The Power of Preconditioning in Overparameterized Low-Rank Matrix Sensing

Xingyu Xu* CMU
Yandi Shen† UChicago
Yuejie Chi* CMU
Cong Ma† UChicago

February 2, 2023

Abstract

We propose ScaledGD(λ), a preconditioned gradient descent method to tackle the low-rank matrix sensing problem when the true rank is unknown, and when the matrix is possibly ill-conditioned. Using overparametrized factor representations, ScaledGD(λ) starts from a small random initialization, and proceeds by gradient descent with a specific form of *damped* preconditioning to combat bad curvatures induced by overparameterization and ill-conditioning. At the expense of light computational overhead incurred by preconditioners, ScaledGD(λ) is remarkably robust to ill-conditioning compared to vanilla gradient descent (GD) even with overprameterization. Specifically, we show that, under the Gaussian design, ScaledGD(λ) converges to the true low-rank matrix at a constant linear rate after a small number of iterations that scales only *logarithmically* with respect to the condition number and the problem dimension. This significantly improves over the convergence rate of vanilla GD which suffers from a polynomial dependency on the condition number. Our work provides evidence on the power of preconditioning in accelerating the convergence without hurting generalization in overparameterized learning.

Keywords: low-rank matrix sensing, overparameterization, preconditioned gradient descent method, random initialization, ill-conditioning

1 Introduction

Low-rank matrix recovery plays an essential role in modern machine learning and signal processing. To fix ideas, let us consider estimating a rank-$r_\star$ positive semidefinite matrix $M_\star \in \mathbb{R}^{n\times n}$ based on a few linear measurements $y \coloneqq \mathcal{A}(M_\star)$, where $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ models the measurement process. Significant research efforts have been devoted to tackling low-rank matrix recovery in a statistically and computationally efficient manner in recent years. Perhaps the most well-known method is convex relaxation (Candès and Plan, 2011; Davenport and Romberg, 2016; Recht et al., 2010), which seeks the matrix with lowest nuclear norm to fit the observed measurements:

$$\min_{M\succeq 0} \quad \|M\|_* \qquad \text{s.t.} \quad y = \mathcal{A}(M).$$

While statistically optimal, convex relaxation is prohibitive in terms of both computation and memory as it directly operates in the ambient matrix domain, i.e., $\mathbb{R}^{n\times n}$. To address this challenge, nonconvex approaches based on low-rank factorization have been proposed (Burer and Monteiro, 2005):

$$\min_{X\in\mathbb{R}^{n\times r}} \quad \frac{1}{4}\big\|\mathcal{A}(XX^\top) - y\big\|_2^2, \tag{1}$$

where r is a user-specified rank parameter. Despite nonconvexity, when the rank is correctly specified, i.e., when $r = r_\star$, the problem (1) admits computationally efficient solvers (Chi et al., 2019), e.g., gradient descent (GD) with spectral initialization or with small random initialization. However, two main challenges remain when applying the factorization-based nonconvex approach (1) in practice.

*Carnegie Mellon University; Email: `{xingyuxu,yuejiec}@andrew.cmu.edu`.

†University of Chicago; Email: `{ydshen,congm}@uchicago.edu`.

parameterization	reference	algorithm	init.	iteration complexity
$r > r_\star$	Stöger and Soltanolkotabi (2021)	GD	random	$\kappa^8 + \kappa^6 \log(\kappa n/\varepsilon)$
	Zhang et al. (2021)	PrecGD	spectral	$\log(1/\varepsilon)$
	Theorem 2	ScaledGD(λ)	random	$\log \kappa \cdot \log(\kappa n) + \log(1/\varepsilon)$
$r = r_\star$	Tong et al. (2021a)	ScaledGD	spectral	$\log(1/\varepsilon)$
	Stöger and Soltanolkotabi (2021)	GD	random	$\kappa^8 \log(\kappa n) + \kappa^2 \log(1/\varepsilon)$
	Theorem 3	ScaledGD(λ)	random	$\log \kappa \cdot \log(\kappa n) + \log(1/\varepsilon)$

Table 1: Comparison of iteration complexity with existing algorithms for low-rank matrix sensing under Gaussian designs. Here, n is the matrix dimension, $r_\star$ is the true rank, r is the overparameterized rank, and κ is the condition number of the problem instance (see Section 2 for a formal problem formulation). It is important to note that in the overparameterized setting ($r > r_\star$), the sample complexity of Zhang et al. (2021) scales polynomially with the overparameterized rank r, while that of Stöger and Soltanolkotabi (2021) and ours only scale polynomially with the true rank $r_\star$.

- **Unknown rank**. First, the true rank $r_\star$ is often unknown, which makes it infeasible to set $r = r_\star$. One necessarily needs to consider an overparametrized setting in which r is set conservatively, i.e., one sets $r \geq r_\star$ or even $r = n$.
- **Poor conditioning**. Second, the ground truth matrix $M_\star$ may well be ill-conditioned, which is commonly encountered in practice. Existing approaches such as gradient descent are still computationally expensive in such settings as the number of iterations necessary for convergence increases with the condition number.

In light of these two challenges, the main goal of this work is to address the following question: *Can one develop an efficient method for solving ill-conditioned matrix recovery in the overparametrized setting?*

1.1 Our contributions: a preview

The main contribution of the current paper is to answer the question affirmatively by developing a *preconditioned* gradient descent method (ScaledGD(λ)) that converges to the (possibly ill-conditioned) low-rank matrix in a fast and global manner, even with overparamterized rank $r \geq r_\star$.

Theorem 1 (Informal)**.** *Under overparameterization $r \geq r_\star$ and mild statistical assumptions,* ScaledGD(λ)*—when starting from a sufficiently small random initialization—achieves a relative ε-accuracy, i.e., $\|X_t X_t^\top - M_\star\|_{\mathsf{F}} \leq \varepsilon \|M_\star\|$, with no more than an order of*

$$\log \kappa \cdot \log(\kappa n) + \log(1/\varepsilon)$$

iterations, where κ is the condition number of the problem.

The above theorem suggests that from a small random initialization, ScaledGD(λ) converges at a constant linear rate—independent of the condition number—after a small logarithmic number of iterations. Overall, the iteration complexity is nearly independent of the condition number and the problem dimension, making it extremely suitable for solving large-scale and ill-conditioned problems. See Table 1 for a summary of comparisons with prior art.

Our algorithm ScaledGD(λ) is closely related to scaled gradient descent (ScaledGD) (Tong et al., 2021a), a recently proposed preconditioned gradient descent method that achieves a κ-independent convergence rate under spectral initialization and exact parameterization. Preserving its low computational overhead, we modify the preconditioner design by introducing a fixed damping term, which prevents the preconditioner itself from being ill-conditioned due to overparameterization. In the exact parameterization setting, our

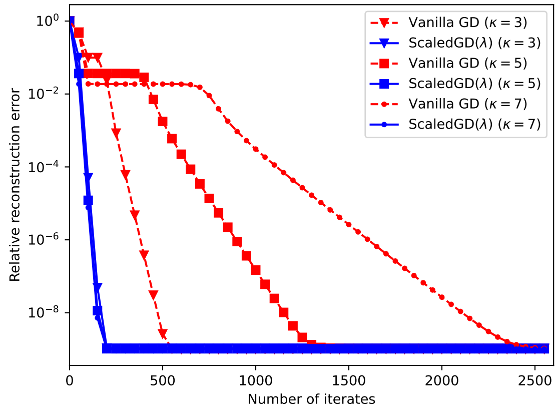

Figure 1: Comparison between ScaledGD(λ) and GD. The learning rate of GD has been fine-tuned to achieve fastest convergence for each κ, while that of ScaledGD(λ) is fixed to 0.3. The initialization scale α in each case has been fine-tuned so that the final accuracy is 10^{-9}. The details of the experiment are deferred to Section 5.

result extends ScaledGD beyond local convergence by characterizing the number of iterations it takes to enter the local basin of attraction from a small random initialization.

Moreover, our results shed light on the power of preconditioning in accelerating the optimization process over vanilla GD while still guaranteeing generalization in overparameterized learning models (Amari et al., 2020). Remarkably, despite the existence of an infinite number of global minima in the landscape of (1) that do not generalize, i.e., not corresponding to the ground truth, starting from a small random initialization, GD (Li et al., 2018; Stöger and Soltanolkotabi, 2021) is known to converge to a generalizable solution without explicit regularization. However, GD takes $O(\kappa^8 + \kappa^6 \log(\kappa n/\varepsilon))$ iterations to reach ε-accuracy, which is unacceptable even for moderate condition numbers. On the other hand, while common wisdom suggests that preconditioning accelerates convergence, it is yet unclear if it still converges to a generalizable global minimum. Our work answers this question in the affirmative for overparameterized low-rank matrix sensing, where ScaledGD(λ) significantly accelerates the convergence against the poor condition number—both in the initial phase and in the local phase—without hurting generalization, which is corroborated in Figure 1.

1.2 Related work

Significant efforts have been devoted to understanding nonconvex optimization for low-rank matrix estimation in recent years, see Chi et al. (2019) and Chen and Chi (2018) for recent overviews. By reparameterizing the low-rank matrix into a product of factor matrices, also known as the Burer-Monteiro factorization (Burer and Monteiro, 2005), the focus point has been examining if the factor matrices can be recovered—up to invertible transformations—faithfully using simple iterative algorithms in a provably efficient manner. However, the majority of prior efforts suffer from the limitations that they assume an exact parameterization where the rank of the ground truth is given or estimated somewhat reliably, and rely on a carefully constructed initialization (e.g., using the spectral method (Chen et al., 2021)) in order to guarantee global convergence in a polynomial time. The analyses adopted in the exact parameterization case fail to generalize when overparameterization presents, and drastically new approaches are called for.

Overparameterization in low-rank matrix sensing. Li et al. (2018) made a theoretical breakthrough that showed that gradient descent converges globally to any prescribed accuracy even in the presence of full overparameterization ($r = n$), with a small random initialization, where their analyses were subsequently adapted and extended in Stöger and Soltanolkotabi (2021) and Zhuo et al. (2021). Ding et al. (2021) investigated robust low-rank matrix recovery with overparameterization from a spectral initialization, and Ma and Fattahi (2022) examined the same problem from a small random initialization with noisy measurements.

Zhang et al. (2022, 2021) developed a preconditioned gradient descent method for overparameterized low-rank matrix sensing. Last but not least, a number of other notable works that study overparameterized low-rank models include, but are not limited to, Geyer et al. (2020); Oymak and Soltanolkotabi (2019); Soltanolkotabi et al. (2018); Zhang (2021, 2022).

Global convergence from random initialization without overparameterization. Despite nonconvexity, it has been established recently that several structured learning models admit global convergence via simple iterative methods even when initialized randomly even without overparameterization. For example, Chen et al. (2019) showed that phase retrieval converges globally from a random initialization using a near-minimal number of samples through a delicate leave-one-out analysis. In addition, the efficiency of randomly initialized GD is established for complete dictionary learning (Bai et al., 2018; Gilboa et al., 2019), multi-channel sparse blind deconvolution (Qu et al., 2019; Shi and Chi, 2021), asymmetric low-rank matrix factorization (Ye and Du, 2021), and rank-one matrix completion (Kim and Chung, 2022). Moving beyond GD, Lee and Stöger (2022) showed that randomly initialized alternating least-squares converges globally for rank-one matrix sensing, whereas Chandrasekher et al. (2022) developed sharp recovery guarantees of alternating minimization for generalized rank-one matrix sensing with sample-splitting and random initialization.

Algorithmic or implicit regularization. Our work is related to the phenomenon of algorithmic or implicit regularization (Gunasekar et al., 2017), where the trajectory of simple iterative algorithms follows a path that maintains desirable properties without explicit regularization. Along this line, Chen et al. (2020); Li et al. (2021); Ma et al. (2019) highlighted the implicit regularization of GD for several statistical estimation tasks, Ma et al. (2021) showed that GD automatically balances the factor matrices in asymmetric low-rank matrix sensing, where Jiang et al. (2022) analyzed the algorithmic regularization in overparameterized asymmetric matrix factorization in a model-free setting.

2 Problem formulation

Section 2.1 introduces the problem of low-rank matrix sensing, and Section 2.2 provides background on the proposed ScaledGD(λ) algorithm developed for the possibly overparametrized case.

2.1 Model and assumptions

Suppose that the ground truth $M_\star \in \mathbb{R}^{n\times n}$ is a positive-semidefinite (PSD) matrix of rank $r_\star \ll n$, whose (compact) eigendecomposition is given by

$$M_\star = U_\star \Sigma_\star^2 U_\star^\top.$$

Here, the columns of $U_\star \in \mathbb{R}^{n\times r_\star}$ specify the set of eigenvectors, and $\Sigma_\star \in \mathbb{R}^{r_\star \times r_\star}$ is a diagonal matrix where the diagonal entries are ordered in a non-increasing fashion. Setting $X_\star := U_\star \Sigma_\star \in \mathbb{R}^{n\times r_\star}$, we can rewrite $M_\star$ as

$$M_\star = X_\star X_\star^\top. \tag{2}$$

We call $X_\star$ the ground truth low-rank factor matrix, whose condition number κ is defined as

$$\kappa := \frac{\sigma_{\max}(X_\star)}{\sigma_{\min}(X_\star)}. \tag{3}$$

Here we recall that $\sigma_{\max}(X_\star)$ and $\sigma_{\min}(X_\star)$ are the largest and the smallest singular values of $X_\star$, respectively.

Instead of having access to $M_\star$ directly, we wish to recover $M_\star$ from a set of random linear measurements $\mathcal{A}(M_\star)$, where $\mathcal{A} : \mathrm{Sym}_2(\mathbb{R}^n) \to \mathbb{R}^m$ is a linear map from the space of $n \times n$ symmetric matrices to $\mathbb{R}^m$, namely

$$y = \mathcal{A}(M_\star), \tag{4}$$

or equivalently,

$$y_i = \langle A_i, M_\star \rangle, \qquad 1 \le i \le m.$$

We are interested in recovering $M_\star$ based on the measurements y and the sensing operator $\mathcal{A}$ in a provably efficient manner, even when the true rank $r_\star$ is unknown.

2.2 $\mathsf{ScaledGD}(\lambda)$ for overparameterized low-rank matrix sensing

Inspired by the factorized representation (2), we aim to recover the low-rank matrix $M_\star$ by solving the following optimization problem (Burer and Monteiro, 2005):

$$\min_{X\in\mathbb{R}^{n\times r}} \quad f(X) \coloneqq \frac{1}{4}\big\|\mathcal{A}(XX^\top) - y\big\|_2^2, \tag{5}$$

where r is a predetermined rank parameter, possibly different from $r_\star$. It is evident that for any rotation matrix $O \in \mathcal{O}_r$, it holds that $f(X) = f(XO)$, leading to an infinite number of global minima of the loss function f.

A prelude: exact parameterization. When r is set to be the true rank $r_\star$ of $M_\star$, Tong et al. (2021a) set forth a provable algorithmic approach called scaled gradient descent ($\mathsf{ScaledGD}$)—gradient descent with a specific form of preconditioning—that adopts the following update rule

$$\mathsf{ScaledGD}: \qquad X_{t+1} = X_t - \eta \underbrace{\mathcal{A}^*\mathcal{A}(X_tX_t^\top - M_\star)X_t}_{=:\nabla f(X_t)}(X_t^\top X_t)^{-1}. \tag{6}$$

Here, X_t is the t-th iterate, $\nabla f(X_t)$ is the gradient of f at $X = X_t$, and $\eta > 0$ is the learning rate. Moreover, $\mathcal{A}^* : \mathbb{R}^m \mapsto \mathrm{Sym}_2(\mathbb{R}^n)$ is the adjoint operator of $\mathcal{A}$, that is $\mathcal{A}^*(y) = \sum_{i=1}^m y_i A_i$ for $y \in \mathbb{R}^m$.

At the expense of light computational overhead, $\mathsf{ScaledGD}$ is remarkably robust to ill-conditioning compared with vanilla gradient descent (GD). It is established in Tong et al. (2021a) that $\mathsf{ScaledGD}$, when starting from spectral initialization, converges linearly at a constant rate—*independent* of the condition number κ of $X_\star$ (cf. (3)); in contrast, the iteration complexity of GD (Tu et al., 2016; Zheng and Lafferty, 2015) scales on the order of κ^2 from the same initialization, therefore GD becomes exceedingly slow when the problem instance is even moderately ill-conditioned, a scenario that is quite commonly encountered in practice.

$\mathsf{ScaledGD}(\lambda)$: overparametrization under unknown rank. In this paper, we are interested in the so-called overparameterization regime, where $r_\star \le r \le n$. From an operational perspective, the true rank $r_\star$ is related to model order, e.g., the number of sources or targets in a scene of interest, which is often unavailable and makes it necessary to consider the misspecified setting. Unfortunately, in the presence of overparameterization, the original $\mathsf{ScaledGD}$ algorithm is no longer appropriate, as the preconditioner $(X_t^\top X_t)^{-1}$ might become numerically unstable to calculate. Therefore, we propose a new variant of $\mathsf{ScaledGD}$ by adjusting the preconditioner as

$$\mathsf{ScaledGD}(\lambda): \qquad X_{t+1} = X_t - \eta \underbrace{\mathcal{A}^*\mathcal{A}(X_tX_t^\top - M_\star)X_t}_{=:\nabla f(X_t)}(X_t^\top X_t + \lambda I)^{-1}, \tag{7}$$

where $\lambda > 0$ is a *fixed* damping parameter. The new algorithm is dubbed as $\mathsf{ScaledGD}(\lambda)$, and it recovers the original $\mathsf{ScaledGD}$ when $\lambda = 0$. Similar to $\mathsf{ScaledGD}$, a key property of $\mathsf{ScaledGD}(\lambda)$ is that the iterates $\{X_t\}$ are equivariant with respect to the parameterization of the factor matrix. Specifically, taking a rotationally equivalent factor X_tO with an arbitrary $O \in \mathcal{O}_r$, and feeding it into the update rule (7), the next iterate

$$X_tO - \eta\mathcal{A}^*\mathcal{A}(X_tX_t^\top - M_\star)X_tO(O^\top X_t^\top X_tO + \lambda I)^{-1} = X_{t+1}O$$

is rotated simultaneously by the same rotation matrix O. In other words, the recovered matrix sequence $M_t = X_tX_t^\top$ is invariant with respect to the parameterization of the factor matrix.

***Remark* 1.** We note that a related variant of $\mathsf{ScaledGD}$, called PrecGD, has been proposed recently in Zhang et al. (2022, 2021) for the overparameterized setting, which follows the update rule

$$\mathsf{PrecGD}: \qquad X_{t+1} = X_t - \eta\mathcal{A}^*\mathcal{A}(X_tX_t^\top - M_\star)X_t(X_t^\top X_t + \lambda_t I)^{-1}, \tag{8}$$

where the damping parameters $\lambda_t = \sqrt{f(X_t)}$ are selected in an *iteration-varying* manner assuming the algorithm is initialized properly. In contrast, $\mathsf{ScaledGD}(\lambda)$ assumes a fixed damping parameter λ throughout the iterations. We shall provide more detailed comparisons with PrecGD in Section 3.

3 Main results

Before formally presenting our theorems, let us introduce several key assumptions that will be in effect throughout this paper.

Restricted Isometry Property. A key property of the operator $\mathcal{A}(\cdot)$ is the celebrated Restricted Isometry Property (RIP) (Recht et al., 2010), which says that the operator $\mathcal{A}(\cdot)$ approximately preserves the distances between low-rank matrices. The formal definition is given as follows.

Definition 1 (Restricted Isometry Property)**.** The linear map $\mathcal{A}(\cdot)$ is said to obey rank-r RIP with a constant $\delta_r \in [0, 1)$, if for all matrices $M \in \mathrm{Sym}_2(\mathbb{R}^n)$ of rank at most r, it holds that

$$(1 - \delta_r)\|M\|_{\mathsf{F}}^2 \le \big\|\mathcal{A}(M)\big\|_2^2 \le (1 + \delta_r)\|M\|_{\mathsf{F}}^2. \tag{9}$$

The Restricted Isometry Constant (RIC) is defined to be the smallest positive δ_r such that (9) holds.

The RIP is a standard assumption in low-rank matrix sensing, which has been verified to hold with high probability for a wide variety of measurement operators. For example, if the entries of $\{A_i\}_{i=1}^m$ are independent up to symmetry with diagonal elements sampled from $\mathcal{N}(0, 1/m)$ and off-diagonal elements from $\mathcal{N}(0, 1/2m)$, then with high probability, $\mathcal{A}(\cdot)$ satisfies rank-r RIP with constant δ_r, as long as $m \ge Cnr/\delta_r^2$ for some sufficiently large universal constant $C > 0$ (Candès and Plan, 2011).

Throughout this paper, we make the following assumption about the operator $\mathcal{A}(\cdot)$.

Assumption 1. *The operator $\mathcal{A}(\cdot)$ satisfies the rank-$(r_\star + 1)$ RIP with $\delta_{r_\star+1} =: \delta$. Furthermore, there exist a sufficiently small constant $c_\delta > 0$ and a sufficiently large constant $C_\delta > 0$ such that*

$$\delta \le c_\delta r_\star^{-1/2} \kappa^{-C_\delta}. \tag{10}$$

Small random initialization. Similar to Li et al. (2018); Stöger and Soltanolkotabi (2021), we set the initialization X_0 to be a small random matrix, i.e.,

$$X_0 = \alpha G, \tag{11}$$

where $G \in \mathbb{R}^{n\times r}$ is some matrix considered to be normalized and $\alpha > 0$ controls the magnitude of the initialization. To simplify exposition, we take G to be a standard random Gaussian matrix, that is, G is a random matrix with i.i.d. entries distributed as $\mathcal{N}(0, 1/n)$.

Choice of parameters. Last but not least, the parameters of $\mathsf{ScaledGD}(\lambda)$ are selected according to the following assumption.

Assumption 2. *There exist some universal constants $c_\eta, c_\lambda, C_\alpha > 0$ such that (η, λ, α) in $\mathsf{ScaledGD}(\lambda)$ satisfy the following conditions:*

$$\text{(learning rate)} \qquad \eta \le c_\eta, \tag{12a}$$

$$\text{(damping parameter)} \qquad \frac{1}{100} c_\lambda \sigma_{\min}^2(X_\star) \le \lambda \le c_\lambda \sigma_{\min}^2(X_\star), \tag{12b}$$

$$\text{(initialization size)} \qquad \log \frac{\|X_\star\|}{\alpha} \ge \frac{C_\alpha}{\eta} \log(2\kappa) \cdot \log(2\kappa n). \tag{12c}$$

We are now in place to present the main theorems.

3.1 The overparameterization setting

We begin with our main theorem, which characterizes the performance of $\mathsf{ScaledGD}(\lambda)$ with overparameterization.

Theorem 2. *Suppose Assumptions 1 and 2 hold. With high probability (with respect to the realization of the random initialization G), there exists a universal constant $C_{\min} > 0$ such that for some $T \le T_{\min} \coloneqq \frac{C_{\min}}{\eta} \log \frac{\|X_\star\|}{\alpha}$, we have*

$$\|X_T X_T^\top - M_\star\|_\mathsf{F} \le \alpha^{1/3} \|X_\star\|^{5/3}.$$

In particular, for any prescribed accuracy target $\varepsilon \in (0,1)$, by choosing a sufficiently small α fulfilling both (12c) and $\alpha \le \varepsilon^3 \|X_\star\|$, we have $\|X_T X_T^\top - M_\star\|_\mathsf{F} \le \varepsilon \|M_\star\|$.

A few remarks are in order.

Iteration complexity. Theorem 2 shows that by choosing an appropriate α, $\mathsf{ScaledGD}(\lambda)$ finds an ε-accurate solution, i.e., $\|X_t X_t^\top - M_\star\|_\mathsf{F} \le \varepsilon \|M_\star\|$, in no more than an order of

$$\log \kappa \cdot \log(\kappa n) + \log(1/\varepsilon)$$

iterations. Roughly speaking, this asserts that $\mathsf{ScaledGD}(\lambda)$ converges at a constant linear rate after an initial phase of approximately $O(\log \kappa \cdot \log(\kappa n))$ iterations. Most notably, the iteration complexity is nearly independent of the condition number κ, with a small overhead only through the poly-logarithmic additive term $O(\log \kappa \cdot \log(\kappa n))$. In contrast, GD requires $O(\kappa^8 + \kappa^6 \log(\kappa n/\varepsilon))$ iterations to converge from a small random initialization to ε-accuracy; see Li et al. (2018); Stöger and Soltanolkotabi (2021). Thus, the convergence of GD is much slower than $\mathsf{ScaledGD}(\lambda)$ even for mildly ill-conditioned matrices.

Sample complexity. The sample complexity of $\mathsf{ScaledGD}(\lambda)$ hinges upon the Assumption 1. When the entries of $\{A_i\}_{i=1}^m$ are independent up to symmetry with diagonal elements sampled from $\mathcal{N}(0, 1/m)$ and off-diagonal elements from $\mathcal{N}(0, 1/2m)$, this assumption is fulfilled as long as $m \gtrsim n r_\star^2 \cdot \mathsf{poly}(\kappa)$. Our sample complexity depends only on the true rank $r_\star$, but not on the overparameterized rank r — a crucial feature in order to provide meaningful guarantees when the overparameterized rank r is close to the full dimension n. The dependency on κ in the sample complexity, on the other end, has been generally unavoidable in nonconvex low-rank estimation (Chi et al., 2019).

Comparison with Zhang et al. (2022, 2021). As mentioned earlier, our proposed algorithm $\mathsf{ScaledGD}(\lambda)$ is quite similar to PrecGD proposed in Zhang et al. (2021) that adopts an iteration-varying damping parameter. In terms of theoretical guarantees, Zhang et al. (2021) only provides the local convergence for PrecGD assuming an initialization close to the ground truth; in contrast, we provide global convergence guarantees where a small random initialization is used. More critically, the sample complexity of PrecGD (Zhang et al., 2021) depends on the overparameterized rank r, while ours only depends on the true rank $r_\star$. While Zhang et al. (2022) also studied variants of PrecGD with global convergence guarantees, they require additional operations such as gradient perturbations and switching between different algorithmic stages, which are harder to implement in practice. Our theory suggests that additional perturbation is unnecessary to ensure the global convergence of $\mathsf{ScaledGD}(\lambda)$, as $\mathsf{ScaledGD}(\lambda)$ automatically adapts to different curvatures of the optimization landscape throughout the entire trajectory.

3.2 The exact parameterization setting

We now single out the exact parametrization case, i.e., when $r = r_\star$. In this case, our theory suggests that $\mathsf{ScaledGD}(\lambda)$ converges to the ground truth even from a random initialization with a fixed scale $\alpha > 0$.

Theorem 3. *Assume that $r = r_\star$. Suppose Assumptions 1 and 2 hold. With high probability (with respect to the realization of the random initialization G), there exist some universal constants $C_{\min} > 0$ and $c > 0$ such that for some $T \le T_{\min} = \frac{C_{\min}}{\eta} \log(\|X_\star\|/\alpha)$, we have for any $t \ge T$*

$$\|X_t X_t^\top - M_\star\|_\mathsf{F} \le (1 - c\eta)^{t-T} \|M_\star\|.$$

Theorem 3 shows that with some fixed initialization scale α, $\mathsf{ScaledGD}(\lambda)$ takes at most an order of

$$\log \kappa \cdot \log(\kappa n) + \log(1/\varepsilon)$$

iterations to converge to ε-accuracy for any $\varepsilon > 0$ in the exact parameterization case. Compared with $\mathsf{ScaledGD}$ (Tong et al., 2021a) which takes $O(\log(1/\varepsilon))$ iterations to converge from a spectral initialization, we only pay a logarithmic order $O(\log\kappa\cdot\log(\kappa n))$ of additional iterations to converge from a random initialization. In addition, once the algorithms enter the local regime, both $\mathsf{ScaledGD}(\lambda)$ and $\mathsf{ScaledGD}$ behave similarly and converge at a fast constant linear rate, suggesting the effect of damping is locally negligible. Furthermore, compared with GD (Stöger and Soltanolkotabi, 2021) which requires $O(\kappa^8\log(\kappa n)+\kappa^2\log(1/\varepsilon))$ iterations to achieve ε-accuracy, our theory again highlights the benefit of $\mathsf{ScaledGD}(\lambda)$ in boosting the global convergence even for mildly ill-conditioned matrices.

4 Analysis

In this section, we present the main steps for proving Theorem 2 and Theorem 3. The detailed proofs are collected in the Appendix. All of our statements will be conditioned on the following high probability event regarding the initialization matrix G:

$$\mathcal{E} = \{\|G\| \le C_G\} \cap \{\sigma_{\min}(\widehat{U}^\top G) \ge (2n)^{-C_G}\}, \tag{13}$$

where $\widehat{U} \in \mathbb{R}^{n\times r_\star}$ is an orthonormal basis of the eigenspace associated with the $r_\star$ largest eigenvalues of $\mathcal{A}^*\mathcal{A}(M_\star)$, and $C_G > 0$ is some sufficiently large universal constant. It is a standard result in random matrix theory that $\mathcal{E}$ happens with high probability, as verified by the following lemma.

Lemma 1. *With respect to the randomness in G, the event $\mathcal{E}$ happens with probability at least $1-(cn)^{-C_G(r-r_\star+1)/2}-2\exp(-cn)$, where $c>0$ is some universal constant.*

Proof. See Appendix A.1. ☐

4.1 Preliminaries: decomposition of X_t

Before embarking on the main proof, we present a useful decomposition (cf. (14)) of the iterate X_t into a signal term, a misalignment error term, and an overparametrization error term. Choose some matrix $U_{\star,\perp} \in \mathbb{R}^{n\times(n-r_\star)}$ such that $[U_\star, U_{\star,\perp}]$ is orthonormal. Then we can define

$$S_t := U_\star^\top X_t \in \mathbb{R}^{r_\star\times r}, \quad \text{and} \quad N_t := U_{\star,\perp}^\top X_t \in \mathbb{R}^{(n-r_\star)\times r}.$$

Let the SVD of S_t be

$$S_t = U_t\Sigma_t V_t^\top,$$

where $U_t \in \mathbb{R}^{r_\star\times r_\star}$, $\Sigma_t \in \mathbb{R}^{r_\star\times r_\star}$, and $V_t \in \mathbb{R}^{r\times r_\star}$. Similar to $U_{\star,\perp}$, we define the orthogonal complement of V_t as $V_{t,\perp} \in \mathbb{R}^{r\times(r-r_\star)}$. When $r = r_\star$ we simply set $V_{t,\perp} = 0$.

We are now ready to present the main decomposition of X_t, which we use repeatedly in later analysis.

Proposition 1. *The following decomposition holds:*

$$X_t = \underbrace{U_\star\widetilde{S}_t V_t^\top}_{\text{signal}} + \underbrace{U_{\star,\perp}\widetilde{N}_t V_t^\top}_{\text{misalignment}} + \underbrace{U_{\star,\perp}\widetilde{O}_t V_{t,\perp}^\top}_{\text{overparametrization}}, \tag{14}$$

where

$$\widetilde{S}_t := S_t V_t \in \mathbb{R}^{r_\star\times r_\star}, \quad \widetilde{N}_t := N_t V_t \in \mathbb{R}^{(n-r_\star)\times r_\star}, \quad \textit{and} \quad \widetilde{O}_t := N_t V_{t,\perp} \in \mathbb{R}^{(n-r_\star)\times(r-r_\star)}. \tag{15}$$

Proof. See Appendix A.2. ☐

Several remarks on the decomposition are in order.

- First, since $V_{t,\perp}$ spans the obsolete subspace arising from overparameterization, $\widetilde{O}_t$ naturally represents the error incurred by overparameterization; in particular, in the well-specified case (i.e., $r = r_\star$), one has zero overparameterization error, i.e., $\widetilde{O}_t = 0$.

- Second, apart from the rotation matrix V_t, $\widetilde{S}_t$ documents the projection of the iterates X_t onto the signal space $U_\star$. Similarly, $\widetilde{N}_t$ characterizes the misalignment of the iterates with the signal subspace $U_\star$. It is easy to observe that in order for $X_t X_t^\top \approx M_\star$, one must have $\widetilde{S}_t \widetilde{S}_t^\top \approx \Sigma_\star^2$, and $\widetilde{N}_t \approx 0$.
- Last but not least, the extra rotation induced by V_t is extremely useful in making the signal/misalignment terms rationally invariant. To see this, suppose that we rotate the current iterate by $X_t \mapsto X_t Q$ with some rotational matrix $Q \in \mathcal{O}_r$, then $S_t \mapsto S_t Q$ but $\widetilde{S}_t$ remains unchanged, and similarly for $\widetilde{N}_t$.

4.2 Proof roadmap

Our analysis breaks into a few phases that characterize the dynamics of the key terms in the above decomposition, which we provide a roadmap to facilitate understanding. Denote

$$C_{\max} := \begin{cases} 4C_{\min}, & r > r_\star, \\ \infty, & r = r_\star, \end{cases} \qquad \text{and} \qquad T_{\max} := \frac{C_{\max}}{\eta} \log(\|X_\star\|/\alpha),$$

where $T_{\max}$ represents the largest index of the iterates that we maintain error control. The analysis boils down to the following phases, indicated by time points t_1, t_2, t_3, t_4 that satisfy

$$t_1 \le T_{\min}/16, \quad t_1 \le t_2 \le t_1 + T_{\min}/16, \quad t_2 \le t_3 \le t_2 + T_{\min}/16, \quad t_3 \le t_4 \le t_3 + T_{\min}/16.$$

- *Phase I: approximate power iterations.* In the initial phase, $\mathsf{ScaledGD}(\lambda)$ behaves similarly to GD, which is shown in Stöger and Soltanolkotabi (2021) to approximate the power method in the first few iterations up to t_1. After this phase, namely for $t \in [t_1, T_{\max}]$, although the signal strength is still quite small, it begins to be aligned with the ground truth with the overparameterization error kept relatively small.
- *Phase II: exponential amplification of the signal.* In this phase, $\mathsf{ScaledGD}(\lambda)$ behaves somewhat as a mixture of GD and $\mathsf{ScaledGD}$ with a proper choice of the damping parameter $\lambda \asymp \sigma_{\min}^2(X_\star)$, which ensures the signal strength first grows exponentially fast to reach a constant level no later than t_2, and then reaches the desired level no later than t_3, i.e., $\widetilde{S}_t \widetilde{S}_t^\top \approx \Sigma_\star^2$.
- *Phase III: local linear convergence.* At the last phase, $\mathsf{ScaledGD}(\lambda)$ behaves similarly to $\mathsf{ScaledGD}$, which converges linearly at a rate independent of the condition number. Specifically, for $t \in [t_3, T_{\max}]$, the reconstruction error $\|X_t X_t^\top - M_\star\|_{\mathsf{F}}$ converges at a linear rate up to some small overparameterization error, until reaching the desired accuracy for any $t \in [t_4, T_{\max}]$.

4.3 Phase I: approximate power iterations

It has been observed in Stöger and Soltanolkotabi (2021) that when initialized at a small scaled random matrix, the first few iterations of GD mimic the power iterations on the matrix $\mathcal{A}^*\mathcal{A}(M_\star)$. When it comes to $\mathsf{ScaledGD}(\lambda)$, since the initialization size α is chosen to be much smaller than the damping parameter λ, the preconditioner $(X_t^\top X_t + \lambda I)^{-1}$ behaves like $(\lambda I)^{-1}$ in the beginning. This renders $\mathsf{ScaledGD}(\lambda)$ akin to gradient descent in the initial phase. As a result, we also expect the first few iterations of $\mathsf{ScaledGD}(\lambda)$ to be similar to the power iterations, i.e.,

$$X_t \approx \Big(I + \frac{\eta}{\lambda}\mathcal{A}^*\mathcal{A}(M_\star)\Big)^t X_0, \qquad \text{when } t \text{ is small.}$$

Such proximity between $\mathsf{ScaledGD}(\lambda)$ and power iterations can indeed be justified in the beginning period, which allows us to deduce the following nice properties *after* the initial iterates of $\mathsf{ScaledGD}(\lambda)$.

Lemma 2. *Under the same setting as Theorem 2, there exists an iteration number $t_1 : t_1 \le T_{\min}/16$ such that*

$$\sigma_{\min}(\widetilde{S}_{t_1}) \ge \alpha^2/\|X_\star\|, \tag{16}$$

and that, for any $t \in [t_1, T_{\max}]$, $\widetilde{S}_t$ is invertible and one has

$$\|\widetilde{O}_t\| \le (C_{2.b}\kappa n)^{-C_{2.b}}\|X_\star\|\sigma_{\min}\big((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t\big), \tag{17a}$$

$$\|\widetilde{O}_t\| \le \left(1+\frac{\eta}{12C_{\max}\kappa}\right)^{t-t_1}\alpha^{5/6}\|X_\star\|^{1/6}, \tag{17b}$$

$$\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \le c_2\kappa^{-C_\delta/2}\|X_\star\|, \tag{17c}$$

$$\|\widetilde{S}_t\| \le C_{2.a}\kappa\|X_\star\|, \tag{17d}$$

where $C_{2.a}$, $C_{2.b}$, c_2 are some positive constants satisfying $C_{2.a} \lesssim c_\lambda^{-1/2}$, $c_2 \lesssim c_\delta/c_\lambda^3$, and $C_{2.b}$ can be made arbitrarily large by increasing C_α.

Proof. See Appendix C. ☐

Remark 2. Let us record two immediate consequences of (17), which sometimes are more convenient for later analysis. From (17a), we may deduce

$$\|\widetilde{O}_t\| \le (C_{2.b}\kappa n)^{-C_{2.b}}\|X_\star\|\sigma_{\min}(\Sigma_\star^2+\lambda I)^{-1/2}\sigma_{\min}(\widetilde{S}_t) \le \kappa(C_{2.b}\kappa n)^{-C_{2.b}}\sigma_{\min}(\widetilde{S}_t) \le (C'_{2.b}\kappa n)^{-C'_{2.b}}\sigma_{\min}(\widetilde{S}_t), \tag{18}$$

where $C'_{2.b} = C_{2.b}/2$, provided $C_{2.b} > 4$. It is clear that $C'_{2.b}$ can also be made arbitrarily large by enlarging C_α. Similarly, from (17b), we may deduce

$$\begin{aligned}\|\widetilde{O}_t\| \le \left(1+\frac{\eta}{12C_{\max}\kappa}\right)^{t-t_1}\alpha^{5/6}\|X_\star\|^{1/6} &\le \left(1+\frac{\eta}{12C_{\max}\kappa}\right)^{\frac{C_{\max}}{\eta}\log(\|X_\star\|/\alpha)}\alpha^{5/6}\|X_\star\|^{1/6} \\ &\le (\|X_\star\|/\alpha)^{1/12}\alpha^{5/6}\|X_\star\|^{1/6} = \alpha^{3/4}\|X_\star\|^{1/4}. \end{aligned} \tag{19}$$

Lemma 2 ensures the iterates of $\mathsf{ScaledGD}(\lambda)$ maintain several desired properties after iteration t_1, as summarized in (17). In particular, for any $t \in [t_1, T_{\max}]$: (i) the overparameterization error $\|\widetilde{O}_t\|$ remains small relatively to the signal strength measured in terms of the scaled minimum singular value $\sigma_{\min}\big((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t\big)$, and remains bounded with respect to the size of the initialization α (cf. (17a) and (17b) and their consequences (18) and (19)); (ii) the scaled misalignment-to-signal ratio remains bounded, suggesting the iterates remain aligned with the ground truth signal subspace $U_\star$ (cf. (17c)); (iii) the size of the signal component $\widetilde{S}_t$ remains bounded (cf. (17d)). These properties play an important role in the follow-up analysis.

Remark 3. It is worth noting that, the scaled minimum singular value $\sigma_{\min}((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t)$ plays a key role in our analysis, which is in sharp contrast to the use of the vanilla minimum singular value $\sigma_{\min}(\widetilde{S}_t)$ in the analysis of gradient descent (Stöger and Soltanolkotabi, 2021). This new measure of signal strength is inspired by the scaled distance for $\mathsf{ScaledGD}$ introduced in Tong et al. (2021a, 2022), which carefully takes the preconditioner design into consideration. Similarly, the metrics $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|$ in (17c) and $\big\|\Sigma_\star^{-1}(\widetilde{S}_{t+1}\widetilde{S}_{t+1}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\big\|$ (to be seen momentarily) are also scaled for similar considerations to unveil the fast convergence (almost) independent of the condition number.

4.4 Phase II: exponential amplification of the signal

By the end of Phase I, the signal strength is still quite small (cf. (16)), which is far from the desired level. Fortunately, the properties established in Lemma 2 allow us to establish an exponential amplification of the signal term $\widetilde{S}_t$ thereafter, which can be further divided into two stages.

1. In the first stage, the signal is boosted to a constant level, i.e., $\widetilde{S}_t\widetilde{S}_t^\top \succeq \frac{1}{10}\Sigma_\star^2$;
2. In the second stage, the signal grows further to the desired level, i.e., $\widetilde{S}_t\widetilde{S}_t^\top \approx \Sigma_\star^2$.

We start with the first stage, which again uses $\sigma_{\min}\big((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t\big)$ as a measure of signal strength in the following lemma.

Lemma 3. *For any t such that (17) holds, we have*

$$\sigma_{\min}\big((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_{t+1}\big) \ge (1-2\eta)\sigma_{\min}\big((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t\big).$$

Moreover, if $\sigma_{\min}\big((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t\big) \le 1/3$*, then*

$$\sigma_{\min}\big((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_{t+1}\big) \ge \left(1+\frac{1}{8}\eta\right)\sigma_{\min}\big((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t\big).$$

Proof. See Appendix D.1. □

The second half of Lemma 3 uncovers the exponential growth of the signal strength $\sigma_{\min}\big((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t\big)$ until a constant level after several iterations, which resembles the exponential growth of the signal strength in GD (Stöger and Soltanolkotabi, 2021). This is formally established in the following corollary.

Corollary 1. *There exists an iteration number* $t_2 : t_1 \le t_2 \le t_1 + T_{\min}/16$ *such that for all* $t \in [t_2, T_{\max}]$*, we have*

$$\widetilde{S}_t\widetilde{S}_t^\top \succeq \frac{1}{10}\Sigma_\star^2. \tag{20}$$

Proof. See Appendix D.2. □

We next aim to show that $\widetilde{S}_t\widetilde{S}_t^\top \approx \Sigma_\star^2$ after the signal strength is above the constant level. To this end, the behavior of $\mathsf{ScaledGD}(\lambda)$ becomes closer to that of $\mathsf{ScaledGD}$, and it turns out to be easier to work with $\big\|\Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)\Sigma_\star^{-1}\big\|$ as a measure of the scaled recovery error of the signal component. We establish the approximate exponential shrinkage of this measure in the following lemma.

Lemma 4. *For all* $t \in [t_2, T_{\max}]$ *with* t_2 *given in Corollary 1, one has*

$$\big\|\Sigma_\star^{-1}(\widetilde{S}_{t+1}\widetilde{S}_{t+1}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\big\| \le (1-\eta)\big\|\Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)\Sigma_\star^{-1}\big\| + \frac{1}{100}\eta. \tag{21}$$

Proof. See Appendiex D.3. □

With the help of Lemma 4, it is straightforward to establish the desired approximate recovery guarantee of the signal component, i.e., $\widetilde{S}_t\widetilde{S}_t^\top \approx \Sigma_\star^2$.

Corollary 2. *There exists an iteration number* $t_3 : t_2 \le t_3 \le t_2 + T_{\min}/16$ *such that for any* $t \in [t_3, T_{\max}]$*, one has*

$$\frac{9}{10}\Sigma_\star^2 \preceq \widetilde{S}_t\widetilde{S}_t^\top \preceq \frac{11}{10}\Sigma_\star^2. \tag{22}$$

Proof. See Appendix D.4. □

4.5 Phase III: local convergence

Corollary 2 tells us that after iteration t_3, we enter a local region in which $\widetilde{S}_t\widetilde{S}_t^\top$ is close to the ground truth $\Sigma_\star^2$. In this local region, the behavior of $\mathsf{ScaledGD}(\lambda)$ becomes closer to that of $\mathsf{ScaledGD}$ analyzed in Tong et al. (2021a). We turn attention to the reconstruction error $\|X_tX_t^\top - M_\star\|_{\mathsf{F}}$ that measures the generalization performance, and show it converges at a linear rate independent of the condition number up to some small overparameterization error.

Lemma 5. *There exists some universal constant* $c_5 > 0$ *such that for any* $t : t_3 \le t \le T_{\max}$*, we have*

$$\|X_tX_t^\top - M_\star\|_{\mathsf{F}} \le (1-c_5\eta)^{t-t_3}\sqrt{r_\star}\|M_\star\| + 8c_5^{-1}\|M_\star\| \max_{t_3\le\tau\le t}\left(\frac{\|\widetilde{O}_\tau\|}{\|X_\star\|}\right)^{1/2}. \tag{23}$$

In particular, there exists an iteration number $t_4 : t_3 \le t_4 \le t_3 + T_{\min}/16$ *such that for any* $t \in [t_4, T_{\max}]$*, we have*

$$\|X_tX_t^\top - M_\star\|_{\mathsf{F}} \le \alpha^{1/3}\|X_\star\|^{5/3} \le \varepsilon\|M_\star\|. \tag{24}$$

Here, ε *and* α *are as stated in Theorem 2.*

Proof. See Appendix E. □

4.6 Proofs of main theorems

Now we are ready to collect the results in the preceding sections to prove our main results, i.e., Theorem 2 and Theorem 3.

We start with proving Theorem 2. By Lemma 2, Corollary 1, Corollary 2 and Lemma 5, the final t_4 given by Lemma 5 is no more than $4 \times T_{\min}/16 \le T_{\min}/2$, thus (24) holds for all $t \in [T_{\min}/2, T_{\max}]$, in particular, for some $T \le T_{\min}$, as claimed.

Now we consider Theorem 3. In case that $r = r_\star$, it follows from definition that $\widetilde{O}_t = 0$ vanishes for all t. It follows from Lemma 5, in particular from (23), that

$$\|X_t X_t^\top - M_\star\|_{\mathsf{F}} \le (1 - c_5\eta)^{t-t_3}\sqrt{r_\star}\|M_\star\|,$$

for any $t \ge t_3$ (recall that $T_{\max} = \infty$ by definition when $r = r_\star$). Note that $(1-c_5\eta)^t\sqrt{r_\star} \le (1-c_5\eta)^{t-T+t_3}$ if $T - t_3 \ge 4\log(r_\star)/(c_5\eta)$ given that $\eta \le c_\eta$ is sufficiently small. Thus for any $t \ge T$ we have

$$\|X_t X_t^\top - M_\star\|_{\mathsf{F}} \le (1 - c_5\eta)^{t-T}\|M_\star\|.$$

It is clear that one may choose such T which also satisfies $T \le t_3 + 8/(c_5\eta) \le t_3 + T_{\min}/16$. We have already shown in the proof of Theorem 2 that $t_3 \le 4 \times T_{\min}/16 \le T_{\min}/4$, thus $T \le T_{\min}$ as desired.

Remark 4. In the overparameterized setting, our theory guarantees the reconstruction error to be small until some iteration $T_{\max}$. This is consistent with the phenomenon known as *early stopping* in prior works of learning with overparameterized models (Li et al., 2018; Stöger and Soltanolkotabi, 2021).

5 Numerical experiments

In this section, we conduct numerical experiments to demonstrate the efficacy of ScaledGD(λ) for solving overparametrized low-rank matrix sensing. We set the ground truth matrix $X_\star = U_\star \Sigma_\star \in \mathbb{R}^{n\times r_\star}$ where $U_\star \in \mathbb{R}^{n\times r_\star}$ is a random orthogonal matrix and $\Sigma_\star \in \mathbb{R}^{r_\star \times r_\star}$ is a diagonal matrix whose condition number is set to be κ. We set $n = 150$ and $r_\star = 3$, and use random Gaussian measurements with $m = 10nr_\star$. The overparameterization rank r is set to be 5.

Comparison with overparametrized GD. We run ScaledGD(λ) and GD with random initialization and compare their convergence speeds under different condition numbers κ of the ground truth $X_\star$; the result is depicted in Figure 1. Even for a moderate range of κ, GD slows down significantly while the convergence speed of ScaledGD(λ) remains almost the same with a almost negligible initial phase, which is consistent with our theory. The advantage of ScaledGD(λ) enlarges as κ increase, and is already more than 10x times faster than GD when $\kappa = 7$.

Effect of initialization size. We study the effect of the initialization scale α on the reconstruction accuracy of ScaledGD(λ). We fix the learning rate η to be a constant and vary the initialization scale. We run ScaledGD(λ) until it converges.[1] The resulting reconstruction errors and their corresponding initialization scales are plotted in Figure 2. It can be inferred that the reconstruction error increases with respect to α, which is consistent with our theory.

6 Discussions

This paper demonstrates that an appropriately preconditioned gradient descent method, called ScaledGD(λ), guarantees an accelerated convergence to the ground truth low-rank matrix in overparameterized low-rank matrix sensing, when initialized from a sufficiently small random initialization. Furthermore, in the case of exact parameterization, our analysis guarantees the fast global convergence of ScaledGD(λ) from a small random initialization. Collectively, this complements and represents a major step forward from prior analyses of ScaledGD (Tong et al., 2021a) by allowing overparametrization and small random initialization. This works opens up a few exciting future directions that are worth further exploring.

[1] More precisely, in accordance with our theory which requires early stopping, we stop the algorithm once we detected that the training error no longer decreases significantly for a long time (e.g. 100 iterations).

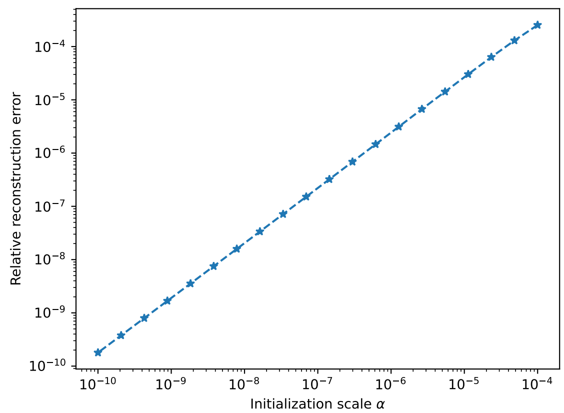

Figure 2: Relative reconstruction error versus initialization scale α. The slope of the dashed line is approximately 1.

- *Asymmetric case.* Our current analysis is confined to the recovery of low-rank positive semidefinite matrices, with only one factor matrix to be recovered. It remains to generalize this analysis to the recovery of general low-rank matrices with overparameterization.
- *Robust setting.* Many applications encounter corrupted measurements that call for robust recovery algorithms that optimize nonsmooth functions such as the least absolute deviation loss. One such example is the scaled subgradient method (Tong et al., 2021b), which is the nonsmooth counterpart of ScaledGD robust to ill-conditioning, and it'll be interesting to study its performance under overparameterization.
- *Noisy case.* For simplicity, our analysis focused on the noise-free case, while in practice, one often deals with noisy data. It is of great importance to understand the statistical error rates of ScaledGD(λ) under a variety of noise models.
- *Other overparameterized learning models.* Our work provides evidence on the power of preconditioning in accelerating the convergence without hurting generalization in overparameterized low-rank matrix sensing, which is one kind of overparameterized learning models. It will be greatly desirable to extend the insights developed herein to other overparameterized learning models, for example tensors (Dong et al., 2022; Tong et al., 2022) and neural networks (Wang et al., 2021).

Acknowledgements

The work of Y. Chi is supported in part by Office of Naval Research under N00014-19-1-2404, and by National Science Foundation under CAREER ECCS-1818571, CCF-1806154, CCF-1901199 and ECCS-2126634.

References

Amari, S.-i., Ba, J., Grosse, R. B., Li, X., Nitanda, A., Suzuki, T., Wu, D., and Xu, J. (2020). When does preconditioning help or hurt generalization? In *International Conference on Learning Representations*.

Bai, Y., Jiang, Q., and Sun, J. (2018). Subgradient descent learns orthogonal dictionaries. In *International Conference on Learning Representations*.

Bhatia, R. (1997). *Matrix analysis*, volume 169. Springer New York, NY.

Burer, S. and Monteiro, R. D. C. (2005). Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444.

Candès, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359.

Chandrasekher, K. A., Lou, M., and Pananjady, A. (2022). Alternating minimization for generalized rank one matrix sensing: Sharp predictions from a random initialization. *arXiv preprint arXiv:2207.09660*.

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

Chen, Y. and Chi, Y. (2018). Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14 – 31.

Chen, Y., Chi, Y., Fan, J., and Ma, C. (2019). Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2):5–37.

Chen, Y., Chi, Y., Fan, J., and Ma, C. (2021). Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806.

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

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

Ding, L., Jiang, L., Chen, Y., Qu, Q., and Zhu, Z. (2021). Rank overspecified robust matrix recovery: Subgradient method and exact recovery. *Advances in Neural Information Processing Systems*, 34:26767–26778.

Dong, H., Tong, T., Ma, C., and Chi, Y. (2022). Fast and provable tensor robust principal component analysis via scaled gradient descent. *arXiv preprint arXiv:2206.09109*.

Geyer, K., Kyrillidis, A., and Kalev, A. (2020). Low-rank regularization and solution uniqueness in over-parameterized matrix sensing. In *International Conference on Artificial Intelligence and Statistics*, pages 930–940. PMLR.

Gilboa, D., Buchanan, S., and Wright, J. (2019). Efficient dictionary learning with gradient descent. In *International Conference on Machine Learning*, pages 2252–2259. PMLR.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2017). Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159.

Jiang, L., Chen, Y., and Ding, L. (2022). Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *arXiv preprint arXiv:2203.02839*.

Kim, D. and Chung, H. W. (2022). Rank-1 matrix completion with gradient descent and small random initialization. *arXiv preprint arXiv:2212.09396*.

Lee, K. and Stöger, D. (2022). Randomly initialized alternating least squares: Fast convergence for matrix sensing. *arXiv preprint arXiv:2204.11516*.

Li, Y., Ma, C., Chen, Y., and Chi, Y. (2021). Nonconvex matrix factorization from rank-one measurements. *IEEE Transactions on Information Theory*, 67(3):1928–1950.

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

Ma, C., Li, Y., and Chi, Y. (2021). Beyond Procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. *IEEE Transactions on Signal Processing*, 69:867–877.

Ma, C., Wang, K., Chi, Y., and Chen, Y. (2019). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, pages 1–182.

Ma, J. and Fattahi, S. (2022). Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *arXiv preprint arXiv:2202.08788.*

Oymak, S. and Soltanolkotabi, M. (2019). Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning*, pages 4951–4960. PMLR.

Qu, Q., Li, X., and Zhu, Z. (2019). A nonconvex approach for exact and efficient multichannel sparse blind deconvolution. *Advances in Neural Information Processing Systems*, 32.

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

Rudelson, M. and Vershynin, R. (2009). Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739.

Shi, L. and Chi, Y. (2021). Manifold gradient descent solves multi-channel sparse blind deconvolution provably and efficiently. *IEEE Transactions on Information Theory*, 67(7):4784–4811.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. (2018). Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769.

Stöger, D. and Soltanolkotabi, M. (2021). Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843.

Tong, T., Ma, C., and Chi, Y. (2021a). Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63.

Tong, T., Ma, C., and Chi, Y. (2021b). Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. *IEEE Transactions on Signal Processing*, 69:2396–2409.

Tong, T., Ma, C., Prater-Bennette, A., Tripp, E., and Chi, Y. (2022). Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements. *Journal of Machine Learning Research*, 23(163):1–77.

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

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In Eldar, Y. C. and Kutyniok, G., editors, *Compressed Sensing*, pages 210–268. Cambridge University Press.

Wang, T., Buchanan, S., Gilboa, D., and Wright, J. (2021). Deep networks provably classify data on curves. *Advances in Neural Information Processing Systems*, 34:28940–28953.

Ye, T. and Du, S. S. (2021). Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34:1429–1439.

Zhang, G., Fattahi, S., and Zhang, R. Y. (2022). Preconditioned gradient descent for overparameterized nonconvex burer–monteiro factorization with global optimality certification. *arXiv preprint arXiv:2206.03345.*

Zhang, J., Fattahi, S., and Zhang, R. Y. (2021). Preconditioned gradient descent for over-parameterized nonconvex matrix factorization. *Advances in Neural Information Processing Systems*, 34:5985–5996.

Zhang, R. Y. (2021). Sharp global guarantees for nonconvex low-rank matrix recovery in the overparameterized regime. *arXiv preprint arXiv:2104.10790.*

Zhang, R. Y. (2022). Improved global guarantees for the nonconvex Burer–Monteiro factorization via rank overparameterization. *arXiv preprint arXiv:2207.01789*.

Zheng, Q. and Lafferty, J. (2015). A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117.

Zhuo, J., Kwon, J., Ho, N., and Caramanis, C. (2021). On the computational and statistical complexity of over-parameterized matrix sensing. *arXiv preprint arXiv:2102.02756*.

A Preliminaries

This section collects several preliminary results that are useful in later proofs. In general, for a matrix A, we will denote by U_A the first factor in its compact SVD $A = U_A \Sigma_A V_A^\top$, unless otherwise specified.

A.1 Proof of Lemma 1

It is a standard result in random matrix theory (Rudelson and Vershynin, 2009; Vershynin, 2012) that an $M \times N$ ($M \ge N$) random matrix G_0 with i.i.d. standard Gaussian entries satisfies

$$\mathbb{P}\left(\|G_0\| \le 4\big(\sqrt{M} + \sqrt{N}\big)\right) \ge 1 - \exp(-M/C), \tag{25a}$$

$$\mathbb{P}\left(\sigma_{\min}(G_0) \ge \varepsilon\big(\sqrt{M} - \sqrt{N-1}\big)\right) \ge 1 - (C\varepsilon)^{M-N+1} - \exp(-M/C), \tag{25b}$$

for some universal constant $C > 0$ and for any $\varepsilon > 0$. Applying (25a) to the random matrix $\sqrt{n}G$ which is an $n \times r$ random matrix with i.i.d. standard Gaussian entries, we have

$$\|G\| \le 4(\sqrt{n} + \sqrt{r})/\sqrt{n} \le 8$$

with probability at least $1 - \exp(-n/C)$.

Turning to the bound on $\sigma_{\min}^{-1}(\widehat{U}^\top G)$, observe that $\sqrt{n}\widehat{U}^\top G$ is a $r_\star \times r$ random matrix with i.i.d. standard Gaussian entries, thus applying (25b) to $\sqrt{n}\widehat{U}^\top G$ with $\varepsilon = (2n)^{-C_G+1}$ yields

$$\sigma_{\min}^{-1}(\widehat{U}^\top G) \le (2n)^{C_G-1}(\sqrt{r} - \sqrt{r_\star - 1})^{-1} \le (2n)^{C_G-1}(2\sqrt{r}) \le (2n)^{C_G}$$

with probability at least $1 - (2n/C)^{-(C_G-1)(r-r_\star+1)} - \exp(-n/C)$. Here, the second inequality follows from

$$\frac{1}{\sqrt{r} - \sqrt{r_\star - 1}} \le \frac{1}{\sqrt{r} - \sqrt{r-1}} = \sqrt{r} + \sqrt{r-1} < 2\sqrt{r}.$$

Combining the above two bounds directly implies the desired probability bound if we choose $c = 1/C$ and choose a large C_G such that $C_G \ge 8$ and $C_G - 1 \ge C_G/2$.

A.2 Proof of Proposition 1

Using the definitions of S_t and N_t, we have

$$\begin{aligned} X_t = (U_\star U_\star^\top + U_{\star,\perp}U_{\star,\perp}^\top)X_t &= U_\star S_t + U_{\star,\perp}N_t \\ &= U_\star \widetilde{S}_t V_t^\top + U_{\star,\perp}N_t(V_t V_t^\top + V_{t,\perp}V_{t,\perp}^\top) \\ &= U_\star \widetilde{S}_t V_t^\top + U_{\star,\perp}\widetilde{N}_t V_t^\top + U_{\star,\perp}\widetilde{O}_t V_{t,\perp}^\top, \end{aligned}$$

where in the second line, we used the relation $\widetilde{S}_t = S_t V_t = U_t \Sigma_t V_t^\top V_t = U_t \Sigma_t$ and thus

$$S_t = \widetilde{S}_t V_t^\top. \tag{26}$$

A.3 Consequences of RIP

The first result is a standard consequence of RIP, see, for example Stöger and Soltanolkotabi (2021, Lemma 7.3).

Lemma 6. *Suppose that the linear map $\mathcal{A} : \mathrm{Sym}_2(\mathbb{R}^n) \to \mathbb{R}^m$ satisfies Assumption 1. Then we have*

$$\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(Z)\| \le \delta \|Z\|_{\mathsf{F}}$$

for any $Z \in \mathrm{Sym}_2(\mathbb{R}^n)$ with rank at most $r_\star$.

We need another straightforward consequence of RIP, given by the following lemma.

Lemma 7. *Under the same setting as in Lemma 6, we have*

$$\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(Z)\| \le 2\delta\sqrt{(r \vee r_\star)/r_\star}\|Z\|_{\mathsf{F}} \le \frac{2(r \vee r_\star)\delta}{\sqrt{r_\star}}\|Z\|$$

for any $Z \in \mathrm{Sym}_2(\mathbb{R}^n)$ with rank at most r.

Proof. Without loss of generality we may assume $r \ge r_\star$, thus $r \vee r_\star = r$. We claim that it is possible to decompose $Z = \sum_{i \le \lceil r/r_\star \rceil} Z_i$ where $Z_i \in \mathrm{Sym}_2(\mathbb{R}^n)$, $\mathrm{rank}(Z_i) \le r_\star$ and $Z_i Z_j = 0$ if $i \ne j$. To see why this is the case, notice the spectral decomposition of Z gives r rank-one components that are mutually orthogonal, thus we may divide them into $\lceil r/r_\star \rceil$ subgroups indexed by $i = 1, \ldots, \lceil r/r_\star \rceil$, such that each subgroup contains at most $r_\star$ components. Let Z_i be the sum of the components in the subgroup i, it is easy to check that Z_i has the desired property.

The property of the decomposition yields

$$\|Z\|_{\mathsf{F}}^2 = \mathrm{tr}(Z^2) = \sum_{i,j \le \lceil r/r_\star \rceil} \mathrm{tr}(Z_i Z_j) = \sum_{i \le \lceil r/r_\star \rceil} \|Z_i\|_{\mathsf{F}}^2. \tag{27}$$

But for each Z_i, Lemma 6 implies

$$\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(Z_i)\| \le \delta \|Z_i\|_{\mathsf{F}}.$$

Summing up for $i \le \lceil r/r_\star \rceil$ yields

$$\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(Z)\| \le \sum_{i \le \lceil r/r_\star \rceil} \|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(Z_i)\| \le \delta \sum_{i \le \lceil r/r_\star \rceil} \|Z_i\|_{\mathsf{F}} \le \delta\sqrt{\lceil r/r_\star \rceil}\,\|Z\|_{\mathsf{F}},$$

where the last inequality follows from (27) and from Cauchy-Schwarz inequality.

The first inequality in Lemma 7 follows from the above inequality by noting that $\lceil r/r_\star \rceil \le 2r/r_\star$ given $r \ge r_\star$ which was assumed in the beginning of the proof. The second inequality in Lemma 7 follows from $\|Z\|_{\mathsf{F}} \le \sqrt{r}\|Z\|$. □

A.4 Matrix perturbation results

The next few results are all on matrix perturbations. We first present a perturbation result on matrix inverse.

Lemma 8. *Assume that A, B are square matrices of the same dimension, and that A is invertible. If $\|A^{-1}B\| \le 1/2$, then*

$$(A + B)^{-1} = A^{-1} + A^{-1}BQA^{-1}, \quad \text{for some } \|Q\| \le 2.$$

Similarly, if $\|BA^{-1}\| \le 1/2$, then we have

$$(A + B)^{-1} = A^{-1} + A^{-1}QBA^{-1}, \quad \text{for some } \|Q\| \le 2.$$

In particular, if $\|B\| \le \sigma_{\min}(A)/2$, then both of the above equations hold.

Proof. The claims follow from the identity

$$(A+B)^{-1} = A^{-1} - A^{-1}B(I + A^{-1}B)^{-1}A^{-1} = A^{-1} - A^{-1}(I + BA^{-1})^{-1}BA^{-1}.$$

For the first claim when $\|A^{-1}B\| \le 1/2$, we set $Q := -(I+A^{-1}B)^{-1}$, which satisfies $\|Q\| = \|(I+A^{-1}B)^{-1}\| \le \frac{1}{1-\|A^{-1}B\|} \le 2$. The second claim follows similarly. Finally, we note that when $\|B\| \le \sigma_{\min}(A)/2$, it holds

$$\|A^{-1}B\| \le \frac{1}{\sigma_{\min}(A)}\|B\| \le \frac{1}{2} \qquad \text{and} \qquad \|BA^{-1}\| \le \|B\|\frac{1}{\sigma_{\min}(A)} \le \frac{1}{2},$$

thus completing the proof. $\square$

Next, we focus on the minimum singular value of certain matrix of form $I + AB$.

Lemma 9. *If A, B are positive definite matrices of the same size, we have*

$$\sigma_{\min}(I + AB) \ge \kappa^{-1/2}(A), \quad \textit{where } \kappa(A) := \frac{\|A\|}{\sigma_{\min}(A)}.$$

Proof. Writing $I + AB = A^{1/2}(I + A^{1/2}BA^{1/2})A^{-1/2}$, we obtain

$$\sigma_{\min}(I + AB) \ge \sigma_{\min}(A^{1/2})\sigma_{\min}(A^{-1/2})\sigma_{\min}(I + A^{1/2}BA^{1/2}).$$

The proof is completed by noting that $\sigma_{\min}(A^{1/2}) = \sigma_{\min}^{1/2}(A)$, $\sigma_{\min}(A^{-1/2}) = \|A\|^{-1/2}$, and that $\sigma_{\min}(I + A^{1/2}BA^{1/2}) \ge 1$ since $A^{1/2}BA^{1/2}$ is positive semidefinite. $\square$

The last result still concerns the minimum singular value of a matrix of interest.

Lemma 10. *There exists a universal constant $c_{10} > 0$ such that if Λ is a positive definite matrix obeying $\|\Lambda\| \le c_{10}$ and $\sigma_{\min}(Y) \le 1/3$, then for any $\eta \le c_{10}$ we have*

$$\sigma_{\min}\Big(\left((1-\eta)I + \eta(YY^\top + \Lambda)^{-1}\right)Y\Big) \ge \Big(1 + \frac{\eta}{6}\Big)\,\sigma_{\min}(Y). \tag{28}$$

Proof. Denote $Z = YY^\top$ and let $U\Sigma U^\top = Z + \Lambda$ be the spectral decomposition of $Z + \Lambda$. By a coordinate transform one may assume $Z + \Lambda = \Sigma$. It suffices to show

$$\lambda_{\min}\Big(\left((1-\eta)I + \eta\Sigma^{-1}\right)Z\left((1-\eta)I + \eta\Sigma^{-1}\right)\Big) \ge \Big(1 + \frac{1}{6}\eta\Big)^2\lambda_{\min}(Z). \tag{29}$$

For simplicity we denote $\zeta = \lambda_{\min}(Z)$, which is by assumption smaller than $1/9$. Fix $K = 1/4$ so that $K \ge 2\zeta + 4c_{10}$ by choosing c_{10} to be small enough. By permuting coordinates we may further assume that the diagonal matrix Σ is of the following form:

$$\Sigma = \begin{bmatrix} \Sigma_{\le K} & \\ & \Sigma_{>K} \end{bmatrix}, \tag{30}$$

where $\Sigma_{\le K}$, $\Sigma_{>K}$ are diagonal matrices such that $\lambda_{\max}(\Sigma_{\le K}) \le K$ and $\lambda_{\min}(\Sigma_{>K}) > K$. It suffices to consider the case where $\Sigma_{>K}$ is not degenerate, because otherwise $\lambda_{\max}(\Sigma) \le K \le 1/2$, and the desired (29) follows as

$$\lambda_{\min}\Big(\left((1-\eta)I + \eta\Sigma^{-1}\right)Z\left((1-\eta)I + \eta\Sigma^{-1}\right)\Big) \ge \left(1 - \eta + \eta\lambda_{\max}^{-1}(\Sigma)\right)^2\lambda_{\min}(Z) \ge (1+\eta)^2\lambda_{\min}(Z).$$

For the rest of the proof, we assume the block corresponding to $\Sigma_{>K}$ is not degenerate.

Divide Z into blocks of the same shape as (30):

$$Z = \begin{bmatrix} Z_0 & A \\ A^\top & Z_1 \end{bmatrix}. \tag{31}$$

The purpose of such division is to facilitate computation of minimum eigenvalues by Schur's complement lemma. For preparation, we make a few simple observations. Since $Z = \Sigma - \Lambda$, we see that A being an off-diagonal submatrix of Z satisfies $\|A\| \le \|\Lambda\| \le c_{10}$, and similarly $\|Z_0 - \Sigma_{\le K}\| \le c_{10}$, $\|Z_1 - \Sigma_{>K}\| \le c_{10}$. In particular, we have

$$\lambda_{\min}(Z_1) \ge \lambda_{\min}(\Sigma_{>K}) - c_{10} > K - c_{10} \ge 2\zeta + 3c_{10} > \zeta, \tag{32}$$

which implies $Z_1 - \zeta I$ is positive definite and invertible. Thus by Schur's complement lemma, $Z \succeq \zeta I$ is equivalent to

$$Z_0 - \zeta I - A(Z_1 - \zeta I)^{-1} A^\top \succeq 0, \tag{33}$$

which provides an analytic characterization for the minimum eigenvalue ζ of Z.

The rest of the proof follows from the following steps: we will first show again by Schur's complement lemma that (29) admits a similar analytic characterization. More precisely, denoting $\zeta' = (1 + \frac{\eta}{6})^2 \zeta$, $\Sigma_0 = (1-\eta)I + \eta \Sigma_{\le K}^{-1}$ and $\Sigma_1 = (1-\eta)I + \eta\Sigma_{>K}^{-1}$, then (29) is equivalent to

$$Z_0 - \zeta' \Sigma_0^{-2} - A(Z_1 - \zeta' \Sigma_1^{-2})^{-1} A^\top \succeq 0. \tag{34}$$

After proving they are equivalent, we will prove that (34) holds as long as the following sufficient condition holds

$$Z_0 - (1+3\eta)^{-2} \zeta' I - A(Z_1 - \zeta I)^{-1} A^\top - 10\eta\zeta A(Z_1 - \zeta I)^{-2} A^\top \succeq 0. \tag{35}$$

In the last step, we establish the above sufficient condition to complete the proof.

Step 1: equivalence between (29) and (34). First notice that

$$\left((1-\eta)I + \eta\Sigma^{-1}\right) Z \left((1-\eta)I + \eta\Sigma^{-1}\right) = \begin{bmatrix} \Sigma_0 Z_0 \Sigma_0 & \Sigma_0 A \Sigma_1 \\ \Sigma_1 A^\top \Sigma_0 & \Sigma_1 Z_1 \Sigma_1 \end{bmatrix}. \tag{36}$$

In order to invoke Schur's complement lemma, we need to verify $\Sigma_1 Z_1 \Sigma_1 - \zeta' I \succ 0$. Observe that by definition we have

$$\Sigma_0 \succeq \left(1 + (K^{-1} - 1)\eta\right) I = (1+3\eta) I, \quad \Sigma_1 \succeq (1-\eta) I. \tag{37}$$

Hence

$$\Sigma_1 Z_1 \Sigma_1 - \zeta' I \succeq (1-\eta)^2 Z_1 - \left(1 + \frac{1}{6}\eta\right)^2 \zeta I \succ 2(1-\eta)^2 \zeta I - \left(1 + \frac{1}{6}\eta\right)^2 \zeta I \succ 0,$$

where in the second inequality we used $Z_1 - 2\zeta I \succ 0$ proved in (32), and in the last inequality we used $\eta \le c_\eta$ with c_η sufficiently small. This completes the verification that $\Sigma_1 Z_1 \Sigma_1 - \zeta' I \succ 0$. Now, invoking Schur's complement lemma yields that (29) is equivalent to

$$\Sigma_0 Z_0 \Sigma_0 - \zeta' I - \Sigma_0 A \Sigma_1 (\Sigma_1 Z_1 \Sigma_1 - \zeta' I)^{-1} \Sigma_1 A^\top \Sigma_0 \succeq 0,$$

which simplifies easily to (34), as claimed.

Step 2: establishing (35) as a sufficient condition for (34). By (37), it follows that

$$\begin{aligned}(Z_1 - \zeta' \Sigma_1^{-2})^{-1} &\preceq (Z_1 - (1-\eta)^{-2} \zeta' I)^{-1} \\ &= \left(Z_1 - \zeta I - \left((1-\eta)^{-2}\zeta' - \zeta\right) I\right)^{-1},\end{aligned} \tag{38}$$

where we used the well-known fact that $A \preceq B$ implies $B^{-1} \preceq A^{-1}$ for positive definite matrices A and B (cf. (Bhatia, 1997, Proposition V.1.6)). We aim to apply Lemma 8 to control the above term, by treating $((1-\eta)^{-2}\zeta' - \zeta) I$ as a perturbation term. For this purpose we need to verify

$$\left|(1-\eta)^{-2}\zeta' - \zeta\right| \le \frac{1}{2} \lambda_{\min}(Z_1 - \zeta I). \tag{39}$$

Given $\eta \le c_\eta$ with sufficiently small c_η, we have $(1-\eta)^{-2} \le 1+3\eta$, $(1+\frac{1}{6}\eta)^2 \le 1+\eta$, and $(1+3\eta)(1+\eta) \le 1+5\eta$, thus

$$0 \le (1-\eta)^{-2}\big(1+\frac{1}{6}\eta\big)^2\zeta - \zeta = (1-\eta)^{-2}\zeta' - \zeta \le (1+3\eta)(1+\eta)\zeta - \zeta \le 5\eta\zeta < \zeta/2,$$

where the last inequality follows from $c_\eta \le 1/10$. On the other hand, invoking (32), we obtain

$$\frac{1}{2}\zeta \le \frac{1}{2}\big(\lambda_{\min}(Z_1) - \zeta\big) = \frac{1}{2}\lambda_{\min}(Z_1 - \zeta I),$$

which verifies (39). Thus we may apply Lemma 8 to show

$$\left\|(Z_1-\zeta I)\left((Z_1-\zeta I)^{-1} - \big(Z_1 - \zeta I - ((1-\eta)^{-2}\zeta' - \zeta)I\big)^{-1}\right)(Z_1-\zeta I)\right\| \le 2\left|(1-\eta)^{-2}\zeta' - \zeta\right| \le 10\eta\zeta,$$

therefore

$$\big(Z_1 - \zeta I - ((1-\eta)^{-2}\zeta' - \zeta)I\big)^{-1} \preceq (Z_1-\zeta I)^{-1} + 10\eta\zeta(Z_1-\zeta I)^{-2}.$$

Together with (38), this implies

$$(Z_1 - \zeta'\Sigma_1^{-2})^{-1} \preceq (Z_1-\zeta I)^{-1} + 10\eta\zeta(Z_1-\zeta I)^{-2}. \tag{40}$$

Combining (37) and (40), we see that a sufficient condition for (34) to hold is (35).

Step 3: establishing (35). It is clear that (35) is implied by

$$\zeta I - (1+3\eta)^{-2}\zeta' I - 10\eta\zeta A(Z_1-\zeta I)^{-2}A^\top \succeq 0, \tag{41}$$

by leveraging the relation $Z_0 \succeq \zeta I + A(Z_1-\zeta I)^{-1}A^\top$ from (33).

Hence, it boils down to prove (41). Recalling $\|A\| \le c_{10}$, and from (32), we know $\lambda_{\min}(Z_1 - \zeta I) \ge K - c_{10} - \zeta \ge \zeta + 3c_{10}$. Thus

$$\|A(Z_1-\zeta I)^{-2}A^\top\| \le \|A\|^2\|(Z_1-\zeta I)^{-2}\| \le c_{10}^2/(\zeta+3c_{10})^2 \le 1/9.$$

Therefore, to prove (41) it suffices to show

$$\zeta - (1+3\eta)^{-2}\zeta' \ge \frac{10}{9}\eta\zeta. \tag{42}$$

It is easy to verify that the above inequality holds for our choice $\zeta' = (1+\frac{1}{6}\eta)^2\zeta$. In fact, given $\eta \le c_\eta$ for sufficiently small c_η, we have $(1+3\eta)^{-2} \le 1-4\eta$, $(1+\frac{1}{6}\eta)^2 \le 1+\eta$. These together yield

$$\zeta - (1+3\eta)^{-2}\big(1+\frac{1}{6}\eta\big)^2\zeta \ge \zeta - (1-4\eta)(1+\eta)\zeta = 3\eta\zeta + 4\eta^2\zeta \ge 3\eta\zeta \ge \frac{10}{9}\eta\zeta,$$

establishing (42) as desired. □

B Decompositions of key terms

In this section, we first present a useful bound of a key error quantity

$$\Delta_t \coloneqq (\mathcal{I} - \mathcal{A}^*\mathcal{A})(X_tX_t^\top - M_\star), \tag{43}$$

where X_t is the iterate of $\mathsf{ScaledGD}(\lambda)$ given in (7).

Lemma 11. *Suppose $\mathcal{A}(\cdot)$ satisfies Assumption 1. For any $t \ge 0$ such that (17) holds, we have*

$$\|\Delta_t\| \le 8\delta\left(\|\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2\|_{\mathsf{F}} + \|\widetilde{S}_t\|\|\widetilde{N}_t\|_{\mathsf{F}} + n\|\widetilde{O}_t\|^2\right). \tag{44}$$

In particular, there exists some constant $c_{11} \lesssim c_\delta/c_\lambda^3$ such that

$$\|\Delta_t\| \le 16(C_{2.a}+1)^2c_\delta\kappa^{-2C_\delta/3}\|X_\star\|^2 \le c_{11}\kappa^{-2C_\delta/3}\|X_\star\|^2. \tag{45}$$

Proof. The decomposition (14) in Proposition 1 yields

$$X_tX_t^\top = U_\star\widetilde{S}_t\widetilde{S}_t^\top U_\star^\top + U_\star\widetilde{S}_t\widetilde{N}_t^\top U_{\star,\perp}^\top + U_{\star,\perp}\widetilde{N}_t\widetilde{S}_t^\top U_\star^\top + U_{\star,\perp}\widetilde{N}_t\widetilde{N}_t^\top U_{\star,\perp}^\top + U_{\star,\perp}\widetilde{O}_t\widetilde{O}_t^\top U_{\star,\perp}^\top.$$

Since $M_\star = U_\star\Sigma_\star^2U_\star^\top$, we have

$$X_tX_t^\top - M_\star = \underbrace{U_\star(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)U_\star^\top}_{=:T_1} + \underbrace{U_\star\widetilde{S}_t\widetilde{N}_t^\top U_{\star,\perp}^\top + U_{\star,\perp}\widetilde{N}_t\widetilde{S}_t^\top U_\star^\top}_{=:T_2} + \underbrace{U_{\star,\perp}\widetilde{N}_t\widetilde{N}_t^\top U_{\star,\perp}^\top}_{=:T_3} + \underbrace{U_{\star,\perp}\widetilde{O}_t\widetilde{O}_t^\top U_{\star,\perp}^\top}_{=:T_4}. \tag{46}$$

Note that $U_\star \in \mathbb{R}^{n\times r_\star}$ is of rank $r_\star$, thus T_1 has rank at most $r_\star$ and T_2 has rank at most $2r_\star$. Similarly, since $\widetilde{N}_t = N_tV_t$ while $V_t \in \mathbb{R}^{r\times r_\star}$ is of rank $r_\star$, T_3 has rank at most $r_\star$. It is also trivial that T_4 as an $n\times n$ matrix has rank at most n. Invoking Lemma 7, we obtain

$$\begin{aligned}
&\|(\mathcal{I}-\mathcal{A}^*\mathcal{A})(T_1)\| \le 2\delta\|U_\star(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)U_\star^\top\|_{\mathsf{F}} \le 2\delta\|\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2\|_{\mathsf{F}},\\
&\|(\mathcal{I}-\mathcal{A}^*\mathcal{A})(T_2)\| \le 2\sqrt{3}\delta\|U_\star\widetilde{S}_t\widetilde{N}_t^\top U_{\star,\perp}^\top + U_{\star,\perp}\widetilde{N}_t\widetilde{S}_t^\top U_\star^\top\|_{\mathsf{F}} \le 4\sqrt{2}\delta\|\widetilde{S}_t\|\|\widetilde{N}_t\|_{\mathsf{F}},\\
&\|(\mathcal{I}-\mathcal{A}^*\mathcal{A})(T_3)\| \le 2\delta\|U_{\star,\perp}\widetilde{N}_t\widetilde{N}_t^\top U_{\star,\perp}^\top\|_{\mathsf{F}} \le 2\delta\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|\|\widetilde{S}_t\|\|\Sigma_\star^{-1}\|\|\widetilde{N}_t\|_{\mathsf{F}} \le \delta\|\widetilde{S}_t\|\|\widetilde{N}_t\|_{\mathsf{F}},\\
&\|(\mathcal{I}-\mathcal{A}^*\mathcal{A})(T_4)\| \le 2\delta n\|U_{\star,\perp}\widetilde{O}_t\widetilde{O}_t^\top U_{\star,\perp}^\top\| \le 2\delta n\|\widetilde{O}_t\|^2,
\end{aligned}$$

where the third line follows from $\|\Sigma_\star^{-1}\| = \kappa\|X_\star\|^{-1}$ and from (17c) in view that C_δ is sufficiently large and c_2 is suffciently small. The conclusion (44) follows from summing up the above inequalities.

For the remaining part of the lemma, note that the following inequalities that bound the individual terms of (44) can be inferred from (17): namely,

$$\|\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star\|_{\mathsf{F}} \le \sqrt{2r_\star}\|\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star\| \le \sqrt{2r_\star}(C_{2.a}^2\kappa^2+1)\|X_\star\|^2$$

by (17d), and

$$\begin{aligned}
\|\widetilde{S}_t\|\|\widetilde{N}_t\|_{\mathsf{F}} &\le \sqrt{r_\star}\|\widetilde{S}_t\|\|\widetilde{N}_t\|\\
&\le \sqrt{r_\star}(C_{2.a}\kappa\|X_\star\|)\cdot\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|\cdot\|\widetilde{S}_t\|\cdot\|\Sigma_\star^{-1}\|\\
&\le \sqrt{r_\star}(C_{2.a}\kappa\|X_\star\|)\cdot(c_2\kappa^{-C_\delta/2}\|X_\star\|)\cdot(C_{2.a}\kappa\|X_\star\|)\cdot\sigma_{\min}^{-1}(\Sigma_\star)\\
&= \sqrt{r_\star}c_2C_{2.a}^2\kappa^3\|X_\star\|^2\kappa^{-C_\delta/2}\\
&\le \sqrt{r_\star}C_{2.a}^2\|X_\star\|^2,
\end{aligned}$$

where the first inequality uses the fact that $\widetilde{N}_t = N_tV_t$ contains a rank-$r_\star$ factor V_t, hence has rank at most $r_\star$; the second line follows from (17d), the third line follows from (17c) and (17d), and the last line follows from choosing c_δ sufficiently small such that $c_2 \le 1$ (which is possible since $c_2 \lesssim c_\delta/c_\lambda^3$) and from choosing $C_\delta \ge 6$ such that $\kappa^3\kappa^{-C_\delta/2} \le 1$. Finally, from (17b) and its corollary (19), we have

$$2n\|\widetilde{O}_t\|^2 \le 2n\alpha^{3/2}\|X_\star\|^{1/2} \le \|X_\star\|^2,$$

since from (12c) it is easy to show that $\alpha \le (2n)^{-2/3}\|X_\star\|$.

Combining these inequalities and (44) yields

$$\|\Delta_t\| \le 8\delta\sqrt{r_\star}(\sqrt{2}C_{2.a}^2\kappa^2 + 1 + C_{2.a}^2+1)\|X_\star\|^2 \le 16\delta\sqrt{r_\star}\kappa^2(C_{2.a}^2+1)\|X_\star\|^2.$$

Recalling that by (10) we have $\delta\sqrt{r_\star}\kappa^2 \le c_\delta\kappa^{-C_\delta+2} \le c_\delta\kappa^{-2C_\delta/3}$ as long as $C_\delta \ge 6$, we obtain the desired conclusion. The bound $c_{11} = 16(C_{2.a}+1)^2c_\delta \lesssim c_\delta/c_\lambda \lesssim c_\delta/c_\lambda^3$ follows from $C_{2.a} \lesssim c_\lambda^{-1/2}$. ☐

We next present several useful decompositions of the signal term S_{t+1} and the noise term N_{t+1}, which are extremely useful in later developments.

Lemma 12. *For any t such that $\widetilde{S}_t$ is invertible and (17) holds, we have*

$$S_{t+1} = \Big((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I + E_t^a)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\Big)\widetilde{S}_tV_t^\top + \eta E_t^b, \tag{47a}$$

$$
\begin{aligned}
N_{t+1} = \widetilde{N}_t\widetilde{S}_t^{-1}\left((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top + \lambda I + \eta E_t^c\right)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t V_t^\top \\
+ \eta E_t^e(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t V_t^\top + \widetilde{O}_t V_{t,\perp}^\top + \eta E_t^d,
\end{aligned} \tag{47b}
$$

where the error terms satisfy

$$
\|E_t^a\| \le 2c_2\kappa^{-4}\|X_\star\| \cdot \|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| + 2\|U_\star^\top \Delta_t\|, \tag{48a}
$$

$$
\|E_t^b\| \le \left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{3/4}\sigma_{\min}(\widetilde{S}_t) \le \frac{1}{20}\kappa^{-10}\sigma_{\min}(\widetilde{S}_t), \tag{48b}
$$

$$
\|E_t^c\| \le \kappa^{-4}\|X_\star\| \cdot \|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|, \tag{48c}
$$

$$
\|E_t^d\| \le \left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{3/4}\sigma_{\min}(\widetilde{S}_t), \tag{48d}
$$

$$
\|E_t^e\| \le 2\|U_\star^\top \Delta_t\| + c_{11}\kappa^{-5}\|X_\star\| \cdot \|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|. \tag{48e}
$$

Moreover, we have

$$
\|E_t^b\| \le \frac{1}{24C_{\max}\kappa}\|\widetilde{O}_t\|, \tag{48f}
$$

$$
\|E_t^d\| \le \frac{1}{24C_{\max}\kappa}\|\widetilde{O}_t\|. \tag{48g}
$$

Here, $\|\cdot\|$ can either be the Frobenius norm or the spectral norm.

To proceed, we would need the approximate update equation of the rotated signal term $\widetilde{S}_{t+1}$, and the rotated misalignment term $\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}$ later in the proof. Since directly analyzing the evolution of these two terms seems challenging, we resort to two surrogate matrices $S_{t+1}V_t + S_{t+1}V_{t,\perp}Q$, and $(N_{t+1}V_t + N_{t+1}V_{t,\perp}Q)(S_{t+1}V_t + S_{t+1}V_{t,\perp}Q)^{-1}$, as documented in the following two lemmas.

Lemma 13. *For any t such that $\widetilde{S}_t$ is invertible and (17) holds, and any matrix $Q \in \mathbb{R}^{(r-r_\star)\times r_\star}$ with $\|Q\| \le 2$, we have*

$$
S_{t+1}V_t + S_{t+1}V_{t,\perp}Q = (I + \eta E_t^{13})\left((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right)\widetilde{S}_t, \tag{49}
$$

where $E_t^{13} \in \mathbb{R}^{r_\star \times r_\star}$ is a matrix (depending on Q) satisfying

$$
\|E_t^{13}\| \le \frac{1}{200(C_{2.a}+1)^4\kappa^5}.
$$

Here, $C_{2.a} > 0$ is given in Lemma 2.

Lemma 14. *For any t such that $\widetilde{S}_t$ is invertible and (17) holds, and any matrix $Q \in \mathbb{R}^{(r-r_\star)\times r_\star}$ with $\|Q\| \le 2$, we have*

$$
\begin{aligned}
&(N_{t+1}V_t + N_{t+1}V_{t,\perp}Q)(S_{t+1}V_t + S_{t+1}V_{t,\perp}Q)^{-1} \\
&= \widetilde{N}_t\widetilde{S}_t^{-1}(1+\eta E_t^{14.a})\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top + \lambda I\big)\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top + \lambda I + \eta\Sigma_\star^2\big)^{-1}(1+\eta E_t^{13})^{-1} + \eta E_t^{14.b}
\end{aligned}
$$

where $E_t^{14.a}$, $E_t^{14.b}$ are matrices (depending on Q) satisfying

$$
\|E_t^{14.a}\| \le \frac{1}{200(C_{2.a}+1)^4\kappa^5}, \tag{50a}
$$

$$
\begin{aligned}
\|E_t^{14.b}\| \le 400c_\lambda^{-1}\kappa^2\|X_\star\|^{-2}\|U_\star^\top\Delta_t\| + \frac{1}{64(C_{2.a}+1)^2\kappa^5\|X_\star\|}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \\
+ \frac{1}{64}\left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{2/3}.
\end{aligned} \tag{50b}
$$

Here, $\|\cdot\|$ can either be the Frobenius norm or the spectral norm, and $C_{2.a} > 0$ is given in Lemma 2.

B.1 Proof of Lemma 12

We split the proof into three steps: (1) provide several useful approximation results regarding the matrix inverses utilizing the facts that $\|\widetilde{O}_t\|$ and $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|$ are small (as shown by Lemma 2); (2) proving the claims (47a), (48a), (48b), and (48f) associated with the signal term S_{t+1}; (3) proving the claims (47b), (48c), (48d), (48e), and (48g) associated with the noise term N_{t+1}. Note that our approximation results in step (1) include choices of some matrices $\{Q_i\}$ with small spectral norms, whose choices may be different from lemma to lemma for simplicity of presentation;

B.1.1 Step 1: preliminaries

We know from (17) that the overparametrization error $\widetilde{O}_t$ is negligible compared to the signals $\widetilde{S}_t$ and $\sigma_{\min}(X_\star)$. This combined with the decomposition (14) reveals a desired approximation $(X_t^\top X_t + \lambda I)^{-1} \approx (V_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t)V_t^\top + \lambda I)^{-1}$. This approximation is formalized in the lemma below.

Lemma 15. *If $\lambda \geq 4(\|\widetilde{O}_t\|^2 \vee 2\|\widetilde{N}_t\|\|\widetilde{O}_t\|)$ for some t, then*

$$\begin{aligned}(X_t^\top X_t + \lambda I)^{-1} &= \Big(V_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t)V_t^\top + \lambda I\Big)^{-1} \\ &\quad + \Big(V_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t)V_t^\top + \lambda I\Big)^{-1} E_t^{15.a} \Big(V_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t)V_t^\top + \lambda I\Big)^{-1} \\ &= \Big(V_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t)V_t^\top + \lambda I\Big)^{-1} (I + E_t^{15.b}) \end{aligned} \tag{51}$$

where the error terms $E_t^{15.a}$, $E_t^{15.b}$ can be expressed as

$$E_t^{15.a} = (V_{t,\perp}\widetilde{O}_t^\top \widetilde{O}_t V_{t,\perp}^\top + V_t\widetilde{N}_t^\top \widetilde{O}_t V_{t,\perp}^\top + V_{t,\perp}\widetilde{O}_t^\top \widetilde{N}_t V_t^\top)Q_1, \tag{52a}$$

$$E_t^{15.b} = \lambda^{-1}E_t^{15.a}Q_2, \tag{52b}$$

for some matrices Q_1, Q_2 such that $\max\{\|Q_1\|, \|Q_2\|\} \leq 2$.

Proof. Expanding $X_t^\top X_t$ according to (14), we have

$$X_t^\top X_t = V_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t)V_t^\top + V_{t,\perp}\widetilde{O}_t^\top \widetilde{O}_t V_{t,\perp}^\top + V_t\widetilde{N}_t^\top \widetilde{O}_t V_{t,\perp}^\top + V_{t,\perp}\widetilde{O}_t^\top \widetilde{N}_t V_t^\top.$$

The conclusion readily follows from Lemma 8 by setting therein $A = V_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t)V_t^\top + \lambda I$ and $B = V_{t,\perp}\widetilde{O}_t^\top \widetilde{O}_t V_{t,\perp}^\top + V_t\widetilde{N}_t^\top \widetilde{O}_t V_{t,\perp}^\top + V_{t,\perp}\widetilde{O}_t^\top \widetilde{N}_t V_t^\top$, where the condition $\|A^{-1}B\| \leq 1/2$ is satisfied since

$$\|A^{-1}B\| \leq \sigma_{\min}(A)^{-1}\|B\| \leq \lambda^{-1} \cdot (\|\widetilde{O}_t\|^2 + 2\|\widetilde{O}_t\|\|\widetilde{N}_t\|) \leq 1/2.$$

□

Moreover, the dominating term on the right hand side of (51) can be equivalently written as

$$\begin{aligned}\Big(V_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t)V_t^\top + \lambda I\Big)^{-1} &= \Big(V_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t + \lambda I)V_t^\top + \lambda V_{t,\perp}V_{t,\perp}^\top\Big)^{-1} \\ &= V_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t + \lambda I)^{-1}V_t^\top + \lambda^{-1}V_{t,\perp}V_{t,\perp}^\top. \end{aligned} \tag{53}$$

When the misalignment error $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|$ is small, we expect $(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t + \lambda I)^{-1} \approx (\widetilde{S}_t^\top \widetilde{S}_t + \lambda I)^{-1}$, which is formalized in the following lemma that establishes $(\widetilde{S}_t\widetilde{S}_t^\top + \widetilde{S}_t\widetilde{N}_t^\top \widetilde{N}_t\widetilde{S}_t^{-1} + \lambda I)^{-1} \approx (\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}$, due to the following approximation

$$\begin{aligned}(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t + \lambda I)^{-1} &= \widetilde{S}_t^{-1}(\widetilde{S}_t\widetilde{S}_t^\top + \widetilde{S}_t\widetilde{N}_t^\top \widetilde{N}_t\widetilde{S}_t^{-1} + \lambda I)^{-1}\widetilde{S}_t \\ &\approx \widetilde{S}_t^{-1}(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t = (\widetilde{S}_t^\top \widetilde{S}_t + \lambda I)^{-1}.\end{aligned}$$

Lemma 16. *If $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \le \sigma_{\min}(X_\star)/16$ for some t, then*

$$(\widetilde{S}_t\widetilde{S}_t^\top + \widetilde{S}_t\widetilde{N}_t^\top\widetilde{N}_t\widetilde{S}_t^{-1} + \lambda I)^{-1} = (I + E_t^{16})(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}, \tag{54}$$

where the error term E_t^{16} is a matrix defined as

$$E_t^{16} = \kappa^2\|X_\star\|^{-2}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|Q_1(\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star)Q_2, \tag{55}$$

where Q_1, Q_2 are matrices of appropriate dimensions satisfying $\|Q_1\| \le 1$, $\|Q_2\| \le 2$. In particular, we have

$$\|\!|E_t^{16}|\!\| \le 2\kappa^2\|X_\star\|^{-2}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \cdot \|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\|, \tag{56}$$

where $\|\!|\cdot|\!\|$ can be either the operator norm or the Frobenius norm.

Proof. In order to apply Lemma 8, setting $A = \widetilde{S}_t\widetilde{S}_t^\top + \lambda I$ and $B = \widetilde{S}_t\widetilde{N}_t^\top\widetilde{N}_t\widetilde{S}_t^{-1}$, it is straightforward to verify that

$$\|A^{-1}B\| = \|(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t\widetilde{N}_t^\top\widetilde{N}_t\widetilde{S}_t^{-1}\| \le \|\widetilde{N}_t\widetilde{S}_t^{-1}\|^2 \le \|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|^2\|\Sigma_\star^{-1}\|^2 \le (1/16)^2,$$

where we use the obvious fact that $\|(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t\widetilde{S}_t^\top\| \le 1$. Applying Lemma 8, we obtain

$$\begin{aligned}
&(\widetilde{S}_t\widetilde{S}_t^\top + \widetilde{S}_t\widetilde{N}_t^\top\widetilde{N}_t\widetilde{S}_t^{-1} + \lambda I)^{-1} - (\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\\
&= (\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t\widetilde{N}_t^\top\widetilde{N}_t\widetilde{S}_t^{-1}Q(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\\
&= (\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t\widetilde{S}_t^\top\Sigma_\star^{-1}(\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star)^\top(\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star)\Sigma_\star^{-1}Q(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}
\end{aligned}$$

for some matrix Q with $\|Q\| \le 2$. Since one may further write

$$\begin{aligned}
&(\widetilde{S}_t\widetilde{S}_t^\top + \widetilde{S}_t\widetilde{N}_t^\top\widetilde{N}_t\widetilde{S}_t^{-1} + \lambda I)^{-1} - (\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\\
&= \|\Sigma_\star^{-1}\|^2\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t\widetilde{S}_t^\top\frac{\Sigma_\star^{-1}}{\|\Sigma_\star^{-1}\|}\frac{(\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star)^\top}{\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|}(\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star)\frac{\Sigma_\star^{-1}}{\|\Sigma_\star^{-1}\|}Q(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1},
\end{aligned}$$

the conclusion follows by setting E_t^{16} as in (55) with

$$Q_1 = (\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t\widetilde{S}_t^\top\frac{\Sigma_\star^{-1}}{\|\Sigma_\star^{-1}\|}\frac{(\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star)^\top}{\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|}, \quad Q_2 = \frac{\Sigma_\star^{-1}}{\|\Sigma_\star^{-1}\|}Q.$$

The last inequality (56) is then a direct consequence of (55). ∎

B.1.2 Step 2: a key recursion

Recall the definition Δ_t in (43), we can rewrite the update equation (7) as

$$X_{t+1} = X_t - \eta(X_tX_t^\top - M_\star)X_t(X_t^\top X_t + \lambda I)^{-1} + \eta\Delta_tX_t(X_t^\top X_t + \lambda I)^{-1}. \tag{57}$$

Multiplying both sides of (57) by $U_\star^\top$ on the left, we obtain

$$\begin{aligned}
S_{t+1} &= S_t - \eta S_tX_t^\top X_t(X_t^\top X_t + \lambda I)^{-1} + \eta\Sigma_\star^2S_t(X_t^\top X_t + \lambda I)^{-1} + \eta U_\star^\top\Delta_tX_t(X_t^\top X_t + \lambda I)^{-1}\\
&= (1-\eta)S_t + \eta(\Sigma_\star^2 + \lambda I + U_\star^\top\Delta_tU_\star)S_t(X_t^\top X_t + \lambda I)^{-1} + \eta U_\star^\top\Delta_tU_{\star,\perp}N_t(X_t^\top X_t + \lambda I)^{-1}.
\end{aligned} \tag{58}$$

Similarly, multiplying both sides of (57) by $U_{\star,\perp}^\top$, we obtain

$$\begin{aligned}
N_{t+1} &= N_t\left(I - \eta X_t^\top X_t(X_t^\top X_t + \lambda I)^{-1}\right) + \eta U_{\star,\perp}^\top\Delta_tX_t(X_t^\top X_t + \lambda I)^{-1}\\
&= (1-\eta)N_t + \eta\lambda N_t(X_t^\top X_t + \lambda I)^{-1} + \eta U_{\star,\perp}^\top\Delta_tU_\star S_t(X_t^\top X_t + \lambda I)^{-1} + \eta U_{\star,\perp}^\top\Delta_tU_{\star,\perp}N_t(X_t^\top X_t + \lambda I)^{-1}.
\end{aligned} \tag{59}$$

These expressions motivate the need to study the terms $S_t(X_t^\top X_t + \lambda I)^{-1}$ and $N_t(X_t^\top X_t + \lambda I)^{-1}$, which we formalize in the following lemma.

Lemma 17. *Under the same setting as Lemma 12, we have*

$$S_t(X_t^\top X_t+\lambda I)^{-1} = (I+E_t^{16})(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\widetilde{S}_tV_t^\top + E_t^{17.a}, \tag{60a}$$

$$N_t(X_t^\top X_t+\lambda I)^{-1} = \widetilde{N}_t\widetilde{S}_t^{-1}(I+E_t^{16})(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\widetilde{S}_tV_t^\top + \lambda^{-1}\widetilde{O}_tV_{t,\perp}^\top + E_t^{17.b}, \tag{60b}$$

where E_t^{16} is given in (55), and the error terms $E_t^{17.a}$, $E_t^{17.b}$ can be expressed as

$$E_t^{17.a} = \kappa\lambda^{-1}\|X_\star\|^{-1}\|\widetilde{O}_t\|Q_1(\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star)^\top Q_2, \tag{61a}$$

$$\begin{aligned} E_t^{17.b} &= \left(\widetilde{N}_t(\widetilde{S}_t^\top\widetilde{S}_t+\widetilde{N}_t^\top\widetilde{N}_t+\lambda I)^{-1}V_t^\top+\lambda^{-1}\widetilde{O}_tV_{t,\perp}^\top\right)E_t^{15.b} \\ &= \lambda^{-1}(\|\widetilde{N}_t\|Q_3+\|\widetilde{O}_t\|Q_4)E_t^{15.b}. \end{aligned} \tag{61b}$$

for some matrices $\{Q_i\}_{1\le i\le 4}$ with spectral norm bounded by 2, and $E_t^{15.b}$ defined in (52b).

Proof. To begin, combining Lemma 15 and the discussion thereafter (cf. (51)–(53)) and the fact that $\widetilde{S}_t = S_tV_t$, we have for some matrix Q with $\|Q\|\le 2$ that

$$\begin{aligned} S_t(X_t^\top X_t+\lambda I)^{-1} &= \widetilde{S}_t(\widetilde{S}_t^\top\widetilde{S}_t+\widetilde{N}_t^\top\widetilde{N}_t+\lambda I)^{-1}V_t^\top\left(I+E_t^{15.b}\right) \\ &= \widetilde{S}_t(\widetilde{S}_t^\top\widetilde{S}_t+\widetilde{N}_t^\top\widetilde{N}_t+\lambda I)^{-1}V_t^\top + \widetilde{S}_t(\widetilde{S}_t^\top\widetilde{S}_t+\widetilde{N}_t^\top\widetilde{N}_t+\lambda I)^{-1}\lambda^{-1}\widetilde{N}_t^\top\widetilde{O}_tQ \\ &= (\widetilde{S}_t\widetilde{S}_t^\top+\widetilde{S}_t\widetilde{N}_t^\top\widetilde{N}_t\widetilde{S}_t^{-1}+\lambda I)^{-1}\widetilde{S}_tV_t^\top \\ &\quad + \widetilde{S}_t(\widetilde{S}_t^\top\widetilde{S}_t+\widetilde{N}_t^\top\widetilde{N}_t+\lambda I)^{-1}\widetilde{S}_t^\top(\widetilde{N}_t\widetilde{S}_t^{-1})^\top(\widetilde{O}_t/\lambda)Q. \end{aligned} \tag{62}$$

Note that the condition of Lemma 15 can be verified as follows: since

$$\|\widetilde{O}_t\| \le C_{2.b}^{-C_{2.b}}\kappa^{-3}\cdot\|X_\star\|\cdot\sigma_{\min}\left((\Sigma_\star^2+\lambda I)^{-1/2}\right)\cdot\|\widetilde{S}_t\| \le C_{2.b}^{-C_{2.b}}C_{2.a}\sigma_{\min}(X_\star),$$

$$\|\widetilde{N}_t\| \le \|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|\cdot\|\Sigma_\star^{-1}\|\cdot\|\widetilde{S}_t\| \le c_2\kappa^{-C_\delta/2}\|X_\star\|\cdot\frac{C_{2.a}\kappa\|X_\star\|}{\sigma_{\min}(X_\star)} \le c_2C_{2.a}\sigma_{\min}(X_\star)$$

provided $C_\delta\ge 6$, the bounds $c_2\lesssim c_\delta/c_\lambda^3$ and $C_{2.a}\lesssim c_\lambda^{-1/2}$ imply that when we choose C_α to be large enough (depending on c_λ, c_δ),

$$2\|\widetilde{N}_t\|\|\widetilde{O}_t\|\vee\|\widetilde{O}_t\|^2\le\lambda/4,$$

as desired.

Now the first term in (62) can be handled by invoking Lemma 16, since its condition is verified by $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|\le c_2\kappa^{-(C_\delta/2-1)}\sigma_{\min}(X_\star)\le\sigma_{\min}(X_\star)/16$ provided $C_\delta\ge 2$ and $c_2\le 1/16$ by choosing c_δ sufficiently small (depending on c_λ). Namely,

$$(\widetilde{S}_t\widetilde{S}_t^\top+\widetilde{S}_t\widetilde{N}_t^\top\widetilde{N}_t\widetilde{S}_t^{-1}+\lambda I)^{-1}\widetilde{S}_tV_t^\top = (I+E_t^{16})(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\widetilde{S}_tV_t^\top.$$

For the second term, by noting that

$$\|\widetilde{S}_t(\widetilde{S}_t^\top\widetilde{S}_t+\widetilde{N}_t^\top\widetilde{N}_t+\lambda I)^{-1}\widetilde{S}_t^\top\|\le\|\widetilde{S}_t(\widetilde{S}_t^\top\widetilde{S}_t+\lambda I)^{-1}\widetilde{S}_t^\top\|\le 1,$$

it can be expressed as

$$\lambda^{-1}\|\widetilde{O}_t\|\widetilde{S}_t(\widetilde{S}_t^\top\widetilde{S}_t+\lambda I)^{-1}\widetilde{S}_t^\top(\widetilde{N}_t\widetilde{S}_t^{-1})^\top(\widetilde{O}_t/\|\widetilde{O}_t\|)Q = \kappa\lambda^{-1}\|X_\star\|^{-1}\|\widetilde{O}_t\|Q_1(\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star)^\top Q_2$$

for $Q_1 = \widetilde{S}_t(\widetilde{S}_t^\top\widetilde{S}_t+\lambda I)^{-1}\widetilde{S}_t^\top\cdot\kappa^{-1}\|X_\star\|\Sigma_\star^{-1}$ with $\|Q_1\|\le 1$ and $Q_2 = (\widetilde{O}_t/\|\widetilde{O}_t\|)Q$ which satisfies $\|Q_2\|\le\|Q\|\le 2$. Applying the above two bounds to (62) yields (60a).

Similarly, moving to (60b), it follows that

$$\begin{aligned} N_t(X_t^\top X_t+\lambda I)^{-1} &= \left(\widetilde{N}_t(\widetilde{S}_t^\top\widetilde{S}_t+\widetilde{N}_t^\top\widetilde{N}_t+\lambda I)^{-1}V_t^\top+\lambda^{-1}\widetilde{O}_tV_{t,\perp}^\top\right)\left(I+E_t^{15.b}\right) \\ &= \widetilde{N}_t(\widetilde{S}_t^\top\widetilde{S}_t+\widetilde{N}_t^\top\widetilde{N}_t+\lambda I)^{-1}V_t^\top+\lambda^{-1}\widetilde{O}_tV_{t,\perp}^\top+E_t^{17.b}, \end{aligned} \tag{63}$$

where we have

$$
\begin{aligned}
E_t^{17.b} &= \left(\widetilde{N}_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t + \lambda I)^{-1} V_t^\top + \lambda^{-1}\widetilde{O}_t V_{t,\perp}^\top\right) E_t^{15.b} \\
&= \lambda^{-1}(\|\widetilde{N}_t\| Q_3 + \|\widetilde{O}_t\| Q_4) E_t^{15.b}
\end{aligned}
$$

for some matrices Q_3, Q_4 with $\|Q_3\|, \|Q_4\| \le 1$. In the last line we used $\|(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t + \lambda I)^{-1}\| \le \lambda^{-1}$. For the first term of (63), we use Lemma 16 and obtain

$$
\begin{aligned}
\widetilde{N}_t(\widetilde{S}_t^\top \widetilde{S}_t + \widetilde{N}_t^\top \widetilde{N}_t + \lambda I)^{-1} V_t^\top &= \widetilde{N}_t \widetilde{S}_t^{-1}(\widetilde{S}_t \widetilde{S}_t^\top + \widetilde{S}_t \widetilde{N}_t^\top \widetilde{N}_t \widetilde{S}_t^{-1} + \lambda I)^{-1} \widetilde{S}_t V_t^\top \\
&= \widetilde{N}_t \widetilde{S}_t^{-1}(I + E_t^{16})(\widetilde{S}_t \widetilde{S}_t^\top + \lambda I)^{-1} \widetilde{S}_t V_t^\top .
\end{aligned}
$$

This yields the representation in (60b).

□

B.1.3 Step 3: proofs associated with S_{t+1}.

With the help of Lemma 17, we are ready to prove (47a) and the associated norm bounds (48a), (48b), and (48f). To begin with, we plug (60a), (60b) into (58) and use $S_t = \widetilde{S}_t V_t^\top$ to obtain

$$
S_{t+1} = \left((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I + E_t^a)(\widetilde{S}_t \widetilde{S}_t^\top + \lambda I)^{-1}\right) \widetilde{S}_t V_t^\top + \eta E_t^b,
$$

where the error terms E_t^a and E_t^b are

$$
\begin{aligned}
E_t^a &\coloneqq U_\star^\top \Delta_t U_\star + (\Sigma_\star^2 + U_\star^\top \Delta_t U_\star + \lambda I) E_t^{16} + U_\star^\top \Delta_t U_{\star,\perp} \widetilde{N}_t \widetilde{S}_t^{-1}(I + E_t^{16}), \\
E_t^b &\coloneqq (\Sigma_\star^2 + U_\star^\top \Delta_t U_\star + \lambda I) E_t^{17.a} + U_\star^\top \Delta_t U_{\star,\perp}(\lambda^{-1}\widetilde{O}_t V_{t,\perp}^\top + E_t^{17.b}).
\end{aligned}
$$

This establishes the identity (47a). To control $|||E_t^a|||$, we observe that

$$
\begin{aligned}
|||E_t^a||| &\le |||U_\star^\top \Delta_t||| + \|\Sigma_\star^2 + U_\star^\top \Delta_t U_\star + \lambda I\| \cdot |||E_t^{16}||| + |||U_\star^\top \Delta_t||| \cdot \|\widetilde{N}_t \widetilde{S}_t^{-1} \Sigma_\star\| \cdot \|\Sigma_\star^{-1}\| \cdot (1 + \|E_t^{16}\|) \\
&\le \left(1 + c_{11}\kappa^{-2C_\delta/3} + c_\lambda\right) \|X_\star\|^2 \cdot |||E_t^{16}||| + |||U_\star^\top \Delta_t||| + c_2 \kappa^{-C_\delta/2} \|X_\star\| \cdot \sigma_{\min}^{-1}(X_\star) \cdot (1 + \|E_t^{16}\|) \cdot |||U_\star^\top \Delta_t||| \\
&\le 2\|X_\star\|^2 \cdot |||E_t^{16}||| + \left(1 + c_2(1 + \|E_t^{16}\|)\right) |||U_\star^\top \Delta_t|||,
\end{aligned}
$$

where the second line follows from Lemma 11 and Equations (12b), (17c); the last line holds since c_{11}, c_λ are sufficiently small and C_δ is sufficiently large. Now we invoke the bound (56) in Lemma 16 to see

$$
\begin{aligned}
|||E_t^{16}||| \le 2\kappa^2 \|X_\star\|^{-2} \|\widetilde{N}_t \widetilde{S}_t^{-1} \Sigma_\star\| |||\widetilde{N}_t \widetilde{S}_t^{-1} \Sigma_\star||| &\le 2c_2 \kappa^2 \kappa^{-C_\delta/2} \|X_\star\|^{-1} |||\widetilde{N}_t \widetilde{S}_t^{-1} \Sigma_\star||| \\
&\le 2c_2 \kappa^{-4} \|X_\star\|^{-1} |||\widetilde{N}_t \widetilde{S}_t^{-1} \Sigma_\star|||,
\end{aligned}
$$

where the last line follows again by choosing sufficiently large $C_\delta \ge 12$. Furthermore, since $\|\widetilde{N}_t \widetilde{S}_t^{-1} \Sigma_\star\| \le c_2 \kappa^{-C_\delta/2} \|X_\star\|$ for small enough c_2, we obtain $\|E_t^{16}\| \le 1$. Combining these inequalities yields the claimed bound

$$
|||E_t^a||| \le 2c_2 \kappa^{-4} \|X_\star\| \cdot |||\widetilde{N}_t \widetilde{S}_t^{-1} \Sigma_\star||| + 2|||U_\star^\top \Delta_t|||.
$$

The bound of $|||E_t^b|||$ and $\|E_t^b\|$ can be proved in a similar way, utilizing the bound for $\|\widetilde{O}_t\|$ in (19). In fact, a computation similar to the above shows

$$
\begin{aligned}
|||E_t^b||| &\le 2\|X_\star\|^2 \cdot |||E^{17.a}||| + \lambda^{-1}\|\Delta_t\| \cdot |||\widetilde{O}_t||| + \|\Delta_t\| \cdot |||E^{17.b}||| \\
&\le 2\kappa\lambda^{-1} \cdot \|X_\star\| \cdot \|\widetilde{O}_t\| \cdot \|Q_1\| \cdot \|Q_2\| \cdot |||\widetilde{N}_t \widetilde{S}_t^{-1} \Sigma_\star||| + 100 c_\lambda^{-1} \sigma_{\min}^{-1}(M_\star) c_{11} \kappa^{-2C_\delta/3} \|X_\star\|^2 \cdot |||\widetilde{O}_t||| \\
&\quad + 8\lambda^{-2} c_{11} \kappa^{-2C_\delta/3}(\|\widetilde{N}_t\| + \|\widetilde{O}_t\|) \|\widetilde{N}_t\| \cdot |||\widetilde{O}_t||| \\
&\le 800\kappa^3 c_\lambda^{-1} \|X_\star\|^{-1} \|\widetilde{O}_t\| \cdot |||\widetilde{N}_t \widetilde{S}_t^{-1} \Sigma_\star||| + \frac{1}{48(C_{\max}+1)\kappa} |||\widetilde{O}_t|||.
\end{aligned}
$$

Here, $C_{\max}$ is the constant given by Lemma 2. Similarly, we have

$$\|E_t^b\| \le 800\kappa^3 c_\lambda^{-1}\|X_\star\|^{-1}\|\widetilde{O}_t\| \cdot \|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| + \frac{1}{48(C_{\max}+1)\kappa}\|\widetilde{O}_t\|.$$

The bound (48f) now follows directly from the bound of $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|$ in Lemma 2, provided c_δ is sufficiently small and C_δ is sufficiently large. To prove (48b), we note that

$$|||A||| \le n\|A\| \tag{64}$$

for any unitarily invariant norm $|||\cdot|||$ and real matrix $A \in \mathbb{R}^{p\times q}$ with $p \vee q \le n$ (which can be easily verified when $|||\cdot||| = \|\cdot\|$ or $\|\cdot\|_{\mathsf{F}}$). Thus

$$|||E_t^b||| \le \left(800\kappa^3 c_\lambda^{-1} c_2 \kappa^{-C_\delta/2} + \frac{1}{24(C_{\max}+1)\kappa}\right) n\|\widetilde{O}_t\| \le \left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{3/4} \sigma_{\min}(\widetilde{S}_t)$$

where the last inequality follows from the control of $\|\widetilde{O}_t\|$ given by (18) provided c_2 is sufficiently small and $C_{2.b}$ therein is sufficiently large. This establishes the first inequality in (48b), and the second inequality therein follows directly from (18).

B.1.4 Step 4: proofs associated with $\widetilde{N}_{t+1}$.

Now we move on to prove the identity (47b), and the norm controls (48c), (48d), (48e), and (48g) associated with the misalignment term $\widetilde{N}_{t+1}$. Plugging (60a), (60b) into (59) and using the decomposition $N_t = \widetilde{N}_t V_t^\top + \widetilde{O}_t V_{t,\perp}^\top$, we have

$$\begin{aligned} N_{t+1} &= \widetilde{N}_t\widetilde{S}_t^{-1}\left((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top + \lambda I + \eta E_t^c\right)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t V_t^\top \\ &\quad + \eta E_t^e(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t V_t^\top + \widetilde{O}_t V_{t,\perp}^\top + \eta E_t^d, \end{aligned}$$

where the error terms are defined to be

$$\begin{aligned} E_t^c &:= \lambda E_t^{16}, \\ E_t^d &:= (\lambda I + U_{\star,\perp}^\top \Delta_t U_{\star,\perp})E_t^{17.b} + \lambda^{-1}U_{\star,\perp}^\top\Delta_t U_{\star,\perp}\widetilde{O}_t V_{t,\perp}^\top + U_{\star,\perp}^\top \Delta_t U_\star E_t^{17.a}, \\ E_t^e &:= U_{\star,\perp}^\top\Delta_t U_\star(I + E_t^{16}) + U_{\star,\perp}^\top\Delta_t U_{\star,\perp}\widetilde{N}_t\widetilde{S}_t^{-1}(I + E_t^{16}). \end{aligned}$$

This establishes the decomposition (47b). The remaining norm controls follow from the expressions above and similar computation as we have done for S_{t+1}. For the sake of brevity, we omit the details.

B.2 Proof of Lemma 13

Use the identity (47a) in Lemma 12 and the fact that V_t and $V_{t,\perp}$ have orthogonal columns to obtain

$$\begin{aligned} S_{t+1}V_t + S_{t+1}V_{t,\perp}Q &= \left((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I + E_t^a)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right)\widetilde{S}_t + \eta E_t^b(V_t + V_{t,\perp}Q) \\ &= (I + \eta E_t^{13})\left((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right)\widetilde{S}_t \\ &= (I + \eta E_t^{13})\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top + \lambda I + \eta\Sigma_\star^2\big)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_t, \end{aligned} \tag{65}$$

where E_t^{13} is defined to be

$$\begin{aligned} E_t^{13} &:= \left(E_t^a(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1} + E_t^b(V_t + V_{t,\perp}Q)\widetilde{S}_t^{-1}\right)\left((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right)^{-1} \\ &= E_t^a\left((1-\eta)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I) + \eta(\Sigma_\star^2 + \lambda I)\right)^{-1} \end{aligned}$$

$$
\begin{aligned}
&\quad + E_t^b(V_t + V_{t,\perp}Q)\widetilde{S}_t^{-1}\Big((1-\eta)I + \eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\Big)^{-1} \\
&=: T_1 + T_2,
\end{aligned}
$$

where the invertibility of $\widetilde{S}_t$ follows from Lemma 2, and the invertibility of $(1-\eta)I+\eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}$ follows from (106).

Since $(1-\eta)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)+\eta(\Sigma_\star^2+\lambda I)\succeq \lambda I$ and $\lambda \ge \frac{1}{100}c_\lambda\sigma_{\min}(M_\star)$ by (12b), we have

$$
\|T_1\| \le \lambda^{-1}\|E_t^a\| \le 100c_\lambda^{-1}\sigma_{\min}^{-1}(M_\star)\|E_t^a\|.
$$

In view of the bound (48a) on $\|E_t^a\|$ in Lemma 12, we further have

$$
\begin{aligned}
\|T_1\| &\le 100c_\lambda^{-1}\sigma_{\min}^{-2}(X_\star)(\kappa^{-4}\|X_\star\|\cdot\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| + \|\Delta_t\|) \\
&\le 100c_\lambda^{-1}\kappa^2\|X_\star\|^{-2}(\kappa^{-4}c_2\kappa^{-C_\delta/2} + c_{11}\kappa^{-2C_\delta/3})\|X_\star\|^2 \\
&\le \frac{1}{400(C_{2.a}+1)^4\kappa^5},
\end{aligned}
$$

where the second inequality follows from (17c) in Lemma 2 and Lemma 11, and the last inequality holds as long as c_2 and c_{11} are sufficiently small and C_δ is sufficiently large (by first fixing c_λ and then choosing c_δ to be sufficiently small).

The term T_2 can be controlled in a similar way. Since $\|AB\|\le\|A\|\cdot\|B\|$, one has

$$
\begin{aligned}
\|T_2\| &\le \|E_t^b\|\cdot(\|V_t\|+\|V_{t,\perp}\|\|Q\|)\cdot\|\widetilde{S}_t^{-1}\|\cdot\sigma_{\min}^{-1}\Big((1-\eta)I+\eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\Big) \\
&\overset{\text{(i)}}{\le} 3\|E_t^b\|\cdot\sigma_{\min}^{-1}(\widetilde{S}_t)\cdot\frac{\kappa}{1-\eta} \overset{\text{(ii)}}{\le} 6\kappa\left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{3/4} \overset{\text{(iii)}}{\le} \frac{1}{400(C_{2.a}+1)^4\kappa^5}.
\end{aligned}
$$

Here, (i) follows from the bound (106) and the facts that $\|V_t\|\vee\|V_{t,\perp}\|\le 1$, $\|Q\|\le 2$; (ii) arises from the control (48b) on $\|E_t^b\|$ in Lemma 12 as well as the condition $\eta\le c_\eta\le 1/2$; and (iii) follows from the implication (18) of Lemma 2.

The proof is completed by summing up the bounds on $\|T_1\|$ and $\|T_2\|$.

B.3 Proof of Lemma 14

Similar to the proof of Lemma 13, we can use the identity (47b) in Lemma 12 and the fact that V_t and $V_{t,\perp}$ have orthogonal columns to obtain

$$
\begin{aligned}
N_{t+1}V_t + N_{t+1}V_{t,\perp}Q &= \widetilde{N}_t\widetilde{S}_t^{-1}\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top+\lambda I+\eta E_t^c\big)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\widetilde{S}_t + \eta E_t^{14.c} \\
&= \widetilde{N}_t\widetilde{S}_t^{-1}(I+\eta E_t^{14.a})\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top+\lambda I\big)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\widetilde{S}_t + \eta E_t^{14.c},
\end{aligned} \tag{66}
$$

where the error terms are defined to be

$$
E_t^{14.c} := E_t^e(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\widetilde{S}_t + \eta^{-1}\widetilde{O}_tQ + E_t^d(V_t+V_{t,\perp}Q), \tag{67}
$$

$$
E_t^{14.a} := E_t^c\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top+\lambda I\big)^{-1}. \tag{68}
$$

Combine (66) and (65) to arrive at

$$
\begin{aligned}
&(N_{t+1}V_t+N_{t+1}V_{t,\perp}Q)(S_{t+1}V_t+S_{t+1}V_{t,\perp}Q)^{-1} \\
&\quad = \widetilde{N}_t\widetilde{S}_t^{-1}(I+\eta E_t^{14.a})\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top+\lambda I\big)\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top+\lambda I+\eta\Sigma_\star^2\big)^{-1}(I+\eta E_t^{13})^{-1} + \eta E_t^{14.b},
\end{aligned} \tag{69}
$$

where, using

$$
(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top+\lambda I+\eta\Sigma_\star^2\big)^{-1} = \big((1-\eta)I+\eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\big)^{-1},
$$

we have

$$
\begin{aligned}
E_t^{14.b} &:= E_t^{14.c}\widetilde{S}_t^{-1}(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top + \lambda I + \eta\Sigma_\star^2\big)^{-1}(I+\eta E_t^{13})^{-1} \\
&= E_t^e\big((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top + \lambda I + \eta\Sigma_\star^2\big)^{-1}(I+\eta E_t^{13})^{-1} \\
&\quad + \eta^{-1}\widetilde{O}_tQ\widetilde{S}_t^{-1}\Big((1-\eta)I + \eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\Big)^{-1}(I+\eta E_t^{13})^{-1} \\
&\quad + E_t^d(V_t + V_{t,\perp}Q)\widetilde{S}_t^{-1}\Big((1-\eta)I + \eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\Big)^{-1}(I+\eta E_t^{13})^{-1} \\
&=: T_1 + T_2 + T_3.
\end{aligned}
$$

It remains to bound $\|E^{14.a}\|$ and $|||E^{14.b}|||$. By (48c), we have

$$
\begin{aligned}
\|E^{14.a}\| \le \lambda^{-1}\|E_t^c\| &\le 100c_\lambda^{-1}\sigma_{\min}^{-2}(X_\star)\cdot\kappa^{-4}\|X_\star\|\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \\
&\le 100c_\lambda^{-1}c_2\kappa^{-2}\kappa^{-C_\delta/2} \\
&\le \frac{1}{200(C_{2.a}+1)^4\kappa^5},
\end{aligned}
$$

where the penultimate inequality follows from (17c) and the last inequality holds with the proviso that c_2 is sufficiently small and C_δ is sufficiently large.

Now we move to bound $|||E^{14.b}|||$. To this end, the relation $\|(I+\eta E_t^{13})^{-1}\| \le 2$ is quite helpful. This follows from Lemma 13 in which we have established that $\|E_t^{13}\| \le 1/2$. As a result of this relation, we obtain

$$
\begin{aligned}
&|||T_1||| \le 2\lambda^{-1}|||E_t^e|||, \\
&|||T_2||| \le 2|||\widetilde{O}_t|||\cdot\|Q\|\cdot\|\widetilde{S}_t^{-1}\|\cdot\left\|\Big((1-\eta)I + \eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\Big)^{-1}\right\|, \\
&|||T_3||| \le 2|||E_t^d|||\cdot(1+\|Q\|)\cdot\|\widetilde{S}_t^{-1}\|\cdot\left\|\Big((1-\eta)I + \eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\Big)^{-1}\right\|.
\end{aligned}
$$

Similar to the control of T_1 in the proof of Lemma 13, we can take the condition $\lambda \ge \frac{1}{100}c_\lambda\sigma_{\min}^2(X_\star)$ and the bound (48e) collectively to see that

$$
|||T_1||| \le 400c_\lambda^{-1}\kappa^2\|X_\star\|^{-2}|||U_\star^\top\Delta_t||| + \frac{1}{64(C_{2.a}+1)^2\kappa^3\|X_\star\|}|||\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|||.
$$

Regarding the terms T_2 and T_3, we see from (106) that

$$
\left\|\Big((1-\eta)I + \eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\Big)^{-1}\right\| \le \frac{\kappa}{1-\eta} \le 2\kappa,
$$

as long η is sufficiently small. Recalling the assumption $\|Q\| \le 2$, this allows us to obtain

$$
\begin{aligned}
&|||T_2||| \le 8\eta^{-1}\kappa\frac{|||\widetilde{O}_t|||}{\sigma_{\min}(\widetilde{S}_t)} \le 8\eta^{-1}\kappa n\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}, \\
&|||T_3||| \le 12\kappa|||E_t^d|||/\sigma_{\min}(\widetilde{S}_t),
\end{aligned}
$$

where the first inequality again uses the elementary fact $|||\widetilde{O}_t||| \le n\|\widetilde{O}_t\|$ in (64).

The desired bounds then follow from plugging in the bounds (48d) and (19).

C Proofs for Phase I

The goal of this section is to prove Lemma 2 in an inductive manner. We achieve this goal in two steps. In Section C.1, we find an iteration number $t_1 \le T_{\min}/16$ such that the claim (17) is true at t_1. This establishes the base case. Then in Section C.2, we prove the induction step, namely if the claim (17) holds for some iteration $t \ge t_1$, we aim to show that (17) continues to hold for the iteration $t+1$. These two steps taken collectively finishes the proof of Lemma 2.

C.1 Establishing the base case: Finding a valid t_1

The following lemma ensures the existence of such an iteration number t_1.

Lemma 18. *Under the same setting as Theorem 2, we have for some $t_1 \le T_{\min}/16$ such that (16) holds and that (17) hold with $t = t_1$.*

The rest of this subsection is devoted to the proof of this lemma.

Define an auxiliary sequence

$$\widehat{X}_t \coloneqq \Big(I + \frac{\eta}{\lambda}\mathcal{A}^*\mathcal{A}(M_\star)\Big)^t X_0, \tag{70}$$

which can be viewed as power iterations on the matrix $\mathcal{A}^*\mathcal{A}(M_\star)$ from the initialization X_0.

In what follows, we first establish that the true iterates $\{X_t\}$ stay close to the auxiliary iterates $\{\widehat{X}_t\}$ as long as the initialization size α is small; see Lemma 19. This proximity then allows us to invoke the result in Stöger and Soltanolkotabi (2021) (see Lemma 20) to establish Lemma 18. For the rest of the appendices, we work on the following event given in (13):

$$\mathcal{E} = \{\|G\| \le C_G\} \cap \{\sigma_{\min}^{-1}(\widehat{U}^\top G) \le (2n)^{C_G}\}.$$

Step 1: controlling distance between X_t and $\widehat{X}_t$. The following lemma guarantees the closeness between the two iterates $\{X_t\}$ and $\{\widehat{X}_t\}$, with the proof deferred to Appendix C.1.1. Recall that C_G is the constant defined in the event $\mathcal{E}$ in (13), and c_λ is the constant given in Theorem 2.

Lemma 19. *Suppose that $\lambda \ge \frac{1}{100}c_\lambda \sigma_{\min}^2(X_\star)$. For any $\theta \in (0,1)$, there exists a large enough constant $K = K(\theta, c_\lambda, C_G) > 0$ such that the following holds: As long as α obeys*

$$\log\frac{\|X_\star\|}{\alpha} \ge \frac{K}{\eta}\log(2\kappa n)\cdot\Big(1 + \log\Big(1 + \frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big)\Big), \tag{71}$$

one has for all $t \le \frac{1}{\theta\eta}\log(\kappa n)$:

$$\big\|X_t - \widehat{X}_t\big\| \le t\Big(1 + \frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big)^t \frac{\alpha^2}{\|X_\star\|}. \tag{72}$$

Moreover, $\|X_t\| \le \|X_\star\|$ for all such t.

Step 2: borrowing a lemma from Stöger and Soltanolkotabi (2021). Compared to the original sequence X_t, the behavior of the power iterates $\widehat{X}_t$ is much easier to analyze. Now that we have sufficient control over $\|X_t - \widehat{X}_t\|$, it is possible to show that X_t has the desired properties in Lemma 18 by first establishing the corresponding property of $\widehat{X}_t$ and then invoking a standard matrix perturbation argument. Fortunately, such a strategy has been implemented by Stöger and Soltanolkotabi (2021) and wrapped into the following helper lemma.

Denote

$$s_j \coloneqq \sigma_j\Big(I + \frac{\eta}{\lambda}\mathcal{A}^*\mathcal{A}(M_\star)\Big) = 1 + \frac{\eta}{\lambda}\sigma_j\big(\mathcal{A}^*\mathcal{A}(M_\star)\big), \qquad j = 1, 2, \dots, n$$

and recall that $\widehat{U}$ (resp. $U_{\widetilde{X}_t}$) is an orthonormal basis of the eigenspace associated with the $r_\star$ largest eigenvalues of $\mathcal{A}^*\mathcal{A}(M_\star)$ (resp. $\widetilde{X}_t$).

Lemma 20. *There exists some small universal $c_{20} > 0$ such that the following hold. Assume that for some $\gamma \le c_{20}$,*

$$\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(M_\star)\| \le \gamma\sigma_{\min}^2(X_\star), \tag{73}$$

and furthermore,

$$\phi := \frac{\alpha\|G\| s_{r_\star+1}^t + \|X_t - \widehat{X}_t\|}{\alpha\sigma_{\min}(\widehat{U}^\top G) s_{r_\star}^t} \le c_{20}\kappa^{-2}. \tag{74}$$

Then there exists some universal $C_{20} > 0$ such that the following hold:

$$\sigma_{\min}(\widetilde{S}_t) \ge \frac{\alpha}{4}\sigma_{\min}(\widehat{U}^\top G) s_{r_\star}^t, \tag{75a}$$

$$\|\widetilde{O}_t\| \le C_{20}\phi\alpha\sigma_{\min}(\widehat{U}^\top G) s_{r_\star}^t, \tag{75b}$$

$$\|U_{\star,\perp}^\top U_{\widetilde{X}_t}\| \le C_{20}(\gamma + \phi), \tag{75c}$$

where $\widetilde{X}_t := X_t V_t \in \mathbb{R}^{n\times r_\star}$.

Proof of Lemma 20. This follows from the claims of Stöger and Soltanolkotabi (2021, Lemma 8.5) by noting that $\|\widetilde{O}_t\| = \|U_{\star,\perp}^\top X_t V_{t,\perp}\| \le \|X_t V_{t,\perp}\|$ for (75b).[2] □

Step 3: completing the proof. Now, with the help of Lemma 20, we are ready to prove Lemma 18. We start with verifying the two assumptions in Lemma 20.

Verifying assumption (73). By the RIP in (9), Lemma 7, and the condition of δ in (10), we have

$$\big\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(M_\star)\big\| \le \sqrt{r_\star}\delta\|M_\star\| \le c_\delta\kappa^{-(C_\delta-2)}\sigma_{\min}^2(X_\star) =: \gamma\sigma_{\min}^2(X_\star). \tag{76}$$

Here $\gamma = c_\delta\kappa^{-(C_\delta-2)} \le c_{20}$, as c_δ is assumed to be sufficiently small.

Verifying assumption (74). By Weyl's inequality and (76), we have

$$\left|s_j - 1 - \frac{\eta}{\lambda}\sigma_j(M_\star)\right| \le \frac{\eta}{\lambda}\big\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(M_\star)\big\| \le \frac{\eta}{\lambda}c_\delta\kappa^{-(C_\delta-2)}\sigma_{\min}^2(X_\star) \le \frac{100c_\delta}{c_\lambda}\eta,$$

where the last inequality follows from the condition $\lambda \ge \frac{1}{100}c_\lambda\sigma_{\min}^2(X_\star)$. Furthermore, using the condition $\lambda \le c_\lambda\sigma_{\min}^2(X_\star)$ assumed in (12b), the above bound implies that, for some $C = C(c_\lambda, c_\delta) > 0$,

$$s_1 \le 1 + \frac{\eta}{\lambda}\|M_\star\| + \frac{100c_\delta}{c_\lambda}\eta \le 1 + C\eta\kappa^2, \tag{77a}$$

$$s_{r_\star} \ge 1 + \frac{\eta}{\lambda}\sigma_{\min}^2(X_\star) - \frac{100c_\delta}{c_\lambda}\eta \ge 1 + \frac{\eta}{2c_\lambda}, \tag{77b}$$

$$s_{r_\star} \le 1 + \frac{\eta}{\lambda}\sigma_{\min}^2(X_\star) + \frac{100c_\delta}{c_\lambda}\eta \le 1 + \frac{2\eta}{c_\lambda}, \tag{77c}$$

$$s_{r_\star+1} \le 1 + \frac{100c_\delta}{c_\lambda}\eta \le 1 + \frac{\eta}{4c_\lambda}, \tag{77d}$$

where we use the fact that $\sigma_{r_\star+1}(M_\star) = 0$, and $c_\delta \le 1/400$. Consequently we have $s_{r_\star}/s_{r_\star+1} \ge 1 + c'\eta$ for some $c' = c'(c_\lambda) > 0$, assuming $c_\eta \le c_\lambda$. Thus for any large constant $L > 0$, there is some constant $c'' = c''(c') > 0$ such that, setting $L' = c''L\log(L)$ we have

$$(s_{r_\star}/s_{r_\star+1})^t \ge (L\kappa n)^L, \quad \forall t \ge \frac{L'}{\eta}\log(\kappa n).$$

On the event $\mathcal{E}$ given in (13), we can choose L large enough so that $L \ge 2C_G$, hence $\|G\| \le L$ and $\sigma_{\min}^{-1}(\widehat{U}^\top G) \le (2n)^{L/2}$. Summarizing these inequalities, we see for $t \ge \frac{L'}{\eta}\log(\kappa n)$,

$$\frac{\alpha\|G\| s_{r_\star+1}^t}{\alpha\sigma_{\min}(\widehat{U}^\top G) s_{r_\star}^t} \le L\sigma_{\min}^{-1}(\widehat{U}^\top G)(s_{r_\star+1}/s_{r_\star})^t$$

[2] The equation (31) in Stöger and Soltanolkotabi (2021, Lemma 8.5) is stated in a weaker form than what they actually proved, and our (75b) indeed follows from the penultimate inequality in the proof of Stöger and Soltanolkotabi (2021, Lemma 8.5).

$$\leq L(2n)^{L/2}(L\kappa n)^{-L} \leq (L\kappa n)^{-L/2}. \tag{78}$$

Furthermore, invoking Lemma 19 with $\theta = 1/(2L')$ (note that (71) is implied by the assumption (12c), where C_α is assumed sufficiently large, considering $\lambda \geq \frac{1}{100}c_\lambda \sigma^2_{\min}(X_\star)$ and $\|\mathcal{A}^*\mathcal{A}(M_\star)\| \leq \|M_\star\| + \gamma\sigma^2_{\min}(X_\star) \leq 2\|X_\star\|^2$ by (76)), we obtain for any $t \leq \frac{1}{\theta\eta}\log(\kappa n) = \frac{2L'}{\eta}\log(\kappa n)$ that $\|X_t - \widehat{X}_t\| \leq ts_1^t\alpha^2/\|X_\star\|$. This implies

$$\begin{aligned}
\frac{\|X_t - \widehat{X}_t\|}{\alpha\sigma_{\min}(\widehat{U}^\top G)s^t_{r_\star}} &\leq (s_1/s_{r_\star})^t \sigma^{-1}_{\min}(\widehat{U}^\top G)\alpha/\|X_\star\| \\
&\leq s_1^t \sigma^{-1}_{\min}(\widehat{U}^\top G)\alpha/\|X_\star\| \\
&\leq \exp(t\log(s_1) + L\log(L\kappa n))\alpha/\|X_\star\| \leq (L\kappa n)^{-L/2}
\end{aligned} \tag{79}$$

where the second inequality follows from (77b), the penultimate inequality follows from our choice of L which ensured $\sigma^{-1}_{\min}(\widehat{U}^\top G) \leq (2n)^{L/2}$, and the last inequality follows from (77a), our choice $t \leq \frac{2L'}{\eta}\log(\kappa n)$ and our assumption (12c) on α which implies $\alpha/\|X_\star\| \leq (2\kappa n)^{-C_\alpha}$, given that C_α is sufficiently large, e.g. $C_\alpha \geq C(L, c_\lambda, c_\eta)$. It may also be inferred from the above arguments that L can be made arbitrarily large by increasing C_α.

Combining the above arguments, we conclude that for any $t \in [(L'/\eta)\log(\kappa n), (2L'/\eta)\log(\kappa n)]$, both of (78), (79) hold, hence the condition in (74) can be verified by

$$\begin{aligned}
\phi = \frac{\alpha\|G\|s^t_{r_\star+1} + \|X_t - \widehat{X}_t\|}{\alpha\sigma_{\min}(\widehat{U}^\top G)s^t_{r_\star}} &\leq 2(L\kappa n)^{-L/2} \\
&\leq c_{20}\kappa^{-2},
\end{aligned} \tag{80}$$

by choosing L sufficiently large.

This completes the verification of both assumptions of Lemma 20. Upon noting that the upper threshold of t satisfies $(2L'/\eta)\log(\kappa n) \leq T_{\min}/16$, we will now invoke the conclusions of Lemma 20 to prove Lemma 18 for some $t \in [(L'/\eta)\log(\kappa n), T_{\min}/16]$.

Proof of bound (16). This can be inferred from (75a) in the following way. Recalling that $\sigma_{\min}(\widehat{U}^\top G) \geq (2n)^{-C_G}$ on the event $\mathcal{E}$, and $s_{r_\star} \geq 1$ by (77b), we obtain from (75a) that

$$\sigma_{\min}(\widetilde{S}_{t_1}) \geq \frac{1}{4}\alpha(2n)^{-C_G} \geq \alpha^2/\|X_\star\|,$$

given the condition (12c) which guarantees

$$\frac{\alpha}{\|X_\star\|} \leq (2n)^{-C_\alpha/\eta} \leq \frac{1}{4}(2n)^{-C_G},$$

as long as $\eta \leq c_\eta \leq 1$ and $C_\alpha \geq C_G + 2$. The proof is complete.

Proof of bound (17a). We combine (75a), (75b), and (80) to obtain

$$\frac{\|\widetilde{O}_{t_1}\|}{\sigma_{\min}(\widetilde{S}_{t_1})} \leq 4C_{20}\phi \leq 4C_{20}(L\kappa n)^{-L/2} \leq (L\kappa n/2)^{-L/2},$$

where the last inequality follows from taking L sufficiently large. We further note that (12b) implies

$$\begin{aligned}
\sigma_{\min}(\widetilde{S}_{t_1}) \leq \|\Sigma^2_\star + \lambda I\|^{1/2}\sigma_{\min}\left((\Sigma^2_\star + \lambda I)^{-1/2}\widetilde{S}_{t_1}\right) &\leq (c_\lambda + 1)^{1/2}\|X_\star\|\sigma_{\min}\left((\Sigma^2_\star + \lambda I)^{-1/2}\widetilde{S}_{t_1}\right) \\
&\leq 2\|X_\star\|\sigma_{\min}\left((\Sigma^2_\star + \lambda I)^{-1/2}\widetilde{S}_{t_1}\right),
\end{aligned}$$

assuming $c_\lambda \leq 1$, hence

$$\frac{\|\widetilde{O}_{t_1}\|}{\sigma_{\min}\left((\Sigma^2_\star + \lambda I)^{-1/2}\widetilde{S}_{t_1}\right)} \leq 2\|X_\star\|(L\kappa n/2)^{-L/2} \leq (C_{2.b}\kappa n)^{-C_{2.b}}\|X_\star\|,$$

as desired, with $C_{2.b} = L/4$ as long as L is sufficiently large. It is also clear that $C_{2.b}$ can be made arbitrarily large by enlarging C_α as L can be.

Proof of bound (17b). We apply (75b) to yield

$$\|\widetilde{O}_{t_1}\| \le C_{20}\phi\alpha\sigma_{\min}(\widehat{U}^\top G)s_{r_\star}^{t_1} \le C_G C_{20}(L\kappa n)^{-L/2}\left(1+\frac{2\eta}{c_\lambda}\right)^{t_1}\alpha \le \alpha^{5/6}\|X_\star\|^{1/6},$$

where the second inequality follows from $\sigma_{\min}(\widehat{U}^\top G) \le \|G\| \le C_G$ by assumption and from (77c); the last inequality follows from $t_1 \le (2L'/\eta)\log(\kappa n)$ and from the condition (12c) on α, provided that C_α is sufficiently large.

Proof of bound (17c). We apply (75c) to yield that

$$\|U_{\star,\perp}^\top U_{\widetilde{X}_{t+1}}\| \le C_{20}(\gamma+\phi) \le \frac{c_\delta}{c_\lambda^3}\kappa^{-2C_\delta/3},$$

using the bounds of γ and ϕ in (76) and (80), provided that $c_\lambda^3 \le \frac{1}{2}\min(1, C_{20})$ and $L \ge 2(C_\delta+1)$. To further bound $\|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star\|$ we need the following lemma.

Lemma 21. *Assume $\widetilde{S}_t$ is invertible, and at least one of the following is true: (i) $\|U_{\star,\perp}^\top U_{\widetilde{X}_t}\| \le 1/4$; (ii) $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \le \kappa^{-1}\|X_\star\|/4$. Then*

$$\kappa^{-1}\|X_\star\|\|U_{\star,\perp}^\top U_{\widetilde{X}_t}\| \le \|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \le 2\|X_\star\|\|U_{\star,\perp}^\top U_{\widetilde{X}_t}\|.$$

The proof is postponed to Section C.1.2. Returning to the proof of bound (17c), the above lemma yields

$$\|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star\| \le \frac{2c_\delta}{c_\lambda^3}\|X_\star\|\kappa^{-2C_\delta/3} \le c_2\|X_\star\|\kappa^{-2C_\delta/3},$$

for some $c_2 \lesssim c_\delta/c_\lambda^3$, as desired.

Proof of bound (17d). We have

$$\|\widetilde{S}_{t_1}\| = \|U_\star^\top X_{t_1}V_{t_1}\| \le \|X_{t_1}\| \le \|X_\star\|,$$

where the last step follows from Lemma 19.

C.1.1 Proof of Lemma 19

We prove the claim (72) by induction and also show that $\|X_t\| \le \|X_\star\|$ follows from (72). For the base case $t=0$, it holds by definition. Assume that (72) holds for some $t \le \frac{1}{\theta\eta}\log(\kappa n) - 1$. We aim to prove that (i) $\|X_t\| \le \|X_\star\|$ and that (ii) the inequality (72) continues to hold for $t+1$.

Proof of $\|X_t\| \le \|X_\star\|$. By the induction hypothesis we know

$$\|X_t - \widehat{X}_t\| \le t\left(1+\frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\right)^t \frac{\alpha^2}{\|X_\star\|}.$$

In view of the constraint (71) on α and the restriction $t \le \frac{1}{\theta\eta}\log(\kappa n)$, we have

$$t\frac{\alpha}{\|X_\star\|} \le \frac{1}{\theta\eta}\log(\kappa n)\cdot\frac{\eta}{K}\frac{1}{\log(\kappa n)} = \frac{1}{K\theta} \le 1$$

as long as $K = K(\theta, c_\lambda, C_G)$ is sufficiently large. This further implies

$$\|X_t - \widehat{X}_t\| \le \left(t\frac{\alpha}{\|X_\star\|}\right)\left(1+\frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\right)^t\alpha \le \left(1+\frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\right)^t\alpha.$$

On the other hand, since $\|X_0\| \le C_G\alpha$ under the event $\mathcal{E}$ (cf. (13)), in view of (70), we have

$$\|\widehat{X}_t\| \le \Big(1 + \frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big)^t \|X_0\| \le C_G\Big(1 + \frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big)^t \alpha.$$

Thus for a large enough $K = K(\theta, c_\lambda, C_G)$, we have

$$\|X_t\| \le \|X_t - \widehat{X}_t\| + \|\widehat{X}_t\| \le \Big(1 + \frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big)^t (C_G + 1)\alpha \le \sqrt{c_\lambda/200} \cdot \kappa^{-1}\|X_\star\|, \tag{81}$$

where the last inequality follows from the condition on t and the choice of α in (71):

$$\log\frac{\|X_\star\|}{\alpha} \ge \log\frac{\sqrt{200}(C_G+1)\kappa}{\sqrt{c_\lambda}} + t\log\Big(1 + \frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big).$$

The inequality (81) clearly implies $\|X_t\| \le \|X_\star\|$.

Proof of (72) at the induction step. The proof builds on a key recursive relation on $\big\|X_{t+1} - \widehat{X}_{t+1}\big\|$, from which the induction follows readily from our assumption.

Step 1: building a recursive relation on $\big\|X_{t+1} - \widehat{X}_{t+1}\big\|$. By definition (70), we have $\widehat{X}_{t+1} = \big(I + \frac{\eta}{\lambda}\mathcal{A}^*\mathcal{A}(M_\star)\big)\widehat{X}_t$, which implies the following decomposition:

$$X_{t+1} - \widehat{X}_{t+1} = \underbrace{\Big[X_{t+1} - \Big(I + \frac{\eta}{\lambda}\mathcal{A}^*\mathcal{A}(M_\star)\Big)X_t\Big]}_{=:T_1} + \underbrace{\Big(I + \frac{\eta}{\lambda}\mathcal{A}^*\mathcal{A}(M_\star)\Big)(X_t - \widehat{X}_t)}_{=:T_2}. \tag{82}$$

We shall control each term separately.

- The second term T_2 can be trivially bounded as

$$\|T_2\| = \Big\|\Big(I + \frac{\eta}{\lambda}\mathcal{A}^*\mathcal{A}(M_\star)\Big)(X_t - \widehat{X}_t)\Big\| \le \Big(1 + \frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big)\big\|X_t - \widehat{X}_t\big\|. \tag{83}$$

- Turning to the first term T_1, by the update rule (7) of X_{t+1} and the triangle inequality, we further have

$$\begin{aligned}\|T_1\| = \Big\|X_{t+1} - \Big(I + \frac{\eta}{\lambda}\mathcal{A}^*\mathcal{A}(M_\star)\Big)X_t\Big\| &\le \big\|\eta\mathcal{A}^*\mathcal{A}(X_tX_t^\top)X_t(X_t^\top X_t + \lambda I)^{-1}\big\| \\ &\quad + \big\|\eta\mathcal{A}^*\mathcal{A}(M_\star)X_t\big((X_t^\top X_t + \lambda I)^{-1} - \lambda^{-1}I\big)\big\|. \end{aligned}\tag{84}$$

  Since $\|(X_t^\top X_t + \lambda I)^{-1}\| \le \lambda^{-1}$, it follows that the first term in (84) can be bounded by

$$\big\|\eta\mathcal{A}^*\mathcal{A}(X_tX_t^\top)X_t(X_t^\top X_t + \lambda I)^{-1}\big\| \le \frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(X_t^\top X_t)\|\|X_t\|.$$

  In addition, since $\sqrt{c_\lambda/200}\cdot\kappa^{-1}\|X_\star\| = \sqrt{c_\lambda\sigma_{\min}^2(X_\star)/200} \le \sqrt{\lambda/2}$ by the condition $\lambda \ge \frac{1}{100}c_\lambda\sigma_{\min}^2(X_\star)$, we have by (81) that $\|X_t\| \le \sqrt{\lambda/2}$. Therefore, invoking Lemma 8 implies that

$$(X_t^\top X_t + \lambda I)^{-1} - \lambda^{-1}I = \lambda^{-2}X_t^\top X_t Q, \quad \text{for some } Q \text{ with } \|Q\| \le 2.$$

  As a result, the second term in (84) can be bounded by

$$\big\|\eta\mathcal{A}^*\mathcal{A}(M_\star)X_t\big((X_t^\top X_t + \lambda I)^{-1} - \lambda^{-1}I\big)\big\| \le 2\frac{\eta}{\lambda^2}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\|X_t\|^3.$$

  Combining the above two inequalities leads to

$$\|T_1\| \le \frac{\eta}{\lambda}\left(\|\mathcal{A}^*\mathcal{A}(X_t^\top X_t)\| + \frac{2}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\|X_t\|^2\right)\|X_t\|.$$

In view of Lemma 7, we know $\|\mathcal{A}^*\mathcal{A}(M_\star)\| \lesssim r_\star\|M_\star\|$ and $\|\mathcal{A}^*\mathcal{A}(X_tX_t^\top)\| \lesssim r\|X_t\|^2$. Plugging these relations into the previous bound leads to

$$\|T_1\| \lesssim \frac{\eta r}{\lambda}\Big(1+\frac{\|M_\star\|}{\lambda}\Big)\|X_t\|^3 \lesssim \frac{\eta\kappa^2 r}{\|M_\star\|}\kappa^2\|X_t\|^3, \tag{85}$$

where the last inequality follows from $\lambda \gtrsim \sigma_{\min}^2(X_\star) = \kappa^{-2}\|M_\star\|$ (cf. (12b)).

Putting the bounds on T_1 and T_2 together leads to

$$\big\|X_{t+1} - \widehat{X}_{t+1}\big\| \le \Big(1+\frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big)\big\|X_t - \widehat{X}_t\big\| + \frac{C\eta\kappa^4 r}{\|M_\star\|}\|X_t\|^3 \tag{86}$$

for some universal constant $C = C(c_\lambda) > 0$.

Step 2: finishing the induction. By the bound of $\|X_t\|$ in (81), it suffices to prove

$$\begin{aligned}
&t\Big(1+\frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big)^{t+1}\frac{\alpha^2}{\|X_\star\|} + \frac{C(C_G+1)^3\eta\kappa^4 r}{\|X_\star\|^2}\Big(1+\frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big)^{3t}\alpha^3 \\
&\le (t+1)\Big(1+\frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\Big)^{t+1}\frac{\alpha^2}{\|X_\star\|}.
\end{aligned}$$

This is equivalent to

$$C(C_G+1)^3\eta\kappa^4 r\left(1+\frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\right)^{2t-1} \le \frac{\|X_\star\|}{\alpha},$$

which again follows readily from our assumption $t \le \frac{1}{\theta\eta}\log(\kappa n)$ and the assumption (71) on α which implies

$$\begin{aligned}
\log\left(\frac{\|X_\star\|}{\alpha}\right) &\ge (2t-1)\log\left(1+\frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\right) + 4\log\kappa + \log n + K \\
&\ge (2t-1)\log\left(1+\frac{\eta}{\lambda}\|\mathcal{A}^*\mathcal{A}(M_\star)\|\right) + \log(\eta\kappa^4 r) + \log(C(C_G+1)^3)
\end{aligned}$$

provided $K = K(\theta, c_\lambda, C_G)$ is sufficiently large. The proof is complete.

C.1.2 Proof of Lemma 21

We begin with the following observation:

$$\begin{aligned}
\widetilde{N}_t\widetilde{S}_t^{-1} &= U_{\star,\perp}^\top U_{\widetilde{X}_t}\Sigma_{\widetilde{X}_t}V_{\widetilde{X}_t}^\top V_{\widetilde{X}_t}\Sigma_{\widetilde{X}_t}^{-1}(U_\star^\top U_{\widetilde{X}_t})^{-1} \\
&= U_{\star,\perp}^\top U_{\widetilde{X}_t}(U_\star^\top U_{\widetilde{X}_t})^{-1}
\end{aligned} \tag{87}$$

where we use: (i) $\widetilde{N}_t = U_{\star,\perp}^\top(U_{\widetilde{X}_t}\Sigma_{\widetilde{X}_t}V_{\widetilde{X}_t}^\top)$ and $\widetilde{S}_t = U_\star^\top U_{\widetilde{X}_t}\Sigma_{\widetilde{X}_t}V_{\widetilde{X}_t}^\top$; (ii) $\widetilde{X}_t$ is invertible since $\widetilde{S}_t$ is invertible, and hence $V_{\widetilde{X}_t}$ has rank $r_\star$ and $\Sigma_{\widetilde{X}_t}, U_\star^\top U_{\widetilde{X}_t}$ are also invertible. We will show that the above quantity is small if (and only if) $U_{\star,\perp}^\top U_{\widetilde{X}_t}$ is small.

Turning to the proof, we first show that (ii) implies (i), thus it suffices to prove the lemma under the condition (i). In fact, in virtue of (87) we have

$$\|U_{\star,\perp}^\top U_{\widetilde{X}_t}\| \le \|\widetilde{N}_t\widetilde{S}_t^{-1}\|\|U_\star^\top U_{\widetilde{X}_t}\| \le \|\widetilde{N}_t\widetilde{S}_t^{-1}\| \le \sigma_{\min}(X_\star)^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|,$$

where we used $\|U_\star^\top U_{\widetilde{X}_t}\| \le \|U_\star\|\|U_{\widetilde{X}_t}\| \le 1$. Consequently, $\|U_{\star,\perp}^\top U_{\widetilde{X}_t}\| \le 1/4$ if $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \le \kappa^{-1}\|X_\star\|/4$, as claimed.

We proceed to show that the conclusion holds assuming condition (i). The first inequality has already been established above. For the second inequality, using (87) again, it suffices to prove $\|(U_\star^\top U_{\widetilde{X}_t})^{-1}\| \le 2$, which is in turn equivalent to $\sigma_{\min}(U_\star^\top U_{\widetilde{X}_t}) \ge 1/2$. Now note that $U_{\widetilde{X}_t} = U_\star U_\star^\top U_{\widetilde{X}_t} + U_{\star,\perp}U_{\star,\perp}^\top U_{\widetilde{X}_t}$, thus

$$\sigma_{\min}(U_\star^\top U_{\widetilde{X}_t}) = \sigma_{r_\star}(U_\star^\top U_{\widetilde{X}_t})$$

$$
\begin{aligned}
&\geq \sigma_{r_\star}(U_\star U_\star^\top U_{\widetilde{X}_t}) \\
&\geq \sigma_{r_\star}(U_{\widetilde{X}_t}) - \|U_{\star,\perp} U_{\star,\perp}^\top U_{\widetilde{X}_t}\| \\
&\geq 1 - \|U_{\star,\perp}^\top U_{\widetilde{X}_t}\| \geq 3/4.
\end{aligned}
$$

In the last line, we used $\sigma_{r_\star}(U_{\widetilde{X}_t}) = 1$, which follows from $U_{\widetilde{X}_t}$ being a $n \times r_\star$ orthonormal matrix, and the assumption (i). This completes the proof.

C.2 Establishing the induction step

The claimed invertibility of $\widetilde{S}_t$ follows from induction and from Lemma 3. In fact, by (16) we know $\widetilde{S}_{t_1}$ is invertible, and by Lemma 3 we know that if $\widetilde{S}_t$ is invertible, $\widetilde{S}_{t+1}$ would also be invertible since $\widetilde{S}_t$ (resp. $\widetilde{S}_{t+1}$) has the same invertibility as $(\Sigma_\star^2 + \lambda I)^{-1}\widetilde{S}_t$ (resp. $(\Sigma_\star^2 + \lambda I)^{-1}\widetilde{S}_{t+1}$). For the rest of the proof we focus on establishing (17) by induction.

For the induction step we need to understand the one-step behaviors of $\|\widetilde{O}_t\|$, $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|$, and $\|\widetilde{S}_t\|$, which are supplied by the following lemmas.

Lemma 22. *For any t such that (17) holds,*

$$
\|\widetilde{O}_{t+1}\| \leq \left(1 + \frac{1}{12 C_{\max}\kappa}\eta\right)\|\widetilde{O}_t\|. \tag{88}
$$

Lemma 23. *For any t such that (17) holds, setting $Z_t = \Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)\Sigma_\star^{-1}$, there exists some universal constant $C_{23} > 0$ such that*

$$
\|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star\| \leq \left(1 - \frac{\eta}{3(\|Z_t\| + \eta)}\right)\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| + \eta\frac{C_{23}\kappa^2}{c_\lambda\|X_\star\|}\|U_\star^\top\Delta_t\| + \eta\left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{1/2}\|X_\star\|. \tag{89}
$$

In particular, if $c_2 = 100 C_{23}(C_{2.a}+1)^4 c_\delta / c_\lambda$, then $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \leq c_2\kappa^{-C_\delta/2}\|X_\star\|$ implies $\|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star\| \leq c_2\kappa^{-C_\delta/2}\|X_\star\|$.

Lemma 24. *For any t such that (17) holds,*

$$
\|\widetilde{S}_{t+1}\| \leq \left(1 - \frac{\eta}{2}\right)\|\widetilde{S}_t\| + 100 c_\lambda^{-1/2}\eta\kappa\|X_\star\|. \tag{90}
$$

In particular, if $C_{2.a} = 200 c_\lambda^{-1/2}$, then $\|\widetilde{S}_t\| \leq C_{2.a}\kappa\|X_\star\|$ implies $\|\widetilde{S}_{t+1}\| \leq C_{2.a}\kappa\|X_\star\|$.

We now return to the induction step. Recall that we need to show (17a)–(17d) hold for $t+1$. It is obvious that (17b)–(17d) hold for $t+1$ by the induction hypothesis and the above lemmas. It remains to prove (17a). To this end we distinguish two cases: $\sigma_{\min}((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t) \leq 1/3$ and $\sigma_{\min}((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t) > 1/3$. In the former case, (17a) for $t+1$ follows from Lemma 22 and Lemma 3 (to be proved in Appendix D.1), which imply (provided $C_{\max} \geq 2$)

$$
\frac{\|\widetilde{O}_{t+1}\|}{\sigma_{\min}((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_{t+1})} \leq \frac{\left(1 + \frac{\eta}{4 C_{\max}\kappa}\right)}{(1 + \eta/8)}\frac{\|\widetilde{O}_t\|}{\sigma_{\min}((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t)} \leq \frac{\|\widetilde{O}_t\|}{\sigma_{\min}((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t)},
$$

as desired. In the latter case where $\sigma_{\min}((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t) > 1/3$, one may apply the first part of Lemma 3 to deduce that $\sigma_{\min}((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_{t+1}) \geq 1/10$ (given that $\eta \leq c_\eta$ for some sufficiently small constant c_η). This combined with (17b) for $t+1$ (already proved) yields desired inequality (17a) for $t+1$, given our assumption (12c) on the smallness of α. This completes the proof.

C.2.1 Proof of Lemma 22

If $r = r_\star$, then we have $\|\widetilde{O}_t\| = 0$ for all $t \geq 0$. The conclusion follows trivially. Therefore, we only consider the case when $r > r_\star$. By definition, we have

$$\begin{aligned}\widetilde{O}_{t+1} = N_{t+1}V_{t+1,\perp} &= N_{t+1}V_tV_t^\top V_{t+1,\perp} + N_{t+1}V_{t,\perp}V_{t,\perp}^\top V_{t+1,\perp} \\ &= -N_{t+1}V_t(S_{t+1}V_t)^{-1}S_{t+1}V_{t,\perp}V_{t,\perp}^\top V_{t+1,\perp} + N_{t+1}V_{t,\perp}V_{t,\perp}^\top V_{t+1,\perp},\end{aligned}$$

where the last inequality uses the fact that $V_t^\top V_{t+1,\perp} = -(S_{t+1}V_t)^{-1}S_{t+1}V_{t,\perp}V_{t,\perp}^\top V_{t+1,\perp}$. To see this, note that

$$S_{t+1}V_{t+1,\perp} = 0 \qquad \Longrightarrow \qquad S_{t+1}V_tV_t^\top V_{t+1,\perp} = -S_{t+1}V_{t,\perp}V_{t,\perp}^\top V_{t+1,\perp}.$$

Left-multiplying both sides by $(S_{t+1}V_t)^{-1}$ yields the desired identity. Note that the invertibility of $S_{t+1}V_t$ follows from the invertibility of $\widetilde{S}_t$ by inserting $Q = 0$ in Lemma 13.

By Lemma 12, we immediately obtain that $S_{t+1}V_{t,\perp} = \eta E_t^b V_{t,\perp}$, and $N_{t+1}V_{t,\perp} = \widetilde{O}_t + \eta E_t^d V_{t,\perp}$, where $\|E_t^b\| \vee \|E_t^d\| \leq \frac{1}{24C_{\max}\kappa}\|\widetilde{O}_t\|$. Assume for now that

$$\|N_{t+1}V_t(S_{t+1}V_t)^{-1}\| \leq 1. \tag{91}$$

In addition, notice that $\|V_{t,\perp}^\top V_{t+1,\perp}\| \leq 1$ since both factors are orthonormal matrices, we have

$$\begin{aligned}\|\widetilde{O}_{t+1}\| &\leq \|\widetilde{O}_t\| + \eta\|N_{t+1}V_t(S_{t+1}V_t)^{-1}\|\|E_t^b\| + \eta\|E_t^d\| \\ &\leq \left(1 + \frac{1}{12C_{\max}\kappa}\eta\right)\|\widetilde{O}_t\|,\end{aligned}$$

as desired. It remains to prove (91).

Proof of bound (91). This can be done by plugging $Q = 0$ into Lemma 14 and bounding the resulting expression. This (in fact, a much stronger inequality) will be done in detail in the proof of Lemma 23, to be presented soon in Section C.2.2. In fact, the resulting expression is the same as (96) there (albeit with different values of $E_t^{13.a}$, $E_t^{14.a}$, $E_t^{14.b}$, which do not affect the proof). Following the same strategy to control (96) there, we may show that $\|N_{t+1}V_t(S_{t+1}V_t)^{-1}\Sigma_\star\|$ enjoys the same bound (101) as $\|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star\|$, the right hand side of which is less than $\kappa^{-1}\|X_\star\| = \|\Sigma_\star^{-1}\|^{-1}$ given (17c) and (17d). Thus $\|N_{t+1}V_t(S_{t+1}V_t)^{-1}\| \leq \|N_{t+1}V_t(S_{t+1}V_t)^{-1}\Sigma_\star\|\|\Sigma_\star^{-1}\| \leq 1$ as claimed.

C.2.2 Proof of Lemma 23

Denoting $\widetilde{X}_t \coloneqq X_tV_t$, we have $\widetilde{N}_t = U_{\star,\perp}^\top \widetilde{X}_t$ and $\widetilde{S}_t = U_\star^\top \widetilde{X}_t$. Suppose for the moment that

$$\|(V_t^\top V_{t+1})^{-1}\| \leq 2, \tag{92}$$

whose proof is deferred to the end of this section. We can write the update equation of $\widetilde{X}_t$ as

$$\begin{aligned}\widetilde{X}_{t+1} = X_{t+1}V_{t+1} &= X_{t+1}V_tV_t^\top V_{t+1} + X_{t+1}V_{t,\perp}V_{t,\perp}^\top V_{t+1} \\ &= \left(X_{t+1}V_t + X_{t+1}V_{t,\perp}V_{t,\perp}^\top V_{t+1}(V_t^\top V_{t+1})^{-1}\right)V_t^\top V_{t+1}.\end{aligned} \tag{93}$$

Left-multiplying both sides of (93) with $U_{\star,\perp}$ (or $U_\star$), we obtain

$$\widetilde{N}_{t+1} = (N_{t+1}V_t + N_{t+1}V_{t,\perp}Q)V_t^\top V_{t+1}, \tag{94a}$$
$$\widetilde{S}_{t+1} = (S_{t+1}V_t + S_{t+1}V_{t,\perp}Q)V_t^\top V_{t+1}, \tag{94b}$$

where we define $Q \coloneqq V_{t,\perp}^\top V_{t+1}(V_t^\top V_{t+1})^{-1}$. Consequently, we arrive at

$$\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1} = (N_{t+1}V_t + N_{t+1}V_{t,\perp}Q)(S_{t+1}V_t + S_{t+1}V_{t,\perp}Q)^{-1}. \tag{95}$$

Since $\|Q\| \le 2$ (which is an immediate implication of (92)), we can invoke Lemma 14 to obtain

$$
\begin{aligned}
\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star =& \widetilde{N}_t\widetilde{S}_t^{-1}(I+\eta E_t^{14.a})A_t(A_t+\eta\Sigma_\star^2)^{-1}(I+\eta E_t^{13})^{-1}\Sigma_\star + \eta E_t^{14.b}\Sigma_\star \\
=& \widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star(I+\eta\Sigma_\star^{-1}E_t^{14.a}\Sigma_\star)H_t(H_t+\eta I)^{-1}(I+\eta\Sigma_\star^{-1}E_t^{13}\Sigma_\star)^{-1} + \eta E_t^{14.b}\Sigma_\star, \qquad (96)
\end{aligned}
$$

where for simplicity of notation, we denote

$$
A_t \coloneqq (1-\eta)\widetilde{S}_t\widetilde{S}_t^\top + \lambda I, \quad \text{and} \quad H_t \coloneqq \Sigma_\star^{-1}A_t\Sigma_\star^{-1}.
$$

In addition, we have

$$
\begin{aligned}
&\|E_t^{13}\| + \|E_t^{14.a}\| \le \frac{1}{64\kappa^5}, \\
&\qquad \|\!|E_t^{14.b}|\!\| \le 800c_\lambda^{-1}\kappa^2\|X_\star\|^{-2}\|\!|U_\star^\top\Delta_t|\!\| + \frac{1}{64(C_{2.a}+1)^2\kappa^5\|X_\star\|}\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| + \frac{1}{64}\left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{2/3}.
\end{aligned}
$$

Moreover, it is clear that $\eta \le c_\eta \le 1 \le \kappa^4$ since $\kappa \ge 1$, and that $\|H_t\| \le \kappa^2(1+\|\widetilde{S}_t\|^2/\|X_\star\|^2) \le (C_{2.a}+1)^2\kappa^4$. Hence we have

$$
\|H_t\| + \eta \le 2(C_{2.a}+1)^2\kappa^4
$$

which implies

$$
\|E_t^{13}\| + \|E_t^{14.a}\| \le \frac{1}{24\kappa}\frac{1}{\|H_t\|+\eta}. \qquad (97)
$$

Similarly we may also show

$$
\|\!|E_t^{14.b}|\!\| \le 800c_\lambda^{-1}\kappa^2\|X_\star\|^{-2}\|\!|U_\star^\top\Delta_t|\!\| + \frac{1}{12(\|H_t\|+\eta)\|X_\star\|}\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| + \frac{1}{2}\left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{2/3}. \qquad (98)
$$

Since H_t is obviously positive definite, we have

$$
\|H_t(H_t+\eta I)^{-1}\| \le 1 - \frac{\eta}{\|H_t\|+\eta}. \qquad (99)
$$

Thus

$$
\begin{aligned}
\|\!|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star|\!\| &\le \left(1-\frac{\eta}{\|H_t\|+\eta}\right)(1-\eta\kappa\|E_t^{13}\|)^{-1}(1+\eta\kappa\|E_t^{14.a}\|)\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| + \eta\|\!|E_t^{14.b}|\!\|\|X_\star\|. \\
&\le \left(1-\frac{\eta}{\|H_t\|+\eta}\right)\left(1+\frac{1}{12}\frac{\eta}{\|H_t\|+\eta}\right)^2\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| \\
&\quad + \eta\frac{800\kappa^2}{c_\lambda\|X_\star\|}\|\!|U_\star^\top\Delta_t|\!\| + \frac{1}{12}\frac{\eta}{\|H_t\|+\eta}\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| + \frac{1}{2}\eta\left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{2/3}\|X_\star\| \\
&\le \left(1-\frac{5}{6}\frac{\eta}{\|H_t\|+\eta}\right)\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| + \frac{1}{12}\frac{\eta}{\|H_t\|+\eta}\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| \\
&\quad + \eta\frac{800\kappa^2}{c_\lambda\|X_\star\|}\|\!|U_\star^\top\Delta_t|\!\| + \frac{1}{2}\eta\left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{2/3}\|X_\star\| \\
&\le \left(1-\frac{3}{4}\frac{\eta}{\|H_t\|+\eta}\right)\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| + \eta\frac{800\kappa^2}{c_\lambda\|X_\star\|}\|\!|U_\star^\top\Delta_t|\!\| + \frac{1}{2}\eta\left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{2/3}\|X_\star\| \\
&\le \left(1-\frac{3}{4}\frac{\eta}{\|Z_t\|+\eta}\right)\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| + \eta\frac{800\kappa^2}{c_\lambda\|X_\star\|}\|\!|U_\star^\top\Delta_t|\!\| + \frac{1}{2}\eta\left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{2/3}\|X_\star\|, \quad (100)
\end{aligned}
$$

where in the second inequality we used $(1-x)^{-1} \le 1+x$ for $x<1$, in the penultimate inequality we used the elementary fact $(1-x)(1+\frac{1}{16}x)^2 \le 1-\frac{5}{6}x$ for $x\in[0,1]$, and in the last inequality we used the obvious fact

$$\|H_t\| = \|\Sigma_\star^{-1}((1-\eta)\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)\Sigma_\star^{-1}\| \le \|\Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)\Sigma_\star^{-1}\| = \|Z_t\|.$$

The desired inequality (89) follows from the above inequality by setting $C_{23} = 800$.

For the remaining claim, we need to apply the conclusion of the first part with $|||\cdot||| = \|\cdot\|$. Then we note the following bounds:

(i) $\|Z_t\| \le \|\Sigma_\star^{-1}\|^2(\|\widetilde{S}_t\|^2+\lambda) \le (C_{2.a}+1)^2\kappa^4$ by (17d) and (12b) (since we may choose $c_\lambda \le 1$);

(ii) $\eta \le c_\eta \le (C_{2.a}+1)^2\kappa^4$;

(iii) $\|U_\star^\top \Delta_t\| \le \|\Delta_t\| \le 16(C_{2.a}+1)^2 c_\delta \kappa^{-2C_\delta/3}\|X_\star\|^2$ by Lemma 11;

(iv) $(\|\widetilde{O}_t\|/\sigma_{\min}(\widetilde{S}_t))^{1/2} \le c_\delta\kappa^{-2C_\delta/3}$ by (17a), if we choose $C_\alpha \ge 3c_\delta^{-1} + 3C_\delta + 3$.

These together imply

$$\|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star\| \le \left(1 - \frac{\eta}{6(C_{2.a}+1)^2\kappa^4}\right)\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| + \eta\frac{16C_{23}\kappa^2}{c_\lambda}(C_{2.a}+1)^2 c_\delta\kappa^{-2C_\delta/3}\|X_\star\| + \eta c_\delta\kappa^{-2C_\delta/3}\|X_\star\|. \tag{101}$$

The conclusion follows easily by plugging in $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \le c_2\kappa^{-C_\delta/2}\|X_\star\|$ and using $\kappa^6\kappa^{-2C_\delta/3} \le \kappa^{-C_\delta/2}$ when C_δ is sufficiently large.

Proof of bound (92). First, we observe that it is equivalent to show that $\sigma_{\min}(V_t^\top V_{t+1}) \ge 1/2$. But from $V_{t+1}V_{t+1}^\top + V_{t+1,\perp}V_{t+1,\perp}^\top = I$ we have

$$\begin{aligned}\sigma_{\min}(V_t^\top V_{t+1}) = \sigma_{r_\star}(V_t^\top V_{t+1}) &\ge \sigma_{r_\star}(V_t^\top V_{t+1}V_{t+1}^\top) = \sigma_{r_\star}(V_t^\top - V_t^\top V_{t+1,\perp}V_{t+1,\perp}^\top)\\ &\ge \sigma_{r_\star}(V_t^\top) - \|V_t^\top V_{t+1,\perp}V_{t+1,\perp}^\top\|\\ &\ge 1 - \|V_t^\top V_{t+1,\perp}\|,\end{aligned}$$

where the last inequality follows from $\sigma_{r_\star}(V_t^\top) = 1$ (since $V_t \in \mathbb{R}^{r\times r_\star}$ is orthonormal) and from that $\|V_t^\top V_{t+1,\perp}V_{t+1,\perp}^\top\| \le \|V_t^\top V_{t+1,\perp}\|$. This implies that, to show $\sigma_{\min}(V_t^\top V_{t+1}) \ge 1/2$, it suffices to prove $\|V_t^\top V_{t+1,\perp}\| \le 1/2$.

Next we prove that $\|V_t^\top V_{t+1,\perp}\| \le 1/2$. Recall that by definition we have $S_{t+1}V_{t+1,\perp} = 0$. Right-multiplying both sides of (47a) by $V_{t+1,\perp}$, we obtain

$$0 = \Big((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I + E_t^a)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\Big)\widetilde{S}_t(V_t^\top V_{t+1,\perp}) + \eta E_t^b V_{t+1,\perp},$$

hence

$$\|V_t^\top V_{t+1,\perp}\| \le \eta\|E_t^b V_{t+1,\perp}\|\|\widetilde{S}_t^{-1}\|\left\|\Big((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I + E_t^a)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\Big)^{-1}\right\|.$$

By (48b) we have

$$\|E_t^b V_{t+1,\perp}\|\|\widetilde{S}_t^{-1}\| \le \frac{\|E_t^b\|}{\sigma_{\min}(\widetilde{S}_t)} \le \frac{1}{10\kappa},$$

thus it suffices to show

$$\eta\left\|\Big((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I + E_t^a)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\Big)^{-1}\right\| \le 5\kappa, \tag{102}$$

or equivalently,

$$\sigma_{\min}\Big((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I + E_t^a)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\Big) \ge \frac{\eta}{5\kappa}. \tag{103}$$

To this end, we write

$$(1-\eta)I + \eta(\Sigma_\star^2 + \lambda I + E_t^a)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}$$

$$= \left(I + \eta E_t^a \left((1-\eta)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I) + \eta(\Sigma_\star^2 + \lambda I)\right)^{-1}\right) \left((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right) \tag{104}$$

and control the two terms separately.

- To control the first factor, starting from (48a) we may deduce

$$\begin{aligned}\|E_t^a\| &\le \kappa^{-4}\|X_\star\|\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| + \|U_\star^\top \Delta_t\| \\ &\le \kappa^{-4}\|X_\star\|c_2\kappa^{-C_\delta/2}\|X_\star\| + c_{11}\kappa^{-2C_\delta/3}\|X_\star\|^2 \\ &\le \kappa^{-2}\|X_\star\|^2/2 = \sigma_{\min}^2(X_\star)/2,\end{aligned}$$

  where the second inequality follows from (17c) and Lemma 11; the last inequality follows from choosing c_δ sufficiently small (recall that $c_2, c_{11} \lesssim c_\delta/c_\lambda^3$) and C_δ sufficiently large. Furthermore, since $\widetilde{S}_t\widetilde{S}_t^\top$ is positive semidefinite, we have

$$\left\|\left((1-\eta)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I) + \eta(\Sigma_\star^2 + \lambda I)\right)^{-1}\right\| \le \eta^{-1}\sigma_{\min}^{-2}(\Sigma_\star) = \eta^{-1}\sigma_{\min}^{-2}(X_\star),$$

  hence

$$\begin{aligned}&\sigma_{\min}\left(1 + \eta E_t^a \left((1-\eta)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I) + \eta(\Sigma_\star^2 + \lambda I)\right)^{-1}\right) \\ &\ge 1 - \eta\|E_t^a\| \left\|\left((1-\eta)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I) + \eta(\Sigma_\star^2 + \lambda I)\right)^{-1}\right\| \\ &\ge 1 - \eta \cdot \frac{\sigma_{\min}^2(X_\star)}{2} \cdot \eta^{-1}\sigma_{\min}^{-2}(X_\star) = 1/2. \end{aligned} \tag{105}$$

- Now we control the second factor. By Lemma 9 we have

$$\begin{aligned}\sigma_{\min}\left(1 - \eta + \eta(\Sigma_\star^2 + \lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right) &= (1-\eta)\sigma_{\min}\left(I + \frac{\eta}{1-\eta}(\Sigma_\star^2 + \lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right) \\ &\ge (1-\eta)\left(\frac{\|\Sigma_\star^2 + \lambda I\|}{\sigma_{\min}(\Sigma_\star^2 + \lambda I)}\right)^{-1/2} \\ &= (1-\eta)\left(\frac{\|X_\star\|^2 + \lambda}{\sigma_{\min}^2(X_\star) + \lambda}\right)^{-1/2}.\end{aligned}$$

  It is easy to check that the function $\lambda \mapsto (a+\lambda)/(b+\lambda)$ is decreasing on $[0, \infty)$ for $a \ge b > 0$, thus

$$\frac{\|X_\star\|^2 + \lambda}{\sigma_{\min}^2(X_\star) + \lambda} \le \frac{\|X_\star\|^2}{\sigma_{\min}^2(X_\star)} = \kappa^2,$$

  which implies

$$\sigma_{\min}\left((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right) \ge \frac{1-\eta}{\kappa}. \tag{106}$$

Plugging (106) and (105) into (104) yields

$$\sigma_{\min}\left((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I + E_t^a)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right) \ge \frac{1-\eta}{2\kappa} \ge \frac{\eta}{5\kappa}, \tag{107}$$

where the last inequality follows from the assumption $\eta \le c_\eta$. This shows (103) as desired, thereby completing the proof.

C.2.3 Proof of Lemma 24

Combine (94b) and Lemma 13 to see that

$$
\begin{aligned}
\|\widetilde{S}_{t+1}\| &\le \|S_{t+1}V_t + S_{t+1}V_{t,\perp}Q\| \\
&\le \|1+\eta E_t^{13}\| \cdot \left\|(1-\eta)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{1/2} + \eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1/2}\right\| \cdot \left\|(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1/2}\widetilde{S}_t\right\| \\
&\le (1+\eta\|E_t^{13}\|)\left((1-\eta)(\|\widetilde{S}_t\|^2+\lambda)^{1/2} + 4\eta\lambda^{-1/2}\|X_\star\|^2\right)(\|\widetilde{S}_t\|^2+\lambda)^{-1/2}\|\widetilde{S}_t\| \\
&\le \left(1+\frac{\eta}{4}\right)\left((1-\eta)\|\widetilde{S}_t\| + 4\eta\frac{\|X_\star\|^2\|\widetilde{S}_t\|}{\sqrt{\lambda(\|\widetilde{S}_t\|^2+\lambda)}}\right) \\
&\le \left(1-\frac{\eta}{2}\right)\|\widetilde{S}_t\| + 5\eta\frac{\|X_\star\|^2}{\sqrt{\lambda}},
\end{aligned} \tag{108}
$$

where the third line follows from $\|\Sigma_\star^2+\lambda I\| \le (1+\lambda)\|X_\star\|^2 \le 2\|X_\star\|^2$ assuming $c_\lambda \le 1$ and from the fact that the singular values of $(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1/2}\widetilde{S}_t$ are $(\sigma_j^2(\widetilde{S}_t)+\lambda)^{-1/2}\sigma_j(\widetilde{S}_t)$, $j=1,\dots,r_\star$,[3] which is bounded by $(\|\widetilde{S}_t\|^2+\lambda)^{-1/2}\|\widetilde{S}_t\|$ since $\sigma \mapsto (\sigma^2+\lambda)^{-1/2}\sigma$ is increasing and since $\|\widetilde{S}_t\|$ is the largest singular value of $\widetilde{S}_t$. In the fourth line, we used the error bound $\|E_t^{13}\| \le 1/4$ and the last line follows from the elementary inequalities $1+\eta/4 \le (1-\eta/2)(1-\eta)^{-1} \le 5/4$ given that $\eta \le c_\eta$ for sufficiently small constant $c_\eta > 0$. The conclusion readily follows from the above inequality and the assumption $\lambda \ge \frac{1}{100}c_\lambda\sigma_{\min}^2(X_\star)$.

D Proofs for Phase II

This section collects the proofs for Phase II.

D.1 Proof of Lemma 3

Since $\|V_{t+1}^\top V_t\| \le 1$, we have

$$
\begin{aligned}
\sigma_{\min}((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_{t+1}) &\ge \sigma_{\min}((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_{t+1}V_{t+1}^\top V_t) \\
&= \sigma_{\min}((\Sigma_\star^2+\lambda I)^{-1/2}S_{t+1}V_t),
\end{aligned}
$$

where the second equality follows from $S_{t+1} = \widetilde{S}_{t+1}V_{t+1}^\top$ (cf. (26)). Apply Lemma 13 with $Q=0$ to see that

$$
S_{t+1}V_t = (I+\eta E_t^{13})\left((1-\eta)I + \eta(\Sigma_\star^2+\lambda I)(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1}\right)\widetilde{S}_t, \tag{109}
$$

where $E_t^{13} \in \mathbb{R}^{r_\star \times r_\star}$ satisfies $\|E_t^{13}\| \le \frac{1}{200(C_{2.a}+1)^4\kappa^5}$. To simplify the notation, we denote

$$
Y_t \coloneqq (\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t,
$$

which allows us to write (109) as

$$
\begin{aligned}
&(\Sigma_\star^2+\lambda I)^{-1/2}S_{t+1}V_t \\
&\quad= \left(I+\eta(\Sigma_\star^2+\lambda I)^{-1/2}E_t^{13}(\Sigma_\star^2+\lambda I)^{1/2}\right)\left((1-\eta)I + \eta\big(Y_tY_t^\top + \lambda(\Sigma_\star^2+\lambda I)^{-1}\big)^{-1}\right)Y_t.
\end{aligned} \tag{110}
$$

Note that

$$
\begin{aligned}
\|(\Sigma_\star^2+\lambda I)^{-1/2}E_t^{13}(\Sigma_\star^2+\lambda I)^{1/2}\| &\le \|(\Sigma_\star^2+\lambda I)^{-1/2}\| \cdot \|(\Sigma_\star^2+\lambda I)^{1/2}\| \cdot \|E_t^{13}\| \\
&\le \kappa\|X_\star\|^{-1} \cdot (2\|X_\star\|) \cdot \|E_t^{13}\|
\end{aligned}
$$

[3]This can be seen from plugging in $\widetilde{S}_t = U_t\Sigma_t$ by definition which implies $(\widetilde{S}_t\widetilde{S}_t^\top+\lambda I)^{-1/2}\widetilde{S}_t = U_t(\Sigma_t+\lambda I)^{-1/2}\Sigma_t$.

$$\leq 2\kappa \cdot \frac{1}{200(C_{2.a}+1)^4\kappa^5} \leq 1/32, \tag{111}$$

where in the second inequality we used $\lambda \leq c_\lambda \|M_\star\| \leq \|X_\star\|^2$ as $c_\lambda \leq 1$, and in the third inequality we used the claimed bound of $\|E_t^{13}\|$. Therefore, it follows that

$$\sigma_{\min}\left(I + \eta(\Sigma_\star^2 + \lambda I)^{-1/2} E_t^{13} (\Sigma_\star^2 + \lambda I)^{1/2}\right) \geq 1 - \eta/32. \tag{112}$$

On the other hand, using $\sigma_{\min}(AB) \geq \sigma_{\min}(A)\sigma_{\min}(B)$ for any matrices A, B, it is obvious that

$$\sigma_{\min}\Big(\big((1-\eta)I + \eta(Y_tY_t^\top + \lambda(\Sigma_\star^2+\lambda I)^{-1})^{-1}\big)\, Y_t\Big) \geq (1-\eta)\sigma_{\min}(Y_t),$$

which in turn implies that

$$\sigma_{\min}\big((\Sigma_\star^2+\lambda I)^{-1/2} S_{t+1} V_t\big) \geq (1-\eta/32)(1-\eta)\sigma_{\min}(Y_t) \geq (1-2\eta)\sigma_{\min}(Y_t),$$

as long as $\eta \leq c_\eta$ for some sufficiently small constant c_η. This proves the first part of Lemma 3.

Now we move to the second part assuming $\sigma_{\min}(Y_t) \leq 1/3$. Using the assumption $\lambda \leq c_\lambda \sigma_{\min}(M_\star)$, we see that

$$\|\lambda(\Sigma_\star^2+\lambda I)^{-1}\| \leq c_\lambda.$$

Given that c_λ is sufficiently small (such that $c_\lambda \leq c_{10}$, where c_{10} is the positive constant in Lemma 10), one may apply Lemma 10 with $Y = Y_t$ and $\Lambda = \lambda(\Sigma_\star^2+\lambda I)^{-1}$ to obtain

$$\begin{aligned}
\sigma_{\min}\big((\Sigma_\star^2+\lambda I)^{-1/2} S_{t+1} V_t\big) &\geq \sigma_{\min}\left(I + \eta(\Sigma_\star^2+\lambda I)^{-1/2} E_t^{13} (\Sigma_\star^2+\lambda I)^{1/2}\right)\left(1+\frac{1}{6}\eta\right)\sigma_{\min}(Y_t) \\
&\overset{\text{(i)}}{\geq} (1-\eta/32)\left(1+\frac{1}{6}\eta\right)\sigma_{\min}(Y_t) \overset{\text{(ii)}}{\geq} \left(1+\frac{1}{8}\eta\right)\sigma_{\min}(Y_t),
\end{aligned}$$

where (i) uses (112), and (ii) follows as long as $\eta \leq c_\eta$ for some sufficiently small constant c_η. The desired conclusion follows.

D.2 Proof of Corollary 1

We will prove a strengthened version of (20), that is

$$\sigma_{\min}\left((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t\right) \geq 1/\sqrt{10}. \tag{113}$$

It is clear that (113) implies (20). Indeed, for each $u \in \mathbb{R}^{r_\star}$, by taking $v = (\Sigma_\star^2+\lambda I)^{1/2}u$, we have

$$u^\top \widetilde{S}_t\widetilde{S}_t^\top u = v^\top(\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t\widetilde{S}_t^\top(\Sigma_\star^2+\lambda I)^{-1/2}v \geq \frac{1}{10}\|v\|^2 \geq \frac{1}{10}u^\top\Sigma_\star^2 u,$$

which implies (20). It then boils down to establish (113).

Step 1: establishing the claim for a midpoint t_2. From Lemma 2 we know that

$$\sigma_{\min}\left((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_{t_1}\right) \geq \|\Sigma_\star^2+\lambda I\|^{-1/2}\sigma_{\min}(\widetilde{S}_{t_1}) \overset{\text{(i)}}{\geq} (c_\lambda+1)^{-1/2}\|X_\star\|^{-1}\cdot\alpha^2/\|X_\star\| \geq \frac{1}{3}(\alpha/\|X_\star\|)^2,$$

where (i) follows from the assumption (12b) and Lemma 2, and the last inequality follows by choosing $c_\lambda \leq 1$. By the second part of Lemma 3, starting from t_1, whenever $\sigma_{\min}((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t) < 1/\sqrt{10} < 1/3$, it would increase exponentially with rate at least $(1+\frac{\eta}{8})$. On the other end, it is easy to verify, given that $\eta \leq c_\eta$ is sufficiently small,

$$\left(1+\frac{\eta}{8}\right)^{\frac{16}{\eta}\log\left(\frac{3}{\sqrt{10}}\frac{\|X_\star\|^2}{\alpha^2}\right)} \geq \frac{3\|X_\star\|^2}{\sqrt{10}\alpha^2} \geq \frac{1}{\sqrt{10}}\frac{1}{\sigma_{\min}\left((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_{t_1}\right)}.$$

Therefore, it takes at most $\frac{16}{\eta}\log\left(\frac{3}{\sqrt{10}}\frac{\|X_\star\|^2}{\alpha^2}\right) \leq T_{\min}/16$ more iterations to make $\sigma_{\min}((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_t)$ grow to at least $1/\sqrt{10}$. Equivalent, for some $t_2 : t_1 \leq t_2 \leq t_1 + T_{\min}/16$, we have

$$\sigma_{\min}\left((\Sigma_\star^2+\lambda I)^{-1/2}\widetilde{S}_{t_2}\right) \geq 1/\sqrt{10}.$$

Step 2: establishing the claim for all $t \in [t_2, T_{\max}]$. It remains to show that (113) continues to hold for all $t \in [t_2, T_{\max}]$. We prove this by induction on t. Assume that (113) holds for some $t \in [t_2, T_{\max} - 1]$. We show that it will also hold for $t+1$. We divide the proof into two cases.

Case 1. If $\sigma_{\min}((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t) \le 1/3$, we deduce from the second part of Lemma 3 that

$$\sigma_{\min}\left((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_{t+1}\right) \ge \left(1 + \frac{\eta}{8}\right)\sigma_{\min}\left((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t\right) \ge \sigma_{\min}\left((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t\right),$$

which by the induction hypothesis is no less than $1/\sqrt{10}$, as desired.

Case 2. If $\sigma_{\min}((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t) > 1/3$, the first part of Lemma 3 yields

$$\sigma_{\min}\left((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_{t+1}\right) \ge (1 - 2\eta)\sigma_{\min}\left((\Sigma_\star^2 + \lambda I)^{-1/2}\widetilde{S}_t\right) \ge (1 - 2\eta)/3,$$

which is greater than $1/\sqrt{10}$ provided $\eta \le c_\eta \le 1/100$, as desired.

Combining the two cases completes the proof.

D.3 Proof of Lemma 4

For simplicity, in this section we denote

$$\Gamma_t \coloneqq \Sigma_\star^{-1}\widetilde{S}_t\widetilde{S}_t^\top\Sigma_\star^{-1} - I = \Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)\Sigma_\star^{-1}. \tag{114}$$

It turns out that Lemma 4 follows naturally from the following technical lemma, whose proof is deferred to the end of this section.

Lemma 25. *For any $t : t_2 \le t \le T_{\max}$, one has*

$$\|\!|\Gamma_{t+1}|\!\| \le (1-\eta)\|\!|\Gamma_t|\!\| + \eta\frac{C_{25}\kappa^4}{\|X_\star\|^2}\|\!|U_\star^\top\Delta_t|\!\| + \frac{1}{16}\eta\|X_\star\|^{-1}\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| + \eta\left(\frac{\|\widetilde{O}_t\|}{\|X_\star\|}\right)^{7/12}, \tag{115}$$

where $C_{25} \lesssim c_\lambda^{-1/2}$ is some positive constant and $\|\!|\cdot|\!\|$ can either be the Frobenius norm or the spectral norm.

From Lemma 11, we know that $\|U_\star^\top\Delta_t\| \le \|\Delta_t\| \le \frac{\|X_\star\|^2}{300C_{25}\kappa^4}$ as c_δ is sufficiently small. Similarly, $\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\| \le \|X_\star\|/100$ and $(\|\widetilde{O}_t\|/\|X_\star\|)^{7/12} \le 1/300$ by Lemma 2. Applying Lemma 25 with the spectral norm, we prove Lemma 4 as desired.

Proof of Lemma 25. We start by rewriting (47a) as

$$\begin{aligned}
S_{t+1} &= \left((1-\eta)I + \eta(\Sigma_\star^2 + \lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right)\widetilde{S}_tV_t^\top + \eta E_t^g \\
&= \left(I - \eta(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1} + \eta(\Sigma_\star^2 + \lambda I)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right)\widetilde{S}_tV_t^\top + \eta E_t^g \\
&= \left(I - \eta(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\right)\widetilde{S}_tV_t^\top + \eta E_t^g,
\end{aligned} \tag{116}$$

where

$$E_t^g = E_t^a(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_tV_t^\top + E_t^b. \tag{117}$$

By Corollary 1, we have $\sigma_{\min}(\widetilde{S}_t)^2 \ge \frac{1}{100}\sigma_{\min}(M_\star)$ for $t \in [t_2, T_{\max}]$, so

$$\|(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\widetilde{S}_tV_t^\top\| \le \|(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1/2}\|\|(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1/2}\widetilde{S}_t\| \le \sigma_{\min}^{-1}(\widetilde{S}_t) \lesssim 1/\sigma_{\min}(X_\star).$$

Combined with the error bounds (48a), (48b), we have for some universal constant $C > 0$ that

$$\|\!|E_t^g|\!\| \le \|\!|E_t^a|\!\| + \eta\|\!|E_t^b|\!\| \le \frac{C\kappa}{\|X_\star\|}\|\!|U_\star^\top\Delta_t|\!\| + Cc_{12}\kappa^{-3}\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| + C\|\widetilde{O}_t\|^{3/4}\|X_\star\|^{1/4}. \tag{118}$$

Step 1: deriving a recursion of Γ_t. Define

$$A_t := \big(I - \eta(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)^{-1}\big)\widetilde{S}_t V_t^\top.$$

Then we can rewrite (116) as $A_t = S_{t+1} - \eta E_t^g$, and by rearranging $A_t A_t^\top = (S_{t+1} - \eta E_t^g)(S_{t+1} - \eta E_t^g)^\top$ in view of (26), it follows that

$$\begin{aligned}\widetilde{S}_{t+1}\widetilde{S}_{t+1}^\top = S_{t+1}S_{t+1}^\top &= A_t A_t^\top + \eta(\|S_{t+1}\| + \|E_t^g\|)(E_t^g Q_1 + Q_2 E_t^{g\top}) \\ &=: A_t A_t^\top + \eta E_t^f\end{aligned}$$

for some matrices Q_1, Q_2 with $\|Q_1\|, \|Q_2\| \le 1$. By mapping both sides of the above equation by $(\cdot) \mapsto \Sigma_\star^{-1}(\cdot)\Sigma_\star^{-1} - I$, we obtain

$$\Gamma_{t+1} = \big(I - \eta\Gamma_t(I + \Gamma_t + \lambda\Sigma_\star^{-2})^{-1}\big)(\Gamma_t + I)\big(I - \eta(I + \Gamma_t + \lambda\Sigma_\star^{-2})^{-1}\Gamma_t\big) - I + \eta\Sigma_\star^{-1}E_t^f\Sigma_\star^{-1}, \tag{119}$$

where we recall the definition of Γ_t in (114).

Step 2: simplify the recursion. Note that $\sigma_{\min}(\Sigma_\star^{-1}\widetilde{S}_t) \ge 1/10$ implies $I + \Gamma_t \succeq \frac{1}{100}I$. From our assumption $\lambda \le c_\lambda\sigma_{\min}(M_\star)$, it follows that $\|\lambda\Sigma_\star^{-2}\| \le c_\lambda \le 1/200 \le \frac{1}{2}\sigma_{\min}(I + \Gamma_t)$, thus in virtue of Lemma 8 we have

$$(I + \Gamma_t + \lambda\Sigma_\star^{-2})^{-1} = (I + \Gamma_t)^{-1} + (I + \Gamma_t)^{-1}(c_\lambda Q')(I + \Gamma_t)^{-1},$$

for some matrix Q' with $\|Q'\| \le 2$. Plugging this into (119) yields

$$\begin{aligned}\Gamma_{t+1} &= \big(I - \eta\Gamma_t(I + \Gamma_t)^{-1}\big)(\Gamma_t + I)\big(I - \eta(I + \Gamma_t)^{-1}\Gamma_t\big) + \eta E_t^h + \eta\Sigma_\star^{-1}E_t^f\Sigma_\star^{-1} \\ &= (1 - 2\eta)\Gamma_t + \eta^2\Gamma_t^2(1 + \Gamma_t)^{-1} + \eta E_t^h + \eta\Sigma_\star^{-1}E_t^f\Sigma_\star^{-1},\end{aligned} \tag{120}$$

where the additional error term E_t^h is defined by

$$\begin{aligned}E_t^h :=& \Gamma_t(I + \Gamma_t)^{-1}(c_\lambda Q')(1 - \eta\Gamma_t(I + \Gamma_t)^{-1}) + (1 - \eta\Gamma_t(I + \Gamma_t)^{-1})(c_\lambda Q')(I + \Gamma_t)^{-1}\Gamma_t \\ &+ \eta\Gamma_t(I + \Gamma_t)^{-1}(c_\lambda Q')(I + \Gamma_t)^{-2}(c_\lambda Q')(I + \Gamma_t)^{-1}\Gamma_t.\end{aligned} \tag{121}$$

Step 3: controlling the error terms. We now control the error terms in (120) separately.

- By (17d) we have $\|S_{t+1}\| \le C_{2.a}\kappa\|X_\star\|$, and by controlling the right hand side of (118) using (17c), (19), and (45) in Lemma 11, it is evident that $\|E_t^g\| \le \kappa\|X_\star\|$. Hence, the term E_t^f obeys

  $$\begin{aligned}\|\!|E_t^f|\!\| &\le (C_{2.a} + 1)\kappa\|X_\star\| \cdot \|\!|E_t^g|\!\| \\ &\le C'C_{2.a}\left(\kappa^2\|\!|U_\star^\top\Delta_t|\!\| + c_{12}\kappa^{-2}\|X_\star\|\|\!|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star|\!\| + \kappa\|\widetilde{O}_t\|^{3/4}\|X_\star\|^{5/4}\right),\end{aligned} \tag{122}$$

  where $C' > 0$ is again some universal constant.

- Since $\Gamma_t \succeq \frac{1}{100}I - I = -\frac{99}{100}I$ as already proved, it is easy to see that $\|(1 + \Gamma_t)^{-1}\| \le C$ and $\|\Gamma_t(1 + \Gamma_t)^{-1}\| \le C$ for some universal constant $C > 0$. Thus,

  $$\|\!|E_t^h|\!\| \le 2c_\lambda C(1 + \eta C)\|Q'\| \cdot \|\!|\Gamma_t|\!\| + \eta c_\lambda^2 C^4\|Q'\|^2\|\!|\Gamma_t|\!\| \le \frac{1}{2}\|\!|\Gamma_t|\!\|, \tag{123}$$

  where the last line follows by using $\|Q'\| \le 2$ and by choosing c_λ, c_η sufficiently small.

- We still need to control $\eta^2\Gamma_t^2(1 + \Gamma_t)^{-1}$. This can be accomplished by invoking $\|\Gamma_t(1 + \Gamma_t)^{-1}\| \le C$ again. In fact, we have

  $$\eta^2\|\!|\Gamma_t^2(1 + \Gamma_t)^{-1}|\!\| \le \eta \cdot \eta\|\Gamma_t(1 + \Gamma_t)^{-1}\| \cdot \|\!|\Gamma_t|\!\| \le \eta \cdot \eta C\|\!|\Gamma_t|\!\| \le \frac{\eta}{2}\|\!|\Gamma_t|\!\| \tag{124}$$

  provided that $\eta \le c_\eta$ is sufficiently small.

Plugging (122), (123), (124) into (120), we readily obtain

$$\begin{aligned}
\|\|\Gamma_{t+1}\|\| &\le (1-2\eta)\|\|\Gamma_t\|\| + \frac{\eta}{2}\|\|\Gamma_t\|\| + \frac{\eta}{2}\|\|\Gamma_t\|\| + \eta\kappa^2\|X_\star\|^{-2}\|\|E_t^f\|\| \\
&\le (1-\eta)\|\|\Gamma_t\|\| + \eta\frac{C'C_{2.a}\kappa^4}{\|X_\star\|^2}\|\|U_\star^\top\Delta_t\|\| + \eta c_{12}C'C_{2.a}\|X_\star\|^{-1}\|\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|\| + \eta C'C_{2.a}\kappa^3\|\widetilde{O}_t\|^{3/4}\|X_\star\|^{-3/4} \\
&\le (1-\eta)\|\|\Gamma_t\|\| + \eta\frac{C_{25}\kappa^4}{\|X_\star\|^2}\|\|U_\star^\top\Delta_t\|\| + \frac{1}{16}\eta\|X_\star\|^{-1}\|\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|\| + \eta\left(\frac{\|\widetilde{O}_t\|}{\|X_\star\|}\right)^{7/12},
\end{aligned}$$

where in the last line we set $C_{25} = C'C_{2.a}$, chose c_{12} sufficiently small and used (19). Finally note that $C_{25} \lesssim C_{2.a} \lesssim c_\lambda^{-1/2}$ as desired.

D.4 Proof of Corollary 2

From Lemma 4, it is elementary (e.g., by induction on t) to show that

$$\left\|\Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)\Sigma_\star^{-1}\right\| \le (1-\eta)^{t-t_2}\left\|\Sigma_\star^{-1}(\widetilde{S}_{t_2}\widetilde{S}_{t_2}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\right\| + \frac{1}{100}, \quad \forall t \in [t_2, T_{\max}]. \tag{125}$$

Suppose for the moment that

$$\left\|\Sigma_\star^{-1}(\widetilde{S}_{t_2}\widetilde{S}_{t_2}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\right\| \le C_{2.a}^2\kappa^4, \tag{126}$$

where $C_{2.a}$ is given in Lemma 2. Then given that $\eta \le c_\eta$ for some sufficiently small c_η, we have $\log(1-\eta) \ge -\eta/2$. As a result, if $t_3 - t_2 \ge 8\log(10C_{2.a}\kappa)/\eta \ge \log(C_{2.a}^{-2}\kappa^{-4}/100)/\log(1-\eta)$, we have $(1-\eta)^{t_3-t_2} \le C_{2.a}^{-2}\kappa^{-4}/100$. When $C_{\min}$ is sufficiently large we may choose such t_3 which simultaneously satisfies $t_3 \le t_2 + T_{\min}/16 \le T_{\max}$ since $8\log(10C_{2.a}\kappa)/\eta \le \frac{C_{\min}}{32\eta}\log(\|X_\star\|/\alpha) = T_{\min}/32$. Invoking (125), we obtain

$$\left\|\Sigma_\star^{-1}(\widetilde{S}_{t_3}\widetilde{S}_{t_3}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\right\| \le (C_{2.a}^{-2}\kappa^{-4}/100)(C_{2.a}^2\kappa^4) + \frac{1}{100} = \frac{1}{50} \le \frac{1}{10}, \tag{127}$$

which implies the desired bound (22).

Proof of inequality (126). It is straightforward to verify that

$$\left\|\Sigma_\star^{-1}(\widetilde{S}_{t_2}\widetilde{S}_{t_2}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\right\| \le \max\left(\|\Sigma_\star^{-1}\widetilde{S}_{t_2}\|^2 - 1, 1 - \sigma_{\min}^2(\Sigma_\star^{-1}\widetilde{S}_{t_2})\right),$$

which combined with (17d) implies that

$$\|\Sigma_\star^{-1}\widetilde{S}_{t_2}\|^2 - 1 \le \|\Sigma_\star^{-1}\|^2\|\widetilde{S}_{t_2}\|^2 \le \sigma_{\min}^{-2}(X_\star)C_{2.a}^2\kappa^2\|X_\star\|^2 = C_{2.a}^2\kappa^4.$$

In addition, by Corollary 1 we have

$$1 - \sigma_{\min}^2(\Sigma_\star^{-1}\widetilde{S}_{t_2}) \le 1 - \frac{1}{10} = \frac{9}{10}.$$

Choosing $C_{2.a}$ sufficiently large (say $C_{2.a} \ge 1$) yields $C_{2.a}^2\kappa^4 \ge 9/10$, and hence the claim (126).

E Proofs for Phase III

To characterize the behavior of $\|X_tX_t^\top - M_\star\|_{\mathsf{F}}$, it is particularly helpful to consider the following decomposition into three error terms related to the signal term, the misalignment term, and the overparametrization term.

Lemma 26. *For all $t \ge t_3$, as long as $\|\Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)\Sigma_\star^{-1}\| \le 1/10$, one has*

$$\|X_tX_t^\top - M_\star\|_{\mathsf{F}} \le 4\|X_\star\|^2\left(\|\Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)\Sigma_\star^{-1}\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}}\right) + 4\|X_\star\|\|\widetilde{O}_t\|.$$

Note that the overparametrization error $\|\widetilde{O}_t\|$ stays small, as stated in (17b) and (19). Therefore we only need to focus on the shrinkage of the first two terms $\|\Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)\Sigma_\star^{-1}\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}}$, which is the focus of the lemma below.

Lemma 27. *For any $t : t_3 \le t \le T_{\max}$, one has*

$$\begin{aligned}&\|\Sigma_\star^{-1}(\widetilde{S}_{t+1}\widetilde{S}_{t+1}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star\|_{\mathsf{F}} \\ &\quad\le \left(1-\frac{\eta}{10}\right)\left(\|\Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)\Sigma_\star^{-1}\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}}\right) + \eta\left(\frac{\|\widetilde{O}_t\|}{\|X_\star\|}\right)^{1/2}. \end{aligned} \tag{128}$$

In particular, $\|\Sigma_\star^{-1}(\widetilde{S}_{t+1}\widetilde{S}_{t+1}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\| \le 1/10$ for all t such that $t_3 \le t \le T_{\max}$.

We now show how Lemma 5 is implied by the above two lemmas. To begin with, we apply Lemma 27 repeatedly to obtain the following bound for all $t \in [t_3, T_{\max}]$:

$$\begin{aligned}&\|\Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)\Sigma_\star^{-1}\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}} \\ &\le \left(1-\frac{\eta}{10}\right)^{t-t_3}\left(\|\Sigma_\star^{-1}(\widetilde{S}_{t_3}\widetilde{S}_{t_3}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_{t_3}\widetilde{S}_{t_3}^{-1}\Sigma_\star\|_{\mathsf{F}}\right) + 10\max_{t_3\le\tau\le t}\left(\frac{\|\widetilde{O}_\tau\|}{\|X_\star\|}\right)^{1/2}, \end{aligned} \tag{129}$$

which motivates us to control the error at time t_3.

We know from Corollary 2 that $\|\Sigma_\star^{-1}(\widetilde{S}_{t_3}\widetilde{S}_{t_3}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\| \le 1/10$. Since $\Sigma_\star^{-1}(\widetilde{S}_{t_3}\widetilde{S}_{t_3}^\top - \Sigma_\star^2)\Sigma_\star^{-1}$ is a $r_\star \times r_\star$ matrix, we have $\|\Sigma_\star^{-1}(\widetilde{S}_{t_3}\widetilde{S}_{t_3}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\|_{\mathsf{F}} \le \sqrt{r_\star}/10$. In addition, we infer from (17c) that

$$\|\widetilde{N}_{t_3}\widetilde{S}_{t_3}^{-1}\Sigma_\star\|_{\mathsf{F}} \le \sqrt{r_\star}\|\widetilde{N}_{t_3}\widetilde{S}_{t_3}^{-1}\Sigma_\star\| \le \sqrt{r_\star}c_2\kappa^{-C_\delta/2}\|X_\star\| \le \sqrt{r_\star}\|X_\star\|/10,$$

as long as c_2 is sufficiently small. Combine the above two bounds to arrive at the conclusion that

$$\|\Sigma_\star^{-1}(\widetilde{S}_{t_3}\widetilde{S}_{t_3}^\top - \Sigma_\star^2)\Sigma_\star^{-1}\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_{t_3}\widetilde{S}_{t_3}^{-1}\Sigma_\star\|_{\mathsf{F}} \le \frac{\sqrt{r_\star}}{10} + \|X_\star\|^{-1}\frac{\sqrt{r_\star}\|X_\star\|}{10} = \frac{\sqrt{r_\star}}{5}. \tag{130}$$

Combining the two inequalities (129) and (130) yields for all $t \in [t_3, T_{\max}]$

$$\|\Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2)\Sigma_\star^{-1}\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}} \le \frac{1}{5}\left(1-\frac{\eta}{10}\right)^{t-t_3}\sqrt{r_\star} + 10\max_{t_3\le\tau\le t}\left(\frac{\|\widetilde{O}_\tau\|}{\|X_\star\|}\right)^{1/2}.$$

We can then invoke Lemma 26 to see that

$$\begin{aligned}\|X_tX_t^\top - M_\star\|_{\mathsf{F}} &\le \frac{4\|X_\star\|^2}{5}\left(1-\frac{\eta}{10}\right)^{t-t_3}\sqrt{r_\star} + 40\|X_\star\|^2\max_{t_3\le\tau\le t}\left(\frac{\|\widetilde{O}_\tau\|}{\|X_\star\|}\right)^{1/2} + 4\|X_\star\|\|\widetilde{O}_t\| \\ &\le \left(1-\frac{\eta}{10}\right)^{t-t_3}\sqrt{r_\star}\|M_\star\| + 80\|M_\star\|\max_{t_3\le\tau\le t}\left(\frac{\|\widetilde{O}_\tau\|}{\|X_\star\|}\right)^{1/2},\end{aligned}$$

where in the last line we use $\|\widetilde{O}_t\| \le \|X_\star\|$—an implication of (19). To see this, the assumption (12c) implies that $\alpha \le \|X_\star\|$ as long as $\eta \le 1/2$ and $C_\alpha \ge 4$, which in turn implies $\|\widetilde{O}_t\| \le \alpha^{2/3}\|X_\star\|^{1/3} \le \|X_\star\|$. This completes the proof for the first part of Lemma 5 with $c_5 = 1/10$.

For the second part of Lemma 5, notice that

$$8c_5^{-1}\max_{t_3\le\tau\le T_{\max}}(\|\widetilde{O}_\tau\|/\|X_\star\|)^{1/2} \le \frac{1}{2}\left(\frac{\alpha}{\|X_\star\|}\right)^{1/3}$$

by (19), thus

$$\|X_tX_t^\top - M_\star\|_{\mathsf{F}} \le (1-c_5\eta)^{t-t_3}\sqrt{r_\star}\|M_\star\| + \frac{1}{2}\left(\frac{\alpha}{\|X_\star\|}\right)^{1/3}$$

for $t_3 \le t \le T_{\max}$. There exists some iteration number $t_4 : t_3 \le t_4 \le t_3 + \frac{2}{c_5\eta}\log(\|X_\star\|/\alpha) \le t_3 + T_{\min}/16$ such that

$$(1-c_5\eta)^{t_4-t_3} \le \left(\frac{\alpha}{\|X_\star\|}\right)^2 \le \frac{1}{2\sqrt{r_\star}}\left(\frac{\alpha}{\|X_\star\|}\right)^{1/3},$$

where the last inequality is due to (12c). It is then clear that t_4 has the property claimed in the lemma.

E.1 Proof of Lemma 26

Starting from (46), we may deduce

$$\begin{aligned}
\|X_tX_t^\top - M_\star\|_{\mathsf{F}} &\le \|\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2\|_{\mathsf{F}} + 2\|\widetilde{S}_t\|\|\widetilde{N}_t\|_{\mathsf{F}} + \|\widetilde{N}_t\|\|\widetilde{N}_t\|_{\mathsf{F}} + \|\widetilde{O}_t\|\|\widetilde{O}_t\|_{\mathsf{F}} \\
&\le \|X_\star\|^2\left(\|\Sigma_\star^{-1}\widetilde{S}_t\widetilde{S}_t^\top\Sigma_\star^{-1} - I\|_{\mathsf{F}} + 2\|\Sigma_\star^{-1}\widetilde{S}_t\|^2\|X_\star\|^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}} + \sqrt{n}\left(\frac{\|\widetilde{O}_t\|}{\|X_\star\|}\right)^2\right) \\
&\le 4\|X_\star\|^2\left(\|\Sigma_\star^{-1}\widetilde{S}_t\widetilde{S}_t^\top\Sigma_\star^{-1} - I\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}} + \frac{\|\widetilde{O}_t\|}{\|X_\star\|}\right),
\end{aligned} \tag{131}$$

where the penultimate line used $\|\widetilde{O}_t\|_{\mathsf{F}} \le \sqrt{n}\|\widetilde{O}_t\|$, and the last line follows from $\|\Sigma_\star^{-1}\widetilde{S}_t\|^2 = \|\Sigma_\star^{-1}\widetilde{S}_t\widetilde{S}_t^\top\Sigma_\star^{-1}\| \le 1 + \|\Sigma_\star^{-1}\widetilde{S}_t\widetilde{S}_t^\top\Sigma_\star^{-1} - I\| \le 2$ (recall that $\|\Sigma_\star^{-1}\widetilde{S}_t\widetilde{S}_t^\top\Sigma_\star^{-1} - I\| \le 1/10$ by assumption) and from (19).

E.2 Proof of Lemma 27

Recall the definition of Γ_t from (114):

$$\Gamma_t := \Sigma_\star^{-1}\widetilde{S}_t\widetilde{S}_t^\top\Sigma_\star^{-1} - I.$$

Fix any $t \in [t_3, T_{\max}]$, if (128) were true for all $\tau \in [t_3, t]$, taking into account that $\|\widetilde{O}_\tau\|/\|X_\star\| \le 1/10000$ for all $\tau \in [t_3, T_{\max}]$ by (19), we could show by induction that $\|\Gamma_\tau\| \le 1/10$ for all $\tau \in [t_3, t]$. Thus it suffices to assume $\|\Gamma_t\| \le 1/10$ and prove (128).

Apply Lemma 25 with Frobenius norm to obtain

$$\|\Gamma_{t+1}\|_{\mathsf{F}} \le (1-\eta)\|\Gamma_t\|_{\mathsf{F}} + \eta\frac{C_{25}\kappa^4}{\|X_\star\|^2}\|U_\star^\top\Delta_t\|_{\mathsf{F}} + \frac{1}{16}\eta\|X_\star\|^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}} + \eta\left(\frac{\|\widetilde{O}_t\|}{\|X_\star\|}\right)^{7/12}, \tag{132}$$

In addition, Lemma 23 tells us that

$$\|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star\|_{\mathsf{F}} \le \left(1 - \frac{\eta}{3(\|Z_t\|+\eta)}\right)\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}} + \eta\frac{C_{23}\kappa^2}{c_\lambda\|X_\star\|}\|U_\star^\top\Delta_t\|_{\mathsf{F}} + \eta\left(\frac{\|\widetilde{O}_t\|}{\sigma_{\min}(\widetilde{S}_t)}\right)^{2/3}\|X_\star\|,$$

where $Z_t = \Sigma_\star^{-1}(\widetilde{S}_t\widetilde{S}_t^\top + \lambda I)\Sigma_\star^{-1}$. It is easy to check that $\|Z_t\| \le 1 + \|\Gamma_t\| + c_\lambda \le 2$ as $\|\Gamma_t\| \le 1/10$ and c_λ is sufficiently small. In addition, one has $\sigma_{\min}(\widetilde{S}_t)^2 \ge (1 - \|\Gamma_t\|)\sigma_{\min}(X_\star)^2$ and $\|\widetilde{O}_t\|/\sigma_{\min}(\widetilde{S}_t) \le (2\kappa)^{-24}$. Combine these relationships together to arrive at

$$\|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star\|_{\mathsf{F}} \le \left(1 - \frac{\eta}{8}\right)\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}} + \eta\frac{C_{23}\kappa^2}{c_\lambda\|X_\star\|}\|U_\star^\top\Delta_t\|_{\mathsf{F}} + \frac{1}{2}\eta\|X_\star\|\left(\frac{\|\widetilde{O}_t\|}{\|X_\star\|}\right)^{7/12}. \tag{133}$$

Summing up (132), (133), we obtain

$$\begin{aligned}
&\|\Gamma_{t+1}\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_{t+1}\widetilde{S}_{t+1}^{-1}\Sigma_\star\|_{\mathsf{F}} \\
&\quad\le \left(1 - \frac{\eta}{8}\right)(\|\Gamma_t\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}}) + \eta\frac{2(C_{23}+C_{25}c_\lambda)\kappa^4}{c_\lambda\|X_\star\|^2}\|U_\star^\top\Delta_t\|_{\mathsf{F}} + 2\eta\left(\frac{\|\widetilde{O}_t\|}{\|X_\star\|}\right)^{7/12}.
\end{aligned} \tag{134}$$

This is close to our desired conclusion, but we would need to eliminate $\|U_\star^\top \Delta_t\|_{\mathsf{F}}$. To this end we observe

$$\begin{aligned}
\|U_\star^\top \Delta_t\|_{\mathsf{F}} &\le \sqrt{r_\star}\|\Delta_t\| \\
&\le 8\delta\sqrt{r_\star}\left(\|\widetilde{S}_t\widetilde{S}_t^\top - \Sigma_\star^2\|_{\mathsf{F}} + \|\widetilde{S}_t\|\|\widetilde{N}_t\|_{\mathsf{F}} + n\|\widetilde{O}_t\|^2\right) \\
&\le 16c_\delta\kappa^{-4}\|X_\star\|^2\left(\|\Gamma_t\|_{\mathsf{F}} + \|X_\star\|^{-1}\|\widetilde{N}_t\widetilde{S}_t^{-1}\Sigma_\star\|_{\mathsf{F}} + \left(\frac{\|\widetilde{O}_t\|}{\|X_\star\|}\right)^{2/3}\right),
\end{aligned}$$

where the first line follows from $U_\star$ being of rank $r_\star$, the second line follows from Lemma 11, and the last line follows from (10) and from controlling the sum inside the brackets in a similar way as (131).

The conclusion follows from plugging the above inequality into (134), noting that c_δ can be chosen sufficiently small and that $\|\widetilde{O}_t\|/\|X_\star\|$ is sufficiently small due to (19).