Non-convex matrix sensing: Breaking the quadratic rank barrier in the sample complexity

Anonymous CPAL submission

For the problem of reconstructing a low-rank matrix from a few linear measurements, two classes of algorithms have been widely studied in the literature: convex approaches based on nuclear norm minimization, and non-convex approaches that use factorized gradient descent. Under certain statistical model assumptions, it is known that nuclear norm minimization recovers the ground truth as soon as the number of samples scales linearly with the number of degrees of freedom of the ground-truth. In contrast, while non-convex approaches are computationally less expensive, existing recovery guarantees assume that the number of samples scales at least quadratically with the rank r of the ground-truth matrix. In this paper, we close this gap by showing that the non-convex approaches can be as efficient as nuclear norm minimization in terms of sample complexity. Namely, we consider the problem of reconstructing a positive semidefinite matrix from a few Gaussian measurements. We show that factorized gradient descent with spectral initialization converges to the ground truth with a linear rate as soon as the number of samples scales with $\Omega(rd\kappa^2)$, where d is the dimension, and κ is the condition number of the ground truth matrix. This improves the previous rank-dependence in the sample complexity of non-convex matrix factorization from quadratic to linear. Our proof relies on a probabilistic decoupling argument, where we show that the gradient descent iterates are only weakly dependent on the individual entries of the measurement matrices. We expect that our proof technique is of independent interest for other non-convex problems.

1. Introduction

Low-rank matrix recovery refers to the problem of reconstructing an unknown matrix $\mathbf{X}_\star \in \mathbb{R}^{d_1 \times d_2}$ with $\text{rank}(\mathbf{X}_\star) =: r \ll \min\{d_1; d_2\}$ from an underdetermined linear set of equations of the form

$$\mathbf{y} = \mathcal{A}(\mathbf{X}_\star) \in \mathbb{R}^m,$$

where $\mathcal{A}$ represents a known linear measurement operator and $\mathbf{y} \in \mathbb{R}^m$ are the observations. This ill-posed inverse problem has been the topic of intense study over many years, given its relevance to a variety of questions in machine learning, signal processing, and statistics. Notable applications include matrix completion [1], phase retrieval [2], robust PCA [3], blind deconvolution [4] and its extension to blind demixing [5]. A major goal has been to develop methods which are *sample-efficient*; that is, they can reconstruct the low-rank matrix $\mathbf{X}_\star$ if the number of observations m is roughly of the same order as the number of degrees of freedom of $\mathbf{X}_\star$. In addition, these methods should also be scalable, meaning they remain computationally efficient as the problem dimensions are increasing.

Several different algorithmic approaches to solve this problem have been proposed. One line of research revolves around the idea of convex relaxation. Here, the nuclear norm $\|\cdot\|_*$, i.e., the sum of singular values, is considered as a convex proxy for the rank function. For many problem classes, including matrix sensing [6], matrix completion [7, 8], and blind deconvolution and demixing [9], it has been shown that this approach is able to recover the unknown matrix $\mathbf{X}_\star$ as soon as the number of samples m scales, up to logarithmic factors, with the information-theoretically optimal sample

Submitted to Second Conference on Parsimony and Learning (CPAL 2025). Do not distribute.

complexity $r(d_1 + d_2)$. However, a drawback of these convex approaches is that they tend to be computationally prohibitive.

For this reason, many studies have considered non-convex heuristics where one minimizes an objective of the form

$$f(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{m} \ell\left(\mathbf{y}_i, \left(\mathcal{A}(\mathbf{U}\mathbf{V}^\top)\right)_i\right), \tag{1}$$

with low-rank factors $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ and a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. To minimize the objective function, local search methods such as gradient descent or alternating minimization with a suitable initialization are used. An advantage of these approaches is that they are computationally less demanding since there are only $r(d_1 + d_2)$ optimization variables instead of at least $d_1 d_2$ optimization variables in the convex approaches. However, due to the non-convexity of the objective function, it might initially seem unclear that local search methods can find the global minimum of the objective (1) efficiently.

Nevertheless, in recent years a large body of literature has demonstrated that under certain statistical assumptions, these methods converge to the global minimum and are thus able to recover the unknown low-rank matrix $\mathbf{X}_\star$. For instance, gradient descent with spectral initialization [10] and other variants of gradient descent [11–13] have been studied for matrix sensing and related problems. Similarly, numerous works have established convergence and recovery guarantees for matrix completion [14–19] and blind deconvolution and demixing [20, 21]. In addition, recent studies also analyzed overparameterized models, where the exact rank r is either not known or where the number of parameters exceeds the number of samples [22–28]. Beyond gradient descent, also alternating minimization [29] and other non-convex methods based on matrix factorization such as GNMR [30] have been proposed and studied. For a more extensive overview of the literature, we refer the reader to [19].

Despite this significant body of literature, the existing theoretical guarantees for non-convex methods based on matrix factorization in the literature are weaker than the corresponding guarantees for nuclear norm minimization in terms of sample complexity. Namely, in all these results, it is required that the number of samples m scales at least quadratically with the rank r and thus the total number of samples scales at least with $r^2(d_1+d_2)$. This raises the question of whether this quadratic rank-dependence is just an artifact of the proof or whether it is inherent to the problem, see, e.g., [31, p. 5264].

In this paper, we resolve this question in the context of symmetric matrix sensing. Under the assumption that $\mathcal{A}$ is a Gaussian measurement operator and $\mathbf{X}_\star \in \mathbb{R}^{d \times d}$ is symmetric and positive semidefinite, we show that factorized gradient descent with spectral initialization is able to recover the unknown matrix $\mathbf{X}_\star$ if the number of samples scales with rd, which, in particular, is linear in the rank of $\mathbf{X}_\star$. Our proof is based on a novel probabilistic decoupling argument. Namely, we show that the trajectory of the gradient descent iterates depends only weakly on any given generalized entry of the measurement matrices in a suitable sense. This allows us to prove stronger concentration bounds than what would be possible if one were to rely solely on uniform concentration bounds (such as the Restricted Isometry Property, for example). To establish this weak dependence, we construct auxiliary virtual sequences and combine this with an ε-net argument. Our novel proof approach paves the way to improved sample complexity bounds for other non-convex algorithms and beyond.

Finally, we note that there are also several non-convex algorithms for low-rank matrix recovery that are not explicitly based on matrix factorization formulation as in equation (1). This includes, for example, Singular Value Projection [32, 33], Normalized Iterative Hard Thresholding [34], Iteratively Reweighted Least Squares (IRLS), see, e.g., [35–38], and Atomic Decomposition for Minimum Rank Approximation (ADMiRA) [39]. However, since many of these algorithms operate in the full matrix space they are less computationally efficient than algorithms based on matrix factorization. In the case of IRLS, only local convergence guarantees (with explicit convergence rates) are known. There have also been algorithms studied that are based on Riemannian optimization, see, e.g., [40–

42]. However, these algorithms require that the sample complexity scales quadratically in the rank r. We believe our work can lead to improved sample size guarantees for these methods as well.

Organization of the paper: This paper is structured as follows. In the remainder of Section 1, we will describe the formal setting and the algorithm, and we will state our main theoretical result, which is Theorem 1.2. In Section 2, we discuss some technical preliminaries regarding the Restricted Isometry Property and perturbation bounds for eigenspaces. In Section 3, we discuss the proof strategy, and we introduce the virtual sequences, which are the main ingredient to establish that the sample complexity depends only linearly on the rank. Section 4 contains the proof of the main result of this paper, Theorem 1.2. We discuss interesting directions for future research in Section 5.

Notation: Before we state the problem formulation, we introduce some basic notation. For a matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$, we denote its transpose by $\mathbf{A}^\top$ and its trace by $\operatorname{trace}(\mathbf{A})$. For matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d_1 \times d_2}$, we define their inner product via $\langle \mathbf{A}, \mathbf{B} \rangle := \operatorname{trace}\left(\mathbf{A}\mathbf{B}^\top\right)$. The Frobenius norm $\|\cdot\|_F$ denotes the norm induced by this inner product, i.e., $\|\mathbf{A}\|_F := \sqrt{\langle \mathbf{A}, \mathbf{A} \rangle}$. By $\|\mathbf{A}\|$ we denote the spectral norm of the matrix $\mathbf{A}$, i.e., the largest singular value of the matrix $\mathbf{A}$. By $\|\mathbf{v}\|_2 := \sqrt{\sum_{i=1}^d \mathbf{v}_i^2}$ we denote the Euclidean norm of a vector $\mathbf{v} \in \mathbb{R}^d$. The set $\mathcal{S}^d \subset \mathbb{R}^{d \times d}$ represents the set of all symmetric matrices. The matrix $\mathbf{Id} \in \mathcal{S}^d$ denotes the identity matrix. Moreover, $\mathcal{I} : \mathcal{S}^d \to \mathcal{S}^d$ represents the identity mapping.

Furthermore, for a matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$ of rank r we denote its singular value decomposition by $\mathbf{A} = \mathbf{V}_\mathbf{A} \mathbf{\Sigma}_\mathbf{A} \mathbf{W}_\mathbf{A}^\top$. The matrices $\mathbf{V}_\mathbf{A} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{W}_\mathbf{A} \in \mathbb{R}^{d_2 \times r}$ contain the left-singular and right-singular vectors of the matrix $\mathbf{A}$. The matrix $\mathbf{\Sigma}_\mathbf{A} \in \mathbb{R}^{r \times r}$ contains the singular values of $\mathbf{A}$. Moreover, $\mathbf{V}_{\mathbf{A},\perp} \in \mathbb{R}^{(d_1 - r) \times r}$ represents an orthogonal matrix whose column span is orthogonal to the column span of $\mathbf{V}_\mathbf{A}$.

1.1. Problem formulation

In this paper, we focus on symmetric matrix sensing. More precisely, we study the problem of reconstructing a symmetric, positive semidefinite matrix $\mathbf{X}_\star \in \mathbb{R}^{d \times d}$ with rank r from m linear observations of the form

$$\mathbf{y}_i = \frac{1}{\sqrt{m}} \langle \mathbf{A}_i, \mathbf{X}_\star \rangle := \frac{1}{\sqrt{m}} \operatorname{trace}\left(\mathbf{A}_i \mathbf{X}_\star\right) \qquad \text{for } i = 1, 2, \ldots, m. \tag{2}$$

Definition 1.1 (Measurement operator)**.** *We define the linear measurement operator $\mathcal{A} : \mathcal{S}^d \to \mathbb{R}^m$ by*

$$[\mathcal{A}(\mathbf{X})]_i := \frac{1}{\sqrt{m}} \langle \mathbf{A}_i, \mathbf{X} \rangle \qquad \textit{for } i = 1, 2, \ldots, m$$

for any matrix $\mathbf{X} \in \mathcal{S}^d$. Recall that $\mathcal{S}^d \subset \mathbb{R}^{d \times d}$ denotes the set of symmetric matrices. The matrices $\{\mathbf{A}_i\}_{i=1}^m \subset \mathbb{R}^{d \times d}$ represent known, symmetric measurement matrices. We assume that their entries are i.i.d. with distribution $\mathcal{N}(0, 1)$ on the diagonal and $\mathcal{N}(0, 1/2)$ on the off-diagonal entries. Each $\mathbf{A}_i$ is also known as a Gaussian orthogonal ensemble [43].

This measurement model has been considered before in, e.g., [10, 22]. With this notation in place, equation (2) can be written more compactly as

$$\mathbf{y} = \mathcal{A}\left(\mathbf{X}_\star\right).$$

To recover the ground-truth matrix $\mathbf{X}_\star$, we consider the non-convex objective function

$$\mathcal{L}(\mathbf{U}) := \frac{1}{4} \left\| \mathbf{y} - \mathcal{A}\left(\mathbf{U}\mathbf{U}^\top\right) \right\|_2^2 = \frac{1}{4} \left\| \mathcal{A}\left(\mathbf{X}_\star - \mathbf{U}\mathbf{U}^\top\right) \right\|_2^2, \tag{3}$$

where $\mathbf{U} \in \mathbb{R}^{d \times r}$ is a matrix and $\|\cdot\|_2$ denotes the ℓ_2-norm of a vector. To minimize this objective, we follow the two-stage approach introduced in [14] for matrix completion, which then subsequently was studied for matrix sensing in [10]. In the first stage, an initialization $\mathbf{U}_0$ is constructed via a so-called spectral initialization. This initialization is subsequently used as a starting point for

the gradient descent scheme in the second stage. To precisely define the spectral initialization, we denote by $\mathcal{A}^* : \mathbb{R}^m \to \mathcal{S}^d$ the adjoint operator of $\mathcal{A}$ with respect to the trace inner product defined in equation (2).

With this definition in place, we can consider the eigendecomposition of the matrix

$$\mathcal{A}^*(\mathbf{y}) =: \widetilde{\mathbf{V}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^\top,$$

where $\widetilde{\mathbf{V}} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix and the matrix $\widetilde{\mathbf{\Lambda}} \in \mathbb{R}^{d\times d}$ is diagonal matrix which contains the eigenvalues of $\mathcal{A}^*(\mathbf{y})$ sorted by their magnitude, i.e., $|\lambda_1(\mathcal{A}^*(\mathbf{y}))| \geq |\lambda_2(\mathcal{A}^*(\mathbf{y}))| \geq \ldots \geq |\lambda_d(\mathcal{A}^*(\mathbf{y}))|$.

Since the measurement matrices $\mathbf{A}_i$ are Gaussian we have that

$$\mathbb{E}\left[\mathcal{A}^*(\mathbf{y})\right] = \mathbb{E}\left[\left(\mathcal{A}^*\mathcal{A}\right)(\mathbf{X}_\star)\right] = \mathbf{X}_\star.$$

Since $\mathbf{X}_\star$ has rank r for a large enough enough sample size m, one has that the truncated rank-r eigendecomposition of $\mathcal{A}^*(\mathbf{y})$ fulfills $\widetilde{\mathbf{V}}_r\widetilde{\mathbf{\Lambda}}_r\widetilde{\mathbf{V}}_r \approx \mathbf{X}_\star$. Here, by $\widetilde{\mathbf{V}}_r \in \mathbb{R}^{d\times r}$ we denote a matrix which contains the first r columns of $\widetilde{\mathbf{V}}$ and by $\widetilde{\mathbf{\Lambda}}_r$ we denote a diagonal matrix which contains the largest r eigenvalues of $\mathcal{A}^*(\mathbf{y})$ in decreasing order. Motivated by this observation, the spectral initialization $\mathbf{U}_0$ is defined as

$$\mathbf{U}_0 := \widetilde{\mathbf{V}}_r\widetilde{\mathbf{\Lambda}}_r^{1/2}.$$

Here, the entries of the diagonal matrix $\widetilde{\mathbf{\Lambda}}_r^{1/2}$ are given by $\sqrt{|\lambda_i(\mathcal{A}^*(\mathbf{y}))|}$. As we will see, all entries of $\widetilde{\mathbf{\Lambda}}_r$ are positive with high probability.

After having computed the initialization $\mathbf{U}_0$, we use $\mathbf{U}_0$ as a starting point of the gradient descent scheme in the second stage, which is defined as follows

$$\mathbf{U}_{t+1} := \mathbf{U}_t - \mu\nabla\mathcal{L}(\mathbf{U}_t) \quad \text{for } t = 0, 1, \ldots,$$

where $\mu > 0$ denotes the step size. A direct computation shows that

$$\begin{aligned}\mathbf{U}_{t+1} &= \mathbf{U}_t + \mu\left[\left(\mathcal{A}^*\mathcal{A}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{U}_t \\ &= \mathbf{U}_t + \frac{\mu}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\rangle \mathbf{A}_i\mathbf{U}_t.\end{aligned} \tag{4}$$

All steps of the two-stage approach are summarized below in Algorithm 1.1.

Algorithm 1 Two-Stage Approach for Low-Rank Matrix Recovery

Input: Measurement operator $\mathcal{A} : \mathcal{S}^d \to \mathbb{R}^m$, observations $\mathbf{y} \in \mathbb{R}^m$, step size $\mu > 0$

Stage 1 (Spectral Initialization): Compute the truncated eigendecomposition $\widetilde{\mathbf{V}}_r\widetilde{\mathbf{\Lambda}}_r\widetilde{\mathbf{V}}_r^\top$ of the data matrix $\mathbf{D} := \mathcal{A}^*(\mathbf{y}) = \frac{1}{\sqrt{m}}\sum_{i=1}^m y_i\mathbf{A}_i$. Here, $\widetilde{\mathbf{\Lambda}}_r \in \mathbb{R}^{d\times d}$ is the diagonal matrix which contains the r largest eigenvalues of the data matrix $\mathbf{D}$ (in absolute value). The columns of $\widetilde{\mathbf{\Lambda}}_r \in \mathbb{R}^{d\times r}$ contain the corresponding eigenvectors. Define the initialization $\mathbf{U}_0 \in \mathbb{R}^{d\times r}$ via

$$\mathbf{U}_0 := \widetilde{\mathbf{V}}_r\widetilde{\mathbf{\Lambda}}_r^{1/2}.$$

Stage 2 (Gradient descent):
for $t = 0, 1, 2, \ldots$ **do**

$$\mathbf{U}_{t+1} := \mathbf{U}_t - \mu\nabla\mathcal{L}\left(\mathbf{U}_t\right)$$

end for

1.2. Main result

To formulate our main result, we need to introduce the condition number of $\mathbf{X}_\star$, which is defined as

$$\kappa := \frac{\|\mathbf{X}_\star\|}{\sigma_{\min}(\mathbf{X}_\star)}.$$

Here, $\sigma_{\min}(\mathbf{X}_\star)$ denotes the smallest non-zero singular value of $\mathbf{X}_\star$.

Next, let $\mathbf{U}_\star \in \mathbb{R}^{d\times r}$ be a matrix such that $\mathbf{X}_\star = \mathbf{U}_\star\mathbf{U}_\star^\top$. The matrix $\mathbf{U}_\star$ is uniquely defined only up to an orthogonal transformation $\mathbf{R} \in \mathbb{R}^{r\times r}$, which is why we can only expect to be able to reconstruct $\mathbf{U}_\star$ up to this ambiguity. To account for this, we will introduce the error metric

$$\operatorname{dist}(\mathbf{U}_t, \mathbf{U}_\star) := \min_{\mathbf{R}\in\mathbb{R}^{r\times r},\, \mathbf{R}^\top\mathbf{R}=\mathbf{Id}_r} \|\mathbf{U}_t\mathbf{R} - \mathbf{U}_\star\|_F. \tag{5}$$

With this notation in place, we can state the main result of this paper.

Theorem 1.2. *Let $\mathcal{A} : \mathcal{S}^d \to \mathbb{R}^m$ be a linear measurement operator as in Definition 1.1 with Gaussian measurement matrices. Moreover, let $\mathbf{X}_\star \in \mathcal{S}^d$ be a positive semidefinite matrix of rank r. Given observations $\mathbf{y} = \mathcal{A}(\mathbf{X}_\star) \in \mathbb{R}^m$, let $\mathbf{U}_0, \mathbf{U}_1, \mathbf{U}_2, \ldots$ be the sequence of gradient descent iterates which are obtained via the two-stage approach described in Algorithm 1. Assume that the number of observations m satisfies*

$$m \geq Crd\kappa^2,$$

and that the step size $\mu > 0$ satisfies

$$\frac{32}{6^d\sigma_{\min}(\mathbf{X}_\star)}\log(16r) \leq \mu \leq \frac{c_1}{\kappa\|\mathbf{X}_\star\|}. \tag{6}$$

Then, with probability at least $1 - 7\exp(-d)$, it holds for all iterations $t \geq 0$ that

$$dist^2(\mathbf{U}_t, \mathbf{U}_\star) \leq c_2 r\left(1 - c_3\mu\sigma_{\min}(\mathbf{X}_\star)\right)^t \sigma_{\min}(\mathbf{X}_\star).$$

Here, $C, c_1, c_2, c_3 > 0$ denote absolute constants.

Remark 1.3. *The lower bound in assumption (6) is rather mild since the left-hand side in this inequality converges to 0 exponentially as the dimension d increases. If the dimension d is larger than an absolute constant, then condition (6) can always be satisfied for some step size μ.*

Theorem 1.2 shows that factorized gradient descent with spectral initialization converges to the ground truth with a linear rate as soon as the number of samples scales at least with $rd\kappa^2$. In particular, the bound on the sample complexity is linear in the rank r. This improves over previous results in the matrix sensing literature, which have a sample complexity of order at least $r^2d\kappa^2$, see, e.g., [10] or [11]. In particular, the sample complexity in Theorem 1.2 is optimal with respect to the rank r and dimension d. To the best of our knowledge, this is the first result in the literature which achieves this optimal dependence in the rank for the non-convex low-rank matrix recovery.

Compared to approaches based on nuclear norm or trace minimization, which only need $\Omega(rd)$ samples in the matrix sensing scenario, our result is still suboptimal by a factor of κ^2. However, all previous results in the literature on non-convex low-rank matrix recovery based on factorized gradient descent require having at least this quadratic dependence on the condition number. It remains an interesting open problem whether the dependence of the sample complexity on the condition number is necessary or an artifact of the proof.

Our main result implies that $\operatorname{dist}(\mathbf{U}_t, \mathbf{U}_\star) \leq \varepsilon$ after $O\left(\frac{\log(r/(\varepsilon\sigma_{\min}(\mathbf{X}_\star)))}{\mu\sigma_{\min}(\mathbf{X}_\star)}\right)$ iterations. Thus, if we choose the largest possible step size $\mu \asymp 1/(\kappa\|\mathbf{X}_\star\|)$ we obtain that we reach ε-accuracy after $O\left(\kappa^2\log(r/(\varepsilon\sigma_{\min}(\mathbf{X}_\star)))\right)$ iterations. Previous work [10] allows for a larger step size $\mu \lesssim 1/(\kappa\|\mathbf{X}_\star\|)$ which yields that one can reach ε-accuracy after $O\left(\kappa\log(r/(\varepsilon\sigma_{\min}(\mathbf{X}_\star)))\right)$ iterations, whereas Theorem 1.2 requires $\mu \lesssim 1/(\kappa\|\mathbf{X}_\star\|)$. It remains an open problem whether this additional condition number in the step size bound can be removed.

Remark 1.4 (Landscape Analysis). *Several works [44–47] have shown that if $m \gtrsim rd$, then the loss landscape of the objective function $\mathcal{L}$ in (3) is benign in the sense that $\mathcal{L}$ has no spurious local minima and all saddle points have at least one direction of strictly negative curvature. It has been established that in such a scenario gradient descent starting from random initialization will converge to the ground truth [48]. However, these results do not imply any guarantees on the convergence rate or on the computational complexity. In fact, there exist examples [49] where gradient descent may take exponential time to escape saddle points. For this reason, the results mentioned above are not directly comparable to our results.*

2. Preliminaries

In the following, we will discuss several technical preliminaries, which are needed in our proof.

2.1. The Restricted Isometry Property

We first recall the Restricted Isometry Property (RIP).

Definition 2.1 (Restricted Isometry Property). *The linear measurement operator $\mathcal{A} : \mathcal{S}^d \to \mathbb{R}^m$ satisfies the Restricted Isometry Property (RIP), of rank r with RIP-constant $\delta_r > 0$, if it holds for all symmetric matrices $\mathbf{Z} \in \mathbb{R}^{d\times d}$ of rank at most r that*

$$(1-\delta_r)\left\|\mathbf{Z}\right\|_F^2 \leq \left\|\mathcal{A}(\mathbf{Z})\right\|_2^2 \leq (1+\delta_r)\left\|\mathbf{Z}\right\|_F^2.$$

In previous works, it was shown that as soon as the measurement operator $\mathcal{A}$ has the RIP, then convex approaches based on nuclear norm minimization as well as non-convex approaches are able to recover the ground truth matrix, see, e.g., [6, 10].

It is well known that as soon as the number of samples m satisfies $m \gtrsim rd$ then the measurement operator $\mathcal{A}$ has the RIP of order r with high probability. This fact is stated in the following lemma.

Lemma 2.2. *Let $\mathcal{A} : \mathcal{S}^d \to \mathbb{R}^m$ be a Gaussian measurement operator as described in Section 1.1. Then the RIP constant δ_r satisfies $\delta_r \leq \delta \leq 1$ with probability $1-\varepsilon$ when*

$$m \geq C\delta^{-2}(rd + \log(2\varepsilon^{-1})),$$

where $C > 0$ is a universal constant. In particular, we have with probability at least $1 - \exp(-d)$, $m \geq C\delta^{-2}rd$.

This lemma differs from similar lemmas in the literature (see, e.g., [50]) by specifying how m depends on the RIP-constant δ. A proof of this lemma is provided in Appendix D.1 together with a more detailed discussion of how this lemma relates to previous work.

Remark 2.3. *The works mentioned in Remark 1.4 have shown that the RIP implies that the optimization landscape of $\mathcal{L}$ is benign (in the sense of Remark 1.4). Moreover, previous work such as [10] or [11], which analyzed gradient descent with spectral initialization similar to the paper at hand, relied on their analysis of gradient descent exclusively on the RIP property of the measurement operator $\mathcal{A}$. As we will explain in Section 3, purely relying on the RIP will not suffice to establish Theorem 1.2. For this reason, in addition to the RIP, we will use the orthogonal invariance of the Gaussian measurement operator $\mathcal{A}$.*

The RIP has several important consequences, which we will need throughout our proof. We recall them in the following lemma.

Lemma 2.4. *Let $\mathcal{A} : \mathcal{S}^d \to \mathbb{R}^m$ be a linear measurement operator on the set of symmetric matrices as defined above. Denote by δ_r the RIP constant of the operator $\mathcal{A}$ of order r. Then the following statements hold.*

1. *Let $\mathbf{V} \in \mathbb{R}^{d\times r'}$ be any matrix with orthonormal columns, i.e., $\mathbf{V}^\top\mathbf{V} = \boldsymbol{Id}$. Then it holds for any symmetric matrix $\mathbf{Z} \in \mathbb{R}^{d\times d}$ of rank at most r that*

$$\left\|\left(\mathcal{I} - \mathcal{A}^*\mathcal{A}\right)(\mathbf{Z})\mathbf{V}\right\|_F \leq \delta_{r+2r'}\left\|\mathbf{Z}\right\|_F. \tag{7}$$

   In particular, it holds that

$$\left\|\left(\mathcal{I} - \mathcal{A}^*\mathcal{A}\right)(\mathbf{Z})\right\| \leq \delta_{r+2}\left\|\mathbf{Z}\right\|_F. \tag{8}$$

2. *Let* $\mathbf{w} \in \mathbb{R}^d$ *such that* $\|\mathbf{w}\|_2 = 1$. *Define the orthogonal projection operators*

$$\mathcal{P}_{\mathbf{w}\mathbf{w}^\top}(\mathbf{Z}) := \langle \mathbf{w}\mathbf{w}^\top, \mathbf{Z}\rangle \mathbf{w}\mathbf{w}^\top, \tag{9}$$
$$\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z}) := \mathbf{Z} - \langle \mathbf{w}\mathbf{w}^\top, \mathbf{Z}\rangle \mathbf{w}\mathbf{w}^\top.$$

Then it holds for any symmetric matrix $\mathbf{Z} \in \mathbb{R}^{d\times d}$ *of rank at most* r *that*

$$|\langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\right)\rangle| \leq \delta_{r+2}\|\mathbf{Z}\|_F. \tag{10}$$

Some variants of these inequalities appeared in the literature already before; see, e.g., [23]. For completeness, we decided to include a proof in Appendix D.2.

Remark 2.5. *To keep the notation more concise, we will sometimes drop the subscript and just use the notation* δ *for the RIP constant. For all results below, the choices of* δ *satisfy* $\delta \leq \delta_{6r}$ *due to the monotonicity of the RIP constant with respect to the rank.*

2.2. Perturbation bounds for eigenspaces

The Davis-Kahan $\sin\theta$-theorem [51] states that the eigenspaces of a symmetric matrix are stable under perturbations of that matrix. Among others, we will need this result in order to show that the spectral initialization recovers the eigenspace of the ground truth matrix sufficiently well. We also will need it in order to show that $\mathbf{U}_{0,\mathbf{w}}$ is sufficiently close to $\mathbf{U}_0$.

To state this theorem, recall that for a symmetric matrix $\mathbf{Z} \in \mathbb{R}^{n\times n}$ with eigendecomposition $\mathbf{Z} = \mathbf{U}_\mathbf{Z}\mathbf{\Lambda}_\mathbf{Z}\mathbf{U}_\mathbf{Z}^\top$ the matrix $\mathbf{U}_{\mathbf{Z},r} \in \mathbb{R}^{n\times r}$ consists of the first r columns of $\mathbf{U}_\mathbf{Z}$ and the matrix $\mathbf{U}_{\mathbf{Z},r,\perp} \in \mathbb{R}^{n\times(n-r)}$ consists of the remaining $n-r$ columns. Moreover, recall that the eigenvalues of $\mathbf{Z}$ are ordered such that their magnitude is decreasing, i.e., $|\lambda_1(\mathbf{Z})| \geq |\lambda_2(\mathbf{Z})| \geq \ldots \geq |\lambda_n(\mathbf{Z})|$.

Lemma 2.6 (Davis-Kahan inequality, Corollary 2.8 in [52])**.** *Set* $|||\cdot||| = \|\cdot\|$ *or* $|||\cdot||| = \|\cdot\|_F$. *Let* $\mathbf{Z}_1 \in \mathbb{R}^{d\times d}$ *and* $\mathbf{Z}_2 \in \mathbb{R}^{d\times d}$ *be two symmetric matrices, such that the eigenvalues of* $\mathbf{Z}_1$ *satisfy* $|\lambda_r(\mathbf{Z}_1)| > |\lambda_{r+1}(\mathbf{Z}_1)|$ *for an integer* $1 \leq r < d$. *Let the eigendecompositions of* $\mathbf{Z}_1$ *and* $\mathbf{Z}_2$ *be given by* $\mathbf{Z}_1 = \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^\top$, *respectively* $\mathbf{Z}_2 = \mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^\top$. *Then, if the assumption*

$$\|\mathbf{Z}_1 - \mathbf{Z}_2\| \leq \left(1 - 1/\sqrt{2}\right)\left(|\lambda_r(\mathbf{Z}_1)| - |\lambda_{r+1}(\mathbf{Z}_1)|\right)$$

is fulfilled, it holds that

$$|||\mathbf{U}_{2,r,\perp}^\top \mathbf{U}_{1,r}||| \leq \frac{\sqrt{2}\,|||(\mathbf{Z}_1 - \mathbf{Z}_2)\mathbf{U}_{1,r}|||}{|\lambda_r(\mathbf{Z}_1)| - |\lambda_{r+1}(\mathbf{Z}_1)|}.$$

3. Outline of the proof

3.1. A fundamental barrier in previous work

Before we give an outline of our proof approach, we want to explain why in previous work the additional r-factor appeared in the sample complexity. As Lemma 4.1 below shows, it holds for the spectral initialization $\mathbf{U}_0$ with high probability that

$$\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\| \leq C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}.$$

In particular, for $m \gg \kappa^2 rd$ we have that

$$\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\| \ll \sigma_{\min}(\mathbf{X}_\star).$$

Thus, the spectral initialization ensures that the initialization $\mathbf{U}_0$ is in a neighborhood of the ground truth. We aim to establish that within this neighborhood, gradient descent converges with a linear rate. To show this, we note first that the gradient of our objective function $\mathcal{L}$ depends on the random

matrices $(\mathbf{A}_i)_{i=1}^m$. To deal with this, a common technique that has been used in previous works is to decompose the gradient of the objective function $\mathcal{L}$ into a sum of two terms:

$$\nabla \mathcal{L}(\mathbf{U}) = \mathbb{E}_{(\mathbf{A}_i)_{i=1}^m}[\nabla \mathcal{L}(\mathbf{U})] + \left[\nabla \mathcal{L}(\mathbf{U}) - \mathbb{E}_{(\mathbf{A}_i)_{i=1}^m}[\nabla \mathcal{L}(\mathbf{U})]\right].$$

The first term is the gradient of the population risk, i.e., the objective function one obtains in the limit case that the sample size m goes to infinity. The second term can be interpreted as a perturbation term that measures the deviation of the gradient of the empirical risk from the gradient of the population risk. In particular, this term converges to zero as the sample size m increases. For this reason, a major task in our proof is to show that the second summand is small with respect to a suitable norm as soon as the sample size m is sufficiently large. A direct computation shows that

$$\begin{aligned}\nabla \mathcal{L}(\mathbf{U}) - \mathbb{E}_{(\mathbf{A}_i)_{i=1}^m}[\nabla \mathcal{L}(\mathbf{U})] &= \left[(\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\mathbf{U}\mathbf{U}^\top - \mathbf{X}_\star\right)\right]\mathbf{U} \\ &= \frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\rangle \mathbf{A}_i - \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right).\end{aligned}$$

To deal with this deviation term, in previous works, bounds of the type

$$\left\| (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\| \ll \left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \tag{11}$$

needed to be established. A major challenge in establishing such bounds is that the gradient descent iterates $(\mathbf{U}_t)_t$ depend on the measurement matrices $(\mathbf{A}_i)_{i=1}^m$ in an intricate way. For this reason, standard matrix concentration inequalities are not directly applicable. To circumvent this issue, previous work establishes *uniform* bounds for the quantity

$$\sup_{\mathbf{Z}\in\mathcal{T}_{2r}} \left\| (\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathbf{Z})\right\|$$

where

$$\mathcal{T}_r := \left\{\mathbf{Z}\in\mathbb{R}^{d\times d} : \mathbf{Z} = \mathbf{Z}^\top, \operatorname{rank}(\mathbf{Z}) \le r, \left\|\mathbf{Z}\right\| \le 1\right\},$$

denotes the collection of matrices with rank at most r and bounded operator norm. Indeed, such a bound can be directly derived from the Restricted Isometry Property. Namely, when $\mathcal{A}$ has the RIP of order $2r+2$ with constant δ_{2r+2} then Lemma 2.4 implies that

$$\sup_{\mathbf{Z}\in\mathcal{T}_{2r}} \left\| (\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathbf{Z})\right\| \le \delta_{2r+2} \sup_{\mathbf{Z}\in\mathcal{T}_{2r}} \left\|\mathbf{Z}\right\|_F \le \delta_{2r+2}\sqrt{2r},$$

where in the second inequality, we used that the matrix $\mathbf{Z}$ has rank at most $2r$ and that $\left\|\mathbf{Z}\right\| = 1$. Thus, it follows from Lemma 2.2 that whenever $m \gg rd$ that with high probability we have that

$$\sup_{\mathbf{Z}\in\mathcal{T}_{2r}} \left\| (\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathbf{Z})\right\| \lesssim \sqrt{\frac{r^2 d}{m}}. \tag{12}$$

This shows that if we want to deduce inequality (11) from the uniform bound (12) we must assume that $m \gg r^2 d$. Indeed, several works, e.g., [22, 23, 53], relied precisely on this bound.

This leads to the question of whether the bound (12) can be sharpened. For example, in [53, p. 9], it was conjectured that using more refined techniques from empirical process theory, one may be able to refine (12). However, as the following result shows, inequality (12) is tight up to absolute numerical constants and thus cannot be improved further.

Theorem 3.1. *Let $(\mathbf{A}_i)_{i\in[m]}$ be independent $d\times d$ symmetric random matrices, where each $\mathbf{A}_i$ has independent entries with distribution $\mathcal{N}(0,1)$ on the diagonal and $\mathcal{N}(0,1/2)$ on the off-diagonal entries. Assume $d \ge 6$, $m \ge C_0$ for some universal constant $C_0 > 0$, and $r \le \frac{d}{16}$. Then, with probability at least $1 - 2\exp(-\frac{m}{32}) - 2\exp(-\frac{d}{32})$, it holds that*

$$\sup_{\mathbf{Z}\in\mathcal{T}_r} \left\| (\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathbf{Z})\right\| \ge \frac{1}{16}\sqrt{\frac{r^2 d}{m}}.$$

Theorem 3.1 shows that we will need to use different proof techniques to establish a bound similar to (11). In particular, we cannot rely on uniform concentration inequalities. These novel techniques will be introduced in Section 3.2 below. Before that, we want to prove Theorem 3.1.

Proof. First, we note that

$$\sup_{\mathbf{Z}\in\mathcal{T}_r} \left\| \left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)(\mathbf{Z}) \right\| = \sup_{\mathbf{Z}\in\mathcal{T}_r} \left\| \frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{Z}\rangle \mathbf{A}_i - \mathbf{Z} \right\| = \sup_{\|\mathbf{u}\|=1} \sup_{\mathbf{Z}\in\mathcal{T}_r} \left| \langle \frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{Z}\rangle \mathbf{A}_i - \mathbf{Z}, \mathbf{u}\mathbf{u}^\top \rangle \right|.$$

Now for any fixed $\mathbf{u}\in\mathbb{R}^d$ with $\left\|\mathbf{u}\right\|_2 = 1$, define

$$\mathcal{T}_{\mathbf{u}} := \left\{ \mathbf{Z}\in\mathbb{R}^{d\times d} : \mathbf{Z}=\mathbf{Z}^\top, \operatorname{rank}(\mathbf{Z})\le r, \left\|\mathbf{Z}\right\|\le 1, \mathbf{Z}\mathbf{u}=0 \right\},$$

i.e., the set consisting of matrices in $\mathcal{T}_r$, whose row space is orthogonal to $\mathbf{u}$. It follows that

$$\begin{aligned}
\sup_{\mathbf{Z}\in\mathcal{T}_r} \left\| \frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{Z}\rangle \mathbf{A}_i - \mathbf{Z} \right\| &\ge \sup_{\mathbf{Z}\in\mathcal{T}_{\mathbf{u}}} \langle \frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{Z}\rangle \mathbf{A}_i - \mathbf{Z}, \mathbf{u}\mathbf{u}^\top \rangle \\
&= \sup_{\mathbf{Z}\in\mathcal{T}_{\mathbf{u}}} \langle \frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{Z}\rangle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top \rangle \\
&= \sup_{\mathbf{Z}\in\mathcal{T}_{\mathbf{u}}} \frac{1}{m}\sum_{i=1}^m \langle \langle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top\rangle \mathbf{A}_i, \mathbf{Z}\rangle.
\end{aligned}$$

Now note that $\langle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top\rangle$ is independent of $(\langle \mathbf{A}_i, \mathbf{Z}\rangle)_{\mathbf{Z}\in\mathcal{T}_{\mathbf{u}}}$. Let $\mathbf{A}\in\mathbb{R}^{d\times d}$ be a matrix with the same distribution as $\mathbf{A}_i$ and which is independent of $(\mathbf{A}_i)_{i=1}^m$. We claim that conditional on $\left\{\langle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top\rangle\right\}_{i=1}^m$ we have that the following two random variables are equal in distribution:

$$\sup_{\mathbf{Z}\in\mathcal{T}_{\mathbf{u}}} \frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top\rangle \langle \mathbf{A}_i, \mathbf{Z}\rangle \overset{d}{=} \frac{1}{\sqrt{m}} \sqrt{\frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top\rangle^2} \sup_{\mathbf{Z}\in\mathcal{T}_{\mathbf{u}}} \langle \mathbf{A}, \mathbf{Z}\rangle. \tag{13}$$

To show (13), one can check that conditional on $\left\{\langle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top\rangle\right\}_{i=1}^m$, the random variables on both sides of (13) are the supremum of Gaussian processes indexed by $\mathcal{T}_{\mathbf{u}}$ with the same covariance structure, so they have the same distribution.

In the following, we set

$$\mathbf{u} := (0,\dots,0,1)^\top \in \mathbb{R}^d. \tag{14}$$

It follows that

$$\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top\rangle^2 = \sum_{i=1}^m (\mathbf{A}_i)_{d,d}^2 .$$

By Lipschitz concentration for Gaussian random variables [54, Theorem 5.6], we obtain

$$\mathbb{P}\left(\left| \sqrt{\sum_{i=1}^m (\mathbf{A}_i)_{d,d}^2} - \mathbb{E}\sqrt{\sum_{i=1}^m (\mathbf{A}_i)_{d,d}^2} \right| \ge \sqrt{m}/4 \right) \le 2\exp(-m/32).$$

This shows that with probability at least $1-2\exp(-m/32)$,

$$\sqrt{\sum_{i=1}^m (\mathbf{A}_i)_{d,d}^2} \ge \mathbb{E}\sqrt{\sum_{i=1}^m (\mathbf{A}_i)_{d,d}^2} - \frac{\sqrt{m}}{4} \ge \sqrt{m}/2 \tag{15}$$

for sufficiently large m, where we have used that the expectation of chi-distribution with parameter m has asymptotic value $\sqrt{m-\frac{1}{2}}$ (see, e.g., [55]). In addition, with $\mathbf{u}$ given in (14), all entries in the d-th row and d-th column of the matrix $\mathbf{Z}\in\mathcal{T}_{\mathbf{u}}$ are equal to zero. Let $\tilde{\mathbf{A}}\in\mathbb{R}^{(d-1)\times(d-1)}$ be the

submatrix $\mathbf{A}$ where the last row and column of $\mathbf{A}$ are removed, and define $\tilde{\mathbf{Z}}$ in the same way. Then we have

$$\sup_{\mathbf{Z}\in\mathcal{T}_{\mathbf{u}}} \langle \mathbf{A}, \mathbf{Z}\rangle = \sup_{\|\tilde{\mathbf{Z}}\|\leq 1, \tilde{\mathbf{Z}}=\tilde{\mathbf{Z}}^\top,\, \mathrm{rank}(\tilde{\mathbf{Z}})\leq r} \langle \tilde{\mathbf{A}}, \tilde{\mathbf{Z}}\rangle = \sum_{i=1}^{r} \sigma_i(\tilde{\mathbf{A}}).$$

Our goal is to bound the sum of singular values on the right-hand side from below. For that, we define the matrix

$$\hat{\mathbf{A}} := \begin{pmatrix} \mathbf{0}_{(\lceil (d-1)/2\rceil -1)\times r} & \mathbf{0}_{\lceil (d-1)/2\rceil \times (d-r)} \\ \tilde{\mathbf{A}}_{\lceil (d-1)/2\rceil :(d-1),1:r} & \mathbf{0}_{(d-1-\lceil (d-1)/2\rceil)\times (d-r)} \end{pmatrix} \in \mathbb{R}^{(d-1)\times(d-1)}.$$

Here, $\tilde{\mathbf{A}}_{\lceil (d-1)/2\rceil :(d-1),1:r}$ denotes the submatrix of $\mathbf{A}$ obtained by restricting $\mathbf{A}$ to the $\lceil (d-1)/2\rceil$-th to $(d-1)$-th rows and the first r columns. By $\mathbf{0}_{a\times b}$ we denote the zero matrix of size a times b. To relate the singular values of $\tilde{\mathbf{A}}$ with the singular values of $\hat{\mathbf{A}}$, we will use the following lemma.

Lemma 3.2 (Corollary 3.1.3 in [56]). *Let $\mathbf{A}\in\mathbb{R}^{(d-1)\times(d-1)}$ and let $\mathbf{B}\in\mathbb{R}^{(d-1)\times(d-1)}$ be a matrix which is obtained from the matrix $\mathbf{A}$ by setting the entries of one row or one column to zero. Then it holds that $\sigma_i(\mathbf{B})\leq\sigma_i(\mathbf{A})$ for all $i=1,\dots,d-1$.*

By repeatedly applying Lemma 3.2, we find

$$\sum_{i=1}^{r}\sigma_i(\hat{\mathbf{A}}) \leq \sum_{i=1}^{r}\sigma_i(\tilde{\mathbf{A}}).$$

On the other hand, we can identify the r largest singular singular values of $\hat{\mathbf{A}}$ with the singular values of a Gaussian matrix of size $\lfloor \frac{d-1}{2}\rfloor \times r$. By standard concentration inequalities for the singular values of Gaussian matrices, see, e.g., [57, Corollary 5.35], we find that with probability at least $1-2\exp(-t^2/2)$,

$$\sigma_r(\hat{\mathbf{A}}) \geq \sqrt{\Big\lfloor \frac{d-1}{2}\Big\rfloor} - \sqrt{r} - t.$$

Taking $t=\frac{\sqrt{d}}{8}$, and using the assumption that $r\leq\frac{d}{16}$, we find for $d\geq 6$,

$$\sum_{i=1}^{r}\sigma_i(\tilde{\mathbf{A}}) \geq \frac{r\sqrt{d}}{8} \tag{16}$$

with probability at least $1-2\exp(-d/32)$. Combining (16) and (15) finishes the proof. □

Note that the key idea in this proof was to fix a vector $\mathbf{u}\in\mathbb{R}^d$ and to pick a matrix $\mathbf{Z}\in\mathcal{T}_r$ based on eigenvectors corresponding to the largest eigenvalues (of a submatrix) of

$$\mathbf{A} = \frac{1}{m}\sum_{i=1}^{m}\langle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top\rangle \mathbf{A}_i.$$

By design, this implies that the matrix $\mathbf{Z}$ was chosen in a way which strongly depends on $\big(\langle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top\rangle\big)_{i=1}^m$. This observation leads to the key idea in our proof. Namely, we will show that our gradient descent iterates $\mathbf{U}_t$ depend, in a suitable sense, only weakly $\big(\langle \mathbf{A}_i, \mathbf{u}\mathbf{u}^\top\rangle\big)_{i=1}^m$ for fixed $\mathbf{u}\in\mathbb{R}^d$. This will allow us to prove stronger upper bounds for the term $\big\|\,(\mathcal{A}^*\mathcal{A}-\mathcal{I})\,(\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top)\,\big\|$ than what can be achieved using uniform concentration inequalities.

3.2. Virtual sequences

As explained at the end of Section 3.1, we aim to establish that the gradient descent iterates $(\mathbf{U}_t)_t$ depend only weakly on $\big(\langle \mathbf{A}_i, \mathbf{w}\mathbf{w}^\top\rangle\big)_{i=1}^m$ in a suitable sense. For this aim, we will use so-called *virtual sequences* $(\mathbf{U}_{t,\mathbf{w}})_{t\in\mathbb{N}}\subset\mathcal{S}^d$. The central idea is to introduce for $\mathbf{w}\in S^{d-1}:=\big\{\mathbf{x}\in\mathbb{R}^d : \|\mathbf{x}\|_2=1\big\}$ a sequence with the following two properties.

1. The sequence $(\mathbf{U}_{t,\mathbf{w}})_{t\in\mathbb{N}}$ is stochastically independent of $\big(\langle \mathbf{A}_i, \mathbf{w}\mathbf{w}^\top\rangle\big)_{i=1}^m$.
2. The sequence $(\mathbf{U}_{t,\mathbf{w}})_{t\in\mathbb{N}}$ stays sufficiently close to the sequence $(\mathbf{U}_t)_{t\in\mathbb{N}}$. More precisely, we require that $\big\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|_F$ stays sufficiently small.

The sequences $(\mathbf{U}_{t,\mathbf{w}})_{t\in\mathbb{N}}$ are called *virtual* since they are introduced solely for proof purposes.

Remark 3.3 (Related work)**.** *In the context of non-convex optimization, the use of virtual sequences has been pioneered in the influential works [18] and [33]. In these works, leave-one-out sequences, which can be seen as a special case of virtual sequences, were introduced to show that the gradient descent iterates depend only weakly on the individual samples or measurements. These works lead to a number of follow-up works. For example, several works used virtual sequences to establish convergence from random initialization for gradient descent in phase retrieval [58] or for alternating minimization in rank-one matrix sensing [59]. In [27], leave-one-out sequences were used to establish that in overparameterized matrix completion gradient descent with small random initialization converges to the ground truth. Similar to the paper at hand, the virtual sequence argument was combined with an ε-net argument. However, the technical details are arguably quite different.*

Before defining the virtual sequences we recall the notion of an ε-net.

Definition 3.4 (ε-net)**.** *Let $A \subset \mathbb{R}^d$. A subset $B \subset A$ is called ε-net of A if for every $\mathbf{x} \in A$ there is a point $\mathbf{x}_0 \in B$ such that $\big\|\mathbf{x} - \mathbf{x}_0\big\|_2 \leq \varepsilon$.*

It is well-known that for $S^{d-1} = \big\{\mathbf{x} \in \mathbb{R}^d : \big\|\mathbf{x}\big\|_2 = 1\big\}$ there exists an ε-net $\mathcal{N}_\varepsilon \subset S^{d-1}$ with cardinality $|\mathcal{N}_\varepsilon| \leq (3/\varepsilon)^d$ [60]. In the remainder of this paper, we will assume that $\mathcal{N}_\varepsilon$ is a fixed ε-net of S^{d-1} with $\varepsilon = 1/2$ such that $|\mathcal{N}_\varepsilon| \leq 6^d$. We will define one virtual sequence $(\mathbf{U}_{t,\mathbf{w}})_t$ for each $\mathbf{w} \in \mathcal{N}_\varepsilon$.

Recall from equation (9) that for $\mathbf{w} \in \mathcal{N}_\varepsilon$ the orthogonal projection operators $\mathcal{P}_{\mathbf{w}\mathbf{w}^\top}$ and $\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}$ were defined for $\mathbf{Z} \in \mathcal{S}^d$ via

$$\mathcal{P}_{\mathbf{w}\mathbf{w}^\top}(\mathbf{Z}) = \langle \mathbf{w}\mathbf{w}^\top, \mathbf{Z}\rangle \mathbf{w}\mathbf{w}^\top, \quad \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z}) = \mathbf{Z} - \langle \mathbf{w}\mathbf{w}^\top, \mathbf{Z}\rangle \mathbf{w}\mathbf{w}^\top.$$

Next, for $\mathbf{w} \in \mathcal{N}_\varepsilon$ we define the modified measurement matrices via

$$\mathbf{A}_{i,\mathbf{w}} := \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{A}_i) = \mathbf{A}_i - \langle \mathbf{w}\mathbf{w}^\top, \mathbf{A}_i\rangle \mathbf{w}\mathbf{w}^\top.$$

Thus, the matrix $\mathbf{A}_{i,\mathbf{w}}$ is obtained from the matrix $\mathbf{A}_i$ by setting the generalized entry $\langle \mathbf{A}_i, \mathbf{w}\mathbf{w}^\top\rangle$ equal to 0. We observe that by definition the matrices $(\mathbf{A}_{i,\mathbf{w}})_{i=1}^m$ are stochastically independent of $\big(\langle \mathbf{A}_i, \mathbf{w}\mathbf{w}^\top\rangle\big)_{i=1}^m$. We define the virtual measurement operator $\mathcal{A}_\mathbf{w} : \mathcal{S}^d \to \mathbb{R}^{m+1}$ via

$$[\mathcal{A}_\mathbf{w}(\mathbf{Z})]_i := \frac{1}{\sqrt{m}} \langle \mathbf{A}_i, \mathbf{X}\rangle$$

for $i \in [m]$ and

$$[\mathcal{A}_\mathbf{w}(\mathbf{Z})]_{m+1} := \langle \mathbf{w}\mathbf{w}^\top, \mathbf{Z}\rangle.$$

Again, we observe that by construction, the measurement operator $\mathcal{A}_\mathbf{w}$ is independent of $\big(\langle \mathbf{A}_i, \mathbf{w}\mathbf{w}^\top\rangle\big)_{i=1}^m$. As a next step, analogously to the definition of the objective function $\mathcal{L}$, we can define the modified objective function $\mathcal{L}_\mathbf{w} : \mathcal{S}^d \to \mathbb{R}$ via

$$\mathcal{L}_\mathbf{w}(\mathbf{U}) := \frac{1}{4}\big\|\mathcal{A}_\mathbf{w}\left(\mathbf{X}_\star - \mathbf{U}\mathbf{U}^\top\right)\big\|_2^2.$$

With these definitions in place, the virtual sequence $(\mathbf{U}_{t,\mathbf{w}})_t$ can be defined analogously to the original sequence $(\mathbf{U}_t)_t$. Namely, to define the spectral initialization, we consider the eigendecomposition

$$(\mathcal{A}_\mathbf{w}^*\mathcal{A}_\mathbf{w})(\mathbf{X}_\star) =: \widetilde{\mathbf{V}}_\mathbf{w}\widetilde{\mathbf{\Lambda}}_\mathbf{w}\widetilde{\mathbf{V}}_\mathbf{w}^\top. \tag{17}$$

Then, analogously as for the original spectral initialization $\mathbf{U}_0$, the matrix $\mathbf{U}_{0,\mathbf{w}}$ is defined as

$$\mathbf{U}_{0,\mathbf{w}} =: \widetilde{\mathbf{V}}_{r,\mathbf{w}}\widetilde{\mathbf{\Lambda}}_{r,\mathbf{w}}^{1/2}. \tag{18}$$

Then the virtual sequence $\{\mathbf{U}_{t,\mathbf{w}}\}_{t\in\mathbb{N}}$ via

$$\mathbf{U}_{t+1,\mathbf{w}} := \mathbf{U}_{t,\mathbf{w}} - \mu\nabla\mathcal{L}_{\mathbf{w}}\left(\mathbf{U}_{t,\mathbf{w}}\right) = \mathbf{U}_{t,\mathbf{w}} + \mu\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}.$$

It follows directly from the definition of $\left(\mathbf{U}_{t,\mathbf{w}}\right)_t$ that this sequence is stochastically independent of $\left(\langle\mathbf{A}_i, \mathbf{w}\mathbf{w}^\top\rangle\right)_{i=1}^m$. At the end of this section, we state and prove the following lemma, which is a direct consequence of the definition of $\mathcal{A}_{\mathbf{w}}$. This lemma will be useful in the convergence analysis where we establish that $\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F$ stays sufficiently small.

Lemma 3.5. *For any symmetric matrix $\mathbf{Z}\in\mathbb{R}^{d\times d}$ it holds that*

$$\begin{aligned}\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top}(\mathbf{Z})\right) &= \mathcal{P}_{\mathbf{w}\mathbf{w}^\top}(\mathbf{Z}),\\ \left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\right) &= \left(\mathcal{A}^*\mathcal{A}\right)\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\right) - \langle\mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\right)\rangle\mathbf{w}\mathbf{w}^\top.\end{aligned}$$

Proof of Lemma 3.5. To prove the first inequality we note first that it follows directly from the definition of $\mathbf{A}_{i,\mathbf{w}}$ that $\langle\mathbf{A}_{i,\mathbf{w}}, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top}(\mathbf{Z})\rangle = 0$. It follows that

$$\begin{aligned}\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top}(\mathbf{Z})\right) &= \frac{1}{\sqrt{m}}\sum_{i=1}^m\left[\mathcal{A}_{\mathbf{w}}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top}(\mathbf{Z})\right)\right]_i\mathbf{A}_{i,\mathbf{w}} + \left(\mathcal{A}_{\mathbf{w}}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top}(\mathbf{Z})\right)\right)_{m+1}\mathbf{w}\mathbf{w}^\top\\ &= \frac{1}{m}\sum_{i=1}^m\langle\mathbf{A}_{i,\mathbf{w}}, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top}(\mathbf{Z})\rangle\mathbf{A}_{i,\mathbf{w}} + \langle\mathbf{w}\mathbf{w}^\top, \mathbf{Z}\rangle\mathbf{w}\mathbf{w}^\top\\ &= \langle\mathbf{w}\mathbf{w}^\top, \mathbf{Z}\rangle\mathbf{w}\mathbf{w}^\top.\end{aligned}$$

This proves the first equation. In order to prove the second equation, we note that

$$\begin{aligned}\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\right) &= \frac{1}{m}\sum_{i=1}^m\langle\mathbf{A}_{i,\mathbf{w}}, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\rangle\mathbf{A}_{i,\mathbf{w}} + \langle\mathbf{w}\mathbf{w}^\top, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\rangle\mathbf{w}\mathbf{w}^\top\\ &= \frac{1}{m}\sum_{i=1}^m\langle\mathbf{A}_{i,\mathbf{w}}, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\rangle\mathbf{A}_{i,\mathbf{w}}\\ &= \frac{1}{m}\sum_{i=1}^m\langle\mathbf{A}_i, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\rangle\mathbf{A}_{i,\mathbf{w}}\\ &= \frac{1}{m}\sum_{i=1}^m\langle\mathbf{A}_i, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\rangle\mathbf{A}_i - \frac{1}{m}\sum_{i=1}^m\langle\mathbf{A}_i, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\rangle\langle\mathbf{w}\mathbf{w}^\top, \mathbf{A}_i\rangle\mathbf{w}\mathbf{w}^\top\\ &= \left(\mathcal{A}^*\mathcal{A}\right)\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\right) - \langle\mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}))\rangle\mathbf{w}\mathbf{w}^\top.\end{aligned}$$

This proves the second equation. □

3.3. Upper bounds for the spectral norm of the deviation term

Recall that by construction, it holds for any $\mathbf{w}\in\mathcal{N}_\varepsilon$ that the sequence $\left(\mathbf{U}_{t,\mathbf{w}}\right)_{t=0,1,\ldots,T}$ is independent of $\left(\langle\mathbf{w}\mathbf{w}^\top, \mathbf{A}_i\rangle\right)_{i=1}^m$. This property allows us to establish the following key lemma which we will use several times throughout our proof.

Lemma 3.6. *Let $\mathcal{N}_\varepsilon$ be the ε-net with $\varepsilon = 1/2$ introduced in Section 3.2 which we used to construct the virtual sequences $\left(\mathbf{U}_{t,\mathbf{w}}\right)_t$. Assume that for the cardinality of $\mathcal{N}_\varepsilon$, we have that $|\mathcal{N}_\varepsilon| \leq 6^d$. Moreover, let $T\in\mathbb{N}$ such that $2T\leq 6^d$. Then, with probability at least $1 - 2\exp\left(-10d\right)$, it holds for all $\mathbf{w}\in\mathcal{N}_\varepsilon$ and all $1\leq t\leq T$ that*

$$\left|\langle\mathbf{w}\mathbf{w}^\top, \left(\mathcal{A}^*\mathcal{A}\right)\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right)\rangle\right| \leq 4\sqrt{\frac{d}{m}}\left\|\mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right)\right\|_2.$$

Proof. We introduce the shorthand

$$\mathbf{\Delta}_{t,\mathbf{w}} := \mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top.$$

Due to the definition of $\mathbf{A}_{i,\mathbf{w}}$ and due to the rotation invariance of the Gaussian distribution, $\{\mathbf{A}_{i,\mathbf{w}}\}_{i=1}^m$ and $\{\langle \mathbf{w}\mathbf{w}^\top, \mathbf{A}_i\rangle\}_{i=1}^m$ are independent. Moreover, note that by construction $\mathbf{\Delta}_{t,\mathbf{w}}$ is independent of $\{\langle \mathbf{w}\mathbf{w}^\top, \mathbf{A}_i\rangle\}_{i=1}^m$. Thus, it follows that $\{\langle \mathbf{w}\mathbf{w}^\top, \mathbf{A}_i\rangle\}_{i=1}^m$ is independent of $\{\langle \mathbf{A}_i, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{\Delta}_{t,\mathbf{w}})\rangle\}_{i=1}^m$. Moreover, the vector $(\langle \mathbf{w}\mathbf{w}^\top, \mathbf{A}_i\rangle)_{i=1}^m$ has i.i.d. entries with distribution $\mathcal{N}(0,1)$. Thus, we have for all $x>0$ with probability at least $1-2\exp(-x^2/2)$ (see [60, Proposition 2.1.2]) that

$$\begin{aligned}\left|\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{\Delta}_{t,\mathbf{w}})\right)\rangle\right| &= \left|\frac{1}{m}\sum_{i=1}^m \langle \mathbf{w}\mathbf{w}^\top, \mathbf{A}_i\rangle\langle \mathbf{A}_i, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{\Delta}_{t,\mathbf{w}})\rangle\right| \\ &\le \frac{x}{m}\sqrt{\sum_{i=1}^m \langle \mathbf{A}_i, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{\Delta}_{t,\mathbf{w}})\rangle^2} \\ &= \frac{x}{\sqrt{m}}\left\|\mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{\Delta}_{t,\mathbf{w}})\right)\right\|_2. \end{aligned} \tag{19}$$

Then, by applying inequality (19) with $x = C\sqrt{d}$ and by taking a union bound, it follows that with probability at least $1-\xi$ (over the whole probability space), we have for all $\mathbf{w}\in\mathcal{N}_\varepsilon$ and all $t\in[T]$ that

$$\left|\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{\Delta}_{t,\mathbf{w}})\right)\rangle\right| \le \frac{C\sqrt{d}}{\sqrt{m}}\left\|\mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{\Delta}_{t,\mathbf{w}})\right)\right\|_2,$$

where

$$\xi \le 2T|\mathcal{N}_\varepsilon|\exp\left(-C^2 d\right) \le 6^{2d}\exp\left(-C^2 d\right) = \exp\left(2d\log(6) - C^2 d\right).$$

The claim follows from choosing $C=4$. □

Recall that our goal was to derive an upper bound for the expression $\left\|\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|$. The following lemma provides such a bound for $1\le t\le T$. Here, $T\in\mathbb{N}$ is some fixed number of iterations, which will be specified later in the proof of our main result.

Proposition 3.7. *Let $\mathcal{N}_\varepsilon$ be the ε-net from above with $\varepsilon = 1/2$ which we used to construct the virtual sequences $(\mathbf{U}_{t,\mathbf{w}})_{t=0,1,\dots,T}$. Assume that the conclusion of Lemma 3.6 holds. Moreover, assume that the linear measurement operator $\mathcal{A}$ has the Restricted Isometry Property of order $2r+2$ with constant $\delta = \delta_{2r+2}\le 1$. Then it holds that for all $0\le t\le T$,*

$$\begin{aligned}\left\|\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\| \le{}& \left(16\sqrt{\frac{2rd}{m}} + 2\delta\right)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \\ &+ 4\left(\delta + 4\sqrt{\frac{d}{m}}\right)\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F.\end{aligned}$$

As already mentioned, in previous literature, the quantity $\left\|\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|$ was controlled via an upper bound of $\sup_{\mathbf{Z}\in\mathcal{T}_{2r}}\left\|\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)(\mathbf{Z})\right\|$, where $\mathcal{T}_{2r}$ is a set of all rank-$2r$ matrices with bounded operator norm. This requires a uniform concentration bound for all matrices of rank at most $2r$ with bounded spectral norm. As we have seen in Theorem 3.1, this argument necessarily leads to a multiplicative factor of $\sqrt{r^2d/m}$.

In contrast, Proposition 3.7 bounds $\left\|\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|$ by a sum of two terms. The first term can be controlled with sample complexity $m \gtrsim rd\kappa^2$ since we also have $\delta \lesssim \sqrt{rd/m}$, see Lemma 2.2. The second term is a uniform bound on the deviation of the "true" sequence from the "virtual" sequences. This term can be interpreted as a measure of how stable the sequence $(\mathbf{U}_t)_t$ are under perturbation of the generalized entries $(\langle \mathbf{A}_i, \mathbf{w}\mathbf{w}^\top\rangle)_{i=1}^m$ of the symmetric measurement matrices.

Proof of Proposition 3.7. We use the shorthand notation

$$\begin{aligned}\boldsymbol{\Delta}_t &:= \mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top,\\ \boldsymbol{\Delta}_{t,\mathbf{w}} &:= \mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top.\end{aligned}$$

Since $\mathcal{N}_\varepsilon$ is an ε-net of S^{d-1} with $\varepsilon = 1/2$ we obtain that

$$\left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_t)\right\| \le 2\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}|\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_t)\rangle|, \tag{20}$$

(see, e.g. [60, Lemma 4.4.1]). Then, for every $\mathbf{w}\in\mathcal{N}_\varepsilon$ using the triangle inequality we obtain that

$$\begin{aligned}|\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_t)\rangle| \le& |\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_{t,\mathbf{w}})\rangle| + |\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_{t,\mathbf{w}}-\boldsymbol{\Delta}_t)\rangle|\\ \le& |\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_{t,\mathbf{w}})\rangle| + \left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_{t,\mathbf{w}}-\boldsymbol{\Delta}_t)\right\|\\ \le& |\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_{t,\mathbf{w}})\rangle| + \delta\left\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}_{t,\mathbf{w}}\right\|_F.\end{aligned} \tag{21}$$

The last line is a consequence of the Restricted Isometry Property and Lemma 2.4, see inequality (8). To estimate the first summand further, we use the triangle inequality again, and we obtain that

$$\begin{aligned}&|\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_{t,\mathbf{w}})\rangle|\\ \le& |\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A}-\mathcal{I})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\boldsymbol{\Delta}_{t,\mathbf{w}})\right)\rangle| + |\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A}-\mathcal{I})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top}(\boldsymbol{\Delta}_{t,\mathbf{w}})\right)\rangle|\\ \overset{(a)}{=}& |\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\boldsymbol{\Delta}_{t,\mathbf{w}})\right)\rangle| + \left|\left(\left\|\mathcal{A}\left(\mathbf{w}\mathbf{w}^\top\right)\right\|_2^2-1\right)\langle \mathbf{w}\mathbf{w}^\top, \boldsymbol{\Delta}_{t,\mathbf{w}}\rangle\right|\\ \overset{(b)}{\le}& |\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\boldsymbol{\Delta}_{t,\mathbf{w}})\right)\rangle| + \delta|\langle \mathbf{w}\mathbf{w}^\top, \boldsymbol{\Delta}_{t,\mathbf{w}}\rangle|\\ \le& |\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\boldsymbol{\Delta}_{t,\mathbf{w}})\right)\rangle| + \delta\left\|\boldsymbol{\Delta}_{t,\mathbf{w}}\right\|.\end{aligned}$$

Equation (a) follows from the definition of $\mathcal{P}_{\mathbf{w}\mathbf{w}^\top}$ and $\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}$ and in inequality (b) we used the Restricted Isometry Property; see Definition 2.1. Thus, by combining the last estimate with inequalities (20) and (21) and taking the supremum over all $\mathbf{w}\in\mathcal{N}_\varepsilon$ we obtain that

$$\begin{aligned}&\left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_t)\right\|\\ \le& 2\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}|\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\boldsymbol{\Delta}_{t,\mathbf{w}})\right)\rangle| + 2\delta\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}\left\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}_{t,\mathbf{w}}\right\|_F + 2\delta\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}\left\|\boldsymbol{\Delta}_{t,\mathbf{w}}\right\|\\ \le& 2\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}|\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\boldsymbol{\Delta}_{t,\mathbf{w}})\right)\rangle| + 4\delta\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}\left\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}_{t,\mathbf{w}}\right\|_F + 2\delta\left\|\boldsymbol{\Delta}_t\right\|.\end{aligned} \tag{22}$$

Since we assumed that the conclusion of Lemma 3.6 holds we obtain for the first summand that

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}|\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}\left(\boldsymbol{\Delta}_{t,\mathbf{w}}\right)\right)\rangle| &\le 4\sqrt{\frac{d}{m}}\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}\left\|\mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\boldsymbol{\Delta}_{t,\mathbf{w}})\right)\right\|_2\\ &\overset{(a)}{\le} 8\sqrt{\frac{d}{m}}\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}\left\|\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\boldsymbol{\Delta}_{t,\mathbf{w}})\right\|_F\\ &\le 8\sqrt{\frac{d}{m}}\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}\left\|\boldsymbol{\Delta}_{t,\mathbf{w}}\right\|_F\\ &\le 8\sqrt{\frac{d}{m}}\left\|\boldsymbol{\Delta}_t\right\|_F + 8\sqrt{\frac{d}{m}}\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}\left\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}_{t,\mathbf{w}}\right\|_F\\ &\overset{(b)}{\le} 8\sqrt{\frac{2rd}{m}}\left\|\boldsymbol{\Delta}_t\right\| + 8\sqrt{\frac{d}{m}}\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}\left\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}_{t,\mathbf{w}}\right\|_F.\end{aligned}$$

Inequality (a) follows from the assumption that the operator $\mathcal{A}$ has the Restricted Isometry Property of order $2r+2$ with an RIP-constant $\delta\le 1$. To obtain inequality (b), we have used that the rank of $\boldsymbol{\Delta}_t$ is at most $2r$. Inserting the last estimate into (22), we obtain

$$\left\|\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)(\boldsymbol{\Delta}_t)\right\| \le \left(16\sqrt{\frac{2rd}{m}}+2\delta\right)\left\|\boldsymbol{\Delta}_t\right\| + 4\left(\delta+4\sqrt{\frac{d}{m}}\right)\sup_{\mathbf{w}\in\mathcal{N}_\varepsilon}\left\|\boldsymbol{\Delta}_t-\boldsymbol{\Delta}_{t,\mathbf{w}}\right\|_F.$$

Inserting the definition of $\boldsymbol{\Delta}_t$ and $\boldsymbol{\Delta}_{t,\mathbf{w}}$ yields the claim. □

4. Proof of the main result

4.1. Spectral Initialization

We provide the following lemma to show that both the original sequence and the virtual sequences are close to the ground truth $\mathbf{X}_\star$ at the spectral initialization. Moreover, this lemma guarantees that $\left\|\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right\|_F$ is sufficiently small. The proof of Lemma 4.1 is deferred to Appendix A.

Lemma 4.1. *There exists an absolute constant $C > 0$ such that the following holds:*

1. *With probability at least $1 - \exp(-4d)$, if $m > C^2\kappa^2 rd$ is satisfied, it holds that*

$$\left\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right\| \le C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}. \tag{23}$$

2. *With probability at least $1 - \exp(-2d)$, if $m > 4C^2\kappa^2 rd$ is satisfied, it holds for every $\mathbf{w} \in \mathcal{N}_\varepsilon$ that*

$$\left\|\mathbf{X}_\star - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right\| \le 2C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}. \tag{24}$$

   Consequently, if $m > 4C^2\kappa^2 rd$, with probability at least $1 - 2\exp(-2d)$, it holds for every $\mathbf{w} \in \mathcal{N}_\varepsilon$ that

$$\left\|\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right\| \le 3C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}. \tag{25}$$

3. *For any $\alpha \in (0,1)$, assume $m \ge \left(51C^2 + \frac{C_1}{\alpha^2}\right)\kappa^2 rd$ for an absolute constant $C_1 > 0$. With probability at least $1 - 4\exp(-d)$, for every $\mathbf{w} \in \mathcal{N}_\varepsilon$,*

$$\left\|\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right\|_F \le \left(2\alpha + C\kappa\sqrt{\frac{rd}{m}}\right)\left(2\sigma_{\min}(\mathbf{X}_\star) + 3\sqrt{2}C\kappa\sqrt{\frac{rd}{m}}\sigma_{\min}(\mathbf{X}_\star)\right) \tag{26}$$

4.2. Convergence Analysis

4.2.1. Outline of proof strategy

Before we explain our proof strategy, we want to recall the following convergence lemma which was proven in [10, Theorem 3.2] and [61]. It states that as soon as $\mathrm{dist}(\mathbf{U}_t, \mathbf{U}_\star)$ is small enough then $\mathrm{dist}(\mathbf{U}_t, \mathbf{U}_\star)$ converges to zero with linear rate. We state it in the version of the overview article [31, Theorem 4].

Lemma 4.2. *Assume that the measurement operator $\mathcal{A}$ satisfies the Restricted Isometry Property for all matrices of rank at most $6r$ with constant $\delta_{6r} < 1/10$. Let $\mathbf{U}_0, \mathbf{U}_1, \mathbf{U}_2, \ldots$ be a sequence of gradient descent iterates defined via equation (4). Assume that the step size satisfies $\mu \le \frac{c_1}{\|\mathbf{X}_\star\|}$ and*

$$dist^2\left(\mathbf{U}_T, \mathbf{U}_\star\right) \le \frac{1}{16}\sigma_{\min}(\mathbf{X}_\star) \tag{27}$$

for some iteration number T. Then it holds for all $t \ge T$ that

$$dist^2\left(\mathbf{U}_t, \mathbf{U}_\star\right) \le \left(1 - c_2\mu\sigma_{\min}(\mathbf{X}_\star)\right)^{t-T} dist^2(\mathbf{U}_T, \mathbf{U}_\star).$$

Here, $c_1, c_2 > 0$ are absolute numerical constants chosen small enough.

Note that the condition $\delta_{6r} < 1/10$ holds with high probability if the sample size satisfies $m \gtrsim rd$. However, condition (27) cannot be guaranteed for the spectral initialization, i.e., for $T = 0$, when $m \asymp rd\kappa^2$. For this reason, Lemma 4.2 is not directly applicable in our proof. To deal with this, we consider two different phases in our convergence analysis. Namely, we set

$$T := \left\lceil \frac{8}{\mu\sigma_{\min}\left(\mathbf{X}_\star\right)} \log\left(16r\right) \right\rceil.$$

We will show that at the end of the first phase, which consists of the iterations $t = 0, 1, \ldots, T$, condition (27) holds. The second phase starts at iteration T. For the second phase, we have established that condition (27) already holds we can directly apply Lemma 4.2 and we obtain linear convergence. Thus, our main focus in this section will be to analyze the first convergence phase.

In the following, we will give an outline of the analysis of this first phase. As is typical in the analysis of non-convex optimization algorithms, we will control several quantities simultaneously in each iteration via an induction argument. The following list contains an overview of these.

a) We will show that $\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F$ and $\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F$ stay sufficiently small for each $\mathbf{w} \in \mathcal{N}_\varepsilon$. Together with Proposition 3.7, this allows us to control the deviation term $\left\|\left(\mathcal{I} - \mathcal{A}^*\mathcal{A}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|$.

b) We will show that for each iteration $t \in [T]$ it holds that $\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \leq c\sigma_{\min}(\mathbf{X}_\star)$ for some small constant $c > 0$. This ensures that the gradient descent iterates stay in the basin of attraction, in which we can establish linear convergence.

c) We will establish that $\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F$ decays linearly in each iteration. Combined with the result from b) this will allow us to establish linear convergence of $\operatorname{dist}(\mathbf{U}_t, \mathbf{U}_\star)$.

The remainder of this section is structured as follows. In Section 4.2.2 we will provide the technical lemmas to control $\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F$ and $\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F$ as described in a) above. In Section 4.2.3, we will provide the technical lemmas which allow us to control the quantities described above in b) and c). In Section 4.2.4, we will combine these ingredients to prove Proposition 4.10, which is our main result describing the convergence of the iterates $(\mathbf{U}_t)_{0\leq t\leq T}$ in the first convergence phase.

4.2.2. Lemmas for controlling the distance between the virtual sequences and the original sequence

The goal of this section is to show that the virtual sequence iterates $(\mathbf{U}_{t,\mathbf{w}})_t$ stay sufficiently close to the original sequence $(\mathbf{U}_t)_t$. This will be established via induction. In the following, we will state all key lemmas. To keep the presentation concise, we have moved the proofs, which may be of independent interest, to Section B.

The first lemma in this section provides an a priori estimate. Its proof can be found in Section B.2.

Lemma 4.3. *For absolute constants $c_1, c_2, c_3 > 0$ chosen small enough the following statement is true. Let $\mathbf{w} \in \mathcal{N}_\varepsilon$ and assume that*

$$\left\|\mathbf{U}_t\right\| \leq \sqrt{2\left\|\mathbf{X}_\star\right\|}, \tag{28}$$

$$\left\|\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\| \leq c_1\sigma_{\min}\left(\mathbf{X}_\star\right), \tag{29}$$

$$\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \leq \sigma_{\min}(\mathbf{X}_\star), \tag{30}$$

$$\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \leq \frac{\sigma_{\min}\left(\mathbf{X}_\star\right)}{80}, \tag{31}$$

and that the step size $\mu > 0$ satisfies $\mu \leq \frac{c_2}{\kappa\left\|\mathbf{X}_\star\right\|}$. In addition, assume that the conclusions of Lemma 3.6 hold and that

$$\max\left\{\delta; 8\sqrt{\frac{rd}{m}}\right\} \leq \frac{c_3}{\kappa}, \tag{32}$$

where $\delta = \delta_{4r+1}$ denotes the Restricted Isometry Property of rank $4r+1$. Then it holds that

$$\left\|\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\right\|_F \leq \frac{\sqrt{\sqrt{2}-1}}{40}\sigma_{\min}(\mathbf{X}_\star).$$

Under the assumption that this a priori estimate holds, the next lemma shows that the quantity $\left\|\mathbf{U}_t\mathbf{U}_t - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F$ can be bounded from above by the quantity $\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F$. The proof of this lemma has been deferred to Section B.3.

Lemma 4.4. *Let $\mathbf{w} \in \mathcal{N}_\varepsilon$ and assume that*

$$\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right\| \leq \frac{\sigma_{\min}\left(\mathbf{X}_\star\right)}{1600}, \tag{33}$$

$$\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \leq \frac{\sqrt{3\left(\sqrt{2}-1\right)}\cdot\sigma_{\min}\left(\mathbf{X}_\star\right)}{40}. \tag{34}$$

Then it holds that

$$\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|_F \leq \frac{3\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F}{5}. \tag{35}$$

Moreover, it holds that

$$\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \leq 3\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F. \tag{36}$$

The following key lemma allows us to control $\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F$ iteratively. Its proof can be found in Section B.4.

Lemma 4.5. *For sufficiently small absolute constants $c_1, c_2, c_3, c_4, c_5, c_6 > 0$ the following statement holds. Let $\mathbf{w} \in \mathcal{N}_\varepsilon$ and assume that*

$$\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\right\| \leq c_1, \tag{37}$$

$$\|\mathbf{U}_t\| \leq \sqrt{2\|\mathbf{X}_\star\|}, \tag{38}$$

$$\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right\| \leq c_2\sigma_{\min}(\mathbf{X}_\star), \tag{39}$$

$$\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \leq c_3\sigma_{\min}\left(\mathbf{X}_\star\right). \tag{40}$$

Moreover, assume that the step size satisfies $\mu \leq \frac{c_4}{\kappa\left\|\mathbf{X}_\star\right\|}$. Furthermore, assume that the conclusion of Lemma 3.6 holds and that

$$\left\|\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\| \leq c_5\sigma_{\min}(\mathbf{X}_\star), \tag{41}$$

$$\max\left\{\delta; 8\sqrt{\frac{2rd}{m}}\right\} \leq \frac{c_6}{\kappa}, \tag{42}$$

where $\delta = \delta_{4r+2}$ denotes the Restricted Isometry Constant of rank $4r+2$. Then, it holds that

$$\begin{aligned}&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\right)\right\|_F\\ &\leq \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + \mu\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|.\end{aligned}$$

4.2.3. Lemmas controlling the distance between $\mathbf{X}_\star$ and $\mathbf{U}_t\mathbf{U}_t^\top$

In the following, let $|||\cdot|||$ denote any matrix norm, which satisfies the inequality

$$|||\mathbf{XYZ}||| \leq \|\mathbf{X}\|\,|||\mathbf{Y}|||\,\|\mathbf{Z}\| \tag{43}$$

for all matrices $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ with dimensions such that the matrix product $\mathbf{XYZ}$ is well-defined. Note that all Schatten-p norms have this property. In particular, this includes the spectral norm $\|\cdot\|$ and the Frobenius norm $\|\cdot\|_F$.

In the following, we are interested in establishing upper bounds for $\left|\left|\left|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right|\right|\right|$, where either $|||\cdot||| = \|\cdot\|_F$ or $|||\cdot||| = \|\cdot\|$. Instead of estimating these quantities directly, we will instead derive upper bounds for the quantity

$$\left|\left|\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right|\right|\right|.$$

To be able to relate this quantity with $\left|\left|\left|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right|\right|\right|$ one can then use the following lemma.

Lemma 4.6. *Let $|||\cdot|||$ be a norm for which inequality (43) holds. Assume that*

$$\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \mathbf{V}_{\mathbf{U}_t}\right\| \leq \frac{1}{\sqrt{2}}. \tag{44}$$

Then the following inequalities hold:

$$\left|||\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \mathbf{U}_t\mathbf{U}_t^\top \mathbf{V}_{\mathbf{X}_\star,\perp}|||\right| \leq 2\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \mathbf{V}_{\mathbf{U}_t}\right\| \left|||\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star,\perp}|||\right|, \tag{45}$$

$$\left|||\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star|||\right| \leq 2\left(1 + \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \mathbf{V}_{\mathbf{U}_t}\right\|\right) \left|||\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)|||\right|. \tag{46}$$

A comparable lemma was proven in [23] in a more general setting but with less explicit constants. For the sake of completeness, we included in Appendix C.1.

The following lemma allows us to control the quantity $\left|||\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)|||\right|$ iteratively. We note that a similar lemma has already been proven in [23] in a more general setting with less explicit constants. For the sake of completeness, we again included a proof in Appendix C.2.

Lemma 4.7. *Let $|||\cdot|||$ be a norm which is submultiplicative in the sense of inequality (43). Assume that*

$$\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \mathbf{V}_{\mathbf{U}_t}\right\| \leq \frac{1}{2}, \tag{47}$$

$$\left\|\mathbf{U}_t\right\| \leq \sqrt{2\left\|\mathbf{X}_\star\right\|},$$

$$\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \leq \frac{\sigma_{\min}(\mathbf{X}_\star)}{48}, \tag{48}$$

$$\left\| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) \right\| \leq \frac{1}{48}\sigma_{\min}\left(\mathbf{X}_\star\right), \tag{49}$$

and that the step size satisfies $\mu \leq \frac{1}{1024\kappa\left\|\mathbf{X}_\star\right\|}$. Then it holds that

$$\begin{aligned}
&\left|||\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{X}_\star\right)|||\right| \\
&\leq \left(1 - \frac{\mu}{8}\sigma_{\min}\left(\mathbf{X}_\star\right)\right) \left|||\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)|||\right| + 5\mu\left\|\mathbf{X}_\star\right\| \left|||\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_t}|||\right|.
\end{aligned}$$

Given an upper bound for $\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|_F$ we can obtain an estimate for $\operatorname{dist}\left(\mathbf{U}_t, \mathbf{U}_\star\right)$ by using the following technical lemma.

Lemma 4.8 (Lemma 5.4 in [10])**.** *Let $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d\times r}$ be two matrices and assume that rank$(\mathbf{U}) = \min\{r; d\}$. Then it holds that*

$$\operatorname{dist}^2\left(\mathbf{U}, \mathbf{V}\right) \leq \frac{1}{2(\sqrt{2}-1)\sigma_{\min}^2(\mathbf{U})}\left\|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\right\|_F^2,$$

where $\operatorname{dist}\left(\mathbf{U}, \mathbf{V}\right)$ is defined in (5).

To check the prerequisite of the Davis-Kahan inequality (Lemma 2.6) in our proof, we will also need the following auxiliary lemma, which provides us with an a priori bound for $\left\|\mathbf{X}_\star - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top\right\|$. Its proof can be found in Appendix C.3.

Lemma 4.9. *There are absolute constants $c_1, c_2, c_3 > 0$ such that the following holds. Assume that $\mu \leq \frac{c_1}{\left\|\mathbf{X}_\star\right\|}$ and*

$$\left\|\mathbf{U}_t\right\| \leq \sqrt{2\left\|\mathbf{X}_\star\right\|}, \tag{50}$$

$$\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \leq c_2\sigma_{\min}(\mathbf{X}_\star), \tag{51}$$

$$\left\| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) \right\| \leq c_3\sigma_{\min}\left(\mathbf{X}_\star\right). \tag{52}$$

Then it holds that

$$\left\|\mathbf{X}_\star - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top\right\| \leq \left(1 - \frac{1}{\sqrt{2}}\right)\sigma_{\min}\left(\mathbf{X}_\star\right).$$

4.2.4. Statement and proof of the main convergence lemma

We now have all the ingredients in place to prove the main lemma in this section, which is stated below.

Lemma 4.10. *There are absolute constants $c_1, c_2, c_3, c_4 > 0$ chosen sufficiently small such that the following statement holds. Assume that the spectral initialization $\mathbf{U}_0$ satisfies*

$$\left\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right\| \leq c_1\sigma_{\min}\left(\mathbf{X}_\star\right) \tag{53}$$

and that for every $\mathbf{w} \in \mathcal{N}_\varepsilon$ we have that

$$\left\|\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right\|_F \leq c_2\sigma_{\min}\left(\mathbf{X}_\star\right). \tag{54}$$

Moreover, we assume that the conclusion of Lemma 3.6 holds for

$$T = \left\lceil \frac{8}{\mu\sigma_{\min}\left(\mathbf{X}_\star\right)} \log\left(16r\right) \right\rceil.$$

Furthermore, we assume that

$$\max\left\{\delta; 8\sqrt{\frac{2rd}{m}}\right\} \leq \frac{c_3}{\kappa}, \tag{55}$$

where $\delta = \delta_{4r+2}$ denotes the Restricted Isometry Property of order $4r+2$. In addition, assume that $\mu \leq \frac{c_4}{\kappa\left\|\mathbf{X}_\star\right\|}$. Then for every iteration t with $0 \leq t \leq T$ it holds that

$$\mathit{dist}^2\left(\mathbf{U}_t, \mathbf{U}_\star\right) \leq r\left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)^{2t} \left\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right\|. \tag{56}$$

In particular, we have that

$$\mathit{dist}^2\left(\mathbf{U}_T, \mathbf{U}_\star\right) \leq \frac{1}{16}\sigma_{\min}(\mathbf{X}_\star), \tag{57}$$

where $\mathbf{U}_\star \in \mathbb{R}^{n\times r}$ denotes a matrix which satisfies $\mathbf{U}_\star\mathbf{U}_\star^\top = \mathbf{X}_\star$.

Proof of Lemma 4.10. We prove by induction that for all iterations t with $0 \leq t \leq T$ the following inequalities hold:

$$\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F \leq \left(1 - \frac{\mu}{16}\sigma_{\min}(\mathbf{X}_\star)\right)^t \left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right)\right\|_F, \tag{58}$$

$$\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\| \leq c_1\sigma_{\min}\left(\mathbf{X}_\star\right), \tag{59}$$

$$\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\right\| \leq \sqrt{2}c_1, \tag{60}$$

$$\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \leq 3c_1\sigma_{\min}\left(\mathbf{X}_\star\right), \tag{61}$$

and, for every $\mathbf{w} \in \mathcal{N}_\varepsilon$,

$$\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \leq c_2\sigma_{\min}\left(\mathbf{X}_\star\right), \tag{62}$$

$$\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \leq 3c_2\sigma_{\min}(\mathbf{X}_\star). \tag{63}$$

The constants $c_1, c_2 > 0$ are the same as in assumptions (53) and (54) and are thus, in particular, independent of the iteration number t.

First, we check that these inequalities hold for $t = 0$. Inequality (58) is immediate. Inequalities (59) and (61) follow from assumption (53). Inequalities (62) and (63) are due to assumption (54). It remains to establish inequality (60) for $t = 0$. Using the Davis-Kahan inequality (see Lemma 2.6) and assumption (53) it follows that

$$\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_0}\right\| \leq \frac{\sqrt{2}\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right)\right\|}{\sigma_{\min}\left(\mathbf{X}_\star\right)} \leq \sqrt{2}c_1.$$

This shows that the above inequalities hold for $t = 0$.

For the induction step, assume now that these inequalities hold for some t. First, we observe that it follows from the induction assumption (61) and Weyl's inequalities that $\|\mathbf{U}_t\| \leq \sqrt{2\|\mathbf{X}_\star\|}$ for $c_1 < 1/3$. Moreover, note that since we assumed that the conclusion of Lemma 3.6 holds we obtain from Proposition 3.7 that

$$\begin{aligned}
&\left\| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) \right\| \\
&\leq \left(16\sqrt{\frac{2rd}{m}} + 2\delta\right)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + 4\left(\delta + 4\sqrt{\frac{d}{m}}\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\
&\overset{(a)}{\leq} \frac{4c_3}{\kappa}\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \frac{6c_3}{\kappa}\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\
&\overset{(b)}{\leq} \frac{10c_3}{\kappa}\sigma_{\min}\left(\mathbf{X}_\star\right), \qquad (64)
\end{aligned}$$

where inequality (a) follows from assumption (55). Inequality (b) is due to the induction hypotheses (61) and (63) with $c_1 \leq 1/3$ and $c_2 \leq 1/3$. Next, we note that from Lemma 4.7 applied with $|||\cdot||| = \|\cdot\|_F$ it follows that

$$\begin{aligned}
&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{X}_\star\right)\right\|_F \\
&\leq \left(1 - \frac{\mu}{8}\sigma_{\min}\left(\mathbf{X}_\star\right)\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F + 5\mu\|\mathbf{X}_\star\|\left\|\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_t}\right\|_F \\
&\overset{(a)}{\leq} \left(1 - \frac{\mu}{8}\sigma_{\min}\left(\mathbf{X}_\star\right)\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F + 5\mu\delta\|\mathbf{X}_\star\|\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|_F \\
&\overset{(b)}{\leq} \left(1 - \frac{\mu}{8}\sigma_{\min}\left(\mathbf{X}_\star\right)\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F + 15\mu\delta\|\mathbf{X}_\star\|\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F \\
&\overset{(c)}{\leq} \left(1 - \frac{\mu}{8}\sigma_{\min}\left(\mathbf{X}_\star\right)\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F + \frac{15\mu c_3\|\mathbf{X}_\star\|}{\kappa}\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F \\
&\overset{(d)}{\leq} \left(1 - \frac{\mu}{16}\sigma_{\min}\left(\mathbf{X}_\star\right)\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F.
\end{aligned}$$

Inequality (a) follows from the Restricted Isometry Property combined with Lemma 2.4. Inequality (b) is due to Lemma 4.6 and inequality (60). Inequality (c) follows from assumption (55) and inequality (d) is due to the fact we can choose $c_3 \leq \frac{1}{240}$. Thus, using the induction assumption, we see that inequality (58) holds for $t+1$.

Next, our goal is to prove inequality (59) for $t+1$. For that, we note that it follows from Lemma 4.7 with $|||\cdot||| = \|\cdot\|$ that

$$\begin{aligned}
&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{X}_\star\right)\right\| \\
&\leq \left(1 - \frac{\mu}{8}\sigma_{\min}\left(\mathbf{X}_\star\right)\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\| + 5\mu\|\mathbf{X}_\star\|\left\|\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\| \\
&\overset{(a)}{\leq} \left(1 - \frac{\mu}{8}\sigma_{\min}\left(\mathbf{X}_\star\right)\right)c_1\sigma_{\min}(\mathbf{X}_\star) + 50c_3\mu\sigma_{\min}^2\left(\mathbf{X}_\star\right) \\
&\overset{(b)}{\leq} c_1\sigma_{\min}\left(\mathbf{X}_\star\right), \qquad (65)
\end{aligned}$$

where inequality (a) follows from the induction hypothesis (59) and inequality (64). Inequality (b) holds since we can choose c_1 and c_3 in such a way that $c_3 \leq \frac{c_1}{400}$. This proves inequality (59) for $t+1$.

We observe that Lemma 4.9 yields the a-priori bound

$$\left\|\mathbf{X}_\star - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top\right\| \leq \left(1 - \frac{1}{\sqrt{2}}\right)\sigma_{\min}\left(\mathbf{X}_\star\right).$$

Thus, we can apply the Davis-Kahan inequality (see Lemma 2.6) which together with inequality (65) yields that

$$\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \mathbf{V}_{\mathbf{U}_{t+1}}\right\| \leq \frac{\sqrt{2}\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{X}_\star\right)\right\|}{\sigma_{\min}\left(\mathbf{X}_\star\right)} \leq \sqrt{2}c_1.$$

This proves inequality (60) for $t+1$. Next, we apply Lemma 4.6 and (65) to obtain that

$$\begin{aligned}\left\|\mathbf{X}_\star - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top\right\| &\leq 2\left(1 + \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \mathbf{V}_{\mathbf{U}_{t+1}}\right\|\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top\right)\right\| \\ &\leq 3\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top\right)\right\| \leq 3c_1\sigma_{\min}(\mathbf{X}_\star),\end{aligned}$$

which proves inequality (61) for $t+1$.

Next, we can apply Lemma 4.5 since all assumptions are satisfied and it follows that

$$\begin{aligned}&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\right)\right\|_F \\ &\leq \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + \mu\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \\ &\overset{(a)}{\leq} \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)c_2\sigma_{\min}\left(\mathbf{X}_\star\right) + 3c_1\mu\sigma_{\min}^2(\mathbf{X}_\star) \\ &\overset{(b)}{\leq} c_2\sigma_{\min}\left(\mathbf{X}_\star\right). \end{aligned} \tag{66}$$

Inequality (a) is due to inequalities (61) and (62). Inequality (b) holds since we can choose that $c_1 \leq \frac{c_2}{48}$. This proves inequality (62).

Next, we want to prove inequality (63) for $t+1$. First, we apply Lemma 4.3 and we obtain for all $\mathbf{w} \in \mathcal{N}_\varepsilon$ the a-priori bound

$$\left\|\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\right\|_F \leq \frac{\sqrt{\sqrt{2}-1}}{40}\cdot\sigma_{\min}\left(\mathbf{X}_\star\right).$$

This allows us to apply Lemma 4.4 and we obtain for all $\mathbf{w} \in \mathcal{N}_\varepsilon$ the sharper bound

$$\left\|\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\right\|_F \leq 3\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\right)\right\|_F \overset{(66)}{\leq} 3c_2\sigma_{\min}\left(\mathbf{X}_\star\right),$$

which shows inequality (63) for $t+1$. This completes the induction step.

To complete the proof of Lemma 4.10 it remains to prove inequalities (56) and (57). For that, we first observe that

$$\begin{aligned}\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|_F &\overset{(a)}{\leq} 3\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F \\ &\overset{(b)}{\leq} 3\left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)^t\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right)\right\|_F \\ &\overset{(c)}{\leq} 3\sqrt{2r}\left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)^t\left\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right\|.\end{aligned}$$

Inequality (a) follows from Lemma 4.6 with $|||\cdot||| = \left\|\cdot\right\|_F$ which is applicable since we have shown by induction that (60) holds for $0 \leq t \leq T$. Inequality (b) holds since we have proven (58) for all $0 \leq t \leq T$. Inequality (c) holds since $\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top$ has rank at most $2r$. Thus, we can apply Lemma 4.8 and obtain that

$$\begin{aligned}\operatorname{dist}^2\left(\mathbf{U}_t, \mathbf{U}_\star\right) &\leq \frac{\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|_F^2}{2\left(\sqrt{2}-1\right)\sigma_{\min}\left(\mathbf{X}_\star\right)} \\ &\leq 18r\left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)^{2t}\cdot\frac{\left\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right\|^2}{2\left(\sqrt{2}-1\right)\sigma_{\min}(\mathbf{X}_\star)} \\ &\leq \frac{9c_1 r}{\left(\sqrt{2}-1\right)}\left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)^{2t}\left\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right\|,\end{aligned}$$

where in the last inequality, we have used assumption (53). This proves inequality (56) since $c_1 \leq \frac{\sqrt{2}-1}{9}$. Next, we note that for $t = T$, the above inequality yields that

$$\begin{aligned}\operatorname{dist}^2(\mathbf{U}_T, \mathbf{U}_\star) &\overset{(a)}{\leq} \frac{9c_1^2 r}{(\sqrt{2}-1)} \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)^{2T} \sigma_{\min}(\mathbf{X}_\star) \\ &\overset{(b)}{\leq} \frac{9c_1^2 r}{(\sqrt{2}-1)} \exp\left(\frac{-T\mu\sigma_{\min}(\mathbf{X}_\star)}{8}\right) \sigma_{\min}(\mathbf{X}_\star) \\ &\overset{(c)}{\leq} \frac{\sigma_{\min}(\mathbf{X}_\star)}{16}.\end{aligned}$$

In inequality (a), we have used again assumption (53). Inequality (b) is due to the elementary inequality $\ln(1+x) \leq x$ for $-1 < x$ and the assumption $\mu < \frac{c_4}{\kappa \|\mathbf{X}_\star\|}$ for sufficiently small $c_4 > 0$. Inequality (c) follows from $T = \left\lceil \frac{8}{\mu\sigma_{\min}(\mathbf{X}_\star)} \log(16r) \right\rceil$ (and from the fact that we can choose $c_1 \leq \frac{\sqrt{\sqrt{2}-1}}{3}$). This proves inequality (57). Thus, the proof of Lemma 4.10 is complete. □

4.3. Proof of Theorem 1.2

Now we have all the ingredients in place to prove the main result of this paper, Theorem 1.2.

Proof of Theorem 1.2. In the following $c > 0$ denotes a sufficiently small absolute constant. First, by Lemma 2.2 we know that due to our assumption $m \gtrsim rd\kappa^2$, with probability $1 - \exp(-d)$ the measurement operator $\mathcal{A}$ satisfies the Restricted Isometry Property of order $6r$ with a constant $\delta = \delta_{6r} \leq \frac{c}{\kappa}$, where $c > 0$ is a sufficiently small absolute constant.

Set

$$T := \left\lceil \frac{8}{\mu\sigma_{\min}(\mathbf{X}_\star)} \log(16r) \right\rceil.$$

Note that since $r \geq 1$ and the assumption $\mu \leq \frac{c_1}{\sigma_{\min}(\mathbf{X}_\star)}$ for small $c_1 > 0$, we have $T \geq 1$. Let $\mathcal{N}_\varepsilon$ be an ε-net of the unit sphere in $\mathbb{R}^d$ with $\varepsilon = 1/2$ such that $|\mathcal{N}_\varepsilon| \leq 6^d$. Now note that $2T \leq 6^d$, where we have used the assumption $\mu \geq \frac{32}{\sigma_{\min}(\mathbf{X}_\star)6^d} \log(16r)$. Thus, it follows from Lemma 3.6 that with probability at least $1 - 2\exp(-10d)$ it holds that

$$|\langle \mathbf{w}\mathbf{w}^\top, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right)\rangle| \leq 4\sqrt{\frac{d}{m}} \left\|\mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right)\right\|_2$$

for all $\mathbf{w} \in \mathcal{N}_\varepsilon$ and for all $0 \leq t \leq T$. Next, we know from Lemma 4.1 and due to our assumption $m \gtrsim rd\kappa^2$ that with probability at least $1 - 5\exp(-d)$, the inequalities

$$\begin{aligned}\left\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right\| &\leq c\sigma_{\min}(\mathbf{X}_\star), \\ \left\|\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right\|_F &\leq c\sigma_{\min}(\mathbf{X}_\star)\end{aligned} \tag{67}$$

hold for a sufficiently small constant $c > 0$. Thus, all the assumptions of Lemma 4.10 are fulfilled. It follows that

$$\operatorname{dist}^2(\mathbf{U}_t, \mathbf{U}_\star) \leq r\left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)^{2t} \left\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right\| \tag{68}$$

for all $0 \leq t \leq T$ and

$$\operatorname{dist}(\mathbf{U}_T, \mathbf{U}_\star) \leq \frac{\sigma_{\min}(\mathbf{X}_\star)}{16}. \tag{69}$$

Due to inequality (69) and since $\delta_{6r} < 1/10$ we can apply Lemma 4.2 which yields that for $t \geq T$,

$$\operatorname{dist}^2(\mathbf{U}_t, \mathbf{U}_\star) \leq (1 - c\mu\sigma_{\min}(\mathbf{X}_\star))^{t-T} \operatorname{dist}^2(\mathbf{U}_T, \mathbf{U}_\star). \tag{70}$$

Thus, by combining (67), (68), and (70) we obtain the conclusion of Theorem 1.2. □

5. Discussions

In this paper, we have shown that for symmetric matrix sensing, factorized gradient descent can recover the ground truth matrix as soon as the number of samples satisfies $m \gtrsim rd\kappa^2$. This improves over previous results in the literature with a quadratic rank dependence. The key ingredient in our proof is a combination of a virtual sequence argument with an ε-net argument.

Going forward, our work opens up a number of exciting research directions. In the following, we highlight a few of these.

- *Breaking the quadratic rank barrier in related non-convex matrix sensing problems:* We expect that our novel proof technique will pave the way to break the quadratic rank barrier in the sample complexity in various related non-convex matrix sensing problems. This includes matrix sensing with an asymmetric ground truth matrix or overparameterized matrix sensing with small random initialization [22]. One might also examine whether our new proof technique can be used to remove the additional rank factor in the sample complexity in related algorithms such as *scaled gradient descent* [11] or *GSMR* [30].
- *Removing the condition number dependence in the sample complexity:* Compared to the nuclear norm minimization approach, the sample complexity in Theorem 1.2 is still suboptimal since it depends quadratically on the condition number of the ground truth matrix $\mathbf{X}_\star$. Indeed, all related results in the non-convex low-rank matrix recovery also have such a dependency on the condition number. It would be interesting to examine whether this dependence on the condition number is actually needed.
- *Beyond Gaussian measurement matrices:* It would also be interesting to examine whether the argument in this paper can be adapted to scenarios where the measurement matrices are no longer Gaussian, e.g., the matrix completion problem. Since the proof presented in this paper heavily relies on the orthogonal invariance of the Gaussian distribution, new insights are likely required to handle scenarios where this property is no longer available. We believe that this is an exciting research direction.

References

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

[2] Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. Phaselift: exact and stable signal recovery from magnitude measurements via convex programming. *Commun. Pure Appl. Math.*, 66(8):1241–1274, 2013. ISSN 0010-3640. doi: 10.1002/cpa.21432.

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

[4] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Trans. Inf. Theory*, 60(3):1711–1732, 2014. ISSN 0018-9448. doi: 10.1109/TIT.2013.2294644.

[5] Shuyang Ling and Thomas Strohmer. Blind deconvolution meets blind demixing: algorithms and performance bounds. *IEEE Trans. Inf. Theory*, 63(7):4497–4520, 2017. ISSN 0018-9448. doi: 10.1109/TIT.2017.2701342.

[6] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, 2010. ISSN 0036-1445. doi: 10.1137/070697835. URL hdl.handle.net/1721.1/60575.

[7] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theory*, 56(5):2053–2080, 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2044061.

[8] David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theory*, 57(3):1548–1566, 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2104999.

[9] Peter Jung, Felix Krahmer, and Dominik Stöger. Blind demixing and deconvolution at near-optimal rate. *IEEE Trans. Inf. Theory*, 64(2):704–727, 2018. ISSN 0018-9448. doi: 10.1109/TIT.2017.2784481.

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

[11] Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *J. Mach. Learn. Res.*, 22:63, 2021. ISSN 1532-4435. URL jmlr.csail.mit.edu/papers/v22/20-1067.html. Id/No 150.

[12] Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and René Vidal. Nonconvex robust low-rank matrix recovery. *SIAM J. Optim.*, 30(1):660–686, 2020. ISSN 1052-6234. doi: 10.1137/18M1224738.

[13] Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Found. Comput. Math.*, 21(6):1505–1593, 2021. ISSN 1615-3375. doi: 10.1007/s10208-020-09490-9.

[14] Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Trans. Inf. Theory*, 56(6):2980–2998, 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2046205.

[15] Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inf. Theory*, 62(11):6535–6579, 2016. ISSN 0018-9448. doi: 10.1109/TIT.2016.2598574.

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

[17] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in Neural Information Processing Systems*, 29, 2016.

[18] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Found. Comput. Math.*, 20(3):451–632, 2020. ISSN 1615-3375. doi: 10.1007/s10208-019-09429-9.

[19] Ji Chen, Dekai Liu, and Xiaodong Li. Nonconvex rectangular matrix completion via gradient descent without ℓ_2, ∞ regularization. *IEEE Trans. Inf. Theory*, 66(9):5806–5841, 2020. ISSN 0018-9448. doi: 10.1109/TIT.2020.2992234.

[20] Shuyang Ling and Thomas Strohmer. Regularized gradient descent: a non-convex recipe for fast joint blind deconvolution and demixing. *Inf. Inference*, 8(1):1–49, 2019. ISSN 2049-8764. doi: 10.1093/imaiai/iax022.

[21] Jialin Dong and Yuanming Shi. Nonconvex demixing from bilinear measurements. *IEEE Trans. Signal Process.*, 66(19):5152–5166, 2018. ISSN 1053-587X. doi: 10.1109/TSP.2018.2864660.

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

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

[24] Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon Shaolei Du, and Jason D Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *International Conference on Machine Learning*, pages 15200–15238. PMLR, 2023.

[25] Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. In *International Conference on Machine Learning*, pages 38611–38654. PMLR, 2023.

[26] Mahdi Soltanolkotabi, Dominik Stöger, and Changzhi Xie. Implicit balancing and regularization: Generalization and convergence guarantees for overparameterized asymmetric matrix sensing. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5140–5142. PMLR, 2023.

[27] Jianhao Ma and Salar Fattahi. Convergence of gradient descent with small initialization for unregularized matrix completion. *arXiv preprint arXiv:2402.06756*, 2024.

[28] Johan S Wind. Asymmetric matrix sensing by gradient descent with small random initialization. *arXiv preprint arXiv:2309.01796*, 2023.

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

[30] Pini Zilber and Boaz Nadler. GNMR: a provable one-line algorithm for low rank matrix recovery. *SIAM J. Math. Data Sci.*, 4(2):909–934, 2022. ISSN 2577-0187. doi: 10.1137/21M1433812.

[31] Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: an overview. *IEEE Trans. Signal Process.*, 67(20):5239–5269, 2019. ISSN 1053-587X. doi: 10.1109/TSP.2019.2937282.

[32] Prateek Jain, Raghu Meka, and Inderjit Dhillon. Guaranteed rank minimization via singular value projection. *Advances in Neural Information Processing Systems*, 23, 2010.

[33] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: primal and dual analysis. *IEEE Trans. Inf. Theory*, 66(11):7274–7301, 2020. ISSN 0018-9448. doi: 10.1109/TIT.2020.2992769.

[34] Jared Tanner and Ke Wei. Normalized iterative hard thresholding for matrix completion. *SIAM J. Sci. Comput.*, 35(5):s104–s125, 2013. ISSN 1064-8275. doi: 10.1137/120876459. URL semanticscholar.org/paper/9b785002627fd2066fce004199758ce137a1ce61.

[35] Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *J. Mach. Learn. Res.*, 13:3441–3473, 2012. ISSN 1532-4435. URL www.jmlr.org/papers/v13/mohan12a.html.

[36] Massimo Fornasier, Holger Rauhut, and Rachel Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM J. Optim.*, 21(4):1614–1640, 2011. ISSN 1052-6234. doi: 10.1137/100811404.

[37] Christian Kümmerle and Juliane Sigl. Harmonic mean iteratively reweighted least squares for low-rank matrix recovery. *J. Mach. Learn. Res.*, 19:49, 2018. ISSN 1532-4435. URL jmlr.csail.mit.edu/papers/v19/17-244.html. Id/No 47.

[38] Christian Kümmerle and Claudio M Verdun. A scalable second order method for ill-conditioned matrix completion from few samples. In *International Conference on Machine Learning*, pages 5872–5883. PMLR, 2021.

[39] Kiryung Lee and Yoram Bresler. ADMiRA: atomic decomposition for minimum rank approximation. *IEEE Trans. Inf. Theory*, 56(9):4402–4416, 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2054251.

[40] Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM J. Matrix Anal. Appl.*, 37(3):1198–1222, 2016.

[41] Bart Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM J. Optim.*, 23(2):1214–1236, 2013. ISSN 1052-6234. doi: 10.1137/110845768. URL semanticscholar.org/paper/feb9713f4e7614aecdb4778c0bc8c2dced60a325.

[42] Guillaume Olikier, André Uschmajew, and Bart Vandereycken. Gauss-southwell type descent methods for low-rank matrix optimization. *arXiv preprint arXiv:2306.00897*, 2023.

[43] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Number 118. Cambridge university press, 2010.

[44] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. *Advances in Neural Information Processing Systems*, 29, 2016.

[45] Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74. PMLR, 2017.

[46] André Uschmajew and Bart Vandereycken. On critical points of quadratic low-rank matrix optimization problems. *IMA J. Numer. Anal.*, 40(4):2626–2651, 2020. ISSN 0272-4979. doi: 10.1093/imanum/drz061.

[47] Richard Y. Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *J. Mach. Learn. Res.*, 20: 34, 2019. ISSN 1532-4435. URL jmlr.csail.mit.edu/papers/v20/19-020.html. Id/No 114.

[48] Jason D. Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Math. Program.*, 176(1-2 (B)):311–337, 2019. ISSN 0025-5610. doi: 10.1007/s10107-019-01374-3.

[49] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. *Advances in neural information processing systems*, 30, 2017.

[50] Emmanuel J. Candès and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory*, 57(4):2342–2359, 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2111771.

[51] Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970. ISSN 0036-1429. doi: 10.1137/0707001.

[52] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral methods for data science: a statistical perspective. *Found. Trends Mach. Learn.*, 14(5):1–246, 2021. ISSN 1935-8237. doi: 10.1561/2200000079.

[53] Jiacheng Zhuo, Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the computational and statistical complexity of over-parameterized matrix sensing. *Journal of Machine Learning Research*, 25(169):1–47, 2024.

[54] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[55] NL Johnson, S Kotz, and N Balakrishnan. Chi-squared distributions including Chi and Rayleigh. *Continuous univariate distributions*, pages 415–493, 1994.

[56] Roger A Horn and Charles R Johnson. *Topics in matrix analysis*. Cambridge university press, 1994.

[57] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[58] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: fast global convergence for nonconvex phase retrieval. *Math. Program.*, 176(1-2 (B)): 5–37, 2019. ISSN 0025-5610. doi: 10.1007/s10107-019-01363-6.

[59] Kiryung Lee and Dominik Stöger. Randomly initialized alternating least squares: Fast convergence for matrix sensing. *SIAM Journal on Mathematics of Data Science*, 5(3):774–799, 2023. doi: 10.1137/22M1506456. URL https://doi.org/10.1137/22M1506456.

[60] Roman Vershynin. *High-dimensional probability. An introduction with applications in data science*, volume 47 of *Camb. Ser. Stat. Probab. Math.* Cambridge: Cambridge University Press, 2018. ISBN 978-1-108-41519-4; 978-1-108-23159-6. doi: 10.1017/9781108231596.

[61] Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *Advances in Neural Information Processing Systems*, 28, 2015.

[62] Felix Krahmer, Shahar Mendelson, and Holger Rauhut. Suprema of chaos processes and the restricted isometry property. *Commun. Pure Appl. Math.*, 67(11):1877–1904, 2014. ISSN 0010-3640. doi: 10.1002/cpa.21504.

[63] Michel Talagrand. *The generic chaining. Upper and lower bounds of stochastic processes*. Springer Monogr. Math. Berlin: Springer, 2005. ISBN 3-540-24518-9; 978-3-642-06386-2; 978-3-540-27499-5. doi: 10.1007/3-540-27499-5.

A. Proof for the Spectral Initialization (Proof of Lemma 4.1)

Proof of Lemma 4.1. (1) We write

$$\left(\mathcal{A}^*\mathcal{A}\right)\left(\mathbf{X}_\star\right)-\mathbf{X}_\star=\frac{1}{m}\sum_{i=1}^m\left(\langle\mathbf{A}_i,\mathbf{X}_\star\rangle\mathbf{A}_i-\mathbf{X}_\star\right).$$

Let $\widetilde{\mathcal{N}_\varepsilon}$ be any ε-net on S^{d-1} with $\varepsilon=\frac{1}{2}$ of size at most 6^d. Then we have

$$\begin{aligned}\left\|\left(\mathcal{A}^*\mathcal{A}\right)\left(\mathbf{X}_\star\right)-\mathbf{X}_\star\right\|\leq&2\sup_{\mathbf{x}\in\widetilde{\mathcal{N}_\varepsilon}}\frac{1}{m}\sum_{i=1}^m\mathbf{x}^\top\left(\langle\mathbf{A}_i,\mathbf{X}_\star\rangle\mathbf{A}_i-\mathbf{X}_\star\right)\mathbf{x}\\=&2\sup_{\mathbf{x}\in\widetilde{\mathcal{N}_\varepsilon}}\frac{1}{m}\sum_{i=1}^m\left(\langle\mathbf{A}_i,\mathbf{X}_\star\rangle\mathbf{x}^\top\mathbf{A}_i\mathbf{x}-\mathbf{x}^\top\mathbf{X}_\star\mathbf{x}\right).\end{aligned}$$

For each $i\in[m]$, we have that $\mathbb{E}\left[\langle\mathbf{A}_i,\mathbf{X}_\star\rangle\mathbf{x}^\top\mathbf{A}_i\mathbf{x}\right]=\mathbf{x}^\top\mathbf{X}_\star\mathbf{x}$. Moreover, the inner product $\langle\mathbf{A}_i,\mathbf{X}_\star\rangle$ is a centered Gaussian random variable with variance $\|\mathbf{X}_\star\|_F^2$ and $\mathbf{x}^\top\mathbf{A}_i\mathbf{x}$ is a centered Gaussian random variable with variance 1. Thus, for each fixed $\mathbf{x}$, $\sum_{i=1}^m\left(\langle\mathbf{A}_i,\mathbf{X}_\star\rangle\mathbf{x}^\top\mathbf{A}_i\mathbf{x}-\mathbf{x}^\top\mathbf{X}_\star\mathbf{x}\right)$ is a sum of m independent and centered sub-exponential random variables with subexponential norm bounded by $K\|\mathbf{X}_\star\|_F$, where K is an absolute constant (see [60, Lemma 2.7.7]). Therefore, by Bernstein's inequality (see, for example, [60, Theorem 2.8.1]), it holds that

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m\left(\langle\mathbf{A}_i,\mathbf{X}_\star\rangle\mathbf{x}^\top\mathbf{A}_i\mathbf{x}-\mathbf{x}^\top\mathbf{X}_\star\mathbf{x}\right)\right|\geq t\right)\leq\exp\left(-C'\min\left\{\frac{mt^2}{\|\mathbf{X}_\star\|_F^2},\frac{mt}{\|\mathbf{X}_\star\|_F}\right\}\right),$$

where $C'>0$ is some absolute constant. Taking $t=\frac{1}{8}C\|\mathbf{X}_\star\|_F\left(\sqrt{\frac{d}{m}}+\frac{d}{m}\right)$ and a union bound over all points $\mathbf{x}$ on $\widetilde{\mathcal{N}_\varepsilon}$, we obtain

$$\left\|\left(\mathcal{A}^*\mathcal{A}\right)\left(\mathbf{X}_\star\right)-\mathbf{X}_\star\right\|\leq\frac{1}{4}C\|\mathbf{X}_\star\|_F\left(\sqrt{\frac{d}{m}}+\frac{d}{m}\right)\leq\frac{1}{4}C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{r}\left(\sqrt{\frac{d}{m}}+\frac{d}{m}\right)\tag{71}$$

with probability at least $1-\exp(d\log(6)-C'C^2d) \geq 1-\exp(-4d)$ for some sufficiently large constant $C > 0$.

We assume that (71) holds and that $m > C^2\kappa^2 rd$. Then Weyl's inequalities imply that

$$\lambda_r((\mathcal{A}^*\mathcal{A})(\mathbf{X}_\star)) > \frac{1}{2}\sigma_{\min}(\mathbf{X}_\star), \quad |\lambda_{r+1}((\mathcal{A}^*\mathcal{A})(\mathbf{X}_\star))| < \frac{1}{2}\sigma_{\min}(\mathbf{X}_\star).$$

Since $\widetilde{\mathbf{\Lambda}}_r$ is a diagonal matrix with entries $\lambda_1((\mathcal{A}^*\mathcal{A})(\mathbf{X}_\star)), \ldots, \lambda_r((\mathcal{A}^*\mathcal{A})(\mathbf{X}_\star))$, it follows from the definition of $\mathbf{U}_0 = \widetilde{\mathbf{V}}_r\widetilde{\mathbf{\Lambda}}_r^{1/2}$ that $\mathbf{U}_0\mathbf{U}_0^\top$ is the best rank-r approximation of $(\mathcal{A}^*\mathcal{A})(\mathbf{X}_\star)$. Consequently, we obtain that

$$\begin{aligned}\left\|\mathbf{X}_\star - \mathbf{U}_0\mathbf{U}_0^\top\right\| &\leq \left\|\mathbf{X}_\star - (\mathcal{A}^*\mathcal{A})(\mathbf{X}_\star)\right\| + \left\|(\mathcal{A}^*\mathcal{A})(\mathbf{X}_\star) - \mathbf{U}_0\mathbf{U}_0^\top\right\| \\ &\leq \left\|\mathbf{X}_\star - (\mathcal{A}^*\mathcal{A})(\mathbf{X}_\star)\right\| + \left\|(\mathcal{A}^*\mathcal{A})(\mathbf{X}_\star) - \mathbf{X}_\star\right\| \leq C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}},\end{aligned}$$

where in the second inequality, we used the Eckart-Young-Mirsky theorem.

(2) Due to Lemma 3.5 we have

$$(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I})(\mathbf{X}_\star) = (\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star)) - \langle\mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star)\right)\rangle\mathbf{w}\mathbf{w}^\top. \tag{72}$$

It follows that

$$\|(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I})(\mathbf{X}_\star)\| \leq \|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star))\| + |\langle\mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star)\right)\rangle|. \tag{73}$$

For a fixed $\mathbf{w} \in \mathcal{N}_\varepsilon$, we obtain with an analogous argument as for (71) that with probability at least $1-\exp(-4d)$,

$$\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star))\| \leq C\left\|\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star)\right\|_F\left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right) \leq \frac{1}{4}C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{r}\left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right).$$

The second term in (73) can be rewritten as

$$\langle\mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star)\right)\rangle = \frac{1}{m}\sum_{i=1}^m \langle\mathbf{w}\mathbf{w}^\top, \mathbf{A}_i\rangle\langle\mathbf{A}_i, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star)\rangle.$$

Here, $\sum_{i=1}^m \langle\mathbf{w}\mathbf{w}^\top, \mathbf{A}_i\rangle\langle\mathbf{A}_i, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star)\rangle$ is a sum of m independent sub-exponential random variables with mean zero due to the rotation invariance of the Gaussian measure. Moreover, each term has sub-exponential norm $K\left\|\mathbf{X}_\star\right\|_F$. Applying Bernstein's inequality as in the proof of (71), we obtain that for each fixed $\mathbf{w}$ with probability at least $1-\exp(-4d)$,

$$\langle\mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star)\right)\rangle \leq \frac{1}{4}C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{r}\left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right). \tag{74}$$

Then, by taking a union bound over $\mathbf{w} \in \mathcal{N}_\varepsilon$, it follows from (73) that with probability at least $1-\exp(-2d)$ that for all $\mathbf{w} \in \mathcal{N}_\varepsilon$ it holds that

$$\left\|(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I})(\mathbf{X}_\star)\right\| \leq \frac{1}{2}C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{r}\left(\sqrt{\frac{d}{m}} + \frac{d}{m}\right). \tag{75}$$

We now assume that (75) holds and that $m > 4C^2\kappa^2 rd$. Then it follows from Weyl's inequalities that

$$\lambda_r((\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})(\mathbf{X}_\star)) > \frac{1}{2}\sigma_{\min}(\mathbf{X}_\star), \quad |\lambda_{r+1}((\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})(\mathbf{X}_\star))| < \frac{1}{2}\sigma_{\min}(\mathbf{X}_\star).$$

It follows from the Eckart-Mirsky-Young theorem and the definition of $\mathbf{U}_{0,\mathbf{w}}$ that $\mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top$ is the best rank-r approximation of $(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})(\mathbf{X}_\star)$. Therefore,

$$\begin{aligned}\left\|\mathbf{X}_\star - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right\| &\leq \left\|\mathbf{X}_\star - (\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})(\mathbf{X}_\star)\right\| + \left\|(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})(\mathbf{X}_\star) - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right\| \\ &\leq 2\left\|\mathbf{X}_\star - (\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})(\mathbf{X}_\star)\right\| \leq 2C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}.\end{aligned}$$

This finishes the proof of inequality (24). Finally, (25) follows from (23) and (24) via the triangle inequality.

(3) From (72), we have

$$\begin{aligned}\left(\mathcal{A}^*\mathcal{A}\right)(\mathbf{X}_\star) - \left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)(\mathbf{X}_\star) &= (\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathbf{X}_\star) - (\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I})(\mathbf{X}_\star) \\ &= \langle \mathbf{w}\mathbf{w}^\top, \mathbf{X}_\star\rangle(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathbf{w}\mathbf{w}^\top) + \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star))\rangle \mathbf{w}\mathbf{w}^\top.\end{aligned}$$

It follows from Lemma 2.2 that there exists an absolute constant $C_1 > 0$ such that for any $\alpha \in (0,1)$ and $m \geq \frac{C_1}{\alpha^2}\kappa^2 rd$, with probability at least $1 - \exp(-d)$, the measurement operator $\mathcal{A}$ satisfies the Restricted Isometry Property of order $6r$ with constant

$$\delta := \delta_{6r} \leq \frac{\alpha}{\kappa}. \tag{76}$$

Then for any $\mathbf{V} \in \mathbb{R}^{d\times r}$ with orthonormal columns and for all $\mathbf{w} \in \mathcal{N}_\varepsilon$, when $m \geq \frac{C_1}{\alpha^2}\kappa^2 rd$, with probability at least $1 - 2\exp(-d)$,

$$\begin{aligned}&\left\|(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})(\mathbf{X}_\star)\mathbf{V}\right\|_F \\ \leq& |\langle \mathbf{w}\mathbf{w}^\top, \mathbf{X}_\star\rangle| \left\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathbf{w}\mathbf{w}^\top)\mathbf{V}\right\|_F + |\langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star))\rangle| \left\|\mathbf{w}\mathbf{w}^\top\mathbf{V}\right\|_F \\ \overset{(a)}{\leq}& \delta \|\mathbf{X}_\star\| \left\|\mathbf{w}\mathbf{w}^\top\right\|_F + |\langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{X}_\star))\rangle| \\ \overset{(b)}{\leq}& \alpha\sigma_{\min}(\mathbf{X}_\star) + \frac{1}{2}C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}.\end{aligned} \tag{77}$$

Here in (a) we use property (7) in Lemma 2.4 and the fact that $\mathbf{w}\mathbf{w}^\top\mathbf{V}$ is of rank 1, and in (b) we use (76) and, moreover, (74) with a union bound over $\mathbf{w} \in \mathcal{N}_\varepsilon$.

We now proceed under the assumption that the inequalities in parts (1) and (2) hold. We use the following notations for spectral initialization:

$$\begin{aligned}\left(\mathcal{A}^*\mathcal{A}\right)(\mathbf{X}_\star) &= \widetilde{\mathbf{V}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{V}}^\top, \quad \mathbf{U}_0 = \widetilde{\mathbf{V}}_r\widetilde{\mathbf{\Lambda}}_r^{1/2}, \\ \left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)(\mathbf{X}_\star) &= \widetilde{\mathbf{V}}_{\mathbf{w}}\widetilde{\mathbf{\Lambda}}_{\mathbf{w}}\widetilde{\mathbf{V}}_{\mathbf{w}}^\top, \quad \mathbf{U}_{0,\mathbf{w}} = \widetilde{\mathbf{V}}_{r,\mathbf{w}}\widetilde{\mathbf{\Lambda}}_{r,\mathbf{w}}^{1/2}.\end{aligned} \tag{78}$$

Denote

$$\mathbf{Z}_1 := (\mathcal{A}^*\mathcal{A})(\mathbf{X}_\star), \quad \mathbf{Z}_2 := (\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})(\mathbf{X}_\star),$$

and

$$\mathbf{Z}_{1,r} := \mathbf{U}_0\mathbf{U}_0^\top, \quad \mathbf{Z}_{2,r} := \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top.$$

Recall the definition of $\widetilde{\mathbf{V}}_r$ and $\widetilde{\mathbf{V}}_{r,\mathbf{w}}$ in (78) and (17). We have

$$\begin{aligned}\|\mathbf{Z}_{1,r} - \mathbf{Z}_{2,r}\|_F &= \|\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\|_F \\ &\leq \|\left(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right)\widetilde{\mathbf{V}}_r\|_F + \|\left(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right)\widetilde{\mathbf{V}}_{r,\perp}\|_F.\end{aligned} \tag{79}$$

For the first term in (79), we have

$$\begin{aligned}&\|\left(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right)\widetilde{\mathbf{V}}_r\|_F \\ =&\|(\mathbf{Z}_1 - \mathbf{Z}_{2,r})\widetilde{\mathbf{V}}_r\|_F \\ \leq&\|(\mathbf{Z}_1 - \mathbf{Z}_2)\widetilde{\mathbf{V}}_r\|_F + \|(\mathbf{Z}_2 - \mathbf{Z}_{2,r})\widetilde{\mathbf{V}}_r\|_F \\ =&\|(\mathbf{Z}_1 - \mathbf{Z}_2)\widetilde{\mathbf{V}}_r\|_F + \|(\widetilde{\mathbf{V}}_{r,\mathbf{w},\perp}\mathbf{\Lambda}_{r,\mathbf{w},\perp}\widetilde{\mathbf{V}}_{r,\mathbf{w},\perp}^\top)\widetilde{\mathbf{V}}_r\|_F \\ \leq&\|(\mathbf{Z}_1 - \mathbf{Z}_2)\widetilde{\mathbf{V}}_r\|_F + \sigma_{r+1}(\mathbf{Z}_2)\|\widetilde{\mathbf{V}}_{r,\mathbf{w},\perp}^\top\widetilde{\mathbf{V}}_r\|_F \\ \leq&\|(\mathbf{Z}_1 - \mathbf{Z}_2)\widetilde{\mathbf{V}}_r\|_F + C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}\|\widetilde{\mathbf{V}}_{r,\mathbf{w},\perp}^\top\widetilde{\mathbf{V}}_r\|_F,\end{aligned} \tag{80}$$

where in the last inequality we used Weyl's inequality and (75), which implies

$$\sigma_{r+1}(\mathbf{Z}_2) = |\sigma_{r+1}(\mathbf{Z}_2) - \sigma_{r+1}(\mathbf{X}_\star)| \leq \|\mathbf{Z}_2 - \mathbf{X}_\star\| \leq C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}\|\widetilde{\mathbf{V}}_{r,\mathbf{w},\perp}^\top\widetilde{\mathbf{V}}_r\|_F. \tag{81}$$

From (75) and (71), it follows that when $m \geq C^2\kappa^2 rd$,

$$\|\mathbf{Z}_1 - \mathbf{Z}_2\| \leq \frac{3C}{2}\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}. \tag{82}$$

Similar to (81), using (75) and Weyl's inequalities we obtain that

$$|\sigma_r(\mathbf{Z}_1) - \sigma_{\min}(\mathbf{X}_\star)| \leq C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}},$$
$$\sigma_{r+1}(\mathbf{Z}_1) \leq C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}.$$

Therefore, if $m > 16C^2\kappa^2 rd$, the spectral gap between $\sigma_r(\mathbf{Z}_1)$ and $\sigma_{r+1}(\mathbf{Z}_2)$ can be bounded from below by

$$\sigma_r(\mathbf{Z}_1) - \sigma_{r+1}(\mathbf{Z}_1) \geq \left(1 - 2C\kappa\sqrt{\frac{rd}{m}}\right)\sigma_{\min}(\mathbf{X}_\star) \geq \frac{1}{2}\sigma_{\min}(\mathbf{X}_\star). \tag{83}$$

When $m \geq 51C^2\kappa^2 rd$, we have from (82) and (83),

$$\begin{aligned}\|\mathbf{Z}_1 - \mathbf{Z}_2\| &\leq \frac{3C}{2}\kappa\sqrt{\frac{rd}{m}}\sigma_{\min}(\mathbf{X}_\star)\\ &\leq \left(1 - \frac{1}{\sqrt{2}}\right)\left(1 - 2C\kappa\sqrt{\frac{rd}{m}}\right)\sigma_{\min}(\mathbf{X}_\star)\\ &\leq \left(1 - \frac{1}{\sqrt{2}}\right)(\sigma_r(\mathbf{Z}_1) - \sigma_{r+1}(\mathbf{Z}_1)).\end{aligned}$$

Thus, the prerequisites of Lemma 2.6 (Davis-Kahan inequality) are satisfied. It follows that when $m \geq 51C^2\kappa^2 rd$,

$$\|\widetilde{\mathbf{V}}_{r,\mathbf{w},\perp}^\top \widetilde{\mathbf{V}}_r\|_F \leq \frac{2\sqrt{2}\|(\mathbf{Z}_1 - \mathbf{Z}_2)\widetilde{\mathbf{V}}_r\|_F}{\sigma_{\min}(\mathbf{X}_\star)}. \tag{84}$$

Hence, when $m \geq \left(51C^2 + \frac{C_1}{\alpha^2}\right)\kappa^2 rd$, we obtain from (80) and (77) that

$$\begin{aligned}\left\|\left(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right)\widetilde{\mathbf{V}}_r\right\|_F &\leq \left(1 + 2\sqrt{2}C\kappa\sqrt{\frac{rd}{m}}\right)\left\|(\mathbf{Z}_1 - \mathbf{Z}_2)\widetilde{\mathbf{V}}_r\right\|_F\\ &\leq 2\left\|(\mathbf{Z}_1 - \mathbf{Z}_2)\widetilde{\mathbf{V}}_r\right\|_F \leq \left(2\alpha + C\kappa\sqrt{\frac{rd}{m}}\right)\sigma_{\min}(\mathbf{X}_\star).\end{aligned} \tag{85}$$

For the second term in (79), we have when $m \geq \left(51C^2 + \frac{C_1}{\alpha^2}\right)\kappa^2 rd$,

$$\begin{aligned}&\left\|\left(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right)\widetilde{\mathbf{V}}_{r,\perp}\right\|_F\\ \leq&\|\widetilde{\mathbf{V}}_r^\top\left(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right)\widetilde{\mathbf{V}}_{r,\perp}\|_F + \|\widetilde{\mathbf{V}}_{r,\perp}^\top\left(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right)\widetilde{\mathbf{V}}_{r,\perp}\|_F\\ \leq&\|\widetilde{\mathbf{V}}_r^\top\left(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right)\|_F + \|\widetilde{\mathbf{V}}_{r,\perp}^\top\mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\widetilde{\mathbf{V}}_{r,\perp}\|_F\\ \leq&\left(2\alpha + C\kappa\sqrt{\frac{rd}{m}}\right)\sigma_{\min}(\mathbf{X}_\star) + \|\widetilde{\mathbf{V}}_{r,\perp}^\top\mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\widetilde{\mathbf{V}}_{r,\perp}\|_F,\end{aligned} \tag{86}$$

where the last inequality is due to (85).

We now consider the second term in (86). Recall the definition of $\mathbf{U}_{0,\mathbf{w}}$ in (18). We have for $m \geq \left(51C^2 + \frac{C_1}{\alpha^2}\right)\kappa^2 rd$,

$$\begin{aligned}
\|\widetilde{\mathbf{V}}_{r,\perp}^\top \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top \widetilde{\mathbf{V}}_{r,\perp}\|_F &= \left\|\widetilde{\mathbf{V}}_{r,\perp}^\top \widetilde{\mathbf{V}}_{r,\mathbf{w}}\mathbf{\Lambda}_{r,\mathbf{w}}\widetilde{\mathbf{V}}_{r,\mathbf{w}}^\top \widetilde{\mathbf{V}}_{r,\perp}\right\|_F \\
&\leq \left\|\widetilde{\mathbf{V}}_{r,\perp}^\top \widetilde{\mathbf{V}}_{r,\mathbf{w}}\mathbf{\Lambda}_{r,\mathbf{w}}\right\|\left\|\widetilde{\mathbf{V}}_{r,\mathbf{w}}^\top \widetilde{\mathbf{V}}_{r,\perp}\right\|_F \\
&= \sqrt{\left\|\widetilde{\mathbf{V}}_{r,\perp}^\top \widetilde{\mathbf{V}}_{r,\mathbf{w}}\mathbf{\Lambda}_{r,\mathbf{w}}^2\widetilde{\mathbf{V}}_{r,\mathbf{w}}^\top \widetilde{\mathbf{V}}_{r,\perp}\right\|}\left\|\widetilde{\mathbf{V}}_{r,\mathbf{w}}^\top \widetilde{\mathbf{V}}_{r,\perp}\right\|_F \\
&= \sqrt{\left\|\widetilde{\mathbf{V}}_{r,\perp}^\top (\mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top)^2\widetilde{\mathbf{V}}_{r,\perp}\right\|}\left\|\widetilde{\mathbf{V}}_{r,\mathbf{w}}^\top \widetilde{\mathbf{V}}_{r,\perp}\right\|_F \\
&= \left\|\widetilde{\mathbf{V}}_{r,\perp}^\top \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right\|\left\|\widetilde{\mathbf{V}}_{r,\mathbf{w}}^\top \widetilde{\mathbf{V}}_{r,\perp}\right\|_F \\
&= \left\|\widetilde{\mathbf{V}}_{r,\perp}^\top (\mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top - \mathbf{U}_0\mathbf{U}_0^\top)\right\|\left\|\widetilde{\mathbf{V}}_{r,\mathbf{w}}^\top \widetilde{\mathbf{V}}_{r,\perp}\right\|_F \\
&\leq \left\|\mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top - \mathbf{U}_0\mathbf{U}_0^\top\right\|\left\|\widetilde{\mathbf{V}}_{r,\mathbf{w}}^\top \widetilde{\mathbf{V}}_{r,\perp}\right\|_F \\
&\overset{(a)}{\leq} 3C\kappa\sigma_{\min}(\mathbf{X}_\star)\sqrt{\frac{rd}{m}}\cdot\frac{2\sqrt{2}\|(\mathbf{Z}_1 - \mathbf{Z}_2)\widetilde{\mathbf{V}}_r\|_F}{\sigma_{\min}(\mathbf{X}_\star)} \\
&\overset{(b)}{\leq} 6\sqrt{2}C\kappa\left(\alpha + \frac{1}{2}C\kappa\sqrt{\frac{rd}{m}}\right)\sqrt{\frac{rd}{m}}\sigma_{\min}(\mathbf{X}_\star),
\end{aligned} \tag{87}$$

where (a) is due to (25) and (84), and (b) is due to (77). Therefore from (86) and (87), we obtain for $m \geq \left(51C^2 + \frac{C_1}{\alpha^2}\right)\kappa^2 rd$,

$$\begin{aligned}
&\|\left(\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right)\widetilde{\mathbf{V}}_{r,\perp}\|_F \\
&\leq \left(2\alpha + C\kappa\sqrt{\frac{rd}{m}}\right)\sigma_{\min}(\mathbf{X}_\star) + 6\sqrt{2}C\kappa\left(\alpha + \frac{1}{2}C\kappa\sqrt{\frac{rd}{m}}\right)\sqrt{\frac{rd}{m}}\sigma_{\min}(\mathbf{X}_\star).
\end{aligned} \tag{88}$$

From (85), (88), and (79), we conclude that if $m \geq \left(51C^2 + \frac{C_1}{\alpha^2}\right)\kappa^2 rd$,

$$\left\|\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}_{0,\mathbf{w}}\mathbf{U}_{0,\mathbf{w}}^\top\right\|_F \leq \left(2\alpha + C\kappa\sqrt{\frac{rd}{m}}\right)\left(2\sigma_{\min}(\mathbf{X}_\star) + 3\sqrt{2}C\kappa\sqrt{\frac{rd}{m}}\sigma_{\min}(\mathbf{X}_\star)\right).$$

This finishes the proof of (26). □

B. Proofs of lemmas concerning the distance between the virtual sequences and the original sequence

B.1. Some auxiliary estimates

In order to prove Lemma 4.3 and Lemma 4.5 we will need several auxiliary estimates. These are summarized in the following lemma.

Lemma B.1. *Assume that the measurement operator $\mathcal{A}$ has the Restricted Isometry Property with constant $\delta = \delta_{4r+1} \leq 1$. Moreover, assume that the conclusion of Lemma 3.6 holds. Then, the following inequalities hold.*

1.

$$\begin{aligned}
&\left\|\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \\
&\leq \left(\delta + \frac{8\sqrt{rd}}{\sqrt{m}}\right)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \left(\delta + \frac{4\sqrt{2d}}{\sqrt{m}}\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F,
\end{aligned} \tag{89}$$

2.

$$\left\|\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \leq 2\delta\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right\|_F, \tag{90}$$

3.

$$\left\| \left[\left(\mathcal{A}^* \mathcal{A} - \mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} \right) \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right) \right] \mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}} \right\|_F \leq \left(\delta + 8\sqrt{\frac{rd}{m}} \right) \left\| \mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\| \quad (91)$$

4. and

$$\begin{aligned} \left\| \left(\mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} - \mathcal{I} \right) \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right) \right\| \leq & \left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{I} \right) \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) \right\| + \left(\delta + 8\sqrt{\frac{rd}{m}} \right) \left\| \mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right\| \\ & + \left(2\delta + 4\sqrt{\frac{2d}{m}} \right) \left\| \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top \right\|_F. \end{aligned} \quad (92)$$

Proof of Lemma B.1. To prove inequality (89), we compute that

$$\begin{aligned} \left(\mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} \right) \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) = & \left(\mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} \right) \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top} \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) \right) + \left(\mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} \right) \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) \right) \\ \overset{(a)}{=} & \left(\mathcal{A}^* \mathcal{A} \right) \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) \right) + \mathcal{P}_{\mathbf{w}\mathbf{w}^\top} \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) \\ & - \langle \mathcal{A} \left(\mathbf{w}\mathbf{w}^\top \right), \mathcal{A} \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) \right) \rangle \mathbf{w}\mathbf{w}^\top, \end{aligned}$$

where in equation (a) we used Lemma 3.5. It follows that

$$\begin{aligned} \left(\mathcal{A}^* \mathcal{A} - \mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} \right) \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) = & \left(\mathcal{A}^* \mathcal{A} - \mathcal{I} \right) \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top} \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) \right) \\ & + \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A} \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top) \right) \rangle \mathbf{w}\mathbf{w}^\top. \end{aligned}$$

By using the triangle inequality, we obtain the estimate

$$\begin{aligned} & \left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} \right) \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) \mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}} \right\|_F \\ \leq & \left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{I} \right) \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top} \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) \right) \mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}} \right\|_F + \left\| \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A} \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top) \right) \rangle \mathbf{w}\mathbf{w}^\top \right\|_F \\ \overset{(a)}{\leq} & \delta \left\| \mathcal{P}_{\mathbf{w}\mathbf{w}^\top} \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right) \right\|_F + \left| \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A} \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top) \right) \rangle \right| \\ \overset{(b)}{\leq} & \delta \left\| \mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right\| + \left| \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A} \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top) \right) \rangle \right| \\ & + \left| \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A} \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{U}_t \mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top) \right) \rangle \right| \\ \overset{(c)}{\leq} & \delta \left\| \mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right\| + \frac{4\sqrt{d}}{\sqrt{m}} \left\| \mathcal{A} \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top) \right) \right\|_2 + \delta \left\| \mathbf{U}_t \mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\|_F \\ \overset{(d)}{\leq} & \delta \left\| \mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right\| + \frac{4\sqrt{2d}}{\sqrt{m}} \left\| \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top) \right\|_F + \delta \left\| \mathbf{U}_t \mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\|_F \\ \overset{(e)}{\leq} & \delta \left\| \mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right\| + \frac{4\sqrt{2d}}{\sqrt{m}} \left\| \mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right\|_F + \left(\delta + \frac{4\sqrt{2d}}{\sqrt{m}} \right) \left\| \mathbf{U}_t \mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\|_F \\ \overset{(f)}{\leq} & \left(\delta + \frac{8\sqrt{rd}}{\sqrt{m}} \right) \left\| \mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right\| + \left(\delta + \frac{4\sqrt{2d}}{\sqrt{m}} \right) \left\| \mathbf{U}_t \mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\|_F. \end{aligned}$$

Inequality (a) follows from the RIP-assumption combined with Lemma 2.4 and from the fact that $\left\| \mathbf{w} \right\|_2 = 1$. Inequality (b) is a consequence of the fact that $\mathcal{P}_{\mathbf{w}\mathbf{w}^\top}$ is a rank-one projection and of the triangle inequality. In inequality (c), we used that the conclusion of Lemma 3.6 holds and Lemma 2.4. In inequality (d), we used the RIP of rank $2r+1$. Inequality (e) is due to the fact that $\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}$ is an orthogonal projection and due to the triangle inequality. In inequality (f), we used that $\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top$ has rank at most $2r$. This proves inequality (89).

To prove inequality (90) we compute first that

$$\begin{aligned} & \left(\mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} - \mathcal{I} \right) \left(\mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top \right) \\ = & \left(\mathcal{A}^* \mathcal{A} - \mathcal{I} \right) \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} \left(\mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top \right) \right) - \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A} \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top) \right) \rangle \mathbf{w}\mathbf{w}^\top. \end{aligned}$$

It follows that

$$
\begin{aligned}
&\left\| \left[\left(\mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} - \mathcal{I}\right) \left(\mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\right) \right] \mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}} \right\|_F \\
\overset{(a)}{\leq} & \delta \left\| \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} \left(\mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\right) \right\|_F + \left| \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top)\right) \rangle \right| \\
\overset{(b)}{\leq} & 2\delta \left\| \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top) \right\|_F \\
\leq & 2\delta \left\| \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top \right\|_F .
\end{aligned}
$$

In inequalities (a) and (b) we used Lemma 2.4. This proves inequality (90).

Next, we prove the third inequality. For that, we observe that using Lemma 3.5 it holds that

$$
\begin{aligned}
\left(\mathcal{A}^* \mathcal{A} - \mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}}\right) \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right) = & \left(\mathcal{A}^* \mathcal{A} - \mathcal{I}\right) \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top} \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right)\right) \\
& + \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top)\right) \rangle \mathbf{w}\mathbf{w}^\top .
\end{aligned}
$$

Then it follows that

$$
\begin{aligned}
&\left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}}\right) \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right) \mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}} \right\|_F \\
\leq & \left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{I}\right) \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top} \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right)\right) \mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}} \right\|_F + \left| \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top)\right) \rangle \right| \\
\overset{(a)}{\leq} & \delta \left\| \mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\| + 4\sqrt{\frac{d}{m}} \left\| \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} (\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top)\right) \right\|_2 \\
\overset{(b)}{\leq} & \delta \left\| \mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\| + 4\sqrt{\frac{2d}{m}} \left\| \mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\|_F \\
\leq & \left(\delta + 8\sqrt{\frac{rd}{m}} \right) \left\| \mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\|,
\end{aligned}
$$

where inequality (a) holds due to Lemma 2.4, since $\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}$ is a rank-one projection, and since we assumed that the conclusion of Lemma 3.6 holds. Inequality (b) is again due to Lemma 2.4 and since $\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}$ is an orthogonal projection. This proves inequality (91).

It remains to prove inequality (92). We note that it holds that

$$
\begin{aligned}
&\left(\mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} - \mathcal{I}\right) \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right) \\
= & \left(\mathcal{A}^* \mathcal{A} - \mathcal{I}\right) \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right)\right) - \langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right)) \rangle \mathbf{w}\mathbf{w}^\top ,
\end{aligned}
$$

where in the last line we applied Lemma 3.5. It follows from the triangle inequality that

$$
\begin{aligned}
&\left\| \left(\mathcal{A}_{\mathbf{w}}^* \mathcal{A}_{\mathbf{w}} - \mathcal{I}\right) \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right) \right\| \\
\leq & \left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{I}\right) \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top\right) \right\| + \left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{I}\right) \left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top} \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right)\right) \right\| \\
& + \left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{I}\right) \left(\mathbf{U}_t \mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right) \right\| + \left| \langle \mathcal{A}\left(\mathbf{w}\mathbf{w}^\top\right), \mathcal{A}\left(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp} \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right)\right) \rangle \right| \\
\overset{(a)}{\leq} & \left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{I}\right) \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top\right) \right\| + \delta \left\| \mathcal{P}_{\mathbf{w}\mathbf{w}^\top} \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top\right) \right\|_F + \delta \left\| \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top \right\|_F \\
& + 4\sqrt{\frac{2d}{m}} \left\| \mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\|_F \\
\leq & \left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{I}\right) \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top\right) \right\| + \delta \left\| \mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right\| \\
& + 2\delta \left\| \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top \right\|_F + 4\sqrt{\frac{2d}{m}} \left\| \mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top \right\|_F \\
\leq & \left\| \left(\mathcal{A}^* \mathcal{A} - \mathcal{I}\right) \left(\mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top\right) \right\| + \left(\delta + 8\sqrt{\frac{rd}{m}} \right) \left\| \mathbf{X}_\star - \mathbf{U}_t \mathbf{U}_t^\top \right\| \\
& + \left(2\delta + 4\sqrt{\frac{2d}{m}} \right) \left\| \mathbf{U}_{t,\mathbf{w}} \mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t \mathbf{U}_t^\top \right\|_F .
\end{aligned}
$$

In inequality (a) we applied Lemma 2.4 and that the conclusion of Lemma 3.6 holds. This proves inequality (92). Thus, the proof of Lemma B.1 is complete. □

B.2. Proof of Lemma 4.3

Proof of Lemma 4.3. We define the shorthand notation

$$\mathbf{M}_t := (\mathcal{A}^*\mathcal{A})\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right),$$
$$\mathbf{M}_{t,\mathbf{w}} := (\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right).$$

It follows that

$$\mathbf{U}_{t+1} = (\mathbf{Id} + \mu\mathbf{M}_t)\,\mathbf{U}_t,$$
$$\mathbf{U}_{t+1,\mathbf{w}} = (\mathbf{Id} + \mu\mathbf{M}_{t,\mathbf{w}})\,\mathbf{U}_{t,\mathbf{w}}.$$

We compute that

$$\begin{aligned}
\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}} =& (\mathbf{Id} + \mu\mathbf{M}_t)\,\mathbf{U}_t\mathbf{U}_t^\top(\mathbf{Id} + \mu\mathbf{M}_t) - (\mathbf{Id} + \mu\mathbf{M}_{t,\mathbf{w}})\,\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top(\mathbf{Id} + \mu\mathbf{M}_{t,\mathbf{w}}) \\
=& \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top + \mu\underbrace{\mathbf{M}_t(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)}_{=:(i)} + \mu\underbrace{(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}})\,\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top}_{=:(ii)} \\
&+ \mu\underbrace{(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\mathbf{M}_t}_{=:(iii)} + \mu\underbrace{\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}})}_{=:(iv)} \\
&+ \mu^2\underbrace{\left(\mathbf{M}_t\mathbf{U}_t\mathbf{U}_t^\top\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{M}_{t,\mathbf{w}}\right)}_{=:(v)}.
\end{aligned}$$

We want to estimate the spectral norm of these terms individually. Before that, we note that

$$\begin{aligned}
\left\|\mathbf{M}_t\right\| &\overset{(a)}{\leq} \left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t\right\| + \left\|\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\| \\
&\overset{(b)}{\leq} \left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + c_1\sigma_{\min}\left(\mathbf{X}_\star\right) \qquad (93) \\
&\overset{(c)}{\leq} 2\sigma_{\min}\left(\mathbf{X}_\star\right). \qquad (94)
\end{aligned}$$

Inequality (a) follows from the triangle inequality and inequality (b) follows from assumption (29). Inequality (c) is a consequence of assumption (30). Moreover, we note that

$$\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}} = (\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) - (\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right).$$

It follows that

$$\begin{aligned}
&\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \\
\leq& \left\|\left[(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}})\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F + \left\|\left[(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I})\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \\
&+ \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\
\overset{(a)}{\leq}& \left(\delta + \frac{8\sqrt{rd}}{\sqrt{m}}\right)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \left(3\delta + \frac{4\sqrt{2d}}{\sqrt{m}} + 1\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\
\overset{(b)}{\leq}& \frac{2c_3}{\kappa}\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \left(\frac{4c_3}{\kappa} + 1\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F, \qquad (95)
\end{aligned}$$

where in inequality (a) we used inequalities (89) and (90) from Lemma B.1. Inequality (b) is due to assumption (32). Note that it also follows from these estimates that

$$\begin{aligned}
\left\|\mathbf{M}_{t,\mathbf{w}}\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\| &\leq \left\|\mathbf{M}_t\right\| + \left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \\
&\overset{(a)}{\leq} 2\sigma_{\min}(\mathbf{X}_\star) + \frac{2c_3}{\kappa}\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \left(\frac{4c_3}{\kappa} + 1\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\
&\overset{(b)}{\leq} 3\sigma_{\min}(\mathbf{X}_\star), \qquad (96)
\end{aligned}$$

where inequality (a) follows from (95). Inequality (b) is a consequence of the assumptions (30) and (31) (and by choosing the absolute constant $c_3 > 0$ small enough).

Now we are in a position to estimate the spectral norms of the terms (i)-(v).

* Estimating term (i): We compute that that

$$\begin{aligned}\left\|\mathbf{M}_t(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right\|_F \leq& \|\mathbf{M}_t\|\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\ \overset{(93)}{\leq}& \left(\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\| + c_1\sigma_{\min}(\mathbf{X}_\star)\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F.\end{aligned}$$

* Estimating term (ii): We compute that

$$\begin{aligned}\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \leq& \left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\| \\ \leq& \left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F\left(\|\mathbf{U}_t\mathbf{U}_t^\top\| + \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|\right) \\ \leq& 3\|\mathbf{X}_\star\|\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F,\end{aligned}$$

where in the last inequality we used assumptions (28) and (30).

* Estimating term (iii): With the same argument as for term (i) we observe that

$$\left\|(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\mathbf{M}_t\right\|_F \leq \left(\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\| + c_1\sigma_{\min}\left(\mathbf{X}_\star\right)\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F.$$

* Estimating term (iv): With the same argument as for term (ii) we compute that

$$\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\right\|_F \leq 3\|\mathbf{X}_\star\|\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F.$$

* Estimating term (v): First, we compute that

$$\begin{aligned}\mathbf{M}_t\mathbf{U}_t\mathbf{U}_t^\top\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{M}_{t,\mathbf{w}} =& \mathbf{M}_t\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{M}_t + \left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{M}_t \\ &+ \mathbf{M}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right).\end{aligned}$$

It follows that

$$\begin{aligned}&\left\|\mathbf{M}_t\mathbf{U}_t\mathbf{U}_t^\top\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{M}_{t,\mathbf{w}}\right\|_F \\ \leq& \|\mathbf{M}_t\|^2\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + \left(\|\mathbf{U}_t\|^2 + \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|\right)\|\mathbf{M}_t\|\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \\ &+ \left\|\mathbf{M}_{t,\mathbf{w}}\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|\left(\|\mathbf{U}_t\|^2 + \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|\right)\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \\ \overset{(a)}{\leq}& \|\mathbf{M}_t\|^2\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + 3\|\mathbf{X}_\star\|\|\mathbf{M}_t\|\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \\ &+ 3\|\mathbf{X}_\star\|\left\|\mathbf{M}_{t,\mathbf{w}}\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \\ \overset{(b)}{\leq}& 4\sigma_{\min}^2(\mathbf{X}_\star)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + 15\sigma_{\min}(\mathbf{X}_\star)\|\mathbf{X}_\star\|\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F.\end{aligned}$$

For inequality (a) we used the assumptions (28) and (31). Inequality (b) is a consequence of inequalities (94) and (96).

* Conclusion: By summing up all terms we obtain that

$$\begin{aligned}&\left\|\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\right\|_F \\ \leq& \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + 2\mu\left(\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\| + c_1\sigma_{\min}(\mathbf{X}_\star)\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\ &+ 6\mu\|\mathbf{X}_\star\|\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \\ &+ \mu^2\left(4\sigma_{\min}^2(\mathbf{X}_\star)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + 15\sigma_{\min}(\mathbf{X}_\star)\|\mathbf{X}_\star\|\left\|\left(\mathbf{M}_t - \mathbf{M}_{t,\mathbf{w}}\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F\right) \\ \overset{(a)}{\leq}& \left(1 + 2\mu\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\| + 2c_1\sigma_{\min}(\mathbf{X}_\star)\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\ &+ 12\mu\sigma_{\min}(\mathbf{X}_\star)c_3\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\| + 6\mu\|\mathbf{X}_\star\|\left(\frac{4c_3}{\kappa} + 1\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\ &+ 4\mu^2\sigma_{\min}^2(\mathbf{X}_\star)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + 30c_3\mu^2\sigma_{\min}^2(\mathbf{X}_\star)\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\| \\ &+ 60c_3\mu^2\sigma_{\min}^2(\mathbf{X}_\star)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + 15\mu^2\sigma_{\min}(\mathbf{X}_\star)\|\mathbf{X}_\star\|\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\ =& \left(1 + 2\mu\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\| + \left(2c_1 + 24c_3\right)\mu\sigma_{\min}(\mathbf{X}_\star) + 6\mu\|\mathbf{X}_\star\| + 4\mu^2\sigma_{\min}^2(\mathbf{X}_\star) + 60c_3\mu^2\sigma_{\min}^2(\mathbf{X}_\star)\right) \\ &\cdot\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + \left(12c_3\mu\sigma_{\min}(\mathbf{X}_\star) + 30c_3\mu^2\sigma_{\min}^2(\mathbf{X}_\star)\right)\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\| \\ \overset{(b)}{\leq}& \frac{\sqrt{\sqrt{2}-1}}{40}\sigma_{\min}(\mathbf{X}_\star).\end{aligned}$$

Inequality (a) follows from inequality (95). Inequality (b) is due to assumptions (30), (31), and the assumption $\mu \leq \frac{c_2}{\kappa \|\mathbf{X}_\star\|}$ for a sufficiently small absolute constant $c_2 > 0$. This completes the proof of Lemma 4.3. □

B.3. Proof of Lemma 4.4

Proof of Lemma 4.4. Let $\mathbf{R} \in \mathbb{R}^{r \times r}$ be an orthogonal matrix. We compute that

$$\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top = \mathbf{U}_t\mathbf{R}\left(\mathbf{U}_t\mathbf{R}\right)^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}} = \mathbf{U}_t\mathbf{R}\left(\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right)^\top - \left(\mathbf{U}_{t,\mathbf{w}} - \mathbf{U}_t\mathbf{R}\right)\mathbf{U}_{t,\mathbf{w}}^\top.$$

It follows that

$$\begin{aligned}
&\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|_F \\
\leq&\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_t\mathbf{R}\right\|\left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|_F + \left\|\mathbf{U}_{t,\mathbf{w}} - \mathbf{U}_t\mathbf{R}\right\|_F\left\|\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{V}_{\mathbf{X}_\star,\perp}\right\| \\
\leq&\left(\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_t\mathbf{R}\right\| + \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_{t,\mathbf{w}}\right\|\right)\left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|_F \\
\leq&\left(2\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_t\right\| + \left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|\right)\left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|_F \\
=&\left(2\sqrt{\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|} + \left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|\right)\left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|_F \\
=&\left(2\sqrt{\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star)\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|} + \left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|\right)\left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|_F \\
\overset{(a)}{\leq}&\left(\frac{1}{20}\sqrt{\sigma_{\min}(\mathbf{X}_\star)} + \left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|_F\right)\left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|_F. \qquad (97)
\end{aligned}$$

In inequality (a) we used Assumption (33). By choosing the orthogonal matrix $\mathbf{R}$ as the minimizer of Procruste's problem, i.e., such that $\left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|_F$ is minimal, we obtain by Lemma 4.8 that

$$\left\|\mathbf{U}_t\mathbf{R} - \mathbf{U}_{t,\mathbf{w}}\right\|_F \leq \frac{\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F}{\sqrt{2\left(\sqrt{2}-1\right)\sigma_{\min}^2\left(\mathbf{U}_t\right)}} \overset{(a)}{\leq} \frac{\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F}{\sqrt{\left(\sqrt{2}-1\right)\frac{3}{2}\sigma_{\min}\left(\mathbf{X}_\star\right)}} \overset{(b)}{\leq} \frac{\sqrt{\sigma_{\min}(\mathbf{X}_\star)}}{20}.$$

Inequality (a) follows from Assumption (33) and Weyl's inequalities for singular values. For inequality (b) we used Assumption (34). Inequality (97) combined with this inequality chain yields that

$$\begin{aligned}
\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|_F \leq& \frac{\sqrt{\sigma_{\min}(\mathbf{X}_\star)}}{10}\cdot\frac{\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F}{\sqrt{\left(\sqrt{2}-1\right)\cdot\frac{3}{2}\sigma_{\min}\left(\mathbf{X}_\star\right)}} \\
\leq& \frac{\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F}{5}. \qquad (98)
\end{aligned}$$

In order to proceed we note that

$$\begin{aligned}
\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \leq&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star}\right\|_F \\
&+ \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|_F \\
\leq&2\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|_F \\
\overset{(a)}{\leq}&2\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + \frac{1}{5}\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F.
\end{aligned}$$

In inequality (a) we have used inequality (98). By rearranging terms we obtain that

$$\begin{aligned}
\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \leq& \frac{2}{1-\frac{1}{5}}\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \\
\leq& 3\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F.
\end{aligned}$$

This shows inequality (36). Then (35) follows directly from inserting the above inequality into (98). □

B.4. Proof of Lemma 4.5

The key idea in the proof of Lemma 4.5 is to decompose $\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\right)$ into a sum of the form

$$\begin{aligned}&\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\right)\\ =&\mathbf{V}_{\mathbf{X}_\star}^\top \left(1+\mu\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right)\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\left(1+\mu\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right)\\ &+\mathbf{V}_{\mathbf{X}_\star}^\top \boldsymbol{\Delta}.\end{aligned} \tag{99}$$

The first summand can be interpreted as a contraction mapping applied to the matrix $\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top$ and thus can be expected to have a smaller Frobenius norm than $\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\|_F$. In contrast, the term $\boldsymbol{\Delta}$, which will be determined explicitly in the proof of Lemma 4.5, can be interpreted as an additive error term which, as we will show, has relatively small Frobenius norm.

To deal with the first summand we need the following auxiliary lemma.

Lemma B.2. *Denote by $\lambda_{\max}(\mathbf{A})$ the largest eigenvalue of a symmetric matrix $\mathbf{A}$ and by $\lambda_{\min}(\mathbf{A})$ the smallest eigenvalue of $\mathbf{A}$. Assume that the assumptions of Lemma 4.5 are satisfied. Then it holds that*

$$\lambda_{\min}\left(\boldsymbol{Id} + \mu\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right) \geq 0, \tag{100}$$

$$\lambda_{\max}\left(\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star}\right) \leq -\frac{\sigma_{\min}\left(\mathbf{X}_\star\right)}{2}, \tag{101}$$

$$\left\|\boldsymbol{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right\| \leq 1 + \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{128}. \tag{102}$$

Proof of Lemma B.2. Note that the assumptions $\mu \leq \frac{c_4}{\kappa\|\mathbf{X}_\star\|}$, (38), and (40) together with Weyl's inequalities imply

$$\begin{aligned}&\lambda_{\min}\left(\mathbf{Id} + \mu\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right)\\ =&\lambda_{\min}\left(\mathbf{Id} + \mu\left((\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top) - \mathbf{U}_t\mathbf{U}_t^\top + \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right)\\ \geq&1 - \mu\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| - \mu\left\|\mathbf{U}_t\mathbf{U}_t^\top\right\| - \mu\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|\\ \geq&0.\end{aligned}$$

for sufficiently small $c_2, c_3, c_4 > 0$. This shows inequality (100).

We observe that

$$\begin{aligned}&\lambda_{\max}\left(\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star}\right)\\ \overset{(a)}{\leq}&\lambda_{\max}\left(-\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star}\right) + \left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|\\ \overset{(b)}{\leq}&\lambda_{\max}\left(-\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star}\right) + (c_2+c_3)\,\sigma_{\min}\left(\mathbf{X}_\star\right)\\ =&-\lambda_{\min}\left(\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_t}\mathbf{V}_{\mathbf{U}_t}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{U}_t}\mathbf{V}_{\mathbf{U}_t}^\top\mathbf{V}_{\mathbf{X}_\star}\right) + (c_2+c_3)\,\sigma_{\min}\left(\mathbf{X}_\star\right)\\ \leq&-\sigma_{\min}\left(\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_t}\right)^2\lambda_{\min}\left(\mathbf{U}_t\mathbf{U}_t^\top\right) + (c_2+c_3)\,\sigma_{\min}\left(\mathbf{X}_\star\right)\\ \overset{(c)}{\leq}&-\frac{\sigma_{\min}\left(\mathbf{X}_\star\right)}{2}.\end{aligned}$$

Inequality (a) follows from Weyl's inequalities. Inequality (b) follows from assumption (39) and (40). For inequality (c) we used assumptions (37), (39) for suffciently small c_1, c_2, c_3, and Weyl's inequalities. This proves inequality (101).

To prove inequality (102), we first establish an upper bound for the largest eigenvalue of $\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top$. For that let $\mathbf{x} \in \mathbb{R}^d$ be arbitrary. We use the orthogonal decomposition $\mathbf{x} =$

$\mathbf{x}_\| + \mathbf{x}_\perp$, where $\mathbf{x}_\|$ is the orthogonal projection of $\mathbf{x}$ onto the column span of $\mathbf{X}_\star$. We compute that

$$\begin{aligned}
&\mathbf{x}^\top \left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{x} \\
=&\mathbf{x}_\|^\top \left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{x}_\| - \mathbf{x}_\perp^\top \left(\mathbf{U}_t\mathbf{U}_t^\top + \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{x}_\perp - 2\mathbf{x}_\perp^\top \left(\mathbf{U}_t\mathbf{U}_t^\top + \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{x}_\| \\
\overset{(101)}{\leq}& -\frac{\sigma_{\min}\left(\mathbf{X}_\star\right)}{2}\left\|\mathbf{x}_\|\right\|_2^2 - 2\mathbf{x}_\perp^\top \left(\mathbf{U}_t\mathbf{U}_t^\top + \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{x}_\|. \qquad (103)
\end{aligned}$$

Next, we observe that

$$\begin{aligned}
-\mathbf{x}_\perp^\top \left(\mathbf{U}_t\mathbf{U}_t^\top + \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{x}_\| &\leq \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top + \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star}\right\|\left\|\mathbf{x}_\|\right\|_2\left\|\mathbf{x}_\perp\right\|_2 \\
&\leq \left(2\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star}\right\| + \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|\right)\left\|\mathbf{x}_\|\right\|_2\left\|\mathbf{x}_\perp\right\|_2 \\
&= \left(2\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star}\right\| + \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|\right)\left\|\mathbf{x}_\|\right\|_2\left\|\mathbf{x}_\perp\right\|_2 \\
&\leq \left(2\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|\right)\left\|\mathbf{x}_\|\right\|_2\left\|\mathbf{x}_\perp\right\|_2 \\
&\leq \frac{\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{x}_\|\right\|_2\left\|\mathbf{x}_\perp\right\|_2}{16}.
\end{aligned}$$

In the last inequality we have used the assumptions (39) and (40) for sufficiently small $c_2, c_3 > 0$. Combining this estimate with (103) we obtain that

$$\begin{aligned}
\mathbf{x}^\top \left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{x} &\leq \sigma_{\min}\left(\mathbf{X}_\star\right)\left(\frac{\left\|\mathbf{x}_\|\right\|_2\left\|\mathbf{x}_\perp\right\|_2}{8} - \frac{\left\|\mathbf{x}_\|\right\|_2^2}{2}\right) \\
&\leq \frac{\sigma_{\min}\left(\mathbf{X}_\star\right)\left\|\mathbf{x}_\perp\right\|_2^2}{128} \leq \frac{\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{x}\right\|_2^2}{128}.
\end{aligned}$$

This implies that

$$\lambda_{\max}\left(\mathbf{Id} + \mu\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right) \leq 1 + \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{128}.$$

This inequality, together with inequality (100), yields inequality (102). Thus, the proof of Lemma B.2 is complete. □

With Lemma B.2 in place, we can show that the first term in the decomposition (99) indeed has a smaller Frobenius norm than the term $\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)$.

Lemma B.3. *Assume that the assumptions of Lemma 4.5 are satisfied. Then, it holds that*

$$\begin{aligned}
&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\boldsymbol{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\left(\boldsymbol{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\right\|_F \\
&\leq \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{8}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F.
\end{aligned}$$

Proof of Lemma B.3. We first compute that

$$\begin{aligned}
&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\right\|_F \\
\leq&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F\left\|\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right\| \\
\leq&\left(1 + \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{128}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F, \qquad (104)
\end{aligned}$$

where in the last line we used inequality (102) from Lemma B.2. In order to proceed, we consider the decomposition

$$\begin{aligned}
&\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right) \\
&\underbrace{\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\mathbf{V}_{\mathbf{X}_\star}\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)}_{=:\mathbf{N}_1} \\
&\underbrace{-\mu\mathbf{V}_{\mathbf{X}_\star}^\top(\mathbf{U}_t\mathbf{U}_t^\top + \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\mathbf{V}_{\mathbf{X}_\star,\perp}\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star}\mathbf{V}_{\mathbf{X}_\star}^\top}_{=:\mathbf{N}_2} \\
&\underbrace{-\mu\mathbf{V}_{\mathbf{X}_\star}^\top(\mathbf{U}_t\mathbf{U}_t^\top + \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\mathbf{V}_{\mathbf{X}_\star,\perp}\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\mathbf{V}_{\mathbf{X}_\star,\perp}^\top}_{=:\mathbf{N}_3}.
\end{aligned}$$

We estimate the Frobenius norm of the three terms individually. For the first term we obtain that

$$\begin{aligned}\left\|\mathbf{N}_1\right\|_F &\leq \left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\mathbf{V}_{\mathbf{X}_\star}\right\|\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \\ &= \left\|\mathbf{Id} + \mu\mathbf{V}_{\mathbf{X}_\star}^\top(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\mathbf{V}_{\mathbf{X}_\star}\right\|\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \\ &\overset{(a)}{\leq} \left(1 + \mu\lambda_{\max}\left(\mathbf{V}_{\mathbf{X}_\star}^\top(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\mathbf{V}_{\mathbf{X}_\star}\right)\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \\ &\overset{(b)}{\leq} \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{2}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F,\end{aligned}$$

where in inequality (a) we have used (100) and in (b) we have used inequality (101) from Lemma B.2. The Frobenius norm of the term $\mathbf{N}_2$ can be estimated by

$$\begin{aligned}\left\|\mathbf{N}_2\right\|_F &\leq \left\|\mathbf{V}_{\mathbf{X}_\star}^\top(\mathbf{U}_t\mathbf{U}_t^\top + \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star}\right\|_F \\ &= \left(\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\left[2\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right) + \left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\right\|\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \\ &\leq \left(2\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star)\right\| + \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \\ &\leq \left(2c_2\sigma_{\min}(\mathbf{X}_\star) + \left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \\ &\leq \left(2c_2 + c_3\right)\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F,\end{aligned}$$

where we have used Assumptions (39) and (40). With similar arguments, we can estimate the Frobenius norm of the term $\mathbf{N}_3$ by

$$\left\|\mathbf{N}_3\right\|_F \leq \left(2c_2 + c_3\right)\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|_F.$$

By using Lemma 4.4 we obtain that

$$\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|_F \leq \frac{3\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F}{5}.$$

It follows that

$$\left\|\mathbf{N}_3\right\|_F \leq \frac{3\left(2c_2 + c_3\right)\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F}{5}.$$

By summing up our estimates for $\left\|\mathbf{N}_1\right\|_F$, $\left\|\mathbf{N}_2\right\|_F$, and $\left\|\mathbf{N}_3\right\|_F$ and choosing the constants $c_1, c_2 > 0$ small enough we obtain that

$$\begin{aligned}&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \\ &\leq \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{4}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F.\end{aligned}$$

Inserting this estimate into (104) yields that

$$\begin{aligned}&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\right\|_F \\ &\leq \left(1 + \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{128}\right)\left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{4}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \\ &\leq \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{8}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F,\end{aligned}$$

where in the last line, we used our assumption on the step size μ. This completes the proof of Lemma B.3. □

With the auxiliary estimates in Lemma B.3 we can give a proof of Lemma 4.5.

Proof of Lemma 4.5. First, we compute that

$$\begin{aligned}
\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top =& \left(\mathbf{Id} + \mu\left[\left(\mathcal{A}^*\mathcal{A}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\right)\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{Id} + \mu\left[\left(\mathcal{A}^*\mathcal{A}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\right)\\
=& \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\\
&+ \mu\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top + \mu\mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\\
&+ \mu^2\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) + \mu^2\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\\
&- \mu^2\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\\
&+ \mu\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\\
&+ \mu\left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\\
&+ \mu^2\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{U}_t\mathbf{U}_t^\top\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right].
\end{aligned}$$

Analogously, we can compute that

$$\begin{aligned}
&\mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\\
=& \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\\
&+ \mu\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top + \mu\mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\\
&+ \mu^2\mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right) + \mu^2\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top\\
&- \mu^2\mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top\\
&+ \mu\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\\
&+ \mu\left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\\
&+ \mu^2\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right].
\end{aligned}$$

Thus, we obtain that

$$\begin{aligned}
&\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\\
=& \mathbf{M}_1 + \mu^2\mathbf{M}_2 + \mu^2\mathbf{M}_3 + \mu^2\mathbf{M}_4 + \mu^2\mathbf{M}_4 + \mu\mathbf{M}_5 + \mu\mathbf{M}_6 + \mu^2\mathbf{M}_7, \qquad (105)
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{M}_1 :=& \left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\left(\mathbf{Id} + \mu(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top)\right)\\
\mathbf{M}_2 :=& \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) - \mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right),\\
\mathbf{M}_3 :=& \left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top,\\
\mathbf{M}_4 :=& \mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top,\\
\mathbf{M}_5 :=& \left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\\
&- \left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right),\\
\mathbf{M}_6 :=& \left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\\
&- \left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right],\\
\mathbf{M}_7 :=& \left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{U}_t\mathbf{U}_t^\top\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\\
&- \left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right].
\end{aligned}$$

Recall that Lemma B.3 shows that

$$\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{M}_1\right\|_F \le \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{8}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F.$$

To complete the proof, we need to derive upper bounds for $\left\|\mathbf{M}_i\right\|_F$, where $i = 2, 3, \ldots, 7$.

Estimating $\left\|\mathbf{M}_2\right\|_F$: We compute that

$$\begin{aligned}
\mathbf{M}_2 =& \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) - \mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\\
=& \left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) + \mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\\
&+ \mathbf{U}_t\mathbf{U}_t^\top\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right).
\end{aligned}$$

Thus, we obtain that

$$\begin{aligned}
&\big\|\mathbf{M}_2\big\|_F\\
\leq& 2\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\big\|_F\big\|\mathbf{U}_t\mathbf{U}_t^\top\big\|\big\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\big\| + \big\|\mathbf{U}_t\mathbf{U}_t^\top\big\|\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\big\|_F\\
\leq& 2\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\big\|_F\big\|\mathbf{U}_t\mathbf{U}_t^\top\big\|\big\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\big\|\\
&+ \big\|\mathbf{U}_t\mathbf{U}_t^\top\big\|\left(\big\|\mathbf{U}_t\mathbf{U}_t^\top\big\| + \big\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|\right)\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\big\|_F\\
\leq& 5\big\|\mathbf{X}_\star\big\|^2\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\big\|_F.
\end{aligned}$$

In the last inequality we used assumptions (38), (39), and (40) for sufficiently small $c_2, c_3 > 0$.

Estimating $\big\|\mathbf{M}_3\big\|_F$: Since $\mathbf{M}_3 = \mathbf{M}_2^\top$ it follows that

$$\big\|\mathbf{M}_3\big\|_F \leq 5\|\mathbf{X}_\star\|^2\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\big\|_F.$$

Estimating $\big\|\mathbf{M}_4\big\|_F$: We compute that

$$\begin{aligned}
\mathbf{M}_4 =& \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top + \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top\\
&+\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right).
\end{aligned}$$

Again, using the assumptions (38) and (40), and the triangle inequality we obtain that

$$\big\|\mathbf{M}_4\big\|_F \leq 20\big\|\mathbf{X}_\star\big\|^2\big\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|_F.$$

Estimating $\big\|\mathbf{M}_5\big\|_F$: We compute

$$\begin{aligned}
\mathbf{M}_5 =& \underbrace{\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)}_{=:\mathbf{O}_1}\\
&+\mu\underbrace{\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)}_{=:\mathbf{O}_2}\\
&+\underbrace{\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)}_{=:\mathbf{O}_3}\\
&+\underbrace{\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left(\mathbf{Id} + \mu\mathbf{X}_\star - \mu\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)}_{=:\mathbf{O}_4}.
\end{aligned}$$

We estimate the Frobenius norm of these summands individually. For the first term we observe that

$$\begin{aligned}
\big\|\mathbf{O}_1\big\|_F \leq& \big\|\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\big\|\big\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|_F\left(1 + \mu\big\|\mathbf{X}_\star\big\| + \mu\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|\right)\\
\overset{(a)}{\leq}& 2\big\|\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\big\|\big\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|_F\\
\overset{(b)}{\leq}& 2c_5\sigma_{\min}(\mathbf{X}_\star)\big\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|_F,
\end{aligned}$$

where in inequality (a) we have used assumptions (38), (40), and the assumption on the step size μ. In inequality (b) we have used assumption (41).

Using again assumptions (38), (40), and (41) we obtain that

$$\big\|\mathbf{O}_2\big\|_F \leq 3c_5\sigma_{\min}(\mathbf{X}_\star)\big\|\mathbf{X}_\star\big\|\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\big\|_F.$$

For the term $\big\|\mathbf{O}_3\big\|_F$ we obtain that

$$\begin{aligned}
\big\|\mathbf{O}_3\big\|_F \leq& \big\|\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\big\|_F\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|\left(1 + \mu\big\|\mathbf{X}_\star\big\| + \mu\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|\right)\\
\leq& \big\|\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\big\|_F\left(\big\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\| + \big\|\mathbf{U}_t\mathbf{U}_t^\top\big\|\right)\\
&\left(1 + \mu\big\|\mathbf{X}_\star\big\| + \mu\big\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\big\| + \mu\big\|\mathbf{U}_t\mathbf{U}_t^\top\big\|\right)\\
\overset{(a)}{\leq}& 4\big\|\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\big\|_F\big\|\mathbf{X}_\star\big\|\\
\overset{(b)}{\leq}& 4\left(\delta + \frac{8\sqrt{rd}}{\sqrt{m}}\right)\big\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\big\|\big\|\mathbf{X}_\star\big\| + 4\left(\delta + \frac{8\sqrt{2d}}{\sqrt{m}}\right)\big\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\big\|_F\big\|\mathbf{X}_\star\big\|.
\end{aligned}$$

Inequality (a) follows from the assumptions (38) and (40), and the assumption on the step size μ. In inequality (b) we used the estimate (89) from Lemma B.1.

For the term $\|\mathbf{O}_4\|_F$ we obtain that

$$\begin{aligned}\|\mathbf{O}_4\|_F \leq& \left\| \left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F \left(\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \left\|\mathbf{U}_t\mathbf{U}_t^\top\right\|\right)\\ &\cdot\left(1+\mu\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \mu\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|\right)\\ \overset{(a)}{\leq}& 3\left\|\mathbf{X}_\star\right\|\left\|\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_F\\ \overset{(b)}{\leq}& 6\delta\left\|\mathbf{X}_\star\right\|\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right\|_F.\end{aligned}$$

Inequality (a) follows from assumptions (39) and (40), and the assumption on the step size μ. Inequality (b) is due to inequality (90) in Lemma B.1. By summing up all terms we obtain that

$$\begin{aligned}&\left\|\mathbf{M}_5\right\|_F \leq \left\|\mathbf{O}_1\right\|_F + \mu\left\|\mathbf{O}_2\right\|_F + \left\|\mathbf{O}_3\right\|_F + \left\|\mathbf{O}_4\right\|_F\\ \leq& 2c_5\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + 3\mu c_5\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star\right\|\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right\|_F\\ &+ 4\left(\delta + \frac{8\sqrt{rd}}{\sqrt{m}}\right)\left\|\mathbf{X}_\star\right\|\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + 4\left(\delta + \frac{4\sqrt{2d}}{\sqrt{m}}\right)\left\|\mathbf{X}_\star\right\|\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F\\ &+ 6\delta\left\|\mathbf{X}_\star\right\|\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right\|_F\\ =& \left[\left(\left(2+3\mu\right)c_5 + 6\kappa\delta\right)\sigma_{\min}(\mathbf{X}_\star) + 4\left(\delta + \frac{4\sqrt{2d}}{\sqrt{m}}\right)\left\|\mathbf{X}_\star\right\|\right]\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F\\ &+ 4\left(\delta + \frac{8\sqrt{rd}}{\sqrt{m}}\right)\left\|\mathbf{X}_\star\right\|\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|\\ \overset{(a)}{\leq}& \left(\left(\left(2+3\mu\right)c_5 + 6c_6\right)\sigma_{\min}(\mathbf{X}_\star) + 8c_6\sigma_{\min}(\mathbf{X}_\star)\right)\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + 8c_6\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|\\ \overset{(b)}{\leq}& \frac{\sigma_{\min}(\mathbf{X}_\star)}{100}\cdot\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F + 8c_6\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|\\ \overset{(c)}{\leq}& \frac{3\sigma_{\min}(\mathbf{X}_\star)}{100}\cdot\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + 8c_6\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|,\end{aligned}$$

where in inequality (a) we used the assumption (42). Inequality (b) follows from choosing the constants c_5 and c_6 small enough. To obtain inequality (c) we applied Lemma 4.4.

Estimating $\|\mathbf{M}_6\|_F$:
Since $\mathbf{M}_6 = \mathbf{M}_5^\top$ we obtain that

$$\left\|\mathbf{M}_6\right\|_F \leq \frac{3\sigma_{\min}(\mathbf{X}_\star)}{100}\cdot\left\|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + 8c_6\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|.$$

Estimating $\|\mathbf{M}_7\|_F$: To deal with the term $\mathbf{M}_7$ we first compute that

$$\begin{aligned}\mathbf{M}_7 =& \underbrace{\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]}_{=:\mathbf{L}_1}\\ &+ \underbrace{\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]}_{=:\mathbf{L}_2}\\ &+ \underbrace{\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]}_{=:\mathbf{L}_3}\\ &+ \underbrace{\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]}_{=:\mathbf{L}_4}\\ &+ \underbrace{\left[\left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]}_{=:\mathbf{L}_5}.\end{aligned}$$

We estimate the Frobenius norm of the summands individually. For $\|\mathbf{L}_1\|_F$ we obtain that

$$\begin{aligned}\|\mathbf{L}_1\|_F &\leq \| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) \| \|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\|_F \| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) \| \\ &\leq c_5^2\sigma_{\min}(\mathbf{X}_\star)^2\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\|_F,\end{aligned}$$

where we have used assumption (41). Next, we note that

$$\begin{aligned}\|\mathbf{L}_2\|_F \leq& \| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\|_F \left(\|\mathbf{U}_t\mathbf{U}_t^\top\| + \|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\|\right) \\ &\cdot \| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) \| \\ \overset{(a)}{\leq}& 3c_5\sigma_{\min}\left(\mathbf{X}_\star\right)\|\mathbf{X}_\star\| \| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\|_F \\ \overset{(b)}{\leq}& 3c_5\delta\sigma_{\min}(\mathbf{X}_\star)\|\mathbf{X}_\star\|\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\|_F \\ \overset{(c)}{\leq}& 3c_5c_6\sigma_{\min}^2(\mathbf{X}_\star)\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\|_F.\end{aligned}$$

Inequality (a) follows from assumptions (38), (40), and (41). Inequality (b) is due to Lemma 2.4 and inequality (c) is due to assumption (42). In order to estimate $\|\mathbf{L}_3\|_F$ we note that

$$\begin{aligned}\|\mathbf{L}_3\|_F &\left(\| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) \| + \| \left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\|_F\right) \\ &\cdot \left(\|\mathbf{U}_t\mathbf{U}_t^\top\| + \|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\|_F\right) \| \left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\|_F \\ \overset{(a)}{\leq}& \left(c_5\sigma_{\min}(\mathbf{X}_\star) + \delta\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\|_F\right)\left(2\|\mathbf{X}_\star\| + c_3\sigma_{\min}(\mathbf{X}_\star)\right)\delta\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\|_F \\ \overset{(b)}{\leq}& 3\left(c_5 + \delta c_3\right)\delta\sigma_{\min}\left(\mathbf{X}_\star\right)\|\mathbf{X}_\star\|\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\|_F \\ \overset{(c)}{\leq}& 3c_6\left(c_5 + \delta c_3\right)\sigma_{\min}^2\left(\mathbf{X}_\star\right)\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\|_F.\end{aligned}$$

In inequality (a) we used the assumptions (38), (40), (41), and Lemma 2.4. Inequality (b) follows from assumption (40) and since the constant $c_3 > 0$ is chosen small enough. Inequality (c) is due to assumption (42).

Next, we can estimate $\|\mathbf{L}_4\|_F$ by

$$\begin{aligned}\|\mathbf{L}_4\|_F \leq& \| \left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\|_F \left(\|\mathbf{U}_t\mathbf{U}_t^\top\| + \|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\|\right) \\ &\cdot \left(\| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) \| + \| \left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right) \|\right) \\ \overset{(a)}{\leq}& \| \left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\|_F \left(2\|\mathbf{X}_\star\| + c_3\sigma_{\min}(\mathbf{X}_\star)\right) \\ &\cdot \left(c_5\sigma_{\min}(\mathbf{X}_\star) + \delta\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\|_F\right) \\ \overset{(b)}{\leq}& 3\left(c_5 + c_3\delta\right)\sigma_{\min}(\mathbf{X}_\star)\|\mathbf{X}_\star\|\| \left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\|_F \\ \overset{(c)}{\leq}& 3\left(c_5 + c_3\delta\right)\left(\delta + 8\sqrt{\frac{rd}{m}}\right)\sigma_{\min}(\mathbf{X}_\star)\|\mathbf{X}_\star\|\|\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\| \\ \leq& 3\left(c_5 + c_3\delta\right)\left(\delta + 8\sqrt{\frac{rd}{m}}\right)\sigma_{\min}(\mathbf{X}_\star)\|\mathbf{X}_\star\|\left(\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\| + \|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\|\right) \\ \overset{(d)}{\leq}& 6c_6\left(c_5 + c_3\delta\right)\sigma_{\min}^2\left(\mathbf{X}_\star\right)\left(\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\| + \|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\|\right).\end{aligned}$$

In inequality (a) we used assumptions (38), (40), and (41) as well as Lemma 2.4. Inequality (b) uses assumption (40). Inequality (c) follows from inequality (91) in Lemma B.1. Inequality (d) is due to assumption (42).

The norm $\|\mathbf{L}_5\|_F$ can be estimated by

$$\begin{aligned}\|\mathbf{L}_5\|_F \leq& \| \left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right) \|\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\|\|\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}^\top\left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\|_F \\ \overset{(a)}{\leq}& 3\|\mathbf{X}_\star\|\| \left(\mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right) \|\| \left[\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}_{\mathbf{w}}^*\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_\star - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\|_F.\end{aligned} \tag{106}$$

In inequality (a) we used the triangle inequality and the assumptions (38), (40). In order to proceed, we note first that

$$
\begin{aligned}
&\left\| \left(\mathcal{A}_{\mathbf{w}}^{*}\mathcal{A}_{\mathbf{w}} - \mathcal{I}\right)\left(\mathbf{X}_{\star} - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top}\right)\right\| \\
&\overset{(a)}{\leq} \left\| \left(\mathcal{A}^{*}\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_{\star} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right)\right\| + \left(\delta + 8\sqrt{\frac{rd}{m}}\right)\left\|\mathbf{X}_{\star} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\| \\
&\quad + \left(2\delta + 4\sqrt{\frac{2d}{m}}\right)\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\|_{F} \\
&\overset{(b)}{\leq} \left\| \left(\mathcal{A}^{*}\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_{\star} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right)\right\| + \frac{2c_6}{\kappa}\left\|\mathbf{X}_{\star} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\| + \frac{3c_6}{\kappa}\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\|_{F} \\
&\overset{(c)}{\leq} \left(c_5 + \frac{2c_2c_6}{\kappa} + \frac{3c_3c_6}{\kappa}\right)\sigma_{\min}(\mathbf{X}_{\star}),
\end{aligned}
$$

where in inequality (a) we used Lemma B.1. Inequality (b) follows from the assumptions (42). Inequality (c) is due to assumption (39), (40), and (41). Moreover, it holds that

$$
\begin{aligned}
\left\| \left[\left(\mathcal{A}^{*}\mathcal{A} - \mathcal{A}_{\mathbf{w}}^{*}\mathcal{A}_{\mathbf{w}}\right)\left(\mathbf{X}_{\star} - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top}\right)\right]\mathbf{V}_{\mathbf{U}_{t,\mathbf{w}}}\right\|_{F} &\overset{(a)}{\leq} \left(\delta + 8\sqrt{\frac{rd}{m}}\right)\left\|\mathbf{X}_{\star} - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top}\right\| \\
&\overset{(b)}{\leq} \frac{2c_6}{\kappa}\left(\left\|\mathbf{X}_{\star} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\| + \left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\|\right).
\end{aligned}
$$

Inequality (a) follows from inequality (91) in Lemma B.1. Inequality (b) is due to assumption (42). Inserting the last two inequality chains into inequality (106) we obtain that

$$
\left\|\mathbf{L}_5\right\|_{F} \leq 6c_6\left(c_5 + \frac{2c_2c_6}{\kappa} + \frac{3c_3c_6}{\kappa}\right)\sigma_{\min}^{2}(\mathbf{X}_{\star})\left(\left\|\mathbf{X}_{\star} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\| + \left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\|\right)
$$

By summing up all terms $\left\|\mathbf{L}_i\right\|_{F}$ for $i = 1, \ldots, 5$ it follows that

$$
\begin{aligned}
\left\|\mathbf{M}_7\right\|_{F} \leq& c_5^2\sigma_{\min}^{2}(\mathbf{X}_{\star})\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\|_{F} \\
&+ 3c_5c_6\sigma_{\min}^{2}(\mathbf{X}_{\star})\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\|_{F} \\
&+ 3c_6\left(c_5 + c_3\delta\right)\sigma_{\min}^{2}\left(\mathbf{X}_{\star}\right)\left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\|_{F} \\
&+ 6c_6\left(c_5 + c_3\delta\right)\sigma_{\min}^{2}\left(\mathbf{X}_{\star}\right)\left(\left\|\mathbf{X}_{\star} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\| + \left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\|\right) \\
&+ 6c_6\left(c_5 + \frac{2c_2c_6}{\kappa} + \frac{3c_3c_6}{\kappa}\right)\sigma_{\min}^{2}(\mathbf{X}_{\star})\left(\left\|\mathbf{X}_{\star} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\| + \left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\|\right) \\
\leq& \sigma_{\min}^{2}(\mathbf{X}_{\star})\left(\left\|\mathbf{X}_{\star} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\| + \left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^{\top} - \mathbf{U}_{t}\mathbf{U}_{t}^{\top}\right\|_{F}\right),
\end{aligned}
$$

where the last inequality holds since the absolute constants $c_3, c_5, c_6 > 0$ are chosen small enough.

Using the decomposition (105), the triangle inequality, combined with our estimates for $\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \mathbf{M}_1\right\|_F$ and for $\left\|\mathbf{M}_i\right\|_F$, where $2 \leq i \leq 7$, we obtain that

$$
\begin{aligned}
&\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top - \mathbf{U}_{t+1,\mathbf{w}}\mathbf{U}_{t+1,\mathbf{w}}^\top\right)\right\|_F \\
&\leq \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{8}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + 30\mu^2\left\|\mathbf{X}_\star\right\|^2\left\|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right\|_F \\
&\quad + \frac{3\mu\sigma_{\min}(\mathbf{X}_\star)}{50}\cdot\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + 16\mu c_6\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \\
&\quad + \mu^2\sigma_{\min}^2(\mathbf{X}_\star)\left(\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \left\|\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right\|_F\right) \\
&\overset{(a)}{\leq} \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{8}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + 90\mu c_4\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F \\
&\quad + \frac{3\mu\sigma_{\min}(\mathbf{X}_\star)}{50}\cdot\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + 16\mu c_6\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \\
&\quad + \mu^2\sigma_{\min}^2(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \frac{3\mu c_4\sigma_{\min}\left(\mathbf{X}_\star\right)}{\kappa}\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\|_F \\
&\overset{(b)}{\leq} \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + \mu\left(16c_6 + \mu\sigma_{\min}(\mathbf{X}_\star)\right)\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \\
&\leq \left(1 - \frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{16}\right)\left\|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_{t,\mathbf{w}}\mathbf{U}_{t,\mathbf{w}}^\top\right)\right\|_F + \mu\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|,
\end{aligned}
$$

where inequality (a) is due to Lemma 4.4 and the assumption on the step size μ. Inequality (b) is obtained by choosing $c_4 < 1/2$, and the last inequality is obtained by choosing $c_6 < \frac{1}{32}$. □

C. Proof of the lemmas controlling the distance between $\mathbf{X}_\star$ and $\mathbf{U}_t\mathbf{U}_t^\top$ (Lemma 4.6, Lemma 4.7, and Lemma 4.9)

C.1. Proof of Lemma 4.6

Proof of Lemma 4.6. We first note that

$$
\begin{aligned}
\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star,\perp} &= \mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\mathbf{V}_{\mathbf{U}_t}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star,\perp} \\
&= \mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\left(\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_t}\right)^{-1}\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_t}\mathbf{V}_{\mathbf{U}_t}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star,\perp} \\
&= \mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\left(\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_t}\right)^{-1}\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star,\perp} \\
&= \mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\left(\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_t}\right)^{-1}\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star,\perp}.
\end{aligned}
$$

Using the submultiplicativity property of the $|||\cdot|||$-norm it follows that

$$
\begin{aligned}
\left|\left|\left|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star,\perp}\right|\right|\right| &\leq \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\right\|\left\|\left(\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_t}\right)^{-1}\right\|\left|\left|\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right|\right|\right| \\
&= \frac{\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\right\|}{\sigma_{\min}(\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_t})}\left|\left|\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right|\right|\right|.
\end{aligned}
$$

Recall that

$$
\sigma_{\min}^2\left(\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_t}\right) = 1 - \left\|\mathbf{V}_{\mathbf{U}_t}^\top\mathbf{V}_{\mathbf{X}_\star,\perp}\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\right\| = 1 - \left\|\mathbf{V}_{\mathbf{U}_t}^\top\mathbf{V}_{\mathbf{X}_\star,\perp}\right\|^2 \geq \frac{1}{4},
$$

where in the last inequality, we used assumption (44). It follows that

$$
\left|\left|\left|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star,\perp}\right|\right|\right| \leq 2\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\right\|\left|\left|\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right|\right|\right|.
$$

This proves inequality (45). To prove inequality (46) we note that

$$
\begin{aligned}
\left|\left|\left|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right|\right|\right| &\leq \left|\left|\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\right|\right|\right| + \left|\left|\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right|\right|\right| \\
&\quad + \left|\left|\left|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right|\right|\right| \\
&\leq 2\left|\left|\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\right|\right|\right| + \left|\left|\left|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star,\perp}\right|\right|\right| \\
&\leq 2\left(1 + \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\right\|\right)\left|\left|\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}_\star\right)\right|\right|\right|,
\end{aligned}
$$

where in the last inequality we used (45). This completes the proof of Lemma 4.6. □

C.2. Proof of Lemma 4.7

Proof of Lemma 4.7. We define the shorthand notation

$$\mathbf{M}_t := (\mathcal{A}^*\mathcal{A})\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) = \mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top + \underbrace{(\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)}_{=:\mathbf{E}_t}.$$

Thus, we have that

$$\mathbf{U}_{t+1} = (\mathbf{Id} + \mu\mathbf{M}_t)\,\mathbf{U}_t.$$

We compute that

$$\begin{aligned}
&\mathbf{X}_\star - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top\\
=&\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mu\mathbf{M}_t\mathbf{U}_t\mathbf{U}_t^\top - \mu\mathbf{U}_t\mathbf{U}_t^\top\mathbf{M}_t - \mu^2\mathbf{M}_t\mathbf{U}_t\mathbf{U}_t^\top\mathbf{M}_t\\
=&\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \mu\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top - \mu\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) - \mu\mathbf{E}_t\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{U}_t\mathbf{U}_t^\top\mathbf{E}_t\\
&- \mu^2\mathbf{M}_t\mathbf{U}_t\mathbf{U}_t^\top\mathbf{M}_t\\
=&\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right) - \mu^2\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top\\
&- \mu\mathbf{E}_t\mathbf{U}_t\mathbf{U}_t^\top - \mu\mathbf{U}_t\mathbf{U}_t^\top\mathbf{E}_t - \mu^2\mathbf{M}_t\mathbf{U}_t\mathbf{U}_t^\top\mathbf{M}_t.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top\right)\\
=&\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{V}_{\mathbf{X}_\star}\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\\
&+ \mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\\
&- \mu^2\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top - \mu\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{E}_t\mathbf{U}_t\mathbf{U}_t^\top - \mu\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{E}_t - \mu^2\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{M}_t\mathbf{U}_t\mathbf{U}_t^\top\mathbf{M}_t\\
=&\underbrace{\left(\mathbf{Id} - \mu\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star}\right)\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)}_{=:(I)}\\
&+ \underbrace{\mu\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star,\perp}\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)}_{=:(II)}\\
&- \underbrace{\left(\mu^2\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top + \mu\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{E}_t\mathbf{U}_t\mathbf{U}_t^\top + \mu\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{E}_t + \mu^2\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{M}_t\mathbf{U}_t\mathbf{U}_t^\top\mathbf{M}_t\right)}_{=:(III)}.
\end{aligned}$$

We estimate the spectral norm of these terms individually.

* Estimating term (I): We obtain that

$$\begin{aligned}
&\left|\!\left|\!\left|\left(\mathbf{Id} - \mu\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star}\right)\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right|\\
\overset{(a)}{\leq}&\left\|\mathbf{Id} - \mu\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star}\right\|\left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right|\left\|\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right\|\\
\overset{(b)}{\leq}&\left\|\mathbf{Id} - \mu\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t\mathbf{U}_t^\top\mathbf{V}_{\mathbf{X}_\star}\right\|\left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right|\\
\overset{(c)}{=}&\left(1 - \mu\sigma_{\min}^2(\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{U}_t)\right)\left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right|\\
\leq&\left(1 - \mu\left(\sigma_{\min}(\mathbf{V}_{\mathbf{X}_\star}^\top\mathbf{V}_{\mathbf{U}_t})\sigma_{\min}\left(\mathbf{U}_t\right)\right)^2\right)\left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right|\\
\overset{(d)}{\leq}&\left(1 - \frac{\mu}{2}\sigma_{\min}^2\left(\mathbf{U}_t\right)\right)\left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right|\\
\overset{(e)}{\leq}&\left(1 - \frac{\mu}{4}\sigma_{\min}\left(\mathbf{X}_\star\right)\right)\left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right|.
\end{aligned}$$

Inequality (a) is due to the submultiplicativity of the $|||\cdot|||$-norm. In inequality (b) and equality (c) we used the assumptions $\left\|\mathbf{U}_t\right\| \leq \sqrt{2\left\|\mathbf{X}_\star\right\|}$ and $\mu \leq \frac{1}{1024\kappa\left\|\mathbf{X}_\star\right\|}$. In inequality (d) we used assumption (47). Inequality (e) follows from assumption (48), which, due to Weyl's inequality, implies $\sigma_{\min}^2\left(\mathbf{U}_t\right) \geq \frac{1}{2}\sigma_{\min}(\mathbf{X}_\star)$.

* Estimating term (II): We note that

$$
\begin{aligned}
&\left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \mathbf{U}_t\mathbf{U}_t^\top \mathbf{V}_{\mathbf{X}_\star,\perp}\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \mathbf{U}_t\mathbf{U}_t^\top \left(\mathbf{Id}-\mu\mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right| \\
&= \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\left(\mathbf{Id}-\mu\mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right| \\
&\overset{(a)}{\leq} \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right\| \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right|\!\right|\!\right| \left\|\mathbf{Id}-\mu\mathbf{U}_t\mathbf{U}_t^\top\right\| \\
&\overset{(b)}{\leq} \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right\| \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right|\!\right|\!\right| \\
&\leq \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right\| \left(\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star}\right\| + \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right|\!\right|\!\right|\right) \\
&\leq \left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right\| \left(\left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right|\!\right|\!\right| + \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\mathbf{V}_{\mathbf{X}_\star,\perp}\right|\!\right|\!\right|\right) \\
&\overset{(c)}{\leq} 2\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right\| \left(1+\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\right\|\right) \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right|\!\right|\!\right| \\
&\overset{(d)}{\leq} 3\left\|\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right\| \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right|\!\right|\!\right| .
\end{aligned}
$$

In inequality (a) we used the submultiplicativity of the $|||\cdot|||$-norm. Inequality (b) follows from the assumption $\left\|\mathbf{U}_t\right\| \leq \sqrt{2\left\|\mathbf{X}_\star\right\|}$ and $\mu \leq \frac{1}{1024\kappa\left\|\mathbf{X}_\star\right\|}$. In inequality (c), we used Lemma 4.6. In inequality (d) we used the assumption $\left\|\mathbf{V}_{\mathbf{X}_\star,\perp}^\top\mathbf{V}_{\mathbf{U}_t}\right\| \leq \frac{1}{2}$. Thus, by using the assumption $\left\|\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right\| \leq \frac{\sigma_{\min}(\mathbf{X}_\star)}{48}$ it follows that

$$
\left|\!\left|\!\left|(II)\right|\!\right|\!\right| \leq \frac{\sigma_{\min}\left(\mathbf{X}_\star\right)}{16} \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{U}_t\mathbf{U}_t^\top-\mathbf{X}_\star\right)\right|\!\right|\!\right| .
$$

* Estimating term (III): We first note that

$$
\begin{aligned}
\left|\!\left|\!\left|\mathbf{M}_t\mathbf{V}_{\mathbf{U}_t}\right|\!\right|\!\right| &\overset{(a)}{\leq} \left|\!\left|\!\left|\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right|\!\right|\!\right| + \left|\!\left|\!\left|\left[\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right)\right]\mathbf{V}_{\mathbf{U}_t}\right|\!\right|\!\right| \\
&\overset{(b)}{\leq} 4\left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right| + \left|\!\left|\!\left|\mathbf{E}_t\mathbf{V}_{\mathbf{U}_t}\right|\!\right|\!\right| , \qquad (107)
\end{aligned}
$$

where (a) follows from the triangle inequality and (b) follows from Lemma 4.6. Moreover, we have that

$$
\left\|\mathbf{M}_t\right\| \leq \left\|\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right\| + \left\|\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right)\right\| \overset{(a)}{\leq} \sigma_{\min}\left(\mathbf{X}_\star\right) . \qquad (108)
$$

Inequality (a) follows from assumptions (48) and (49). Thus, we obtain for term (III) that

$$
\begin{aligned}
\left|\!\left|\!\left|(III)\right|\!\right|\!\right| \leq& \mu^2\left\|\mathbf{U}_t\right\|^4 \left|\!\left|\!\left|\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right|\!\right|\!\right| + 2\mu\left\|\mathbf{U}_t\right\|^2 \left|\!\left|\!\left|\mathbf{E}_t\mathbf{V}_{\mathbf{U}_t}\right|\!\right|\!\right| + \mu^2\left\|\mathbf{U}_t\right\|^2 \left|\!\left|\!\left|\mathbf{M}_t\mathbf{V}_{\mathbf{U}_t}\right|\!\right|\!\right| \left\|\mathbf{M}_t\right\| \\
&\overset{(a)}{\leq} 16\mu^2\left\|\mathbf{X}_\star\right\|^2 \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right| + 4\mu\left\|\mathbf{X}_\star\right\| \left|\!\left|\!\left|\mathbf{E}_t\mathbf{V}_{\mathbf{U}_t}\right|\!\right|\!\right| + 2\mu^2\sigma_{\min}\left(\mathbf{X}_\star\right)\left\|\mathbf{X}_\star\right\| \left|\!\left|\!\left|\mathbf{M}_t\mathbf{V}_{\mathbf{U}_t}\right|\!\right|\!\right| \\
&\overset{(b)}{\leq} \left(16\mu^2\left\|\mathbf{X}_\star\right\|^2 + 8\mu^2\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star\right\|\right) \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right| \\
&\quad + \left(4\mu\left\|\mathbf{X}_\star\right\| + 2\mu^2\sigma_{\min}(\mathbf{X}_\star)\left\|\mathbf{X}_\star\right\|\right) \left|\!\left|\!\left|\mathbf{E}_t\mathbf{V}_{\mathbf{U}_t}\right|\!\right|\!\right| \\
&\overset{(c)}{\leq} \frac{\mu\sigma_{\min}\left(\mathbf{X}_\star\right)}{16} \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right| + 5\mu\left\|\mathbf{X}_\star\right\| \left|\!\left|\!\left|\mathbf{E}_t\mathbf{V}_{\mathbf{U}_t}\right|\!\right|\!\right| .
\end{aligned}
$$

In inequality (a) we used the assumption $\left\|\mathbf{U}_t\right\| \leq \sqrt{2\left\|\mathbf{X}_\star\right\|}$, Lemma 4.6, and inequality (108). Inequality (b) is due to inequalities (107). In inequality (c) we used the assumption that $\mu \leq \frac{1}{1024\kappa\left\|\mathbf{X}_\star\right\|}$.

* Conclusion: By adding up all terms, we obtain that

$$
\begin{aligned}
\left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{X}_\star-\mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top\right)\right|\!\right|\!\right| &\leq \left|\!\left|\!\left|(I)\right|\!\right|\!\right| + \mu\left|\!\left|\!\left|(II)\right|\!\right|\!\right| + \left|\!\left|\!\left|(III)\right|\!\right|\!\right| \\
&\leq \left(1-\frac{\mu\sigma_{\min}(\mathbf{X}_\star)}{8}\right) \left|\!\left|\!\left|\mathbf{V}_{\mathbf{X}_\star}^\top \left(\mathbf{X}_\star-\mathbf{U}_t\mathbf{U}_t^\top\right)\right|\!\right|\!\right| + 5\mu\left\|\mathbf{X}_\star\right\| \left|\!\left|\!\left|\mathbf{E}_t\mathbf{V}_{\mathbf{U}_t}\right|\!\right|\!\right| .
\end{aligned}
$$

This completes the proof. □

C.3. Proof of Lemma 4.9

Proof of Lemma 4.9. Analogously, as in the proof of Lemma 4.7 we define the shorthand notation

$$\mathbf{M}_t := \left(\mathcal{A}^*\mathcal{A}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right) = \mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top - \underbrace{\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)}_{=:\mathbf{E}_t}.$$

We note that

$$\left\|\mathbf{M}_t\right\| \leq \left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + \left\|\left(\mathcal{A}^*\mathcal{A} - \mathcal{I}\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\right\| \leq (c_2 + c_3)\sigma_{\min}(\mathbf{X}_\star).$$

With an analogous computation as in the proof of Lemma 4.7, it follows that

$$\begin{aligned}\mathbf{X}_\star - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top = &\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right)\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\left(\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right) - \mu^2\mathbf{U}_t\mathbf{U}_t^\top\left(\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right)\mathbf{U}_t\mathbf{U}_t^\top \\ &- \mu\mathbf{E}_t\mathbf{U}_t\mathbf{U}_t^\top - \mu\mathbf{U}_t\mathbf{U}_t^\top\mathbf{E}_t - \mu^2\mathbf{M}_t\mathbf{U}_t\mathbf{U}_t^\top\mathbf{M}_t.\end{aligned}$$

When $c_1 \leq 1/2$, we have $\left\|\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right\| \leq 1$ by assumption (50). It follows from the assumptions $\mu \leq \frac{c_1}{\left\|\mathbf{X}_\star\right\|}$, (51), and (52) that for sufficiently small $c_1, c_2, c_3 > 0$

$$\begin{aligned}\left\|\mathbf{X}_\star - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top\right\| \leq &\left\|\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right\|\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\|\left\|\mathbf{Id} - \mu\mathbf{U}_t\mathbf{U}_t^\top\right\| + \mu^2\left\|\mathbf{U}_t\right\|^4\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| \\ &+ 2\mu\left\|\mathbf{E}_t\right\|\left\|\mathbf{U}_t\right\|^2 + \mu^2\left\|\mathbf{M}_t\right\|^2\left\|\mathbf{U}_t\right\|^2 \\ \leq &\left\|\mathbf{X}_\star - \mathbf{U}_t\mathbf{U}_t^\top\right\| + 4\mu^2c_2\left\|\mathbf{X}_\star\right\|^2\sigma_{\min}\left(\mathbf{X}_\star\right) + 4\mu c_3\left\|\mathbf{X}_\star\right\|\sigma_{\min}(\mathbf{X}_\star) \\ &+ 2(c_2 + c_3)^2\mu^2\left\|\mathbf{X}_\star\right\|\sigma_{\min}^2(\mathbf{X}_\star) \\ \leq &\left(c_2 + 4c_1^2c_2 + 4c_1c_3 + 2(c_2 + c_3)^2c_1^2\right)\sigma_{\min}(\mathbf{X}_\star) \\ \leq &\left(1 - \frac{1}{\sqrt{2}}\right)\sigma_{\min}(\mathbf{X}_\star).\end{aligned}$$

This completes the proof. □

D. Proofs regarding the Restricted Isometry Property and its consequences

D.1. Proof of Lemma 2.2

As already mentioned in Section 2.1, there exist similar versions of Lemma 2.1 in the literature (see, e.g., [50]), which, however, do not specify the dependence of the number of samples m on the constant $\delta > 0$. It would be possible to trace the steps of the ε-net argument in [50] and work out the δ-dependence explicitly. However, this would lead to an extra $\log(1/\delta)$-factor, which is unnecessary. The reason is that as δ is decreased, a covering with smaller balls is required, leading to a larger ε-net. This observation suggests a proof strategy based on generic chaining. Indeed, we will use the following general theorem from [62], which is proven via the generic chaining technique. To state it, we define the diameter of a set of matrices $\mathcal{B}$ with respect to some norm $|||\cdot|||$ as

$$d_{|||\cdot|||}(\mathcal{B}) := \sup_{\mathbf{B}\in\mathcal{B}} |||\mathbf{B}|||\,.$$

Moreover, we will also need Talagrand's functional $\gamma_2\left(\mathcal{B}, |||\cdot|||\right)$ [63], where for a precise definition, we refer to [62].

Theorem D.1 (Theorem 3.1 in [62]). *Let $\mathcal{B}$ be a set of matrices, and $\boldsymbol{\xi}$ be a random Gaussian vector, i.e., $\boldsymbol{\xi}$ has i.i.d. entries with distribution $\mathcal{N}(0,1)$. Set*

$$\begin{aligned}E &:= \gamma_2(\mathcal{B}, \|\cdot\|)\left(\gamma_2\left(\mathcal{B}, \|\cdot\|\right) + d_{\|\cdot\|_F}\left(\mathcal{B}\right)\right) + d_{\|\cdot\|_F}(\mathcal{B})d_{\|\cdot\|}(\mathcal{B}), \\ V &:= d_{\|\cdot\|}(\mathcal{B})\left(\gamma_2\left(\mathcal{B}, \|\cdot\|\right) + d_{\|\cdot\|_F}(\mathcal{B})\right), \quad U := d_{\|\cdot\|}^2(\mathcal{B}).\end{aligned}$$

Then, for any $t > 0$,

$$\mathbb{P}\left(\sup_{\mathbf{B}\in\mathcal{B}}\left|\left\|\mathbf{B}\boldsymbol{\xi}\right\|_2^2 - \mathbb{E}\left\|\mathbf{B}\boldsymbol{\xi}\right\|_2^2\right| > c_1E + t\right) \leq 2\exp\left(-c_2\min\left\{\frac{t^2}{V^2}, \frac{t}{U}\right\}\right),$$

where $c_1, c_2 > 0$ denote absolute constants.

With this result in place, we can give a proof of Lemma 2.2. This proof strategy has been used in [62, Section A.3].

Proof of Lemma 2.2. Since $\mathcal{A}$ is a linear operator we can write $\mathcal{A}(\mathbf{X}) = \mathbf{V}_{\mathbf{X}}\boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a Gaussian random vector with independent entries of length $m\binom{d+1}{2}$ and

$$\mathbf{V}_{\mathbf{X}} := \frac{1}{\sqrt{m}} \begin{bmatrix} \mathrm{vec}(\mathbf{X})^\top & & & \\ & \mathrm{vec}(\mathbf{X})^\top & & \\ & & \ddots & \\ & & & \mathrm{vec}(\mathbf{X})^\top \end{bmatrix}$$

is an $m \times (m\binom{d+1}{2})$ block-diagonal matrix. Here, $\mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{\binom{d+1}{2}}$ is a vector indexed by $\{(i,j) \in [d] \times [d] : i \leq j\}$ such that

$$\mathrm{vec}(\mathbf{X})(i,j) = \begin{cases} \sqrt{2}\mathbf{X}_{ij} & i \neq j \\ \mathbf{X}_{ii} & i = j. \end{cases}$$

Let

$$D_r := \{\mathbf{X} \in \mathcal{S}^d : \|\mathbf{X}\|_F = 1,\ \mathrm{rank}(\mathbf{X}) \leq r\}.$$

Then it follows from the identity $\mathcal{A}(\mathbf{X}) = \mathbf{V}_{\mathbf{X}}\boldsymbol{\xi}$ that

$$\delta_r := \sup_{\mathbf{X} \in D_r} \left| \|\mathcal{A}(\mathbf{X})\|_2^2 - \|\mathbf{X}\|_F^2 \right| = \sup_{\mathbf{X} \in D_r} \left| \|\mathbf{V}_{\mathbf{x}}\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|\mathbf{V}_{\mathbf{x}}\boldsymbol{\xi}\|_2^2 \right|.$$

Denote $\mathcal{B} := \{\mathbf{V}_{\mathbf{X}} : \mathbf{X} \in D_r\}$. We now estimate the parameters in Theorem D.1. Note that it follows directly from the definition of $\mathrm{vec}(\mathbf{X})$ that $\|\mathrm{vec}(\mathbf{X})\|_2 = \|\mathbf{X}\|_F = 1$ and hence $\|\mathbf{V}_{\mathbf{X}}\|_F = \|\mathbf{X}\|_F$ for all $X \in \mathcal{S}^d$. Thus, we have $d_F(\mathcal{B}) = 1$ since $\|\mathbf{V}_{\mathbf{X}}\|_F = \|\mathbf{X}\|_F$ for all $\mathbf{X} \in D_r$. On the other hand, for $\mathbf{X} \in D_r$,

$$m\mathbf{V}_{\mathbf{X}}\mathbf{V}_{\mathbf{X}}^T = \mathbf{Id}_m,$$

which implies that

$$\|\mathbf{V}_{\mathbf{X}}\| = \frac{1}{\sqrt{m}}\|\mathrm{vec}(\mathbf{X})\|_2 = \frac{1}{\sqrt{m}}\|\mathbf{X}\|_F \tag{109}$$

and $d_{\|\cdot\|}(\mathcal{B}) = \frac{1}{\sqrt{m}}$. From [50, Lemma 3.1], it follows that the covering number for $d \times d$ symmetric matrices with Frobenius norm 1 and rank at most r satisfies

$$\mathcal{N}(D_r, \|\cdot\|_F, \varepsilon) \leq (1 + 6/\varepsilon)^{(2d+1)r}. \tag{110}$$

Using Dudley's integral estimate (see, e.g., [63]), combined with (109) and (110), we obtain that

$$\gamma_2(\mathcal{B}, \|\cdot\|) = \gamma_2(D_r, \|\cdot\|_F) \leq C\frac{1}{\sqrt{m}} \int_0^1 \sqrt{\log(\mathcal{N}(D_r, \|\cdot\|_F, u))}du \leq C'\sqrt{\frac{dr}{m}}.$$

With the notations in Theorem D.1, we have

$$E = C'\sqrt{\frac{dr}{m}}\left(C'\sqrt{\frac{dr}{m}} + 1\right) + \frac{1}{\sqrt{m}}, \quad V = \frac{1}{\sqrt{m}}\left(C'\sqrt{\frac{dr}{m}} + 1\right), \quad U = \frac{1}{m}.$$

Therefore, applying Theorem D.1, we have $\delta_r \leq \delta$ with probability at least $1 - \varepsilon$ when

$$m \geq C\delta^{-2}(rd + \log(2\varepsilon^{-1})).$$

Here, $C > 0$ denotes some universal constant. This completes the proof of Lemma 2.2. □

D.2. Proof of Lemma 2.4

Proof of Lemma 2.4. We will establish first that for all symmetric matrices $\mathbf{Z}_1, \mathbf{Z}_2 \in \mathbb{R}^{d\times d}$ with rank $\operatorname{rank}(\mathbf{Z}_1) = r$ and $\operatorname{rank}(\mathbf{Z}_1) = r'$ it holds that

$$|\langle (\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathbf{Z}_1), \mathbf{Z}_2\rangle| \leq \delta_{r+r'} \|\mathbf{Z}_1\|_F \|\mathbf{Z}_2\|_F. \tag{111}$$

Let us remark that in the case of $\langle \mathbf{Z}_1, \mathbf{Z}_2\rangle = 0$, this inequality has been proven in [50, Lemma 3.3]. The following proof of this slightly more general statement is analogous.

To prove inequality (111) we assume without loss of generality that $\|\mathbf{Z}_1\|_F = \|\mathbf{Z}_2\|_F = 1$. We note first that from the parallelogram identity, it follows that

$$\begin{aligned}\langle \mathcal{A}(\mathbf{Z}_1), \mathcal{A}(\mathbf{Z}_2)\rangle =& \frac{1}{4}\|\mathcal{A}(\mathbf{Z}_1 + \mathbf{Z}_2)\|_2^2 - \frac{1}{4}\|\mathcal{A}(\mathbf{Z}_1 - \mathbf{Z}_2)\|_2^2 \\ \leq& \frac{1+\delta_{r+r'}}{4}\|\mathbf{Z}_1 + \mathbf{Z}_2\|_F^2 - \frac{1-\delta_{r+r'}}{4}\|\mathbf{Z}_1 - \mathbf{Z}_2\|_F^2 \\ =& \frac{\delta_{r+r'}}{2}\left(\|\mathbf{Z}_1\|_F^2 + \|\mathbf{Z}_2\|_F^2\right) + \langle \mathbf{Z}_1, \mathbf{Z}_2\rangle.\end{aligned}$$

By rearranging terms and using the assumption $\|\mathbf{Z}_1\|_F = \|\mathbf{Z}_2\|_F = 1$ we obtain that

$$\langle (\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathbf{Z}_1), \mathbf{Z}_2\rangle = \langle \mathcal{A}(\mathbf{Z}_1), \mathcal{A}(\mathbf{Z}_2)\rangle - \langle \mathbf{Z}_1, \mathbf{Z}_2\rangle \leq \delta_{r+r'}.$$

Since the reverse bound

$$\langle (\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathbf{Z}_1), \mathbf{Z}_2\rangle \geq -\delta_{r+r'}$$

can be shown analogously, inequality (111) follows.

Next, we prove inequality (7). For that, we note that there exists a matrix $\mathbf{M} \in \mathbb{R}^{d\times r'}$ with $\|\mathbf{M}\|_F = 1$ such that

$$\begin{aligned}\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathbf{Z})\mathbf{V}\|_F =& \langle [(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathbf{Z})]\mathbf{V}, \mathbf{M}\rangle = \langle [(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathbf{Z})], \mathbf{V}\mathbf{M}^\top\rangle \\ =& \langle (\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathbf{Z}), \frac{1}{2}\mathbf{V}\mathbf{M}^\top + \frac{1}{2}\mathbf{M}\mathbf{V}^\top\rangle.\end{aligned}$$

holds. Using inequality (111) we obtain that

$$\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathbf{Z})\mathbf{V}\|_F \leq \delta_{r+2r'}\|\mathbf{Z}\|_F \|\frac{1}{2}\mathbf{V}\mathbf{M}^\top + \frac{1}{2}\mathbf{M}\mathbf{V}^\top\|_F \leq \delta_{r+2r'}\|\mathbf{Z}\|_F \|\mathbf{V}\|\|\mathbf{M}\|_F = \delta_{r+2r'}\|\mathbf{Z}\|_F.$$

This proves inequality (7).

Inequality (8) is a direct consequence of (7). Indeed, let $\mathbf{v} \in \mathbb{R}^d$ with $\|\mathbf{v}\|_2 = 1$ be an eigenvector of $(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathbf{Z})$ corresponding to the largest eigenvalue in absolute value. It then follows from inequality (7) that

$$\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathbf{Z})\| = \|[(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathbf{Z})]\mathbf{v}\|_2 \leq \delta_{r+2}\|\mathbf{Z}\|_F.$$

It remains to prove inequality (10). Note that using the fact $\langle \mathbf{w}\mathbf{w}^\top, \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z}) = 0\rangle$, we have

$$\begin{aligned}|\langle \mathcal{A}(\mathbf{w}\mathbf{w}^\top), \mathcal{A}(\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z}))\rangle| &= |\langle (\mathcal{A}^*\mathcal{A})(\mathbf{w}\mathbf{w}^\top), \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\rangle| \\ &= |\langle (\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathbf{w}\mathbf{w}^\top), \mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\rangle| \\ &\overset{(a)}{\leq} \delta_{(r+1)+1}\|\mathbf{w}\mathbf{w}^\top\|_F \|\mathcal{P}_{\mathbf{w}\mathbf{w}^\top,\perp}(\mathbf{Z})\|_F \\ &\leq \delta_{r+2}\|\mathbf{Z}\|_F,\end{aligned}$$

where in inequality (a) we used (111). This completes the proof of Lemma 2.4. □