boldsymbol{I}+\eta\boldsymbol{M}_t\right)\mathcal{P}_{\boldsymbol{V}_t}^{\perp} = \left(\mathcal{P}_{\boldsymbol{V}_{t+1}} - \mathcal{P}_{\boldsymbol{V}_t}\right)\mathcal{P}_{\boldsymbol{V}_t}^{\perp} + \eta\mathcal{P}_{\boldsymbol{V}_{t+1}}\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^{\perp}. \tag{49}$$

Hence, we can bound (II) by

$$\begin{aligned}\text{(II)} &\le \left\|\mathcal{P}_{\boldsymbol{V}_{t+1}} - \mathcal{P}_{\boldsymbol{V}_t}\right\|\|\boldsymbol{E}_t\| + \eta\left\|\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\right\|\|\boldsymbol{E}_t\| \\ &\overset{(a)}{\le} \left(2\|\boldsymbol{V}_{t+1}-\boldsymbol{V}_t\| + 2\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)\|\boldsymbol{E}_t\|.\end{aligned} \tag{50}$$

Here we use Lemma 43 in (a). It remains to control $\|\boldsymbol{V}_{t+1}-\boldsymbol{V}_t\|$. To this end, upon noticing that $\boldsymbol{V}_{t+1}-\boldsymbol{V}_t = \eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\boldsymbol{V}_t + \boldsymbol{A}_t$, one has

$$\begin{aligned}\|\boldsymbol{V}_{t+1}-\boldsymbol{V}_t\| &\le \eta\left\|\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\right\| + \|\boldsymbol{A}_t\| \\ &\overset{(a)}{\le} 2\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} + 300\eta^2\sigma_1^{\star 2} \\ &\le 3\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.\end{aligned} \tag{51}$$

Here we apply Lemma 12 and Lemma 13 in (a). Hence, we have

$$\text{(II)} \le 5\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\|\boldsymbol{E}_t\|. \tag{52}$$

Putting everything together, we obtain that

$$\begin{aligned}\|\boldsymbol{S}_{t+1}\| &\le \|\boldsymbol{S}_t\|\left(1+\eta\sigma_1^\star-\eta\|\boldsymbol{S}_t\|^2\right) + 3\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\left(\sqrt{\sigma_1^\star}+2\|\boldsymbol{E}_t\|\right) \\ &\le \|\boldsymbol{S}_t\|\left(1+\eta\sigma_1^\star-\eta\|\boldsymbol{S}_t\|^2\right) + 6\Gamma_1\eta\sigma_1^{\star 1.5}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.\end{aligned} \tag{53}$$

Next, we consider two cases separately. First, if $\|\boldsymbol{S}_t\| \le 1.5\sqrt{\sigma_1^\star}$, then we simply have

$$\begin{aligned}\|\boldsymbol{S}_{t+1}\| &\le \|\boldsymbol{S}_t\|\left(1+\eta\sigma_1^\star-\eta\|\boldsymbol{S}_t\|^2\right) + 6\Gamma_1\eta\sigma_1^{\star 1.5}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \\ &\le 1.5\sqrt{\sigma_1^\star}\cdot(1+\eta\sigma_1^\star) + 6\Gamma_1\eta\sigma_1^{\star 1.5}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \\ &\le 2\sqrt{\sigma_1^\star}.\end{aligned} \tag{54}$$

On the other hand, if $1.5\sqrt{\sigma_1^\star} \le \|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$, then we have

$$\begin{aligned}\|\boldsymbol{S}_{t+1}\| &\le \|\boldsymbol{S}_t\|\left(1+\eta\sigma_1^\star-\eta\|\boldsymbol{S}_t\|^2\right)+6\Gamma_1\eta\sigma_1^{\star 1.5}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\\ &\le 2\sqrt{\sigma_1^\star}\left(1-1.25\eta\sigma_1^\star\right)+6\Gamma_1\eta\sigma_1^{\star 1.5}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\\ &\le 2\sqrt{\sigma_1^\star}.\end{aligned} \tag{55}$$

This completes the proof for the maximal signal dynamic.

Minimal signal dynamic. We first provide a lower-bound for $\sigma_r(\boldsymbol{S}_{t+1})$ as follows

$$\sigma_r(\boldsymbol{S}_{t+1}) \ge \underbrace{\sigma_r((\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{S}_t)}_{\text{(I)}} - \underbrace{\left\|\mathcal{P}_{\boldsymbol{V}_{t+1}}(\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{E}_t\right\|}_{\text{(II)}}. \tag{56}$$

For (I), applying Lemma 44, we first obtain that $\text{(I)} \ge \sigma_r\left(\boldsymbol{V}_t^\top(\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{S}_t\right)$. Next, we decompose $\boldsymbol{V}_t^\top(\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{S}_t$ as follows

$$\begin{aligned}&\boldsymbol{V}_t^\top(\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{S}_t\\ &=\underbrace{\left(\boldsymbol{I}+\eta\boldsymbol{V}_t^\top\left(\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\left(\boldsymbol{I}-\eta\boldsymbol{V}_t^\top\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{V}_t\right)^{-1}\right)}_{:=\boldsymbol{B}_t}\underbrace{\boldsymbol{V}_t^\top\boldsymbol{S}_t\left(\boldsymbol{I}-\eta\boldsymbol{S}_t^\top\boldsymbol{S}_t\right)}_{:=\boldsymbol{C}_t}.\end{aligned} \tag{57}$$

According to Lemma 45, we have

$$\sigma_r((\boldsymbol{I}+\eta\boldsymbol{M}_t)\boldsymbol{S}_t) \ge \sigma_r(\boldsymbol{B}_t)\sigma_r(\boldsymbol{C}_t) = \sigma_r(\boldsymbol{B}_t)\left(1-\eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t), \tag{58}$$

where in the last equality, we use the fact that $\boldsymbol{S}_t$ and $\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{S}_t$ share the same singular space and hence

$$\sigma_r(\boldsymbol{C}_t) = \sigma_r(\boldsymbol{S}_t(\boldsymbol{I}-\eta\boldsymbol{S}_t^\top\boldsymbol{S}_t)) = \left(1-\eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t). \tag{59}$$

Now, it suffices to provide a lower-bound for $\sigma_r(\boldsymbol{B}_t)$. To this end, we first notice that

$$\begin{aligned}\sigma_r(\boldsymbol{B}_t) &\ge 1+\eta\sigma_r\left(\boldsymbol{V}_t^\top\left(\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\right)\sigma_r\left(\left(\boldsymbol{I}-\eta\boldsymbol{V}_t^\top\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{V}_t\right)^{-1}\right)\\ &\ge 1+\eta\sigma_r\left(\boldsymbol{V}_t^\top\left(\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\right).\end{aligned} \tag{60}$$

Here the second inequality is due to $\boldsymbol{I}-\eta\boldsymbol{V}_t^\top\boldsymbol{S}_t\boldsymbol{S}_t^\top\boldsymbol{V}_t \preceq \boldsymbol{I}$. To proceed, notice that

$$\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top = \boldsymbol{X}^\star+(\mathcal{R}_\Omega-\mathcal{I})(\boldsymbol{\Delta}_t)-\boldsymbol{\Lambda}_t. \tag{61}$$

Therefore, we have

$$\begin{aligned}\sigma_r\left(\boldsymbol{V}_t^\top\left(\boldsymbol{M}_t+\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\right) &\ge \sigma_r\left(\boldsymbol{V}_t^\top\boldsymbol{X}^\star\boldsymbol{V}_t\right)-\|(\mathcal{R}_\Omega-\mathcal{I})(\boldsymbol{\Delta}_t)\|-\|\boldsymbol{\Lambda}_t\|\\ &\ge \sigma_r\left(\boldsymbol{V}_t^\top\boldsymbol{X}^\star\boldsymbol{V}_t\right)-21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2r^2}{pd}}-5\sqrt{\sigma_1^\star}\|\boldsymbol{E}_t\|\\ &\ge \sigma_r\left(\boldsymbol{V}_t^\top\boldsymbol{X}^\star\boldsymbol{V}_t\right)-22\Gamma\sigma_1^\star\sqrt{\frac{\mu^2r^2}{pd}}.\end{aligned} \tag{62}$$

For the first term in the above inequality, we have

$$\begin{aligned}\sigma_r\Big(\boldsymbol{V}_t^\top \boldsymbol{X}^\star \boldsymbol{V}_t\Big) &\geq \sigma_r\Big(\boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \boldsymbol{V}_t\Big) - \sigma_1^\star \left\|\boldsymbol{V}^\star - \boldsymbol{V}_t\right\| \\ &\geq \sigma_r^\star - 2\sigma_1^\star \left\|\boldsymbol{V}^\star - \boldsymbol{V}_t\right\| \\ &\geq \sigma_r^\star - 2\Gamma_1 \sigma_1^\star \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \\ &\geq 0.95\sigma_r^\star.\end{aligned} \tag{63}$$

Therefore, we obtain

$$\sigma_r\Big(\boldsymbol{V}_t^\top \left(\boldsymbol{M}_t + \boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\Big) \geq 0.95\sigma_r^\star - 22\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}} \geq 0.9\sigma_r^\star. \tag{64}$$

Combining the above arguments, we have

$$\text{(I)} \geq \left(1 + 0.9\eta\sigma_r^\star\right)\left(1 - \eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t) \geq \left(1 + 0.8\eta\sigma_r^\star - \eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t). \tag{65}$$

On the other hand, we have already derived an upper bound for (II) in the maximal signal dynamic, which is

$$\text{(II)} \leq 5\Gamma_1\eta\sigma_1^\star \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \left\|\boldsymbol{E}_t\right\|. \tag{66}$$

Putting everything together, we have

$$\sigma_r(\boldsymbol{S}_{t+1}) \geq \left(1 + 0.8\eta\sigma_r^\star - \eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t) - 5\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\left\|\boldsymbol{E}_t\right\|. \tag{67}$$

C.2. Proof of Residual Dynamic

First, we can expand $\boldsymbol{E}_{t+1}$ as below

$$\boldsymbol{E}_{t+1} = \mathcal{P}_{\boldsymbol{V}_{t+1}}^\perp\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\left(\boldsymbol{S}_t + \boldsymbol{E}_t\right) = \mathcal{P}_{\boldsymbol{V}_{t+1}}^\perp\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{E}_t, \tag{68}$$

where in the second equality, we use the fact that $\mathcal{P}_{\boldsymbol{V}_{t+1}}^\perp\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{S}_t = 0$. Then, by triangle inequality, we obtain

$$\left\|\boldsymbol{E}_{t+1}\right\| \leq \left(1 + \eta\left\|\boldsymbol{M}_t\mathcal{P}_{\boldsymbol{V}_t}^\perp\right\|\right)\left\|\boldsymbol{E}_t\right\| \leq \left(1 + 2\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)\left\|\boldsymbol{E}_t\right\|. \tag{69}$$

Here in the last inequality, we use Lemma 12.

C.3. Proof of Error Dynamic

The core proof idea is adapted from the proof of Proposition 4.3 appeared in (Li et al., 2018). We first expand $\left\|\boldsymbol{\Delta}_{t+1}\right\|_{\mathrm{F}}^2$ as

$$\begin{aligned}\left\|\boldsymbol{\Delta}_{t+1}\right\|_{\mathrm{F}}^2 &= \left\|\boldsymbol{X}^\star - \left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\boldsymbol{U}_t\boldsymbol{U}_t^\top\left(\boldsymbol{I} + \eta\boldsymbol{M}_t\right)\right\|_{\mathrm{F}}^2 \\ &= \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 4\eta\underbrace{\left\langle\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\rangle}_{\text{(I)}} + \text{(II)},\end{aligned} \tag{70}$$

where

$$\begin{aligned}(\mathrm{II}) = 2\left\langle -\boldsymbol{\Delta}_t + \eta \boldsymbol{M}_t \boldsymbol{U}_t \boldsymbol{U}_t^\top, \eta^2 \boldsymbol{M}_t \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{M}_t \right\rangle \\ + \eta^2 \left\| \boldsymbol{M}_t \boldsymbol{U}_t \boldsymbol{U}_t^\top + \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{M}_t \right\|_{\mathrm{F}}^2 + \eta^4 \left\| \boldsymbol{M}_t \boldsymbol{U}_t \boldsymbol{U}_t^\top \boldsymbol{M}_t \right\|_{\mathrm{F}}^2\end{aligned} \tag{71}$$

contains all the higher-order terms. We then provide a lower-bound for (I). To this goal, we notice that

$$\begin{aligned}(\mathrm{I}) &= \left\|\boldsymbol{\Delta}_t \boldsymbol{U}_t\right\|_{\mathrm{F}}^2 - \left\langle \boldsymbol{\Delta}_t, \left(\boldsymbol{I} - \mathcal{R}_\Omega\right)\left(\boldsymbol{\Delta}_t\right) \boldsymbol{U}_t \boldsymbol{U}_t^\top \right\rangle \\ &\geq \left\|\boldsymbol{\Delta}_t \boldsymbol{U}_t\right\|_{\mathrm{F}}^2 - \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \left\|\boldsymbol{U}_t \boldsymbol{U}_t^\top\right\|_{\mathrm{F}} \left\|\left(\mathcal{I} - \mathcal{R}_\Omega\right)\left(\boldsymbol{\Delta}_t\right)\right\| \\ &\geq \left\|\boldsymbol{\Delta}_t \boldsymbol{U}_t\right\|_{\mathrm{F}}^2 - \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \cdot 8\sqrt{r}\sigma_1^\star \cdot \left(5\Gamma \sqrt{\frac{\mu^2 r^2}{pd}} \left\|\boldsymbol{\Delta}_t\right\| + 10\Gamma \sqrt{\frac{\sigma_1^\star \mu r}{p}} \left\|\boldsymbol{E}_t\right\|\right) \\ &= \left\|\boldsymbol{\Delta}_t \boldsymbol{U}_t\right\|_{\mathrm{F}}^2 - 40\Gamma\sigma_1^\star \sqrt{\frac{\mu^2 r^3}{pd}} \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 80\Gamma \sqrt{\frac{\sigma_1^{\star 3} \mu r^2}{p}} \left\|\boldsymbol{E}_t\right\| \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}.\end{aligned} \tag{72}$$

Here in the last inequality, we apply Lemma 12. Next, we provide a lower-bound for $\left\|\boldsymbol{\Delta}_t \boldsymbol{U}_t\right\|_{\mathrm{F}}^2$. To this goal, we first notice that

$$\begin{aligned}\left\|\boldsymbol{\Delta}_t \boldsymbol{U}_t\right\|_{\mathrm{F}} &= \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t - \boldsymbol{\Lambda}_t \boldsymbol{S}_t + \boldsymbol{\Delta}_t \boldsymbol{E}_t\right\|_{\mathrm{F}} \\ &\geq \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t\right\|_{\mathrm{F}} - \left\|\boldsymbol{\Lambda}_t\right\| \left\|\boldsymbol{S}_t\right\|_{\mathrm{F}} - \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \left\|\boldsymbol{E}_t\right\| \\ &\geq \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t\right\|_{\mathrm{F}} - 10\sqrt{r}\sigma_1^\star \left\|\boldsymbol{E}_t\right\| - \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \left\|\boldsymbol{E}_t\right\|.\end{aligned} \tag{73}$$

In the last inequality we use the fact that $\left\|\boldsymbol{\Lambda}_t\right\| \leq 5\sqrt{\sigma_1^\star} \left\|\boldsymbol{E}_t\right\|$ and $\left\|\boldsymbol{S}_t\right\|_{\mathrm{F}} \leq \sqrt{r} \left\|\boldsymbol{S}_t\right\| \leq 2\sqrt{r\sigma_1^\star}$ from Lemma 11. Applying Lemma 35, we can further provide a lower-bound for $\left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t\right\|_{\mathrm{F}}$ as

$$\left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t\right\|_{\mathrm{F}} \geq \sigma_r(\boldsymbol{S}_t) \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{V}_t\right\|_{\mathrm{F}} \geq \frac{\sqrt{\sigma_r^\star}}{2} \left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{V}_t\right\|_{\mathrm{F}}. \tag{74}$$

Next, we present the following intermediate lemma to control $\left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{V}_t\right\|_{\mathrm{F}}$.

Lemma 15 *Suppose that* $\left\|\boldsymbol{V}^\star - \boldsymbol{V}_t\right\| \leq 0.1$. *Then, we have*

$$\left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{V}_t\right\|_{\mathrm{F}}^2 \geq \frac{2}{5} \left\|\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right\|_{\mathrm{F}}^2. \tag{75}$$

We first use this lemma to finish the proof of the loss dynamic and defer the proof to the end of this section. Applying this lemma to Equation (74) yields

$$\begin{aligned}\left\|\left(\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right) \boldsymbol{S}_t\right\|_{\mathrm{F}} &\geq \sqrt{\frac{\sigma_r^\star}{10}} \left\|\boldsymbol{X}^\star - \boldsymbol{S}_t \boldsymbol{S}_t^\top\right\|_{\mathrm{F}} \\ &\geq \sqrt{\frac{\sigma_r^\star}{10}} \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} - \sqrt{\frac{\sigma_r^\star}{10}} \left\|\boldsymbol{\Lambda}_t\right\|_{\mathrm{F}} \\ &\geq \sqrt{\frac{\sigma_r^\star}{10}} \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} - \sigma_1^\star \sqrt{\frac{5}{2\kappa}} \left\|\boldsymbol{E}_t\right\|.\end{aligned} \tag{76}$$

Combining Equation (76) and Equation (73) leads to

$$\begin{aligned}\left\|\boldsymbol{\Delta}_t\boldsymbol{U}_t\right\|_{\mathrm{F}} &\geq \sqrt{\frac{\sigma_r^\star}{10}}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} - \sigma_1^\star\sqrt{\frac{5}{2\kappa}}\left\|\boldsymbol{E}_t\right\| - 10\sqrt{r}\sigma_1^\star\left\|\boldsymbol{E}_t\right\| - \left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left\|\boldsymbol{E}_t\right\| \\ &\geq \sqrt{\frac{\sigma_r^\star}{10}}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} - 20\sqrt{r}\sigma_1^\star\left\|\boldsymbol{E}_t\right\|.\end{aligned} \tag{77}$$

This implies that

$$\left\|\boldsymbol{\Delta}_t\boldsymbol{U}_t\right\|_{\mathrm{F}}^2 \geq \frac{\sigma_r^\star}{10}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 13\sqrt{\frac{r}{\kappa}}\sigma_1^{\star 1.5}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left\|\boldsymbol{E}_t\right\|. \tag{78}$$

Overall, we obtain

$$\begin{aligned}\text{(I)} &\geq \frac{\sigma_r^\star}{10}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 13\sqrt{\frac{r}{\kappa}}\sigma_1^{\star 1.5}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left\|\boldsymbol{E}_t\right\| - 40\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^3}{pd}}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 80C\sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}}\left\|\boldsymbol{E}_t\right\|\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}} \\ &\geq \frac{\sigma_r^\star}{15}\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 - 81\Gamma\sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}}\left\|\boldsymbol{E}_t\right\|\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}.\end{aligned} \tag{79}$$

Next, we control (II). To this end, we first notice that

$$\begin{aligned}-2\left\langle\boldsymbol{\Delta}_t, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\rangle &\leq 2\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left\|\boldsymbol{M}_t\right\|^2\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}} \\ &\leq 16\sqrt{r}\sigma_1^\star\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left(2\left\|\boldsymbol{\Delta}_t\right\| + 10\Gamma\sqrt{\frac{\sigma_1^\star\mu r}{p}}\left\|\boldsymbol{E}_t\right\|\right)^2 \\ &\leq 128\sqrt{r}\sigma_1^\star\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}\left(\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 + 25\Gamma^2\frac{\sigma_1^\star\mu r}{p}\left\|\boldsymbol{E}_t\right\|^2\right).\end{aligned} \tag{80}$$

Similarly, we have

$$\begin{aligned}\left\langle\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top, \boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\rangle &\leq \left\|\boldsymbol{M}_t\right\|^3\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}}^2 \\ &\leq C_1 r\sigma_1^{\star 2}\left(\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^3 + \frac{(\sigma_1^\star\mu r)^{1.5}}{p^{1.5}}\left\|\boldsymbol{E}_t\right\|^3\right),\end{aligned} \tag{81}$$

$$\begin{aligned}\left\|\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top + \boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\|_{\mathrm{F}}^2 &\leq 4\left\|\boldsymbol{M}_t\right\|^2\left\|\boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}}^2 \\ &\leq C_2 r\sigma_1^{\star 2}\left(\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 + \frac{\sigma_1^\star\mu r}{p}\left\|\boldsymbol{E}_t\right\|^2\right).\end{aligned} \tag{82}$$

$$\left\|\boldsymbol{M}_t\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{M}_t\right\|_{\mathrm{F}}^2 \leq C_3 r\sigma_1^{\star 2}\left(\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^4 + \frac{\sigma_1^{\star 2}\mu^2 r^2}{p^2}\left\|\boldsymbol{E}_t\right\|^4\right). \tag{83}$$

These inequalities lead to

$$\text{(II)} \leq C_4\eta^2 r\sigma_1^{\star 2}\left(\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2 + \Gamma^2\frac{\sigma_1^\star\mu r}{p}\left\|\boldsymbol{E}_t\right\|^2\right). \tag{84}$$

Overall, we have

$$\begin{aligned}\left\|\boldsymbol{\Delta}_{t+1}\right\|_{\mathrm{F}}^2 &\leq \left(1-\frac{1}{5}\eta\sigma_r^\star\right)\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}^2+324\Gamma\eta\sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}}\left\|\boldsymbol{E}_t\right\|\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}+C_5\eta^2 r\frac{\sigma_1^{\star 3}\mu r}{p}\left\|\boldsymbol{E}_t\right\|^2 \\ &\leq \left(\left(1-\frac{1}{10}\eta\sigma_r^\star\right)\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}+C_6\eta\sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}}\left\|\boldsymbol{E}_t\right\|\right)^2,\end{aligned} \tag{85}$$

which implies that

$$\left\|\boldsymbol{\Delta}_{t+1}\right\|_{\mathrm{F}} \leq \left(1-\frac{1}{10}\eta\sigma_r^\star\right)\left\|\boldsymbol{\Delta}_t\right\|_{\mathrm{F}}+C_6\eta\sqrt{\frac{\sigma_1^{\star 3}\mu r^2}{p}}\left\|\boldsymbol{E}_t\right\|. \tag{86}$$

Lastly, we provide the proof of Lemma 15.

Proof of Lemma 15. First, we define $\boldsymbol{P}=\boldsymbol{V}^{\star\top}\boldsymbol{V}_t$ and note that $\boldsymbol{S}_t\boldsymbol{S}_t^\top=\boldsymbol{V}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^\top$. This allows us to write

$$\begin{aligned}\left\|\left(\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right)\boldsymbol{V}_t\right\|_{\mathrm{F}}^2 &= \left\|\boldsymbol{\Sigma}_t\right\|_{\mathrm{F}}^2+\left\|\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\|_{\mathrm{F}}^2-2\left\langle\boldsymbol{\Sigma}_t,\boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\rangle, \\ \left\|\boldsymbol{X}^\star-\boldsymbol{S}_t\boldsymbol{S}_t^\top\right\|_{\mathrm{F}}^2 &= \left\|\boldsymbol{\Sigma}_t\right\|_{\mathrm{F}}^2+\left\|\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}^2-2\left\langle\boldsymbol{\Sigma}_t,\boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\rangle.\end{aligned} \tag{87}$$

Substituting the above equivalent forms into Equation (75), we need to show that

$$3\left\|\boldsymbol{\Sigma}_t\right\|_{\mathrm{F}}^2+5\left\|\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\|_{\mathrm{F}}^2 \geq 2\left\|\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}^2+6\left\langle\boldsymbol{\Sigma}_t,\boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\rangle. \tag{88}$$

To this end, we first apply the Cauchy-Schwartz inequality, which gives us $2\left\langle\boldsymbol{\Sigma}_t,\boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\rangle \leq \left\|\boldsymbol{\Sigma}_t\right\|_{\mathrm{F}}^2+\left\|\boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\|_{\mathrm{F}}^2$. Therefore, it suffices to show that

$$5\left\|\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\|_{\mathrm{F}}^2-2\left\|\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}^2-3\left\|\boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\|_{\mathrm{F}}^2 \geq 0. \tag{89}$$

This follows from

$$5\left\|\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\|_{\mathrm{F}}^2-2\left\|\boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}^2-3\left\|\boldsymbol{P}^\top\boldsymbol{\Sigma}^\star\boldsymbol{P}\right\|_{\mathrm{F}}^2=\operatorname{tr}\left(\left(\boldsymbol{\Sigma}^\star\left(\boldsymbol{I}-\boldsymbol{P}\boldsymbol{P}^\top\right)\boldsymbol{\Sigma}^\star\right)\cdot\left(3\boldsymbol{P}\boldsymbol{P}^\top-2\boldsymbol{I}\right)\right)\geq 0. \tag{90}$$

Here we use the facts that $\boldsymbol{\Sigma}^\star\left(\boldsymbol{I}-\boldsymbol{P}\boldsymbol{P}^\top\right)\boldsymbol{\Sigma}^\star \succeq 0$ since $\|\boldsymbol{P}\| \leq \|\boldsymbol{V}^\star\|\,\|\boldsymbol{V}_t\| \leq 1$, and $3\boldsymbol{P}\boldsymbol{P}^\top-2\boldsymbol{I} \succeq 0$ since $\sigma_r(\boldsymbol{P}) \geq 1-\|\boldsymbol{V}^\star-\boldsymbol{V}_t\| \geq 0.9$. This completes the proof. ■

Appendix D. Proofs for Main Theorems

In this section, we use the one-step dynamics in Proposition 14 to prove our main theorems under the conditions that $\|\boldsymbol{V}_t\|_{2,\infty} \leq 2\sqrt{\frac{\mu r}{d}}$ and $\|\boldsymbol{V}^\star-\boldsymbol{V}_t\|_{\mathrm{F}} \leq \Gamma_1\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}$ for all $0 \leq t \leq T$. These two conditions will be established later in Appendix E.

D.1. Proof of Theorem 2

The proof is divided into three distinct steps.

Step 1. In the first step, we show that $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \leq 2\alpha$ hold for all $0 \leq t \leq T$.

We prove this by induction. First, in the base case where $t = 0$, these two conditions are naturally met because $\|\boldsymbol{S}_0\| \leq \|\boldsymbol{U}_0\| \leq \alpha \leq 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_0\| \leq 2\|\boldsymbol{E}_0\| \leq 2\alpha$. Next, for the induction step, we assume that $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \leq 2\alpha$ hold for all $0 \leq s \leq t$, with $t \leq T-1$. Utilizing Proposition 14, we can directly derive that $\|\boldsymbol{S}_{t+1}\| \leq 2\sqrt{\sigma_1^\star}$. Regarding $\|\boldsymbol{E}_{t+1}\|$, we have

$$\|\boldsymbol{E}_{t+1}\| \leq \left(1 + 2\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)^{t+1}\alpha \overset{(a)}{\leq} \left(1 + 4\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\cdot(t+1)\right)\alpha \leq 2\alpha. \tag{91}$$

Here in (a), we apply Lemma 46. This is valid since $4\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\cdot(t+1) \leq 1$ for any $t \leq T-1 \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$ provided that the sampling rate satisfies $p \gtrsim \frac{\kappa^4\mu^2 r^3\log^4\left(\frac{1}{\alpha}\right)}{d}$. This completes the induction step.

Step 2. This step demonstrates that the minimal signal $\sigma_r(\boldsymbol{S}_t)$ grows linearly to $\frac{\sqrt{\sigma_r^\star}}{2}$.

Given that we have already established $\|\boldsymbol{S}_t\| \leq 2\sqrt{\sigma_1^\star}$ and $\|\boldsymbol{E}_t\| \leq 2\alpha$ for all $0 \leq t \leq T$, we can simplify the minimal signal dynamic in Proposition 14 as

$$\begin{aligned}\sigma_r(\boldsymbol{S}_{t+1}) &\geq \left(1 + 0.8\eta\sigma_r^\star - \eta\sigma_r^2(\boldsymbol{S}_t)\right)\sigma_r(\boldsymbol{S}_t) - 12\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\alpha \\ &\geq \left(1 + 0.4\eta\sigma_r^\star\right)\sigma_r(\boldsymbol{S}_t) - 12\Gamma_1\eta\frac{\sigma_1^\star\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\alpha.\end{aligned} \tag{92}$$

This holds for any t that satisfies $\sigma_r(\boldsymbol{S}_t) \leq \frac{3\sqrt{\sigma_r^\star}}{4}$. By applying Lemma 47, we obtain

$$\sigma_r(\boldsymbol{S}_t) \geq \left(1 + 0.4\eta\sigma_r^\star\right)^t\left(\sigma_r(\boldsymbol{S}_t) - 30\Gamma_1\frac{\kappa^2\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\alpha\right). \tag{93}$$

At initialization, it is observed that

$$\sigma_r(\boldsymbol{S}_0) - 30\Gamma_1\frac{\kappa^2\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\alpha \geq c_0\alpha - 30\Gamma_1\frac{\kappa^2\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\alpha \geq \frac{c_0}{2}\alpha, \tag{94}$$

provided that $p \gtrsim \frac{\kappa^4\mu^2 r^3\log^2\left(\frac{1}{\alpha}\right)}{d}$. Consequently, within $T_1 \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{\sigma_r^\star}{\alpha}\right)$ iterations, $\sigma_r(\boldsymbol{S}_t)$ reaches $\frac{\sqrt{\sigma_r^\star}}{2}$. It is also easy to show that $\sigma_r(\boldsymbol{S}_t) \geq \frac{\sqrt{\sigma_r^\star}}{2}$ holds true for all $t \geq T_1$.

Step 3. This step is dedicated to demonstrating that the error $\left\|\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}}$ converges linearly to $O(\alpha)$ once $\sigma_r(\boldsymbol{S}_t) \geq \frac{\sqrt{\sigma_r^\star}}{2}$.

Based on our arguments in **Step 2**, where we established that $\sigma_r(\boldsymbol{S}_t) \geq \frac{\sqrt{\sigma_r^\star}}{2}$ for $t \geq T_1$, and leveraging Proposition 14 along with Lemma 47, we can derive that

$$\left\|\boldsymbol{X}^\star - \boldsymbol{U}_t\boldsymbol{U}_t^\top\right\|_{\mathrm{F}} \leq \left(1 - \frac{1}{5}\eta\sigma_r^\star\right)^{t-T_1}\left\|\boldsymbol{X}^\star - \boldsymbol{U}_{T_1}\boldsymbol{U}_{T_1}^\top\right\|_{\mathrm{F}} + 3240\Gamma\sqrt{\frac{\sigma_1^\star\kappa^2\mu r^2}{p}}\alpha. \tag{95}$$

Note that $\left\|\boldsymbol{X}^\star - \boldsymbol{U}_{T_1}\boldsymbol{U}_{T_1}^\top\right\|_{\mathrm{F}} \leq \|\boldsymbol{X}^\star\|_{\mathrm{F}} + \left\|\boldsymbol{U}_{T_1}\boldsymbol{U}_{T_1}^\top\right\|_{\mathrm{F}} \leq 9\sqrt{r}\sigma_1^\star$ according to Lemma 10. Hence, within an additional $T_2 = \frac{1}{\eta\sigma_r^\star}\log\left(\frac{r\sigma_1^\star}{\alpha}\right)$ iterations, the error converges to $O\left(\sqrt{\frac{\sigma_1^\star\kappa^2\mu r^2}{p}}\alpha\right)$, thus concluding the proof of Theorem 2.

D.2. Proof of Theorem 3

To prove this result, we first apply Theorem 2 to output a solution $\boldsymbol{U}_{t_0}$ and its leave-one-out versions $\boldsymbol{U}_{t_0}^{(l)}$ that meet the initialization conditions in Theorem 9. Then, we apply Theorem 9 to obtain the desired result.

Establishing Condition (14). By choosing the initialization scale $\alpha = c \cdot \frac{\sigma_r^\star}{\kappa^{1.5} d}$ for sufficiently small $c > 0$ and assuming the sampling rate of $p \gtrsim \frac{\kappa^2 \mu^4 r^9 \log^4(d)}{d}$, Theorem 2 guarantees that, with probability at least $1 - \frac{1}{d^2}$, the iterations of GD with step-size $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$ satisfy

$$\left\|\boldsymbol{U}_{t_0}\boldsymbol{U}_{t_0}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \le 0.9\Gamma_4 \sqrt{\frac{\sigma_1^\star \kappa^2 \mu r^2}{p}} \frac{\sigma_r^\star}{\kappa^{1.5} d}, \quad \text{for some} \quad t_0 \lesssim \frac{1}{\eta \sigma_r^\star} \log\left(\frac{\kappa^{1.5} d}{\sigma_r^\star}\right). \tag{96}$$

On the other hand, Lemma 40 in the appendix implies that

$$\operatorname{dist}\left(\boldsymbol{U}_{t_0}, \boldsymbol{U}^\star\right) \le \frac{1.1}{\sigma_r^\star}\left\|\boldsymbol{U}_{t_0}\boldsymbol{U}_{t_0}^\top - \boldsymbol{X}^\star\right\|_{\mathrm{F}} \le \Gamma_4 \sqrt{\frac{\sigma_r^\star \mu^3 r^3 \log(d)}{pd^2}},$$

which establishes Condition (14).

Establishing Condition (15). The proof of Condition (15) follows a similar logic, recognizing that the leave-one-out sequences exhibit a stronger concentration than the original iterations. Consequently, they fulfill Condition (15) within at most $t_0 \lesssim \frac{1}{\eta\sigma_r^\star} \log\left(\frac{\kappa^{1.5} d}{\sigma_r^\star}\right)$ iterations. Further details of this argument are omitted for brevity.

This shows that the initial conditions of Theorem 9 are satisfied after t_0 iterations. From this iteration onward, Theorem 9 shows that the iterations of GD enter a local linear convergence regime, which readily establishes the final result of Theorem 3.

Appendix E. Proofs for Incoherence Dynamic

In this section, we present our proofs for establishing the incoherence of $\boldsymbol{V}_t$. To simplify the presentation, we omit the "$\sim$" from our notations. Therefore, $\boldsymbol{V}_t^{(l)}, \boldsymbol{Z}_t^{(l)}, \boldsymbol{M}_t^{(l)}, \ldots$ in this section refer to $\widetilde{\boldsymbol{V}}_t^{(l)}, \widetilde{\boldsymbol{Z}}_t^{(l)}, \widetilde{\boldsymbol{M}}_t^{(l)}, \ldots$ defined in Section 3.2.

E.1. Proof of Lemma 6

We first state a finer variant of this lemma here.

Proposition 16 (Controlling $\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_{\mathrm{F}}$) *Suppose that the stepsize satisfies $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$. Moreover, suppose that $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$, $\|\boldsymbol{E}_t\| \le \sqrt{\frac{\sigma_1^\star}{d}}$, $\|\boldsymbol{V}_t\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$ and $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}} \le \Gamma_1 \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}$ for all $t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$. Then, conditioned on E_{good}, we have*

$$\|\boldsymbol{V}_t - \boldsymbol{V}^\star\|_{\mathrm{F}} \le \Gamma_1 \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \quad \forall t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right). \tag{97}$$

Proof First notice that $\boldsymbol{V}_{t+1}$ can be rewritten as

$$\begin{aligned}\boldsymbol{V}_{t+1} &= \boldsymbol{Z}_{t+1}\big(\boldsymbol{Z}_{t+1}^\top \boldsymbol{Z}_{t+1}\big)^{-1/2} \\ &= (\boldsymbol{I}+\eta\boldsymbol{M}_t)\,\boldsymbol{V}_t\big(\boldsymbol{V}_t^\top\,(\boldsymbol{I}+\eta\boldsymbol{M}_t)^2\,\boldsymbol{V}_t\big)^{-1/2} \\ &= (\boldsymbol{I}+\eta\boldsymbol{M}_t)\,\boldsymbol{V}_t\,(\boldsymbol{I}+\boldsymbol{Y}_t)^{-1/2}\end{aligned} \tag{98}$$

where we denote $\boldsymbol{Y}_t = \boldsymbol{V}_t^\top\left(2\eta\boldsymbol{M}_t+\eta^2\boldsymbol{M}_t^2\right)\boldsymbol{V}_t$. Next, we apply Taylor expansion for the matrix-valued function $f(\boldsymbol{X}) = (\boldsymbol{I}+\boldsymbol{X})^{-1/2}$, which states that for any $\boldsymbol{X}$ satisfying $\|\boldsymbol{X}\|<1$,

$$f(\boldsymbol{X}) = (\boldsymbol{I}+\boldsymbol{X})^{-1/2} = \boldsymbol{I}-\frac{1}{2}\boldsymbol{X}+\boldsymbol{R}(\boldsymbol{X}) \text{ where } \boldsymbol{R}(\boldsymbol{X}) = \sum_{k=2}^{\infty}\frac{(-1)^k(2k)!}{4^k(k!)^2}\boldsymbol{X}^k. \tag{99}$$

Then, upon setting $\boldsymbol{X}=\boldsymbol{Y}_t$ in the above equation and plugging it into Equation (98) and rearranging the subterms, we have

$$\boldsymbol{V}_{t+1} = (\boldsymbol{I}+\eta\boldsymbol{M}_t)\,\boldsymbol{V}_t\left(\boldsymbol{I}-\frac{1}{2}\boldsymbol{V}_t^\top\left(2\eta\boldsymbol{M}_t+\eta^2\boldsymbol{M}_t^2\right)\boldsymbol{V}_t+\boldsymbol{R}(\boldsymbol{Y}_t)\right) = \left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\right)\boldsymbol{V}_t+\boldsymbol{A}_t, \tag{100}$$

where

$$\boldsymbol{A}_t = -\eta^2\boldsymbol{M}_t\boldsymbol{V}_t\boldsymbol{V}_t^\top\boldsymbol{M}_t\boldsymbol{V}_t - 0.5\eta^2\boldsymbol{V}_t\boldsymbol{V}_t^\top\boldsymbol{M}_t^2\boldsymbol{V}_t - 0.5\eta^3\boldsymbol{M}_t\boldsymbol{V}_t\boldsymbol{V}^{\star\top}\boldsymbol{M}_t^2\boldsymbol{V}_t + (\boldsymbol{I}+\eta\boldsymbol{M}_t)\,\boldsymbol{V}_t\boldsymbol{R}(\boldsymbol{Y}_t) \tag{101}$$

contains all the higher-order terms. Next, according to triangle inequality, we can provide an upper bound for $\|\boldsymbol{V}^\star-\boldsymbol{V}_{t+1}\|_{\mathrm{F}}$ as follows

$$\|\boldsymbol{V}^\star-\boldsymbol{V}_{t+1}\|_{\mathrm{F}} \le \underbrace{\left\|\boldsymbol{V}^\star-\left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\right)\boldsymbol{V}_t\right\|_{\mathrm{F}}}_{:=(\mathrm{I})} + \|\boldsymbol{A}_t\|_{\mathrm{F}}\,. \tag{102}$$

We first control the leading term (I). To this goal, we apply triangle inequality and obtain

$$\begin{aligned}(\mathrm{I}) &\le \left\|\boldsymbol{V}^\star-\left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{X}^\star-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\right)\boldsymbol{V}_t\right\|_{\mathrm{F}} + \eta\left\|\left(\mathcal{I}-\mathcal{R}_\Omega\right)\left(\boldsymbol{X}^\star-\boldsymbol{U}_t\boldsymbol{U}_t^\top\right)\boldsymbol{V}_t\right\|_{\mathrm{F}} \\ &\le \underbrace{\left\|\boldsymbol{V}^\star-\left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{\Delta}_t\right)\boldsymbol{V}_t\right\|_{\mathrm{F}}}_{:=(\mathrm{I}_1)} + \eta\sqrt{r}\,\underbrace{\left\|\left(\mathcal{I}-\mathcal{R}_\Omega\right)(\boldsymbol{\Delta}_t)\right\|}_{:=(\mathrm{I}_2)}.\end{aligned} \tag{103}$$

To control (I_1), we further decompose it as

$$\begin{aligned}(\mathrm{I}_1) &\le \left\|\boldsymbol{V}^\star-\left(\boldsymbol{I}+\eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{V}^\star\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\right)\boldsymbol{V}_t\right\|_{\mathrm{F}} + \eta\left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{U}_t\boldsymbol{U}_t^\top\boldsymbol{V}_t\right\|_{\mathrm{F}} \\ &\overset{(a)}{\le} \left\|\boldsymbol{V}^\star-\boldsymbol{V}_t-\eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}^\star-\boldsymbol{V}_t\right)\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t\right\|_{\mathrm{F}} + \eta\sqrt{r}\left\|\boldsymbol{E}_t\boldsymbol{U}_t^\top\boldsymbol{V}_t\right\| \\ &\overset{(b)}{\le} \underbrace{\left\|\boldsymbol{V}^\star-\boldsymbol{V}_t-\eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}^\star-\boldsymbol{V}_t\right)\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t\right\|_{\mathrm{F}}}_{:=(\mathrm{I}_{1,1})} + 2\eta\sqrt{r\sigma_1^\star}\,\|\boldsymbol{E}_t\|\,.\end{aligned} \tag{104}$$

Here in (a), we use the fact that $\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{V}_t = 0$ and the definition $\boldsymbol{E}_t = \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{U}_t$. In (b), we use the assumption that $\|\boldsymbol{S}_t\| \le 2\sqrt{\sigma_1^\star}$. Next, according to the orthogonality of $\boldsymbol{V}_t$ and $\boldsymbol{V}_t^\perp$, we can upper-bound $(\mathrm{I}_{1,1})$ as follows

$$\begin{aligned}
(\mathrm{I}_{1,1})^2 &= \|\mathcal{P}_{\boldsymbol{V}_t}(\boldsymbol{V}^\star - \boldsymbol{V}_t)\|_{\mathrm{F}}^2 + \left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}(\boldsymbol{V}^\star - \boldsymbol{V}_t)\left(\boldsymbol{I} - \eta\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t\right)\right\|_{\mathrm{F}}^2 \\
&\le \|\mathcal{P}_{\boldsymbol{V}_t}(\boldsymbol{V}^\star - \boldsymbol{V}_t)\|_{\mathrm{F}}^2 + \left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}(\boldsymbol{V}^\star - \boldsymbol{V}_t)\right\|_{\mathrm{F}}^2 \left\|\boldsymbol{I} - \eta\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t\right\|^2 \\
&\overset{(a)}{\le} \|\mathcal{P}_{\boldsymbol{V}_t}(\boldsymbol{V}^\star - \boldsymbol{V}_t)\|_{\mathrm{F}}^2 + \left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}(\boldsymbol{V}^\star - \boldsymbol{V}_t)\right\|_{\mathrm{F}}^2 \\
&= \|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}}^2 .
\end{aligned} \tag{105}$$

Here (a) is due to the fact that $\left\|\boldsymbol{I} - \eta\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_t\right\| \le \|\boldsymbol{I} - \eta\boldsymbol{\Sigma}^\star\| + \eta\sigma_1^\star\|\boldsymbol{V}^\star - \boldsymbol{V}_t\| \le 1 - \eta(\sigma_r^\star - \sigma_1^\star\|\boldsymbol{V}^\star - \boldsymbol{V}_t\|) \le 1$ since we assume $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\| \le \frac{1}{2\kappa}$. Therefore, we derive that

$$(\mathrm{I}_1) \le \|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}} + 2\eta\sqrt{r\sigma_1^\star}\,\|\boldsymbol{E}_t\| . \tag{106}$$

On the other hand, Lemma 12 tells us that, conditioned on E_{good}, we have $(\mathrm{I}_2) \le 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}$. Therefore, we can conclude that

$$(\mathrm{I}) \le \|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}} + 2\eta\sqrt{r\sigma_1^\star}\,\|\boldsymbol{E}_t\| + \eta\sqrt{r}\cdot 21\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}} \le \|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}} + 22\Gamma\eta\frac{\sigma_1^\star\mu r^{1.5}}{\sqrt{pd}}. \tag{107}$$

Next, according to Lemma 13, we can control $\|\boldsymbol{A}_t\|_{\mathrm{F}}$ as

$$\|\boldsymbol{A}_t\|_{\mathrm{F}} \le \sqrt{r}\,\|\boldsymbol{A}_t\| \le 300\sqrt{r}\eta^2\sigma_1^{\star 2}. \tag{108}$$

Putting everything together, we have

$$\begin{aligned}
\|\boldsymbol{V}^\star - \boldsymbol{V}_{t+1}\|_{\mathrm{F}} &\le \|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}} + 22\Gamma\eta\frac{\sigma_1^\star\mu r^{1.5}}{\sqrt{pd}} + 300\sqrt{r}\eta^2\sigma_1^{\star 2} \\
&\le \|\boldsymbol{V}^\star - \boldsymbol{V}_t\|_{\mathrm{F}} + 23\Gamma\eta\frac{\sigma_1^\star\mu r^{1.5}}{\sqrt{pd}},
\end{aligned} \tag{109}$$

provided that $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$. This completes the proof. ■

E.2. Proof of Proposition 7

We restate the proposition here for clarity.

Proposition 17 (Dynamic of $\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|$) *Under the same conditions as Proposition 16, for all $t \le T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$, we have*

$$\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_{t+1}^{(l)}\right)_{l,\cdot}\right\| \le (1 - 0.5\eta\sigma_r^\star)\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| + \Gamma_2\eta\sigma_1^\star\frac{\kappa\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}}. \tag{110}$$

Proof Similar to Equation (100), we can express $\boldsymbol{V}_{t+1}^{(l)}$ as

$$\boldsymbol{V}_{t+1}^{(l)} = \left(\boldsymbol{I} + \eta \mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} \boldsymbol{M}_t^{(l)}\right) \boldsymbol{V}_t^{(l)} + \boldsymbol{A}_t^{(l)}, \tag{111}$$

where $\boldsymbol{M}_t^{(l)} = \mathcal{R}_{\Omega^{(l)}}\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top}\right)$ and $\boldsymbol{A}_t^{(l)}$ is defined as

$$\begin{aligned}\boldsymbol{A}_t^{(l)} = &-\eta^2 \boldsymbol{M}_t^{(l)} \boldsymbol{V}_t^{(l)} \boldsymbol{V}_t^{(l)\top} \boldsymbol{M}_t^{(l)} \boldsymbol{V}_t^{(l)} - 0.5\eta^2 \boldsymbol{V}_t^{(l)} \boldsymbol{V}_t^{(l)\top} \boldsymbol{M}_t^{(l)2} \boldsymbol{V}_t^{(l)} \\ &- 0.5\eta^3 \boldsymbol{M}_t^{(l)} \boldsymbol{V}_t^{(l)} \boldsymbol{V}^{(l)\top} \boldsymbol{M}_t^{(l)2} \boldsymbol{V}_t^{(l)} + (\boldsymbol{I} + \eta \boldsymbol{M}_t)\, \boldsymbol{V}_t^{(l)} \boldsymbol{R}\left(\boldsymbol{Y}_t^{(l)}\right)\end{aligned} \tag{112}$$

containing all the higher-order terms. Applying triangle inequality yields

$$\begin{aligned}\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_{t+1}^{(l)}\right)_{l,\cdot}\right\| &\le \left\|\left(\boldsymbol{V}^\star - \left(\boldsymbol{I} + \eta \mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} \boldsymbol{M}_t^{(l)}\right) \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| + \left\|\left(\boldsymbol{A}_t^{(l)}\right)_{l,\cdot}\right\| \\ &= \underbrace{\left\|\left(\boldsymbol{V}^\star - \left(\boldsymbol{I} + \eta \mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} \boldsymbol{\Xi}_t^{(l)}\right) \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|}_{:=(\mathrm{I})} + \left\|\left(\boldsymbol{A}_t^{(l)}\right)_{l,\cdot}\right\|.\end{aligned} \tag{113}$$

Here in the last equality, we use the fact that $\mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} \boldsymbol{M}_t^{(l)} = \mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} \boldsymbol{\Xi}_t^{(l)}$ where $\boldsymbol{\Xi}_t^{(l)} = \boldsymbol{M}_t^{(l)} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} + \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top}$. Upon noticing that $\mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} = \boldsymbol{I} - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}$, we further decompose (I) as follows

$$(\mathrm{I}) \le \underbrace{\left\|\left(\boldsymbol{V}^\star - \left(\boldsymbol{I} + \eta \boldsymbol{\Xi}_t^{(l)}\right) \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|}_{:=(\mathrm{I}_1)} + \eta \underbrace{\left\|\left(\mathcal{P}_{\boldsymbol{V}_t^{(l)}} \boldsymbol{\Xi}_t^{(l)} \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|}_{:=(\mathrm{I}_2)}. \tag{114}$$

To control (I_1), notice that the l-th row of $\boldsymbol{M}_t^{(l)}$ is equal to the l-th row of $\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top}$ due to our choice of $\mathcal{R}_{\Omega^{(l)}}$. Therefore, we have

$$\begin{aligned}(\mathrm{I}_1) &= \left\|\left(\boldsymbol{V}^\star - \left(\boldsymbol{I} + \eta\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top}\right)\right) \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \\ &= \left\|\left(\boldsymbol{V}^\star - \left(\boldsymbol{I} + \eta\left(\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right) \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top}\right)\right) \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \\ &= \left\|\left(\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)\left(\boldsymbol{I} - \eta \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \boldsymbol{V}_t^{(l)}\right)\right)_{l,\cdot}\right\| \\ &\le \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \left\|\boldsymbol{I} - \eta \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \boldsymbol{V}_t^{(l)}\right\| \\ &\overset{(a)}{\le} (1 - 0.5\eta\sigma_r^\star) \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|.\end{aligned} \tag{115}$$

Here in (a), we use the fact that $\left\|\boldsymbol{I} - \eta \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} \boldsymbol{V}_t^{(l)}\right\| \le \|\boldsymbol{I} - \eta \boldsymbol{\Sigma}^\star\| + \eta\sigma_1^\star \left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\| \le 1 - \eta(\sigma_r^\star - \sigma_1^\star \left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\|) \le 1 - 0.5\eta\sigma_r^\star$ since we have $\left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\| \le \frac{1}{2\kappa}$ according to the following proposition.

Lemma 18 (Frobenius norm control) *Under the same conditions as Proposition 16, for any* $1 \le t \le T = \frac{100}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$*, for all* $1 \le l \le d$*, we have*

$$\left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \le \Gamma_1 \frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}. \tag{116}$$

The proof of the above lemma is the same as that of Proposition 16 and hence omitted here. Next, for (I_2), we first have

$$(\mathrm{I}_2) = \left\|\left(\boldsymbol{V}_t^{(l)}\boldsymbol{V}_t^{(l)\top}\boldsymbol{\Xi}_t^{(l)}\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \le \left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\left\|\boldsymbol{V}_t^{(l)\top}\boldsymbol{\Xi}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\| \le \left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\left\|\boldsymbol{\Xi}_t^{(l)}\right\|. \tag{117}$$

For the first part, we have

$$\left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \le \left\|\boldsymbol{V}_{l,\cdot}^\star\right\| + \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \le \sqrt{\frac{\mu r}{d}} + \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \le 2\sqrt{\frac{\mu r}{d}}. \tag{118}$$

Here we use the assumption that $\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \le \sqrt{\frac{\mu r}{d}}$. For the second part, we have

$$\begin{aligned}
\left\|\boldsymbol{\Xi}_t^{(l)}\right\| &= \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} - \left(\mathcal{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\right\| \\
&\le \left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\right\| + \left\|\left(\mathcal{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\right\| \\
&\overset{(a)}{\le} \sigma_1^\star\left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\| + 21\Gamma\frac{\sigma_1^\star\mu r}{\sqrt{pd}} \\
&\overset{(b)}{\le} \sigma_1^\star\cdot\Gamma_1\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} + 21\Gamma\frac{\sigma_1^\star\mu r}{\sqrt{pd}} \\
&\le 2\Gamma_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.
\end{aligned} \tag{119}$$

Here in (a), we apply Lemma 12. In (b), we apply Lemma 18. Invoking Equations (118) and (119) in Equation (117), we obtain that

$$(\mathrm{I}_2) \le 2\sqrt{\frac{\mu r}{d}}\cdot 2\Gamma_1\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} = 4\Gamma_1\sigma_1^\star\frac{\kappa\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}}. \tag{120}$$

Next, we control the higher-order term $\left\|\left(\boldsymbol{A}_t^{(l)}\right)_{l,\cdot}\right\|$. To this end, we first notice that the l-th row of $\boldsymbol{A}_t^{(l)}$ is equal to the l-th row of

$$\begin{aligned}
&-\eta^2\boldsymbol{\Delta}_t^{(l)}\mathcal{P}_{\boldsymbol{V}_t^{(l)}}\boldsymbol{V}_t^{(l)\top}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)} - 0.5\eta^2\mathcal{P}_{\boldsymbol{V}_t^{(l)}}\boldsymbol{M}_t^{(l)2}\boldsymbol{V}_t^{(l)} - 0.5\eta^3\boldsymbol{\Delta}_t^{(l)}\mathcal{P}_{\boldsymbol{V}_t^{(l)}}\boldsymbol{M}_t^{(l)2}\boldsymbol{V}_t^{(l)} \\
&\quad + \left(\boldsymbol{I} + \eta\boldsymbol{\Delta}_t^{(l)}\right)\boldsymbol{V}_t^{(l)}\boldsymbol{R}\left(\boldsymbol{Y}_t^{(l)}\right).
\end{aligned} \tag{121}$$

Therefore, we can upper-bound its operator norm by

$$\begin{aligned}
\left\|\left(\boldsymbol{A}_t^{(l)}\right)_{l,\cdot}\right\| \le{}& \left\|\left(\boldsymbol{\Delta}_t^{(l)}\right)_{l,\cdot}\right\|\left(\eta^2\left\|\boldsymbol{M}_t^{(l)}\right\| + 0.5\eta^3\left\|\boldsymbol{M}_t^{(l)}\right\|^2 + \eta\left\|\boldsymbol{R}\left(\boldsymbol{Y}_t^{(l)}\right)\right\|\right) \\
&+ \left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\left(0.5\eta^2\left\|\boldsymbol{M}_t^{(l)}\right\|^2 + \left\|\boldsymbol{R}\left(\boldsymbol{Y}_t^{(l)}\right)\right\|\right).
\end{aligned} \tag{122}$$

Next, we notice that $\left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| \leq 2\sqrt{\frac{\mu r}{d}}$ and

$$\begin{aligned}\left\|\left(\boldsymbol{\Delta}_t^{(l)}\right)_{l,\cdot}\right\| &\leq \left\|\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)_{l,\cdot}\right\| \\ &\leq \sigma_1^\star\sqrt{\frac{\mu r}{d}} + \left\|\left(\boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\|\|\boldsymbol{\Sigma}_t\| \\ &\leq 5\sigma_1^\star\sqrt{\frac{\mu r}{d}}.\end{aligned} \tag{123}$$

On the other hand, according to Lemma 13, we have $\left\|\boldsymbol{M}_t^{(l)}\right\| \leq 6\sqrt{\sigma_1^\star}$ and $\left\|\boldsymbol{R}\left(\boldsymbol{Y}_t^{(l)}\right)\right\| \leq 200\eta^2\sigma_1^{\star 2}$. Therefore, we derive that

$$\left\|\left(\boldsymbol{A}_t^{(l)}\right)_{l,\cdot}\right\| \leq 470\eta^2\sigma_1^{\star 2}\sqrt{\frac{\mu r}{d}}. \tag{124}$$

Overall, we obtain

$$\begin{aligned}\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_{t+1}^{(l)}\right)_{l,\cdot}\right\| &\leq (1-0.5\eta\sigma_r^\star)\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| + 4\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}} + 470\eta^2\sigma_1^{\star 2}\sqrt{\frac{\mu r}{d}} \\ &\leq (1-0.5\eta\sigma_r^\star)\left\|\left(\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right)_{l,\cdot}\right\| + 8\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu^{1.5}r^2\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd^2}},\end{aligned} \tag{125}$$

provided that $\eta \asymp \frac{\mu r}{\sqrt{pd}\sigma_1^\star}$. This completes the proof of Proposition 17. ■

E.3. Proof of Proposition 8

We restate the proposition here for clarity.

Proposition 19 (One-step dynamic of $\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}$) *Suppose that the sampling rate satisfies $p \gtrsim \frac{\kappa^6\mu^4 r^9\log^4\left(\frac{1}{\alpha}\right)\log^2(d)}{d}$. Suppose that $\|\boldsymbol{V}^\star - \boldsymbol{V}_t\| \leq \frac{1}{2\kappa}$ and $\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \leq \sqrt{\frac{\mu r}{4d}}$. With probability at least $1 - \frac{1}{d^3}$, for any $1 \leq t \leq T \lesssim \frac{1}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right)$, we have*

$$\left\|\boldsymbol{V}_{t+1} - \boldsymbol{V}_{t+1}^{(l)}\right\|_{\mathrm{F}} \leq \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} + \Gamma\eta\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}. \tag{126}$$

Proof Note that

$$\boldsymbol{V}_{t+1} = \left(\boldsymbol{I} + \eta\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\right)\boldsymbol{V}_t + \boldsymbol{A}_t \quad \text{and} \quad \boldsymbol{V}_{t+1}^{(l)} = \left(\boldsymbol{I} + \eta\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\right)\boldsymbol{V}_t^{(l)} + \boldsymbol{A}_t^{(l)}. \tag{127}$$

Hence, we can expand $\left\|\boldsymbol{V}_{t+1} - \boldsymbol{V}_{t+1}^{(l)}\right\|_{\mathrm{F}}^2$ as

$$\left\|\boldsymbol{V}_{t+1} - \boldsymbol{V}_{t+1}^{(l)}\right\|_{\mathrm{F}}^2 = \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 + 2\eta\underbrace{\left\langle\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\boldsymbol{M}_t\boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)}\right\rangle}_{:=(\mathrm{I})} + (\mathrm{II}), \tag{128}$$

where

$$\begin{aligned}(\mathrm{II}) = \eta^2 \left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{M}_t \boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{M}_t^{(l)} \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}^2 + \left\| \boldsymbol{A}_t - \boldsymbol{A}_t^{(l)} \right\|_{\mathrm{F}}^2 \\ + 2 \left\langle \left(\boldsymbol{I} + \eta \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{M}_t \right) \boldsymbol{V}_t - \left(\boldsymbol{I} + \eta \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{M}_t^{(l)} \right) \boldsymbol{V}_t^{(l)}, \boldsymbol{A}_t - \boldsymbol{A}_t^{(l)} \right\rangle\end{aligned} \tag{129}$$

contains all the higher-order terms.

We first control (I). Notice that $\mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{M}_t = \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{\Xi}_t$ and $\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{M}_t^{(l)} = \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{\Xi}_t^{(l)}$. Here we define $\boldsymbol{\Xi}_t = \boldsymbol{M}_t - \boldsymbol{V}_t \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} + \boldsymbol{V}_t \boldsymbol{\Sigma}_t \boldsymbol{V}_t^\top$ and $\boldsymbol{\Xi}_t^{(l)} = \boldsymbol{M}_t^{(l)} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} + \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top}$, respectively. Therefore, we can decompose (I) as follows

$$\begin{aligned}(\mathrm{I}) &= \left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{\Xi}_t \boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{\Xi}_t^{(l)} \boldsymbol{V}_t^{(l)} \right\rangle \\ &= \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \left(\mathcal{P}_{\boldsymbol{V}_t}^{\perp} - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \right) \boldsymbol{\Xi}_t^{(l)} \boldsymbol{V}_t^{(l)} \right\rangle}_{:=(\mathrm{I}_1)} + \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{\Xi}_t \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right) \right\rangle}_{:=(\mathrm{I}_2)} \\ &\quad + \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\boldsymbol{\Xi}_t - \boldsymbol{\Xi}_t^{(l)} \right) \boldsymbol{V}_t^{(l)} \right\rangle}_{:=(\mathrm{I}_3)}.\end{aligned} \tag{130}$$

Next, we provide upper bounds for these terms separately. For (I_1), applying Cauchy-Schwarz inequality leads to

$$\begin{aligned}(\mathrm{I}_1) &\le \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} \left\| \left(\mathcal{P}_{\boldsymbol{V}_t}^{\perp} - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \right) \boldsymbol{\Xi}_t^{(l)} \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} \\ &\le \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} \left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \right\|_{\mathrm{F}} \left\| \boldsymbol{\Xi}_t^{(l)} \right\| \\ &\overset{(a)}{\le} 2 \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}^2 \left\| \boldsymbol{\Xi}_t^{(l)} \right\| \\ &\overset{(b)}{\le} 4 \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}^2 \cdot \Gamma_1 \sigma_1^\star \frac{\kappa \mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}.\end{aligned} \tag{131}$$

Here in (a), we apply Lemma 43. In (b), we use the result from Equation (119). Similarly, (I_2), we have

$$(\mathrm{I}_2) \le \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}^2 \|\boldsymbol{\Xi}_t\| \le 2 \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}^2 \cdot \Gamma_1 \sigma_1^\star \frac{\kappa \mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}. \tag{132}$$

Next, for (I_3), we further decompose it as

$$\begin{aligned}
(\mathrm{I}_3) &= \left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\boldsymbol{M}_t - \boldsymbol{M}_t^{(l)} + \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}^{\star} \boldsymbol{V}^{\star\top} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top} \right) \boldsymbol{V}_t^{(l)} \right\rangle \\
&= \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\mathcal{R}_{\Omega} - \mathcal{R}_{\Omega^{(l)}} \right) \left(\boldsymbol{X}^{\star} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top} \right) \boldsymbol{V}_t^{(l)} \right\rangle}_{:=(\mathrm{I}_{3,1})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\mathcal{I} - \mathcal{R}_{\Omega} \right) \left(\boldsymbol{V}_t \boldsymbol{\Sigma}_t \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right)^{\top} \right) \boldsymbol{V}_t^{(l)} \right\rangle}_{:=(\mathrm{I}_{3,2})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\mathcal{I} - \mathcal{R}_{\Omega} \right) \left(\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right) \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top} \right) \boldsymbol{V}_t^{(l)} \right\rangle}_{:=(\mathrm{I}_{3,3})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}^{\star} \boldsymbol{V}^{\star\top} \boldsymbol{V}_t^{(l)} \right\rangle}_{:=(\mathrm{I}_{3,4})}.
\end{aligned} \tag{133}$$

We control these terms separately. First, we present the following key proposition, the proof of which is deferred to the end of this section.

Proposition 20 *For all $0 \le t \le T$, we have*

$$\left\| \left(\mathcal{R}_{\Omega} - \mathcal{R}_{\Omega^{(l)}} \right) \left(\boldsymbol{X}^{\star} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top} \right) \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} \le \Gamma_3 \sigma_1^{\star} \sqrt{\frac{\kappa \mu^3 r^{5.5} \log\left(\frac{1}{\alpha}\right) \log\left(d\right)}{\sqrt{pd} \cdot d}}. \tag{134}$$

For $(\mathrm{I}_{3,1})$, we have

$$\begin{aligned}
(\mathrm{I}_{3,1}) &\le \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} \left\| \left(\mathcal{R}_{\Omega} - \mathcal{R}_{\Omega^{(l)}} \right) \left(\boldsymbol{X}^{\star} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top} \right) \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} \\
&\le \Gamma_3 \sigma_1^{\star} \sqrt{\frac{\kappa \mu^3 r^{5.5} \log\left(\frac{1}{\alpha}\right) \log\left(d\right)}{\sqrt{pd} \cdot d}} \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}.
\end{aligned} \tag{135}$$

Here in the last inequality, we use Proposition 20. Next, we apply Lemma 29 to control $(\mathrm{I}_{3,2})$. Specifically, upon setting $\boldsymbol{A} = \boldsymbol{V}_t$, $\boldsymbol{B} = \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right)$, $\boldsymbol{C} = \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right) \boldsymbol{\Sigma}_t$, and $\boldsymbol{D} = \boldsymbol{V}_t^{(l)}$ in Lemma 29, with probability at least $1 - d^{-3}$, one has

$$\begin{aligned}
(\mathrm{I}_{3,2}) &= \left\langle \left(\mathcal{I} - \mathcal{R}_{\Omega} \right) \left(\boldsymbol{A} \boldsymbol{C}^{\top} \right), \boldsymbol{B} \boldsymbol{D}^{\top} \right\rangle \\
&\le \Gamma \sqrt{\frac{d}{p}} \left\| \boldsymbol{A} \right\|_{2,\infty} \left\| \boldsymbol{B} \right\|_{\mathrm{F}} \cdot \left\| \boldsymbol{C} \right\|_{\mathrm{F}} \left\| \boldsymbol{D} \right\|_{2,\infty} \\
&= \Gamma \sqrt{\frac{d}{p}} \left\| \boldsymbol{V}_t \right\|_{2,\infty} \left\| \boldsymbol{V}_t^{(l)} \right\|_{2,\infty} \left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right) \right\|_{\mathrm{F}} \left\| \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right) \boldsymbol{\Sigma}_t \right\|_{\mathrm{F}} \\
&\le 16 \Gamma \sigma_1^{\star} \sqrt{\frac{\mu^2 r^2}{pd}} \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}^2.
\end{aligned} \tag{136}$$

Similarly, we apply Lemma 29 to control $(\mathrm{I}_{3,3})$. Upon setting $\boldsymbol{A} = \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t$, $\boldsymbol{B} = \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)$, $\boldsymbol{C} = \boldsymbol{V}_t$, and $\boldsymbol{D} = \boldsymbol{V}_t^{(l)}$ in Lemma 29, with probability at least $1 - d^{-3}$, one has

$$\begin{aligned}(\mathrm{I}_{3,3}) &\le \Gamma\sqrt{\frac{d}{p}}\left\|\boldsymbol{V}_t\right\|_{\mathrm{F}}\left\|\boldsymbol{V}_t^{(l)}\right\|_{2,\infty}\left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\right\|_{\mathrm{F}}\left\|\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_t\right\|_{2,\infty} \\ &\le 8\Gamma\sigma_1^{\star}\sqrt{\frac{\mu r^2}{p}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{2,\infty}.\end{aligned} \tag{137}$$

To control $\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{2,\infty}$, we apply Lemma 39:

$$\begin{aligned}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{2,\infty} &\le \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\left\|\boldsymbol{L}_{\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}}\right\|_{2,\infty} \\ &\le \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\sqrt{\|\boldsymbol{L}_{\boldsymbol{V}_t}\|_{2,\infty}^2 + \left\|\boldsymbol{L}_{\boldsymbol{V}_t^{(l)}}\right\|_{2,\infty}^2} \\ &\le 3\sqrt{\frac{\mu r}{d}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}.\end{aligned} \tag{138}$$

Hence, we have

$$(\mathrm{I}_{3,3}) \le 24\Gamma\sigma_1^{\star}\sqrt{\frac{\mu^2 r^3}{pd}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2. \tag{139}$$

Lastly, for $(\mathrm{I}_{3,4})$, we notice that

$$\begin{aligned}(\mathrm{I}_{3,4}) &= -\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^{\star}\boldsymbol{V}^{\star\top}\boldsymbol{V}_t^{(l)}\right\rangle \\ &\le -\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^{\star}\right\rangle + \sigma_1^{\star}\left\|\boldsymbol{V}^{\star} - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 \\ &\overset{(a)}{\le} \sigma_1^{\star}\left\|\boldsymbol{V}^{\star} - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 \\ &\overset{(b)}{\le} \Gamma_1\sigma_1^{\star}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2.\end{aligned} \tag{140}$$

Here in (a), we use the fact that

$$\left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^{\star}\right\rangle = \left\|\mathcal{P}_{\boldsymbol{V}_t}^{\perp}\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}^{\star 1/2}\right\|_{\mathrm{F}}^2 \ge 0. \tag{141}$$

In (b), we apply Proposition 16. Therefore, we obtain

$$(\mathrm{I}_3) \le C_1\sigma_1^{\star}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 + \Gamma_3\sigma_1^{\star}\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}, \tag{142}$$

which implies that

$$(\mathrm{I}) \le C_2\sigma_1^{\star}\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 + \Gamma_3\sigma_1^{\star}\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}. \tag{143}$$

Next, we move on to controlling (II). First, by the basic inequality $2ab \le a^2 + b^2$, we bound (II) as

$$\text{(II)} \le 2\eta^2 \left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{M}_t \boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{M}_t^{(l)} \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}^2 + 2 \left\| \boldsymbol{A}_t - \boldsymbol{A}_t^{(l)} \right\|_{\mathrm{F}}^2 + 2 \left\langle \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}, \boldsymbol{A}_t - \boldsymbol{A}_t^{(l)} \right\rangle. \tag{144}$$

Next, we provide the control over $\left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{M}_t \boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{M}_t^{(l)} \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}$. The remaining two terms can be controlled in a similar fashion, so we omit their analysis. We first apply triangle inequality to obtain

$$\begin{aligned} \left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{M}_t \boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{M}_t^{(l)} \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} &\le 2 \left\| \boldsymbol{M}_t \right\| \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} + \left\| \left(\boldsymbol{M}_t - \boldsymbol{M}_t^{(l)} \right) \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} \\ &\le 12\sigma_1^{\star} \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} + \left\| \left(\boldsymbol{M}_t - \boldsymbol{M}_t^{(l)} \right) \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}. \end{aligned} \tag{145}$$

For the second term, we further decompose it as

$$\begin{aligned} \left\| \left(\boldsymbol{M}_t - \boldsymbol{M}_t^{(l)} \right) \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} \le & \left\| (\mathcal{R}_{\Omega} - \mathcal{R}_{\Omega^{(l)}}) \left(\boldsymbol{X}^{\star} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top} \right) \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} \\ & + \left\| (\mathcal{I} - \mathcal{R}_{\Omega}) \left(\boldsymbol{V}_t \boldsymbol{\Sigma}_t \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right)^{\top} \right) \boldsymbol{V}_t \right\|_{\mathrm{F}} \\ & + \left\| (\mathcal{I} - \mathcal{R}_{\Omega}) \left(\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right) \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{\top} \right) \boldsymbol{V}_t \right\|_{\mathrm{F}} \\ & + 4\sigma_1^{\star} \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}. \end{aligned} \tag{146}$$

The first term can be controlled by Proposition 20. For the second term, we have

$$\begin{aligned} \left\| (\mathcal{I} - \mathcal{R}_{\Omega}) \left(\boldsymbol{V}_t \boldsymbol{\Sigma}_t \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right)^{\top} \right) \boldsymbol{V}_t \right\|_{\mathrm{F}} &= \sup_{\|\boldsymbol{Z}\|_{\mathrm{F}} \le 1} \left\langle (\mathcal{I} - \mathcal{R}_{\Omega}) \left(\boldsymbol{V}_t \boldsymbol{\Sigma}_t \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right)^{\top} \right), \boldsymbol{Z} \boldsymbol{V}_t \right\rangle \\ &\overset{(a)}{\le} \Gamma \sqrt{\frac{d}{p}} \left\| \boldsymbol{V}_t \right\|_{2,\infty}^2 \left\| \left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right) \boldsymbol{\Sigma}_t \right\|_{\mathrm{F}} \\ &\le 8\Gamma\sigma_1^{\star} \sqrt{\frac{\mu^2 r^2}{pd}} \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}. \end{aligned} \tag{147}$$

Here we apply Corollary 29 in (a). Similarly, we can also control the third term as follows

$$\begin{aligned} \left\| (\mathcal{I} - \mathcal{R}_{\Omega}) \left(\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right) \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{\top} \right) \boldsymbol{V}_t \right\|_{\mathrm{F}} &= \sup_{\|\boldsymbol{Z}\|_{\mathrm{F}} \le 1} \left\langle (\mathcal{I} - \mathcal{R}_{\Omega}) \left(\left(\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right) \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{\top} \right), \boldsymbol{Z} \boldsymbol{V}_t \right\rangle \\ &\le 8\Gamma \sqrt{\frac{\mu r^2}{p}} \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{2,\infty} \\ &\le 24\Gamma\sigma_1^{\star} \sqrt{\frac{\mu^2 r^3}{pd}} \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}. \end{aligned} \tag{148}$$

Combining the above inequalities leads to

$$\left\| \mathcal{P}_{\boldsymbol{V}_t}^{\perp} \boldsymbol{M}_t \boldsymbol{V}_t - \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{M}_t^{(l)} \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} \le C\sigma_1^{\star} \left\| \boldsymbol{V}_t - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}}. \tag{149}$$

Similarly, we can derive that $\left\|\boldsymbol{A}_t - \boldsymbol{A}_t^{(l)}\right\|_{\mathrm{F}} \leq C\sigma_1^\star \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} + C\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log(d)}{\sqrt{pd}\cdot d}}$. Therefore, we have

$$\text{(II)} \leq C_3\eta^2\left(\sigma_1^\star\left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 + \sigma_1^{\star 2}\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(d\right)}{\sqrt{pd}\cdot d}\right). \tag{150}$$

Putting everything together, we obtain

$$\begin{aligned}\left\|\boldsymbol{V}_{t+1}-\boldsymbol{V}_{t+1}^{(l)}\right\|_{\mathrm{F}} &\leq \left(1+\Gamma_1\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)\left\|\boldsymbol{V}_t-\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} + \Gamma_3\eta\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(d\right)}{\sqrt{pd}\cdot d}} \\ &\leq \left\|\boldsymbol{V}_t - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} + 2\Gamma_3\eta\sigma_1^\star\sqrt{\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(d\right)}{\sqrt{pd}\cdot d}}.\end{aligned} \tag{151}$$

■

E.4. Proof of Proposition 20

To prove this proposition, we propose a novel argument based on *adaptive ϵ-nets*. Upon fixing an ϵ-net $\mathcal{N}_\epsilon$ with respect to Frobenius norm for the operator norm ball $\mathcal{B}_{\mathrm{op}}^{r\times r}(4\sigma_1^\star)$ with $\epsilon = \frac{c}{d}$[2], we first construct a series of adaptive ϵ-nets $\left\{\mathcal{V}_{\epsilon,t}^{(l)}\right\}_{t=0}^{T}$ in the following recursive manner.

$$\begin{aligned}\mathcal{V}_{\epsilon,t+1}^{(l)} = \Bigg\{\boldsymbol{Y}_{\epsilon,t+1}^{(l)} : \left\|\boldsymbol{Y}_{\epsilon,t+1}^{(l)}\right\|_{2,\infty} \leq 2\sqrt{\frac{\mu r}{d}},\ \text{and } \boldsymbol{Y}_{\epsilon,t+1}^{(l)} = \boldsymbol{Z}_{\epsilon,t+1}^{(l)}\left(\boldsymbol{Z}_{\epsilon,t+1}^{(l)\top}\boldsymbol{Z}_{\epsilon,t+1}^{(l)}\right)^{-1/2} \text{ where} \\ \boldsymbol{Z}_{\epsilon,t+1}^{(l)} = \left(\boldsymbol{I} + \eta\mathcal{R}_{\Omega^{(l)}}\left(\boldsymbol{X}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\right)\boldsymbol{V}_{\epsilon,t}^{(l)}, \forall \boldsymbol{V}_{\epsilon,t}^{(l)} \in \mathcal{V}_{\epsilon,t}^{(l)}, \boldsymbol{\Sigma}_{\epsilon,t} \in \mathcal{N}_\epsilon\Bigg\}.\end{aligned} \tag{152}$$

2. We call a set $\mathcal{N}_\epsilon$ an ϵ-net with respect to a norm $\|\cdot\|$ for a given set S if for any element $s \in S$, there exists an element $s' \in \mathcal{N}_\epsilon$ such that $\|s - s'\| \leq \epsilon$.

Moreover, we set $\mathcal{V}_{\epsilon,0}^{(l)} = \{\boldsymbol{V}^\star\}$. Then, we denote the best approximation of $\boldsymbol{V}_t^{(l)}$ in $\mathcal{V}_{\epsilon,t}^{(l)}$ as $\boldsymbol{V}_{\epsilon,t}^{(l)} = \arg\min_{\boldsymbol{V}\in\mathcal{V}_{\epsilon,t}^{(l)}} \left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}\right\|_{\mathrm{F}}$. Hence, we have the following decomposition

$$
\begin{aligned}
&\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}} \\
&\le \underbrace{\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}}_{\text{(I)}} \\
&\quad + \underbrace{\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right)\right\|_{\mathrm{F}}}_{\text{(II)}} \\
&\quad + \underbrace{\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right)^\top\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}}_{\text{(III)}} \\
&\quad + \underbrace{\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{V}_t^{(l)}\left(\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_{\epsilon,t}\right)\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}}_{\text{(IV)}} \\
&\quad + \underbrace{\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)\left(\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right)\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}}_{\text{(V)}}.
\end{aligned}
\tag{153}
$$

To proceed, we require the following key lemmas.

Lemma 21 *For any $1 \le i \le j \le d$, let $p_{i,j}$ be a Bernoulli random variable taking the value $p_{i,j} = 1$ if and only if $(i,j) \in \Omega$. Let $r_{i,j} = \frac{1}{2p}(p_{i,j} + p_{j,i})$. The following statements hold:*

- **Independent case:** *Suppose the sampling rate $p \gtrsim \frac{\mu r}{d}\log\left(\frac{4r}{\delta}\right)$. Suppose that $\boldsymbol{X} \in \mathcal{S}_{d\times d}$ and $\boldsymbol{V} \in \mathcal{O}_{d\times r}$ with $\|\boldsymbol{V}\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$ are independent of $r_{l,1},\ldots,r_{l,d}$. Then, with probability at least $1-\delta$, we have*

$$
\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)(\boldsymbol{X})\,\boldsymbol{V}\right\|_{\mathrm{F}}^2 \le \frac{32\mu r\log(4r/\delta)}{p}\left\|\boldsymbol{X}\right\|_{\max}^2. \tag{154}
$$

- **General case:** *Suppose the sampling rate $p \gtrsim \frac{\log(d)}{d}$. For arbitrary $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times r}$, with probability at least $1 - \frac{1}{d^3}$, we have*

$$
\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)(\boldsymbol{X})\,\boldsymbol{V}\right\|_{\mathrm{F}}^2 \le \frac{2d}{p}\left\|\boldsymbol{X}\right\|_{\max}^2\left\|\boldsymbol{V}\right\|_{\mathrm{F}}^2. \tag{155}
$$

Lemma 22 *Suppose the sampling rate satisfies $p \gtrsim \frac{\mu r}{d}\log\left(\frac{4r}{\delta}\right)$. For any $0 \le t \le T = \frac{100}{\eta\sigma_r^\star}\log\left(\frac{1}{\alpha}\right), 1 \le l \le d$, with probability at least $1 - \frac{1}{d^3}$, we have*

$$
\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 \le \Gamma\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{r\sigma_1^\star}{\epsilon}\right)}{\sqrt{pd^3}}. \tag{156}
$$

Lemma 23 *Under the same conditions as Lemma 22, we have*

$$\left\|\boldsymbol{V}_{t+1}^{(l)}-\boldsymbol{V}_{\epsilon,t+1}^{(l)}\right\|_{\mathrm{F}} \leq \left(1+\Gamma\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)\left\|\boldsymbol{V}_{t}^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}+C_3\eta\frac{\sigma_1^\star\mu r}{\sqrt{pd}}\epsilon. \tag{157}$$

The proof of these lemmas is deferred to the end of this section. Now we use Lemma 23 to control the dynamic of $\left\|\boldsymbol{V}_{t}^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}$. Note that $\left\|\boldsymbol{V}_{0}^{(l)}-\boldsymbol{V}_{\epsilon,0}^{(l)}\right\|_{\mathrm{F}}=0$. Hence, applying Lemma 47, we have

$$\begin{aligned}\left\|\boldsymbol{V}_{t}^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} &\leq C\left(\left(1+\Gamma\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}\right)^t-1\right)\frac{\eta\frac{\sigma_1^\star\mu r}{\sqrt{pd}}\epsilon}{\eta\sigma_1^\star\frac{\kappa\mu r^{1.5}\log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}}}\\ &\leq C\eta\frac{\sigma_1^\star\mu r}{\sqrt{pd}}\epsilon\cdot t.\end{aligned} \tag{158}$$

Now we are ready to control (I) to (V) separately. First, Lemma 22 directly implies that

$$\text{(I)} \leq \Gamma\frac{\sqrt{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{r\sigma_1^\star}{\epsilon}\right)}}{\sqrt[4]{pd^3}}. \tag{159}$$

On the other hand, applying Lemma 21 to (II) leads to

$$\text{(II)} \leq \sqrt{\frac{2d}{p}}\left\|\boldsymbol{X}^\star-\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right\|_{\max}\left\|\boldsymbol{V}_t^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} \leq \Gamma\sqrt{\frac{\mu^2r^2}{pd}}\left\|\boldsymbol{V}_t^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}. \tag{160}$$

Similarly, we have

$$\begin{aligned}\text{(III)} &\leq \sqrt{\frac{2d}{p}}\left\|\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\left(\boldsymbol{V}_t^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right)^\top\right\|_{\max}\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} \leq \Gamma\sqrt{\frac{\mu r^2}{p}}\left\|\boldsymbol{V}_t^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}},\\ \text{(IV)} &\leq \sqrt{\frac{2d}{p}}\left\|\boldsymbol{V}_t^{(l)}\left(\boldsymbol{\Sigma}_t-\boldsymbol{\Sigma}_{\epsilon,t}\right)\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right\|_{\max}\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} \leq \Gamma\sqrt{\frac{\mu^2r^3}{pd}}\epsilon,\\ \text{(V)} &\leq \sqrt{\frac{2d}{p}}\left\|\left(\boldsymbol{V}_t^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right)\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right\|_{\max}\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}} \leq \Gamma\sqrt{\frac{\mu r^2}{p}}\left\|\boldsymbol{V}_t^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}.\end{aligned} \tag{161}$$

Overall, we derive that

$$\begin{aligned}&\left\|\left(\mathcal{R}_\Omega-\mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star-\boldsymbol{V}_t^{(l)}\boldsymbol{\Sigma}_t\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}\\ &\leq \Gamma\frac{\sqrt{\kappa\mu^3r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{r\sigma_1^\star}{\epsilon}\right)}}{\sqrt[4]{pd^3}}+\Gamma\sqrt{\frac{\mu r^2}{p}}\left\|\boldsymbol{V}_t^{(l)}-\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}\\ &\leq \Gamma\frac{\sqrt{\kappa\mu^3r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{r\sigma_1^\star}{\epsilon}\right)}}{\sqrt[4]{pd^3}}+\Gamma\sqrt{\frac{\mu r^2}{p}}\cdot\eta\frac{\sigma_1^\star\mu r}{\sqrt{pd}}\epsilon\cdot t\\ &\leq 2\Gamma\sigma_1^\star\frac{\sqrt{\kappa\mu^3r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(d\right)}}{\sqrt[4]{pd^3}}.\end{aligned} \tag{162}$$

The last inequality follows from the fact that $\epsilon = \frac{c}{d}$. This completes the proof of Proposition 20. ■

Next, we proceed to present the proofs of Lemma 21, Lemma 22, and Lemma 23.

Proof of Lemma 21. We first expand $\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})(\boldsymbol{X})\boldsymbol{V}\|_{\mathrm{F}}^2$ as follows

$$\|(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}})(\boldsymbol{X})\boldsymbol{V}\|_{\mathrm{F}}^2 = \sum_{j=1}^{r}\left(\sum_{k=1}^{d}(r_{l,k}-1)X_{l,k}V_{k,j}\right)^2 + \sum_{i\neq l}(r_{i,l}-1)^2 X_{i,l}^2\|\boldsymbol{V}_{l,\cdot}\|^2. \quad (163)$$

Next, we prove these two cases separately.

Independent case. First, we control $\sum_{k=1}^{d}(r_{l,k}-1)X_{l,k}V_{k,j}$ for all $1\le j\le r$ via Bernstein's inequality (Lemma 32). To this goal, for a fixed $1\le j\le r$, upon defining $Z_k = (r_{l,k}-1)X_{l,k}V_{k,j}$ with $\mathbb{E}[Z_k]=0$, we have

$$\begin{aligned} M &:= \max_k |Z_k| \le \frac{1}{p}\|\boldsymbol{X}\|_{\max}\|\boldsymbol{V}\|_{2,\infty}, \\ \nu^2 &:= \sum_{k=1}^{d}\mathrm{Var}\left[Z_k^2\right] \le \frac{1}{p}\sum_{k=1}^{d}X_{l,k}^2V_{k,j}^2 \le \frac{1}{p}\|\boldsymbol{X}\|_{\max}^2\sum_{k=1}^{d}V_{k,j}^2 = \frac{1}{p}\|\boldsymbol{X}\|_{\max}^2. \end{aligned} \quad (164)$$

Here in the last equality we use the fact that $\sum_{k=1}^{d}V_{k,j}^2 = 1$ since $\boldsymbol{V}\in\mathcal{O}_{d\times r}$. Therefore, due to Bernstein's inequality, with probability at least $1-\frac{\delta}{2r}$, one has

$$\begin{aligned} \left|\sum_{k=1}^{d}Z_k\right| &\le 2\nu\sqrt{\log\left(\frac{4r}{\delta}\right)} + \frac{4}{3}M\log\left(\frac{4r}{\delta}\right) \\ &\le \frac{2\|\boldsymbol{X}\|_{\max}}{\sqrt{p}}\sqrt{\log\left(\frac{4r}{\delta}\right)} + \frac{4\|\boldsymbol{X}\|_{\max}\|\boldsymbol{V}\|_{2,\infty}}{3p}\log\left(\frac{4r}{\delta}\right) \\ &\le \frac{4\|\boldsymbol{X}\|_{\max}}{\sqrt{p}}\sqrt{\log\left(\frac{4r}{\delta}\right)}, \end{aligned} \quad (165)$$

where we use the assumption $p \gtrsim \frac{\mu r}{d}\log\left(\frac{4r}{\delta}\right)$. Hence, via a union bound, we know that with probability at least $1-\frac{\delta}{2}$, we have

$$\sum_{j=1}^{r}\left(\sum_{k=1}^{d}(r_{l,k}-1)X_{l,k}V_{k,j}\right)^2 \le \frac{16r\|\boldsymbol{X}\|_{\max}^2}{p}\log\left(\frac{4r}{\delta}\right). \quad (166)$$

Next, we can control $\sum_{i\neq l}(r_{i,l}-1)^2X_{i,l}^2\|\boldsymbol{V}_{l,\cdot}\|^2$ as

$$\sum_{i\neq l}(r_{i,l}-1)^2X_{i,l}^2\|\boldsymbol{V}_{l,\cdot}\|^2 \le \sum_{i=1}^{d}(r_{i,l}-1)^2\|\boldsymbol{X}\|_{\max}^2\|\boldsymbol{V}\|_{2,\infty}^2. \quad (167)$$

Then, we apply Bernstein's inequality to control $\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2$. To this end, notice that

$$\begin{aligned}
&\mathbb{E}\left[\left(r_{l,k}-1\right)^2\right]=\frac{1-p}{2p}\leq\frac{1}{2p},\\
&M:=\max_k\left(r_{l,k}-1\right)^2\leq\left(\frac{1}{p}-1\right)^2\leq\frac{1}{p^2},\\
&\nu^2:=\sum_{k=1}^{d}\mathrm{Var}\left[\left(r_{l,k}-1\right)^2\right]\leq\frac{d}{p^3}.
\end{aligned}\tag{168}$$

Therefore, with probability at least $1-\frac{\delta}{2}$, we have

$$\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2\leq\frac{d}{2p}+\frac{4}{3}\frac{1}{p^2}\log\left(\frac{2}{\delta}\right)+2\sqrt{\frac{d}{p^3}}\sqrt{\log\left(\frac{2}{\delta}\right)}\leq\frac{d}{p}\tag{169}$$

since we set $p\gtrsim\frac{\mu r}{d}\log\left(\frac{4r}{\delta}\right)$. This implies that with probability at least $1-\frac{\delta}{2}$, we have

$$\sum_{i\neq l}\left(r_{i,l}-1\right)^2X_{i,l}^2\left\|\boldsymbol{V}_{l,\cdot}\right\|^2\leq\frac{d}{p}\left\|\boldsymbol{X}\right\|_{\max}^2\left\|\boldsymbol{V}\right\|_{2,\infty}^2.\tag{170}$$

Finally, taking a union bound, we conclude that with probability at least $1-\delta$, we have

$$\begin{aligned}
\left\|\left(\mathcal{R}_{\Omega}-\mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}\right)\boldsymbol{V}\right\|_{\mathrm{F}}^2&\leq\frac{16r\left\|\boldsymbol{X}\right\|_{\max}^2\log(4r/\delta)}{p}+\frac{d}{p}\left\|\boldsymbol{X}\right\|_{\max}^2\left\|\boldsymbol{V}\right\|_{2,\infty}^2\\
&\leq\frac{32\mu r\log(4r/\delta)}{p}\left\|\boldsymbol{X}\right\|_{\max}^2.
\end{aligned}\tag{171}$$

General case. First, we apply Cauchy-Schwarz inequality to obtain

$$\begin{aligned}
\sum_{j=1}^{r}\left(\sum_{k=1}^{d}\left(r_{l,k}-1\right)X_{l,k}V_{k,j}\right)^2&\leq\sum_{j=1}^{r}\left(\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2\right)\left(\sum_{k=1}^{d}X_{l,k}^2V_{k,j}^2\right)\\
&\leq\left(\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2\right)\cdot\sum_{j=1}^{r}\left\|\boldsymbol{X}\right\|_{\max}^2\sum_{k=1}^{d}V_{k,j}^2\\
&=\left(\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2\right)\cdot\left\|\boldsymbol{X}\right\|_{\max}^2\left\|\boldsymbol{V}\right\|_{\mathrm{F}}^2.
\end{aligned}\tag{172}$$

Next, we can control $\sum_{i\neq l}\left(r_{i,l}-1\right)^2X_{i,l}^2\left\|\boldsymbol{V}_{l,\cdot}\right\|^2$ as follows

$$\begin{aligned}
\sum_{i\neq l}\left(r_{i,l}-1\right)^2X_{i,l}^2\left\|\boldsymbol{V}_{l,\cdot}\right\|^2&\leq\sum_{i=1}^{d}\left(r_{i,l}-1\right)^2\left\|\boldsymbol{X}\right\|_{\max}^2\left\|\boldsymbol{V}\right\|_{\mathrm{F}}^2\\
&=\left(\sum_{k=1}^{d}\left(r_{l,k}-1\right)^2\right)\left\|\boldsymbol{X}\right\|_{\max}^2\left\|\boldsymbol{V}\right\|_{\mathrm{F}}^2.
\end{aligned}\tag{173}$$

Here we use the fact that $r_{i,j} = r_{j,i}, \forall i, j \in [d]$. Next, we apply Bernstein's inequality to control $\sum_{k=1}^{d} (r_{l,k} - 1)^2$. To this end, notice that

$$\begin{aligned} &\mathbb{E}\left[(r_{l,k} - 1)^2\right] = \frac{1-p}{2p} \leq \frac{1}{2p}, \\ &M := \max_k (r_{l,k} - 1)^2 \leq \left(\frac{1}{p} - 1\right)^2 \leq \frac{1}{p^2}, \\ &\nu^2 := \sum_{k=1}^{d} \mathrm{Var}\left[(r_{l,k} - 1)^2\right] \leq \frac{d}{p^3}. \end{aligned} \tag{174}$$

Therefore, with probability at least $1 - \delta$, we have

$$\sum_{k=1}^{d} (r_{l,k} - 1)^2 \leq \frac{d}{2p} + \frac{4}{3}\frac{1}{p^2} \log\left(\frac{2}{\delta}\right) + 2\sqrt{\frac{d}{p^3}}\sqrt{\log\left(\frac{2}{\delta}\right)}. \tag{175}$$

Specifically, upon setting $\delta = \frac{1}{d^3}$, we obtain that with probability at least $1 - \frac{1}{d^3}$, one has

$$\sum_{k=1}^{d} (r_{l,k} - 1)^2 \leq \frac{d}{p}, \tag{176}$$

since we set $p \gtrsim \frac{\log(d)}{d}$. Overall, we have

$$\left\|\left(\mathcal{R}_{\Omega} - \mathcal{R}_{\Omega^{(l)}}\right)(\boldsymbol{X})\,\boldsymbol{V}\right\|_{\mathrm{F}}^2 \leq 2\left(\sum_{k=1}^{d} (r_{l,k} - 1)^2\right)\|\boldsymbol{X}\|_{\max}^2 \|\boldsymbol{V}\|_{\mathrm{F}}^2 \leq \frac{2d}{p}\|\boldsymbol{X}\|_{\max}^2 \|\boldsymbol{V}\|_{\mathrm{F}}^2, \tag{177}$$

with probability at least $1 - \frac{1}{d^3}$. This completes the proof. ■

Proof of Lemma 22. First, for fixed $\boldsymbol{\Sigma}_{\epsilon,t} \in \mathcal{N}_\epsilon$ and $\boldsymbol{V}_{\epsilon,t}^{(l)} \in \mathcal{V}_{\epsilon,t}^{(l)}$, Lemma 21 implies that with probability at least $1 - \delta$, we have

$$\left\|\left(\mathcal{R}_{\Omega} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 \leq \frac{32\mu r \log(4r/\delta)}{p}\left\|\boldsymbol{X}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right\|_{\max}^2. \tag{178}$$

Note that

$$\begin{aligned} \left\|\boldsymbol{X}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right\|_{\max} &\leq \|\boldsymbol{X}^\star\|_{\max} + \left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right\|_{\max} \\ &\overset{(a)}{\leq} \|\boldsymbol{\Sigma}^\star\|\,\|\boldsymbol{V}^\star\|_{2,\infty}^2 + \|\boldsymbol{\Sigma}_{\epsilon,t}\|\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{2,\infty}^2 \\ &\leq 9\sigma_1^\star\frac{\mu r}{d}. \end{aligned} \tag{179}$$

Therefore, with probability at least $1 - \delta$, we have

$$\left\|\left(\mathcal{R}_{\Omega} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 \leq \frac{288\mu^3 r^3 \log(4r/\delta)}{pd^2}. \tag{180}$$

Lastly, we apply the union bound to finalize the desired result. To this end, note that $\mathcal{B}_{\mathrm{op}}^{r\times r}(4\sigma_1^\star) \subset \mathcal{B}_{\mathrm{F}}^{r\times r}(4\sqrt{r}\sigma_1^\star)$. Hence, according to Lemma 42, we know that $|\mathcal{N}_\epsilon| \le \left(\frac{12\sqrt{r}\sigma_1^\star}{\epsilon}\right)^{r^2}$ which implies that $\left|\mathcal{V}_{\epsilon,t}^{(l)}\right| \le \left(\frac{6\sqrt{r}\sigma_1^\star}{\epsilon}\right)^{r^2 t}$. Therefore, the total cardinality of $\bigcup_{l=1}^d \bigcup_{t=0}^T \mathcal{V}_{\epsilon,t}^{(l)}$ is upper-bounded by

$$\left|\bigcup_{l=1}^{d}\bigcup_{t=0}^{T}\mathcal{V}_{\epsilon,t}^{(l)}\right| \le d\cdot\sum_{t=0}^{T}\left(\frac{12\sqrt{r}\sigma_1^\star}{\epsilon}\right)^{r^2 t} \le 2d\left(\frac{12\sqrt{r}\sigma_1^\star}{\epsilon}\right)^{r^2 T}. \tag{181}$$

Hence, once we set $\delta = \frac{1}{2d^4}\left(\frac{6\sqrt{r}\sigma_1^\star}{\epsilon}\right)^{-r^2T}$, we obtain that with probability at least $1-d^{-3}$, for any $0\le t\le T, 1\le l\le d$ and $\boldsymbol{V}_{\epsilon,t}^{(l)} \in \mathcal{V}_{\epsilon,t}^{(l)}$,

$$\begin{aligned}\left\|\left(\mathcal{R}_\Omega - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{X}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 &\le \Gamma\frac{\mu^3 r^3}{pd^2}\left(\log(d) + r^2 T\left(\frac{r\sigma_1^\star}{\epsilon}\right)\right)\\ &\le \Gamma\frac{\kappa\mu^3 r^{5.5}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{r\sigma_1^\star}{\epsilon}\right)}{\sqrt{pd^3}}.\end{aligned} \tag{182}$$

Here we use the fact that $T \le \frac{100}{\eta\sigma_r^\star}\log(1/\alpha)$ and $\eta = \Theta\left(\frac{\mu}{\sigma_1^\star}\sqrt{\frac{r}{pd}}\right)$. This completes the proof. ■

Proof of Lemma 23. The derivation of $\left\|\boldsymbol{V}_{t+1}^{(l)} - \boldsymbol{V}_{\epsilon,t+1}^{(l)}\right\|_{\mathrm{F}}$ is nearly the same as that of $\left\|\boldsymbol{V}_{t+1} - \boldsymbol{V}_{t+1}^{(l)}\right\|_{\mathrm{F}}$. First, note that

$$\boldsymbol{V}_{\epsilon,t+1}^{(l)} = \left(\boldsymbol{I} + \eta\mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp}\boldsymbol{M}_{\epsilon,t}^{(l)}\right)\boldsymbol{V}_{\epsilon,t}^{(l)} + \boldsymbol{A}_{\epsilon,t}^{(l)}, \tag{183}$$

where $\boldsymbol{M}_{\epsilon,t}^{(l)}$ and $\boldsymbol{A}_{\epsilon,t}^{(l)}$ are defined similar to $\boldsymbol{M}_t^{(l)}$ and $\boldsymbol{A}_t^{(l)}$. Hence, we can expand $\left\|\boldsymbol{V}_{t+1}^{(l)} - \boldsymbol{V}_{\epsilon,t+1}^{(l)}\right\|_{\mathrm{F}}^2$ as

$$\left\|\boldsymbol{V}_{t+1}^{(l)} - \boldsymbol{V}_{\epsilon,t+1}^{(l)}\right\|_{\mathrm{F}}^2 = \left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 + 2\eta\underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)} - \mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp}\boldsymbol{M}_{\epsilon,t}^{(l)}\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle}_{:=(\mathrm{I})} + (\mathrm{II}). \tag{184}$$

where

$$\begin{aligned}(\mathrm{II}) = \eta^2\left\|\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\boldsymbol{V}_t^{(l)} - \mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp}\boldsymbol{M}_{\epsilon,t}^{(l)}\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{A}_t^{(l)} - \boldsymbol{A}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2\\ + 2\left\langle\left(\boldsymbol{I} + \eta\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{M}_t^{(l)}\right)\boldsymbol{V}_t^{(l)} - \left(\boldsymbol{I} + \eta\mathcal{P}_{\boldsymbol{V}_{\epsilon,t}^{(l)}}^{\perp}\boldsymbol{M}_{\epsilon,t}^{(l)}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}, \boldsymbol{A}_t^{(l)} - \boldsymbol{A}_{\epsilon,t}^{(l)}\right\rangle\end{aligned} \tag{185}$$

contains all the higher-order terms. Next, we further decompose (I) as follows

$$\begin{aligned}
(\mathrm{I}) &= \left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} \boldsymbol{\Xi}_t^{(l)} \boldsymbol{V}_t^{(l)} - \mathcal{P}^{\perp}_{\boldsymbol{V}_{\epsilon,t}^{(l)}} \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \boldsymbol{V}_{\epsilon,t}^{(l)} \right\rangle \\
&= \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \left(\mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} - \mathcal{P}^{\perp}_{\boldsymbol{V}_{\epsilon,t}^{(l)}} \right) \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \boldsymbol{V}_{\epsilon,t}^{(l)} \right\rangle}_{:=(\mathrm{I}_1)} + \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} \boldsymbol{\Xi}_t^{(l)} \left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right) \right\rangle}_{:=(\mathrm{I}_2)} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} \left(\boldsymbol{\Xi}_t^{(l)} - \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \right) \boldsymbol{V}_{\epsilon,t}^{(l)} \right\rangle}_{:=(\mathrm{I}_3)}.
\end{aligned} \tag{186}$$

Here we define $\boldsymbol{\Xi}_t^{(l)} = \boldsymbol{M}_t^{(l)} - \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} + \boldsymbol{V}_t^{(l)} \boldsymbol{\Sigma}_t \boldsymbol{V}_t^{(l)\top}$ and $\boldsymbol{\Xi}_{\epsilon,t}^{(l)} = \boldsymbol{M}_{\epsilon,t}^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}^\star \boldsymbol{V}^{\star\top} + \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}_{\epsilon,t} \boldsymbol{V}_{\epsilon,t}^{(l)\top}$. We control (I_1) as follows

$$(\mathrm{I}_1) \le \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \left\| \mathcal{P}^{\perp}_{\boldsymbol{V}_t^{(l)}} - \mathcal{P}^{\perp}_{\boldsymbol{V}_{\epsilon,t}^{(l)}} \right\|_{\mathrm{F}} \left\| \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \right\| \le 2 \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 \left\| \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \right\|. \tag{187}$$

Here we apply Lemma 43 in the second inequality. For $\left\| \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \right\|$, following the same analysis as in Equation (119), we have

$$\begin{aligned}
\left\| \boldsymbol{\Xi}_{\epsilon,t}^{(l)} \right\| &\le \sigma_1^\star \left\| \boldsymbol{V}^\star - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\| + 10\Gamma \frac{\sigma_1^\star \mu r}{\sqrt{pd}} \\
&\le \sigma_1^\star \left(\left\| \boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)} \right\| + \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\| \right) + 10\Gamma \frac{\sigma_1^\star \mu r}{\sqrt{pd}} \\
&\le \sigma_1^\star \left\| \boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)} \right\| + 20\Gamma \frac{\sigma_1^\star \mu r}{\sqrt{pd}}.
\end{aligned} \tag{188}$$

Here in the last inequality, we use the fact that $\left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\| \le \epsilon \le 10\Gamma \frac{\mu r}{\sqrt{pd}}$. Plugging this inequality into Equation (187), we obtain

$$(\mathrm{I}_1) \le 2\sigma_1^\star \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 \left(\sigma_1^\star \left\| \boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)} \right\| + 20\Gamma \frac{\sigma_1^\star \mu r}{\sqrt{pd}} \right). \tag{189}$$

Similar to our derivation for (I_1), it follows that $(\mathrm{I}_2) \le \sigma_1^\star \left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}^2 \left(\sigma_1^\star \left\|\boldsymbol{V}^\star - \boldsymbol{V}_t^{(l)}\right\| + 10\Gamma\frac{\sigma_1^\star \mu r}{\sqrt{pd}}\right)$. Lastly, to bound (I_3), we further decompose it as

$$
\begin{aligned}
(\mathrm{I}_3) &= \left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\boldsymbol{M}_t^{(l)} - \boldsymbol{M}_{\epsilon,t}^{(l)} + \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top} - \boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_t^{(l)}\boldsymbol{V}_{\epsilon,t}^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle \\
&= \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\mathcal{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}_{\epsilon,t}^{(l)}\left(\boldsymbol{V}_{\epsilon,t}^{(l)} - \boldsymbol{V}_t^{(l)}\right)^\top\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,1})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\mathcal{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\left(\boldsymbol{V}_{\epsilon,t}^{(l)} - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_{\epsilon,t}^{(l)}\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,2})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\mathcal{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{V}_t^{(l)}\left(\boldsymbol{\Sigma}_{\epsilon,t} - \boldsymbol{\Sigma}_t\right)\boldsymbol{V}_t^{(l)\top}\right)\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,3})} \\
&\quad + \underbrace{\left\langle \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\boldsymbol{V}_{\epsilon,t}^{(l)}\boldsymbol{\Sigma}^\star\boldsymbol{V}^{\star\top}\boldsymbol{V}_{\epsilon,t}^{(l)}\right\rangle}_{:=(\mathrm{I}_{3,4})}.
\end{aligned}
\tag{190}
$$

Next, we control these terms separately. First, we apply Lemma 29 to control $(\mathrm{I}_{3,1})$. Specifically, upon setting $\boldsymbol{A} = \boldsymbol{V}_{\epsilon,t}^{(l)}, \boldsymbol{B} = \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right), \boldsymbol{C} = \left(\boldsymbol{V}_{\epsilon,t}^{(l)} - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_{\epsilon,t}, \boldsymbol{D} = \boldsymbol{V}_{\epsilon,t}^{(l)}$ in Lemma 29, with probability at least $1 - d^{-3}$, we have

$$
\begin{aligned}
(\mathrm{I}_{3,1}) &= \left\langle \left(\mathcal{I} - \mathcal{R}_\Omega\right)\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top\right\rangle \\
&\le \Gamma\sqrt{\frac{d}{p}}\left\|\boldsymbol{A}\right\|_{2,\infty}\left\|\boldsymbol{B}\right\|_{\mathrm{F}} \cdot \left\|\boldsymbol{C}\right\|_{\mathrm{F}}\left\|\boldsymbol{D}\right\|_{2,\infty} \\
&= \Gamma\sqrt{\frac{d}{p}}\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{2,\infty}^2\left\|\mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp}\left(\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right)\right\|_{\mathrm{F}}\left\|\left(\boldsymbol{V}_{\epsilon,t}^{(l)} - \boldsymbol{V}_t^{(l)}\right)\boldsymbol{\Sigma}_{\epsilon,t}\right\|_{\mathrm{F}} \\
&\le 16\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2.
\end{aligned}
\tag{191}
$$

Here in the last inequality, we use the fact that $\|\boldsymbol{\Sigma}_{\epsilon,t}\| \le 4\sigma_1^\star$ and $\left\|\boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{2,\infty} \le 2\sqrt{\frac{\mu r}{d}}$. In a manner akin to our derivation for $(\mathrm{I}_{3,1})$, we can also show that

$$
\begin{aligned}
(\mathrm{I}_{3,2}) &\le 16\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}\left\|\boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)}\right\|_{\mathrm{F}}^2 \\
(\mathrm{I}_{3,3}) &\le 4\Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}\epsilon\left\|\boldsymbol{V}_{\epsilon,t}^{(l)} - \boldsymbol{V}_t^{(l)}\right\|_{\mathrm{F}}.
\end{aligned}
\tag{192}
$$

Lastly, for $(\mathrm{I}_{3,4})$, we observe that

$$\begin{aligned}
(\mathrm{I}_{3,4}) &= -\left\langle \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}^{\star} \boldsymbol{V}^{\star\top} \boldsymbol{V}_{\epsilon,t}^{(l)} \right\rangle \\
&\leq -\left\langle \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}^{\star} \right\rangle + \sigma_1^{\star} \left\| \boldsymbol{V}^{\star} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 \\
&\overset{(a)}{\leq} \sigma_1^{\star} \left\| \boldsymbol{V}^{\star} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 \\
&\leq \sigma_1^{\star} \left(\left\| \boldsymbol{V}^{\star} - \boldsymbol{V}_t^{(l)} \right\|_{\mathrm{F}} + \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \right) \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 \\
&\overset{(b)}{\leq} \sigma_1^{\star} \left(\Gamma_1 \frac{\kappa \mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} + \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \right) \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 .
\end{aligned} \tag{193}$$

Here in (a), we use the fact that $\left\langle \boldsymbol{V}_{\epsilon,t}^{(l)}, \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}^{\star} \right\rangle = \left\| \mathcal{P}_{\boldsymbol{V}_t^{(l)}}^{\perp} \boldsymbol{V}_{\epsilon,t}^{(l)} \boldsymbol{\Sigma}^{\star 1/2} \right\|_{\mathrm{F}}^2 \geq 0$. In (b), we apply Proposition 16.

Putting everything together, we obtain that

$$(\mathrm{I}_3) \leq \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \left(32\Gamma\sigma_1^{\star} \sqrt{\frac{\mu^2 r^2}{pd}} \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} + 4\Gamma\sigma_1^{\star} \sqrt{\frac{\mu^2 r^2}{pd}} \epsilon \right), \tag{194}$$

which in turn leads to

$$(\mathrm{I}) \leq \Gamma\sigma_1^{\star} \frac{\kappa \mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}}^2 + \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} \cdot 4\Gamma\sigma_1^{\star} \sqrt{\frac{\mu^2 r^2}{pd}} \epsilon. \tag{195}$$

Therefore, our final bound is established as

$$\left\| \boldsymbol{V}_{t+1}^{(l)} - \boldsymbol{V}_{\epsilon,t+1}^{(l)} \right\|_{\mathrm{F}} \leq \left(1 + \Gamma\eta\sigma_1^{\star} \frac{\kappa \mu r^{1.5} \log\left(\frac{1}{\alpha}\right)}{\sqrt{pd}} \right) \left\| \boldsymbol{V}_t^{(l)} - \boldsymbol{V}_{\epsilon,t}^{(l)} \right\|_{\mathrm{F}} + C_3 \eta \frac{\sigma_1^{\star} \mu r}{\sqrt{pd}} \epsilon. \tag{196}$$

This completes the proof. ■

Appendix F. Proofs for Different Initialization Schemes

Gaussian initialization. Suppose the search rank $0.5d \leq r' \leq d$. Let $\boldsymbol{Z} = \boldsymbol{G}/\left\|\boldsymbol{G}\right\|$ where $\boldsymbol{G} \in \mathbb{R}^{d \times r'}$ is a standard Gaussian matrix. Before proceeding, we first introduce the following lemma, which characterizes the concentration of the largest and smallest singular values of a standard Gaussian matrix.

Lemma 24 (Adapted from Theorem 6.1 in (Wainwright, 2019)) *Suppose that $\boldsymbol{G} \in \mathbb{R}^{d_1 \times d_2}$ is a standard Gaussian matrix where $d_1 \geq d_2$. Then, for any $\delta > 0$, we have*

$$\begin{aligned}
\mathbb{P}\left(\left\|\boldsymbol{G}\right\| \geq (2+\delta)\sqrt{d_1} \right) &\leq \exp\left\{ -\frac{d_1 \delta^2}{2} \right\}, \\
\mathbb{P}\left(\sigma_{\min}(\boldsymbol{G}) \geq (1-\delta)\sqrt{d_1} - \sqrt{d_2} \right) &\leq \exp\left\{ -\frac{d_1 \delta^2}{2} \right\}.
\end{aligned} \tag{197}$$

Notice that $\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{G}) = \sigma_{\min}(\boldsymbol{V}^{\star\top}\boldsymbol{G})$ where $\boldsymbol{V}^{\star\top}\boldsymbol{G} \in \mathbb{R}^{r\times r'}$ is another standard Gaussian matrix. Hence, upon setting $\delta = \frac{1}{2}$ and noting that $r' \geq \frac{d}{2}$, Lemma 24 implies that with probability at least $1 - \exp\{d/16\}$, we have

$$\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{G}) \geq \frac{1}{2}\sqrt{r'} - \sqrt{r} \geq \frac{1}{4}\sqrt{d}. \tag{198}$$

On the other hand, upon setting $\delta = \frac{1}{2}$, Lemma 24 implies that with probability at least $1-\exp\{d/8\}$, we have

$$\boldsymbol{G} \leq \frac{5}{2}\sqrt{d}. \tag{199}$$

Via a union bound, we know that with probability at least $1 - 2\exp\{d/16\}$, we have

$$\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{Z}) = \frac{\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{G})}{\|\boldsymbol{G}\|} \geq \frac{\sqrt{d}/4}{5\sqrt{d}/2} = \frac{1}{10}. \tag{200}$$

Hence, with probability at least $1 - 2\exp\{d/16\}$, Condition 1 is satisfied with $c_0 = 0.1$.

Orthogonal initialization. Suppose the search rank satisfies $r' = d$. Upon choosing $\boldsymbol{Z} = \boldsymbol{O}$ for some $\boldsymbol{O} \in \mathcal{O}_{d\times d}$, we have

$$\sigma_r(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{O}) = \sigma_r(\boldsymbol{V}^\star\boldsymbol{V}^{\star\top}) = 1. \tag{201}$$

Hence, Condition 1 is satisfied with $c_0 = 1$.

Spectral initialization. Let $\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ be the eigendecomposition of the best rank-r' approximation of $\mathcal{R}_\Omega(\boldsymbol{X}^\star)$ measured in Frobenius norm. Suppose $r \leq r' \leq d$ and $\boldsymbol{Z} = \boldsymbol{U}/\|\boldsymbol{U}\|$, where $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$. Corollary 30 tells us that, with probability at least $1 - \frac{1}{d^3}$, we have

$$\|\mathcal{R}_\Omega(\boldsymbol{X}^\star) - \boldsymbol{X}^\star\| \leq \Gamma\sigma_1^\star\sqrt{\frac{\mu^2 r^2}{pd}}. \tag{202}$$

Conditioned on this event, we have

$$\|\boldsymbol{U}\|^2 = \|\mathcal{R}_\Omega(\boldsymbol{X}^\star)\| \leq \|\boldsymbol{X}^\star\| + \|\mathcal{R}_\Omega(\boldsymbol{X}^\star) - \boldsymbol{X}^\star\| \leq 2\sigma_1^\star. \tag{203}$$

On the other hand, by Weyl's inequality, we have

$$\begin{aligned}
\sigma_r^2(\mathcal{P}_{\boldsymbol{V}^\star}\boldsymbol{U}) &= \sigma_r(\boldsymbol{V}^{\star\top}\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{V}^\star) \\
&\geq \sigma_r(\boldsymbol{V}^{\star\top}\mathcal{R}_\Omega(\boldsymbol{X}^\star)\boldsymbol{V}^\star) - \sigma_{r+1}(\mathcal{R}_\Omega(\boldsymbol{X}^\star)) \\
&\geq \sigma_r(\boldsymbol{V}^{\star\top}\boldsymbol{X}^\star\boldsymbol{V}^\star) - \|\mathcal{R}_\Omega(\boldsymbol{X}^\star) - \boldsymbol{X}^\star\| - \sigma_{r+1}(\mathcal{R}_\Omega(\boldsymbol{X}^\star)) \\
&\geq \sigma_r^\star - 2\|\mathcal{R}_\Omega(\boldsymbol{X}^\star) - \boldsymbol{X}^\star\| \\
&\geq 0.5\sigma_r^\star.
\end{aligned} \tag{204}$$

Combining the above two inequalities, we conclude that, with probability at least $1 - \frac{1}{d^3}$, Condition 1 is satisfied with $c_0 = \frac{1}{2\kappa}$.

Appendix G. Concentration Inequalities for Matrix Completion

Recall that the sampling matrix $\mathbf{\Omega} \in \mathbb{R}^{d\times d}$ is defined as

$$\Omega_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \tag{205}$$

The following lemma characterizes the concentration behavior of $\mathbf{\Omega}$.

Lemma 25 (Adapted from (Vu, 2018, Lemma 8)) *Suppose the sampling rate satisfies $p \gtrsim \frac{\log(d)}{d}$. There is a universal constant $\Gamma > 0$ such that, with probability at least $1 - \frac{1}{d^3}$, we have*

$$\left\| \frac{\mathbf{\Omega} + \mathbf{\Omega}^\top}{2p} - \boldsymbol{J} \right\| \leq \Gamma \sqrt{\frac{d}{p}}. \tag{206}$$

Here $\boldsymbol{J}$ is the all-one matrix.

The original result appeared in (Vu, 2018, Lemma 8) only holds for symmetric Bernoulli model, i.e., $\mathbf{\Omega} = \mathbf{\Omega}^\top$. However, we can easily extend it to the asymmetric case via the dilation trick (Tropp et al., 2015). Hence, we omit the details here. Next, we have the following extension to the leave-one-out sequences.

Corollary 26 *Suppose the sampling rate satisfies $p \gtrsim \frac{\log(d)}{d}$. Then, with probability at least $1 - \frac{1}{d^3}$, for all $1 \leq l \leq d$, we have*

$$\left\| \frac{\mathbf{\Omega}^{(l)} + \mathbf{\Omega}^{(l)\top}}{2p} - \boldsymbol{J} \right\| \leq \Gamma \sqrt{\frac{d}{p}}. \tag{207}$$

Proof of Corollary 26. Note that for any $1 \leq l \leq d$, the matrix $\frac{\mathbf{\Omega}^{(l)} + \mathbf{\Omega}^{(l)\top}}{2p} - \boldsymbol{J}$ can be derived from $\frac{\mathbf{\Omega} + \mathbf{\Omega}^\top}{2p} - \boldsymbol{J}$ by zeroing out the l-th row and column. Hence, the proof follows by invoking Lemma 34 in Lemma 25. ■

Lemma 27 ((Chen et al., 2020, Lemma A.1) and (Chen and Li, 2019, Lemma 8)) *For all $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D} \in \mathbb{R}^{d\times r}$, we have*

$$\left| \left\langle \left(\mathcal{I} - \mathcal{R}_\Omega\right)\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top \right\rangle \right| \leq \left\| \frac{\mathbf{\Omega} + \mathbf{\Omega}^\top}{2p} - \boldsymbol{J} \right\| \cdot \|\boldsymbol{A}\|_{2,\infty} \|\boldsymbol{B}\|_{\mathrm{F}} \cdot \|\boldsymbol{C}\|_{\mathrm{F}} \|\boldsymbol{D}\|_{2,\infty}. \tag{208}$$

Moreover, for all $1 \leq l \leq d$, we have

$$\left| \left\langle \left(\mathcal{I} - \mathcal{R}_{\Omega^{(l)}}\right)\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top \right\rangle \right| \leq \left\| \frac{\mathbf{\Omega}^{(l)} + \mathbf{\Omega}^{(l)\top}}{2p} - \boldsymbol{J} \right\| \cdot \|\boldsymbol{A}\|_{2,\infty} \|\boldsymbol{B}\|_{\mathrm{F}} \cdot \|\boldsymbol{C}\|_{\mathrm{F}} \|\boldsymbol{D}\|_{2,\infty}. \tag{209}$$

As a special case, we have

Lemma 28 *For any matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ with the form $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$, we have*

$$\|(\mathcal{I} - \mathcal{R}_{\Omega^{(l)}})(\boldsymbol{X})\| \le \|(\mathcal{I} - \mathcal{R}_{\Omega})(\boldsymbol{X})\| \le \left\|\frac{\boldsymbol{\Omega}+\boldsymbol{\Omega}^\top}{2p} - \boldsymbol{J}\right\| \|\boldsymbol{U}\|_{2,\infty} \|\boldsymbol{V}\|_{2,\infty}, \quad \forall 1 \le l \le d. \tag{210}$$

Note that the above two results are deterministic. Combining them with Lemma 25 leads to the following high-probability results.

Corollary 29 ((Chen et al., 2020, Lemma 4.3 and Lemma A.1)) *Suppose that the sampling rate satisfies $p \gtrsim \frac{\log(d)}{d}$. There exists a universal constant $C > 0$ such that, for any $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D} \in \mathbb{R}^{d\times r}$, with probability at least $1 - \frac{1}{d^3}$, we have*

$$\left|\left\langle (\mathcal{I} - \mathcal{R}_{\Omega})\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top \right\rangle\right| \le \Gamma\sqrt{\frac{d}{p}} \cdot \|\boldsymbol{A}\|_{2,\infty} \|\boldsymbol{B}\|_{\mathrm{F}} \cdot \|\boldsymbol{C}\|_{\mathrm{F}} \|\boldsymbol{D}\|_{2,\infty}. \tag{211}$$

Moreover, with the same probability, for any $1 \le l \le d$, we have

$$\left|\left\langle (\mathcal{I} - \mathcal{R}_{\Omega^{(l)}})\left(\boldsymbol{A}\boldsymbol{C}^\top\right), \boldsymbol{B}\boldsymbol{D}^\top \right\rangle\right| \le \Gamma\sqrt{\frac{d}{p}} \cdot \|\boldsymbol{A}\|_{2,\infty} \|\boldsymbol{B}\|_{\mathrm{F}} \cdot \|\boldsymbol{C}\|_{\mathrm{F}} \|\boldsymbol{D}\|_{2,\infty}. \tag{212}$$

Corollary 30 ((Chen and Li, 2019, Lemma 9)) *Consider an arbitrary matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ decomposed as $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$. There exists a universal constant C such that, with probability at least $1 - \frac{1}{d^3}$, we have*

$$\left\|\left(\mathcal{I} - \mathcal{R}_{\Omega}^{(l)}\right)(\boldsymbol{X})\right\| \le \|(\mathcal{I} - \mathcal{R}_{\Omega})(\boldsymbol{X})\| \le \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{U}\|_{2,\infty} \|\boldsymbol{V}\|_{2,\infty}. \tag{213}$$

Lastly, we provide a finer result for a matrix of the form $\boldsymbol{X} - \boldsymbol{Y}$.

Lemma 31 *For two arbitrary symmetric matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d\times d}$, with probability at least $1 - \frac{1}{d^3}$, we have*

$$\left\|\left(\mathcal{I} - \mathcal{R}_{\Omega}^{(l)}\right)(\boldsymbol{X}-\boldsymbol{Y})\right\| \le \|(\mathcal{I} - \mathcal{R}_{\Omega})(\boldsymbol{X}-\boldsymbol{Y})\| \le \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{X}-\boldsymbol{Y}\| \left(\|\boldsymbol{U}_{\boldsymbol{X}}\|_{2,\infty}^2 + \|\boldsymbol{U}_{\boldsymbol{Y}}\|_{2,\infty}^2\right). \tag{214}$$

Proof We denote $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{Y}$ with an SVD $\boldsymbol{Z} = \boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{L}_{\boldsymbol{Z}}^\top$ (recall that $\boldsymbol{Z}$ is symmetric). Hence, with probability at least $1 - \frac{1}{d^3}$, we have

$$\begin{aligned}
\|(\mathcal{I} - \mathcal{R}_{\Omega})(\boldsymbol{Z})\| &\overset{(a)}{\le} \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\|_{2,\infty} \|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty} \\
&\overset{(b)}{\le} \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{\Sigma}_{\boldsymbol{Z}}\| \|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty}^2 \\
&= \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{Z}\| \|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty}^2 \\
&\overset{(c)}{\le} \Gamma\sqrt{\frac{d}{p}} \|\boldsymbol{Z}\| \left(\|\boldsymbol{L}_{\boldsymbol{X}}\|_{2,\infty}^2 + \|\boldsymbol{L}_{\boldsymbol{Y}}\|_{2,\infty}^2\right).
\end{aligned} \tag{215}$$

Here in (a), we apply Lemma 30 upon setting $\boldsymbol{U} = \boldsymbol{L_Z}\boldsymbol{\Sigma_Z}$ and $\boldsymbol{V} = \boldsymbol{L_Z}$. In (b), we apply Lemma 33. Lastly, in (c), we apply Lemma 39. This completes the proof. ■

Appendix H. Auxiliary Lemmas

H.1. Concentration Inequalities

Lemma 32 (Bernstein's inequality) *Let $X_1, \cdots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \le M$ almost surely, for all i and set $\nu^2 = \sum_{i=1}^n \mathrm{Var}\left[X_i^2\right]$. Then, for all positive t,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \ge t\right) \le 2\exp\left(-\frac{t^2}{2\nu^2 + \frac{2}{3}Mt}\right). \tag{216}$$

Or equivalently, with probability at least $1-\delta$, one has

$$\left|\sum_{i=1}^n X_i\right| \le 2\nu\sqrt{\log\left(2/\delta\right)} + \frac{4}{3}M\log\left(2/\delta\right). \tag{217}$$

H.2. Matrix Norm Inequalities

Lemma 33 ((Cape et al., 2019, Proposition 6.5)) *For $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$, and $\boldsymbol{B} \in \mathbb{R}^{d_2\times d_3}$, we have*

$$\|\boldsymbol{AB}\|_{2,\infty} \le \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\| \quad \text{and} \quad \|\boldsymbol{AB}\|_{2,\infty} \le \|\boldsymbol{A}\|_{\infty}\|\boldsymbol{B}\|_{2,\infty}. \tag{218}$$

Lemma 34 (Adapted from (Sun and Luo, 2016, Proposition A.3)) *For any matrix $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}$, denote $\boldsymbol{A}_{-i,\cdot}$ ($\boldsymbol{A}_{\cdot,-i}$) as the matrix obtained by replacing the i-th row (column) of $\boldsymbol{A}$ by zeros, respectively. Then, we have*

$$\|\boldsymbol{A}_{-i,\cdot}\| \le \|\boldsymbol{A}\| \quad \text{and} \quad \|\boldsymbol{A}_{\cdot,-j}\| \le \|\boldsymbol{A}\|, \forall i \in [d_1], j \in [d_2]. \tag{219}$$

Lemma 35 ((Sun and Luo, 2016, Proposition A.4)) *For any two matrices $\boldsymbol{A} \in \mathbb{R}^{d_1\times d_2}, \boldsymbol{B} \in \mathbb{R}^{d_2\times d_3}$, we have*

$$\|\boldsymbol{AB}\| \le \|\boldsymbol{A}\|\|\boldsymbol{B}\| \quad \text{and} \quad \|\boldsymbol{AB}\|_{\mathrm{F}} \le \|\boldsymbol{A}\|\|\boldsymbol{B}\|_{\mathrm{F}}. \tag{220}$$

Furthermore, if $d_1 \ge d_2$, then

$$\sigma_{\min}(\boldsymbol{A})\|\boldsymbol{B}\|_{\mathrm{F}} \le \|\boldsymbol{AB}\|_{\mathrm{F}} \quad \text{and} \quad \sigma_{\min}(\boldsymbol{A})\|\boldsymbol{B}\| \le \|\boldsymbol{AB}\|. \tag{221}$$

Lemma 36 *For arbitrary matrices $\boldsymbol{U} \in \mathbb{R}^{d_1\times d_2}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{d_2\times d_3}$ and $\boldsymbol{V} \in \mathbb{R}^{d_4\times d_3}$, we have*

$$\left\|\boldsymbol{U\Sigma V}^\top\right\|_{\max} \le \|\boldsymbol{\Sigma}\|\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}. \tag{222}$$

Proof By Cauchy-Schwartz inequality, we have

$$\begin{aligned}\left\|\boldsymbol{U\Sigma V}^\top\right\|_{\max} &= \max_{i,j}\left|\sum_k [\boldsymbol{U\Sigma}]_{i,k}V_{j,k}\right| \\ &\le \max_{i,j}\|(\boldsymbol{U\Sigma})_{i,\cdot}\|\,\|\boldsymbol{V}_{j,\cdot}\| \\ &= \|\boldsymbol{U\Sigma}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty} \\ &\le \|\boldsymbol{\Sigma}\|\|\boldsymbol{U}\|_{2,\infty}\|\boldsymbol{V}\|_{2,\infty}.\end{aligned} \tag{223}$$

Here we apply Lemma 33 to derive the last inequality. This completes the proof. ■

Lemma 37 *For any matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$, we have*

$$\|\boldsymbol{A}\|_{2,\infty}^2 = \left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max}. \tag{224}$$

Proof We write $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_{d_1}$ as the row vectors of $\boldsymbol{A}$. Then, we have

$$\|\boldsymbol{A}\|_{2,\infty}^2 = \max_{1\le i\le d_1} \|\boldsymbol{a}_i\|^2. \tag{225}$$

On the other hand, we can write $\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max}$ as

$$\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max} = \max_{1\le i,j\le d_1} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle|. \tag{226}$$

First, we have $\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max} \le \|\boldsymbol{A}\|_{2,\infty}^2$ since

$$\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max} = \max_{1\le i,j\le d_1} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle| \le \max_{1\le i,j\le d_1} \|\boldsymbol{a}_i\|\,\|\boldsymbol{a}_j\| = \max_{1\le i\le d_1} \|\boldsymbol{a}_i\|^2 = \|\boldsymbol{A}\|_{2,\infty}^2. \tag{227}$$

Second, we have $\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max} \ge \|\boldsymbol{A}\|_{2,\infty}^2$ upon noting that

$$\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max} = \max_{1\le i,j\le d_1} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle| \ge \max_{i=j} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle| = \max_{1\le i\le d_1} \|\boldsymbol{a}_i\|^2 = \|\boldsymbol{A}\|_{2,\infty}^2. \tag{228}$$

Therefore, we derive that $\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\|_{\max} = \|\boldsymbol{A}\|_{2,\infty}^2$, which completes the proof. ■

Lemma 38 *For two PSD matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$ with $\boldsymbol{A} \preceq \boldsymbol{B}$, we have*

$$\|\boldsymbol{A}\|_{\max} \le \|\boldsymbol{B}\|_{\max}. \tag{229}$$

Proof The proof follows by the fact that for any PSD matrix $\boldsymbol{A}$, we have $\|\boldsymbol{A}\|_{\max} = \max_i\{A_{i,i}\}$. According to this fact, we immediately have

$$\|\boldsymbol{A}\|_{\max} = \max_i\{A_{i,i}\} \le \max_i\{B_{i,i}\} = \|\boldsymbol{B}\|_{\max} \tag{230}$$

since $\boldsymbol{A} \preceq \boldsymbol{B}$. Now we turn to prove this fact. Note that we can write any PSD matrix $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{P}^\top$. Then, according to Lemma 37, we have

$$\|\boldsymbol{A}\|_{\max} = \|\boldsymbol{P}\|_{2,\infty}^2 = \max_i \|\boldsymbol{p}_i\|^2 = \max_i\{A_{i,i}\}. \tag{231}$$

Here we write $\{\boldsymbol{p}_i\}$ as the row vectors of $\boldsymbol{P}$. This completes the proof. ■

Lemma 39 *For $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{Y}$, we have*

$$\|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty}^2 \le \|\boldsymbol{L}_{\boldsymbol{X}}\|_{2,\infty}^2 + \|\boldsymbol{L}_{\boldsymbol{Y}}\|_{2,\infty}^2. \tag{232}$$

Proof We bound $\|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty}^2$ as follows

$$\begin{aligned}\|\boldsymbol{L}_{\boldsymbol{Z}}\|_{2,\infty}^2 &\overset{(a)}{=} \left\|\boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{L}_{\boldsymbol{Z}}^\top\right\|_{\max}\\ &\overset{(b)}{\leq} \left\|\boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{L}_{\boldsymbol{X}}^\top + \boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{L}_{\boldsymbol{Y}}^\top\right\|_{\max}\\ &\leq \left\|\boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{L}_{\boldsymbol{X}}^\top\right\|_{\max} + \left\|\boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{L}_{\boldsymbol{Y}}^\top\right\|_{\max}\\ &\overset{(c)}{=} \|\boldsymbol{L}_{\boldsymbol{X}}\|_{2,\infty}^2 + \|\boldsymbol{L}_{\boldsymbol{Y}}\|_{2,\infty}^2\,.\end{aligned} \tag{233}$$

Here (a) and (c) follow from Lemma 37. In (b), we apply Lemma 38 since $\boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{L}_{\boldsymbol{Z}}^\top \preceq \boldsymbol{L}_{\boldsymbol{X}}\boldsymbol{L}_{\boldsymbol{X}}^\top + \boldsymbol{L}_{\boldsymbol{Y}}\boldsymbol{L}_{\boldsymbol{Y}}^\top$. This is due to the fact that $\mathrm{col}(\boldsymbol{Z}) \subseteq \mathrm{col}(\boldsymbol{X}) \oplus \mathrm{col}(\boldsymbol{Y})$ which leads to $\boldsymbol{L}_{\boldsymbol{Z}}\boldsymbol{L}_{\boldsymbol{Z}}^\top = \mathcal{P}_{\boldsymbol{Z}} \preceq \mathcal{P}_{\boldsymbol{X}} + \mathcal{P}_{\boldsymbol{Y}}$. This completes the proof. ■

Lemma 40 ((Tu et al., 2016, Lemma 5.4)) *For any $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d\times r}$ with $\sigma_r(\boldsymbol{X}) > 0$, we have*

$$\mathrm{dist}^2(\boldsymbol{X}, \boldsymbol{Y}) \leq \frac{1}{2\left(\sqrt{2}-1\right)\sigma_r^2(\boldsymbol{X})}\left\|\boldsymbol{X}\boldsymbol{X}^\top - \boldsymbol{Y}\boldsymbol{Y}^\top\right\|_{\mathrm{F}}^2. \tag{234}$$

Lemma 41 *Consider a matrix $\boldsymbol{U} \in \mathbb{R}^{d_1\times d_2}$ and a diagonal matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d_2\times d_2}$. We have*

$$\|\boldsymbol{U}\boldsymbol{\Sigma}\|_{2,\infty} \geq \sigma_r(\boldsymbol{\Sigma})\,\|\boldsymbol{U}\|_{2,\infty}\,. \tag{235}$$

Proof We first write $\boldsymbol{\Sigma} = \mathrm{Diag}\{\sigma_1, \cdots, \sigma_{d_2}\}$. Next, note that

$$\|(\boldsymbol{U}\boldsymbol{\Sigma})_{l,\cdot}\|^2 = \sum_{j=1}^{d_2} U_{l,j}^2\sigma_j^2 \geq \sigma_r^2(\boldsymbol{\Sigma})\sum_{j=1}^{d_2} U_{l,j}^2 = \sigma_r^2(\boldsymbol{\Sigma})\,\|\boldsymbol{U}_{l,\cdot}\|^2\,. \tag{236}$$

Hence, taking the maximum over index l on both sides, we immediately obtain

$$\|\boldsymbol{U}\boldsymbol{\Sigma}\|_{2,\infty} \geq \sigma_r(\boldsymbol{\Sigma})\,\|\boldsymbol{U}\|_{2,\infty}\,. \tag{237}$$

■

H.3. Other Useful Inequalities

Lemma 42 (Adapted from (Vershynin, 2018, Corollary 4.2.13)) *The covering number $\mathcal{N}_\epsilon$ of $\mathcal{B}_{\mathrm{F}}^{r\times r}(R)$ satisfies the following inequality for any $0 < \epsilon \leq 1$:*

$$\mathcal{N}_\epsilon \leq \left(\frac{3R}{\epsilon}\right)^{r^2}. \tag{238}$$

Lemma 43 *For two orthogonal matrices $\boldsymbol{V}_1, \boldsymbol{V}_2 \in \mathcal{O}_{d\times r}$, we have*

$$\left\|\boldsymbol{V}_1\boldsymbol{V}_1^\top - \boldsymbol{V}_2\boldsymbol{V}_2^\top\right\| \leq 2\,\|\boldsymbol{V}_1 - \boldsymbol{V}_2\| \quad \textit{and} \quad \left\|\boldsymbol{V}_1\boldsymbol{V}_1^\top - \boldsymbol{V}_2\boldsymbol{V}_2^\top\right\|_{\mathrm{F}} \leq 2\,\|\boldsymbol{V}_1 - \boldsymbol{V}_2\|_{\mathrm{F}}\,. \tag{239}$$

Proof Note that $\boldsymbol{V}_1\boldsymbol{V}_1^\top - \boldsymbol{V}_2\boldsymbol{V}_2^\top = \boldsymbol{V}_1\left(\boldsymbol{V}_1 - \boldsymbol{V}_2\right)^\top + \left(\boldsymbol{V}_1 - \boldsymbol{V}_2\right)\boldsymbol{V}_2^\top$. Hence, we have

$$\left\|\boldsymbol{V}_1\boldsymbol{V}_1^\top - \boldsymbol{V}_2\boldsymbol{V}_2^\top\right\| \le \left\|\boldsymbol{V}_1 - \boldsymbol{V}_2\right\|\left(\left\|\boldsymbol{V}_1\right\| + \left\|\boldsymbol{V}_2\right\|\right) \le 2\left\|\boldsymbol{V}_1 - \boldsymbol{V}_2\right\|. \tag{240}$$

Similarly, for the Frobenius norm, we also have

$$\left\|\boldsymbol{V}_1\boldsymbol{V}_1^\top - \boldsymbol{V}_2\boldsymbol{V}_2^\top\right\|_{\mathrm{F}} \le \left\|\boldsymbol{V}_1 - \boldsymbol{V}_2\right\|_{\mathrm{F}}\left(\left\|\boldsymbol{V}_1\right\| + \left\|\boldsymbol{V}_2\right\|\right) \le 2\left\|\boldsymbol{V}_1 - \boldsymbol{V}_2\right\|_{\mathrm{F}}. \tag{241}$$

This completes the proof. ■

Lemma 44 *For arbitrary matrix $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2}$ with $\mathrm{rank}(\boldsymbol{X}) = r$ and $\boldsymbol{O} \in \mathcal{O}_{d_1\times r}$, we have*

$$\sigma_r(\boldsymbol{X}) \ge \sigma_r\Big(\boldsymbol{O}^\top\boldsymbol{X}\Big). \tag{242}$$

Moreover, if $\boldsymbol{O} = \boldsymbol{U}_{\boldsymbol{X}}$, we have $\sigma_i(\boldsymbol{X}) = \sigma_i\left(\boldsymbol{O}^\top\boldsymbol{X}\right)$ for all $1 \le i \le r$.

Proof We first prove the special case. Suppose $\boldsymbol{O} = \boldsymbol{U}_{\boldsymbol{X}}$, then we have $\boldsymbol{O}^\top\boldsymbol{X} = \boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{V}_{\boldsymbol{X}}^\top$. Note that this is the SVD of $\boldsymbol{O}^\top\boldsymbol{X}$ with the singular value matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. Hence, $\boldsymbol{O}^\top\boldsymbol{X}$ has the same singluar values as $\boldsymbol{X}$. For the general case, it follows by

$$\begin{aligned}
\sigma_r\Big(\boldsymbol{O}^\top\boldsymbol{X}\Big) &= \inf_{\boldsymbol{Y}\in\mathcal{M}_{\le r-1}}\left\|\boldsymbol{O}^\top\boldsymbol{X} - \boldsymbol{Y}\right\|_{\mathrm{F}} \\
&\le \left\|\boldsymbol{O}^\top\boldsymbol{X} - \boldsymbol{O}^\top\boldsymbol{U}_{\boldsymbol{X}}\boldsymbol{\Sigma}_{\boldsymbol{X},-1}\boldsymbol{V}_{\boldsymbol{X}}^\top\right\|_{\mathrm{F}} \\
&\le \left\|\boldsymbol{O}^\top\sigma_r(\boldsymbol{X})\boldsymbol{u}_{\boldsymbol{X},-1}\boldsymbol{v}_{\boldsymbol{X},-1}^\top\right\|_{\mathrm{F}} \\
&\le \sigma_r(\boldsymbol{X}).
\end{aligned} \tag{243}$$

Here we define $\mathcal{M}_{\le r-1}$ as the set of all matrices of rank at most $r-1$. We also denote $\boldsymbol{\Sigma}_{\boldsymbol{X},-1} = \mathrm{Diag}\{\sigma_1^\star(\boldsymbol{X}), \cdots, \sigma_{r-1}(\boldsymbol{X}), 0\}$. Finally, $\boldsymbol{u}_{\boldsymbol{X},-1}, \boldsymbol{v}_{\boldsymbol{X},-1}$ refer to the last columns of $\boldsymbol{U}_{\boldsymbol{X}}, \boldsymbol{V}_{\boldsymbol{X}}$, respectively.

■

Lemma 45 *Consider two matrices $\boldsymbol{A} \in \mathbb{R}^{r\times r}, \boldsymbol{B} \in \mathbb{R}^{r\times d}$ where $\boldsymbol{A}$ is invertible and $\boldsymbol{B} \neq \boldsymbol{0}$. We have*

$$\sigma_r(\boldsymbol{A}\boldsymbol{B}) \ge \sigma_r(\boldsymbol{A})\sigma_r(\boldsymbol{B}). \tag{244}$$

Proof It directly follows by

$$\begin{aligned}
\sigma_r(\boldsymbol{A}\boldsymbol{B}) &= \min_{\|\boldsymbol{x}\|=1, \boldsymbol{x}\in\mathrm{range}(\boldsymbol{B}^\top)}\|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\| \\
&\ge \sigma_r(\boldsymbol{A})\min_{\|\boldsymbol{x}\|=1, \boldsymbol{x}\in\mathrm{range}(\boldsymbol{B}^\top)}\|\boldsymbol{B}\boldsymbol{x}\| \\
&= \sigma_r(\boldsymbol{A})\sigma_r(\boldsymbol{B}).
\end{aligned} \tag{245}$$

■

Lemma 46 *For any $x \in [0, \frac{1}{2(r-1)})$ and $r > 1$, we have*

$$(1+x)^r \leq 1 + 2rx. \tag{246}$$

Lemma 47 *For the series $\{x_t\}_{t=0}^{\infty}$, the following two statements hold:*

- *Suppose that $x_{t+1} \leq Ax_t + B, \forall t \geq 0$ where $A > 0, A \neq 1$ and $x_0 + \frac{B}{A-1} \geq 0$. Then, we have*
$$x_t \leq A^t \left(x_0 + \frac{B}{A-1} \right) - \frac{B}{A-1}. \tag{247}$$
- *Suppose that $x_{t+1} \geq Ax_t - B, \forall t \geq 0$ where $A > 0$ and $x_0 - \frac{B}{A-1} \geq 0$. Then, we have*
$$x_t \geq A^t \left(x_0 - \frac{B}{A-1} \right) + \frac{B}{A-1}. \tag{248}$$

Proof We first prove the first statement. Note that we can rewrite $x_{t+1} \leq Ax_t + B$ as $y_{t+1} \leq Ay_t$ where $y_t = x_t + \frac{B}{A-1}$. Note that $y_0 \geq 0, A > 0$ by our assumption. Hence, we derive $y_t \leq A^t y_0 = A^t \left(x_0 + \frac{B}{A-1} \right)$, which implies that $x_t = y_t - \frac{B}{A-1} \leq A^t \left(x_0 + \frac{B}{A-1} \right) - \frac{B}{A-1}$.

For the second statement, we first rewrite the condition as $y_{t+1} \geq Ay_t$ where $y_t = x_t - \frac{B}{A-1}$. Then, we have $y_t \geq A^t y_0$, which implies that $x_t = y_t + \frac{B}{A-1} \geq A^t y_0 + \frac{B}{A-1} = A^t \left(x_0 - \frac{B}{A-1} \right) + \frac{B}{A-1}$. ■