Accelerating Ill-Conditioned Low-Rank Matrix Estimation via Scaled Gradient Descent

Tian Tong* Carnegie Mellon University

Cong Ma† UC Berkeley

Yuejie Chi* Carnegie Mellon University

May 2020; Revised June 2021

Abstract

Low-rank matrix estimation is a canonical problem that finds numerous applications in signal processing, machine learning and imaging science. A popular approach in practice is to factorize the matrix into two compact low-rank factors, and then optimize these factors directly via simple iterative methods such as gradient descent and alternating minimization. Despite nonconvexity, recent literatures have shown that these simple heuristics in fact achieve linear convergence when initialized properly for a growing number of problems of interest. However, upon closer examination, existing approaches can still be computationally expensive especially for ill-conditioned matrices: the convergence rate of gradient descent depends linearly on the condition number of the low-rank matrix, while the per-iteration cost of alternating minimization is often prohibitive for large matrices.

The goal of this paper is to set forth a competitive algorithmic approach dubbed *Scaled Gradient Descent* (`ScaledGD`) which can be viewed as preconditioned or diagonally-scaled gradient descent, where the preconditioners are adaptive and iteration-varying with a minimal computational overhead. With tailored variants for low-rank matrix sensing, robust principal component analysis and matrix completion, we theoretically show that `ScaledGD` achieves the best of both worlds: it converges linearly at a rate independent of the condition number of the low-rank matrix similar as alternating minimization, while maintaining the low per-iteration cost of gradient descent. Our analysis is also applicable to general loss functions that are restricted strongly convex and smooth over low-rank matrices. To the best of our knowledge, `ScaledGD` is the first algorithm that provably has such properties over a wide range of low-rank matrix estimation tasks. At the core of our analysis is the introduction of a new distance function that takes account of the preconditioners when measuring the distance between the iterates and the ground truth. Finally, numerical examples are provided to demonstrate the effectiveness of `ScaledGD` in accelerating the convergence rate of ill-conditioned low-rank matrix estimation in a wide number of applications.

Keywords: low-rank matrix factorization, scaled gradient descent, ill-conditioned matrix recovery, matrix sensing, robust PCA, matrix completion.

Contents

*Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Emails: {ttong1,yuejiec}@andrew.cmu.edu.

†Department of Electrical Engineering and Computer Science, UC Berkeley, Berkeley, CA 94720, USA; Email: congm@berkeley.edu.

1 Introduction

Low-rank matrix estimation plays a critical role in fields such as machine learning, signal processing, imaging science, and many others. Broadly speaking, one aims to recover a rank-r matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ from a set of observations $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\star)$, where the operator $\mathcal{A}(\cdot)$ models the measurement process. It is natural to minimize the least-squares loss function subject to a rank constraint:

$$\underset{\boldsymbol{X}\in\mathbb{R}^{n_1\times n_2}}{\text{minimize}} \ f(\boldsymbol{X}) \coloneqq \frac{1}{2}\|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{y}\|_2^2 \qquad \text{s.t.} \quad \text{rank}(\boldsymbol{X}) \le r, \tag{1}$$

which is, however, computationally intractable in general due to the rank constraint. Moreover, as the size of the matrix increases, the costs involved in optimizing over the full matrix space (i.e. $\mathbb{R}^{n_1\times n_2}$) are prohibitive in terms of both memory and computation. To cope with these challenges, one popular approach is to parametrize $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{R}^\top$ by two low-rank factors $\boldsymbol{L} \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{R} \in \mathbb{R}^{n_2\times r}$ that are more memory-efficient, and then to optimize over the factors instead:

$$\underset{\boldsymbol{L}\in\mathbb{R}^{n_1\times r},\boldsymbol{R}\in\mathbb{R}^{n_2\times r}}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{L},\boldsymbol{R}) \coloneqq f(\boldsymbol{L}\boldsymbol{R}^\top). \tag{2}$$

Although this leads to a nonconvex optimization problem over the factors, recent breakthroughs have shown that simple algorithms (e.g. gradient descent, alternating minimization), when properly initialized (e.g. via the spectral method), can provably converge to the true low-rank factors under mild statistical assumptions. These benign convergence guarantees hold for a growing number of problems such as low-rank matrix sensing, matrix completion, robust principal component analysis (robust PCA), phase synchronization, and so on.

However, upon closer examination, existing approaches such as gradient descent and alternating minimization are still computationally expensive, especially for ill-conditioned matrices. Take low-rank matrix sensing as an example: although the per-iteration cost is small, the iteration complexity of gradient descent scales linearly with respect to the condition number of the low-rank matrix $\boldsymbol{X}_\star$ [TBS+16]; on the other end, while the iteration complexity of alternating minimization [JNS13] is independent of the condition number, each iteration requires inverting a linear system whose size is proportional to the dimension of the matrix and thus the per-iteration cost is prohibitive for large-scale problems. These together raise an important open question: *can one design an algorithm with a comparable per-iteration cost as gradient descent, but converges much faster at a rate that is independent of the condition number as alternating minimization in a provable manner for a wide variety of low-rank matrix estimation tasks?*

1.1 Preconditioning helps: scaled gradient descent

In this paper, we answer this question affirmatively by studying the following scaled gradient descent (`ScaledGD`) algorithm to optimize (2). Given an initialization $(\boldsymbol{L}_0, \boldsymbol{R}_0)$, `ScaledGD` proceeds as follows

$$\begin{aligned} \boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta \nabla_{\boldsymbol{L}}\mathcal{L}(\boldsymbol{L}_t,\boldsymbol{R}_t)(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1},\\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta \nabla_{\boldsymbol{R}}\mathcal{L}(\boldsymbol{L}_t,\boldsymbol{R}_t)(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}, \end{aligned} \tag{3}$$

where $\eta > 0$ is the step size and $\nabla_{\boldsymbol{L}}\mathcal{L}(\boldsymbol{L}_t,\boldsymbol{R}_t)$ (resp. $\nabla_{\boldsymbol{R}}\mathcal{L}(\boldsymbol{L}_t,\boldsymbol{R}_t)$) is the gradient of the loss function $\mathcal{L}$ with respect to the factor $\boldsymbol{L}_t$ (resp. $\boldsymbol{R}_t$) at the t-th iteration. Comparing to vanilla gradient descent, the search directions of the low-rank factors $\boldsymbol{L}_t, \boldsymbol{R}_t$ in (3) are *scaled* by $(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}$ and $(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}$ respectively. Intuitively, the scaling serves as a preconditioner as in quasi-Newton type algorithms, with the hope of improving the quality of the search direction to allow larger step sizes. Since the computation of the Hessian is extremely expensive, it is necessary to design preconditioners that are both theoretically sound and practically cheap to compute. Such requirements are met by `ScaledGD`, where the preconditioners are computed by inverting two $r \times r$ matrices, whose size is much smaller than the dimension of matrix factors. Therefore, each iteration of `ScaledGD` adds minimal overhead to the gradient computation and has the order-wise same per-iteration cost as gradient descent. Moreover, the preconditioners are adaptive and iteration-varying. Another key property of `ScaledGD` is that it ensures the iterates are covariant with respect to the parameterization of low-rank factors up to invertible transforms.

While `ScaledGD` and its alternating variants have been proposed in [MAS12, MS16, TW16] for a subset of the problems we studied, none of these prior art provides any theoretical validations to the empirical success. In this work, we confirm *theoretically* that `ScaledGD` achieves linear convergence at a rate *independent of* the condition number of the matrix when initialized properly, e.g. using the standard spectral method, for several canonical problems: low-rank matrix sensing, robust PCA, and matrix completion. Table 1 summarizes the performance guarantees of `ScaledGD` in terms of both statistical and computational complexities with comparisons to prior algorithms using the vanilla gradient method.

- *Low-rank matrix sensing.* As long as the measurement operator satisfies the standard restricted isometry property (RIP) with an RIP constant $\delta_{2r} \lesssim 1/(\sqrt{r}\kappa)$, where κ is the condition number of $\boldsymbol{X}_\star$, `ScaledGD` reaches ϵ-accuracy in $O(\log(1/\epsilon))$ iterations when initialized by the spectral method. This strictly improves

	Matrix sensing		Robust PCA		Matrix completion	
Algorithms	sample complexity	iteration complexity	corruption fraction	iteration complexity	sample complexity	iteration complexity
GD	$nr^2\kappa^2$	$\kappa\log\frac{1}{\epsilon}$	$\frac{1}{\mu r^{3/2}\kappa^{3/2}\vee\mu r\kappa^2}$	$\kappa\log\frac{1}{\epsilon}$	$(\mu\vee\log n)\mu nr^2\kappa^2$	$\kappa\log\frac{1}{\epsilon}$
ScaledGD (this paper)	$nr^2\kappa^2$	$\log\frac{1}{\epsilon}$	$\frac{1}{\mu r^{3/2}\kappa}$	$\log\frac{1}{\epsilon}$	$(\mu\kappa^2\vee\log n)\mu nr^2\kappa^2$	$\log\frac{1}{\epsilon}$

Table 1: Comparisons of `ScaledGD` with gradient descent (`GD`) when tailored to various problems (with spectral initialization) [TBS+16, YPCC16, ZL16], where they have comparable per-iteration costs. Here, we say that the output $\boldsymbol{X}$ of an algorithm reaches ϵ-accuracy, if it satisfies $\|\boldsymbol{X}-\boldsymbol{X}_\star\|_{\mathsf{F}}\le\epsilon\sigma_r(\boldsymbol{X}_\star)$. Here, $n\coloneqq n_1\vee n_2=\max\{n_1,n_2\}$, κ and μ are the condition number and incoherence parameter of $\boldsymbol{X}_\star$.

the iteration complexity $O(\kappa\log(1/\epsilon))$ of gradient descent in [TBS+16] under the same sample complexity requirement.

- *Robust PCA.* Under the deterministic corruption model [CSPW11], as long as the fraction α of corruptions per row / column satisfies $\alpha\lesssim 1/(\mu r^{3/2}\kappa)$, where μ is the incoherence parameter of $\boldsymbol{X}_\star$, `ScaledGD` in conjunction with hard thresholding reaches ϵ-accuracy in $O(\log(1/\epsilon))$ iterations when initialized by the spectral method. This strictly improves the iteration complexity of projected gradient descent [YPCC16].
- *Matrix completion.* Under the random Bernoulli observation model, as long as the sample complexity satisfies $n_1n_2p\gtrsim(\mu\kappa^2\vee\log n)\mu nr^2\kappa^2$ with $n=n_1\vee n_2$, `ScaledGD` in conjunction with a properly designed projection operator reaches ϵ-accuracy in $O(\log(1/\epsilon))$ iterations when initialized by the spectral method. This improves the iteration complexity of projected gradient descent [ZL16] at the expense of requiring a larger sample size.

In addition, `ScaledGD` does not require any explicit regularizations that balance the norms of two low-rank factors as required in [TBS+16, YPCC16, ZL16], and removed the additional projection that maintains the incoherence properties in robust PCA [YPCC16], thus unveiling the implicit regularization property of `ScaledGD`. To the best of our knowledge, this is the first factored gradient descent algorithm that achieves a fast convergence rate that is independent of the condition number of the low-rank matrix at near-optimal sample complexities without increasing the per-iteration computational cost. Our analysis is also applicable to general loss functions that are restricted strongly convex and smooth over low-rank matrices.

At the core of our analysis, we introduce a new distance metric (i.e. Lyapunov function) that accounts for the preconditioners, and carefully show the contraction of the `ScaledGD` iterates under the new distance metric. We expect that the `ScaledGD` algorithm can accelerate the convergence for other low-rank matrix estimation problems, as well as facilitate the design and analysis of other quasi-Newton first-order algorithms. As a teaser, Figure 1 illustrates the relative error of completing a 1000×1000 incoherent matrix of rank 10 with varying condition numbers from 20% of its entries, using either `ScaledGD` or vanilla GD with spectral initialization. Even for moderately ill-conditioned matrices, the convergence rate of vanilla GD slows down dramatically, while it is evident that `ScaledGD` converges at a rate independent of the condition number and therefore is much more efficient.

Remark 1 (`ScaledGD` for PSD matrices). When the low-rank matrix of interest is positive semi-definite (PSD), we factorize the matrix $\boldsymbol{X}\in\mathbb{R}^{n\times n}$ as $\boldsymbol{X}=\boldsymbol{L}\boldsymbol{L}^\top$, with $\boldsymbol{L}\in\mathbb{R}^{n\times r}$. The update rule of `ScaledGD` simplifies to

$$\boldsymbol{L}_{t+1}=\boldsymbol{L}_t-\eta\nabla_{\boldsymbol{L}}\mathcal{L}(\boldsymbol{L}_t)(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}. \tag{4}$$

We focus on the asymmetric case since the analysis is more involved with two factors. Our theory applies to the PSD case without loss of generality.

1.2 Related work

Our work contributes to the growing literature of design and analysis of provable nonconvex optimization procedures for high-dimensional signal estimation; see e.g. [JK17, CC18, CLC19] for recent overviews. A

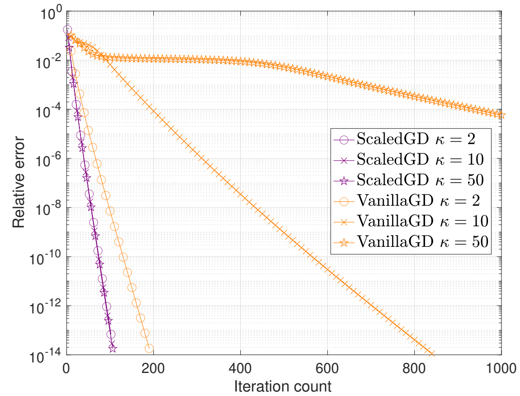


Figure 1: Performance of `ScaledGD` and vanilla GD for completing a 1000×1000 incoherent matrix of rank 10 with different condition numbers $\kappa = 2, 10, 50$, where each entry is observed independently with probability 0.2. Here, both methods are initialized via the spectral method. It can be seen that `ScaledGD` converges much faster than vanilla GD even for moderately large condition numbers.

growing number of problems have been demonstrated to possess benign geometry that is amenable for optimization [MBM18] either globally or locally under appropriate statistical models. On one end, it is shown that there are no spurious local minima in the optimization landscape of matrix sensing and completion [GLM16, BNS16, PKCS17, GJZ17], phase retrieval [SQW18, DDP17], dictionary learning [SQW15], kernel PCA [CL19] and linear neural networks [BH89, Kaw16]. Such landscape analysis facilitates the adoption of generic saddle-point escaping algorithms [NP06, GHJY15, JGN+17] to ensure global convergence. However, the resulting iteration complexity is typically high. On the other end, local refinements with carefully-designed initializations often admit fast convergence, for example in phase retrieval [CLS15, MWCC19], matrix sensing [JNS13, ZL15, WCCL16], matrix completion [SL16, CW15, MWCC19, CLL20, ZL16, CCF+20], blind deconvolution [LLSW19, MWCC19], and robust PCA [NNS+14, YPCC16, CFMY20], to name a few.

Existing approaches for asymmetric low-rank matrix estimation often require additional regularization terms to balance the two factors, either in the form of $\frac{1}{2}\|\boldsymbol{L}^\top \boldsymbol{L} - \boldsymbol{R}^\top \boldsymbol{R}\|_{\mathsf{F}}^2$ [TBS+16, PKCS17] or $\frac{1}{2}\|\boldsymbol{L}\|_{\mathsf{F}}^2 + \frac{1}{2}\|\boldsymbol{R}\|_{\mathsf{F}}^2$ [ZLTW18, CCF+20, CFMY20], which ease the theoretical analysis but are often unnecessary for the practical success, as long as the initialization is balanced. Some recent work studies the unregularized gradient descent for low-rank matrix factorization and sensing including [CCD+21, DHL18, MLC21]. However, the iteration complexity of all these approaches scales at least linearly with respect to the condition number κ of the low-rank matrix, e.g. $O(\kappa \log(1/\epsilon))$, to reach ϵ-accuracy, therefore they converge slowly when the underlying matrix becomes ill-conditioned. In contrast, `ScaledGD` enjoys a local convergence rate of $O(\log(1/\epsilon))$, therefore incurring a much smaller computational footprint when κ is large. Last but not least, alternating minimization [JNS13, HW14] (which alternatively updates $\boldsymbol{L}_t$ and $\boldsymbol{R}_t$) or singular value projection [NNS+14, JMD10] (which operates in the matrix space) also converge at the rate $O(\log(1/\epsilon))$, but the per-iteration cost is much higher than `ScaledGD`. Another notable algorithm is the Riemannian gradient descent algorithm in [WCCL16], which also converges at the rate $O(\log(1/\epsilon))$ under the same sample complexity for low-rank matrix sensing, but requires a higher memory complexity since it operates in the matrix space rather than the factor space.

From an algorithmic perspective, our approach is closely related to the alternating steepest descent (ASD) method in [TW16] for low-rank matrix completion, which performs the proposed updates (3) for the low-rank factors in an alternating manner. Furthermore, the scaled gradient updates were also introduced in [MAS12, MS16] for low-rank matrix completion from the perspective of Riemannian optimization. However, none of [TW16, MAS12, MS16] offered any statistical nor computational guarantees for global convergence. Our analysis of `ScaledGD` can be viewed as providing justifications to these precursors. Moreover, we have systematically extended the framework of `ScaledGD` to work in a large number of low-rank matrix estimation tasks such as robust PCA.

1.3 Paper organization and notation

The rest of this paper is organized as follows. Section 2 describes the proposed `ScaledGD` method and details its application to low-rank matrix sensing, robust PCA and matrix completion with theoretical guarantees in terms of both statistical and computational complexities, highlighting the role of a new distance metric. The convergence guarantee of `ScaledGD` under the general loss function is also presented. In Section 3, we outline the proof for our main results. Section 4 illustrates the excellent empirical performance of `ScaledGD` in a variety of low-rank matrix estimation problems. Finally, we conclude in Section 5.

Before continuing, we introduce several notation used throughout the paper. First of all, we use boldfaced symbols for vectors and matrices. For a vector $\boldsymbol{v}$, we use $\|\boldsymbol{v}\|_0$ to denote its ℓ_0 counting norm, and $\|\boldsymbol{v}\|_2$ to denote the ℓ_2 norm. For any matrix $\boldsymbol{A}$, we use $\sigma_i(\boldsymbol{A})$ to denote its i-th largest singular value, and let $\boldsymbol{A}_{i,\cdot}$ and $\boldsymbol{A}_{\cdot,j}$ denote its i-th row and j-th column, respectively. In addition, $\|\boldsymbol{A}\|_{\mathsf{op}}$, $\|\boldsymbol{A}\|_{\mathsf{F}}$, $\|\boldsymbol{A}\|_{1,\infty}$, $\|\boldsymbol{A}\|_{2,\infty}$, and $\|\boldsymbol{A}\|_{\infty}$ stand for the spectral norm (i.e. the largest singular value), the Frobenius norm, the $\ell_{1,\infty}$ norm (i.e. the largest ℓ_1 norm of the rows), the $\ell_{2,\infty}$ norm (i.e. the largest ℓ_2 norm of the rows), and the entrywise ℓ_∞ norm (the largest magnitude of all entries) of a matrix $\boldsymbol{A}$. We denote

$$\mathcal{P}_r(\boldsymbol{A}) = \min_{\widetilde{\boldsymbol{A}}:\mathrm{rank}(\widetilde{\boldsymbol{A}})\le r} \|\boldsymbol{A}-\widetilde{\boldsymbol{A}}\|_{\mathsf{F}}^2 \tag{5}$$

as the rank-r approximation of $\boldsymbol{A}$, which is given by the top-r SVD of $\boldsymbol{A}$ by the Eckart-Young-Mirsky theorem. We also use $\mathrm{vec}(\boldsymbol{A})$ to denote the vectorization of a matrix $\boldsymbol{A}$. For matrices $\boldsymbol{A},\boldsymbol{B}$ of the same size, we use $\langle \boldsymbol{A},\boldsymbol{B}\rangle = \sum_{i,j}\boldsymbol{A}_{i,j}\boldsymbol{B}_{i,j} = \mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{B})$ to denote their inner product. The set of invertible matrices in $\mathbb{R}^{r\times r}$ is denoted by $\mathrm{GL}(r)$. Let $a\vee b=\max\{a,b\}$ and $a\wedge b=\min\{a,b\}$. Throughout, $f(n)\lesssim g(n)$ or $f(n)=O(g(n))$ means $|f(n)|/|g(n)|\le C$ for some constant $C>0$ when n is sufficiently large; $f(n)\gtrsim g(n)$ means $|f(n)|/|g(n)|\ge C$ for some constant $C>0$ when n is sufficiently large. Last but not least, we use the terminology "with overwhelming probability" to denote the event happens with probability at least $1-c_1 n^{-c_2}$, where $c_1,c_2>0$ are some universal constants, whose values may vary from line to line.

2 Scaled Gradient Descent for Low-Rank Matrix Estimation

This section is devoted to introducing `ScaledGD` and establishing its statistical and computational guarantees for various low-rank matrix estimation problems. Before we instantiate tailored versions of `ScaledGD` on concrete low-rank matrix estimation problems, we first pause to provide more insights of the update rule of `ScaledGD`, by connecting it to the quasi-Newton method. Note that the update rule (3) for `ScaledGD` can be equivalently written in a vectorization form as

$$\begin{aligned}\mathrm{vec}(\boldsymbol{F}_{t+1}) &= \mathrm{vec}(\boldsymbol{F}_t) - \eta\begin{bmatrix}(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}\otimes\boldsymbol{I}_{n_1} & \boldsymbol{0}\\ \boldsymbol{0} & (\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}\otimes\boldsymbol{I}_{n_2}\end{bmatrix}\mathrm{vec}(\nabla_{\boldsymbol{F}}\mathcal{L}(\boldsymbol{F}_t))\\ &= \mathrm{vec}(\boldsymbol{F}_t) - \eta\boldsymbol{H}_t^{-1}\,\mathrm{vec}(\nabla_{\boldsymbol{F}}\mathcal{L}(\boldsymbol{F}_t)),\end{aligned} \tag{6}$$

where we denote $\boldsymbol{F}_t=[\boldsymbol{L}_t^\top,\boldsymbol{R}_t^\top]^\top\in\mathbb{R}^{(n_1+n_2)\times r}$, and by $\otimes$ the Kronecker product. Here, the block diagonal matrix $\boldsymbol{H}_t$ is set to be

$$\boldsymbol{H}_t := \begin{bmatrix}(\boldsymbol{R}_t^\top\boldsymbol{R}_t)\otimes\boldsymbol{I}_{n_1} & \boldsymbol{0}\\ \boldsymbol{0} & (\boldsymbol{L}_t^\top\boldsymbol{L}_t)\otimes\boldsymbol{I}_{n_2}\end{bmatrix}.$$

The form (6) makes it apparent that `ScaledGD` can be interpreted as a quasi-Newton algorithm, where the inverse of $\boldsymbol{H}_t$ can be cheaply computed through inverting two rank-r matrices.

2.1 Assumptions and error metric

Denote by $\boldsymbol{U}_\star\boldsymbol{\Sigma}_\star\boldsymbol{V}_\star^\top$ the compact singular value decomposition (SVD) of the rank-r matrix $\boldsymbol{X}_\star\in\mathbb{R}^{n_1\times n_2}$. Here $\boldsymbol{U}_\star\in\mathbb{R}^{n_1\times r}$ and $\boldsymbol{V}_\star\in\mathbb{R}^{n_2\times r}$ are composed of r left and right singular vectors, respectively, and $\boldsymbol{\Sigma}_\star\in\mathbb{R}^{r\times r}$ is a diagonal matrix consisting of r singular values of $\boldsymbol{X}_\star$ organized in a non-increasing order, i.e. $\sigma_1(\boldsymbol{X}_\star)\ge\cdots\ge\sigma_r(\boldsymbol{X}_\star)>0$. Define

$$\kappa := \sigma_1(\boldsymbol{X}_\star)/\sigma_r(\boldsymbol{X}_\star) \tag{7}$$

as the condition number of $\boldsymbol{X}_\star$. Define the ground truth low-rank factors as

$$\boldsymbol{L}_\star \coloneqq \boldsymbol{U}_\star \boldsymbol{\Sigma}_\star^{1/2}, \qquad \text{and} \qquad \boldsymbol{R}_\star \coloneqq \boldsymbol{V}_\star \boldsymbol{\Sigma}_\star^{1/2}, \tag{8}$$

so that $\boldsymbol{X}_\star = \boldsymbol{L}_\star \boldsymbol{R}_\star^\top$. Correspondingly, denote the stacked factor matrix as

$$\boldsymbol{F}_\star \coloneqq \begin{bmatrix} \boldsymbol{L}_\star \\ \boldsymbol{R}_\star \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}. \tag{9}$$

Next, we are in need of a right metric to measure the performance of the `ScaledGD` iterates $\boldsymbol{F}_t \coloneqq [\boldsymbol{L}_t^\top, \boldsymbol{R}_t^\top]^\top$. Obviously, the factored representation is not unique in that for any invertible matrix $\boldsymbol{Q} \in \mathrm{GL}(r)$, one has $\boldsymbol{L}\boldsymbol{R}^\top = (\boldsymbol{L}\boldsymbol{Q})(\boldsymbol{R}\boldsymbol{Q}^{-\top})^\top$. Therefore, the reconstruction error metric needs to take into account this identifiability issue. More importantly, we need a diagonal scaling in the distance error metric to properly account for the effect of preconditioning. To provide intuition, note that the update rule (3) can be viewed as finding the best local quadratic approximation of $\mathcal{L}(\cdot)$ in the following sense:

$$\begin{aligned}(\boldsymbol{L}_{t+1}, \boldsymbol{R}_{t+1}) = \operatorname*{argmin}_{\boldsymbol{L},\boldsymbol{R}} \; & \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t) + \langle \nabla_{\boldsymbol{L}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t), \boldsymbol{L} - \boldsymbol{L}_t \rangle + \langle \nabla_{\boldsymbol{R}} \mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t), \boldsymbol{R} - \boldsymbol{R}_t \rangle \\ & + \frac{1}{2\eta}\left(\left\| (\boldsymbol{L} - \boldsymbol{L}_t)(\boldsymbol{R}_t^\top \boldsymbol{R}_t)^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R} - \boldsymbol{R}_t)(\boldsymbol{L}_t^\top \boldsymbol{L}_t)^{1/2} \right\|_{\mathsf{F}}^2 \right),\end{aligned}$$

where it is different from the common interpretation of gradient descent in the way the quadratic approximation is taken by a scaled norm. When $\boldsymbol{L}_t \approx \boldsymbol{L}_\star$ and $\boldsymbol{R}_t \approx \boldsymbol{R}_\star$ are approaching the ground truth, the additional scaling factors can be approximated by $\boldsymbol{L}_t^\top \boldsymbol{L}_t \approx \boldsymbol{\Sigma}_\star$ and $\boldsymbol{R}_t^\top \boldsymbol{R}_t \approx \boldsymbol{\Sigma}_\star$, leading to the following error metric

$$\mathrm{dist}^2(\boldsymbol{F}, \boldsymbol{F}_\star) \coloneqq \inf_{\boldsymbol{Q}\in\mathrm{GL}(r)} \left\| (\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2. \tag{10}$$

Correspondingly, we define the optimal alignment matrix $\boldsymbol{Q}$ between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$ as

$$\boldsymbol{Q} \coloneqq \operatorname*{argmin}_{\boldsymbol{Q}\in\mathrm{GL}(r)} \left\| (\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2, \tag{11}$$

whenever the minimum is achieved.[1] It turns out that for the `ScaledGD` iterates $\{\boldsymbol{F}_t\}$, the optimal alignment matrices $\{\boldsymbol{Q}_t\}$ always exist (at least when properly initialized) and hence are well-defined. The design and analysis of this new distance metric are of crucial importance in obtaining the improved rate of `ScaledGD`; see Appendix A.1 for a collection of its properties. In comparison, the previously studied distance metrics (proposed mainly for GD) either do not include the diagonal scaling [MLC21, TBS+16], or only consider the ambiguity class up to orthonormal transforms [TBS+16], which fail to unveil the benefit of `ScaledGD`.

2.2 Matrix sensing

Assume that we have collected a set of linear measurements about a rank-r matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1\times n_2}$, given as

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}_\star) \in \mathbb{R}^m, \tag{12}$$

where $\mathcal{A}(\boldsymbol{X}) = \{\langle \boldsymbol{A}_k, \boldsymbol{X}\rangle\}_{k=1}^m : \mathbb{R}^{n_1\times n_2} \mapsto \mathbb{R}^m$ is the linear map modeling the measurement process. The goal of low-rank matrix sensing is to recover $\boldsymbol{X}_\star$ from $\boldsymbol{y}$, especially when the number of measurements $m \ll n_1 n_2$, by exploiting the low-rank property. This problem has wide applications in medical imaging, signal processing, and data compression [CP11].

[1] If there are multiple minimizers, we can arbitrarily take one to be $\boldsymbol{Q}$.

Algorithm. Writing $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ into a factored form $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{R}^\top$, we consider the following optimization problem:

$$\underset{\boldsymbol{F} \in \mathbb{R}^{(n_1+n_2)\times r}}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{F}) = \frac{1}{2}\left\|\mathcal{A}(\boldsymbol{L}\boldsymbol{R}^\top) - \boldsymbol{y}\right\|_2^2. \tag{13}$$

Here as before, $\boldsymbol{F}$ denotes the stacked factor matrix $[\boldsymbol{L}^\top, \boldsymbol{R}^\top]^\top$. We suggest running `ScaledGD` (3) with the spectral initialization to solve (13), which performs the top-r SVD on $\mathcal{A}^*(\boldsymbol{y})$, where $\mathcal{A}^*(\cdot)$ is the adjoint operator of $\mathcal{A}(\cdot)$. The full algorithm is stated in Algorithm 1. The low-rank matrix can be estimated as $\boldsymbol{X}_T = \boldsymbol{L}_T\boldsymbol{R}_T^\top$ after running T iterations of `ScaledGD`.

Algorithm 1 `ScaledGD` for low-rank matrix sensing with spectral initialization

---

Spectral initialization: Let $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$ be the top-r SVD of $\mathcal{A}^*(\boldsymbol{y})$, and set

$$\boldsymbol{L}_0 = \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{1/2}, \quad \text{and} \quad \boldsymbol{R}_0 = \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{1/2}. \tag{14}$$

Scaled gradient updates: **for** $t = 0, 1, 2, \dots, T-1$ **do**

$$\begin{aligned} \boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta\mathcal{A}^*(\mathcal{A}(\boldsymbol{L}_t\boldsymbol{R}_t^\top) - \boldsymbol{y})\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}, \\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta\mathcal{A}^*(\mathcal{A}(\boldsymbol{L}_t\boldsymbol{R}_t^\top) - \boldsymbol{y})^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}. \end{aligned} \tag{15}$$

Theoretical guarantees. To understand the performance of `ScaledGD` for low-rank matrix sensing, we adopt a standard assumption on the sensing operator $\mathcal{A}(\cdot)$, namely the Restricted Isometry Property (RIP).

Definition 1 (RIP [RFP10])**.** The linear map $\mathcal{A}(\cdot)$ is said to obey the rank-r RIP with a constant $\delta_r \in [0, 1)$, if for all matrices $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ of rank at most r, one has

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

It is well-known that many measurement ensembles satisfy the RIP property [RFP10,CP11]. For example, if the entries of $\boldsymbol{A}_i$'s are composed of i.i.d. Gaussian entries $\mathcal{N}(0, 1/m)$, then the RIP is satisfied for a constant δ_r as long as m is on the order of $(n_1 + n_2)r/\delta_r^2$. With the RIP condition in place, the following theorem demonstrates that `ScaledGD` converges linearly — in terms of the new distance metric (cf. (10)) — at a constant rate as long as the sensing operator $\mathcal{A}(\cdot)$ has a sufficiently small RIP constant.

Theorem 1. *Suppose that $\mathcal{A}(\cdot)$ obeys the $2r$-RIP with $\delta_{2r} \le 0.02/(\sqrt{r}\kappa)$. If the step size obeys $0 < \eta \le 2/3$, then for all $t \ge 0$, the iterates of the `ScaledGD` method in Algorithm 1 satisfy*

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1 - 0.6\eta)^t 0.1\sigma_r(\boldsymbol{X}_\star), \quad \textit{and} \quad \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le (1 - 0.6\eta)^t 0.15\sigma_r(\boldsymbol{X}_\star).$$

Theorem 1 establishes that the distance $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ contracts linearly at a constant rate, as long as the sample size satisfies $m = O(nr^2\kappa^2)$ with Gaussian random measurements [RFP10], where we recall that $n = n_1 \vee n_2$. To reach ϵ-accuracy, i.e. $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$, `ScaledGD` takes at most $T = O(\log(1/\epsilon))$ iterations, which is *independent* of the condition number κ of $\boldsymbol{X}_\star$. In comparison, alternating minimization with spectral initialization (`AltMinSense`) converges in $O(\log(1/\epsilon))$ iterations as long as $m = O(nr^3\kappa^4)$ [JNS13], where the per-iteration cost is much higher.[2] On the other end, gradient descent with spectral initialization in [TBS+16] converges in $O(\kappa\log(1/\epsilon))$ iterations as long as $m = O(nr^2\kappa^2)$. Therefore, `ScaledGD` converges at a much faster rate than GD at the same sample complexity while requiring a significantly lower per-iteration cost than `AltMinSense`.

Remark 2. [TBS+16] suggested that one can employ a more expensive initialization scheme, e.g. performing multiple projected gradient descent steps over the low-rank matrix, to reduce the sample complexity. By seeding `ScaledGD` with the output of updates of the form $\boldsymbol{X}_{\tau+1} = \mathcal{P}_r\left(\boldsymbol{X}_\tau - \mathcal{A}^*(\mathcal{A}(\boldsymbol{X}_\tau) - \boldsymbol{y})\right)$ after $T_0 \gtrsim \log(\sqrt{r}\kappa)$ iterations, where $\mathcal{P}_r(\cdot)$ is defined in (5), `ScaledGD` succeeds with the sample size $O(nr)$ which is information theoretically optimal.

[2] The exact per-iteration complexity of `AltMinSense` depends on how the least-squares subproblems are solved with m equations and nr unknowns; see [LHLZ20, Table 1] for detailed comparisons.

2.3 Robust PCA

Assume that we have observed the data matrix

$$\boldsymbol{Y} = \boldsymbol{X}_\star + \boldsymbol{S}_\star,$$

which is a superposition of a rank-r matrix $\boldsymbol{X}_\star$, modeling the clean data, and a sparse matrix $\boldsymbol{S}_\star$, modeling the corruption or outliers. The goal of robust PCA [CLMW11, CSPW11] is to separate the two matrices $\boldsymbol{X}_\star$ and $\boldsymbol{S}_\star$ from their mixture $\boldsymbol{Y}$. This problem finds numerous applications in video surveillance, image processing, and so on.

Following [CSPW11, NNS+14, YPCC16], we consider a deterministic sparsity model for $\boldsymbol{S}_\star$, in which $\boldsymbol{S}_\star$ contains at most α-fraction of nonzero entries per row and column for some $\alpha \in [0,1)$, i.e. $\boldsymbol{S}_\star \in \mathcal{S}_\alpha$, where we denote

$$\mathcal{S}_\alpha \coloneqq \{\boldsymbol{S} \in \mathbb{R}^{n_1 \times n_2} : \|\boldsymbol{S}_{i,\cdot}\|_0 \le \alpha n_2 \text{ for all } i, \text{ and } \|\boldsymbol{S}_{\cdot,j}\|_0 \le \alpha n_1 \text{ for all } j\}. \tag{16}$$

Algorithm. Writing $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ into the factored form $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{R}^\top$, we consider the following optimization problem:

$$\underset{\boldsymbol{F}\in\mathbb{R}^{(n_1+n_2)\times r}, \boldsymbol{S}\in\mathcal{S}_\alpha}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{F}, \boldsymbol{S}) = \frac{1}{2}\left\|\boldsymbol{L}\boldsymbol{R}^\top + \boldsymbol{S} - \boldsymbol{Y}\right\|_{\mathsf{F}}^2. \tag{17}$$

It is thus natural to alternatively update $\boldsymbol{F} = [\boldsymbol{L}^\top, \boldsymbol{R}^\top]^\top$ and $\boldsymbol{S}$, where $\boldsymbol{F}$ is updated via the proposed `ScaledGD` algorithm, and $\boldsymbol{S}$ is updated by hard thresholding, which trims the small entries of the residual matrix $\boldsymbol{Y} - \boldsymbol{L}\boldsymbol{R}^\top$. More specifically, for some truncation level $0 \le \bar{\alpha} \le 1$, we define the sparsification operator that only keeps $\bar{\alpha}$ fraction of largest entries in each row and column:

$$(\mathcal{T}_{\bar{\alpha}}[\boldsymbol{A}])_{i,j} = \begin{cases} \boldsymbol{A}_{i,j}, & \text{if } |\boldsymbol{A}|_{i,j} \ge |\boldsymbol{A}|_{i,(\bar{\alpha}n_2)}, \text{ and } |\boldsymbol{A}|_{i,j} \ge |\boldsymbol{A}|_{(\bar{\alpha}n_1),j} \\ 0, & \text{otherwise} \end{cases}, \tag{18}$$

where $|\boldsymbol{A}|_{i,(k)}$ (resp. $|\boldsymbol{A}|_{(k),j}$) denote the k-th largest element in magnitude in the i-th row (resp. j-th column).

The `ScaledGD` algorithm with the spectral initialization for solving robust PCA is formally stated in Algorithm 2. Note that, comparing with [YPCC16], we do not require a balancing term $\|\boldsymbol{L}^\top\boldsymbol{L} - \boldsymbol{R}^\top\boldsymbol{R}\|_{\mathsf{F}}^2$ in the loss function (17), nor the projection of the low-rank factors onto the $\ell_{2,\infty}$ ball in each iteration.

Algorithm 2 `ScaledGD` for robust PCA with spectral initialization

Spectral initialization: Let $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$ be the top-r SVD of $\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}]$, and set

$$\boldsymbol{L}_0 = \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{1/2}, \quad \text{and} \quad \boldsymbol{R}_0 = \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{1/2}. \tag{19}$$

Scaled gradient updates: **for** $t = 0, 1, 2, \dots, T-1$ **do**

$$\begin{aligned} \boldsymbol{S}_t &= \mathcal{T}_{2\alpha}[\boldsymbol{Y} - \boldsymbol{L}_t\boldsymbol{R}_t^\top], \\ \boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta(\boldsymbol{L}_t\boldsymbol{R}_t^\top + \boldsymbol{S}_t - \boldsymbol{Y})\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}, \\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta(\boldsymbol{L}_t\boldsymbol{R}_t^\top + \boldsymbol{S}_t - \boldsymbol{Y})^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}. \end{aligned} \tag{20}$$

Theoretical guarantee. Before stating our main result for robust PCA, we introduce the incoherence condition which is known to be crucial for reliable estimation of the low-rank matrix $\boldsymbol{X}_\star$ in robust PCA [Che15].

Definition 2 (Incoherence)**.** A rank-r matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n_1 \times n_2}$ with compact SVD as $\boldsymbol{X}_\star = \boldsymbol{U}_\star\boldsymbol{\Sigma}_\star\boldsymbol{V}_\star^\top$ is said to be μ-incoherent if

$$\|\boldsymbol{U}_\star\|_{2,\infty} \le \sqrt{\frac{\mu}{n_1}}\|\boldsymbol{U}_\star\|_{\mathsf{F}} = \sqrt{\frac{\mu r}{n_1}}, \quad \text{and} \quad \|\boldsymbol{V}_\star\|_{2,\infty} \le \sqrt{\frac{\mu}{n_2}}\|\boldsymbol{V}_\star\|_{\mathsf{F}} = \sqrt{\frac{\mu r}{n_2}}.$$

The following theorem establishes that ScaledGD converges linearly at a constant rate as long as the fraction α of corruptions is sufficiently small.

Theorem 2. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent and that the corruption fraction α obeys $\alpha \le c/(\mu r^{3/2}\kappa)$ for some sufficiently small constant $c > 0$. If the step size obeys $0.1 \le \eta \le 2/3$, then for all $t \ge 0$, the iterates of ScaledGD in Algorithm 2 satisfy*

$$\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1-0.6\eta)^t 0.02\sigma_r(\boldsymbol{X}_\star), \quad \text{and} \quad \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le (1-0.6\eta)^t 0.03\sigma_r(\boldsymbol{X}_\star).$$

Theorem 2 establishes that the distance $\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ contracts linearly at a constant rate, as long as the fraction of corruptions satisfies $\alpha \lesssim 1/(\mu r^{3/2}\kappa)$. To reach ϵ-accuracy, i.e. $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$, ScaledGD takes at most $T = O(\log(1/\epsilon))$ iterations, which is *independent* of κ. In comparison, the AltProj algorithm[3] with spectral initialization converges in $O(\log(1/\epsilon))$ iterations as long as $\alpha \lesssim 1/(\mu r)$ [NNS+14], where the per-iteration cost is much higher both in terms of computation and memory as it requires the computation of the low-rank SVD of the full matrix. On the other hand, projected gradient descent with spectral initialization in [YPCC16] converges in $O(\kappa\log(1/\epsilon))$ iterations as long as $\alpha \lesssim 1/(\mu r^{3/2}\kappa^{3/2} \vee \mu r\kappa^2)$. Therefore, ScaledGD converges at a much faster rate than GD while requesting a significantly lower per-iteration cost than AltProj. In addition, our theory suggests that ScaledGD maintains the incoherence and balancedness of the low-rank factors without imposing explicit regularizations, which is not captured in previous analysis [YPCC16].

2.4 Matrix completion

Assume that we have observed a subset Ω of entries of $\boldsymbol{X}_\star$ given as $\mathcal{P}_\Omega(\boldsymbol{X}_\star)$, where $\mathcal{P}_\Omega : \mathbb{R}^{n_1\times n_2} \mapsto \mathbb{R}^{n_1\times n_2}$ is a projection such that

$$(\mathcal{P}_\Omega(\boldsymbol{X}))_{i,j} = \begin{cases} \boldsymbol{X}_{i,j}, & \text{if } (i,j)\in\Omega \\ 0, & \text{otherwise} \end{cases}. \tag{21}$$

Here Ω is generated according to the Bernoulli model in the sense that each $(i,j) \in \Omega$ independent with probability p. The goal of matrix completion is to recover the matrix $\boldsymbol{X}_\star$ from its partial observation $\mathcal{P}_\Omega(\boldsymbol{X}_\star)$. This problem has many applications in recommendation systems, signal processing, sensor network localization, and so on [CR09].

Algorithm. Again, writing $\boldsymbol{X} \in \mathbb{R}^{n_1\times n_2}$ into the factored form $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{R}^\top$, we consider the following optimization problem:

$$\underset{\boldsymbol{F}\in\mathbb{R}^{(n_1+n_2)\times r}}{\text{minimize}} \ \mathcal{L}(\boldsymbol{F}) \coloneqq \frac{1}{2p}\left\|\mathcal{P}_\Omega(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\right\|_{\mathsf{F}}^2. \tag{22}$$

Similarly to robust PCA, the underlying low-rank matrix $\boldsymbol{X}_\star$ needs to be incoherent (cf. Definition 2) to avoid ill-posedness. One typical strategy to ensure the incoherence condition is to perform projection after the gradient update, by projecting the iterates to maintain small $\ell_{2,\infty}$ norms of the factor matrices. However, the standard projection operator [CW15] is not covariant with respect to invertible transforms, and consequently, needs to be modified when using scaled gradient updates. To that end, we introduce the following new projection operator: for every $\widetilde{\boldsymbol{F}} \in \mathbb{R}^{(n_1+n_2)\times r} = [\widetilde{\boldsymbol{L}}^\top, \widetilde{\boldsymbol{R}}^\top]^\top$,

$$\begin{aligned} \mathcal{P}_B(\widetilde{\boldsymbol{F}}) = \underset{\boldsymbol{F}\in\mathbb{R}^{(n_1+n_2)\times r}}{\operatorname{argmin}} \ & \left\|(\boldsymbol{L}-\widetilde{\boldsymbol{L}})(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}-\widetilde{\boldsymbol{R}})(\widetilde{\boldsymbol{L}}^\top\widetilde{\boldsymbol{L}})^{1/2}\right\|_{\mathsf{F}}^2, \\ \text{s.t.} \quad & \sqrt{n_1}\left\|\boldsymbol{L}(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}\right\|_{2,\infty} \vee \sqrt{n_2}\left\|\boldsymbol{R}(\widetilde{\boldsymbol{L}}^\top\widetilde{\boldsymbol{L}})^{1/2}\right\|_{2,\infty} \le B \end{aligned} \tag{23}$$

which finds a factored matrix that is closest to $\widetilde{\boldsymbol{F}}$ and stays incoherent in a weighted sense. Luckily, the solution to the above scaled projection admits a simple closed-form solution, as stated below.

[3] AltProj employs a multi-stage strategy to remove the dependence on κ in α, which we do not consider here. The same strategy might also improve the dependence on κ for ScaledGD, which we leave for future work.

Proposition 1. *The solution to* (23) *is given by*

$$\mathcal{P}_B(\widetilde{\boldsymbol{F}}) := \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{R} \end{bmatrix}, \quad \text{where} \quad \begin{aligned} \boldsymbol{L}_{i,\cdot} &:= \left(1 \wedge \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}\right)\widetilde{\boldsymbol{L}}_{i,\cdot}, \quad 1 \le i \le n_1, \\ \boldsymbol{R}_{j,\cdot} &:= \left(1 \wedge \frac{B}{\sqrt{n_2}\|\widetilde{\boldsymbol{R}}_{j,\cdot}\widetilde{\boldsymbol{L}}^\top\|_2}\right)\widetilde{\boldsymbol{R}}_{j,\cdot}, \quad 1 \le j \le n_2. \end{aligned} \tag{24}$$

Proof. See Appendix E.1.1. □

With the new projection operator in place, we propose the scaled projected gradient descent (`ScaledPGD`) method with the spectral initialization for solving matrix completion, formally stated in Algorithm 3.

Algorithm 3 `ScaledPGD` for matrix completion with spectral initialization

---

Spectral initialization: Let $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$ be the top-r SVD of $\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{X}_\star)$, and set

$$\begin{bmatrix} \boldsymbol{L}_0 \\ \boldsymbol{R}_0 \end{bmatrix} = \mathcal{P}_B\left(\begin{bmatrix} \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{1/2} \\ \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{1/2} \end{bmatrix}\right). \tag{25}$$

Scaled projected gradient updates: **for** $t = 0, 1, 2, \dots, T-1$ **do**

$$\begin{bmatrix} \boldsymbol{L}_{t+1} \\ \boldsymbol{R}_{t+1} \end{bmatrix} = \mathcal{P}_B\left(\begin{bmatrix} \boldsymbol{L}_t - \frac{\eta}{p}\mathcal{P}_\Omega(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1} \\ \boldsymbol{R}_t - \frac{\eta}{p}\mathcal{P}_\Omega(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1} \end{bmatrix}\right). \tag{26}$$

Theoretical guarantee. Consider a random observation model, where each index (i,j) belongs to the index set Ω independently with probability $0 < p \le 1$. The following theorem establishes that `ScaledPGD` converges linearly at a constant rate as long as the number of observations is sufficiently large.

Theorem 3. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent, and that p satisfies $p \ge C(\mu\kappa^2 \vee \log(n_1 \vee n_2))\mu r^2\kappa^2/(n_1 \wedge n_2)$ for some sufficiently large constant C. Set the projection radius as $B = C_B\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star)$ for some constant $C_B \ge 1.02$. If the step size obeys $0 < \eta \le 2/3$, then with probability at least $1 - c_1(n_1 \vee n_2)^{-c_2}$, for all $t \ge 0$, the iterates of `ScaledPGD` in* (26) *satisfy*

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1 - 0.6\eta)^t 0.02\sigma_r(\boldsymbol{X}_\star), \quad \text{and} \quad \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le (1 - 0.6\eta)^t 0.03\sigma_r(\boldsymbol{X}_\star).$$

Here $c_1, c_2 > 0$ are two universal constants.

Theorem 3 establishes that the distance $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ contracts linearly at a constant rate, as long as the probability of observation satisfies $p \gtrsim (\mu\kappa^2 \vee \log(n_1 \vee n_2))\mu r^2\kappa^2/(n_1 \wedge n_2)$. To reach ϵ-accuracy, i.e. $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$, `ScaledPGD` takes at most $T = O(\log(1/\epsilon))$ iterations, which is *independent* of κ. In comparison, projected gradient descent [ZL16] with spectral initialization converges in $O(\kappa\log(1/\epsilon))$ iterations as long as $p \gtrsim (\mu \vee \log(n_1 \vee n_2))\mu r^2\kappa^2/(n_1 \wedge n_2)$. Therefore, `ScaledPGD` achieves much faster convergence than its unscaled counterpart, at an expense of higher sample complexity. We believe this higher sample complexity is an artifact of our proof techniques, as numerically we do not observe a degradation in terms of sample complexity.

2.5 Optimizing general loss functions

Last but not least, we generalize our analysis of `ScaledGD` to minimize a general loss function in the form of (2), where the update rule of `ScaledGD` is given by

$$\begin{aligned} \boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta\nabla f(\boldsymbol{L}_t\boldsymbol{R}_t^\top)\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}, \\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta\nabla f(\boldsymbol{L}_t\boldsymbol{R}_t^\top)^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}. \end{aligned} \tag{27}$$

Two important properties of the loss function $f : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}$ play a key role in the analysis.

Definition 3 (Restricted smoothness)**.** A differentiable function $f : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}$ is said to be rank-r restricted L-smooth for some $L > 0$ if

$$f(\boldsymbol{X}_2) \le f(\boldsymbol{X}_1) + \langle \nabla f(\boldsymbol{X}_1), \boldsymbol{X}_2 - \boldsymbol{X}_1 \rangle + \frac{L}{2} \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_{\mathsf{F}}^2,$$

for any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n_1 \times n_2}$ with rank at most r.

Definition 4 (Restricted strong convexity)**.** A differentiable function $f : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}$ is said to be rank-r restricted μ-strongly convex for some $\mu \ge 0$ if

$$f(\boldsymbol{X}_2) \ge f(\boldsymbol{X}_1) + \langle \nabla f(\boldsymbol{X}_1), \boldsymbol{X}_2 - \boldsymbol{X}_1 \rangle + \frac{\mu}{2} \|\boldsymbol{X}_2 - \boldsymbol{X}_1\|_{\mathsf{F}}^2,$$

for any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n_1 \times n_2}$ with rank at most r. When $\mu = 0$, we simply say $f(\cdot)$ is rank-r restricted convex.

Further, when $\mu > 0$, define the condition number of the loss function $f(\cdot)$ over rank-r matrices as

$$\kappa_f \coloneqq L/\mu. \tag{28}$$

Encouragingly, many problems can be viewed as a special case of optimizing this general loss (27), including but not limited to:

- *low-rank matrix factorization*, where the loss function $f(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}}^2$ in (29) satisfies $\kappa_f = 1$;
- *low-rank matrix sensing*, where the loss function $f(\boldsymbol{X}) = \frac{1}{2}\|\mathcal{A}(\boldsymbol{X} - \boldsymbol{X}_\star)\|_2^2$ in (13) satisfies $\kappa_f \approx 1$ when $\mathcal{A}(\cdot)$ obeys the rank-r RIP with a sufficiently small RIP constant;
- *quadratic sampling*, where the loss function $f(\boldsymbol{X}) = \frac{1}{2}\sum_{i=1}^m |\langle \boldsymbol{a}_i \boldsymbol{a}_i^\top, \boldsymbol{X} - \boldsymbol{X}_\star \rangle|^2$ satisfies restricted strong convexity and smoothness when $\boldsymbol{a}_i$'s are i.i.d. Gaussian vectors for sufficiently large m [SWW17, LMCC21];
- *exponential-family PCA*, where the loss function $f(\boldsymbol{X}) = -\sum_{i,j} \log p(\boldsymbol{Y}_{i,j}|\boldsymbol{X}_{i,j})$, where $p(\boldsymbol{Y}_{i,j}|\boldsymbol{X}_{i,j})$ is the probability density function of $\boldsymbol{Y}_{i,j}$ conditional on $\boldsymbol{X}_{i,j}$, following an exponential-family distribution such as Bernoulli and Poisson distributions. The resulting loss function satisfies restricted strong convexity and smoothness with a condition number $\kappa_f > 1$ depending on the property of the specific distribution [GRG14, Laf15].

Indeed, the treatment of a general loss function brings the condition number of $f(\cdot)$ under the spotlight, since in our earlier case studies $\kappa_f \approx 1$. Our purpose is thus to understand the interplay of two types of conditioning numbers in the convergence of first-order methods. For simplicity, we assume that $f(\cdot)$ is minimized at the ground truth rank-r matrix $\boldsymbol{X}_\star$.[4] The following theorem establishes that as long as properly initialized, then `ScaledGD` converges linearly at a constant rate.

Theorem 4. *Suppose that $f(\cdot)$ is rank-$2r$ restricted L-smooth and μ-strongly convex, of which $\boldsymbol{X}_\star$ is a minimizer, and that the initialization $\boldsymbol{F}_0$ satisfies $\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.1\sigma_r(\boldsymbol{X}_\star)/\sqrt{\kappa_f}$. If the step size obeys $0 < \eta \le 0.4/L$, then for all $t \ge 0$, the iterates of `ScaledGD` in (27) satisfy*

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le (1 - 0.7\eta\mu)^t 0.1\sigma_r(\boldsymbol{X}_\star)/\sqrt{\kappa_f}, \quad \text{and} \quad \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le (1 - 0.7\eta\mu)^t 0.15\sigma_r(\boldsymbol{X}_\star)/\sqrt{\kappa_f}.$$

Theorem 4 establishes that the distance $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ contracts linearly at a constant rate, as long as the initialization $\boldsymbol{F}_0$ is sufficiently close to $\boldsymbol{F}_\star$. To reach ϵ-accuracy, i.e. $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$, `ScaledGD` takes at most $T = O(\kappa_f \log(1/\epsilon))$ iterations, which depends only on the condition number κ_f of $f(\cdot)$, but is independent of the condition number κ of the matrix $\boldsymbol{X}_\star$. In contrast, prior theory of vanilla gradient descent [PKCS18, BKS16] requires $O(\kappa_f\kappa \log(1/\epsilon))$ iterations, which is worse than our rate by a factor of κ.

3 Proof Sketch

In this section, we sketch the proof of the main theorems, highlighting the role of the scaled distance metric (cf. (10)) in these analyses.

[4] In practice, due to the presence of statistical noise, the minimizer of $f(\cdot)$ might be only approximately low-rank, to which our analysis can be extended in a straightforward fashion.

3.1 A warm-up analysis: matrix factorization

Let us consider the problem of factorizing a matrix $\boldsymbol{X}_\star$ into two low-rank factors:

$$\underset{\boldsymbol{F}\in\mathbb{R}^{(n_1+n_2)\times r}}{\text{minimize}} \quad \mathcal{L}(\boldsymbol{F}) = \frac{1}{2}\left\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}}^2. \tag{29}$$

For this toy problem, the update rule of `ScaledGD` is given as

$$\begin{aligned}\boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1},\\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}.\end{aligned} \tag{30}$$

To shed light on why `ScaledGD` is robust to ill-conditioning, it is worthwhile to think of `ScaledGD` as a quasi-Newton algorithm: the following proposition (proven in Appendix B.1) reveals that `ScaledGD` is equivalent to approximating the Hessian of the loss function in (29) by only keeping its diagonal blocks.

Proposition 2. *For the matrix factorization problem (29), `ScaledGD` is equivalent to the following update rule*

$$\mathrm{vec}(\boldsymbol{F}_{t+1}) = \mathrm{vec}(\boldsymbol{F}_t) - \eta\begin{bmatrix}\nabla^2_{\boldsymbol{L},\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}_t) & \boldsymbol{0}\\ \boldsymbol{0} & \nabla^2_{\boldsymbol{R},\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}_t)\end{bmatrix}^{-1}\mathrm{vec}(\nabla_{\boldsymbol{F}}\mathcal{L}(\boldsymbol{F}_t)).$$

Here, $\nabla^2_{\boldsymbol{L},\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}_t)$ (resp. $\nabla^2_{\boldsymbol{R},\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}_t)$) denotes the second order derivative w.r.t. $\boldsymbol{L}$ (resp. $\boldsymbol{R}$) at $\boldsymbol{F}_t$.

The following theorem, whose proof can be found in Appendix B.2, formally establishes that as long as `ScaledGD` is initialized close to the ground truth, $\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)$ will contract at a constant linear rate for the matrix factorization problem.

Theorem 5. *Suppose that the initialization $\boldsymbol{F}_0$ satisfies $\mathrm{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star)\le 0.1\sigma_r(\boldsymbol{X}_\star)$. If the step size obeys $0<\eta\le 2/3$, then for all $t\ge 0$, the iterates of the `ScaledGD` method in (30) satisfy*

$$\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\le(1-0.7\eta)^t 0.1\sigma_r(\boldsymbol{X}_\star), \quad \textit{and} \quad \left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top-\boldsymbol{X}_\star\right\|_{\mathsf{F}}\le(1-0.7\eta)^t 0.15\sigma_r(\boldsymbol{X}_\star).$$

Comparing to the rate of contraction $(1-1/\kappa)$ of gradient descent for matrix factorization [MLC21, CLC19], Theorem 5 demonstrates that the preconditioners indeed allow better search directions in the local neighborhood of the ground truth, and hence a faster convergence rate.

3.2 Proof outline for matrix sensing

It can be seen that the update rule (15) of `ScaledGD` in Algorithm 1 closely mimics (30) when $\mathcal{A}(\cdot)$ satisfies the RIP. Therefore, leveraging the RIP of $\mathcal{A}(\cdot)$ and Theorem 5, we can establish the following local convergence guarantee of Algorithm 1, which has a weaker requirement on δ_{2r} than the main theorem (cf. Theorem 1).

Lemma 1. *Suppose that $\mathcal{A}(\cdot)$ obeys the $2r$-RIP with $\delta_{2r}\le 0.02$. If the t-th iterate satisfies $\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\le 0.1\sigma_r(\boldsymbol{X}_\star)$, then $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}\le 1.5\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)$. In addition, if the step size obeys $0<\eta\le 2/3$, then the $(t+1)$-th iterate $\boldsymbol{F}_{t+1}$ of the `ScaledGD` method in (15) of Algorithm 1 satisfies*

$$\mathrm{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\le(1-0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star).$$

It then boils to down to finding a good initialization, for which we have the following lemma on the quality of the spectral initialization.

Lemma 2. *Suppose that $\mathcal{A}(\cdot)$ obeys the $2r$-RIP with a constant δ_{2r}. Then the spectral initialization in (14) for low-rank matrix sensing satisfies*

$$\mathrm{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star)\le 5\delta_{2r}\sqrt{r}\kappa\sigma_r(\boldsymbol{X}_\star).$$

Therefore, as long as δ_{2r} is small enough, say $\delta_{2r}\le 0.02/(\sqrt{r}\kappa)$ as specified in Theorem 1, the initial distance satisfies $\mathrm{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star)\le 0.1\sigma_r(\boldsymbol{X}_\star)$, allowing us to invoke Lemma 1 recursively. The proof of Theorem 1 is then complete. The proofs of Lemmas 1-2 can be found in Appendix C.

3.3 Proof outline for robust PCA

As before, we begin with the following local convergence guarantee of Algorithm 2, which has a weaker requirement on α than the main theorem (cf. Theorem 2). The difference with low-rank matrix sensing is that local convergence for robust PCA requires a further incoherence condition on the iterates (cf. (31)), where we recall from (11) that $\boldsymbol{Q}_t$ is the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$.

Lemma 3. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent and $\alpha \le 10^{-4}/(\mu r)$. If the t-th iterate satisfies $\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$ and the incoherence condition*

$$\sqrt{n_1}\left\|(\boldsymbol{L}_t\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \vee \sqrt{n_2}\left\|(\boldsymbol{R}_t\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star), \tag{31}$$

then $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le 1.5\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$. In addition, if the step size obeys $0.1 \le \eta \le 2/3$, then the $(t+1)$-th iterate $\boldsymbol{F}_{t+1}$ of the `ScaledGD` method in (20) of Algorithm 2 satisfies

$$\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \le (1 - 0.6\eta)\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

and the incoherence condition

$$\sqrt{n_1}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \vee \sqrt{n_2}\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star).$$

As long as the initialization is close to the ground truth and satisfies the incoherence condition, Lemma 3 ensures that the iterates of `ScaledGD` remain incoherent and converge linearly. This allows us to remove the unnecessary projection step in [YPCC16], whose main objective is to ensure the incoherence of the iterates.

We are left with checking the initial conditions. The following lemma ensures that the spectral initialization in (19) is close to the ground truth as long as α is sufficiently small.

Lemma 4. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent. Then the spectral initialization (19) for robust PCA satisfies*

$$\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 20\alpha\mu r^{3/2}\kappa\sigma_r(\boldsymbol{X}_\star).$$

As a result, setting $\alpha \le 10^{-3}/(\mu r^{3/2}\kappa)$, the spectral initialization satisfies $\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$. In addition, we need to make sure that the spectral initialization satisfies the incoherence condition, which is provided in the following lemma.

Lemma 5. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent and $\alpha \le 0.1/(\mu r\kappa)$, and that $\operatorname{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$. Then the spectral initialization (19) satisfies the incoherence condition*

$$\sqrt{n_1}\left\|(\boldsymbol{L}_0\boldsymbol{Q}_0 - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \vee \sqrt{n_2}\left\|(\boldsymbol{R}_0\boldsymbol{Q}_0^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star).$$

Combining Lemmas 3-5 finishes the proof of Theorem 2. The proofs of the the three supporting lemmas can be found in Section D.

3.4 Proof outline for matrix completion

A key property of the new projection operator. We start with the following lemma that entails a key property of the scaled projection (24), which ensures the scaled projection satisfies both non-expansiveness and incoherence under the scaled metric.

Lemma 6. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent, and $\operatorname{dist}(\widetilde{\boldsymbol{F}}, \boldsymbol{F}_\star) \le \epsilon\sigma_r(\boldsymbol{X}_\star)$ for some $\epsilon < 1$. Set $B \ge (1+\epsilon)\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star)$, then $\mathcal{P}_B(\widetilde{\boldsymbol{F}})$ satisfies the non-expansiveness*

$$\operatorname{dist}(\mathcal{P}_B(\widetilde{\boldsymbol{F}}), \boldsymbol{F}_\star) \le \operatorname{dist}(\widetilde{\boldsymbol{F}}, \boldsymbol{F}_\star),$$

and the incoherence condition

$$\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{R}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{L}^\top\|_{2,\infty} \le B.$$

It is worth noting that the incoherence condition adopts a slightly different form than that of robust PCA, which is more convenient for matrix completion. The next lemma guarantees the fast local convergence of Algorithm 3 as long as the sample complexity is large enough and the parameter B is set properly.

Lemma 7. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent, and $p \geq C(\mu r\kappa^4 \vee \log(n_1 \vee n_2))\mu r/(n_1 \wedge n_2)$ for some sufficiently large constant C. Set the projection radius as $B = C_B\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star)$ for some constant $C_B \geq 1.02$. Under an event $\mathcal{E}$ which happens with overwhelming probability (i.e. at least $1 - c_1(n_1 \vee n_2)^{-c_2}$), if the t-th iterate satisfies $\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \leq 0.02\sigma_r(\boldsymbol{X}_\star)$, and the incoherence condition*

$$\sqrt{n_1}\|\boldsymbol{L}_t\boldsymbol{R}_t^\top\|_{2,\infty} \vee \sqrt{n_1}\|\boldsymbol{R}_t\boldsymbol{L}_t^\top\|_{2,\infty} \leq B,$$

then $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \leq 1.5\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$. In addition, if the step size obeys $0 < \eta \leq 2/3$, then the $(t+1)$-th iterate $\boldsymbol{F}_{t+1}$ of the `ScaledPGD` method in (26) of Algorithm 3 satisfies

$$\operatorname{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \leq (1 - 0.6\eta)\operatorname{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

and the incoherence condition

$$\sqrt{n_1}\|\boldsymbol{L}_{t+1}\boldsymbol{R}_{t+1}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}_{t+1}\boldsymbol{L}_{t+1}^\top\|_{2,\infty} \leq B.$$

As long as we can find an initialization that is close to the ground truth and satisfies the incoherence condition, Lemma 7 ensures that the iterates of `ScaledPGD` remain incoherent and converge linearly. The follow lemma ensures that such an initialization can be ensured via the spectral method.

Lemma 8. *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent, then with overwhelming probability, the spectral initialization before projection $\widetilde{\boldsymbol{F}}_0 := \begin{bmatrix} \boldsymbol{U}_0\boldsymbol{\Sigma}_0^{1/2} \\ \boldsymbol{V}_0\boldsymbol{\Sigma}_0^{1/2} \end{bmatrix}$ in (25) satisfies*

$$\operatorname{dist}(\widetilde{\boldsymbol{F}}_0, \boldsymbol{F}_\star) \leq C_0\left(\frac{\mu r\log(n_1 \vee n_2)}{p\sqrt{n_1 n_2}} + \sqrt{\frac{\mu r\log(n_1 \vee n_2)}{p(n_1 \wedge n_2)}}\right)5\sqrt{r}\kappa\sigma_r(\boldsymbol{X}_\star).$$

Therefore, as long as $p \geq C\mu r^2\kappa^2\log(n_1 \vee n_2)/(n_1 \wedge n_2)$ for some sufficiently large constant C, the initial distance satisfies $\operatorname{dist}(\widetilde{\boldsymbol{F}}_0, \boldsymbol{F}_\star) \leq 0.02\sigma_r(\boldsymbol{X}_\star)$. One can then invoke Lemma 6 to see that $\boldsymbol{F}_0 = \mathcal{P}_B(\widetilde{\boldsymbol{F}}_0)$ meets the requirements of Lemma 7 due to the non-expansiveness and incoherence properties of the projection operator. The proofs of the the the supporting lemmas can be found in Section E.

4 Numerical Experiments

In this section, we provide numerical experiments to corroborate our theoretical findings, with the codes available at

https://github.com/Titan-Tong/ScaledGD.

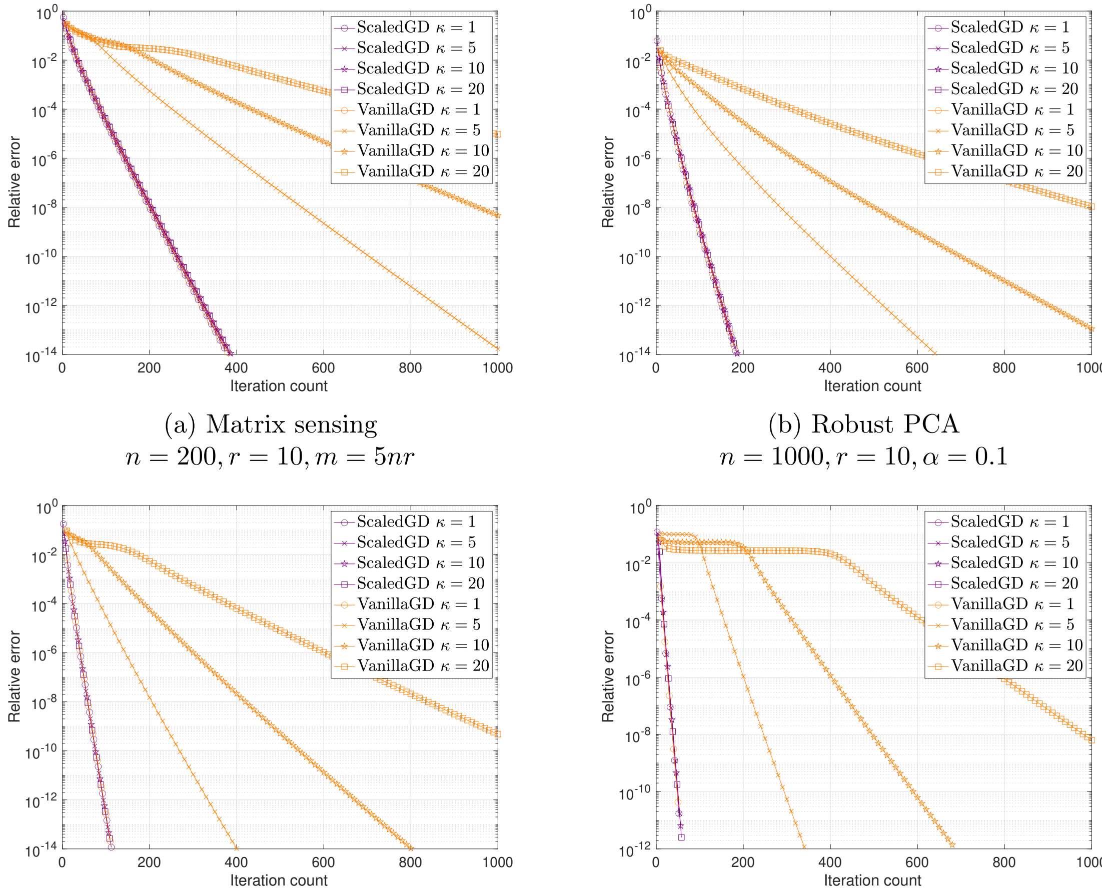
The simulations are performed in Matlab with a 3.6 GHz Intel Xeon Gold 6244 CPU.

4.1 Comparison with vanilla GD

To begin, we compare the iteration complexity of `ScaledGD` with vanilla gradient descent (GD). The update rule of vanilla GD for solving (2) is given as

$$\begin{aligned} \boldsymbol{L}_{t+1} &= \boldsymbol{L}_t - \eta_{\mathsf{GD}}\nabla_{\boldsymbol{L}}\mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t), \\ \boldsymbol{R}_{t+1} &= \boldsymbol{R}_t - \eta_{\mathsf{GD}}\nabla_{\boldsymbol{R}}\mathcal{L}(\boldsymbol{L}_t, \boldsymbol{R}_t), \end{aligned} \tag{32}$$

where $\eta_{\mathsf{GD}} = \eta/\sigma_1(\boldsymbol{X}_\star)$ stands for the step size for gradient descent. This choice is often recommended by the theory of vanilla GD [TBS+16, YPCC16, MWCC19] and the scaling by $\sigma_1(\boldsymbol{X}_\star)$ is needed for its convergence. For ease of comparison, we fix $\eta = 0.5$ for both `ScaledGD` and vanilla GD (see Figure 4 for justifications). Both algorithms start from the same spectral initialization. To avoid notational clutter, we work on square *asymmetric* matrices with $n_1 = n_2 = n$. We consider four low-rank matrix estimation tasks:



(a) Matrix sensing
$n = 200, r = 10, m = 5nr$

(b) Robust PCA
$n = 1000, r = 10, \alpha = 0.1$

(c) Matrix completion
$n = 1000, r = 10, p = 0.2$

(d) Hankel matrix completion
$n = 1000, r = 10, p = 0.2$

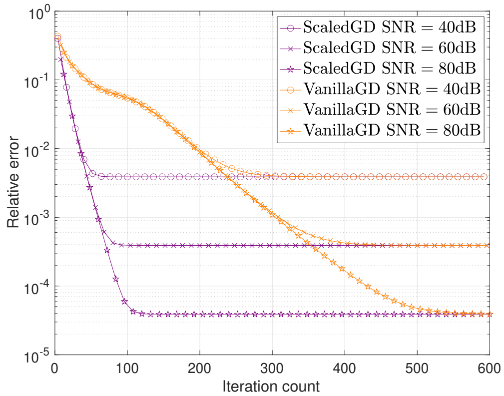
Figure 2: The relative errors of `ScaledGD` and vanilla GD with respect to the iteration count under different condition numbers $\kappa = 1, 5, 10, 20$ for (a) matrix sensing, (b) robust PCA, (c) matrix completion, and (d) Hankel matrix completion.

- *Low-rank matrix sensing.* The problem formulation is detailed in Section 2.2. Here, we collect $m = 5nr$ measurements in the form of $\boldsymbol{y}_k = \langle \boldsymbol{A}_k, \boldsymbol{X}_\star \rangle + \boldsymbol{w}_k$, in which the measurement matrices $\boldsymbol{A}_k$ are generated with i.i.d. Gaussian entries with zero mean and variance $1/m$, and $\boldsymbol{w}_k \sim \mathcal{N}(0, \sigma_w^2)$ are i.i.d. Gaussian noises.

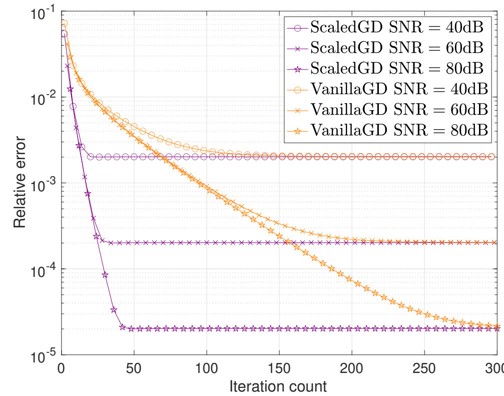
- *Robust PCA.* The problem formulation is stated in Section 2.3. We generate the corruption with a sparse matrix $\boldsymbol{S}_\star \in \mathcal{S}_\alpha$ with $\alpha = 0.1$. More specifically, we generate a matrix with standard Gaussian entries and pass it through $\mathcal{T}_\alpha[\cdot]$ to obtain $\boldsymbol{S}_\star$. The observation is $\boldsymbol{Y} = \boldsymbol{X}_\star + \boldsymbol{S}_\star + \boldsymbol{W}$, where $\boldsymbol{W}_{i,j} \sim \mathcal{N}(0, \sigma_w^2)$ are i.i.d. Gaussian noises.

- *Matrix completion.* The problem formulation is stated in Section 2.4. We assume random Bernoulli observations, where each entry of $\boldsymbol{X}_\star$ is observed with probability $p = 0.2$ independently. The observation is $\boldsymbol{Y} = \mathcal{P}_\Omega(\boldsymbol{X}_\star + \boldsymbol{W})$, where $\boldsymbol{W}_{i,j} \sim \mathcal{N}(0, \sigma_w^2)$ are i.i.d. Gaussian noises. Moreover, we perform the scaled gradient updates without projections.

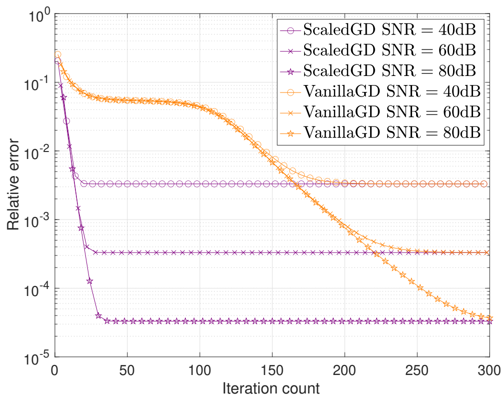
- *Hankel matrix completion.* Briefly speaking, a Hankel matrix shares the same value along each skew-diagonal, and we aim at recovering a low-rank Hankel matrix from observing a few skew-diagonals [CC14, CWW18]. We assume random Bernoulli observations, where each skew-diagonal of $\boldsymbol{X}_\star$ is observed with



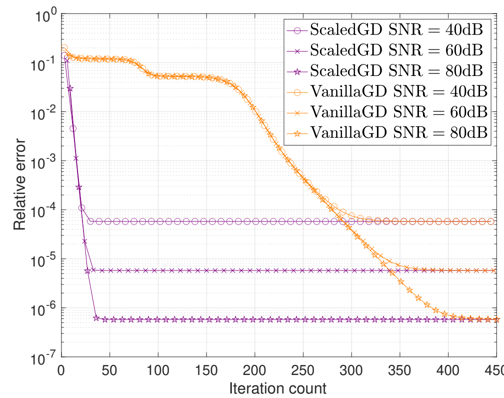
(a) Matrix Sensing
$n = 200, r = 10, m = 5nr$



(b) Robust PCA
$n = 1000, r = 10, \alpha = 0.1$



(c) Matrix completion
$n = 1000, r = 10, p = 0.2$



(d) Hankel matrix completion
$n = 1000, r = 10, p = 0.2$

Figure 3: The relative errors of `ScaledGD` and vanilla GD with respect to the iteration count under the condition number $\kappa = 10$ and signal-to-noise ratios SNR $= 40, 60, 80$dB for (a) matrix sensing, (b) robust PCA, (c) matrix completion, and (d) Hankel matrix completion.

probability $p = 0.2$ independently. The loss function is

$$\mathcal{L}(\boldsymbol{L}, \boldsymbol{R}) = \frac{1}{2p} \left\| \mathcal{H}_{\Omega}(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{Y}) \right\|_{\mathsf{F}}^2 + \frac{1}{2} \left\| (\mathcal{I} - \mathcal{H})(\boldsymbol{L}\boldsymbol{R}^\top) \right\|_{\mathsf{F}}^2, \tag{33}$$

where $\mathcal{I}(\cdot)$ denotes the identity operator, and the Hankel projection is defined as $\mathcal{H}(\boldsymbol{X}) \coloneqq \sum_{k=1}^{2n-1} \langle \boldsymbol{H}_k, \boldsymbol{X} \rangle \boldsymbol{H}_k$, which maps $\boldsymbol{X}$ to its closest Hankel matrix. Here, the Hankel basis matrix $\boldsymbol{H}_k$ is the $n \times n$ matrix with the entries in the k-th skew diagonal as $\frac{1}{\sqrt{\omega_k}}$, and all other entries as 0, where ω_k is the length of the k-th skew diagonal. Note that $\boldsymbol{X}$ is a Hankel matrix if and only if $(\mathcal{I} - \mathcal{H})(\boldsymbol{X}) = \boldsymbol{0}$. The Hankel projection on the observation index set Ω is defined as $\mathcal{H}_{\Omega}(\boldsymbol{X}) \coloneqq \sum_{k \in \Omega} \langle \boldsymbol{H}_k, \boldsymbol{X} \rangle \boldsymbol{H}_k$. The observation is $\boldsymbol{Y} = \mathcal{H}_{\Omega}(\boldsymbol{X}_\star + \boldsymbol{W})$, where $\boldsymbol{W}$ is a Hankel matrix whose entries along each skew-diagonal are i.i.d. Gaussian noises $\mathcal{N}(0, \sigma_w^2)$.

For the first three problems, we generate the ground truth matrix $\boldsymbol{X}_\star \in \mathbb{R}^{n \times n}$ in the following way. We first generate an $n \times r$ matrix with i.i.d. random signs, and take its r left singular vectors as $\boldsymbol{U}_\star$, and similarly for $\boldsymbol{V}_\star$. The singular values are set to be linearly distributed from 1 to $1/\kappa$. The ground truth is then defined as $\boldsymbol{X}_\star = \boldsymbol{U}_\star \boldsymbol{\Sigma}_\star \boldsymbol{V}_\star^\top$ which has the specified condition number κ and rank r. For Hankel matrix completion, we generate $\boldsymbol{X}_\star$ as an $n \times n$ Hankel matrix with entries given as

$$(\boldsymbol{X}_\star)_{i,j} = \sum_{\ell=1}^{r} \frac{\sigma_\ell}{n} e^{2\pi\imath(i+j-2)f_\ell}, \quad i,j = 1,\dots,n,$$

where f_ℓ, $\ell = 1,\dots,r$ are randomly chosen from $1/n, 2/n, \dots, 1$, and σ_ℓ are linearly distributed from 1 to $1/\kappa$. The Vandermonde decomposition lemma tells that $\boldsymbol{X}_\star$ has rank r and singular values σ_ℓ, $\ell = 1,\dots,r$.

We first illustrate the convergence performance under noise-free observations, i.e. $\sigma_w = 0$. We plot the relative reconstruction error $\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_{\mathsf{F}}/\|\boldsymbol{X}_\star\|_{\mathsf{F}}$ with respect to the iteration count t in Figure 2 for the four problems under different condition numbers $\kappa = 1, 5, 10, 20$. For all these models, we can see that `ScaledGD` has a convergence rate independent of κ, with all curves almost overlay on each other. Under good conditioning $\kappa = 1$, `ScaledGD` converges at the same rate as vanilla GD; under ill conditioning, i.e. when κ is large, `ScaledGD` converges much faster than vanilla GD and leads to significant computational savings.

We next move to demonstrate that `ScaledGD` is robust to small additive noises. Denote the signal-to-noise ratio as $\mathsf{SNR} \coloneqq 10\log_{10}\frac{\|\boldsymbol{X}_\star\|_{\mathsf{F}}^2}{n^2\sigma_w^2}$ in dB. We plot the reconstruction error $\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_{\mathsf{F}}/\|\boldsymbol{X}_\star\|_{\mathsf{F}}$ with respect to the iteration count t in Figure 3 under the condition number $\kappa = 10$ and various $\mathsf{SNR} = 40, 60, 80$dB. We can see that `ScaledGD` and vanilla GD achieve the same statistical error eventually, but `ScaledGD` converges much faster. In addition, the convergence speeds are not influenced by the noise levels.

Careful readers might wonder how sensitivity our comparisons are with respect to the choice of step sizes. To address this, we illustrate the convergence speeds of both `ScaledGD` and vanilla GD under different step sizes η for matrix completion (under the same setting as Figure 2 (c)), where similar plots can be obtained for other problems as well. We run both algorithms for at most 80 iterations, and terminate if the relative error exceeds 10^2 (which happens if the step size is too large and the algorithm diverges). Figure 4 plots the relative error with respect to the step size η for both algorithms, where we can see that `ScaledGD` outperforms vanilla GD over a large range of step sizes, even under optimized values for performance. Hence, our choice of $\eta = 0.5$ in previous experiments renders a typical comparison between `ScaledGD` and vanilla GD.

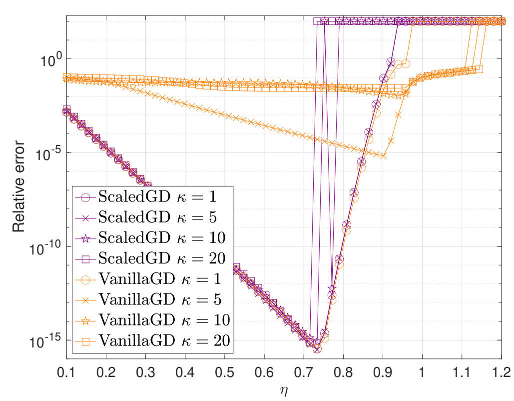


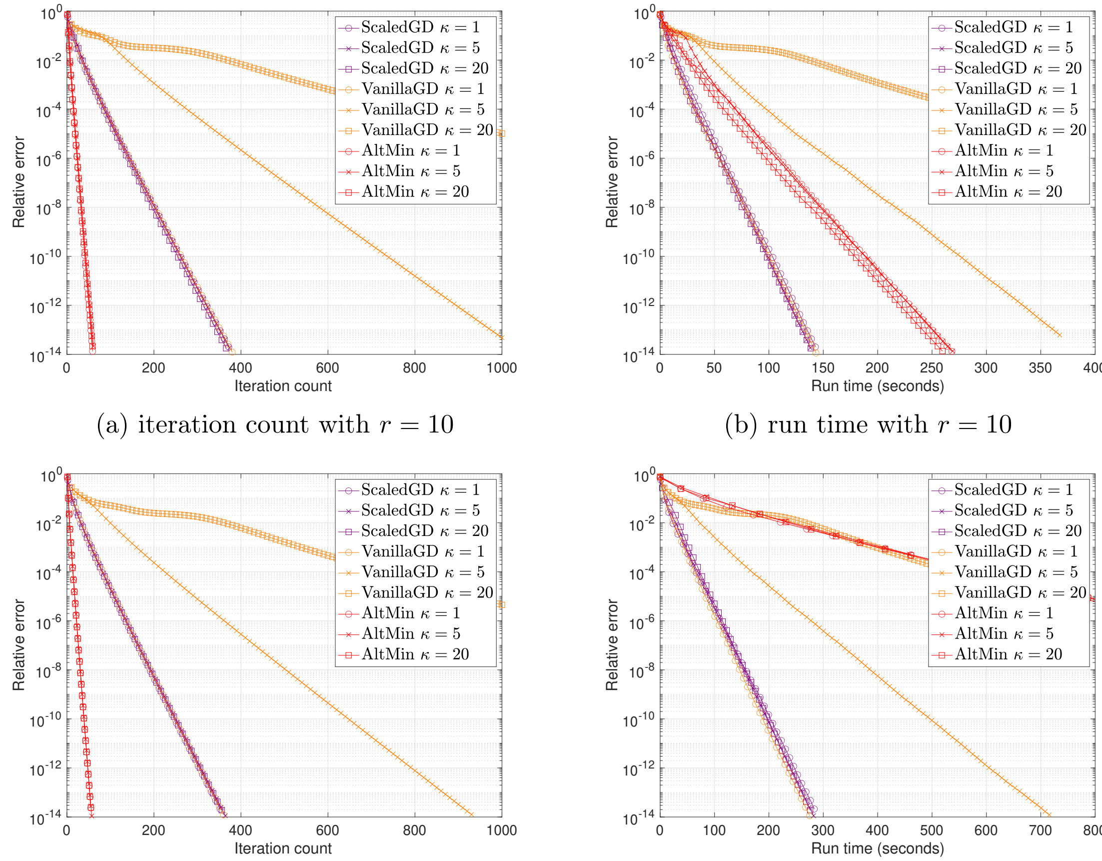
Figure 4: The relative errors of `ScaledGD` and vanilla GD after 80 iterations with respect to different step sizes η from 0.1 to 1.2, for matrix completion with $n = 1000, r = 10, p = 0.2$.

4.2 Run time comparisons

We now compare the run time of `ScaledGD` with vanilla GD and alternating minimization (`AltMin`) [JNS13]. Specifically, for matrix sensing, alternating minimization (`AltMinSense`) updates the factors alternatively as

$$\boldsymbol{L}_{t+1} = \operatorname*{argmin}_{\boldsymbol{L}} \left\|\mathcal{A}(\boldsymbol{L}\boldsymbol{R}_t^\top) - \boldsymbol{y}\right\|_2^2,$$

$$\boldsymbol{R}_{t+1} = \operatorname*{argmin}_{\boldsymbol{R}} \left\|\mathcal{A}(\boldsymbol{L}_{t+1}\boldsymbol{R}^\top) - \boldsymbol{y}\right\|_2^2,$$



(a) iteration count with $r = 10$ (b) run time with $r = 10$

(c) iteration count with $r = 20$ (d) run time with $r = 20$

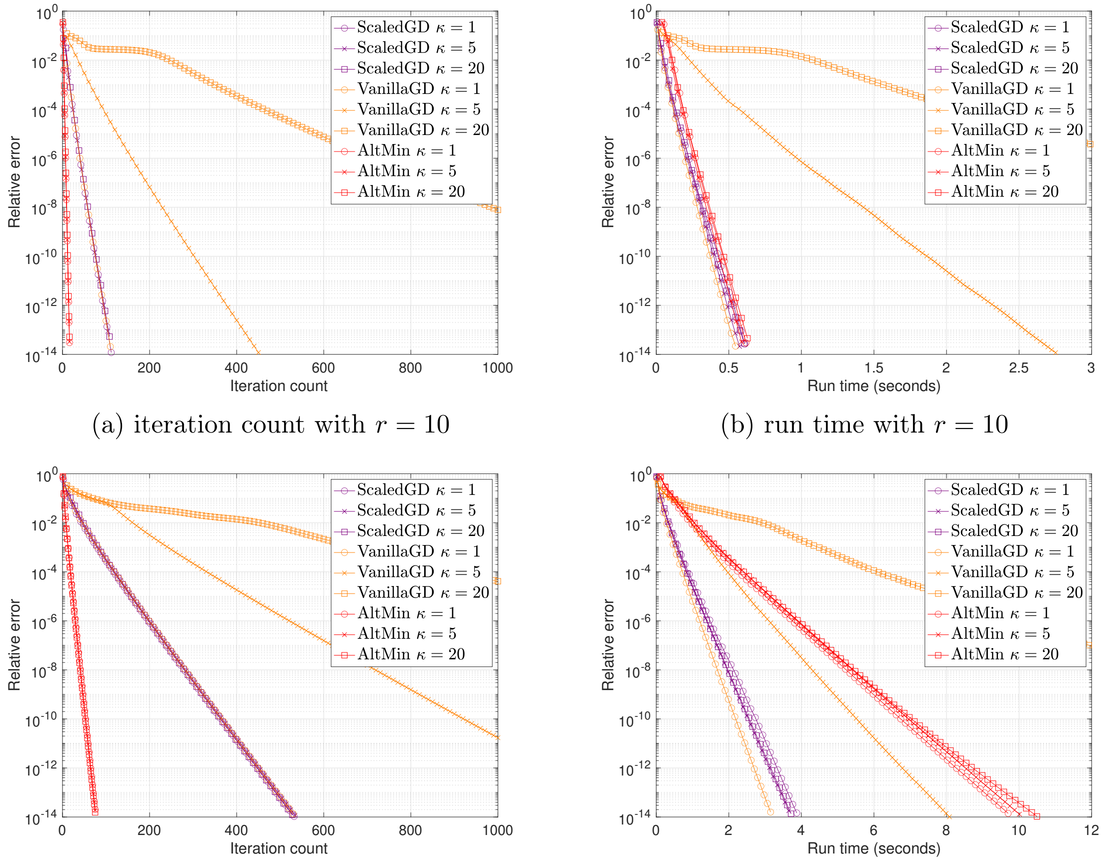
Figure 5: The relative errors of `ScaledGD`, vanilla GD and `AltMin` with respect to the iteration count and run time (in seconds) under different condition numbers $\kappa = 1, 5, 20$ for matrix sensing with $n = 200$, and $m = 5nr$. (a, b): $r = 10$; (c, d): $r = 20$.

which corresponds to solving two least-squares problems. For matrix completion, the update rule of alternating minimization proceeds as

$$\boldsymbol{L}_{t+1} = \operatorname*{argmin}_{\boldsymbol{L}} \left\| \mathcal{P}_\Omega(\boldsymbol{L}\boldsymbol{R}_t^\top - \boldsymbol{Y}) \right\|_2^2,$$
$$\boldsymbol{R}_{t+1} = \operatorname*{argmin}_{\boldsymbol{R}} \left\| \mathcal{P}_\Omega(\boldsymbol{L}_{t+1}\boldsymbol{R}^\top - \boldsymbol{Y}) \right\|_2^2,$$

which can be implemented more efficiently since each row of $\boldsymbol{L}$ (resp. $\boldsymbol{R}$) can be updated independently via solving a much smaller least-squares problem due to the decomposable structure of the objective function. It is worth noting that, to the best of our knowledge, this most natural variant of alternating minimization for matrix completion still eludes from a provable performance guarantee, nonetheless, we choose it to compare against due to its popularity and excellent empirical performance.

Figure 5 plots the relative errors of `ScaledGD`, vanilla GD and alternating minimization (`AltMin`) with respect to the iteration count and run time (in seconds) under different condition numbers $\kappa = 1, 5, 20$; and similarly, Figure 6 plots the corresponding results for matrix completion. It can be seen that, both `ScaledGD` and `AltMin` admit a convergence rate that is independent of the condition number, where the per-iteration complexity of `AltMin` is much higher than that of `ScaledGD`. As expected, the run time of `ScaledGD` only adds a minimal overhead to vanilla GD while being much more robust to ill-conditioning. Noteworthily, `AltMin` takes much more time and becomes significantly slower than `ScaledGD` when the rank r is larger. Nonetheless, we emphasize that since the run time is impacted by many factors in terms of problem



(a) iteration count with $r = 10$

(b) run time with $r = 10$

(c) iteration count with $r = 50$

(d) run time with $r = 50$

Figure 6: The relative errors of `ScaledGD`, vanilla GD and `AltMin` with respect to the iteration count and run time (in seconds) under different condition numbers $\kappa = 1, 5, 20$ for matrix completion with $n = 1000$, and $p = 0.2$. (a, b): $r = 10$; (c, d): $r = 50$.

parameters as well as implementation details, our purpose is to demonstrate the competitive performance of `ScaledGD` over alternatives, rather than claiming it as the state-of-the-art.

5 Conclusions

This paper proposes scaled gradient descent (`ScaledGD`) for factored low-rank matrix estimation, which maintains the low per-iteration computational complexity of vanilla gradient descent, but offers significant speed-up in terms of the convergence rate with respect to the condition number κ of the low-rank matrix. In particular, we rigorously establish that for low-rank matrix sensing, robust PCA, and matrix completion, to reach ϵ-accuracy, `ScaledGD` only takes $O(\log(1/\epsilon))$ iterations without the dependency on the condition number when initialized via the spectral method, under standard assumptions. The key to our analysis is the introduction of a new distance metric that takes into account the preconditioning and unbalancedness of the low-rank factors, and we have developed new tools to analyze the trajectory of `ScaledGD` under this new metric. This work opens up many venues for future research, as we discuss below.

- *Improved analysis.* In this paper, we have focused on establishing the fast local convergence rate. It is interesting to study if the theory developed herein can be further strengthened in terms of sample complexity and the size of basin of attraction. For matrix completion, it will be interesting to see if a similar guarantee continues to hold in the absence of the projection, which will generalize recent works [MWCC19, CLL20] that successfully removed these projections for vanilla gradient descent.

- *Other low-rank recovery problems.* Besides the problems studied herein, there are many other applications involving the recovery of an ill-conditioned low-rank matrix, such as robust PCA with missing data, quadratic sampling, and so on. It is of interest to establish fast convergence rates of `ScaledGD` that are independent of the condition number for these problems as well. In addition, it is worthwhile to explore if a similar preconditioning trick can be useful to problems beyond low-rank matrix estimation. One recent attempt is to generalize `ScaledGD` for low-rank tensor estimation [TMPB+21].
- *Acceleration schemes?* As it is evident from our analysis of the general loss case, `ScaledGD` may still converge slowly when the loss function is ill-conditioned over low-rank matrices, i.e. κ_f is large. In this case, it might be of interest to combine techniques such as momentum [KC12] from the optimization literature to further accelerate the convergence. In our companion paper [TMC21], we have extended `ScaledGD` to nonsmooth formulations, which possess better curvatures than their smooth counterparts for certain problems.

Acknowledgements

The work of T. Tong and Y. Chi is supported in part by ONR under the grants N00014-18-1-2142 and N00014-19-1-2404, by ARO under the grant W911NF-18-1-0303, and by NSF under the grants CAREER ECCS-1818571, CCF-1806154 and CCF-1901199.

References

[BH89] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

[BKS16] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582. PMLR, 2016.

[BNS16] S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[CC14] Y. Chen and Y. Chi. Robust spectral compressed sensing via structured matrix completion. *IEEE Transactions on Information Theory*, 60(10):6576–6601, 2014.

[CC18] Y. Chen and Y. Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14 – 31, 2018.

[CCD+21] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, pages 1–89, 2021.

[CCF+20] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020.

[CFMY20] Y. Chen, J. Fan, C. Ma, and Y. Yan. Bridging convex and nonconvex optimization in robust PCA: Noise, outliers, and missing data. *arXiv preprint arXiv:2001.05484*, 2020.

[Che15] Y. Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

[CL19] J. Chen and X. Li. Model-free nonconvex matrix completion: Local minima analysis and applications in memory-efficient kernel PCA. *Journal of Machine Learning Research*, 20(142):1–39, 2019.

[CLC19] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[CLL20] J. Chen, D. Liu, and X. Li. Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 66(9):5806–5841, 2020.

[CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011.

[CLS15] E. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007, 2015.

[CP11] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[CSPW11] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[CW15] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[CWW18] J.-F. Cai, T. Wang, and K. Wei. Spectral compressed sensing via projected gradient descent. *SIAM Journal on Optimization*, 28(3):2625–2653, 2018.

[DDP17] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *arXiv preprint arXiv:1711.03247*, 2017.

[DHL18] S. S. Du, W. Hu, and J. D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, pages 384–395, 2018.

[GHJY15] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points-online stochastic gradient for tensor decomposition. In *Conference on Learning Theory (COLT)*, pages 797–842, 2015.

[GJZ17] R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242, 2017.

[GLM16] R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[GRG14] S. Gunasekar, P. Ravikumar, and J. Ghosh. Exponential family matrix completion under structural constraints. In *International Conference on Machine Learning*, pages 1917–1925, 2014.

[HW14] M. Hardt and M. Wootters. Fast matrix completion without the condition number. In *Proceedings of The 27th Conference on Learning Theory*, pages 638–678, 2014.

[JGN+17] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732, 2017.

[JK17] P. Jain and P. Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.

[JMD10] P. Jain, R. Meka, and I. S. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.

[JNS13] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[Kaw16] K. Kawaguchi. Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016.

[KC12] A. Kyrillidis and V. Cevher. Matrix ALPS: Accelerated low rank and sparse matrix reconstruction. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 185–188. IEEE, 2012.

[Laf15] J. Lafond. Low rank matrix completion with exponential family noise. In *Conference on Learning Theory*, pages 1224–1243, 2015.

[LHLZ20] Y. Luo, W. Huang, X. Li, and A. R. Zhang. Recursive importance sketching for rank constrained least squares: Algorithms and high-order convergence. *arXiv preprint arXiv:2011.08360*, 2020.

[LLSW19] X. Li, S. Ling, T. Strohmer, and K. Wei. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, 47(3):893–934, 2019.

[LMCC21] Y. Li, C. Ma, Y. Chen, and Y. Chi. Nonconvex matrix factorization from rank-one measurements. *IEEE Transactions on Information Theory*, 67(3):1928–1950, 2021.

[MAS12] B. Mishra, K. A. Apuroop, and R. Sepulchre. A Riemannian geometry for low-rank matrix completion. *arXiv preprint arXiv:1211.1550*, 2012.

[Maz16] M. Mazeika. The singular value decomposition and low rank approximation. Technical report, University of Chicago, 2016.

[MBM18] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

[MLC21] C. Ma, Y. Li, and Y. Chi. Beyond Procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. *IEEE Transactions on Signal Processing*, 69:867–877, 2021.

[MS16] B. Mishra and R. Sepulchre. Riemannian preconditioning. *SIAM Journal on Optimization*, 26(1):635–660, 2016.

[MWCC19] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, pages 1–182, 2019.

[NNS+14] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[NP06] Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[PKCS17] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

[PKCS18] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably. *SIAM Journal on Imaging Sciences*, 11(4):2165–2204, 2018.

[RFP10] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[SL16] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

[SQW15] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery using nonconvex optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2351–2360, 2015.

[SQW18] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[SWW17] S. Sanghavi, R. Ward, and C. D. White. The local convexity of solving systems of quadratic equations. *Results in Mathematics*, 71(3-4):569–608, 2017.

[TBS+16] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *International Conference Machine Learning*, pages 964–973, 2016.

[TMC21] T. Tong, C. Ma, and Y. Chi. Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number. *IEEE Transactions on Signal Processing*, 2021.

[TMPB+21] T. Tong, C. Ma, A. Prater-Bennette, E. Tripp, and Y. Chi. Scaling and scalability: Provable nonconvex low-rank tensor estimation from incomplete measurements. *arXiv preprint arXiv:2104.14526*, 2021.

[TW16] J. Tanner and K. Wei. Low rank matrix completion by alternating steepest descent methods. *Applied and Computational Harmonic Analysis*, 40(2):417–429, 2016.

[WCCL16] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung. Guarantees of Riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

[YPCC16] X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust PCA via gradient descent. In *Advances in neural information processing systems*, pages 4152–4160, 2016.

[ZL15] Q. Zheng and J. Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

[ZL16] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[ZLTW18] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

A Technical Lemmas

This section gathers several useful lemmas that will be used in the appendix. Throughout all lemmas, we use $\boldsymbol{X}_\star$ to denote the ground truth low-rank matrix, with its compact SVD as $\boldsymbol{X}_\star = \boldsymbol{U}_\star\boldsymbol{\Sigma}_\star\boldsymbol{V}_\star^\top$, and the stacked factor matrix is defined as $\boldsymbol{F}_\star = \begin{bmatrix}\boldsymbol{L}_\star \\ \boldsymbol{R}_\star\end{bmatrix} = \begin{bmatrix}\boldsymbol{U}_\star\boldsymbol{\Sigma}_\star^{1/2} \\ \boldsymbol{V}_\star\boldsymbol{\Sigma}_\star^{1/2}\end{bmatrix}$.

A.1 New distance metric

We begin with the investigation of the new distance metric (10), where the matrix $\boldsymbol{Q}$ that attains the infimum, if exists, is called the optimal alignment matrix between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$; see (11). Notice that (10) involves a minimization problem over an open set (the set of invertible matrices). Hence the minimizer, i.e. the optimal alignment matrix between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$ is not guaranteed to be attained. Fortunately, a simple sufficient condition guarantees the existence of the minimizer; see the lemma below.

Lemma 9. *Fix any factor matrix* $\boldsymbol{F} = \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{R} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$. *Suppose that*

$$\mathrm{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) = \sqrt{\inf_{\boldsymbol{Q}\in\mathrm{GL}(r)} \left\| (\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2} < \sigma_r(\boldsymbol{X}_\star), \tag{34}$$

then the minimizer of the above minimization problem is attained at some $\boldsymbol{Q} \in \mathrm{GL}(r)$, *i.e. the optimal alignment matrix* $\boldsymbol{Q}$ *between* $\boldsymbol{F}$ *and* $\boldsymbol{F}_\star$ *exists.*

Proof. In view of the condition (34) and the definition of infimum, one knows that there must exist a matrix $\bar{\boldsymbol{Q}} \in \mathrm{GL}(r)$ such that

$$\sqrt{\left\| (\boldsymbol{L}\bar{\boldsymbol{Q}} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top} - \boldsymbol{R}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2} \le \epsilon\sigma_r(\boldsymbol{X}_\star),$$

for some ϵ obeying $0 < \epsilon < 1$. It further implies that

$$\left\| (\boldsymbol{L}\bar{\boldsymbol{Q}} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{-1/2} \right\|_{\mathsf{op}} \vee \left\| (\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top} - \boldsymbol{R}_\star)\, \boldsymbol{\Sigma}_\star^{-1/2} \right\|_{\mathsf{op}} \le \epsilon.$$

Invoke Weyl's inequality $|\sigma_r(\boldsymbol{A}) - \sigma_r(\boldsymbol{B})| \le \|\boldsymbol{A} - \boldsymbol{B}\|_{\mathsf{op}}$, and use that $\sigma_r(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}) = \sigma_r(\boldsymbol{U}_\star) = 1$ to obtain

$$\sigma_r(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{-1/2}) \ge \sigma_r(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}) - \left\| (\boldsymbol{L}\bar{\boldsymbol{Q}} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{-1/2} \right\|_{\mathsf{op}} \ge 1 - \epsilon. \tag{35}$$

In addition, it is straightforward to verify that

$$\begin{aligned} &\inf_{\boldsymbol{Q}\in\mathrm{GL}(r)} \left\| (\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 &(36)\\ &= \inf_{\boldsymbol{H}\in\mathrm{GL}(r)} \left\| (\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}\boldsymbol{H}^{-\top} - \boldsymbol{R}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2. &(37) \end{aligned}$$

Indeed, if the minimizer of the second optimization problem (cf. (37)) is attained at some $\boldsymbol{H}$, then $\bar{\boldsymbol{Q}}\boldsymbol{H}$ must be the minimizer of the first problem (36). Therefore, from now on, we focus on proving that the minimizer of the second problem (37) is attained at some $\boldsymbol{H}$. In view of (36) and (37), one has

$$\begin{aligned} &\inf_{\boldsymbol{H}\in\mathrm{GL}(r)} \left\| (\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}\boldsymbol{H}^{-\top} - \boldsymbol{R}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 \\ &\le \left\| (\boldsymbol{L}\bar{\boldsymbol{Q}} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top} - \boldsymbol{R}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2, \end{aligned}$$

Clearly, for any $\bar{\boldsymbol{Q}}\boldsymbol{H}$ to yield a smaller distance than $\bar{\boldsymbol{Q}}$, $\boldsymbol{H}$ must obey

$$\sqrt{\left\| (\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| (\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}\boldsymbol{H}^{-\top} - \boldsymbol{R}_\star)\, \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2} \le \epsilon\sigma_r(\boldsymbol{X}_\star).$$

It further implies that

$$\left\| (\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{-1/2} \right\|_{\mathsf{op}} \vee \left\| (\boldsymbol{R}\bar{\boldsymbol{Q}}^{-\top}\boldsymbol{H}^{-\top} - \boldsymbol{R}_\star)\, \boldsymbol{\Sigma}_\star^{-1/2} \right\|_{\mathsf{op}} \le \epsilon.$$

Invoke Weyl's inequality $|\sigma_1(\boldsymbol{A}) - \sigma_1(\boldsymbol{B})| \le \|\boldsymbol{A} - \boldsymbol{B}\|_{\mathsf{op}}$, and use that $\sigma_1(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}) = \sigma_1(\boldsymbol{U}_\star) = 1$ to obtain

$$\sigma_1(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H}\boldsymbol{\Sigma}_\star^{-1/2}) \le \sigma_1(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}) + \left\| (\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H} - \boldsymbol{L}_\star)\, \boldsymbol{\Sigma}_\star^{-1/2} \right\|_{\mathsf{op}} \le 1 + \epsilon. \tag{38}$$

Combine (35) and (38), and use the relation $\sigma_r(\boldsymbol{A})\sigma_1(\boldsymbol{B}) \le \sigma_1(\boldsymbol{A}\boldsymbol{B})$ to obtain

$$\sigma_r(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{-1/2})\sigma_1(\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{H}\boldsymbol{\Sigma}_\star^{-1/2}) \le \sigma_1(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{H}\boldsymbol{\Sigma}_\star^{-1/2}) \le \frac{1+\epsilon}{1-\epsilon}\sigma_r(\boldsymbol{L}\bar{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{-1/2}).$$

As a result, one has $\sigma_1(\mathbf{\Sigma}_\star^{1/2}\mathbf{H}\mathbf{\Sigma}_\star^{-1/2}) \le \frac{1+\epsilon}{1-\epsilon}$.

Similarly, one can show that $\sigma_1(\mathbf{\Sigma}_\star^{1/2}\mathbf{H}^{-\top}\mathbf{\Sigma}_\star^{-1/2}) \le \frac{1+\epsilon}{1-\epsilon}$, equivalently, $\sigma_r(\mathbf{\Sigma}_\star^{1/2}\mathbf{H}\mathbf{\Sigma}_\star^{-1/2}) \ge \frac{1-\epsilon}{1+\epsilon}$. Combining the above two arguments reveals that the minimization problem (37) is equivalent to the constrained problem:

$$\begin{aligned} \underset{\mathbf{H}\in\mathrm{GL}(r)}{\operatorname{minimize}} \quad & \left\| \left(\mathbf{L}\bar{\mathbf{Q}}\mathbf{H} - \mathbf{L}_\star\right)\mathbf{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 + \left\| \left(\mathbf{R}\bar{\mathbf{Q}}^{-\top}\mathbf{H}^{-\top} - \mathbf{R}_\star\right)\mathbf{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 \\ \text{s.t.} \quad & \frac{1-\epsilon}{1+\epsilon} \le \sigma_r(\mathbf{\Sigma}_\star^{1/2}\mathbf{H}\mathbf{\Sigma}_\star^{-1/2}) \le \sigma_1(\mathbf{\Sigma}_\star^{1/2}\mathbf{H}\mathbf{\Sigma}_\star^{-1/2}) \le \frac{1+\epsilon}{1-\epsilon}. \end{aligned}$$

Notice that this is a continuous optimization problem over a compact set. Apply the Weierstrass extreme value theorem to finish the proof. □

With the existence of the optimal alignment matrix in place, the following lemma provides the first-order necessary condition for the minimizer.

Lemma 10. *For any factor matrix* $\mathbf{F} = \begin{bmatrix}\mathbf{L}\\ \mathbf{R}\end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$, *suppose that the optimal alignment matrix*

$$\mathbf{Q} = \underset{\mathbf{Q}\in\mathrm{GL}(r)}{\operatorname{argmin}} \ \left\|(\mathbf{L}\mathbf{Q} - \mathbf{L}_\star)\mathbf{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\mathbf{R}\mathbf{Q}^{-\top} - \mathbf{R}_\star)\mathbf{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2$$

between $\mathbf{F}$ *and* $\mathbf{F}_\star$ *exists, then* $\mathbf{Q}$ *obeys*

$$(\mathbf{L}\mathbf{Q})^\top(\mathbf{L}\mathbf{Q} - \mathbf{L}_\star)\mathbf{\Sigma}_\star = \mathbf{\Sigma}_\star(\mathbf{R}\mathbf{Q}^{-\top} - \mathbf{R}_\star)^\top\mathbf{R}\mathbf{Q}^{-\top}. \tag{39}$$

Proof. Expand the squares in the definition of $\mathbf{Q}$ to obtain

$$\mathbf{Q} = \underset{\mathbf{Q}\in\mathrm{GL}(r)}{\operatorname{argmin}} \ \operatorname{tr}\left((\mathbf{L}\mathbf{Q} - \mathbf{L}_\star)^\top(\mathbf{L}\mathbf{Q} - \mathbf{L}_\star)\mathbf{\Sigma}_\star\right) + \operatorname{tr}\left((\mathbf{R}\mathbf{Q}^{-\top} - \mathbf{R}_\star)^\top(\mathbf{R}\mathbf{Q}^{-\top} - \mathbf{R}_\star)\mathbf{\Sigma}_\star\right).$$

Clearly, the first order necessary condition (i.e. the gradient is zero) yields

$$2\mathbf{L}^\top(\mathbf{L}\mathbf{Q} - \mathbf{L}_\star)\mathbf{\Sigma}_\star - 2\mathbf{Q}^{-\top}\mathbf{\Sigma}_\star(\mathbf{R}\mathbf{Q}^{-\top} - \mathbf{R}_\star)^\top\mathbf{R}\mathbf{Q}^{-\top} = \mathbf{0},$$

which implies the optimal alignment criterion (39). □

Last but not least, we connect the newly proposed distance to the usual Frobenius norm in Lemma 11, the proof of which is a slight modification to [TBS+16, Lemma 5.4] and [GJZ17, Lemma 41].

Lemma 11. *For any factor matrix* $\mathbf{F} = \begin{bmatrix}\mathbf{L}\\ \mathbf{R}\end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$, *the distance between* $\mathbf{F}$ *and* $\mathbf{F}_\star$ *satisfies*

$$\operatorname{dist}(\mathbf{F}, \mathbf{F}_\star) \le \left(\sqrt{2}+1\right)^{1/2}\|\mathbf{L}\mathbf{R}^\top - \mathbf{X}_\star\|_{\mathsf{F}}.$$

Proof. Suppose that $\mathbf{X} := \mathbf{L}\mathbf{R}^\top$ has compact SVD as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Without loss of generality, we can assume that $\mathbf{F} = \begin{bmatrix}\mathbf{U}\mathbf{\Sigma}^{1/2}\\ \mathbf{V}\mathbf{\Sigma}^{1/2}\end{bmatrix}$, since any factorization of $\mathbf{L}\mathbf{R}^\top$ yields the same distance. Introduce two auxiliary matrices $\bar{\mathbf{F}} := \begin{bmatrix}\mathbf{U}\mathbf{\Sigma}^{1/2}\\ -\mathbf{V}\mathbf{\Sigma}^{1/2}\end{bmatrix}$ and $\bar{\mathbf{F}}_\star := \begin{bmatrix}\mathbf{U}_\star\mathbf{\Sigma}_\star^{1/2}\\ -\mathbf{V}_\star\mathbf{\Sigma}_\star^{1/2}\end{bmatrix}$. Apply the dilation trick to obtain

$$2\begin{bmatrix}\mathbf{0} & \mathbf{X}\\ \mathbf{X}^\top & \mathbf{0}\end{bmatrix} = \mathbf{F}\mathbf{F}^\top - \bar{\mathbf{F}}\bar{\mathbf{F}}^\top, \quad 2\begin{bmatrix}\mathbf{0} & \mathbf{X}_\star\\ \mathbf{X}_\star^\top & \mathbf{0}\end{bmatrix} = \mathbf{F}_\star\mathbf{F}_\star^\top - \bar{\mathbf{F}}_\star\bar{\mathbf{F}}_\star^\top.$$

As a result, the squared Frobenius norm of $\mathbf{X} - \mathbf{X}_\star$ is given by

$$8\|\mathbf{X} - \mathbf{X}_\star\|_{\mathsf{F}}^2 = \left\|\mathbf{F}\mathbf{F}^\top - \bar{\mathbf{F}}\bar{\mathbf{F}}^\top - \mathbf{F}_\star\mathbf{F}_\star^\top + \bar{\mathbf{F}}_\star\bar{\mathbf{F}}_\star^\top\right\|_{\mathsf{F}}^2$$

$$
\begin{aligned}
&= \left\|\boldsymbol{F}\boldsymbol{F}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top\right\|_{\mathsf{F}}^2 + \left\|\bar{\boldsymbol{F}}\bar{\boldsymbol{F}}^\top - \bar{\boldsymbol{F}}_\star\bar{\boldsymbol{F}}_\star^\top\right\|_{\mathsf{F}}^2 - 2\,\mathrm{tr}\left((\boldsymbol{F}\boldsymbol{F}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top)(\bar{\boldsymbol{F}}\bar{\boldsymbol{F}}^\top - \bar{\boldsymbol{F}}_\star\bar{\boldsymbol{F}}_\star^\top)\right) \\
&= 2\left\|\boldsymbol{F}\boldsymbol{F}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top\right\|_{\mathsf{F}}^2 + 2\|\boldsymbol{F}^\top\bar{\boldsymbol{F}}_\star\|_{\mathsf{F}}^2 + 2\|\boldsymbol{F}_\star^\top\bar{\boldsymbol{F}}\|_{\mathsf{F}}^2 \\
&\geq 2\left\|\boldsymbol{F}\boldsymbol{F}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top\right\|_{\mathsf{F}}^2,
\end{aligned}
$$

where we use the facts that $\left\|\boldsymbol{F}\boldsymbol{F}^\top - \boldsymbol{F}_\star\boldsymbol{F}_\star^\top\right\|_{\mathsf{F}}^2 = \left\|\bar{\boldsymbol{F}}\bar{\boldsymbol{F}}^\top - \bar{\boldsymbol{F}}_\star\bar{\boldsymbol{F}}_\star^\top\right\|_{\mathsf{F}}^2$ and $\boldsymbol{F}^\top\bar{\boldsymbol{F}} = \boldsymbol{F}_\star^\top\bar{\boldsymbol{F}}_\star = \boldsymbol{0}$.

Let $\boldsymbol{O} \coloneqq \mathrm{sgn}(\boldsymbol{F}^\top\boldsymbol{F}_\star)$[5] be the optimal orthonormal alignment matrix between $\boldsymbol{F}$ and $\boldsymbol{F}_\star$. Denote $\boldsymbol{\Delta} \coloneqq \boldsymbol{F}\boldsymbol{O} - \boldsymbol{F}_\star$. Follow the same argument as [TBS+16, Lemma 5.14] and [GJZ17, Lemma 41] to obtain

$$
\begin{aligned}
4\|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}}^2 &\geq \left\|\boldsymbol{F}_\star\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{F}_\star^\top + \boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right\|_{\mathsf{F}}^2 \\
&= \mathrm{tr}\left(2\boldsymbol{F}_\star^\top\boldsymbol{F}_\star\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + (\boldsymbol{\Delta}^\top\boldsymbol{\Delta})^2 + 2(\boldsymbol{F}_\star^\top\boldsymbol{\Delta})^2 + 4\boldsymbol{F}_\star^\top\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\right) \\
&= \mathrm{tr}\left(2\boldsymbol{F}_\star^\top\boldsymbol{F}_\star\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + (\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + \sqrt{2}\boldsymbol{F}_\star^\top\boldsymbol{\Delta})^2 + (4-2\sqrt{2})\boldsymbol{F}_\star^\top\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\right) \\
&= \mathrm{tr}\left(2(\sqrt{2}-1)\boldsymbol{F}_\star^\top\boldsymbol{F}_\star\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + (\boldsymbol{\Delta}^\top\boldsymbol{\Delta} + \sqrt{2}\boldsymbol{F}_\star^\top\boldsymbol{\Delta})^2 + (4-2\sqrt{2})\boldsymbol{F}_\star^\top\boldsymbol{F}\boldsymbol{O}\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\right) \\
&\geq \mathrm{tr}\left(4(\sqrt{2}-1)\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}^\top\boldsymbol{\Delta}\right) = 4(\sqrt{2}-1)\left\|(\boldsymbol{F}\boldsymbol{O} - \boldsymbol{F}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2,
\end{aligned}
$$

where the last inequality follows from the facts that $\boldsymbol{F}_\star^\top\boldsymbol{F}_\star = 2\boldsymbol{\Sigma}_\star$ and that $\boldsymbol{F}_\star^\top\boldsymbol{F}\boldsymbol{O}$ is positive semi-definite. Therefore we obtain

$$
\left\|(\boldsymbol{F}\boldsymbol{O} - \boldsymbol{F}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}} \leq \left(\sqrt{2}+1\right)^{1/2}\|\boldsymbol{X} - \boldsymbol{X}_\star\|_{\mathsf{F}}.
$$

This in conjunction with $\mathrm{dist}(\boldsymbol{F}, \boldsymbol{F}_\star) \leq \|(\boldsymbol{F}\boldsymbol{O} - \boldsymbol{F}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}$ yields the claimed result. □

A.2 Matrix perturbation bounds

Lemma 12. *For any $\boldsymbol{L} \in \mathbb{R}^{n_1\times r}, \boldsymbol{R} \in \mathbb{R}^{n_2\times r}$, denote $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$ and $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$. Suppose that $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} < 1$, then one has*

$$
\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \leq \frac{1}{1 - \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}}; \tag{40a}
$$

$$
\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \leq \frac{1}{1 - \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}}; \tag{40b}
$$

$$
\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{U}_\star\right\|_{\mathsf{op}} \leq \frac{\sqrt{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}}{1 - \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}}; \tag{40c}
$$

$$
\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{V}_\star\right\|_{\mathsf{op}} \leq \frac{\sqrt{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}}{1 - \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}}. \tag{40d}
$$

Proof. We only prove claims (40a) and (40c) on the factor $\boldsymbol{L}$, while the claims on the factor $\boldsymbol{R}$ follow from a similar argument. We start to prove (40a). Notice that

$$
\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} = \frac{1}{\sigma_r(\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2})}.
$$

In addition, invoke Weyl's inequality to obtain

$$
\sigma_r(\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2}) \geq \sigma_r(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}) - \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} = 1 - \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}},
$$

where we have used the fact that $\boldsymbol{U}_\star = \boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}$ satisfies $\sigma_r(\boldsymbol{U}_\star) = 1$. Combine the preceding two relations to prove (40a).

[5] Let $\boldsymbol{A}\boldsymbol{S}\boldsymbol{B}^\top$ be the SVD of $\boldsymbol{F}^\top\boldsymbol{F}_\star$, then the matrix sign is $\mathrm{sgn}(\boldsymbol{F}^\top\boldsymbol{F}_\star) \coloneqq \boldsymbol{A}\boldsymbol{B}^\top$.

We proceed to prove (40c). Combine $\boldsymbol{L}_\star^\top \boldsymbol{U}_\star = \boldsymbol{\Sigma}_\star^{1/2}$ and $(\boldsymbol{I}_{n_1} - \boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top)\boldsymbol{L} = \boldsymbol{0}$ to obtain the decomposition

$$\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{U}_\star = -\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Delta}_L^\top\boldsymbol{U}_\star + (\boldsymbol{I}_{n_1} - \boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}.$$

The fact that $\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Delta}_L^\top\boldsymbol{U}_\star$ and $(\boldsymbol{I}_{n_1} - \boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}$ are orthogonal implies

$$\begin{aligned}
\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{U}_\star\right\|_{\mathsf{op}}^2 &\le \left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Delta}_L^\top\boldsymbol{U}_\star\right\|_{\mathsf{op}}^2 + \left\|(\boldsymbol{I}_{n_1} - \boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\right\|_{\mathsf{op}}^2 \\
&\le \left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}^2 \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}^2 + \left\|\boldsymbol{I}_{n_1} - \boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top\right\|_{\mathsf{op}}^2 \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}^2 \\
&\le \frac{\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}^2}{(1 - \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}})^2} + \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}^2 \\
&\le \frac{2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}^2}{(1 - \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}})^2},
\end{aligned}$$

where we have used (40a) and the fact that $\|\boldsymbol{I}_{n_1} - \boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top\|_{\mathsf{op}} \le 1$ in the third line. □

Lemma 13. *For any $\boldsymbol{L} \in \mathbb{R}^{n_1\times r}, \boldsymbol{R} \in \mathbb{R}^{n_2\times r}$, denote $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$ and $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, then one has*

$$\begin{aligned}
\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} &\le \|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} \\
&\le \left(1 + \frac{1}{2}(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}})\right)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).
\end{aligned}$$

Proof. In light of the decomposition $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top + \boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top$ and the triangle inequality, one has

$$\begin{aligned}
\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} &\le \|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} \\
&= \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}},
\end{aligned}$$

where we have used the facts that

$$\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} = \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{V}_\star^\top\|_{\mathsf{F}} = \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}, \quad \text{and} \quad \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} = \|\boldsymbol{U}_\star\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} = \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.$$

This together with the simple upper bound

$$\begin{aligned}
\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} &= \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2})^\top\|_{\mathsf{F}} + \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2})^\top\|_{\mathsf{F}} \\
&\le \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} + \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \\
&\le \frac{1}{2}(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}})\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)
\end{aligned}$$

finishes the proof. □

Lemma 14. *For any $\boldsymbol{L} \in \mathbb{R}^{n_1\times r}, \boldsymbol{R} \in \mathbb{R}^{n_2\times r}$ and any invertible matrices $\boldsymbol{Q}, \bar{\boldsymbol{Q}} \in \mathrm{GL}(r)$, suppose that $\|(\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \vee \|(\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} < 1$, then one has*

$$\begin{aligned}
\left\|\boldsymbol{\Sigma}_\star^{1/2}\bar{\boldsymbol{Q}}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{\Sigma}_\star\right\|_{\mathsf{op}} &\le \frac{\|\boldsymbol{R}(\bar{\boldsymbol{Q}}^{-\top} - \boldsymbol{Q}^{-\top})\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}}}{1 - \|(\boldsymbol{R}\boldsymbol{Q}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}}; \\
\left\|\boldsymbol{\Sigma}_\star^{1/2}\bar{\boldsymbol{Q}}^{\top}\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{\Sigma}_\star\right\|_{\mathsf{op}} &\le \frac{\|\boldsymbol{L}(\bar{\boldsymbol{Q}} - \boldsymbol{Q})\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}}}{1 - \|(\boldsymbol{L}\boldsymbol{Q} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}}.
\end{aligned}$$

Proof. Insert $\boldsymbol{R}^\top \boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1}$, and use the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathsf{op}} \le \|\boldsymbol{A}\|_{\mathsf{op}}\|\boldsymbol{B}\|_{\mathsf{op}}$ to obtain

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_\star^{1/2}\bar{\boldsymbol{Q}}^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{\Sigma}_\star\right\|_{\mathsf{op}} &= \left\|\boldsymbol{\Sigma}_\star^{1/2}(\bar{\boldsymbol{Q}}^{-1}-\boldsymbol{Q}^{-1})\boldsymbol{R}^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \\ &\le \left\|\boldsymbol{R}(\bar{\boldsymbol{Q}}^{-\top}-\boldsymbol{Q}^{-\top})\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \\ &= \left\|\boldsymbol{R}(\bar{\boldsymbol{Q}}^{-\top}-\boldsymbol{Q}^{-\top})\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}\left\|\boldsymbol{R}\boldsymbol{Q}^{-\top}((\boldsymbol{R}\boldsymbol{Q}^{-\top})^\top\boldsymbol{R}\boldsymbol{Q}^{-\top})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \\ &\le \frac{\|\boldsymbol{R}(\bar{\boldsymbol{Q}}^{-\top}-\boldsymbol{Q}^{-\top})\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}}}{1-\|(\boldsymbol{R}\boldsymbol{Q}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}},\end{aligned}$$

where the last line uses Lemma 12.

Similarly, insert $\boldsymbol{L}^\top \boldsymbol{L}(\boldsymbol{L}^\top \boldsymbol{L})^{-1}$, and use the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathsf{op}} \le \|\boldsymbol{A}\|_{\mathsf{op}}\|\boldsymbol{B}\|_{\mathsf{op}}$ to obtain

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_\star^{1/2}\bar{\boldsymbol{Q}}^{\top}\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{\Sigma}_\star\right\|_{\mathsf{op}} &= \left\|\boldsymbol{\Sigma}_\star^{1/2}(\bar{\boldsymbol{Q}}^{\top}-\boldsymbol{Q}^{\top})\boldsymbol{L}^\top\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \\ &\le \left\|\boldsymbol{L}(\bar{\boldsymbol{Q}}-\boldsymbol{Q})\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}\left\|\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{Q}^{-\top}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \\ &= \left\|\boldsymbol{L}(\bar{\boldsymbol{Q}}-\boldsymbol{Q})\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}\left\|\boldsymbol{L}\boldsymbol{Q}((\boldsymbol{L}\boldsymbol{Q})^\top\boldsymbol{L}\boldsymbol{Q})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \\ &\le \frac{\|\boldsymbol{L}(\bar{\boldsymbol{Q}}-\boldsymbol{Q})\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}}}{1-\|(\boldsymbol{L}\boldsymbol{Q}-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}},\end{aligned}$$

where the last line uses Lemma 12. □

A.3 Partial Frobenius norm

We introduce the partial Frobenius norm

$$\|\boldsymbol{X}\|_{\mathsf{F},r} \coloneqq \sqrt{\sum_{i=1}^{r}\sigma_i^2(\boldsymbol{X})} = \|\mathcal{P}_r(\boldsymbol{X})\|_{\mathsf{F}} \tag{41}$$

as the ℓ_2 norm of the vector composed of the top-r singular values of the matrix $\boldsymbol{X}$, or equivalently as the Frobenius norm of the rank-r approximation $\mathcal{P}_r(\boldsymbol{X})$ defined in (5). It is straightforward to verify that $\|\cdot\|_{\mathsf{F},r}$ is a norm; see also [Maz16]. The following lemma provides several equivalent and useful characterizations of this partial Frobenius norm.

Lemma 15. *For any $\boldsymbol{X} \in \mathbb{R}^{n_1\times n_2}$, one has*

$$\begin{aligned}\|\boldsymbol{X}\|_{\mathsf{F},r} &= \max_{\widetilde{\boldsymbol{V}}\in\mathbb{R}^{n_2\times r}:\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}}=\boldsymbol{I}_r}\|\boldsymbol{X}\widetilde{\boldsymbol{V}}\|_{\mathsf{F}} &\text{(42a)}\\ &= \max_{\widetilde{\boldsymbol{X}}\in\mathbb{R}^{n_1\times n_2}:\|\widetilde{\boldsymbol{X}}\|_{\mathsf{F}}\le 1,\operatorname{rank}(\widetilde{\boldsymbol{X}})\le r}|\langle\boldsymbol{X},\widetilde{\boldsymbol{X}}\rangle| &\text{(42b)}\\ &= \max_{\widetilde{\boldsymbol{R}}\in\mathbb{R}^{n_2\times r}:\|\widetilde{\boldsymbol{R}}\|_{\mathsf{op}}\le 1}\|\boldsymbol{X}\widetilde{\boldsymbol{R}}\|_{\mathsf{F}}. &\text{(42c)}\end{aligned}$$

Proof. The first representation (42a) follows immediately from the extremal partial trace identity; see [Maz16, Proposition 4.4], by noticing the following relation

$$\sum_{i=1}^{r}\sigma_i^2(\boldsymbol{X}) = \max_{\mathbb{V}\subseteq\mathbb{R}^{n_2}:\dim(\mathbb{V})=r}\operatorname{tr}\left(\boldsymbol{X}^\top\boldsymbol{X}\mid\mathbb{V}\right) = \max_{\widetilde{\boldsymbol{V}}\in\mathbb{R}^{n_2\times r}:\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}}=\boldsymbol{I}_r}\|\boldsymbol{X}\widetilde{\boldsymbol{V}}\|_{\mathsf{F}}^2.$$

Here the partial trace over a vector space $\mathbb{V}$ is defined as

$$\operatorname{tr}(\boldsymbol{X}^\top\boldsymbol{X}\mid\mathbb{V}) \coloneqq \sum_{i=1}^{r}\widetilde{\boldsymbol{v}}_i^\top\boldsymbol{X}^\top\boldsymbol{X}\widetilde{\boldsymbol{v}}_i,$$

where $\{\widetilde{\boldsymbol{v}}_i\}_{1\le i\le r}$ is any orthonormal basis of $\mathbb{V}$. The partial trace is invariant to the choice of orthonormal basis and therefore well-defined.

To prove the second representation (42b), for any $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n_1\times n_2}$ obeying $\mathrm{rank}(\widetilde{\boldsymbol{X}}) \le r$ and $\|\widetilde{\boldsymbol{X}}\|_{\mathsf{F}} \le 1$, denoting $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^\top$ as its compact SVD, one has

$$|\langle \boldsymbol{X}, \widetilde{\boldsymbol{X}}\rangle| = |\langle \boldsymbol{X}, \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^\top\rangle| = |\langle \boldsymbol{X}\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\rangle| \le \|\boldsymbol{X}\widetilde{\boldsymbol{V}}\|_{\mathsf{F}}\|\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\|_{\mathsf{F}} \le \|\boldsymbol{X}\|_{\mathsf{F},r},$$

where the last inequality follows from (42a). In addition, the maximum in (42b) is attained at $\widetilde{\boldsymbol{X}} = \mathcal{P}_r(\boldsymbol{X})/\|\mathcal{P}_r(\boldsymbol{X})\|_{\mathsf{F}}$.

To prove the third representation (42c), for any $\widetilde{\boldsymbol{R}} \in \mathbb{R}^{n_2\times r}$ obeying $\|\widetilde{\boldsymbol{R}}\|_{\mathsf{op}} \le 1$, combine the variational representation of the Frobenius norm and (42b) to obtain

$$\begin{aligned}\|\boldsymbol{X}\widetilde{\boldsymbol{R}}\|_{\mathsf{F}} &= \max_{\widetilde{\boldsymbol{L}}\in\mathbb{R}^{n_1\times n_2}:\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}}\le 1} |\langle \boldsymbol{X}\widetilde{\boldsymbol{R}}, \widetilde{\boldsymbol{L}}\rangle| \\ &= \max_{\widetilde{\boldsymbol{L}}\in\mathbb{R}^{n_1\times n_2}:\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}}\le 1} |\langle \boldsymbol{X}, \widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{R}}^\top\rangle| \le \|\boldsymbol{X}\|_{\mathsf{F},r},\end{aligned}$$

where the last inequality follows from (42b). In addition, the maximum in (42c) is attained at $\widetilde{\boldsymbol{R}} = \boldsymbol{V}$, where $\boldsymbol{V}$ denotes the top-r right singular vectors of $\boldsymbol{X}$. □

Remark 3. For self-completeness, we also provide a detailed proof of the first representation (42a). This proof is inductive on r. When $r = 1$, we have

$$\sigma_1(\boldsymbol{X}) = \|\boldsymbol{X}\boldsymbol{v}_1\|_2 = \max_{\widetilde{\boldsymbol{v}}\in\mathbb{R}^{n_2}:\|\widetilde{\boldsymbol{v}}\|_2=1} \|\boldsymbol{X}\widetilde{\boldsymbol{v}}\|_2,$$

where $\boldsymbol{v}_1$ denotes the top right singular vector of $\boldsymbol{X}$. Assume that the statement holds for $\|\cdot\|_{\mathsf{F},r-1}$. Now consider $\|\cdot\|_{\mathsf{F},r}$. For any $\widetilde{\boldsymbol{V}} \in \mathbb{R}^{n_2\times r}$ such that $\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}} = \boldsymbol{I}_r$, we can first pick $\widetilde{\boldsymbol{v}}_2, \dots, \widetilde{\boldsymbol{v}}_r$ as a set of orthonormal vectors in the column space of $\widetilde{\boldsymbol{V}}$ that are orthogonal to $\boldsymbol{v}_1$, and then pick $\widetilde{\boldsymbol{v}}_1$ via the Gram-Schmidt process, so that $\{\widetilde{\boldsymbol{v}}_i\}_{i=1}^r$ provides an orthonormal basis of the column space of $\widetilde{\boldsymbol{V}}$. Further, by the orthogonality of $\widetilde{\boldsymbol{V}}$, there exists an orthonormal matrix $\boldsymbol{O}$ such that

$$\widetilde{\boldsymbol{V}} = [\widetilde{\boldsymbol{v}}_1, \dots, \widetilde{\boldsymbol{v}}_r]\boldsymbol{O}.$$

Combining this formula with the induction hypothesis yields

$$\begin{aligned}\|\boldsymbol{X}\widetilde{\boldsymbol{V}}\|_{\mathsf{F}}^2 &= \|\boldsymbol{X}[\widetilde{\boldsymbol{v}}_1, \dots, \widetilde{\boldsymbol{v}}_r]\|_{\mathsf{F}}^2 \\ &= \|\boldsymbol{X}\widetilde{\boldsymbol{v}}_1\|_2^2 + \|\boldsymbol{X}[\widetilde{\boldsymbol{v}}_2, \dots, \widetilde{\boldsymbol{v}}_r]\|_{\mathsf{F}}^2 \\ &= \|\boldsymbol{X}\widetilde{\boldsymbol{v}}_1\|_2^2 + \|(\boldsymbol{X} - \mathcal{P}_1(\boldsymbol{X}))[\widetilde{\boldsymbol{v}}_2, \dots, \widetilde{\boldsymbol{v}}_r]\|_{\mathsf{F}}^2 \\ &\le \sigma_1^2(\boldsymbol{X}) + \|\boldsymbol{X} - \mathcal{P}_1(\boldsymbol{X})\|_{\mathsf{F},r-1}^2 \\ &= \sum_{i=1}^r \sigma_i^2(\boldsymbol{X}) = \|\boldsymbol{X}\|_{\mathsf{F},r}^2,\end{aligned}$$

where the first line holds since $\boldsymbol{O}$ is orthonormal, the third line holds since $\mathcal{P}_1(\boldsymbol{X})[\widetilde{\boldsymbol{v}}_2, \dots, \widetilde{\boldsymbol{v}}_r] = \boldsymbol{0}$, the fourth line follows from the induction hypothesis, and the last line follows from the definition (41). In addition, the maximum in (42a) is attained at $\widetilde{\boldsymbol{V}} = \boldsymbol{V}$, where $\boldsymbol{V}$ denotes the top-r right singular vectors of $\boldsymbol{X}$. This finishes the proof.

Recall that $\mathcal{P}_r(\boldsymbol{X})$ denotes the best rank-r approximation of $\boldsymbol{X}$ under the Frobenius norm. It turns out that $\mathcal{P}_r(\boldsymbol{X})$ is also the best rank-r approximation of $\boldsymbol{X}$ under the partial Frobenius norm $\|\cdot\|_{\mathsf{F},r}$. This claim is formally stated below; see also [Maz16, Theorem 4.21].

Lemma 16. *Fix any $\boldsymbol{X} \in \mathbb{R}^{n_1\times n_2}$ and recall the definition of $\mathcal{P}_r(\boldsymbol{X})$ in (5). One has*

$$\mathcal{P}_r(\boldsymbol{X}) = \operatorname*{argmin}_{\widetilde{\boldsymbol{X}}\in\mathbb{R}^{n_1\times n_2}:\mathrm{rank}(\widetilde{\boldsymbol{X}})\le r} \|\boldsymbol{X} - \widetilde{\boldsymbol{X}}\|_{\mathsf{F},r}.$$

Proof. For any $\widetilde{\boldsymbol{X}}$ of rank at most r, invoke Weyl's inequality to obtain $\sigma_{r+i}(\boldsymbol{X}) \le \sigma_i(\boldsymbol{X}-\widetilde{\boldsymbol{X}})+\sigma_{r+1}(\widetilde{\boldsymbol{X}}) = \sigma_i(\boldsymbol{X}-\widetilde{\boldsymbol{X}})$, for $i=1,\dots,r$. Thus one has

$$\|\boldsymbol{X}-\mathcal{P}_r(\boldsymbol{X})\|_{\mathsf{F},r}^2 = \sum_{i=1}^{r}\sigma_{r+i}^2(\boldsymbol{X}) \le \sum_{i=1}^{r}\sigma_i^2(\boldsymbol{X}-\widetilde{\boldsymbol{X}}) = \|\boldsymbol{X}-\widetilde{\boldsymbol{X}}\|_{\mathsf{F},r}^2.$$

The proof is finished by observing that the rank of $\mathcal{P}_r(\boldsymbol{X})$ is at most r. $\square$

B Proof for Low-Rank Matrix Factorization

B.1 Proof of Proposition 2

The gradients of $\mathcal{L}(\boldsymbol{F})$ in (29) with respect to $\boldsymbol{L}$ and $\boldsymbol{R}$ are given as

$$\nabla_{\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}) = (\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}, \quad \nabla_{\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}) = (\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)^\top\boldsymbol{L},$$

which can be used to compute the Hessian with respect to $\boldsymbol{L}$ and $\boldsymbol{R}$. Writing for the vectorized variables, the Hessians are given as

$$\nabla^2_{\boldsymbol{L},\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}) = (\boldsymbol{R}^\top\boldsymbol{R})\otimes\boldsymbol{I}_{n_1}, \quad \nabla^2_{\boldsymbol{R},\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}) = (\boldsymbol{L}^\top\boldsymbol{L})\otimes\boldsymbol{I}_{n_2}.$$

Viewed in the vectorized form, the `ScaledGD` update in (3) can be rewritten as

$$\begin{aligned}
\operatorname{vec}(\boldsymbol{L}_{t+1}) &= \operatorname{vec}(\boldsymbol{L}_t) - \eta((\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}\otimes\boldsymbol{I}_{n_1})\operatorname{vec}((\boldsymbol{L}_t\boldsymbol{R}_t^\top-\boldsymbol{X}_\star)\boldsymbol{R}_t)\\
&= \operatorname{vec}(\boldsymbol{L}_t) - \eta(\nabla^2_{\boldsymbol{L},\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}_t))^{-1}\operatorname{vec}(\nabla_{\boldsymbol{L}}\mathcal{L}(\boldsymbol{F}_t)),\\
\operatorname{vec}(\boldsymbol{R}_{t+1}) &= \operatorname{vec}(\boldsymbol{R}_t) - \eta((\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}\otimes\boldsymbol{I}_{n_2})\operatorname{vec}((\boldsymbol{L}_t\boldsymbol{R}_t^\top-\boldsymbol{X}_\star)^\top\boldsymbol{L}_t)\\
&= \operatorname{vec}(\boldsymbol{R}_t) - \eta(\nabla^2_{\boldsymbol{R},\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}_t))^{-1}\operatorname{vec}(\nabla_{\boldsymbol{R}}\mathcal{L}(\boldsymbol{F}_t)).
\end{aligned}$$

B.2 Proof of Theorem 5

The proof is inductive in nature. More specifically, we intend to show that for all $t \ge 0$,

1. $\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star) \le (1-0.7\eta)^t\operatorname{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star) \le 0.1(1-0.7\eta)^t\sigma_r(\boldsymbol{X}_\star)$, and
2. the optimal alignment matrix $\boldsymbol{Q}_t$ between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists.

For the base case, i.e. $t=0$, the first induction hypothesis trivially holds, while the second also holds true in view of Lemma 9 and the assumption that $\operatorname{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star) \le 0.1\sigma_r(\boldsymbol{X}_\star)$. We therefore concentrate on the induction step. Suppose that the t-th iterate $\boldsymbol{F}_t$ obeys the aforementioned induction hypotheses. Our goal is to show that $\boldsymbol{F}_{t+1}$ continues to satisfy those.

For notational convenience, denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_t\boldsymbol{Q}_t$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_t\boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L}-\boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R}-\boldsymbol{R}_\star$, and $\epsilon \coloneqq 0.1$. By the definition of $\operatorname{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)$, one has

$$\operatorname{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star) \le \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2, \tag{43}$$

where we recall that $\boldsymbol{Q}_t$ is the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$. Utilize the `ScaledGD` update rule (30) and the decomposition $\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$ to obtain

$$\begin{aligned}
(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} &= \left(\boldsymbol{L}-\eta(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\\
&= \left(\boldsymbol{\Delta}_L-\eta(\boldsymbol{\Delta}_L\boldsymbol{R}^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\right)\boldsymbol{\Sigma}_\star^{1/2}\\
&= (1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.
\end{aligned}$$

As a result, one can expand the first square in (43) as

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 = (1-\eta)^2\operatorname{tr}\left(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right) - 2\eta(1-\eta)\underbrace{\operatorname{tr}\left(\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)}_{\mathfrak{M}_1}$$

$$+\eta^2 \underbrace{\left\| \boldsymbol{L}_\star \boldsymbol{\Delta}_R^\top \boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2}_{\mathfrak{M}_2}. \tag{44}$$

The first term $\mathrm{tr}(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top)$ is closely related to $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$, and hence our focus will be on relating $\mathfrak{M}_1$ and $\mathfrak{M}_2$ to $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$. We start with the term $\mathfrak{M}_1$. Since $\boldsymbol{L}$ and $\boldsymbol{R}$ are aligned with $\boldsymbol{L}_\star$ and $\boldsymbol{R}_\star$, Lemma 10 tells that $\boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \boldsymbol{L} = \boldsymbol{R}^\top \boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star$. This together with $\boldsymbol{L}_\star = \boldsymbol{L} - \boldsymbol{\Delta}_L$ allows us to rewrite $\mathfrak{M}_1$ as

$$\begin{aligned}
\mathfrak{M}_1 &= \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \boldsymbol{L}_\star \boldsymbol{\Delta}_R^\top\right) \\
&= \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \boldsymbol{L} \boldsymbol{\Delta}_R^\top\right) - \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \boldsymbol{\Delta}_L \boldsymbol{\Delta}_R^\top\right) \\
&= \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{R}^\top \boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_R^\top\right) - \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \boldsymbol{\Delta}_L \boldsymbol{\Delta}_R^\top\right).
\end{aligned}$$

Moving on to $\mathfrak{M}_2$, we can utilize the fact $\boldsymbol{L}_\star^\top \boldsymbol{L}_\star = \boldsymbol{\Sigma}_\star$ and the decomposition $\boldsymbol{\Sigma}_\star = \boldsymbol{R}^\top \boldsymbol{R} - (\boldsymbol{R}^\top \boldsymbol{R} - \boldsymbol{\Sigma}_\star)$ to obtain

$$\begin{aligned}
\mathfrak{M}_2 &= \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star (\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{R}^\top \boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_R^\top\right) \\
&= \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{R}^\top \boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_R^\top\right) - \mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} (\boldsymbol{R}^\top \boldsymbol{R} - \boldsymbol{\Sigma}_\star)(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{R}^\top \boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_R^\top\right).
\end{aligned}$$

Putting $\mathfrak{M}_1$ and $\mathfrak{M}_2$ back to (44) yields

$$\begin{aligned}
\left\| (\boldsymbol{L}_{t+1} \boldsymbol{Q}_t - \boldsymbol{L}_\star) \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{F}}^2 &= (1-\eta)^2 \,\mathrm{tr}\left(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top\right) - \eta(2-3\eta) \underbrace{\mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{R}^\top \boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_R^\top\right)}_{\mathfrak{F}_1} \\
&\quad + 2\eta(1-\eta) \underbrace{\mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \boldsymbol{\Delta}_L \boldsymbol{\Delta}_R^\top\right)}_{\mathfrak{F}_2} \\
&\quad - \eta^2 \underbrace{\mathrm{tr}\left(\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} (\boldsymbol{R}^\top \boldsymbol{R} - \boldsymbol{\Sigma}_\star)(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{R}^\top \boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_R^\top\right)}_{\mathfrak{F}_3}.
\end{aligned}$$

In what follows, we will control the three terms $\mathfrak{F}_1, \mathfrak{F}_2$ and $\mathfrak{F}_3$ separately.

1. Notice that $\mathfrak{F}_1$ is the inner product of two positive semi-definite matrices $\boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{R}^\top$ and $\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_R^\top$. Consequently we have $\mathfrak{F}_1 \geq 0$.

2. To control $\mathfrak{F}_2$, we need certain control on $\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}$ and $\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}$. The first induction hypothesis

$$\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) = \sqrt{\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{-1/2} \boldsymbol{\Sigma}_\star\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{-1/2} \boldsymbol{\Sigma}_\star\|_{\mathsf{F}}^2} \leq \epsilon \sigma_r(\boldsymbol{X}_\star)$$

together with the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathsf{F}} \geq \|\boldsymbol{A}\|_{\mathsf{F}} \sigma_r(\boldsymbol{B})$ tells that

$$\sqrt{\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{F}}^2}\, \sigma_r(\boldsymbol{X}_\star) \leq \epsilon \sigma_r(\boldsymbol{X}_\star).$$

In light of the relation $\|\boldsymbol{A}\|_{\mathsf{op}} \leq \|\boldsymbol{A}\|_{\mathsf{F}}$, this further implies

$$\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \vee \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \leq \epsilon. \tag{45}$$

Invoke Lemma 12 to see

$$\left\| \boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{op}} \leq \frac{1}{1-\epsilon}.$$

With these consequences, one can bound $|\mathfrak{F}_2|$ by

$$\begin{aligned}
|\mathfrak{F}_2| &= \left| \mathrm{tr}\left(\boldsymbol{\Sigma}_\star^{-1/2} \boldsymbol{\Delta}_R^\top \boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top \boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\right) \right| \\
&\leq \left\| \boldsymbol{\Sigma}_\star^{-1/2} \boldsymbol{\Delta}_R^\top \boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{op}} \mathrm{tr}\left(\boldsymbol{\Sigma}_\star^{1/2} \boldsymbol{\Delta}_L^\top \boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\right) \\
&\leq \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \left\| \boldsymbol{R}(\boldsymbol{R}^\top \boldsymbol{R})^{-1} \boldsymbol{\Sigma}_\star^{1/2} \right\|_{\mathsf{op}} \mathrm{tr}\left(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top\right) \\
&\leq \frac{\epsilon}{1-\epsilon} \mathrm{tr}\left(\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star \boldsymbol{\Delta}_L^\top\right).
\end{aligned}$$

3. Similarly, one can bound $|\mathfrak{F}_3|$ by

$$\begin{aligned}|\mathfrak{F}_3| &\le \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}(\boldsymbol{R}^\top\boldsymbol{R}-\boldsymbol{\Sigma}_\star)(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\|_{\mathsf{op}}\operatorname{tr}\left(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right)\\ &\le \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}^2\left\|\boldsymbol{\Sigma}_\star^{-1/2}(\boldsymbol{R}^\top\boldsymbol{R}-\boldsymbol{\Sigma}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\|_{\mathsf{op}}\operatorname{tr}\left(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right)\\ &\le \frac{1}{(1-\epsilon)^2}\left\|\boldsymbol{\Sigma}_\star^{-1/2}(\boldsymbol{R}^\top\boldsymbol{R}-\boldsymbol{\Sigma}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\|_{\mathsf{op}}\operatorname{tr}\left(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right).\end{aligned}$$

Further notice that

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_\star^{-1/2}(\boldsymbol{R}^\top\boldsymbol{R}-\boldsymbol{\Sigma}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\|_{\mathsf{op}} &= \left\|\boldsymbol{\Sigma}_\star^{-1/2}(\boldsymbol{R}_\star^\top\boldsymbol{\Delta}_R+\boldsymbol{\Delta}_R^\top\boldsymbol{R}_\star+\boldsymbol{\Delta}_R^\top\boldsymbol{\Delta}_R)\boldsymbol{\Sigma}_\star^{-1/2}\right\|_{\mathsf{op}}\\ &\le 2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}^2\\ &\le 2\epsilon+\epsilon^2.\end{aligned}$$

Take the preceding two bounds together to arrive at

$$|\mathfrak{F}_3| \le \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\operatorname{tr}\left(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right).$$

Combining the bounds for $\mathfrak{F}_1,\mathfrak{F}_2,\mathfrak{F}_3$, one has

$$\begin{aligned}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &= \left\|(1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}-\eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2\\ &\le \left((1-\eta)^2+\frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\operatorname{tr}\left(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\operatorname{tr}\left(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top\right).\end{aligned}\tag{46}$$

A similarly bound holds for the second square $\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$ in (43). Therefore we obtain

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \le \rho^2(\eta;\epsilon)\operatorname{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star),$$

where we identify

$$\operatorname{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star)=\operatorname{tr}(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top)+\operatorname{tr}(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_R^\top)\tag{47}$$

and the contraction rate $\rho^2(\eta;\epsilon)$ is given by

$$\rho^2(\eta;\epsilon) \coloneqq (1-\eta)^2+\frac{2\epsilon}{1-\epsilon}\eta(1-\eta)+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2.$$

With $\epsilon=0.1$ and $0<\eta\le 2/3$, one has $\rho(\eta;\epsilon)\le 1-0.7\eta$. Thus we conclude that

$$\begin{aligned}\operatorname{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star) &\le \sqrt{\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}\\ &\le (1-0.7\eta)\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\\ &\le (1-0.7\eta)^{t+1}\operatorname{dist}(\boldsymbol{F}_0,\boldsymbol{F}_\star)\le(1-0.7\eta)^{t+1}0.1\sigma_r(\boldsymbol{X}_\star).\end{aligned}$$

This proves the first induction hypothesis. The existence of the optimal alignment matrix $\boldsymbol{Q}_{t+1}$ between $\boldsymbol{F}_{t+1}$ and $\boldsymbol{F}_\star$ is assured by Lemma 9, which finishes the proof for the second hypothesis.

So far, we have demonstrated the first conclusion in the theorem. The second conclusion is an easy consequence of Lemma 13 as

$$\begin{aligned}\left\|\boldsymbol{L}_t\boldsymbol{R}_t^\top-\boldsymbol{X}_\star\right\|_{\mathsf{F}} &\le \left(1+\frac{\epsilon}{2}\right)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\\ &\le \left(1+\frac{\epsilon}{2}\right)\sqrt{2}\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\\ &\le 1.5\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star).\end{aligned}\tag{48}$$

Here, the second line follows from the elementary inequality $a+b\le\sqrt{2(a^2+b^2)}$ and the expression of $\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)$ in (47). The proof is now completed.

C Proof for Low-Rank Matrix Sensing

We start by recording a useful lemma.

Lemma 17 ([CP11]). *Suppose that $\mathcal{A}(\cdot)$ obeys the $2r$-RIP with a constant δ_{2r}. Then for any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{n_1\times n_2}$ of rank at most r, one has*

$$|\langle \mathcal{A}(\boldsymbol{X}_1), \mathcal{A}(\boldsymbol{X}_2)\rangle - \langle \boldsymbol{X}_1, \boldsymbol{X}_2\rangle| \le \delta_{2r}\|\boldsymbol{X}_1\|_{\mathsf{F}}\|\boldsymbol{X}_2\|_{\mathsf{F}},$$

which can be stated equivalently as

$$\left|\operatorname{tr}\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_1)\boldsymbol{X}_2^\top\right)\right| \le \delta_{2r}\|\boldsymbol{X}_1\|_{\mathsf{F}}\|\boldsymbol{X}_2\|_{\mathsf{F}}. \tag{49}$$

As a simple corollary, one has that for any matrix $\boldsymbol{R}\in\mathbb{R}^{n_2\times r}$:

$$\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_1)\boldsymbol{R}\|_{\mathsf{F}} \le \delta_{2r}\|\boldsymbol{X}_1\|_{\mathsf{F}}\|\boldsymbol{R}\|_{\mathsf{op}}. \tag{50}$$

This is due to the fact that

$$\begin{aligned}
\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_1)\boldsymbol{R}\|_{\mathsf{F}} &= \max_{\widetilde{\boldsymbol{L}}:\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}}\le 1} \operatorname{tr}\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{X}_1)\boldsymbol{R}\widetilde{\boldsymbol{L}}^\top\right)\\
&\le \max_{\widetilde{\boldsymbol{L}}:\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}}\le 1} \delta_{2r}\|\boldsymbol{X}_1\|_{\mathsf{F}}\|\widetilde{\boldsymbol{L}}\boldsymbol{R}^\top\|_{\mathsf{F}}\\
&\le \delta_{2r}\|\boldsymbol{X}_1\|_{\mathsf{F}}\|\boldsymbol{R}\|_{\mathsf{op}}.
\end{aligned}$$

Here, the first line follows from the variational representation of the Frobenius norm, the second line follows from (49), and the last line follows from the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathsf{F}} \le \|\boldsymbol{A}\|_{\mathsf{F}}\|\boldsymbol{B}\|_{\mathsf{op}}$.

C.1 Proof of Lemma 1

The proof mostly mirrors that in Section B.2. First, in view of the condition $\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star) \le 0.1\sigma_r(\boldsymbol{X}_\star)$ and Lemma 9, one knows that $\boldsymbol{Q}_t$, the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists. Therefore, for notational convenience, denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_t\boldsymbol{Q}_t$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_t\boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L}-\boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R}-\boldsymbol{R}_\star$, and $\epsilon \coloneqq 0.1$. Similar to the derivation in (45), we have

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \le \epsilon. \tag{51}$$

The conclusion $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \le 1.5\operatorname{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)$ is a simple consequence of Lemma 13; see (48) for a detailed argument. From now on, we focus on proving the distance contraction.

With these notations in place, we have by the definition of $\operatorname{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)$ that

$$\operatorname{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star) \le \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2. \tag{52}$$

Apply the update rule (15) and the decomposition $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$ to obtain

$$\begin{aligned}
(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} &= \left(\boldsymbol{L} - \eta\mathcal{A}^*\mathcal{A}(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\\
&= \left(\boldsymbol{\Delta}_L - \eta(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \eta(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\right)\boldsymbol{\Sigma}_\star^{1/2}\\
&= (1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \eta(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.
\end{aligned}$$

This allows us to expand the first square in (52) as

$$\begin{aligned}
\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &= \underbrace{\left\|(1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{S}_1}\\
&\quad - 2\eta(1-\eta)\underbrace{\operatorname{tr}\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)}_{\mathfrak{S}_2}
\end{aligned}$$

$$
+ 2\eta^2 \underbrace{\mathrm{tr}\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\right)}_{\mathfrak{S}_3}
$$

$$
+ \eta^2 \underbrace{\left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{S}_4}.
$$

In what follows, we shall control the four terms separately, of which $\mathfrak{S}_1$ is the main term, and $\mathfrak{S}_2, \mathfrak{S}_3$ and $\mathfrak{S}_4$ are perturbation terms.

1. Notice that the main term $\mathfrak{S}_1$ has already been controlled in (46) under the condition (51). It obeys

$$
\mathfrak{S}_1 \le \left((1-\eta)^2 + \frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.
$$

2. For the second term $\mathfrak{S}_2$, decompose $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top + \boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top$ and apply the triangle inequality to obtain

$$
\begin{aligned}
|\mathfrak{S}_2| &= \left|\mathrm{tr}\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top + \boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)\right| \\
&\le \left|\mathrm{tr}\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)\right| \\
&\quad + \left|\mathrm{tr}\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)\right| \\
&\quad + \left|\mathrm{tr}\left((\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)\right|.
\end{aligned}
$$

   Invoke Lemma 17 to further obtain

$$
\begin{aligned}
|\mathfrak{S}_2| &\le \delta_{2r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\right)\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right\|_{\mathsf{F}} \\
&\le \delta_{2r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\right)\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}},
\end{aligned}
$$

   where the second line follows from the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathsf{F}} \le \|\boldsymbol{A}\|_{\mathsf{op}}\|\boldsymbol{B}\|_{\mathsf{F}}$. Take the condition (51) and Lemmas 12 and 13 together to obtain

$$
\begin{aligned}
\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} &\le \frac{1}{1-\epsilon}; \\
\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}} + \|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}} &\le (1+\frac{\epsilon}{2})\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).
\end{aligned}
$$

   These consequences further imply that

$$
\begin{aligned}
|\mathfrak{S}_2| &\le \frac{\delta_{2r}(2+\epsilon)}{2(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \\
&= \frac{\delta_{2r}(2+\epsilon)}{2(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).
\end{aligned}
$$

   For the term $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}$, we can apply the elementary inequality $2ab \le a^2 + b^2$ to see

$$
\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \le \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{1}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.
$$

   The preceding two bounds taken collectively yield

$$
|\mathfrak{S}_2| \le \frac{\delta_{2r}(2+\epsilon)}{2\,(1-\epsilon)}\left(\frac{3}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).
$$

3. The third term $\mathfrak{S}_3$ can be similarly bounded as

$$\begin{aligned}
|\mathfrak{S}_3| &\le \delta_{2r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}}+\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\right)\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\right\|_{\mathsf{F}}\\
&\le \delta_{2r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}}+\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\right)\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}^2\|\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\|_{\mathsf{F}}\\
&\le \frac{\delta_{2r}(2+\epsilon)}{2(1-\epsilon)^2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\\
&\le \frac{\delta_{2r}(2+\epsilon)}{2(1-\epsilon)^2}\left(\frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{3}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).
\end{aligned}$$

4. We are then left with the last term $\mathfrak{S}_4$, for which we have

$$\begin{aligned}
\sqrt{\mathfrak{S}_4} &= \left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\\
&\le \left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\\
&\quad+\left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\\
&\quad+\left\|(\mathcal{A}^*\mathcal{A}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}},
\end{aligned}$$

where once again we use the decomposition $\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star=\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top+\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top$. Use (50) to see that

$$\sqrt{\mathfrak{S}_4}\le\delta_{2r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}}+\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\|_{\mathsf{F}}\right)\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}.$$

Repeating the same argument in bounding $\mathfrak{S}_2$ yields

$$\sqrt{\mathfrak{S}_4}\le\frac{\delta_{2r}\,(2+\epsilon)}{2\,(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).$$

We can then take the squares of both sides and use $(a+b)^2\le 2a^2+2b^2$ to reach

$$\mathfrak{S}_4\le\frac{\delta_{2r}^2(2+\epsilon)^2}{2(1-\epsilon)^2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).$$

Taking the bounds for $\mathfrak{S}_1,\mathfrak{S}_2,\mathfrak{S}_3,\mathfrak{S}_4$ collectively yields

$$\begin{aligned}
\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &\le \left((1-\eta)^2+\frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\\
&\quad+\frac{\delta_{2r}(2+\epsilon)}{1-\epsilon}\eta(1-\eta)\left(\frac{3}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{1}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right)\\
&\quad+\frac{\delta_{2r}(2+\epsilon)}{(1-\epsilon)^2}\eta^2\left(\frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{3}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right)\\
&\quad+\frac{\delta_{2r}^2(2+\epsilon)^2}{2(1-\epsilon)^2}\eta^2\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).
\end{aligned}$$

Similarly, we can expand the second square in (52) and obtain a similar bound. Combine both to obtain

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2\le\rho^2(\eta;\epsilon,\delta_{2r})\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star),$$

where the contraction rate is given by

$$\rho^2(\eta;\epsilon,\delta_{2r}):=(1-\eta)^2+\frac{2\epsilon+\delta_{2r}(4+2\epsilon)}{1-\epsilon}\eta(1-\eta)+\frac{2\epsilon+\epsilon^2+\delta_{2r}(4+2\epsilon)+\delta_{2r}^2(2+\epsilon)^2}{(1-\epsilon)^2}\eta^2.$$

With $\epsilon=0.1$, $\delta_{2r}\le 0.02$, and $0<\eta\le 2/3$, one has $\rho(\eta;\epsilon,\delta_{2r})\le 1-0.6\eta$. Thus we conclude that

$$\begin{aligned}
\mathrm{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star) &\le \sqrt{\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}\\
&\le (1-0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star).
\end{aligned}$$

C.2 Proof of Lemma 2

With the knowledge of partial Frobenius norm $\|\cdot\|_{\mathsf{F},r}$, we are ready to establish the claimed result. Invoke Lemma 11 to relate $\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star)$ to $\|\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star\|_\mathsf{F}$, and use that $\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star$ has rank at most $2r$ to obtain

$$\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le \sqrt{\sqrt{2}+1}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star\right\|_\mathsf{F} \le \sqrt{2(\sqrt{2}+1)}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F},r}.$$

Note that $\boldsymbol{L}_0\boldsymbol{R}_0^\top$ is the best rank-r approximation of $\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_\star)$, and apply the triangle inequality combined with Lemma 16 to obtain

$$\begin{aligned}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F},r} &\le \left\|\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_\star) - \boldsymbol{L}_0\boldsymbol{R}_0^\top\right\|_{\mathsf{F},r} + \|\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_\star) - \boldsymbol{X}_\star\|_{\mathsf{F},r} \\ &\le 2\left\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{X}_\star)\right\|_{\mathsf{F},r} \le 2\delta_{2r}\|\boldsymbol{X}_\star\|_\mathsf{F}.\end{aligned}$$

Here, the last inequality follows from combining Lemma 15 and (50) as

$$\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{X}_\star)\|_{\mathsf{F},r} = \max_{\widetilde{\boldsymbol{R}}\in\mathbb{R}^{n_2\times r}:\|\widetilde{\boldsymbol{R}}\|_{\mathsf{op}}\le 1}\left\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{X}_\star)\widetilde{\boldsymbol{R}}\right\|_\mathsf{F} \le \delta_{2r}\|\boldsymbol{X}_\star\|_\mathsf{F}.$$

As a result, one has

$$\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 2\sqrt{2(\sqrt{2}+1)}\delta_{2r}\|\boldsymbol{X}_\star\|_\mathsf{F} \le 5\delta_{2r}\sqrt{r}\kappa\sigma_r(\boldsymbol{X}_\star).$$

D Proof for Robust PCA

We first establish a useful property regarding the truncation operator $\mathcal{T}_{2\alpha}[\cdot]$.

Lemma 18. *Given $\boldsymbol{S}_\star \in \mathcal{S}_\alpha$ and $\boldsymbol{S} = \mathcal{T}_{2\alpha}[\boldsymbol{X}_\star + \boldsymbol{S}_\star - \boldsymbol{L}\boldsymbol{R}^\top]$, one has*

$$\|\boldsymbol{S} - \boldsymbol{S}_\star\|_\infty \le 2\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_\infty. \tag{53}$$

In addition, for any low-rank matrix $\boldsymbol{M} = \boldsymbol{L}_M\boldsymbol{R}_M^\top \in \mathbb{R}^{n_1\times n_2}$ with $\boldsymbol{L}_M \in \mathbb{R}^{n_1\times r}, \boldsymbol{R}_M \in \mathbb{R}^{n_2\times r}$, one has

$$\begin{aligned}|\langle \boldsymbol{S} - \boldsymbol{S}_\star, \boldsymbol{M}\rangle| &\le \sqrt{3\alpha}\nu\left(\|(\boldsymbol{L} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_\mathsf{F} + \|(\boldsymbol{R} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_\mathsf{F}\right)\|\boldsymbol{M}\|_\mathsf{F} \\ &\quad + 2\sqrt{\alpha}\left(\sqrt{n_1}\|\boldsymbol{L}_M\|_{2,\infty}\|\boldsymbol{R}_M\|_\mathsf{F} \wedge \sqrt{n_2}\|\boldsymbol{L}_M\|_\mathsf{F}\|\boldsymbol{R}_M\|_{2,\infty}\right)\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_\mathsf{F},\end{aligned} \tag{54}$$

where ν obeys

$$\nu \ge \frac{\sqrt{n_1}}{2}\left(\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\right) \vee \frac{\sqrt{n_2}}{2}\left(\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\right).$$

Proof. Denote $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, and $\boldsymbol{\Delta}_X \coloneqq \boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star$. Let $\Omega, \Omega_\star$ be the support of $\boldsymbol{S}$ and $\boldsymbol{S}_\star$, respectively. As a result, $\boldsymbol{S} - \boldsymbol{S}_\star$ is supported on $\Omega \cup \Omega_\star$.

We start with proving the first claim, i.e. (53). For $(i,j) \in \Omega$, by the definition of $\mathcal{T}_{2\alpha}[\cdot]$, we have $(\boldsymbol{S} - \boldsymbol{S}_\star)_{i,j} = (-\boldsymbol{\Delta}_X)_{i,j}$. For $(i,j) \in \Omega_\star \setminus \Omega$, one necessarily has $\boldsymbol{S}_{i,j} = 0$ and therefore $(\boldsymbol{S} - \boldsymbol{S}_\star)_{i,j} = (-\boldsymbol{S}_\star)_{i,j}$. Again by the definition of the operator $\mathcal{T}_{2\alpha}[\cdot]$, we know $|\boldsymbol{S}_\star - \boldsymbol{\Delta}_X|_{i,j}$ is either smaller than $|\boldsymbol{S}_\star - \boldsymbol{\Delta}_X|_{i,(2\alpha n_2)}$ or $|\boldsymbol{S}_\star - \boldsymbol{\Delta}_X|_{(2\alpha n_1),j}$. Furthermore, we know that $\boldsymbol{S}_\star$ contains at most α-fraction nonzero entries per row and column. Consequently, one has $|\boldsymbol{S}_\star - \boldsymbol{\Delta}_X|_{i,j} \le |\boldsymbol{\Delta}_X|_{i,(\alpha n_2)} \vee |\boldsymbol{\Delta}_X|_{(\alpha n_1),j}$. Combining the two cases above, we conclude that

$$|\boldsymbol{S} - \boldsymbol{S}_\star|_{i,j} \le \begin{cases}|\boldsymbol{\Delta}_X|_{i,j}, & (i,j) \in \Omega \\ |\boldsymbol{\Delta}_X|_{i,j} + \left(|\boldsymbol{\Delta}_X|_{i,(\alpha n_2)} \vee |\boldsymbol{\Delta}_X|_{(\alpha n_1),j}\right), & (i,j) \in \Omega_\star \setminus \Omega\end{cases}. \tag{55}$$

This immediately implies the ℓ_∞ norm bound (53).

Next, we prove the second claim (54). Recall that $\boldsymbol{S} - \boldsymbol{S}_\star$ is supported on $\Omega \cup \Omega_\star$. We then have

$$|\langle \boldsymbol{S} - \boldsymbol{S}_\star, \boldsymbol{M}\rangle| \le \sum_{(i,j)\in\Omega}|\boldsymbol{S} - \boldsymbol{S}_\star|_{i,j}|\boldsymbol{M}|_{i,j} + \sum_{(i,j)\in\Omega_\star\setminus\Omega}|\boldsymbol{S} - \boldsymbol{S}_\star|_{i,j}|\boldsymbol{M}|_{i,j}$$

$$\leq \sum_{(i,j)\in\Omega\cup\Omega_\star} |\boldsymbol{\Delta}_X|_{i,j}|\boldsymbol{M}|_{i,j} + \sum_{(i,j)\in\Omega_\star\setminus\Omega} \left(|\boldsymbol{\Delta}_X|_{i,(\alpha n_2)} + |\boldsymbol{\Delta}_X|_{(\alpha n_1),j}\right)|\boldsymbol{M}|_{i,j},$$

where the second line follows from (55). Let $\beta > 0$ be some positive number, whose value will be determined later. Use $2ab \leq \beta^{-1}a^2 + \beta b^2$ to further obtain

$$|\langle \boldsymbol{S}-\boldsymbol{S}_\star, \boldsymbol{M}\rangle| \leq \underbrace{\sum_{(i,j)\in\Omega\cup\Omega_\star} |\boldsymbol{\Delta}_X|_{i,j}|\boldsymbol{M}|_{i,j}}_{\mathfrak{A}_1} + \frac{1}{2\beta}\underbrace{\sum_{(i,j)\in\Omega_\star\setminus\Omega} \left(|\boldsymbol{\Delta}_X|^2_{i,(\alpha n_2)} + |\boldsymbol{\Delta}_X|^2_{(\alpha n_1),j}\right)}_{\mathfrak{A}_2} + \beta\underbrace{\sum_{(i,j)\in\Omega_\star\setminus\Omega} |\boldsymbol{M}|^2_{i,j}}_{\mathfrak{A}_3}.$$

In regard to the three terms $\mathfrak{A}_1, \mathfrak{A}_2$ and $\mathfrak{A}_3$, we have the following claims, whose proofs are deferred to the end.

Claim 1. *The first term $\mathfrak{A}_1$ satisfies*

$$\mathfrak{A}_1 \leq \sqrt{3\alpha}\nu\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{M}\|_{\mathsf{F}}.$$

Claim 2. *The second term $\mathfrak{A}_2$ satisfies*

$$\mathfrak{A}_2 \leq 2\|\boldsymbol{\Delta}_X\|_{\mathsf{F}}^2.$$

Claim 3. *The third term $\mathfrak{A}_3$ satisfies*

$$\mathfrak{A}_3 \leq \alpha\left(n_1\|\boldsymbol{L}_M\|_{2,\infty}^2\|\boldsymbol{R}_M\|_{\mathsf{F}}^2 \wedge n_2\|\boldsymbol{L}_M\|_{\mathsf{F}}^2\|\boldsymbol{R}_M\|_{2,\infty}^2\right).$$

Combine the pieces to reach

$$\begin{aligned}|\langle \boldsymbol{S}-\boldsymbol{S}_\star, \boldsymbol{M}\rangle| &\leq \sqrt{3\alpha}\nu\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{M}\|_{\mathsf{F}} \\ &\quad + \frac{\|\boldsymbol{\Delta}_X\|_{\mathsf{F}}^2}{\beta} + \beta\alpha\left(n_1\|\boldsymbol{L}_M\|_{2,\infty}^2\|\boldsymbol{R}_M\|_{\mathsf{F}}^2 \wedge n_2\|\boldsymbol{L}_M\|_{\mathsf{F}}^2\|\boldsymbol{R}_M\|_{2,\infty}^2\right).\end{aligned}$$

One can then choose β optimally to yield

$$\begin{aligned}|\langle \boldsymbol{S}-\boldsymbol{S}_\star, \boldsymbol{M}\rangle| &\leq \sqrt{3\alpha}\nu\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{M}\|_{\mathsf{F}} \\ &\quad + 2\sqrt{\alpha}\left(\sqrt{n_1}\|\boldsymbol{L}_M\|_{2,\infty}\|\boldsymbol{R}_M\|_{\mathsf{F}} \wedge \sqrt{n_2}\|\boldsymbol{L}_M\|_{\mathsf{F}}\|\boldsymbol{R}_M\|_{2,\infty}\right)\|\boldsymbol{\Delta}_X\|_{\mathsf{F}}.\end{aligned}$$

This finishes the proof. □

Proof of Claim 1. Use the decomposition $\boldsymbol{\Delta}_X = \boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top = \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top + \boldsymbol{L}\boldsymbol{\Delta}_R^\top$ to obtain

$$\begin{aligned}|\boldsymbol{\Delta}_X|_{i,j} &\leq \|(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2, \quad \text{and} \\ |\boldsymbol{\Delta}_X|_{i,j} &\leq \|(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2\|\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2.\end{aligned}$$

Take the average to yield

$$|\boldsymbol{\Delta}_X|_{i,j} \leq \frac{\nu}{\sqrt{n_2}}\|(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2 + \frac{\nu}{\sqrt{n_1}}\|(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2,$$

where we have used the assumption on ν. With this upper bound on $|\boldsymbol{\Delta}_X|_{i,j}$ in place, we can further control $\mathfrak{A}_1$ as

$$\begin{aligned}\mathfrak{A}_1 &\leq \sum_{(i,j)\in\Omega\cup\Omega_\star} \frac{\nu}{\sqrt{n_2}}\|(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2|\boldsymbol{M}|_{i,j} + \sum_{(i,j)\in\Omega\cup\Omega_\star} \frac{\nu}{\sqrt{n_1}}\|(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2|\boldsymbol{M}|_{i,j} \\ &\leq \left(\sqrt{\sum_{(i,j)\in\Omega\cup\Omega_\star} \|(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2^2/n_2} + \sqrt{\sum_{(i,j)\in\Omega\cup\Omega_\star} \|(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2^2/n_1}\right)\nu\|\boldsymbol{M}\|_{\mathsf{F}}.\end{aligned}$$

Regarding the first term, one has

$$\begin{aligned}\sum_{(i,j)\in\Omega\cup\Omega_\star}\|(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2^2 &= \sum_{i=1}^{n_1}\sum_{j:(i,j)\in\Omega\cup\Omega_\star}\|(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2^2\\ &\le 3\alpha n_2\sum_{i=1}^{n_1}\|(\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}\|_2^2\\ &= 3\alpha n_2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2,\end{aligned}$$

where the second line follows from the fact that $\Omega\cup\Omega_\star$ contains at most $3\alpha n_2$ non-zero entries in each row. Similarly, we can show that

$$\sum_{(i,j)\in\Omega\cup\Omega_\star}\|(\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot}\|_2^2 \le 3\alpha n_1\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

In all, we arrive at

$$\mathfrak{A}_1 \le \sqrt{3\alpha}\nu\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{M}\|_{\mathsf{F}},$$

which is the desired claim. □

Proof of Claim 2. Recall that $(\boldsymbol{\Delta}_X)_{i,(\alpha n_2)}$ denotes the (αn_2)-th largest entry in the i-th row of $\boldsymbol{\Delta}_X$. One necessarily has

$$\alpha n_2|\boldsymbol{\Delta}_X|_{i,(\alpha n_2)}^2 \le \|(\boldsymbol{\Delta}_X)_{i,\cdot}\|_2^2.$$

As a result, we obtain

$$\begin{aligned}\sum_{(i,j)\in\Omega_\star\setminus\Omega}|\boldsymbol{\Delta}_X|_{i,(\alpha n_2)}^2 &\le \sum_{(i,j)\in\Omega_\star}|\boldsymbol{\Delta}_X|_{i,(\alpha n_2)}^2\\ &\le \sum_{i=1}^{n_1}\sum_{j:(i,j)\in\Omega_\star}\frac{\|(\boldsymbol{\Delta}_X)_{i,\cdot}\|_2^2}{\alpha n_2}\\ &\le \sum_{i=1}^{n_1}\|(\boldsymbol{\Delta}_X)_{i,\cdot}\|_2^2 = \|\boldsymbol{\Delta}_X\|_{\mathsf{F}}^2,\end{aligned}$$

where the last line follows from the fact that $\Omega_\star$ contains at most αn_2 nonzero entries in each row. Similarly one can show that

$$\sum_{(i,j)\in\Omega_\star\setminus\Omega}|\boldsymbol{\Delta}_X|_{(\alpha n_1),j}^2 \le \|\boldsymbol{\Delta}_X\|_{\mathsf{F}}^2.$$

Combining the above two bounds with the definition of $\mathfrak{A}_2$ completes the proof. □

Proof of Claim 3. By definition, $\boldsymbol{M} = \boldsymbol{L}_M\boldsymbol{R}_M^\top$, and hence one has

$$\mathfrak{A}_3 = \sum_{(i,j)\in\Omega_\star\setminus\Omega}|(\boldsymbol{L}_M)_{i,\cdot}(\boldsymbol{R}_M)_{j,\cdot}^\top|^2 \le \sum_{(i,j)\in\Omega_\star}|(\boldsymbol{L}_M)_{i,\cdot}(\boldsymbol{R}_M)_{j,\cdot}^\top|^2.$$

We can further upper bound $\mathfrak{A}_3$ as

$$\begin{aligned}\mathfrak{A}_3 &\le \sum_{(i,j)\in\Omega_\star}\|(\boldsymbol{L}_M)_{i,\cdot}\|_2^2\|(\boldsymbol{R}_M)_{j,\cdot}\|_2^2\\ &\le \sum_{i=1}^{n_1}\sum_{j:(i,j)\in\Omega_\star}\|(\boldsymbol{L}_M)_{i,\cdot}\|_2^2\|\boldsymbol{R}_M\|_{2,\infty}^2\end{aligned}$$

$$\leq \sum_{i=1}^{n_1} \alpha n_2 \|(\boldsymbol{L}_M)_{i,\cdot}\|_2^2 \|\boldsymbol{R}_M\|_{2,\infty}^2 = \alpha n_2 \|\boldsymbol{L}_M\|_{\mathsf{F}}^2 \|\boldsymbol{R}_M\|_{2,\infty}^2,$$

where the last line follows from the fact that $\Omega_\star$ contains at most αn_2 non-zero entries in each row. Similarly, one can obtain

$$\mathfrak{A}_3 \leq \alpha n_1 \|\boldsymbol{L}_M\|_{2,\infty}^2 \|\boldsymbol{R}_M\|_{\mathsf{F}}^2,$$

which completes the proof. $\square$

D.1 Proof of Lemma 3

We begin with introducing several useful notations and facts. In view of the condition $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \leq 0.02\sigma_r(\boldsymbol{X}_\star)$ and Lemma 9, one knows that $\boldsymbol{Q}_t$, the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists. Therefore, for notational convenience, denote $\boldsymbol{L} := \boldsymbol{L}_t\boldsymbol{Q}_t$, $\boldsymbol{R} := \boldsymbol{R}_t\boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L := \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R := \boldsymbol{R} - \boldsymbol{R}_\star$, $\boldsymbol{S} := \boldsymbol{S}_t = \mathcal{T}_{2\alpha}[\boldsymbol{X}_\star + \boldsymbol{S}_\star - \boldsymbol{L}\boldsymbol{R}^\top]$, and $\epsilon := 0.02$. Similar to the derivation in (45), we have

$$\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \vee \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \leq \epsilon. \tag{56}$$

Moreover, the incoherence condition

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \leq \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star) \tag{57}$$

implies

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \leq \sqrt{\mu r}, \tag{58}$$

which combined with the triangle inequality further implies

$$\sqrt{n_1}\|\boldsymbol{L} \boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R} \boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \leq 2\sqrt{\mu r}. \tag{59}$$

The conclusion $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \leq 1.5\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ is a simple consequence of Lemma 13; see (48) for a detailed argument. In what follows, we shall prove the distance contraction and the incoherence condition separately.

D.1.1 Distance contraction

By the definition of $\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star)$, one has

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) \leq \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2. \tag{60}$$

From now on, we focus on controlling the first square $\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$. In view of the update rule (20), one has

$$\begin{aligned}
(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} &= \left(\boldsymbol{L} - \eta(\boldsymbol{L}\boldsymbol{R}^\top + \boldsymbol{S} - \boldsymbol{X}_\star - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= \left(\boldsymbol{\Delta}_L - \eta(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \eta(\boldsymbol{S} - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= (1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \eta(\boldsymbol{S} - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.
\end{aligned} \tag{61}$$

Here, we use the notation introduced above and the decomposition $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$. Take the squared Frobenius norm of both sides of (61) to obtain

$$\begin{aligned}
\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &= \underbrace{\left\|(1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{R}_1} \\
&\quad - 2\eta(1-\eta)\underbrace{\mathrm{tr}\left((\boldsymbol{S} - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)}_{\mathfrak{R}_2}
\end{aligned}$$

$$
+ 2\eta^2 \underbrace{\operatorname{tr}\left((\boldsymbol{S}-\boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\right)}_{\mathfrak{R}_3}
+ \eta^2 \underbrace{\left\|(\boldsymbol{S}-\boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{R}_4}.
$$

In the sequel, we shall bound the four terms separately, of which $\mathfrak{R}_1$ is the main term, and $\mathfrak{R}_2, \mathfrak{R}_3$ and $\mathfrak{R}_4$ are perturbation terms.

1. Notice that the main term $\mathfrak{R}_1$ has already been controlled in (46) under the condition (56). It obeys

$$
\mathfrak{R}_1 \le \left((1-\eta)^2 + \frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.
$$

2. For the second term $\mathfrak{R}_2$, set $\boldsymbol{M} \coloneqq \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top$ with $\boldsymbol{L}_M \coloneqq \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}$, $\boldsymbol{R}_M \coloneqq \boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}$, and then invoke Lemma 18 with $\nu \coloneqq 3\sqrt{\mu r}/2$ to see

$$
\begin{aligned}
|\mathfrak{R}_2| &\le \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\left\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\|_{\mathsf{F}} \\
&\quad + 2\sqrt{\alpha n_2}\left\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \\
&\le \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \\
&\quad + 2\sqrt{\alpha n_2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}}.
\end{aligned}
$$

Take the condition (56) and Lemmas 12 and 13 together to obtain

$$
\begin{aligned}
\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} &\le \frac{1}{1-\epsilon}; \\
\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} = \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}^2 &\le \frac{1}{(1-\epsilon)^2}; \\
\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} &\le (1+\frac{\epsilon}{2})\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).
\end{aligned}
\tag{62}
$$

These consequences combined with the condition (59) yield

$$
\begin{aligned}
|\mathfrak{R}_2| &\le \frac{3\sqrt{3\alpha\mu r}}{2(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \\
&\quad + \frac{4\sqrt{\alpha\mu r}}{(1-\epsilon)^2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}(1+\frac{\epsilon}{2})\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right) \\
&\le \sqrt{\alpha\mu r}\frac{3\sqrt{3}+\frac{4(2+\epsilon)}{1-\epsilon}}{2(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right) \\
&\le \sqrt{\alpha\mu r}\frac{3\sqrt{3}+\frac{4(2+\epsilon)}{1-\epsilon}}{2(1-\epsilon)}\left(\frac{3}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{1}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right),
\end{aligned}
$$

where the last inequality holds since $2ab \le a^2+b^2$.

3. The third term $\mathfrak{R}_3$ can be controlled similarly. Set $\boldsymbol{M} \coloneqq \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top$ with $\boldsymbol{L}_M \coloneqq \boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}$ and $\boldsymbol{R}_M \coloneqq \boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}$, and invoke Lemma 18 with $\nu \coloneqq 3\sqrt{\mu r}/2$ to arrive at

$$
\begin{aligned}
|\mathfrak{R}_3| &\le \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\left\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\|_{\mathsf{F}} \\
&\quad + 2\sqrt{\alpha n_1}\|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\|\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star\|_{\mathsf{F}}
\end{aligned}
$$

$$\leq \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}^2$$
$$+2\sqrt{\alpha n_1}\|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}.$$

Use the consequences (62) again to obtain

$$\begin{aligned}
|\mathfrak{R}_3| &\leq \frac{3\sqrt{3\alpha\mu r}}{2(1-\epsilon)^2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \\
&\quad+\frac{2\sqrt{\alpha\mu r}}{(1-\epsilon)^2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}(1+\frac{\epsilon}{2})\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right) \\
&\leq \sqrt{\alpha\mu r}\frac{3\sqrt{3}+2(2+\epsilon)}{2(1-\epsilon)^2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\
&\leq \sqrt{\alpha\mu r}\frac{3\sqrt{3}+2(2+\epsilon)}{2(1-\epsilon)^2}\left(\frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{3}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).
\end{aligned}$$

4. For the last term $\mathfrak{R}_4$, utilize the variational representation of the Frobenius norm to see

$$\sqrt{\mathfrak{R}_4}=\mathrm{tr}\left((\boldsymbol{S}-\boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\widetilde{\boldsymbol{L}}^\top\right)$$

for some $\widetilde{\boldsymbol{L}}\in\mathbb{R}^{n_1\times r}$ obeying $\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}}=1$. Setting $\boldsymbol{M}:=\widetilde{\boldsymbol{L}}\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top=\boldsymbol{L}_M\boldsymbol{R}_M^\top$ with $\boldsymbol{L}_M:=\widetilde{\boldsymbol{L}}\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}$ and $\boldsymbol{R}_M:=\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}$, we are ready to apply Lemma 18 again with $\nu:=3\sqrt{\mu r}/2$ to see

$$\begin{aligned}
\sqrt{\mathfrak{R}_4} &\leq \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\left\|\widetilde{\boldsymbol{L}}\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\|_{\mathsf{F}} \\
&\quad+2\sqrt{\alpha n_2}\left\|\widetilde{\boldsymbol{L}}\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_{\mathsf{F}} \\
&\leq \frac{3}{2}\sqrt{3\alpha\mu r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \\
&\quad+2\sqrt{\alpha n_2}\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}.
\end{aligned}$$

This combined with the consequences (62) and condition (59) yields

$$\sqrt{\mathfrak{R}_4}\leq\sqrt{\alpha\mu r}\frac{3\sqrt{3}+\frac{4(2+\epsilon)}{1-\epsilon}}{2(1-\epsilon)}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right).$$

Take the square, and use the elementary inequality $(a+b)^2\leq 2a^2+2b^2$ to reach

$$\mathfrak{R}_4\leq\alpha\mu r\frac{(3\sqrt{3}+\frac{4(2+\epsilon)}{1-\epsilon})^2}{2(1-\epsilon)^2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).$$

Taking collectively the bounds for $\mathfrak{R}_1,\mathfrak{R}_2,\mathfrak{R}_3$ and $\mathfrak{R}_4$ yields the control of $\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$ as

$$\begin{aligned}
\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &\leq \left((1-\eta)^2+\frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 \\
&\quad+\sqrt{\alpha\mu r}\frac{3\sqrt{3}+\frac{4(2+\epsilon)}{1-\epsilon}}{1-\epsilon}\eta(1-\eta)\left(\frac{3}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{1}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\
&\quad+\sqrt{\alpha\mu r}\frac{3\sqrt{3}+2(2+\epsilon)}{(1-\epsilon)^2}\eta^2\left(\frac{1}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{3}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\
&\quad+\alpha\mu r\frac{(3\sqrt{3}+\frac{4(2+\epsilon)}{1-\epsilon})^2}{2(1-\epsilon)^2}\eta^2\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).
\end{aligned}$$

Similarly, we can obtain the control of $\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$. Combine them together and identify $\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) = \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$ to reach

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \le \rho^2(\eta;\epsilon,\alpha\mu r)\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

where the contraction rate $\rho^2(\eta;\epsilon,\alpha\mu r)$ is given by

$$\begin{aligned}\rho^2(\eta;\epsilon,\alpha\mu r) &\coloneqq (1-\eta)^2 + \frac{2\epsilon + \sqrt{\alpha\mu r}(6\sqrt{3} + \frac{8(2+\epsilon)}{1-\epsilon})}{1-\epsilon}\eta(1-\eta) \\ &\quad + \frac{2\epsilon + \epsilon^2 + \sqrt{\alpha\mu r}(6\sqrt{3} + 4(2+\epsilon)) + \alpha\mu r(3\sqrt{3} + \frac{4(2+\epsilon)}{1-\epsilon})^2}{(1-\epsilon)^2}\eta^2.\end{aligned}$$

With $\epsilon = 0.02$, $\alpha\mu r \le 10^{-4}$, and $0 < \eta \le 2/3$, one has $\rho(\eta;\epsilon,\alpha\mu r) \le 1 - 0.6\eta$. Thus we conclude that

$$\begin{aligned}\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) &\le \sqrt{\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2} \\ &\le (1-0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star).\end{aligned} \tag{63}$$

D.1.2 Incoherence condition

We start by controlling the term $\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}$. We know from (61) that

$$(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} = (1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \eta(\boldsymbol{S} - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.$$

Apply the triangle inequality to obtain

$$\begin{aligned}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} &\le (1-\eta)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} + \eta\underbrace{\left\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}}_{\mathfrak{T}_1} \\ &\quad + \eta\underbrace{\left\|(\boldsymbol{S} - \boldsymbol{S}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}}_{\mathfrak{T}_2}.\end{aligned}$$

The first term $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}$ follows from the incoherence condition (57) as

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star).$$

In the sequel, we shall bound the terms $\mathfrak{T}_1$ and $\mathfrak{T}_2$.

1. For the term $\mathfrak{T}_1$, use the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{2,\infty} \le \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|_{\mathsf{op}}$, and combine the condition (56) with the consequences (62) to obtain

$$\begin{aligned}\mathfrak{T}_1 &\le \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \\ &\le \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \\ &\le \frac{\epsilon}{1-\epsilon}\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star),\end{aligned}$$

2. For the term $\mathfrak{T}_2$, use the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{2,\infty} \le \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|_{\mathsf{op}}$ to obtain

$$\mathfrak{T}_2 \le \|\boldsymbol{S} - \boldsymbol{S}_\star\|_{2,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}.$$

We know from Lemma 18 that $\boldsymbol{S}-\boldsymbol{S}_\star$ has at most $3\alpha n_2$ non-zero entries in each row, and $\|\boldsymbol{S}-\boldsymbol{S}_\star\|_\infty \le 2\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_\infty$. Upper bound the $\ell_{2,\infty}$ norm by the ℓ_∞ norm as

$$\|\boldsymbol{S}-\boldsymbol{S}_\star\|_{2,\infty} \le \sqrt{3\alpha n_2}\|\boldsymbol{S}-\boldsymbol{S}_\star\|_\infty \le 2\sqrt{3\alpha n_2}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_\infty.$$

Split $\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star=\boldsymbol{\Delta}_L\boldsymbol{R}^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$, and take the conditions (57) and (59) to obtain

$$\begin{aligned}\|\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\|_\infty &\le \|\boldsymbol{\Delta}_L\boldsymbol{R}^\top\|_\infty+\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\|_\infty\\ &\le \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}+\|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}\\ &\le \sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star)2\sqrt{\frac{\mu r}{n_2}}+\sqrt{\frac{\mu r}{n_1}}\sqrt{\frac{\mu r}{n_2}}\sigma_r(\boldsymbol{X}_\star)\\ &=\frac{3\mu r}{\sqrt{n_1 n_2}}\sigma_r(\boldsymbol{X}_\star).\end{aligned}$$

This combined with the consequences (62) yields

$$\mathfrak{T}_2 \le \frac{6\sqrt{3\alpha\mu r}}{1-\epsilon}\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star).$$

Taking collectively the bounds for $\mathfrak{T}_1,\mathfrak{T}_2$ yields the control

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \left(1-\eta+\frac{\epsilon+6\sqrt{3\alpha\mu r}}{1-\epsilon}\eta\right)\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star). \tag{64}$$

The last step is to switch the alignment matrix from $\boldsymbol{Q}_t$ to $\boldsymbol{Q}_{t+1}$. (63) together with Lemma 9 demonstrates the existence of $\boldsymbol{Q}_{t+1}$. Apply the triangle inequality to obtain

$$\begin{aligned}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} &\le \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}+\left\|\boldsymbol{L}_{t+1}(\boldsymbol{Q}_{t+1}-\boldsymbol{Q}_t)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}\\ &\le \left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}+\|\boldsymbol{L}_{t+1}\boldsymbol{Q}_t\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{Q}_t^{-1}\boldsymbol{Q}_{t+1}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{\Sigma}_\star\right\|_{\mathsf{op}}.\end{aligned}$$

We deduct from (64) that

$$\|\boldsymbol{L}_{t+1}\boldsymbol{Q}_t\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \le \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}+\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\right\|_{2,\infty} \le \left(2-\eta+\frac{\epsilon+6\sqrt{3\alpha\mu r}}{1-\epsilon}\eta\right)\sqrt{\frac{\mu r}{n_1}}.$$

Regarding the alignment matrix term, invoke Lemma 14 to obtain

$$\begin{aligned}\left\|\boldsymbol{\Sigma}_\star^{1/2}\boldsymbol{Q}_t^{-1}\boldsymbol{Q}_{t+1}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{\Sigma}_\star\right\|_{\mathsf{op}} &\le \frac{\|(\boldsymbol{R}_{t+1}(\boldsymbol{Q}_t^{-\top}-\boldsymbol{Q}_{t+1}^{-\top})\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}}}{1-\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}}\\ &\le \frac{\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}}+\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}}}{1-\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}}\\ &\le \frac{2\epsilon}{1-\epsilon}\sigma_r(\boldsymbol{X}_\star),\end{aligned}$$

where we deduct from (63) that the distances using either $\boldsymbol{Q}_t$ or $\boldsymbol{Q}_{t+1}$ are bounded by

$$\begin{aligned}\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}} &\le \epsilon\sigma_r(\boldsymbol{X}_\star);\\ \|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}} &\le \epsilon\sigma_r(\boldsymbol{X}_\star);\\ \|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} &\le \epsilon.\end{aligned}$$

Combine all pieces to reach

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \left(\frac{1+\epsilon}{1-\epsilon}\left(1-\eta+\frac{\epsilon+6\sqrt{3\alpha\mu r}}{1-\epsilon}\eta\right)+\frac{2\epsilon}{1-\epsilon}\right)\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star).$$

With $\epsilon = 0.02$, $\alpha\mu r \le 10^{-4}$, and $0.1 \le \eta \le 2/3$, we get the desired incoherence condition

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_{t+1} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star).$$

Similarly, we can prove the other part

$$\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_{t+1}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_2}}\sigma_r(\boldsymbol{X}_\star).$$

D.2 Proof of Lemma 4

We first record two lemmas from [YPCC16], which are useful for studying the properties of the initialization.

Lemma 19 ([YPCC16, Section 6.1]). *Given $\boldsymbol{S}_\star \in \mathcal{S}_\alpha$, one has $\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_\infty \le 2\|\boldsymbol{X}_\star\|_\infty$.*

Lemma 20 ([YPCC16, Lemma 1]). *For any matrix $\boldsymbol{M} \in \mathcal{S}_\alpha$, one has $\|\boldsymbol{M}\|_{\mathsf{op}} \le \alpha\sqrt{n_1 n_2}\|\boldsymbol{M}\|_\infty$.*

With these two lemmas in place, we are ready to establish the claimed result. Invoke Lemma 11 to obtain

$$\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le \sqrt{\sqrt{2}+1}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le \sqrt{(\sqrt{2}+1)2r}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star\right\|_{\mathsf{op}},$$

where the last relation uses the fact that $\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star$ has rank at most $2r$. We can further apply the triangle inequality to see

$$\begin{aligned}\left\|\boldsymbol{L}_0\boldsymbol{R}_0^\top - \boldsymbol{X}_\star\right\|_{\mathsf{op}} &\le \left\|\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}] - \boldsymbol{L}_0\boldsymbol{R}_0^\top\right\|_{\mathsf{op}} + \left\|\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}] - \boldsymbol{X}_\star\right\|_{\mathsf{op}} \\ &\le 2\left\|\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}] - \boldsymbol{X}_\star\right\|_{\mathsf{op}} = 2\left\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\right\|_{\mathsf{op}}.\end{aligned}$$

Here the second inequality hinges on the fact that $\boldsymbol{L}_0\boldsymbol{R}_0^\top$ is the best rank-r approximation of $\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}]$, and the last identity arises from $\boldsymbol{Y} = \boldsymbol{X}_\star + \boldsymbol{S}_\star$. Follow the same argument as [YPCC16, Section 6.1], combining Lemmas 19 and 20 to reach

$$\begin{aligned}\left\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\right\|_{\mathsf{op}} &\le 2\alpha\sqrt{n_1 n_2}\left\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\right\|_\infty \\ &\le 4\alpha\sqrt{n_1 n_2}\|\boldsymbol{X}_\star\|_\infty \le 4\alpha\mu r\kappa\sigma_r(\boldsymbol{X}_\star),\end{aligned}$$

where the last inequality follows from the incoherence assumption

$$\|\boldsymbol{X}_\star\|_\infty \le \|\boldsymbol{U}_\star\|_{2,\infty}\|\boldsymbol{\Sigma}_\star\|_{\mathsf{op}}\|\boldsymbol{V}_\star\|_{2,\infty} \le \frac{\mu r}{\sqrt{n_1 n_2}}\kappa\sigma_r(\boldsymbol{X}_\star). \tag{65}$$

Take the above inequalities together to arrive at

$$\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 8\sqrt{2(\sqrt{2}+1)}\alpha\mu r^{3/2}\kappa\sigma_r(\boldsymbol{X}_\star) \le 20\alpha\mu r^{3/2}\kappa\sigma_r(\boldsymbol{X}_\star).$$

D.3 Proof of Lemma 5

In view of the condition $\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le 0.02\sigma_r(\boldsymbol{X}_\star)$ and Lemma 9, one knows that $\boldsymbol{Q}_0$, the optimal alignment matrix between $\boldsymbol{F}_0$ and $\boldsymbol{F}_\star$ exists. Therefore, for notational convenience, denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_0\boldsymbol{Q}_0$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_0\boldsymbol{Q}_0^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, and $\epsilon \coloneqq 0.02$. Our objective is then translated to demonstrate

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star).$$

From now on, we focus on bounding $\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}$. Since $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$ is the top-r SVD of $\boldsymbol{Y} - \mathcal{T}_\alpha[\boldsymbol{Y}]$, and recall that $\boldsymbol{Y} = \boldsymbol{X}_\star + \boldsymbol{S}_\star$, we have the relation

$$(\boldsymbol{X}_\star + \boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star])\boldsymbol{V}_0 = \boldsymbol{U}_0\boldsymbol{\Sigma}_0,$$

which further implies the following decomposition of $\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}$.

Claim 4. *One has*

$$\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} = (\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star])\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.$$

Combining Claim 4 with the triangle inequality yields

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \underbrace{\left\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}}_{\mathfrak{J}_1} + \underbrace{\left\|(\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star])\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty}}_{\mathfrak{J}_2}.$$

In what follows, we shall control $\mathfrak{J}_1$ and $\mathfrak{J}_2$ in turn.

1. For the term $\mathfrak{J}_1$, use the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{2,\infty} \le \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|_{\mathsf{op}}$ to obtain

$$\mathfrak{J}_1 \le \|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}.$$

The incoherence assumption tells $\|\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} = \|\boldsymbol{U}_\star\|_{2,\infty} \le \sqrt{\mu r/n_1}$. In addition, the assumption $\mathrm{dist}(\boldsymbol{F}_0, \boldsymbol{F}_\star) \le \epsilon\sigma_r(\boldsymbol{X}_\star)$ entails the bound $\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{op}} \le \epsilon\sigma_r(\boldsymbol{X}_\star)$. Finally, repeating the argument for obtaining (56) yields $\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \le \epsilon$, which together with Lemma 12 reveals

$$\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \le \frac{1}{1-\epsilon}.$$

In all, we arrive at

$$\mathfrak{J}_1 \le \frac{\epsilon}{1-\epsilon}\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star).$$

2. Proceeding to the term $\mathfrak{J}_2$, use the relations $\|\boldsymbol{A}\boldsymbol{B}\|_{2,\infty} \le \|\boldsymbol{A}\|_{1,\infty}\|\boldsymbol{B}\|_{2,\infty}$ and $\|\boldsymbol{A}\boldsymbol{B}\|_{2,\infty} \le \|\boldsymbol{A}\|_{2,\infty}\|\boldsymbol{B}\|_{\mathsf{op}}$ to obtain

$$\begin{aligned}\mathfrak{J}_2 &\le \|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_{1,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{2,\infty} \\ &\le \|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_{1,\infty}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}.\end{aligned}$$

Regarding $\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]$, Lemma 19 tells that $\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]$ has at most $2\alpha n_2$ non-zero entries in each row, and $\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_\infty \le 2\|\boldsymbol{X}_\star\|_\infty$. Consequently, we can upper bound the $\ell_{1,\infty}$ norm by the ℓ_∞ norm as

$$\begin{aligned}\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_{1,\infty} &\le 2\alpha n_2\|\boldsymbol{S}_\star - \mathcal{T}_\alpha[\boldsymbol{X}_\star + \boldsymbol{S}_\star]\|_\infty \\ &\le 4\alpha n_2\|\boldsymbol{X}_\star\|_\infty \\ &\le 4\alpha n_2\frac{\mu r}{\sqrt{n_1 n_2}}\kappa\sigma_r(\boldsymbol{X}_\star).\end{aligned}$$

Here the last inequality follows from the incoherence assumption (65). For the term $\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}$, one can apply the triangle inequality to see

$$\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \le \|\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_2}} + \frac{\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}}{\sigma_r(\boldsymbol{X}_\star)}.$$

Last but not least, repeat the argument for (62) to obtain

$$\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} = \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}^2 \le \frac{1}{(1-\epsilon)^2}.$$

Taking together the above bounds yields

$$\mathfrak{J}_2 \le \frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\sqrt{\frac{\mu r}{n_1}}\sigma_r(\boldsymbol{X}_\star) + \frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\sqrt{\frac{n_2}{n_1}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}.$$

Combine the bounds on $\mathfrak{I}_1$ and $\mathfrak{I}_2$ to reach

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \left(\frac{\epsilon}{1-\epsilon}+\frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\right)\sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star)+\frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\sqrt{n_2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}.$$

Similarly, we have

$$\sqrt{n_2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \left(\frac{\epsilon}{1-\epsilon}+\frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\right)\sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star)+\frac{4\alpha\mu r\kappa}{(1-\epsilon)^2}\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}.$$

Taking the maximum and solving for $\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}\vee\sqrt{n_2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}$ yield the relation

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}\vee\sqrt{n_2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \frac{\epsilon(1-\epsilon)+4\alpha\mu r\kappa}{(1-\epsilon)^2-4\alpha\mu r\kappa}\sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star).$$

With $\epsilon=0.02$ and $\alpha\mu r\kappa\le 0.1$, we get the desired conclusion

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}\vee\sqrt{n_2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \sqrt{\mu r}\sigma_r(\boldsymbol{X}_\star).$$

Proof of Claim 4. Identify $\boldsymbol{U}_0$ (resp. $\boldsymbol{V}_0$) with $\boldsymbol{L}_0\boldsymbol{\Sigma}_0^{-1/2}$ (resp. $\boldsymbol{R}_0\boldsymbol{\Sigma}_0^{-1/2}$) to yield

$$(\boldsymbol{X}_\star+\boldsymbol{S}_\star-\mathcal{T}_\alpha[\boldsymbol{X}_\star+\boldsymbol{S}_\star])\boldsymbol{R}_0\boldsymbol{\Sigma}_0^{-1}=\boldsymbol{L}_0,$$

which is equivalent to $(\boldsymbol{X}_\star+\boldsymbol{S}_\star-\mathcal{T}_\alpha[\boldsymbol{X}_\star+\boldsymbol{S}_\star])\boldsymbol{R}_0(\boldsymbol{R}_0^\top\boldsymbol{R}_0)^{-1}=\boldsymbol{L}_0$ since $\boldsymbol{\Sigma}_0=\boldsymbol{R}_0^\top\boldsymbol{R}_0$. Multiply both sides by $\boldsymbol{Q}_0\boldsymbol{\Sigma}_\star^{1/2}$ to obtain

$$(\boldsymbol{X}_\star+\boldsymbol{S}_\star-\mathcal{T}_\alpha[\boldsymbol{X}_\star+\boldsymbol{S}_\star])\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}=\boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2},$$

where we recall that $\boldsymbol{L}=\boldsymbol{L}_0\boldsymbol{Q}_0$ and $\boldsymbol{R}=\boldsymbol{R}_0\boldsymbol{Q}_0^{-\top}$. In the end, subtract $\boldsymbol{X}_\star\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}$ from both sides to reach

$$\begin{aligned}(\boldsymbol{S}_\star-\mathcal{T}_\alpha[\boldsymbol{X}_\star+\boldsymbol{S}_\star])\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}&=\boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2}-\boldsymbol{L}_\star\boldsymbol{R}_\star^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\\&=(\boldsymbol{L}-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}+\boldsymbol{L}_\star(\boldsymbol{R}-\boldsymbol{R}_\star)^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\\&=\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.\end{aligned}$$

This finishes the proof. □

E Proof for Matrix Completion

E.1 New projection operator

E.1.1 Proof of Proposition 1

First, notice that the optimization of $\boldsymbol{L}$ and $\boldsymbol{R}$ in (23) can be decomposed and done in parallel, hence we focus on the optimization of $\boldsymbol{L}$ below:

$$\boldsymbol{L}=\operatorname*{argmin}_{\boldsymbol{L}\in\mathbb{R}^{n_1\times r}}\left\|(\boldsymbol{L}-\widetilde{\boldsymbol{L}})(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}\right\|_{\mathsf{F}}^2\qquad\text{s.t.}\quad\sqrt{n_1}\left\|\boldsymbol{L}(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}\right\|_{2,\infty}\le B.$$

By a change of variables as $\boldsymbol{G}\coloneqq\boldsymbol{L}(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}$ and $\widetilde{\boldsymbol{G}}\coloneqq\widetilde{\boldsymbol{L}}(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{1/2}$, we rewrite the above problem equivalently as

$$\boldsymbol{G}=\operatorname*{argmin}_{\boldsymbol{G}\in\mathbb{R}^{n_1\times r}}\|\boldsymbol{G}-\widetilde{\boldsymbol{G}}\|_{\mathsf{F}}^2\qquad\text{s.t.}\quad\sqrt{n_1}\,\|\boldsymbol{G}\|_{2,\infty}\le B,$$

whose solution is given as [CW15]

$$\boldsymbol{G}_{i,\cdot}=\left(1\wedge\frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{G}}_{i,\cdot}\|_2}\right)\widetilde{\boldsymbol{G}}_{i,\cdot},\quad 1\le i\le n_1.$$

By applying again the change of variable $\boldsymbol{L}=\boldsymbol{G}(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{-1/2}$ and $\widetilde{\boldsymbol{L}}=\widetilde{\boldsymbol{G}}(\widetilde{\boldsymbol{R}}^\top\widetilde{\boldsymbol{R}})^{-1/2}$, we obtain the claimed solution.

E.1.2 Proof of Lemma 6

We begin with proving the non-expansiveness property. Denote the optimal alignment matrix between $\widetilde{\boldsymbol{F}}$ and $\boldsymbol{F}_\star$ as $\widetilde{\boldsymbol{Q}}$, whose existence is guaranteed by Lemma 9. Denoting $\mathcal{P}_B(\widetilde{\boldsymbol{F}}) = [\boldsymbol{L}^\top, \boldsymbol{R}^\top]^\top$, by the definition of $\mathrm{dist}(\mathcal{P}_B(\widetilde{\boldsymbol{F}}), \boldsymbol{F}_\star)$, we know that

$$\mathrm{dist}^2(\mathcal{P}_B(\widetilde{\boldsymbol{F}}), \boldsymbol{F}_\star) \le \sum_{i=1}^{n_1} \left\| \boldsymbol{L}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot} \right\|_2^2 + \sum_{j=1}^{n_2} \left\| \boldsymbol{R}_{j,\cdot}\widetilde{\boldsymbol{Q}}^{-\top}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot} \right\|_2^2. \tag{66}$$

Recall that the condition $\mathrm{dist}(\widetilde{\boldsymbol{F}}, \boldsymbol{F}_\star) \le \epsilon\sigma_r(\boldsymbol{X}_\star)$ implies

$$\left\| (\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{Q}} - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{-1/2} \right\|_{\mathsf{op}} \vee \left\| (\widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{Q}}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2} \right\|_{\mathsf{op}} \le \epsilon,$$

which, together with $\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2} = \boldsymbol{V}_\star$, further implies that

$$\begin{aligned}
\left\| \widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top \right\|_2 &\le \left\| \widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} \right\|_2 \left\| \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{Q}}^{-\top}\boldsymbol{\Sigma}_\star^{-1/2} \right\|_{\mathsf{op}} \\
&\le \left\| \widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} \right\|_2 \left(\|\boldsymbol{V}_\star\|_{\mathsf{op}} + \left\| (\widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{Q}}^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{-1/2} \right\|_{\mathsf{op}} \right) \le (1+\epsilon) \left\| \widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} \right\|_2.
\end{aligned}$$

In addition, the μ-incoherence of $\boldsymbol{X}_\star$ yields

$$\sqrt{n_1} \left\| (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot} \right\|_2 \le \sqrt{n_1}\|\boldsymbol{U}_\star\|_{2,\infty}\|\boldsymbol{\Sigma}_\star\|_{\mathsf{op}} \le \sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star) \le \frac{B}{1+\epsilon},$$

where the last inequality follows from the choice of B. Take the above two relations collectively to reach

$$\frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2} \ge \frac{\left\| (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot} \right\|_2}{\left\| \widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} \right\|_2}.$$

We claim that performing the following projection yields a contraction on each row; see also [ZL16, Lemma 11].

Claim 5. *For vectors $\boldsymbol{u}, \boldsymbol{u}_\star \in \mathbb{R}^n$ and $\lambda \ge \|\boldsymbol{u}_\star\|_2/\|\boldsymbol{u}\|_2$, it holds that*

$$\|(1 \wedge \lambda)\boldsymbol{u} - \boldsymbol{u}_\star\|_2 \le \|\boldsymbol{u} - \boldsymbol{u}_\star\|_2.$$

Apply Claim 5 with $\boldsymbol{u} := \widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2}$, $\boldsymbol{u}_\star := (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot}$, and $\lambda := B/(\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2)$ to obtain

$$\begin{aligned}
\left\| \boldsymbol{L}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot} \right\|_2^2 &= \left\| \left(1 \wedge \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}\right) \widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot} \right\|_2^2 \\
&\le \left\| \widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot} \right\|_2^2.
\end{aligned}$$

Following a similar argument for $\boldsymbol{R}$, and plugging them back to (66), we conclude that

$$\mathrm{dist}^2(\mathcal{P}_B(\widetilde{\boldsymbol{F}}), \boldsymbol{F}_\star) \le \sum_{i=1}^{n_1} \left\| \widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{Q}}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{1/2})_{i,\cdot} \right\|_2^2 + \sum_{j=1}^{n_2} \left\| \widetilde{\boldsymbol{R}}_{j,\cdot}\widetilde{\boldsymbol{Q}}^{-\top}\boldsymbol{\Sigma}_\star^{1/2} - (\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{1/2})_{j,\cdot} \right\|_2^2 = \mathrm{dist}^2(\widetilde{\boldsymbol{F}}, \boldsymbol{F}_\star).$$

We move on to the incoherence condition. For any $1 \le i \le n_1$, one has

$$\begin{aligned}
\|\boldsymbol{L}_{i,\cdot}\boldsymbol{R}^\top\|_2^2 = \sum_{j=1}^{n_2}\langle \boldsymbol{L}_{i,\cdot}, \boldsymbol{R}_{j,\cdot}\rangle^2 &= \sum_{j=1}^{n_2} \left(1 \wedge \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}\right)^2 \langle \widetilde{\boldsymbol{L}}_{i,\cdot}, \widetilde{\boldsymbol{R}}_{j,\cdot}\rangle^2 \left(1 \wedge \frac{B}{\sqrt{n_2}\|\widetilde{\boldsymbol{R}}_{j,\cdot}\widetilde{\boldsymbol{L}}^\top\|_2}\right)^2 \\
&\overset{\text{(i)}}{\le} \left(1 \wedge \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}\right)^2 \sum_{j=1}^{n_2}\langle \widetilde{\boldsymbol{L}}_{i,\cdot}, \widetilde{\boldsymbol{R}}_{j,\cdot}\rangle^2 = \left(1 \wedge \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}\right)^2 \|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2^2
\end{aligned}$$

$$\overset{\text{(ii)}}{\leq} \frac{B^2}{n_1}.$$

where (i) follows from $1 \wedge \frac{B}{\sqrt{n_2}\|\widetilde{\boldsymbol{R}}_{j,\cdot}\widetilde{\boldsymbol{L}}^\top\|_2} \leq 1$, and (ii) follows from $1 \wedge \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2} \leq \frac{B}{\sqrt{n_1}\|\widetilde{\boldsymbol{L}}_{i,\cdot}\widetilde{\boldsymbol{R}}^\top\|_2}$. Similarly, one has $\|\boldsymbol{R}_{j,\cdot}\boldsymbol{L}^\top\|_2^2 \leq B^2/n_2$. Combining these two bounds completes the proof.

Proof of Claim 5. When $\lambda > 1$, the claim holds as an identity. Otherwise $\lambda \leq 1$. Denote $h(\bar{\lambda}) \coloneqq \|\bar{\lambda}\boldsymbol{u} - \boldsymbol{u}_\star\|_2^2$. Calculate its derivative to conclude that $h(\bar{\lambda})$ is monotonically increasing when $\bar{\lambda} \geq \lambda_\star \coloneqq \langle \boldsymbol{u}, \boldsymbol{u}_\star\rangle/\|\boldsymbol{u}\|_2^2$. Note that $\lambda \geq \|\boldsymbol{u}_\star\|_2/\|\boldsymbol{u}\|_2 \geq \lambda_\star$, thus $h(\lambda) \leq h(1)$, i.e. the claim holds. $\square$

E.2 Proof of Lemma 7

We first record two useful lemmas regarding the projector $\mathcal{P}_\Omega(\cdot)$.

Lemma 21 ([ZL16, Lemma 10])**.** *Suppose that $\boldsymbol{X}_\star$ is μ-incoherent, and $p \gtrsim \mu r\log(n_1 \vee n_2)/(n_1 \wedge n_2)$. With overwhelming probability, one has*

$$\begin{aligned}
&\left|\left\langle (p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}_\star\boldsymbol{R}_A^\top + \boldsymbol{L}_A\boldsymbol{R}_\star^\top), \boldsymbol{L}_\star\boldsymbol{R}_B^\top + \boldsymbol{L}_B\boldsymbol{R}_\star^\top\right\rangle\right| \\
&\qquad \leq C_1\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{L}_\star\boldsymbol{R}_A^\top + \boldsymbol{L}_A\boldsymbol{R}_\star^\top\|_{\mathsf{F}}\|\boldsymbol{L}_\star\boldsymbol{R}_B^\top + \boldsymbol{L}_B\boldsymbol{R}_\star^\top\|_{\mathsf{F}},
\end{aligned}$$

simultaneously for all $\boldsymbol{L}_A, \boldsymbol{L}_B \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{R}_A, \boldsymbol{R}_B \in \mathbb{R}^{n_2\times r}$, where $C_1 > 0$ is some universal constant.

Lemma 22 ([CL19, Lemma 8], [CLL20, Lemma 12])**.** *Suppose that $p \gtrsim \log(n_1 \vee n_2)/(n_1 \wedge n_2)$. With overwhelming probability, one has*

$$\begin{aligned}
&\left|\left\langle (p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}_A\boldsymbol{R}_A^\top), \boldsymbol{L}_B\boldsymbol{R}_B^\top\right\rangle\right| \\
&\qquad \leq C_2\sqrt{\frac{n_1\vee n_2}{p}}\left(\|\boldsymbol{L}_A\|_{\mathsf{F}}\|\boldsymbol{L}_B\|_{2,\infty} \wedge \|\boldsymbol{L}_A\|_{2,\infty}\|\boldsymbol{L}_B\|_{\mathsf{F}}\right)\left(\|\boldsymbol{R}_A\|_{\mathsf{F}}\|\boldsymbol{R}_B\|_{2,\infty} \wedge \|\boldsymbol{R}_A\|_{2,\infty}\|\boldsymbol{R}_B\|_{\mathsf{F}}\right),
\end{aligned}$$

simultaneously for all $\boldsymbol{L}_A, \boldsymbol{L}_B \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{R}_A, \boldsymbol{R}_B \in \mathbb{R}^{n_2\times r}$, where $C_2 > 0$ is some universal constant.

In view of the above two lemmas, define the event $\mathcal{E}$ as the intersection of the events that the bounds in Lemmas 21 and 22 hold, which happens with overwhelming probability. The rest of the proof is then performed under the event that $\mathcal{E}$ holds.

By the condition $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \leq 0.02\sigma_r(\boldsymbol{X}_\star)$ and Lemma 9, one knows that $\boldsymbol{Q}_t$, the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists. Therefore, for notational convenience, we denote $\boldsymbol{L} \coloneqq \boldsymbol{L}_t\boldsymbol{Q}_t$, $\boldsymbol{R} \coloneqq \boldsymbol{R}_t\boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L \coloneqq \boldsymbol{L} - \boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R \coloneqq \boldsymbol{R} - \boldsymbol{R}_\star$, and $\epsilon \coloneqq 0.02$. In addition, denote $\widetilde{\boldsymbol{F}}_{t+1}$ as the update before projection as

$$\widetilde{\boldsymbol{F}}_{t+1} \coloneqq \begin{bmatrix}\widetilde{\boldsymbol{L}}_{t+1}\\ \widetilde{\boldsymbol{R}}_{t+1}\end{bmatrix} = \begin{bmatrix}\boldsymbol{L}_t - \eta p^{-1}\mathcal{P}_\Omega(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)\boldsymbol{R}_t(\boldsymbol{R}_t^\top\boldsymbol{R}_t)^{-1}\\ \boldsymbol{R}_t - \eta p^{-1}\mathcal{P}_\Omega(\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star)^\top\boldsymbol{L}_t(\boldsymbol{L}_t^\top\boldsymbol{L}_t)^{-1}\end{bmatrix},$$

and therefore $\boldsymbol{F}_{t+1} = \mathcal{P}_B(\widetilde{\boldsymbol{F}}_{t+1})$. Note that in view of Lemma 6, it suffices to prove the following relation

$$\mathrm{dist}(\widetilde{\boldsymbol{F}}_{t+1}, \boldsymbol{F}_\star) \leq (1 - 0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star). \tag{67}$$

The conclusion $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top - \boldsymbol{X}_\star\|_{\mathsf{F}} \leq 1.5\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star)$ is a simple consequence of Lemma 13; see (48) for a detailed argument. In what follows, we concentrate on proving (67).

To begin with, we list a few easy consequences under the assumed conditions.

Claim 6. *Under conditions $\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star) \leq \epsilon\sigma_r(\boldsymbol{X}_\star)$ and $\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{R}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{L}^\top\|_{2,\infty} \leq C_B\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star)$, one has*

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \vee \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}} \leq \epsilon; \tag{68a}$$

$$\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \leq \frac{1}{1-\epsilon}; \tag{68b}$$

$$\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} \le \frac{1}{(1-\epsilon)^2}; \tag{68c}$$

$$\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \frac{C_B}{1-\epsilon}\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star); \tag{68d}$$

$$\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty} \le \frac{C_B\kappa}{1-\epsilon}\sqrt{\mu r}; \tag{68e}$$

$$\sqrt{n_1}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \le \left(1+\frac{C_B}{1-\epsilon}\right)\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star). \tag{68f}$$

Now we are ready to embark on the proof of (67). By the definition of $\mathrm{dist}(\widetilde{\boldsymbol{F}}_{t+1},\boldsymbol{F}_\star)$, one has

$$\mathrm{dist}^2(\widetilde{\boldsymbol{F}}_{t+1},\boldsymbol{F}_\star) \le \left\|(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\widetilde{\boldsymbol{R}}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2, \tag{69}$$

where we recall that $\boldsymbol{Q}_t$ is the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$. Plug in the update rule (26) and the decomposition $\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star=\boldsymbol{\Delta}_L\boldsymbol{R}^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$ to obtain

$$\begin{aligned}
(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} &= \left(\boldsymbol{L}-\eta p^{-1}\mathcal{P}_\Omega(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}-\boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2}\\
&= \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}-\eta(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}-\eta(p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\\
&= (1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}-\eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}-\eta(p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.
\end{aligned}$$

This allows us to expand the first square in (69) as

$$\begin{aligned}
\left\|(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &= \underbrace{\left\|(1-\eta)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}-\eta\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{P}_1}\\
&\quad -2\eta(1-\eta)\underbrace{\mathrm{tr}\left((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)}_{\mathfrak{P}_2}\\
&\quad +2\eta^2\underbrace{\mathrm{tr}\left((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\right)}_{\mathfrak{P}_3}\\
&\quad +\eta^2\underbrace{\left\|(p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{P}_4}.
\end{aligned}$$

In the sequel, we shall control the four terms separately, of which $\mathfrak{P}_1$ is the main term, and $\mathfrak{P}_2,\mathfrak{P}_3$ and $\mathfrak{P}_4$ are perturbation terms.

1. Notice that the main term $\mathfrak{P}_1$ has already been controlled in (46) under the condition (68a). It obeys

$$\mathfrak{P}_1 \le \left((1-\eta)^2+\frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

2. For the second term $\mathfrak{P}_2$, decompose $\boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star=\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}\boldsymbol{\Delta}_R^\top$ and apply the triangle inequality to obtain

$$\begin{aligned}
|\mathfrak{P}_2| &= \left|\mathrm{tr}\left((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)\right|\\
&\le \underbrace{\left|\mathrm{tr}\left((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top)\boldsymbol{R}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)\right|}_{\mathfrak{P}_{2,1}}\\
&\quad +\underbrace{\left|\mathrm{tr}\left((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top)\boldsymbol{\Delta}_R(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)\right|}_{\mathfrak{P}_{2,2}}
\end{aligned}$$

$$+\underbrace{\left|\operatorname{tr}\left((p^{-1}\mathcal{P}_\Omega-\mathcal{I})(\boldsymbol{L}\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star\boldsymbol{\Delta}_L^\top\right)\right|}_{\mathfrak{P}_{2,3}}.$$

For the first term $\mathfrak{P}_{2,1}$, under the event $\mathcal{E}$, we can invoke Lemma 21 to obtain

$$\begin{aligned}\mathfrak{P}_{2,1}&\le C_1\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top\|_{\mathsf{F}}\left\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}_\star^\top\right\|_{\mathsf{F}}\\&\le C_1\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}},\end{aligned}$$

where the second line follows from the relation $\|\boldsymbol{A}\boldsymbol{B}\|_{\mathsf{F}}\le\|\boldsymbol{A}\|_{\mathsf{op}}\|\boldsymbol{B}\|_{\mathsf{F}}$. Use the condition (68c) to obtain

$$\mathfrak{P}_{2,1}\le\frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

Regarding the remaining terms $\mathfrak{P}_{2,2}$ and $\mathfrak{P}_{2,3}$, our main hammer is Lemma 22. Invoking Lemma 22 under the event $\mathcal{E}$ with $\boldsymbol{L}_A\coloneqq\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}$, $\boldsymbol{R}_A\coloneqq\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2}$, $\boldsymbol{L}_B\coloneqq\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}$, and $\boldsymbol{R}_B\coloneqq\boldsymbol{\Delta}_R(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}$, we arrive at

$$\begin{aligned}\mathfrak{P}_{2,2}&\le C_2\sqrt{\frac{n_1\vee n_2}{p}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{\Delta}_R(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}\\&\le C_2\sqrt{\frac{n_1\vee n_2}{p}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{F}}\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}.\end{aligned}$$

Similarly, with the help of Lemma 22, one has

$$\mathfrak{P}_{2,3}\le C_2\sqrt{\frac{n_1\vee n_2}{p}}\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2}\|_{2,\infty}\left\|\boldsymbol{\Sigma}_\star^{1/2}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}.$$

Utilizing the consequences in Claim 6, we arrive at

$$\begin{aligned}\mathfrak{P}_{2,2}&\le\frac{C_2\kappa}{(1-\epsilon)^2}\left(1+\frac{C_B}{1-\epsilon}\right)\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}};\\\mathfrak{P}_{2,3}&\le\frac{C_2C_B^2\kappa^2}{(1-\epsilon)^4}\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.\end{aligned}$$

We then combine the bounds for $\mathfrak{P}_{2,1},\mathfrak{P}_{2,2}$ and $\mathfrak{P}_{2,3}$ to see

$$\begin{aligned}\mathfrak{P}_2&\le\frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\\&\quad+\frac{C_2\kappa}{(1-\epsilon)^2}\left(1+\frac{C_B}{1-\epsilon}+\frac{C_B^2\kappa}{(1-\epsilon)^2}\right)\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\\&=\delta_1\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\delta_2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\\&\le(\delta_1+\frac{\delta_2}{2})\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\frac{\delta_2}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2,\end{aligned}$$

where we denote

$$\delta_1\coloneqq\frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}},\quad\text{and}\quad\delta_2\coloneqq\frac{C_2\kappa}{(1-\epsilon)^2}\left(1+\frac{C_B}{1-\epsilon}+\frac{C_B^2\kappa}{(1-\epsilon)^2}\right)\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}.$$

3. Following a similar argument for controlling $\mathfrak{P}_2$ (i.e. repeatedly using Lemmas 21 and 22), we can obtain the following bounds for $\mathfrak{P}_3$ and $\mathfrak{P}_4$, whose proof are deferred to the end of this section.

Claim 7. *Under the event $\mathcal{E}$, one has*

$$\mathfrak{P}_3 \le \frac{\delta_2}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + (\delta_1 + \frac{\delta_2}{2})\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2;$$
$$\mathfrak{P}_4 \le \delta_1(\delta_1+\delta_2)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \delta_2(\delta_1+\delta_2)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

Taking the bounds for $\mathfrak{P}_1, \mathfrak{P}_2, \mathfrak{P}_3$ and $\mathfrak{P}_4$ collectively yields

$$\begin{aligned}
\left\|(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \le{} & \left((1-\eta)^2 + \frac{2\epsilon}{1-\epsilon}\eta(1-\eta)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 \\
& + \eta(1-\eta)\left((2\delta_1+\delta_2)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \delta_2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\
& + \eta^2\left(\delta_2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + (2\delta_1+\delta_2)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\
& + \eta^2\left(\delta_1(\delta_1+\delta_2)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \delta_2(\delta_1+\delta_2)\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right).
\end{aligned}$$

A similar upper bound holds for the second square in (69). As a result, we reach the conclusion that

$$\left\|(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\widetilde{\boldsymbol{R}}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \le \rho^2(\eta;\epsilon,\delta_1,\delta_2)\,\mathrm{dist}^2(\boldsymbol{F}_t,\boldsymbol{F}_\star),$$

where the contraction rate $\rho^2(\eta;\epsilon,\delta_1,\delta_2)$ is given by

$$\rho^2(\eta;\epsilon,\delta_1,\delta_2) \coloneqq (1-\eta)^2 + \left(\frac{2\epsilon}{1-\epsilon} + 2(\delta_1+\delta_2)\right)\eta(1-\eta) + \left(\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2} + 2(\delta_1+\delta_2) + (\delta_1+\delta_2)^2\right)\eta^2.$$

As long as $p \ge C(\mu r\kappa^4 \vee \log(n_1 \vee n_2))\mu r/(n_1 \wedge n_2)$ for some sufficiently large constant C, one has $\delta_1+\delta_2 \le 0.1$ under the setting $\epsilon = 0.02$. When $0 < \eta \le 2/3$, one further has $\rho(\eta;\epsilon,\delta_1,\delta_2) \le 1 - 0.6\eta$. Thus we conclude that

$$\begin{aligned}
\mathrm{dist}(\widetilde{\boldsymbol{F}}_{t+1},\boldsymbol{F}_\star) &\le \sqrt{\left\|(\widetilde{\boldsymbol{L}}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\widetilde{\boldsymbol{R}}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2} \\
&\le (1-0.6\eta)\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star),
\end{aligned}$$

which is exactly the upper bound we are after; see (67). This finishes the proof.

Proof of Claim 6. First, repeating the derivation for (45) obtains (68a). Second, take the condition (68a) and Lemma 12 together to obtain (68b) and (68c). Third, take the incoherence condition $\sqrt{n_1}\|\boldsymbol{L}\boldsymbol{R}^\top\|_{2,\infty} \vee \sqrt{n_2}\|\boldsymbol{R}\boldsymbol{L}^\top\|_{2,\infty} \le C_B\sqrt{\mu r}\sigma_1(\boldsymbol{X}_\star)$ together with the relations

$$\begin{aligned}
\|\boldsymbol{L}\boldsymbol{R}^\top\|_{2,\infty} &\ge \sigma_r(\boldsymbol{R}\boldsymbol{\Sigma}_\star^{-1/2})\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \\
&\ge \left(\sigma_r(\boldsymbol{R}_\star\boldsymbol{\Sigma}_\star^{-1/2}) - \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}\right)\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \\
&\ge (1-\epsilon)\|\boldsymbol{L}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}; \\
\|\boldsymbol{R}\boldsymbol{L}^\top\|_{2,\infty} &\ge \sigma_r(\boldsymbol{L}\boldsymbol{\Sigma}_\star^{-1/2})\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \\
&\ge \left(\sigma_r(\boldsymbol{L}_\star\boldsymbol{\Sigma}_\star^{-1/2}) - \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}\right)\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty} \\
&\ge (1-\epsilon)\|\boldsymbol{R}\boldsymbol{\Sigma}_\star^{1/2}\|_{2,\infty}
\end{aligned}$$

to obtain (68d) and (68e). Finally, apply the triangle inequality together with incoherence assumption to obtain (68f). ☐

Proof of Claim 7. We start with the term $\mathfrak{P}_3$, for which we have

$$|\mathfrak{P}_3| \le \underbrace{\left|\mathrm{tr}\left((p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{\Delta}_R\boldsymbol{L}_\star^\top\right)\right|}_{\mathfrak{P}_{3,1}}$$

$$+\underbrace{\left|\mathrm{tr}\left((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{\Delta}_{L}\boldsymbol{R}^{\top})\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_{\star}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{R}^{\top}\boldsymbol{\Delta}_{R}\boldsymbol{L}_{\star}^{\top}\right)\right|}_{\mathfrak{P}_{3,2}}.$$

Invoke Lemma 21 to bound $\mathfrak{P}_{3,1}$ as

$$\begin{aligned}\mathfrak{P}_{3,1}&\le C_1\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{L}_{\star}\boldsymbol{\Delta}_{R}^{\top}\|_{\mathsf{F}}\left\|\boldsymbol{L}_{\star}\boldsymbol{\Delta}_{R}^{\top}\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_{\star}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{R}^{\top}\right\|_{\mathsf{F}}\\&\le C_1\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_{R}\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}^2\left\|\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_{\star}^{1/2}\right\|_{\mathsf{op}}^2.\end{aligned}$$

The condition (68b) allows us to obtain a simplified bound

$$\mathfrak{P}_{3,1}\le\frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_{R}\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}^2.$$

In regard to $\mathfrak{P}_{3,2}$, we apply Lemma 22 with $\boldsymbol{L}_A\coloneqq\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_{\star}^{1/2}$, $\boldsymbol{R}_A\coloneqq\boldsymbol{R}\boldsymbol{\Sigma}_{\star}^{-1/2}$, $\boldsymbol{L}_B\coloneqq\boldsymbol{L}_{\star}\boldsymbol{\Sigma}_{\star}^{-1/2}$, and $\boldsymbol{R}_B\coloneqq\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_{\star}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{R}^{\top}\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_{\star}^{1/2}$ to see

$$\begin{aligned}\mathfrak{P}_{3,2}&\le C_2\sqrt{\frac{n_1\vee n_2}{p}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}\|\boldsymbol{L}_{\star}\boldsymbol{\Sigma}_{\star}^{-1/2}\|_{2,\infty}\|\boldsymbol{R}\boldsymbol{\Sigma}_{\star}^{-1/2}\|_{2,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_{\star}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{R}^{\top}\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_{\star}^{1/2}\right\|_{\mathsf{F}}\\&\le C_2\sqrt{\frac{n_1\vee n_2}{p}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}\|\boldsymbol{L}_{\star}\boldsymbol{\Sigma}_{\star}^{-1/2}\|_{2,\infty}\|\boldsymbol{R}\boldsymbol{\Sigma}_{\star}^{-1/2}\|_{2,\infty}\left\|\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_{\star}^{1/2}\right\|_{\mathsf{op}}^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}.\end{aligned}$$

Again, use the consequences in Claim 6 to reach

$$\begin{aligned}\mathfrak{P}_{3,2}&\le C_2\sqrt{\frac{n_1\vee n_2}{p}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}\sqrt{\frac{\mu r}{n_1}}\frac{C_B\kappa}{1-\epsilon}\sqrt{\frac{\mu r}{n_2}}\frac{1}{(1-\epsilon)^2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}\\&=\frac{C_2C_B\kappa}{(1-\epsilon)^3}\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}.\end{aligned}$$

Combine the bounds of $\mathfrak{P}_{3,1}$ and $\mathfrak{P}_{3,2}$ to reach

$$\begin{aligned}\mathfrak{P}_3&\le\frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r\log(n_1\vee n_2)}{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}^2\\&\quad+\frac{C_2C_B\kappa}{(1-\epsilon)^3}\frac{\mu r}{\sqrt{p(n_1\wedge n_2)}}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}\\&\le\delta_1\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}^2+\delta_2\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}\\&\le\frac{\delta_2}{2}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}^2+(\delta_1+\frac{\delta_2}{2})\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_{\star}^{1/2}\|_{\mathsf{F}}^2.\end{aligned}$$

Moving on to the term $\mathfrak{P}_4$, we have

$$\begin{aligned}\sqrt{\mathfrak{P}_4}&=\left\|(p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{L}\boldsymbol{R}^{\top}-\boldsymbol{X}_{\star})\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_{\star}^{1/2}\right\|_{\mathsf{F}}\\&\le\underbrace{\left|\mathrm{tr}\left((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_{\star}^{\top})\boldsymbol{R}_{\star}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_{\star}^{1/2}\widetilde{\boldsymbol{L}}^{\top}\right)\right|}_{\mathfrak{P}_{4,1}}\\&\quad+\underbrace{\left|\mathrm{tr}\left((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{\Delta}_L\boldsymbol{R}_{\star}^{\top})\boldsymbol{\Delta}_R(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_{\star}^{1/2}\widetilde{\boldsymbol{L}}^{\top}\right)\right|}_{\mathfrak{P}_{4,2}}\\&\quad+\underbrace{\left|\mathrm{tr}\left((p^{-1}\mathcal{P}_{\Omega}-\mathcal{I})(\boldsymbol{L}\boldsymbol{\Delta}_R^{\top})\boldsymbol{R}(\boldsymbol{R}^{\top}\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_{\star}^{1/2}\widetilde{\boldsymbol{L}}^{\top}\right)\right|}_{\mathfrak{P}_{4,3}},\end{aligned}$$

where we have used the variational representation of the Frobenius norm for some $\widetilde{\boldsymbol{L}} \in \mathbb{R}^{n_1 \times r}$ obeying $\|\widetilde{\boldsymbol{L}}\|_{\mathsf{F}} = 1$. Note that the decomposition of $\sqrt{\mathfrak{P}_4}$ is extremely similar to that of $\mathfrak{P}_2$. Therefore we can follow a similar argument (i.e. applying Lemmas 21 and 22) to control these terms as

$$\begin{aligned}
\mathfrak{P}_{4,1} &\le \frac{C_1}{(1-\epsilon)^2}\sqrt{\frac{\mu r \log(n_1 \vee n_2)}{p(n_1 \wedge n_2)}}\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}};\\
\mathfrak{P}_{4,2} &\le \frac{C_2\kappa}{(1-\epsilon)^2}\left(1+\frac{C_B}{1-\epsilon}\right)\frac{\mu r}{\sqrt{p(n_1 \wedge n_2)}}\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}};\\
\mathfrak{P}_{4,3} &\le \frac{C_2 C_B^2 \kappa^2}{(1-\epsilon)^4}\frac{\mu r}{\sqrt{p(n_1 \wedge n_2)}}\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.
\end{aligned}$$

For conciseness, we omit the details for bounding each term. Combine them to reach

$$\sqrt{\mathfrak{P}_4} \le \delta_1 \|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \delta_2 \|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.$$

Finally take the square on both sides and use $2ab \le a^2 + b^2$ to obtain the upper bound

$$\mathfrak{P}_4 \le \delta_1(\delta_1+\delta_2)\|\boldsymbol{\Delta}_L \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \delta_2(\delta_1+\delta_2)\|\boldsymbol{\Delta}_R \boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2.$$

□

E.3 Proof of Lemma 8

We start by recording a useful lemma below.

Lemma 23 ([Che15, Lemma 2], [CLL20, Lemma 4]). *For any fixed $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$, with overwhelming probability, one has*

$$\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{X})\right\|_{\mathsf{op}} \le C_0 \frac{\log(n_1 \vee n_2)}{p}\|\boldsymbol{X}\|_\infty + C_0\sqrt{\frac{\log(n_1 \vee n_2)}{p}}(\|\boldsymbol{X}\|_{2,\infty} \vee \|\boldsymbol{X}^\top\|_{2,\infty}),$$

where $C_0 > 0$ is some universal constant that does not depend on $\boldsymbol{X}$.

In view of Lemma 11, one has

$$\mathrm{dist}(\widetilde{\boldsymbol{F}}_0, \boldsymbol{F}_\star) \le \sqrt{\sqrt{2}+1}\left\|\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top - \boldsymbol{X}_\star\right\|_{\mathsf{F}} \le \sqrt{(\sqrt{2}+1)2r}\left\|\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top - \boldsymbol{X}_\star\right\|_{\mathsf{op}}, \tag{70}$$

where the last relation uses the fact that $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top - \boldsymbol{X}_\star$ has rank at most $2r$. Applying the triangle inequality, we obtain

$$\begin{aligned}
\left\|\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top - \boldsymbol{X}_\star\right\|_{\mathsf{op}} &\le \left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star) - \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top\right\|_{\mathsf{op}} + \left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star) - \boldsymbol{X}_\star\right\|_{\mathsf{op}}\\
&\le 2\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{X}_\star)\right\|_{\mathsf{op}}.
\end{aligned} \tag{71}$$

Here the second inequality hinges on the fact that $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$ is the best rank-r approximation to $p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star)$, i.e.

$$\left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star) - \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top\right\|_{\mathsf{op}} \le \left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star) - \boldsymbol{X}_\star\right\|_{\mathsf{op}}.$$

Combining (70) and (71) yields

$$\mathrm{dist}(\widetilde{\boldsymbol{F}}_0, \boldsymbol{F}_\star) \le 2\sqrt{(\sqrt{2}+1)2r}\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{X}_\star)\right\|_{\mathsf{op}} \le 5\sqrt{r}\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{X}_\star)\right\|_{\mathsf{op}}.$$

It then boils down to controlling $\left\|p^{-1}\mathcal{P}_\Omega(\boldsymbol{X}_\star) - \boldsymbol{X}_\star\right\|_{\mathsf{op}}$, which is readily supplied by Lemma 23 as

$$\left\|(p^{-1}\mathcal{P}_\Omega - \mathcal{I})(\boldsymbol{X}_\star)\right\|_{\mathsf{op}} \le C_0 \frac{\log(n_1 \vee n_2)}{p}\|\boldsymbol{X}_\star\|_\infty + C_0\sqrt{\frac{\log(n_1 \vee n_2)}{p}}(\|\boldsymbol{X}_\star\|_{2,\infty} \vee \|\boldsymbol{X}_\star^\top\|_{2,\infty}),$$

which holds with overwhelming probability. The proof is finished by plugging the following bounds from incoherence assumption of $\boldsymbol{X}_\star$:

$$\begin{aligned}
\|\boldsymbol{X}_\star\|_\infty &\le \|\boldsymbol{U}_\star\|_{2,\infty}\|\boldsymbol{\Sigma}_\star\|_{\mathsf{op}}\|\boldsymbol{V}_\star\|_{2,\infty} \le \frac{\mu r}{\sqrt{n_1 n_2}}\kappa\sigma_r(\boldsymbol{X}_\star);\\
\|\boldsymbol{X}_\star\|_{2,\infty} &\le \|\boldsymbol{U}_\star\|_{2,\infty}\|\boldsymbol{\Sigma}_\star\|_{\mathsf{op}}\|\boldsymbol{V}_\star\|_{\mathsf{op}} \le \sqrt{\frac{\mu r}{n_1}}\kappa\sigma_r(\boldsymbol{X}_\star);\\
\|\boldsymbol{X}_\star^\top\|_{2,\infty} &\le \|\boldsymbol{U}_\star\|_{\mathsf{op}}\|\boldsymbol{\Sigma}_\star\|_{\mathsf{op}}\|\boldsymbol{V}_\star\|_{2,\infty} \le \sqrt{\frac{\mu r}{n_2}}\kappa\sigma_r(\boldsymbol{X}_\star).
\end{aligned}$$

F Proof for General Loss Functions

We first present a useful property of restricted smooth and convex functions.

Lemma 24. *Suppose that $f:\mathbb{R}^{n_1\times n_2}\mapsto\mathbb{R}$ is rank-$2r$ restricted L-smooth and rank-$2r$ restricted convex. Then for any $\boldsymbol{X}_1,\boldsymbol{X}_2\in\mathbb{R}^{n_1\times n_2}$ of rank at most r, one has*

$$\langle\nabla f(\boldsymbol{X}_1)-\nabla f(\boldsymbol{X}_2),\boldsymbol{X}_1-\boldsymbol{X}_2\rangle \ge \frac{1}{L}\|\nabla f(\boldsymbol{X}_1)-\nabla f(\boldsymbol{X}_2)\|_{\mathsf{F},r}^2.$$

Proof. Since $f(\cdot)$ is rank-$2r$ restricted L-smooth and convex, it holds for any $\bar{\boldsymbol{X}}\in\mathbb{R}^{n_1\times n_2}$ with rank at most $2r$ that

$$f(\boldsymbol{X}_1)+\langle\nabla f(\boldsymbol{X}_1),\bar{\boldsymbol{X}}-\boldsymbol{X}_1\rangle \le f(\bar{\boldsymbol{X}}) \le f(\boldsymbol{X}_2)+\langle\nabla f(\boldsymbol{X}_2),\bar{\boldsymbol{X}}-\boldsymbol{X}_2\rangle+\frac{L}{2}\|\bar{\boldsymbol{X}}-\boldsymbol{X}_2\|_{\mathsf{F}}^2.$$

Reorganize the terms to yield

$$f(\boldsymbol{X}_1)+\langle\nabla f(\boldsymbol{X}_1),\boldsymbol{X}_2-\boldsymbol{X}_1\rangle \le f(\boldsymbol{X}_2)+\langle\nabla f(\boldsymbol{X}_2)-\nabla f(\boldsymbol{X}_1),\bar{\boldsymbol{X}}-\boldsymbol{X}_2\rangle+\frac{L}{2}\|\bar{\boldsymbol{X}}-\boldsymbol{X}_2\|_{\mathsf{F}}^2.$$

Take $\bar{\boldsymbol{X}}=\boldsymbol{X}_2-\frac{1}{L}\mathcal{P}_r(\nabla f(\boldsymbol{X}_2)-\nabla f(\boldsymbol{X}_1))$, whose rank is at most $2r$, to see

$$f(\boldsymbol{X}_1)+\langle\nabla f(\boldsymbol{X}_1),\boldsymbol{X}_2-\boldsymbol{X}_1\rangle+\frac{1}{2L}\left\|\nabla f(\boldsymbol{X}_2)-\nabla f(\boldsymbol{X}_1)\right\|_{\mathsf{F},r}^2 \le f(\boldsymbol{X}_2).$$

We can further switch the roles of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ to obtain

$$f(\boldsymbol{X}_2)+\langle\nabla f(\boldsymbol{X}_2),\boldsymbol{X}_1-\boldsymbol{X}_2\rangle+\frac{1}{2L}\left\|\nabla f(\boldsymbol{X}_2)-\nabla f(\boldsymbol{X}_1)\right\|_{\mathsf{F},r}^2 \le f(\boldsymbol{X}_1).$$

Adding the above two inequalities yields the desired bound. ◻

F.1 Proof of Theorem 4

Suppose that the t-th iterate $\boldsymbol{F}_t$ obeys the condition $\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)\le 0.1\sigma_r(\boldsymbol{X}_\star)/\sqrt{\kappa_f}$. In view of Lemma 9, one knows that $\boldsymbol{Q}_t$, the optimal alignment matrix between $\boldsymbol{F}_t$ and $\boldsymbol{F}_\star$ exists. Therefore, for notational convenience, denote $\boldsymbol{L}\coloneqq\boldsymbol{L}_t\boldsymbol{Q}_t$, $\boldsymbol{R}\coloneqq\boldsymbol{R}_t\boldsymbol{Q}_t^{-\top}$, $\boldsymbol{\Delta}_L\coloneqq\boldsymbol{L}-\boldsymbol{L}_\star$, $\boldsymbol{\Delta}_R\coloneqq\boldsymbol{R}-\boldsymbol{R}_\star$, and $\epsilon\coloneqq 0.1/\sqrt{\kappa_f}$. Similar to the derivation in (45), we have

$$\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}\vee\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}\le\epsilon. \tag{72}$$

The conclusion $\|\boldsymbol{L}_t\boldsymbol{R}_t^\top-\boldsymbol{X}_\star\|_{\mathsf{F}}\le 1.5\,\mathrm{dist}(\boldsymbol{F}_t,\boldsymbol{F}_\star)$ is a simple consequence of Lemma 13; see (48) for a detailed argument. From now on, we focus on proving the distance contraction.

By the definition of $\mathrm{dist}(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)$, one has

$$\mathrm{dist}^2(\boldsymbol{F}_{t+1},\boldsymbol{F}_\star)\le\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2+\left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2. \tag{73}$$

Introduce an auxiliary function

$$f_\mu(\boldsymbol{X}) = f(\boldsymbol{X}) - \frac{\mu}{2}\|\boldsymbol{X}-\boldsymbol{X}_\star\|_{\mathsf{F}}^2,$$

which is rank-$2r$ restricted $(L-\mu)$-smooth and rank-$2r$ restricted convex. Using the `ScaledGD` update rule (27) and the decomposition $\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star = \boldsymbol{\Delta}_L\boldsymbol{R}^\top + \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top$, we obtain

$$\begin{aligned}
(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2} &= \left(\boldsymbol{L} - \eta\nabla f(\boldsymbol{L}\boldsymbol{R}^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= \left(\boldsymbol{L} - \eta\mu(\boldsymbol{L}\boldsymbol{R}^\top - \boldsymbol{X}_\star)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \eta\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1} - \boldsymbol{L}_\star\right)\boldsymbol{\Sigma}_\star^{1/2} \\
&= (1-\eta\mu)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\mu\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \eta\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}.
\end{aligned}$$

As a result, one can expand the first square in (73) as

$$\begin{aligned}
\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &= \underbrace{\left\|(1-\eta\mu)\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2} - \eta\mu\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{G}_1} \\
&\quad - 2\eta(1-\eta\mu)\underbrace{\left\langle\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top - \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top - \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\rangle}_{\mathfrak{G}_2} \\
&\quad - 2\eta(1-\eta\mu)\left\langle\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top + \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\rangle \\
&\quad + 2\eta^2\mu\underbrace{\left\langle\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\rangle}_{\mathfrak{G}_3} \\
&\quad + \eta^2\underbrace{\left\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2}_{\mathfrak{G}_4}.
\end{aligned}$$

In the sequel, we shall bound the four terms separately.

1. Notice that the main term $\mathfrak{G}_1$ has already been controlled in (46) under the condition (72). It obeys

$$\mathfrak{G}_1 \le \left((1-\eta\mu)^2 + \frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2,$$

   as long as $\eta\mu \le 2/3$.

2. For the second term $\mathfrak{G}_2$, note that $\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top - \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top - \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top$ has rank at most r. Hence we can invoke Lemma 15 to obtain

$$\begin{aligned}
|\mathfrak{G}_2| &\le \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\left\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top - \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top - \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\|_{\mathsf{F}} \\
&\le \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\left(\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{V}_\star\right\|_{\mathsf{op}} + \frac{1}{2}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{-1/2}\|_{\mathsf{op}}\right),
\end{aligned}$$

   where the second line uses $\boldsymbol{R}_\star = \boldsymbol{V}_\star\boldsymbol{\Sigma}_\star^{1/2}$. Take the condition (72) and Lemma 12 together to obtain

$$\begin{aligned}
\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}} &\le \frac{1}{1-\epsilon}; \\
\left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2} - \boldsymbol{V}_\star\right\|_{\mathsf{op}} &\le \frac{\sqrt{2}\epsilon}{1-\epsilon}.
\end{aligned}$$

   These consequences further imply that

$$|\mathfrak{G}_2| \le \left(\frac{\sqrt{2}\epsilon}{1-\epsilon} + \frac{\epsilon}{2}\right)\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.$$

3. As above, the third term $\mathfrak{G}_3$ can be similarly bounded as

$$\begin{aligned}|\mathfrak{G}_3| &\le \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r} \left\|\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{R}^\top\right\|_{\mathsf{F}} \\ &\le \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}^2 \\ &\le \frac{1}{(1-\epsilon)^2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}.\end{aligned}$$

4. For the last term $\mathfrak{G}_4$, invoke Lemma 15 to obtain

$$\mathfrak{G}_4 \le \|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2 \left\|\boldsymbol{R}(\boldsymbol{R}^\top\boldsymbol{R})^{-1}\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{op}}^2 \le \frac{1}{(1-\epsilon)^2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2.$$

Taking collectively the bounds for $\mathfrak{G}_1, \mathfrak{G}_2, \mathfrak{G}_3$ and $\mathfrak{G}_4$ yields

$$\begin{aligned}\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 &\le \left((1-\eta\mu)^2+\frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu)\right)\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 \\ &\quad + 2\eta(\frac{\sqrt{2}\epsilon}{1-\epsilon}+\frac{\epsilon}{2})(1-\eta\mu)\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} \\ &\quad - 2\eta(1-\eta\mu)\left\langle \nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top + \frac{1}{2}\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\right\rangle \\ &\quad + \frac{2\eta^2\mu}{(1-\epsilon)^2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}} + \frac{\eta^2}{(1-\epsilon)^2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2.\end{aligned}$$

Similarly, we can obtain the control of $\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$. Combine them together to reach

$$\begin{aligned}&\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t-\boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \\ &\le \left((1-\eta\mu)^2+\frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu)+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2\right)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\ &\quad + 2\eta\left((\frac{\sqrt{2}\epsilon}{1-\epsilon}+\frac{\epsilon}{2})(1-\eta\mu)+\frac{\eta\mu}{(1-\epsilon)^2}\right)\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right) \\ &\quad - 2\eta(1-\eta\mu)\langle\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top+\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\rangle + \frac{2\eta^2}{(1-\epsilon)^2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2 \\ &\le \left((1-\eta\mu)^2+\frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu)+\frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2\right)\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right) \\ &\quad + 2\eta\underbrace{\left((\frac{\sqrt{2}\epsilon}{1-\epsilon}+\frac{\epsilon}{2})(1-\eta\mu)+\frac{\eta\mu}{(1-\epsilon)^2}\right)}_{\mathfrak{C}_1}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right) \\ &\quad - 2\eta\underbrace{\left(\frac{1-\eta\mu}{L-\mu}-\frac{\eta}{(1-\epsilon)^2}\right)}_{\mathfrak{C}_2}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2,\end{aligned}$$

where the last line follows from Lemma 24 (notice that $\nabla f_\mu(\boldsymbol{X}_\star)=\boldsymbol{0}$) as

$$\langle\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{\Delta}_L\boldsymbol{R}_\star^\top+\boldsymbol{L}_\star\boldsymbol{\Delta}_R^\top+\boldsymbol{\Delta}_L\boldsymbol{\Delta}_R^\top\rangle = \langle\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top), \boldsymbol{L}\boldsymbol{R}^\top-\boldsymbol{X}_\star\rangle \ge \frac{1}{L-\mu}\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2.$$

Notice that $\mathfrak{C}_2>0$ as long as $\eta\le(1-\epsilon)^2/L$. Maximizing the quadratic function of $\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}$ yields

$$\mathfrak{C}_1\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right) - \mathfrak{C}_2\|\nabla f_\mu(\boldsymbol{L}\boldsymbol{R}^\top)\|_{\mathsf{F},r}^2 \le \frac{\mathfrak{C}_1^2}{4\mathfrak{C}_2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}+\|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}\right)^2$$

$$\leq \frac{\mathfrak{C}_1^2}{2\mathfrak{C}_2}\left(\|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2\right),$$

where the last inequality holds since $(a+b)^2 \leq 2(a^2+b^2)$. Identify $\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star) = \|\boldsymbol{\Delta}_L\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2 + \|\boldsymbol{\Delta}_R\boldsymbol{\Sigma}_\star^{1/2}\|_{\mathsf{F}}^2$ to obtain

$$\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 \leq \rho^2(\eta; \epsilon, \mu, L)\,\mathrm{dist}^2(\boldsymbol{F}_t, \boldsymbol{F}_\star),$$

where the contraction rate is given by

$$\rho^2(\eta; \epsilon, \mu, L) \coloneqq (1-\eta\mu)^2 + \frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu) + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2 + \frac{\left((\frac{\sqrt{2}\epsilon}{1-\epsilon} + \frac{\epsilon}{2})(1-\eta\mu) + \frac{\eta\mu}{(1-\epsilon)^2}\right)^2}{1-\eta\mu - \frac{\eta(L-\mu)}{(1-\epsilon)^2}}\eta(L-\mu).$$

With $\epsilon = 0.1/\sqrt{\kappa_f}$ and $0 < \eta \leq 0.4/L$, one has $\rho(\eta; \epsilon, \mu, L) \leq 1 - 0.7\eta\mu$. Thus we conclude that

$$\begin{aligned}\mathrm{dist}(\boldsymbol{F}_{t+1}, \boldsymbol{F}_\star) &\leq \sqrt{\left\|(\boldsymbol{L}_{t+1}\boldsymbol{Q}_t - \boldsymbol{L}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2 + \left\|(\boldsymbol{R}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{R}_\star)\boldsymbol{\Sigma}_\star^{1/2}\right\|_{\mathsf{F}}^2} \\ &\leq (1-0.7\eta\mu)\,\mathrm{dist}(\boldsymbol{F}_t, \boldsymbol{F}_\star),\end{aligned}$$

which is the desired claim.

Remark 4. We provide numerical details for the contraction rate. For simplicity, we shall prove $\rho(\eta; \epsilon, \mu, L) \leq 1 - 0.7\eta\mu$ under a stricter condition $\epsilon = 0.02/\sqrt{\kappa_f}$. The stronger result under the condition $\epsilon = 0.1/\sqrt{\kappa_f}$ can be verified through a subtler analysis.

With $\epsilon = 0.02/\sqrt{\kappa_f}$ and $0 < \eta \leq 0.4/L$, one can bound the terms in $\rho^2(\eta; \epsilon, \mu, L)$ as

$$(1-\eta\mu)^2 + \frac{2\epsilon}{1-\epsilon}\eta\mu(1-\eta\mu) + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\eta^2\mu^2 \leq 1 - 1.959\eta\mu + 1.002\eta^2\mu^2; \tag{74}$$

$$\begin{aligned}\frac{\left((\frac{\sqrt{2}\epsilon}{1-\epsilon} + \frac{\epsilon}{2})(1-\eta\mu) + \frac{\eta\mu}{(1-\epsilon)^2}\right)^2}{1-\eta\mu - \frac{\eta(L-\mu)}{(1-\epsilon)^2}}\eta(L-\mu) &\leq \frac{\frac{0.0016}{\kappa_f} + 0.078\eta\mu + 1.005\eta^2\mu^2}{1-1.042\eta L}\eta L \\ &\leq \frac{0.0016\eta\frac{L}{\kappa_f} + 0.4\times(0.078\eta\mu + 1.005\eta^2\mu^2)}{1-0.4\times 1.042} \\ &\leq 0.057\eta\mu + 0.69\eta^2\mu^2,\end{aligned} \tag{75}$$

where the last line uses the definition (28) of κ_f. Putting (74) and (75) together further implies

$$\rho^2(\eta; \epsilon, \mu, L) \leq 1 - 1.9\eta\mu + 1.7\eta^2\mu^2 \leq (1-0.7\eta\mu)^2,$$

as long as $0 < \eta\mu \leq 0.4$.