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

Tian Tong∗ Cong Ma† Yuejie Chi∗
Carnegie Mellon University UC Berkeley Carnegie Mellon University

May 2020; Revised May 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

1 Introduction 2

1.1 Preconditioning helps: scaled gradient descent ............................................. 3
1.2 Related work .................................................................................................... 4
1.3 Paper organization and notation ................................................................. 6

∗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 $X^\star \in \mathbb{R}^{n_1 \times n_2}$ from a set of observations $y = A(X^\star)$, where the operator $A(\cdot)$ models the measurement process. It is natural to minimize the least-squares loss function subject to a rank constraint:

$$\min_{X \in \mathbb{R}^{n_1 \times n_2}} f(X) := \frac{1}{2} \|A(X) - y\|_2^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r, \quad (1)$$
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 $X = LR^\top$ by two low-rank factors $L \in \mathbb{R}^{n_1 \times r}$ and $R \in \mathbb{R}^{n_2 \times r}$ that are more memory-efficient, and then to optimize over the factors instead:

$$\minimize_{L \in \mathbb{R}^{n_1 \times r}, R \in \mathbb{R}^{n_2 \times r}} \mathcal{L}(L, R) := f(LL^\top).$$

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: though 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 $X$, [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 $(L_0, R_0)$, ScaledGD proceeds as follows

$$L_{t+1} = L_t - \eta \nabla_L \mathcal{L}(L_t, R_t)(R_t^\top R_t)^{-1},$$

$$R_{t+1} = R_t - \eta \nabla_R \mathcal{L}(L_t, R_t)(L_t^\top L_t)^{-1},$$

where $\eta > 0$ is the step size and $\nabla_L \mathcal{L}(L_t, R_t)$ (resp. $\nabla_R \mathcal{L}(L_t, R_t)$) is the gradient of the loss function $\mathcal{L}$ with respect to the factor $L_t$ (resp. $R_t$) at the $t$-th iteration. Comparing to vanilla gradient descent, the search directions of the low-rank factors $L_t, R_t$ in (3) are scaled by $(R_t^\top R_t)^{-1}$ and $(L_t^\top L_t)^{-1}$ 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 $\delta_{2r} \leq 1/(\sqrt{r}\kappa)$, where $\kappa$ is the condition number of $X$, ScaledGD reaches $\epsilon$-accuracy in $O(\log(1/\epsilon))$ iterations when initialized by the spectral method. This strictly improves
RobustPCA.

procedures for high-dimensional signal estimation; see e.g. [JK17, CC18, CLC19] for recent overviews. A

Our work contributes to the growing literature of design and analysis of provable nonconvex optimization

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work

1.2 Related work
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 requires additional regularization terms to balance the two factors, either in the form of $\frac{1}{2}\|L^\top L - R^\top R\|_F^2$ [TBS+16, PKCS17] or $\frac{1}{2}\|L\|_F^2 + \frac{1}{2}\|R\|_F^2$ [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 $\kappa$ of the low-rank matrix, e.g. $O(\kappa \log(1/\epsilon))$, to reach $\epsilon$-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 $\kappa$ is large. Last but not least, alternating minimization [JNS13, HW14] (which alternatively updates $L_t$ and $R_t$) 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 $\kappa$ 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 \( v \), we use \( \| v \|_0 \) to denote its \( \ell_0 \) counting norm, and \( \| v \|_2 \) to denote the \( \ell_2 \) norm. For any matrix \( A \), we use \( \sigma_i(A) \) to denote its \( i \)-th largest singular value, and let \( A_i \), and \( A_{-i} \) denote its \( i \)-th row and \( j \)-th column, respectively. In addition, \( \| A \|_{op} \), \( \| A \|_{F} \), \( \| A \|_1 \), \( \| A \|_2 \), and \( \| A \|_{\infty} \) stand for the spectral norm (i.e. the largest singular value), the Frobenius norm, the \( \ell_1 \) norm (the largest \( \ell_1 \) norm of the rows), the \( \ell_2 \) norm (i.e. the largest \( \ell_2 \) norm of the rows), and the entrywise \( \ell_\infty \) norm (the largest magnitude of all entries) of a matrix \( A \). We denote

\[
\mathcal{P}_r(A) = \min_{\tilde{A}: \text{rank}(A) \leq r} \| A - \tilde{A} \|_F^2
\]

as the rank-\( r \) approximation of \( A \), which is given by the top-\( r \) SVD of \( A \) by the Eckart-Young-Mirscky theorem. We also use \( \text{vec}(A) \) to denote the vectorization of a matrix \( A \). For matrices \( A, B \) of the same size, we use \( \langle A, B \rangle = \sum_{i,j} A_{i,j} B_{i,j} = \text{tr}(A^\top B) \) to denote their inner product. The set of invertible matrices in \( \mathbb{R}^{n \times n} \) is denoted by \( \text{GL}(n) \). 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 \( |f(n)|/|g(n)| \leq C \) for some constant \( C > 0 \) when \( n \) is sufficiently large; \( f(n) \gtrsim g(n) \) means \( |f(n)|/|g(n)| \geq C \) 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

\[
\text{vec}(F_{t+1}) = \text{vec}(F_t) - \eta \begin{bmatrix}
(R_t^\top R_t)^{-1} \otimes I_{n_1} & 0 \\
0 & (L_t^\top L_t)^{-1} \otimes I_{n_2}
\end{bmatrix} \text{vec}(\nabla F \mathcal{L}(F_t))
\]

\[
= \text{vec}(F_t) - \eta H_t^{-1} \text{vec}(\nabla F \mathcal{L}(F_t)),
\]

where we denote \( F_t = [L_t^\top, R_t^\top]^\top \in \mathbb{R}^{(n_1 + n_2) \times r} \), and by \( \otimes \) the Kronecker product. Here, the block diagonal matrix \( H_t \) is set to be

\[
H_t := \begin{bmatrix}
(R_t^\top R_t) \otimes I_{n_1} & 0 \\
0 & (L_t^\top L_t) \otimes I_{n_2}
\end{bmatrix}.
\]

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

2.1 Assumptions and error metric

Denote by \( U_r, \Sigma_r, V_r^\top \) the compact singular value decomposition (SVD) of the rank-\( r \) matrix \( X_r \in \mathbb{R}^{n_1 \times n_2} \). Here \( U_r \in \mathbb{R}^{n_1 \times r} \) and \( V_r \in \mathbb{R}^{n_2 \times r} \) are composed of \( r \) left and right singular vectors, respectively, and \( \Sigma_r \in \mathbb{R}^{r \times r} \) is a diagonal matrix consisting of \( r \) singular values of \( X_r \) organized in a non-increasing order, i.e. \( \sigma_1(X_r) \geq \cdots \geq \sigma_r(X_r) > 0 \). Define

\[
\kappa := \sigma_1(X_r)/\sigma_r(X_r)
\]

\[7\]
as the condition number of $X_*$. Define the ground truth low-rank factors as

$$L_* := U_* \Sigma_*^{1/2}, \quad \text{and} \quad R_* := V_* \Sigma_*^{1/2},$$

so that $X_* = L_* R_*^\top$. Correspondingly, denote the stacked factor matrix as

$$F_* := \begin{bmatrix} L_* \\ R_* \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times r}.$$  

Next, we are in need of a right metric to measure the performance of the ScaledGD iterates $F_t := [L_t^\top, R_t^\top]^\top$. Obviously, the factored representation is not unique in that for any invertible matrix $Q \in \text{GL}(r)$, one has $LR = (LQ)(RQ^{-1})^\top$. 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:

$$(L_{t+1}, R_{t+1}) = \arg\min_{L, R} \mathcal{L}(L_t, R_t) + \langle \nabla L \mathcal{L}(L_t, R_t), L - L_t \rangle + \langle \nabla R \mathcal{L}(L_t, R_t), R - R_t \rangle$$

$$+ \frac{1}{2\eta} \left( \| (L - L_t)(R_t^\top R_t)^{1/2} \|_F^2 + \| (R - R_t)(L_t^\top L_t)^{1/2} \|_F^2 \right),$$

where it is different from the common interpretation of gradient descent in the way the quadratic approximation is taken by a scaled norm. When $L_t \approx L_*$ and $R_t \approx R_*$ are approaching the ground truth, the additional scaling factors can be approximated by $L_t^\top L_t \approx \Sigma_*$ and $R_t^\top R_t \approx \Sigma_*$, leading to the following error metric

$$\text{dist}^2(F, F_*) := \inf_{Q \in \text{GL}(r)} \left( \| LQ - L_* \|_F \Sigma_*^{1/2} \|_F^2 + \| (RQ^{-1} - R_*) \Sigma_*^{1/2} \|_F^2 \right).$$

Correspondingly, we define the optimal alignment matrix $Q$ between $F$ and $F_*$ as

$$Q := \arg\min_{Q \in \text{GL}(r)} \left( \| LQ - L_* \|_F \Sigma_*^{1/2} \|_F^2 + \| (RQ^{-1} - R_*) \Sigma_*^{1/2} \|_F^2 \right),$$

whenever the minimum is achieved. It turns out that for the ScaledGD iterates $\{F_t\}$, the optimal alignment matrices $\{Q_t\}$ 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 $X_* \in \mathbb{R}^{n_1 \times n_2}$, given as

$$y = A(X_*) \in \mathbb{R}^m,$$

where $A(X) = \{A_k, X\}_{k=1}^m : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}^m$ is the linear map modeling the measurement process. The goal of low-rank matrix sensing is to recover $X_*$ from $y$, 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].

---

1If there are multiple minimizers, we can arbitrarily take one to be $Q$.  

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

$$\minimize_{F \in \mathbb{R}^{(n_1+n_2)^2 \times r}} \mathcal{L}(F) = \frac{1}{2} \|A(LR^T) - y\|_2^2.$$  \hspace{1cm} (13)

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

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

Spectral initialization: Let $U_0 \Sigma_0 V_0^T$ be the top-$r$ SVD of $A^*(y)$, and set $L_0 = U_0 \Sigma_0^{1/2}$, and $R_0 = V_0 \Sigma_0^{1/2}$.

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

$$L_{t+1} = L_t - \eta A^*(A(L_t R_t^T) - y) R_t (R_t^T R_t)^{-1},$$

$$R_{t+1} = R_t - \eta A^*(A(L_t R_t^T) - y) L_t (L_t^T L_t)^{-1}.$$ \hspace{1cm} (15)

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

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

$$(1 - \delta_r)\|M\|_F^2 \leq \|A(M)\|_2^2 \leq (1 + \delta_r)\|M\|_F^2.$$

It is well-known that many measurement ensembles satisfy the RIP property [RFP10, CP11]. For example, if the entries of $A_i$’s are composed of i.i.d. Gaussian entries $\mathcal{N}(0, 1/m)$, then the RIP is satisfied for a constant $\delta_r$ 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 $A(\cdot)$ has a sufficiently small RIP constant.

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

$$\text{dist}(F_t, F_0) \leq (1 - 0.6\eta)^t 0.1\sigma_r(X_\star), \quad \text{and} \quad \|L_t R_t^T - X_\star\|_F \leq (1 - 0.6\eta)^t 0.15\sigma_r(X_\star).$$

Theorem 1 establishes that the distance dist$(F_t, F_\star)$ 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 $n = n_1 \lor n_2$. To reach $\epsilon$-accuracy, i.e., $\|L_t R_t^T - X_\star\|_F \leq \epsilon \sigma_r(X_\star)$, ScaledGD takes at most $T = O(\log(1/\epsilon))$ iterations, which is independent of the condition number $\kappa$ of $X_\star$. In comparison, alternating minimization with spectral initialization (AltMinSense) converges in $O(\log(1/\epsilon))$ iterations as long as $m = O(nr^2\kappa^2)$ [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 $X_{r+1} = P_r(X_r - A^*(A(X_r) - y))$ after $T_0 \geq \max\{\log r, \log \kappa\}$ iterations, where $P_r(\cdot)$ 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

\[ \mathbf{Y} = \mathbf{X}_* + \mathbf{S}_*, \]

which is a superposition of a rank-\(r\) matrix \( \mathbf{X}_* \), modeling the clean data, and a sparse matrix \( \mathbf{S}_* \), modeling the corruption or outliers. The goal of robust PCA [CLMW11, CSPW11] is to separate the two matrices \( \mathbf{X}_* \) and \( \mathbf{S}_* \) from their mixture \( \mathbf{Y} \). 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 \( \mathbf{S}_* \), in which \( \mathbf{S}_* \) contains at most \( \alpha \)-fraction of nonzero entries per row and column for some \( \alpha \in [0, 1) \), i.e. \( \mathbf{S}_* \in \mathcal{S}_\alpha \), where we denote

\[ \mathcal{S}_\alpha := \{ \mathbf{S} \in \mathbb{R}^{n_1 \times n_2} : \| \mathbf{S}_i \|_0 \leq \alpha n_2 \text{ for all } i, \text{ and } \| \mathbf{S}_{i,j} \|_0 \leq \alpha n_1 \text{ for all } j \}. \]  

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

\[
\min_{\mathbf{F} \in \mathbb{R}^{n_1 \times n_2} \times \mathbf{S} \in \mathcal{S}_\alpha} \mathcal{L}(\mathbf{F}, \mathbf{S}) = \frac{1}{2} \| \mathbf{L}\mathbf{R}^\top + \mathbf{S} - \mathbf{Y} \|_F^2. \tag{17}
\]

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

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

(18)

where \( |\mathbf{A}|_{i,k} \) (resp. \( |\mathbf{A}|_{k,j} \)) 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 \( \| \mathbf{L}^\top \mathbf{L} - \mathbf{R}^\top \mathbf{R} \|_2^2 \) 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 \( \mathbf{U}_0 \mathbf{\Sigma}_0 \mathbf{V}_0^\top \) be the top-\(r\) SVD of \( \mathbf{Y} - \mathcal{T}_{\bar{\alpha}}[\mathbf{Y}] \), and set

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

**Scaled gradient updates:** for \( t = 0, 1, 2, \ldots, T - 1 \) do

\[
\mathbf{S}_{t+1} = \mathcal{T}_{\bar{\alpha}}[\mathbf{Y} - \mathbf{L}_t \mathbf{R}_t^\top], \\
\mathbf{L}_{t+1} = \mathbf{L}_t - \eta (\mathbf{L}_t \mathbf{R}_t^\top + \mathbf{S}_t - \mathbf{Y}) \mathbf{R}_t (\mathbf{R}_t^\top \mathbf{R}_t)^{-1}, \\
\mathbf{R}_{t+1} = \mathbf{R}_t - \eta (\mathbf{L}_t \mathbf{R}_t^\top + \mathbf{S}_t - \mathbf{Y})^\top \mathbf{L}_t (\mathbf{L}_t^\top \mathbf{L}_t)^{-1}. \tag{20}
\]

**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 \( \mathbf{X}_* \) in robust PCA [Che15].

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

\[
\| \mathbf{U}_* \|_{2,\infty} \leq \sqrt{\frac{\mu}{n_1}} \| \mathbf{U}_* \|_F = \sqrt{\frac{\mu \nu}{n_1}}, \quad \text{and} \quad \| \mathbf{V}_* \|_{2,\infty} \leq \sqrt{\frac{\mu}{n_2}} \| \mathbf{V}_* \|_F = \sqrt{\frac{\mu \nu}{n_2}}.
\]
The following theorem establishes that ScaledGD converges linearly at a constant rate as long as the fraction \( \alpha \) of corruptions is sufficiently small.

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

\[
\text{dist}(\mathbf{F}_t, \mathbf{F}_s) \leq (1 - 0.6\eta)^t 0.02\sigma_r(\mathbf{X}_s), \quad \text{and} \quad \|\mathbf{L}_t \mathbf{R}^\top_t - \mathbf{X}_s\|_F \leq (1 - 0.6\eta)^t 0.03\sigma_r(\mathbf{X}_s).
\]

Theorem 2 establishes that the distance \( \text{dist}(\mathbf{F}_t, \mathbf{F}_s) \) contracts linearly at a constant rate as long as the fraction of corruptions satisfies \( \alpha \leq 1/(\mu r^3/2\kappa) \). To reach \( \epsilon \)-accuracy, i.e. \( \|\mathbf{L}_t \mathbf{R}^\top_t - \mathbf{X}_s\|_F \leq \epsilon \sigma_r(\mathbf{X}_s) \), ScaledGD takes at most \( T = O(\log(1/\epsilon)) \) iterations, which is independent of \( \kappa \). In comparison, the AltProj algorithm\(^3\) with spectral initialization converges in \( O(\log(1/\epsilon)) \) iterations as long as \( \alpha \leq 1/(\mu r^3/2\kappa^2/\sqrt{\nu \kappa^2}) \). 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 \( \Omega \) of entries of \( \mathbf{X} \) given as \( \mathcal{P}_\Omega(\mathbf{X}_s) \), where \( \mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \mapsto \mathbb{R}^{n_1 \times n_2} \) is a projection such that

\[
(\mathcal{P}_\Omega(\mathbf{X}))_{i,j} = \begin{cases} 
\mathbf{X}_{i,j}, & \text{if } (i,j) \in \Omega \\
0, & \text{otherwise}
\end{cases}
\]  

(21)

Here \( \Omega \) 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 \( \mathbf{X}_s \) from its partial observation \( \mathcal{P}_\Omega(\mathbf{X}_s) \). This problem has many applications in recommendation systems, signal processing, sensor network localization, and so on [CR09].

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

\[
\mathbf{F} \in \mathbb{R}^{(n_1 + n_2)^r} \quad \mathcal{L}(\mathbf{F}) := \frac{1}{2p}\|\mathcal{P}_\Omega(\mathbf{L} \mathbf{R}^\top - \mathbf{X}_s)\|_F^2.
\]

(22)

Similarly to robust PCA, the underlying low-rank matrix \( \mathbf{X}_s \) 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 [CWW15] 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 \( \mathbf{F} \in \mathbb{R}^{(n_1 + n_2)^r} \),

\[
\mathcal{P}_B(\mathbf{F}) = \arg\min_{\mathbf{F} \in \mathbb{R}^{(n_1 + n_2)^r}} \|\mathbf{L} - \mathbf{L} (\mathbf{R}^\top \mathbf{R})^{1/2} \|_F^2 + \|\mathbf{R} - \mathbf{R} (\mathbf{L}^\top \mathbf{L})^{1/2} \|_F^2, \\
\text{s.t. } \sqrt{\mu_1} \|\mathbf{L} (\mathbf{R}^\top \mathbf{R})^{1/2} \|_{2,\infty} + \sqrt{\mu_2} \|\mathbf{R} (\mathbf{L}^\top \mathbf{L})^{1/2} \|_{2,\infty} \leq B,
\]

(23)

which finds a factored matrix that is closest to \( \mathbf{F} \) 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 \( \kappa \) in \( \alpha \), which we do not consider here. The same strategy might also improve the dependence on \( \kappa \) for ScaledGD, which we leave for future work.
Proposition 1. The solution to (23) is given by
\[
\mathcal{P}_B(\tilde{F}) := \begin{bmatrix} L \\\n R \end{bmatrix}, \quad \text{where} \quad L_i := \left(1 \wedge \frac{B}{\sqrt{m_1} \|L_i R_i^\top\|_2}\right) \tilde{L}_{i,:}, \quad 1 \leq i \leq n_1,
\]
\[
R_j := \left(1 \wedge \frac{B}{\sqrt{n_2} \|R_j L_j^\top\|_2}\right) \tilde{R}_{j,:), \quad 1 \leq j \leq n_2.
\] (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

\begin{algorithm}
\caption{ScaledPGD for matrix completion with spectral initialization}
\begin{algorithmic}
\STATE \textbf{Spectral initialization:} Let $U_0 \Sigma_0 V_0^\top$ be the top-$r$ SVD of $\frac{1}{p} \mathcal{P}_\Omega(X_*)$, and set
\[
\begin{bmatrix} L_0 \\ R_0 \end{bmatrix} = \mathcal{P}_B \left( \begin{bmatrix} U_0 \Sigma_0^{1/2} \\ V_0 \Sigma_0^{1/2} \end{bmatrix} \right).
\]
(25)
\STATE \textbf{Scaled projected gradient updates:} for $t = 0, 1, 2, \ldots, T-1$ do
\[
\begin{bmatrix} L_{t+1} \\ R_{t+1} \end{bmatrix} = \mathcal{P}_B \left( \begin{bmatrix} L_t - \frac{2}{p} \mathcal{P}_\Omega(L_t R_t^\top - X_*) R_t (R_t^\top R_t)^{-1} \\ R_t - \frac{2}{p} \mathcal{P}_\Omega(L_t R_t^\top - X_*)^\top L_t (L_t^\top L_t)^{-1} \end{bmatrix} \right).
\]
(26)
\end{algorithmic}
\end{algorithm}

Theoretical guarantee. Consider a random observation model, where each index $(i,j)$ belongs to the index set $\Omega$ independently with probability $0 < p \leq 1$. 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 $X_*$ is $\mu$-incoherent, and that $p$ satisfies $p \geq C(\mu \kappa^2 \sqrt{\log(n_1 \vee n_2)}) \mu r^2 \kappa^2 / (n_1 \wedge n_2)$ for some sufficiently large constant $C$. Set the projection radius as $B = C_B \sqrt{\mu r \kappa_1(X_*)}$ for some constant $C_B \geq 1.02$. If the step size obeys $0 < \eta \leq 2/3$, then with probability at least $1 - c_1(n_1 \vee n_2)^{-c_2}$, for all $t \geq 0$, the iterates of ScaledPGD in (26) satisfy
\[
dist(F_t, F_*) \leq (1 - 0.6 \eta)^t (0.02 \sigma_\tau(X_*)), \quad \text{and} \quad \|L_t R_t^\top - X_*\|_F \leq (1 - 0.6 \eta)^t (0.03 \sigma_\tau(X_*)).
\]
Here $c_1, c_2 > 0$ are two universal constants.

Theorem 3 establishes that the distance $\dist(F_t, F_*)$ contracts linearly at a constant rate, as long as the probability of observation satisfies $p \geq (\mu \kappa^2 \sqrt{\log(n_1 \vee n_2)}) \mu r^2 \kappa^2 / (n_1 \wedge n_2)$. To reach $\epsilon$-accuracy, i.e. $\|L_t R_t^\top - X_*\|_F \leq \epsilon \sigma_\tau(X_*)$, ScaledPGD takes at most $T = O(\log(1/\epsilon))$ iterations, which is independent of $\kappa$. In comparison, projected gradient descent [ZL16] with spectral initialization converges in $O(\kappa \log(1/\epsilon))$ iterations as long as $p \geq (\mu \sqrt{\log(n_1 \vee n_2)}) \mu r^2 \kappa^2 / (n_1 \wedge n_2)$. 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
\[
L_{t+1} = L_t - \eta \nabla f(L_t R_t^\top) R_t (R_t^\top R_t)^{-1},
\]
\[
R_{t+1} = R_t - \eta \nabla f(L_t R_t^\top)^\top L_t (L_t^\top L_t)^{-1}.
\] (27)

Two important properties of the loss function $f(\cdot) : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$ play a key role in the analysis.
Definition 3 (Restricted smoothness). A differentiable function \( f(\cdot) : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R} \) is said to be rank-\( r \) restricted \( L \)-smooth for some \( L > 0 \) if
\[
f(X_2) \leq f(X_1) + \langle \nabla f(X_1), X_2 - X_1 \rangle + \frac{L}{2} \| X_2 - X_1 \|^2,
\]
for any \( X_1, X_2 \in \mathbb{R}^{n_1 \times n_2} \) with rank at most \( r \).

Definition 4 (Restricted strong convexity). A differentiable function \( f(\cdot) : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R} \) is said to be rank-\( r \) restricted \( \mu \)-strongly convex for some \( \mu \geq 0 \) if
\[
f(X_2) \geq f(X_1) + \langle \nabla f(X_1), X_2 - X_1 \rangle + \frac{\mu}{2} \| X_2 - X_1 \|^2,
\]
for any \( X_1, X_2 \in \mathbb{R}^{n_1 \times n_2} \) 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
\[
\kappa_f := \frac{L}{\mu}.
\]

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 \( f(X) = \frac{1}{2} \| X - X_* \|^2 \) in \( (29) \) satisfies \( \kappa_f = 1 \);
- **low-rank matrix sensing**, where the loss function \( f(X) = \frac{1}{2} \| AX - X_* \|^2 \) in \( (13) \) satisfies \( \kappa_f \approx 1 \) when \( A \) obeys the rank-\( r \) RIP with a sufficiently small RIP constant;
- **quadratic sampling**, where the loss function \( f(X) = \frac{1}{2} \sum_{i=1}^{m} \| a_i a_i^T, X - X_* \|^2 \) satisfies restricted strong convexity and smoothness when \( a_i \)'s are i.i.d. Gaussian vectors for sufficiently large \( m \) [SWW17,LMCC21];
- **exponential-family PCA**, where the loss function \( f(X) = - \sum_{i,j} \log p(Y_{i,j} | X_{i,j}) \), where \( p(Y_{i,j} | X_{i,j}) \) is the probability density function of \( Y_{i,j} \) conditional on \( X_{i,j} \), 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 \( X_* \).\footnote{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.} The following theorem establishes that as long as properly initialized, then \( \text{ScaledGD} \) converges linearly at a constant rate.

**Theorem 4.** Suppose that \( f(\cdot) : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R} \) is rank-\( 2r \) restricted \( L \)-smooth and \( \mu \)-strongly convex, of which \( X_* \) is a minimizer, and that the initialization \( F_0 \) satisfies \( \text{dist}(F_0, F_*) \leq 0.1 \sigma_r(X_*) / \sqrt{\kappa_f} \). If the step size obeys \( 0 < \eta \leq 0.4 / L \), then for all \( t \geq 0 \), the iterates of \( \text{ScaledGD} \) in \( (27) \) satisfy
\[
\text{dist}(F_t, F_*) \leq (1 - 0.7 \eta \mu)^t \frac{0.1 \sigma_r(X_*)}{\sqrt{\kappa_f}}, \quad \text{and} \quad \| L_t R_t^T - X_* \|_F \leq (1 - 0.7 \eta \mu)^t \frac{0.15 \sigma_r(X_*)}{\sqrt{\kappa_f}}.
\]

Theorem 4 establishes that the distance \( \text{dist}(F_t, F_*) \) contracts linearly at a constant rate, as long as the initialization \( F_0 \) is sufficiently close to \( F_* \). To reach \( \epsilon \)-accuracy, i.e. \( \| L_t R_t^T - X_* \|_F \leq \epsilon \sigma_r(X_*) \), \( \text{ScaledGD} \) takes at most \( T = O(\kappa_f \log(1/\epsilon)) \) iterations, which depends only on the condition number \( \kappa_f \) of \( f(\cdot) \), but is independent of the condition number \( \kappa \) of the matrix \( X_* \). 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 \( \kappa \).

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.
3.1 A warm-up analysis: matrix factorization

Let us consider the problem of factorizing a matrix $X_*$ into two low-rank factors:

$$\min_{F \in \mathbb{R}^{(n_1+n_2) \times r}} \mathcal{L}(F) = \frac{1}{2} \|LR^\top - X_*\|^2_F. \quad (29)$$

For this toy problem, the update rule of $\text{ScaledGD}$ is given as

$$L_{t+1} = L_t - \eta (L_t R_t^\top - X_*) R_t (R_t^\top R_t)^{-1},$$

$$R_{t+1} = R_t - \eta (L_t R_t^\top - X_*)^\top L_t (L_t^\top L_t)^{-1}. \quad (30)$$

To shed light on why $\text{ScaledGD}$ is robust to ill-conditioning, it is worthwhile to think of $\text{ScaledGD}$ as a quasi-Newton algorithm: the following proposition (proven in Appendix B.1) reveals that $\text{ScaledGD}$ 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), $\text{ScaledGD}$ is equivalent to the following update rule

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

Here, $\nabla_{L,L}^2 \mathcal{L}(F_t)$ (resp. $\nabla_{R,R}^2 \mathcal{L}(F_t)$) denotes the second order derivative w.r.t. $L$ (resp. $R$) at $F_t$.

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

**Theorem 5.** Suppose that the initialization $F_0$ satisfies $\text{dist}(F_0, F_*) \leq 0.1\sigma_r(X_*)$. If the step size obeys $0 < \eta \leq 2/3$, then for all $t \geq 0$, the iterates of the $\text{ScaledGD}$ method in (30) satisfy

$$\text{dist}(F_t, F_*) \leq (1 - 0.7\eta)^t 0.1\sigma_r(X_*), \quad \text{and} \quad \|L_t R_t^\top - X_*\|_F \leq (1 - 0.7\eta)^t 0.15\sigma_r(X_*).$$

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 $\text{ScaledGD}$ 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 $\delta_{2r}$ than the main theorem (cf. Theorem 1).

**Lemma 1.** Suppose that $\mathcal{A}(\cdot)$ obeys the $2r$-RIP with $\delta_{2r} \leq 0.02$. If the $t$-th iterate satisfies $\text{dist}(F_t, F_*) \leq 0.1\sigma_r(X_*)$, then $\|L_t R_t^\top - X_*\|_F \leq 1.5 \text{dist}(F_t, F_*)$. In addition, if the step size obeys $0 < \eta \leq 2/3$, then the $(t+1)$-th iterate $F_{t+1}$ of the $\text{ScaledGD}$ method in (15) of Algorithm 1 satisfies

$$\text{dist}(F_{t+1}, F_*) \leq (1 - 0.6\eta) \text{dist}(F_t, F_*).$$

It then boils 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 $\delta_{2r}$. Then the spectral initialization in (14) for low-rank matrix sensing satisfies

$$\text{dist}(F_0, F_*) \leq 5\delta_{2r} \sqrt{\kappa} \sigma_r(X_*).$$

Therefore, as long as $\delta_{2r}$ is small enough, say $\delta_{2r} \leq 0.02/(\sqrt{\kappa})$ as specified in Theorem 1, the initial distance satisfies $\text{dist}(F_0, F_*) \leq 0.1\sigma_r(X_*)$, 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 \( \alpha \) 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 \( Q_t \) is the optimal alignment matrix between \( F_t \) and \( F_* \).

**Lemma 3.** Suppose that \( X_* \) is \( \mu \)-incoherent and \( \alpha \leq 10^{-4}/(\mu \nu) \). If the \( t \)-th iterate satisfies \( \text{dist}(F_t, F_*) \leq 0.02 \sigma_r(X_*) \) and the incoherence condition

\[
\sqrt{n_1} \left\| \left( L_t Q_t - L_* \right) \Sigma_{t}^{1/2} \right\|_{2, \infty} \vee \sqrt{n_2} \left\| \left( R_t Q_t^T - R_* \right) \Sigma_{t}^{1/2} \right\|_{2, \infty} \leq \sqrt{\mu r} \sigma_r(X_*),
\]

then \( \|L_t R_t^T - X_*\|_F \leq 1.5 \text{dist}(F_t, F_*) \). In addition, if the step size obeys \( 0.1 \leq \eta \leq 2/3 \), then the \( (t+1) \)-th iterate \( F_{t+1} \) of the ScaledGD method in (20) of Algorithm 2 satisfies

\[
\text{dist}(F_{t+1}, F_*) \leq (1 - 0.6 \eta) \text{dist}(F_t, F_*),
\]

and the incoherence condition

\[
\sqrt{n_1} \left\| \left( L_{t+1} Q_{t+1} - L_* \right) \Sigma_{t+1}^{1/2} \right\|_{2, \infty} \vee \sqrt{n_2} \left\| \left( R_{t+1} Q_{t+1}^T - R_* \right) \Sigma_{t+1}^{1/2} \right\|_{2, \infty} \leq \sqrt{\mu r} \sigma_r(X_*).
\]

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 \( \alpha \) is sufficiently small.

**Lemma 4.** Suppose that \( X_* \) is \( \mu \)-incoherent. Then the spectral initialization (19) for robust PCA satisfies

\[
\text{dist}(F_0, F_*) \leq 20 \alpha \mu r^{3/2} \kappa \sigma_r(X_*).
\]

As a result, setting \( \alpha \leq 10^{-3}/(\mu r^{3/2} \kappa) \), the spectral initialization satisfies \( \text{dist}(F_0, F_*) \leq 0.02 \sigma_r(X_*) \). 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 \( X_* \) is \( \mu \)-incoherent and \( \alpha \leq 0.1/(\mu \nu \kappa) \), and that \( \text{dist}(F_0, F_*) \leq 0.02 \sigma_r(X_*) \). Then the spectral initialization (19) satisfies the incoherence condition

\[
\sqrt{n_1} \left\| \left( L_0 Q_0 - L_* \right) \Sigma_0^{1/2} \right\|_{2, \infty} \vee \sqrt{n_2} \left\| \left( R_0 Q_0^T - R_* \right) \Sigma_0^{1/2} \right\|_{2, \infty} \leq \sqrt{\mu r} \sigma_r(X_*).
\]

Combining Lemmas 3-5 finishes the proof of Theorem 2. The proofs of 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 \( X_* \) is \( \mu \)-incoherent, and \( \text{dist}(\tilde{F}, F_*) \leq \epsilon \sigma_r(X_*) \) for some \( \epsilon < 1 \). Set \( B \geq (1 + \epsilon)\sqrt{\mu r} \sigma_1(X_*) \), then \( P_B(\tilde{F}) \) satisfies the non-expansiveness

\[
\text{dist}(P_B(\tilde{F}), F_*) \leq \text{dist}(\tilde{F}, F_*),
\]

and the incoherence condition

\[
\sqrt{n_1} \left\| LR^T \right\|_{2, \infty} \vee \sqrt{n_2} \left\| RL^T \right\|_{2, \infty} \leq B.
\]
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 $X_*$ is $\mu$-incoherent, and $p \geq C(\mu r^2 \kappa^2 \log(n_1 \lor n_2)) \mu r/(n_1 \land n_2)$ for some sufficiently large constant $C$. Set the projection radius as $B = C_B \sqrt{\mu r \sigma_1(X_*)}$ for some constant $C_B \geq 1.02$. Under an event $E$ which happens with overwhelming probability (i.e. at least $1 - c_1(n_1 \lor n_2)^{-c_2}$), if the $t$-th iterate satisfies $\text{dist}(F_t, F_*) \leq 0.02\sigma_r(X_*)$, and the incoherence condition

$$\sqrt{n_1} \|L_t R_t^\top\|_{2,\infty} \lor \sqrt{n_1} \|R_t L_t^\top\|_{2,\infty} \leq B,$$

then $\|L_t R_t^\top - X_*\|_F \leq 1.5 \text{dist}(F_t, F_*)$. In addition, if the step size obeys $0 < \eta \leq 2/3$, then the $(t + 1)$-th iterate $F_{t+1}$ of the ScaledPGD method in (26) of Algorithm 3 satisfies

$$\text{dist}(F_{t+1}, F_*) \leq (1 - 0.6\eta) \text{dist}(F_t, F_*),$$

and the incoherence condition

$$\sqrt{n_1} \|L_{t+1} R_{t+1}^\top\|_{2,\infty} \lor \sqrt{n_2} \|R_{t+1} L_{t+1}^\top\|_{2,\infty} \leq B.$$

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 $X_*$ is $\mu$-incoherent, then with overwhelming probability, the spectral initialization before projection $\tilde{F}_0 := \begin{bmatrix} U_0 \Sigma_0^{1/2} \\ V_0 \Sigma_0^{1/2} \end{bmatrix}$ in (25) satisfies

$$\text{dist}(\tilde{F}_0, F_*) \leq C_0 \left( \frac{\mu r \log(n_1 \lor n_2)}{\sqrt{n_1 n_2}} + \sqrt{\frac{\mu r \log(n_1 \lor n_2)}{p(n_1 \land n_2)}} \right) 5\sqrt{r} \kappa \sigma_r(X_*).$$

Therefore, as long as $p \geq C\mu r^2 \kappa^2 \log(n_1 \lor n_2)/(n_1 \land n_2)$ for some sufficiently large constant $C$, the initial distance satisfies $\text{dist}(\tilde{F}_0, F_*) \leq 0.02\sigma_r(X_*)$. One can then invoke Lemma 6 to see that $F_0 = P_B(\tilde{F}_0)$ 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

$$L_{t+1} = L_t - \eta_\text{GD} \nabla_L \mathcal{L}(L_t, R_t),$$

$$R_{t+1} = R_t - \eta_\text{GD} \nabla_R \mathcal{L}(L_t, R_t),$$

where $\eta_\text{GD} = \eta/\sigma_1(X_*)$ 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 $\sigma_1(X_*)$ 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:
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 $y_k = \langle A_k, X_t \rangle + w_k$, in which the measurement matrices $A_k$ are generated with i.i.d. Gaussian entries with zero mean and variance $1/m$, and $w_k \sim N(0, \sigma^2_w)$ 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 $S_\star \in S_\alpha$ with $\alpha = 0.1$. More specifically, we generate a matrix with standard Gaussian entries and pass it through $T_\alpha[\cdot]$ to obtain $S_\star$. The observation is $Y = X_\star + S_\star + W$, where $W_{i,j} \sim N(0, \sigma^2_w)$ 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 $X_\star$ is observed with probability $p = 0.2$ independently. The observation is $Y = P_\Omega(X_\star + W)$, where $W_{i,j} \sim N(0, \sigma^2_w)$ 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 $X_\star$ is observed with
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, 80dB for (a) matrix sensing, (b) robust PCA, (c) matrix completion, and (d) Hankel matrix completion.

The loss function is

$$L(L, R) = \frac{1}{2p}\|\mathcal{H}_\Omega(LR^\top - Y)\|^2_F + \frac{1}{2}\|(I - \mathcal{H})(LR^\top)\|^2_F,$$

where $\mathcal{H}(\cdot)$ denotes the identity operator, and the Hankel projection is defined as $\mathcal{H}(X) := \sum_{k=1}^{2n-1}(H_k, X)H_k$, which maps $X$ to its closest Hankel matrix. Here, the Hankel basis matrix $H_k$ 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 $\omega_k$ is the length of the $k$-th skew diagonal. Note that $X$ is a Hankel matrix if and only if $(I - \mathcal{H})(X) = 0$. The Hankel projection on the observation index set $\Omega$ is defined as $\mathcal{H}_\Omega(X) := \sum_{k \in \Omega}(H_k, X)H_k$. The observation is $Y = \mathcal{H}_\Omega(X_\ast + W)$, where $W$ 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 $X_\ast \in \mathbb{R}^{n \times n}$ 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 $U_\ast$, and similarly for $V_\ast$. The singular values are set to be linearly distributed from $\kappa^{-1}$ to 1. The ground truth is then defined as $X_\ast = U_\ast \Sigma_\ast V_\ast^\top$ which has the specified condition number $\kappa$ and rank $r$. For Hankel matrix completion,
we generate $X_\star$ as an $n \times n$ Hankel matrix with entries given as

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

where $f_\ell, \ell = 1,\ldots,r$ are randomly chosen from $1/n, 2/n, \ldots, 1$, and $\sigma_\ell$ are linearly distributed from $\kappa^{-1}$ to $1$. The Vandermonde decomposition lemma tells that $X_\star$ has rank $r$ and singular values $\sigma_\ell, \ell = 1,\ldots,r$.

We first illustrate the convergence performance under noise-free observations, i.e. $\sigma_w = 0$. We plot the relative reconstruction error $\|X_t - X_\star\|_F / \|X_\star\|_F$ 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 $\kappa$, 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 $\kappa$ 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 $\text{SNR} := 10 \log_{10} \frac{\|X_\star\|_F}{\sigma_w}$ in dB. We plot the reconstruction error $\|X_t - X_\star\|_F / \|X_\star\|_F$ with respect to the iteration count $t$ in Figure 3 under the condition number $\kappa = 10$ and various SNR = 40, 60, 80dB. 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 $\eta$ 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 $\eta$ 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 $\eta$ from 0.1 to 1.2, for matrix completion with $n = 1000, r = 10, p = 0.2$.](image-url)

### 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

$$L_{t+1} = \arg\min_L \|A(L R_t^\top) - y\|_2^2, $$

$$R_{t+1} = \arg\min_R \|A(L_{t+1} R^\top) - y\|_2^2, $$
which corresponds to solving two least-squares problems. For matrix completion, the update rule of alternating minimization proceeds as

\[
L_{t+1} = \arg\min_L \|P_\Omega (LR_t^\top - Y)\|_2^2,
\]

\[
R_{t+1} = \arg\min_R \|P_\Omega (L_{t+1}R^\top - Y)\|_2^2,
\]

which can be implemented more efficiently since each row of \(L\) (resp. \(R\)) 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 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
parameters as well as implementation details, our purpose is to demonstrate the competitive performance of \textit{ScaledGD} over alternatives, rather than claiming it as the state-of-the-art.

5 Conclusions

This paper proposes scaled gradient descent (\textit{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 $\kappa$ of the low-rank matrix. In particular, we rigorously establish that for low-rank matrix sensing, robust PCA, and matrix completion, to reach $\epsilon$-accuracy, \textit{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 \textit{ScaledGD} under this new metric. This work opens up many venues for future research, as we discuss below.

- \textit{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 \cite{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\textsuperscript{+}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. \( \kappa_f \) 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**


A Technical Lemmas

This section gathers several useful lemmas that will be used in the appendix. Throughout all lemmas, we use $X_*$ to denote the ground truth low-rank matrix, with its compact SVD as $X_* = U_* \Sigma_* V_*^T$, and the stacked factor matrix is defined as $F_* = \begin{bmatrix} L_* \\ R_* \end{bmatrix} = \begin{bmatrix} U_* \Sigma_1^{1/2} \\ V_* \Sigma_2^{1/2} \end{bmatrix}$.

A.1 New distance metric

We begin with the investigation of the new distance metric (10), where the matrix $Q$ that attains the infimum, if exists, is called the optimal alignment matrix between $F$ and $F_*$; 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 $F$ and $F_*$, 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 $F = \begin{bmatrix} L \\ R \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times r}$. Suppose that

$$\text{dist}(F, F_* \equiv \begin{bmatrix} L_* \\ R_* \end{bmatrix}) = \sqrt{\inf_{Q \in \text{GL}(r)} \| (LQ - L_*) \Sigma_*^{1/2} \|_F^2 + \| (RQ - R_*) \Sigma_*^{1/2} \|_F^2} < \sigma_r(X_*),$$  \hfill (34)

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

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

$$\sqrt{\| (LQ - L_*) \Sigma_*^{1/2} \|_F^2 + \| (RQ - R_*) \Sigma_*^{1/2} \|_F^2} \leq \epsilon \sigma_r(X_*),$$

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

$$\| (LQ - L_*) \Sigma_*^{1/2} \|_{\text{op}} \vee \| (RQ - R_*) \Sigma_*^{1/2} \|_{\text{op}} \leq \epsilon.$$

Invoke Weyl’s inequality $|\sigma_r(A) - \sigma_r(B)| \leq \| A - B \|_{\text{op}}$ and use that $\sigma_r(L_* \Sigma_*^{-1/2}) = \sigma_r(U_*) = 1$ to obtain

$$\sigma_r(LQ \Sigma_*^{-1/2}) \geq \sigma_r(L_* \Sigma_*^{-1/2}) - \| (LQ - L_*) \Sigma_*^{-1/2} \|_{\text{op}} \geq 1 - \epsilon.$$  \hfill (35)

In addition, it is straightforward to verify that

$$\inf_{Q \in \text{GL}(r)} \| (LQ - L_*) \Sigma_*^{1/2} \|_F^2 \leq \inf_{H \in \text{GL}(r)} \| (LQH - L_*) \Sigma_*^{1/2} \|_F^2,$$  \hfill (36)

$$= \inf_{H \in \text{GL}(r)} \| (LQH - L_*) \Sigma_*^{1/2} \|_F^2 \geq \inf_{H \in \text{GL}(r)} \| (LQH - L_*) \Sigma_*^{1/2} \|_F^2.$$  \hfill (37)

Indeed, if the minimizer of the second optimization problem (cf. (37)) is attained at some $H$, then $QH$ 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 $H$. In view of (36) and (37), one has

$$\inf_{H \in \text{GL}(r)} \| (LQH - L_*) \Sigma_*^{1/2} \|_F^2 \leq \| (LQ - L_*) \Sigma_*^{1/2} \|_F^2 \leq \epsilon \sigma_r(X_*).$$

It further implies that

$$\| (LQH - L_*) \Sigma_*^{1/2} \|_{\text{op}} \vee (RQH - R_* \Sigma_*^{-1/2}) \|_{\text{op}} \leq \epsilon.$$

Invoke Weyl’s inequality $|\sigma_1(A) - \sigma_1(B)| \leq \| A - B \|_{\text{op}}$ and use that $\sigma_1(L_* \Sigma_*^{-1/2}) = \sigma_1(U_*) = 1$ to obtain

$$\sigma_1(LQH \Sigma_*^{-1/2}) \leq \sigma_1(L_* \Sigma_*^{-1/2}) \leq 1 + \epsilon.$$  \hfill (38)

Combine (35) and (38), and use the relation $\sigma_r(A) \sigma_1(B) \leq \sigma_1(AB)$ to obtain

$$\sigma_r(LQ \Sigma_*^{-1/2}) \sigma_1(H \Sigma_*^{-1/2}) \leq \sigma_1(LQH \Sigma_*^{-1/2}) \leq \frac{1 + \epsilon}{1 - \epsilon} \sigma_r(LQ \Sigma_*^{-1/2}).$$
As a result, one has $\sigma_{1}(\Sigma_{1/2}^1 H \Sigma_{-1/2}^1) \leq \frac{1+\epsilon}{1-\epsilon}$.

Similarly, one can show that $\sigma_{1}(\Sigma_{1/2}^1 H^{-1} \Sigma_{-1/2}^1) \leq \frac{1+\epsilon}{1-\epsilon}$, equivalently, $\sigma_{r}(\Sigma_{1/2}^1 H \Sigma_{-1/2}^1) \geq \frac{1-\epsilon}{1+\epsilon}$. Combining the above two arguments reveals that the minimization problem (37) is equivalent to the constrained problem:

$$\text{minimize}_{H \in \text{GL}(r)} \left\| (LQH - L_{*}) \Sigma_{1/2}^1 \right\|_{F}^{2} + \left\| (RQ^{-T} H^{-T} - R_{*}) \Sigma_{1/2}^1 \right\|_{F}^{2}$$

s.t. $\frac{1-\epsilon}{1+\epsilon} \leq \sigma_{r}(\Sigma_{1/2}^1 H \Sigma_{-1/2}^1) \leq \sigma_{1}(\Sigma_{1/2}^1 H \Sigma_{-1/2}^1) \leq \frac{1+\epsilon}{1-\epsilon}$.

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

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 $F = [L \ R] \in \mathbb{R}^{(n_1+n_2) \times r}$, suppose that the optimal alignment matrix $Q = \text{argmin}_{Q \in \text{GL}(r)} \left\| (LQ - L_{*}) \Sigma_{1/2}^1 \right\|_{F}^{2} + \left\| (RQ^{-T} - R_{*}) \Sigma_{1/2}^1 \right\|_{F}^{2}$ between $F$ and $F_{*}$ exists, then $Q$ obeys

$$(LQ)^{T}(LQ - L_{*}) \Sigma_{*} = \Sigma_{*}(RQ^{-T} - R_{*})^{T} RQ^{-T}. \quad (39)$$

**Proof.** Expand the squares in the definition of $Q$ to obtain

$$Q = \text{argmin}_{Q \in \text{GL}(r)} \text{tr} \left( (LQ - L_{*})^{T} (LQ - L_{*}) \Sigma_{*} \right) + \text{tr} \left( (RQ^{-T} - R_{*})^{T} (RQ^{-T} - R_{*}) \Sigma_{*} \right).$$

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

$$2L^{T}(LQ - L_{*}) \Sigma_{*} - 2Q^{-T} \Sigma_{*}(RQ^{-T} - R_{*})^{T} RQ^{-T} = 0,$$

which implies the optimal alignment criterion (39). \( \Box \)

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 $F = [L \ R] \in \mathbb{R}^{(n_1+n_2) \times r}$, the distance between $F$ and $F_{*}$ satisfies

$$\text{dist}(F, F_{*}) \leq \left( \sqrt{2} + 1 \right)^{1/2} \| LR^{T} - X_{*} \|_{F}.$$

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

$$2 \begin{bmatrix} X \ X^{T} \\ 0 \end{bmatrix} = FF^{T} - \bar{F} F_{*}^{T} + \bar{F} F_{*}^{T}, \quad 2 \begin{bmatrix} X \ X^{T} \\ 0 \end{bmatrix} = F F^{T} - \bar{F} F_{*}^{T} + \bar{F} F_{*}^{T}.$$

As a result, the squared Frobenius norm of $X - X_{*}$ is given by

$$8\|X - X_{*}\|_{F}^{2} = \| FF^{T} - \bar{F} F_{*}^{T} - \bar{F} F_{*}^{T} + \bar{F} F_{*}^{T} \|_{F}^{2}.$$
This in conjunction with 

Therefore we obtain

where we use the facts that \( \| FF^T - F^* F^* \|_F^2 = \| F F^T - F^* F^* \|_F^2 \) and \( F^T F = F^* F^* = 0 \).

Let \( O := \text{sgn}(F^* F^*) \)^5 be the optimal orthonormal alignment matrix between \( F \) and \( F^* \). Denote \( \Delta := FO - F^* \). Follow the same argument as [TBS+16, Lemma 5.14] and [GJZ17, Lemma 41] to obtain

\[
4\| X - X^* \|_F^2 \geq \| F \Delta + \Delta F^* + \Delta \Delta^T \|_F^2 \\
= \text{tr} \left( 2(F^* F^T \Delta + (\Delta^T \Delta)^2 + 2(F^* \Delta)^2 + 4F^*_\Delta \Delta^T \Delta) \right) \\
= \text{tr} \left( 2(F^* F^T \Delta + (\Delta^T \Delta + \sqrt{2}F^* \Delta)^2 + (4 - 2\sqrt{2})F^*_\Delta \Delta^T \Delta) \right) \\
\geq \text{tr} \left( 4(\sqrt{2} - 1)\Sigma, \Delta^T \Delta \right) = 4(\sqrt{2} - 1) \| (FO - F^*) \Sigma^1/2 \|_F^2,
\]

where the last inequality follows from the facts that \( F^* F^* = 2\Sigma \) and that \( F^* FO \) is positive semi-definite. Therefore we obtain

\[
\| (FO - F^*) \Sigma^1/2 \|_F |\leq (\sqrt{2} + 1)^{1/2} \| X - X^* \|_F.
\]

This in conjunction with \( \text{dist}(F, F^*) \leq \| (FO - F^*) \Sigma^1/2 \|_F \) yields the claimed result.

\[\square\]

### A.2 Matrix perturbation bounds

**Lemma 12.** For any \( L \in \mathbb{R}^{n \times r}, R \in \mathbb{R}^{p \times r} \), denote \( \Delta_L := L - L^* \) and \( \Delta_R := R - R^* \). Suppose that \( \| \Delta_L \Sigma^{-1/2} \|_{\text{op}} \vee \| \Delta_R \Sigma^{-1/2} \|_{\text{op}} < 1 \), then one has

\[
\begin{align*}
\| L(L^T L)^{-1} \Sigma^{1/2} \|_{\text{op}} &\leq \frac{1}{1 - \| \Delta_L \Sigma^{-1/2} \|_{\text{op}}}; \\
\| R(R^T R)^{-1} \Sigma^{1/2} \|_{\text{op}} &\leq \frac{1}{1 - \| \Delta_R \Sigma^{-1/2} \|_{\text{op}}}; \\
\| (L^T L)^{-1} \Sigma^1/2 - U \|_F &\leq \frac{\sqrt{2} \| \Delta_L \Sigma^{-1/2} \|_{\text{op}}}{1 - \| \Delta_L \Sigma^{-1/2} \|_{\text{op}}}; \\
\| (R^T R)^{-1} \Sigma^1/2 - V \|_F &\leq \frac{\sqrt{2} \| \Delta_R \Sigma^{-1/2} \|_{\text{op}}}{1 - \| \Delta_R \Sigma^{-1/2} \|_{\text{op}}}.
\end{align*}
\]

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

\[
\| L(L^T L)^{-1} \Sigma^{1/2} \|_{\text{op}} = \frac{1}{\sigma_r(L \Sigma^{-1/2})}.
\]

In addition, invoke Weyl's inequality to obtain

\[
\sigma_r(L \Sigma^{-1/2}) \geq \sigma_r(L^* \Sigma^{-1/2}) - \| \Delta_L \Sigma^{-1/2} \|_{\text{op}} = 1 - \| \Delta_L \Sigma^{-1/2} \|_{\text{op}},
\]

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

[^5]: Let \( A \Sigma B^T \) be the SVD of \( F^* F^* \), then the matrix sign is \( \text{sgn}(F^* F^*) := A \Sigma B^T \).
We proceed to prove (40c). Combine \( U_* = L_* \Sigma_*^{-1/2} \) and \((I_{n_1} - L(L^T L)^{-1} L^T) L = 0 \) to obtain the decomposition
\[
U_* = L(L^T L)^{-1} L^T U_* + (I_{n_1} - L(L^T L)^{-1} L^T) L_* \Sigma_*^{-1/2}
= L(L^T L)^{-1} L^T U_* - (I_{n_1} - L(L^T L)^{-1} L^T) \Delta_L \Sigma_*^{-1/2}.
\]
Subtract it by \( L(L^T L)^{-1} \Sigma_*^{-1/2} \) and write \( \Sigma_*^{-1/2} = L^T U_* \) to obtain
\[
L(L^T L)^{-1} \Sigma_*^{-1/2} - U_* = -L(L^T L)^{-1} \Delta_L^T U_* + (I_{n_1} - L(L^T L)^{-1} L^T) \Delta_L \Sigma_*^{-1/2}.
\]
For any \( \tilde{v} \in \mathbb{R}^{n_2} \) with \( \| \tilde{v} \|_2 \leq 1 \), the fact that \( L(L^T L)^{-1} \Delta_L U_* \tilde{v} \) and \((I_{n_1} - L(L^T L)^{-1} L^T) \Delta_L \Sigma_*^{-1/2} \tilde{v} \) are orthogonal implies
\[
\| (L(L^T L)^{-1} \Sigma_*^{-1/2} - U_*) \tilde{v} \|_2^2 = \| L(L^T L)^{-1} \Delta_L U_* \tilde{v} \|_2^2 + \| (I_{n_1} - L(L^T L)^{-1} L^T) \Delta_L \Sigma_*^{-1/2} \tilde{v} \|_2^2
\leq \frac{\| \Delta_L \Sigma_*^{-1/2} \|_{\text{op}}^2}{(1 - \| \Delta_L \Sigma_*^{-1/2} \|_{\text{op}}^2)^2} + \| \Delta_L \Sigma_*^{-1/2} \|_{\text{op}}^2
\leq \frac{2 \| \Delta_L \Sigma_*^{-1/2} \|_{\text{op}}^2}{(1 - \| \Delta_L \Sigma_*^{-1/2} \|_{\text{op}}^2)^2},
\]
where we have used (40a) and the fact that \( \| I_{n_1} - L(L^T L)^{-1} L^T \|_{\text{op}} \leq 1 \) in the third line. Utilize the variational representation of the operator norm to obtain
\[
\| L(L^T L)^{-1} \Sigma_*^{-1/2} - U_* \|_{\text{op}} \leq \max_{\tilde{v} \in \mathbb{R}^{n_2} : \| \tilde{v} \|_2 \leq 1} \| (L(L^T L)^{-1} \Sigma_*^{-1/2} - U_*) \tilde{v} \|_2 \\
\leq \frac{\sqrt{2} \| \Delta_L \Sigma_*^{-1/2} \|_{\text{op}}}{1 - \| \Delta_L \Sigma_*^{-1/2} \|_{\text{op}}^2}.
\]

Lemma 13. For any \( L \in \mathbb{R}^{n_1 \times r}, R \in \mathbb{R}^{n_2 \times r} \), denote \( \Delta_L := L - L_* \) and \( \Delta_R := R - R_* \), then one has
\[
\| LR^T - X_* \|_F \leq \| \Delta_L R_*^T \|_F + \| L_* \Delta_R^T \|_F + \| \Delta_L \Delta_R^T \|_F
\leq \left( 1 + \frac{1}{2} (\| \Delta_L \Sigma_*^{-1/2} \|_{\text{op}} \vee \| \Delta_R \Sigma_*^{-1/2} \|_{\text{op}}) \right) \left( \| \Delta_L \Sigma_*^{-1/2} \|_F + \| \Delta_R \Sigma_*^{-1/2} \|_F \right).
\]

Proof. In light of the decomposition \( LR^T - X_* = \Delta_L R_*^T + L_* \Delta_R^T + \Delta_L \Delta_R^T \) and the triangle inequality, one has
\[
\| LR^T - X_* \|_F \leq \| \Delta_L R_*^T \|_F + \| L_* \Delta_R^T \|_F + \| \Delta_L \Delta_R^T \|_F
= \| \Delta_L \Sigma_*^{1/2} \|_F + \| \Delta_R \Sigma_*^{1/2} \|_F + \| \Delta_L \Delta_R \|_F,
\]
where we have used the facts that
\[
\| \Delta_L R_*^T \|_F = \| \Delta_L \Sigma_*^{1/2} V_*^T \|_F = \| \Delta_L \Sigma_*^{1/2} \|_F, \quad \text{and} \quad \| L_* \Delta_R^T \|_F = \| U_* \Sigma_*^{1/2} \|_F = \| \Delta_R \Sigma_*^{1/2} \|_F.
\]
This together with the simple upper bound
\[
\| \Delta_L \Delta_R \|_F = \frac{1}{2} \| \Delta_L \Sigma_*^{1/2} (\Delta_R \Sigma_*^{-1/2})^T \|_F + \frac{1}{2} \| \Delta_L (\Delta_R \Sigma_*^{-1/2})^T \|_F
\leq \frac{1}{2} \| \Delta_L \Sigma_*^{1/2} \|_F \| \Delta_R \Sigma_*^{-1/2} \|_{\text{op}} + \frac{1}{2} \| \Delta_L \Sigma_*^{-1/2} \|_{\text{op}} \| \Delta_R \Sigma_*^{1/2} \|_F
\leq \frac{1}{2} \| \Delta_L \Sigma_*^{-1/2} \|_{\text{op}} \vee \| \Delta_R \Sigma_*^{-1/2} \|_{\text{op}} \left( \| \Delta_L \Sigma_*^{1/2} \|_F + \| \Delta_R \Sigma_*^{1/2} \|_F \right)
\]
finishes the proof.
Lemma 14. For any \( L \in \mathbb{R}^{n_1 \times r} \), \( R \in \mathbb{R}^{n_2 \times r} \) and any invertible matrices \( Q, \bar{Q} \in \text{GL}(r) \), suppose that
\[
\| (LQ - L_s)\Sigma_s^{-1/2} \|_{\text{op}} \vee \| (RQ^{-\top} - R_s)\Sigma_s^{-1/2} \|_{\text{op}} < 1,
\]
then one has
\[
\begin{align*}
\| \Sigma_s^{1/2} Q \Sigma_s^{1/2} - \Sigma_s \|_{\text{op}} & \leq \frac{\| R(\bar{Q}^{-\top} - Q^{-\top})\Sigma_s^{1/2} \|_{\text{op}}}{1 - \| (RQ^{-\top} - R_s)\Sigma_s^{-1/2} \|_{\text{op}}}, \\
\| \Sigma_s^{1/2} Q^{-\top} Q^{-\top} \Sigma_s^{1/2} - \Sigma_s \|_{\text{op}} & \leq \frac{\| L(Q - Q)\Sigma_s^{1/2} \|_{\text{op}}}{1 - \| (LQ - L_s)\Sigma_s^{-1/2} \|_{\text{op}}}.
\end{align*}
\]

Proof. Insert \( R^\top R(R^\top R)^{-1} \), and use the relation \( \| AB \|_{\text{op}} \leq \| A \|_{\text{op}} \| B \|_{\text{op}} \) to obtain
\[
\begin{align*}
\| \Sigma_s^{1/2} \bar{Q}^{-1} Q \Sigma_s^{1/2} - \Sigma_s \|_{\text{op}} &= \| \Sigma_s^{1/2} (\bar{Q}^{-1} - Q^{-1}) R^\top R(R^\top R)^{-1} Q \Sigma_s^{1/2} \|_{\text{op}} \\
&\leq \| R(\bar{Q}^{-\top} - Q^{-\top})\Sigma_s^{1/2} \|_{\text{op}} \| R(R^\top R)^{-1} Q \Sigma_s^{1/2} \|_{\text{op}} \\
&= \| R(\bar{Q}^{-\top} - Q^{-\top})\Sigma_s^{1/2} \|_{\text{op}} \| RQ^{-\top}((RQ^{-\top})^\top RQ^{-\top})^{-1}\Sigma_s^{1/2} \|_{\text{op}} \\
&\leq \frac{\| R(\bar{Q}^{-\top} - Q^{-\top})\Sigma_s^{1/2} \|_{\text{op}}}{1 - \| (RQ^{-\top} - R_s)\Sigma_s^{-1/2} \|_{\text{op}}},
\end{align*}
\]
where the last line uses Lemma 12.

Similarly, insert \( L^\top L(L^\top L)^{-1} \), and use the relation \( \| AB \|_{\text{op}} \leq \| A \|_{\text{op}} \| B \|_{\text{op}} \) to obtain
\[
\begin{align*}
\| \Sigma_s^{1/2} \bar{Q}^{-\top} Q^{-\top} \Sigma_s^{1/2} - \Sigma_s \|_{\text{op}} &= \| \Sigma_s^{1/2} (\bar{Q}^{-\top} - Q^{-\top}) L^\top L(L^\top L)^{-1} Q^{-\top} \Sigma_s^{1/2} \|_{\text{op}} \\
&\leq \| L(Q - Q)\Sigma_s^{1/2} \|_{\text{op}} \| L(L^\top L)^{-1} Q^{-\top} \Sigma_s^{1/2} \|_{\text{op}} \\
&= \| L(Q - Q)\Sigma_s^{1/2} \|_{\text{op}} \| LQ((LQ)^\top LQ)^{-1}\Sigma_s^{1/2} \|_{\text{op}} \\
&\leq \frac{\| L(Q - Q)\Sigma_s^{1/2} \|_{\text{op}}}{1 - \| (LQ - L_s)\Sigma_s^{-1/2} \|_{\text{op}}},
\end{align*}
\]
where the last line uses Lemma 12. \( \square \)

A.3 Partial Frobenius norm
We introduce the partial Frobenius norm
\[
\| X \|_{F,r} := \sqrt{\sum_{i=1}^{r} \sigma_i^2(X) = \| P_r(X) \|_F}
\]  
(41)
as the \( \ell_2 \) norm of the vector composed of the top-\( r \) singular values of the matrix \( X \), or equivalently as the Frobenius norm of the rank-\( r \) approximation \( P_r(X) \) defined in (5). It is straightforward to verify that \( \| \cdot \|_{F,r} \) is a norm; see also [Maz16]. The following lemma provides several equivalent and useful characterizations of this partial Frobenius norm.

Lemma 15. For any \( X \in \mathbb{R}^{n_1 \times n_2} \), one has
\[
\| X \|_{F,r} = \max_{\tilde{V} \in \mathbb{R}^{n_2 \times r}: \tilde{V}^\top \tilde{V} = I_r} \| X \tilde{V} \|_F = \max_{\tilde{X} \in \mathbb{R}^{n_1 \times n_2}: \| \tilde{X} \|_F \leq \text{rank}(\tilde{X}) \leq r} \| (X, \tilde{X}) \| = \max_{\tilde{R} \in \mathbb{R}^{n_2 \times r}: \| \tilde{R} \|_{\text{op}} \leq 1} \| X \tilde{R} \|_F.
\]  
(42a)  
(42b)  
(42c)
Proof. The first representation (42a) follows immediately from the extremal partial trace identity; see [Maz16, Proposition 4.4], by noticing the following relation
\[
\sum_{i=1}^r \sigma_i^2(X) = \max_{\mathbb{V} \subseteq \mathbb{R}^{n \times n} : \text{dim}(\mathbb{V}) = r} \text{tr}(X^T X | \mathbb{V}) = \max_{\mathbb{V} \in \mathbb{R}^{n \times r} : \mathbb{V}^T \mathbb{V} = I_r} \|X \bar{\mathbb{V}}\|_F^2.
\]
Here the partial trace over a vector space \(\mathbb{V}\) is defined as
\[
\text{tr}(X^T X | \mathbb{V}) := \sum_{i=1}^r \bar{v}_i^T X^T X \bar{v}_i,
\]
where \(\{\bar{v}_i\}_{1 \leq i \leq r}\) 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 \(\bar{X} \in \mathbb{R}^{n_1 \times n_2}\) obeying \(\text{rank}(\bar{X}) \leq r\) and \(\|\bar{X}\|_F \leq 1\), denoting \(\bar{X} = \bar{U} \bar{\Sigma} \bar{V}^T\) as its compact SVD, one has
\[
|\langle X, \bar{X} \rangle| = |\langle X, \bar{U} \bar{\Sigma} \bar{V}^T \rangle| = |\langle X \bar{V}, \bar{U} \Sigma \bar{V}^T \rangle| \leq \|X \bar{V}\|_F \|\bar{U} \Sigma\|_F \leq \|X\|_{F,r},
\]
where the last inequality follows from (42a). In addition, the maximum in (42b) is attained at \(\bar{X} = P_r(X)/\|P_r(X)\|_F\).

To prove the third representation (42c), for any \(\bar{R} \in \mathbb{R}^{r \times r}\) obeying \(\|\bar{R}\|_{op} \leq 1\), combine the variational representation of the Frobenius norm and (42b) to obtain
\[
\|X \bar{R}\|_F = \max_{\bar{L} \in \mathbb{R}^{n \times r} : \|\bar{L}\|_F \leq 1} |\langle X \bar{R}, \bar{L} \rangle| = \max_{\bar{L} \in \mathbb{R}^{n \times r} : \|\bar{L}\|_F \leq 1} |\langle X, \bar{L} \bar{R}^T \rangle| \leq \|X\|_{F,r},
\]
where the last inequality follows from (42b). In addition, the maximum in (42c) is attained at \(\bar{R} = \bar{V}\), where \(\bar{V}\) denotes the top-\(r\) right singular vectors of \(X\). \(\Box\)

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
\[
\sigma_1(X) = \|X v_1\|_2 = \max_{\bar{v} \in \mathbb{R}^{n_2} : \|\bar{v}\|_2 = 1} \|X \bar{v}\|_2,
\]
where \(v_1\) denotes the top right singular vector of \(X\). Assume that the statement holds for \(\|\cdot\|_{F,r-1}\). Now consider \(\|\cdot\|_{F,r}\). For any \(\bar{V} \in \mathbb{R}^{n_2 \times r}\) such that \(\bar{V}^T \bar{V} = I_r\), we can first pick \(\bar{v}_2, \ldots, \bar{v}_r\) as a set of orthonormal vectors in the column space of \(\bar{V}\) that are orthogonal to \(v_1\), and then pick \(\bar{v}_1\) via the Gram-Schmidt process, so that \(\{\bar{v}_i\}_{i=1}^r\) provides an orthonormal basis of the column space of \(\bar{V}\). Further, by the orthogonality of \(\bar{V}\), there exists an orthonormal matrix \(O\) such that
\[
\bar{V} = [\bar{v}_1, \ldots, \bar{v}_r] O.
\]
Combining this formula with the induction hypothesis yields
\[
\|X \bar{V}\|_F^2 = \|X [\bar{v}_1, \ldots, \bar{v}_r]\|_F^2
\]
\[
= \|X \bar{v}_1\|_2^2 + \|X [\bar{v}_2, \ldots, \bar{v}_r]\|_F^2
\]
\[
= \|X \bar{v}_1\|_2^2 + \|X - P_1(X) [\bar{v}_2, \ldots, \bar{v}_r]\|_{F,r-1}^2
\]
\[
\leq \sigma_1^2(X) + \|X - P_1(X)\|_{F,r-1}^2,
\]
where the first line holds since \(O\) is orthonormal, the third line holds since \(P_1(X) [\bar{v}_2, \ldots, \bar{v}_r] = 0\), 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 \(\bar{V} = V\), where \(V\) denotes the top-\(r\) right singular vectors of \(X\). This finishes the proof.
Recall that $\mathcal{P}_r(X)$ denotes the best rank-$r$ approximation of $X$ under the Frobenius norm. It turns out that $\mathcal{P}_r(X)$ is also the best rank-$r$ approximation of $X$ under the partial Frobenius norm $\|\cdot\|_{F,r}$. This claim is formally stated below; see also [Maz16, Theorem 4.21].

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

$$
\mathcal{P}_r(X) = \arg \min_{\hat{X} \in \mathbb{R}^{n_1 \times n_2} : \text{rank}(\hat{X}) \leq r} \|X - \hat{X}\|_{F,r}.
$$

**Proof.** For any $\tilde{X}$ of rank at most $r$, invoke Weyl’s inequality to obtain $\sigma_{r+i}(X) \leq \sigma_i(X - \tilde{X}) + \sigma_{r+1}(\tilde{X}) = \sigma_i(X - \tilde{X})$, for $i = 1, \ldots, r$. Thus one has

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

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

**B Proof for Low-Rank Matrix Factorization**

**B.1 Proof of Proposition 2**

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

$$
\nabla_L \mathcal{L}(F) = (LR^T - X_\star)R, \quad \nabla_R \mathcal{L}(F) = (LR^T - X_\star)^T L,
$$

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

$$
\nabla_{L,L}^2 \mathcal{L}(F) = (R^T R) \otimes I_{n_1}, \quad \nabla_{R,R}^2 \mathcal{L}(F) = (L^T L) \otimes I_{n_2}.
$$

Viewed in the vectorized form, the **ScaledGD** method in (3) can be rewritten as

$$
\begin{align*}
\text{vec}(L_{t+1}) &= \text{vec}(L_t) - \eta (R^T R_t)^{-1} \otimes I_{n_1} \text{vec}((L_t R^T_t - X_\star)R_t) \\
&= \text{vec}(L_t) - \eta (\nabla_{L,L}^2 \mathcal{L}(F_t))^{-1} \text{vec}(\nabla_L \mathcal{L}(F_t)), \\
\text{vec}(R_{t+1}) &= \text{vec}(R_t) - \eta (L^T L_t)^{-1} \otimes I_{n_2} \text{vec}((L_t R^T_t - X_\star)^T L_t) \\
&= \text{vec}(R_t) - \eta (\nabla_{R,R}^2 \mathcal{L}(F_t))^{-1} \text{vec}(\nabla_R \mathcal{L}(F_t)).
\end{align*}
$$

**B.2 Proof of Theorem 5**

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

1. $\text{dist}(F_t, F_\star) \leq (1 - 0.7\eta)^t \text{dist}(F_0, F_\star) \leq 0.1(1 - 0.7\eta)^t \sigma_r(X_\star)$, and

2. the optimal alignment matrix $Q_t$ between $F_t$ and $F_\star$ 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 $\text{dist}(F_0, F_\star) \leq 0.1\sigma_r(X_\star)$. We therefore concentrate on the induction step. Suppose that the $t$-th iterate $F_t$ obeys the aforementioned induction hypotheses. Our goal is to show that $F_{t+1}$ continues to satisfy those.

For notational convenience, denote $L := L_t Q_t$, $R := R_t Q_t^{-T}$, $\Delta_L := L - L_\star$, $\Delta_R := R - R_\star$, and $\epsilon := 0.1$. By the definition of $\text{dist}(F_{t+1}, F_\star)$, one has

$$
\text{dist}^2(F_{t+1}, F_\star) \leq \left\| (L_{t+1} Q_t - L_\star) \Sigma_\star^{1/2} \right\|_F^2 + \left\| (R_{t+1} Q_t^{-T} - R_\star) \Sigma_\star^{1/2} \right\|_F^2,
$$

(43)
where we recall that $Q_t$ is the optimal alignment matrix between $F_t$ and $F_*$. Utilize the ScaledGD update rule (30) and the decomposition $LR^\top - X_* = \Delta_L R^\top + L, \Delta_R^\top$ to obtain

\[
(L_{t+1}Q_t - L_*)\Sigma_1^{1/2} = (L - \eta(LR^\top - X_*)R(R^\top R)^{-1} - L_*)\Sigma_1^{1/2} = (\Delta_L - \eta(\Delta_L R^\top + L, \Delta_R^\top)R(R^\top R)^{-1})\Sigma_1^{1/2} = (1 - \eta)\Delta_L^{1/2} - \eta L, \Delta_L^\top R(R^\top R)^{-1}\Sigma_1^{1/2}.
\]

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

\[
\left\|(L_{t+1}Q_t - L_*)\Sigma_1^{1/2}\right\|_F^2 = (1 - \eta)^2 \text{tr}\left(\Delta_L \Sigma_1 \Delta_L^\top\right) - 2\eta(1 - \eta) \text{tr}\left(L, \Delta_L^\top R(R^\top R)^{-1}\Sigma_1 \Delta_L^\top\right)
\]

\[
+ \eta^2 \left\|L, \Delta_L^\top R(R^\top R)^{-1}\Sigma_1^{1/2}\right\|_F^2.
\]

The first term $\text{tr}(\Delta_L \Sigma_1 \Delta_L^\top)$ is closely related to $\text{dist}(F_t, F_*)$, and hence our focus will be on relating $\mathfrak{M}_1$ and $\mathfrak{M}_2$ to $\text{dist}(F_t, F_*)$. Let $L$ and $R$ be aligned with $L_*$ and $R_*$, Lemma 10 tells that $\Sigma, \Delta_\Sigma^\top L = R^\top \Delta_R \Sigma_*$. This together with $L_* = L - \Delta_L$ allows us to rewrite $\mathfrak{M}_1$ as

\[
\mathfrak{M}_1 = \text{tr}\left(R(R^\top R)^{-1}\Sigma, \Delta_L^\top L, \Delta_R^\top\right) = \text{tr}\left(R(R^\top R)^{-1}\Sigma, \Delta_L^\top L \Delta_R^\top\right) - \text{tr}\left(R(R^\top R)^{-1}\Sigma, \Delta_L^\top \Delta_R \Sigma_*^\top\right).
\]

Moving on to $\mathfrak{M}_2$, we can utilize the fact $L_*^\top L_* = \Sigma_*$ and the decomposition $\Sigma_* = R^\top R - (R^\top R - \Sigma_*)$ to obtain

\[
\mathfrak{M}_2 = \text{tr}\left(R(R^\top R)^{-1}\Sigma, (R^\top R)^{-1}R^\top \Delta_R \Sigma_*^\top\right) = \text{tr}\left(R(R^\top R)^{-1}R^\top \Delta_R \Sigma_*^\top\right) - \text{tr}\left(R(R^\top R)^{-1}(R^\top R - \Sigma_*)(R^\top R)^{-1}R^\top \Delta_R \Sigma_*^\top\right).
\]

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

\[
\left\|(L_{t+1}Q_t - L_*)\Sigma_1^{1/2}\right\|_F^2 = (1 - \eta)^2 \text{tr}\left(\Delta_L \Sigma_1 \Delta_L^\top\right) - \eta(2 - 3\eta) \text{tr}\left(R(R^\top R)^{-1}R^\top \Delta_R \Sigma_*^\top\right) + 2\eta(1 - \eta) \text{tr}\left(R(R^\top R)^{-1}\Sigma, \Delta_L^\top \Delta_L^\top\right) - \eta^2 \text{tr}\left(R(R^\top R)^{-1}(R^\top R - \Sigma_*)(R^\top R)^{-1}R^\top \Delta_R \Sigma_*^\top\right).
\]

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

1. Notice that $\mathfrak{g}_1$ is the inner product of two positive semi-definite matrices $R(R^\top R)^{-1}R^\top$ and $\Delta_R \Sigma_*^\top$. Consequently we have $\mathfrak{g}_1 \geq 0$.

2. To control $\mathfrak{g}_2$, we need certain control on $\|\Delta_L \Sigma_*^{-1/2}\|_{op}$ and $\|\Delta_R \Sigma_*^{-1/2}\|_{op}$. The first induction hypothesis is $\text{dist}(F_t, F_*) = \sqrt{\|\Delta_L \Sigma_*^{-1/2}\|_{op}^2 + \|\Delta_R \Sigma_*^{-1/2}\|_{op}^2} \leq \epsilon_\sigma_r(X_*)$ together with the relation $\|AB\|_F \geq \|A\|_F \sigma_r(B)$ tells that

\[
\sqrt{\|\Delta_L \Sigma_*^{-1/2}\|_F^2 + \|\Delta_R \Sigma_*^{-1/2}\|_F^2} \sigma_r(X_*) \leq \epsilon_\sigma_r(X_*).
\]
In light of the relation $\|A\|_F \leq \|A\|_F$, this further implies
\[
\|\Delta L \Sigma_\ast^{-1/2}\|_\omega \vee \|\Delta R \Sigma_\ast^{-1/2}\|_\omega \leq \epsilon. \tag{45}
\]

Invoke Lemma 12 to see
\[
\|R(R^\top R)^{-1/2} \Sigma_\ast^{1/2}\|_\omega \leq \frac{1}{1 - \epsilon}.
\]

With these consequences, one can bound $|\tilde{F}_2|$ by
\[
|\tilde{F}_2| = \left| \text{tr} \left( \Sigma_\ast^{-1/2} \Delta R R(R^\top R)^{-1} \Sigma_\ast \Delta L \Sigma_\ast^{1/2} \right) \right|
\leq \left\| \Sigma_\ast^{-1/2} \Delta R R(R^\top R)^{-1} \Sigma_\ast^{1/2} \right\|_\omega \text{tr} \left( \Sigma_\ast^{-1/2} \Delta L \Sigma_\ast^{1/2} \right)
\leq \|\Delta R \Sigma_\ast^{-1/2}\|_\omega \|R(R^\top R)^{-1/2} \Sigma_\ast^{1/2}\|_\omega \text{tr} \left( \Delta L \Sigma_\ast \Delta L^\top \right)
\leq \frac{1}{1 - \epsilon} \text{tr} \left( \Delta L \Sigma_\ast \Delta L^\top \right).
\]

Further notice that
\[
\|\Sigma_\ast^{-1/2} (R^\top R - \Sigma_\ast) \Sigma_\ast^{-1/2}\|_\omega = \left\| \Sigma_\ast^{-1/2} (R^\top R \Delta R + \Delta R^\top R_\ast + \Delta R^\top \Delta R) \Sigma_\ast^{-1/2} \right\|_\omega
\leq 2 \|\Delta R \Sigma_\ast^{-1/2}\|_\omega + \|\Delta R \Sigma_\ast^{-1/2}\|_\omega^2
\leq 2 \epsilon + \epsilon^2.
\]

Take the preceding two bounds together to arrive at
\[
|\tilde{F}_2| \leq \frac{2 \epsilon + \epsilon^2}{1 - \epsilon} \text{tr} \left( \Delta R \Sigma_\ast \Delta R^\top \right).
\]

Combining the bounds for $\tilde{F}_1, \tilde{F}_2, \tilde{F}_3$, one has
\[
\|(L_{t+1} Q_t - L_\ast) \Sigma_\ast^{1/2}\|_F = \|(1 - \eta) \Delta L \Sigma_\ast^{1/2} - \eta \Delta L R(R^\top R)^{-1} \Sigma_\ast^{1/2}\|_F
\leq \left( (1 - \eta)^2 + \frac{2 \epsilon}{1 - \epsilon} \eta (1 - \eta) \right) \text{tr} \left( \Delta L \Sigma_\ast \Delta L^\top \right) + \frac{2 \epsilon + \epsilon^2}{(1 - \epsilon)^2} \eta^2 \text{tr} \left( \Delta R \Sigma_\ast \Delta R^\top \right) \tag{46}
\]

A similarly bound holds for the second square $\|(R_{t+1} Q_t^\top - R_\ast) \Sigma_\ast^{1/2}\|_F^2$ in (43). Therefore we obtain
\[
\left\| (L_{t+1} Q_t - L_\ast) \Sigma_\ast^{1/2} \right\|_F^2 + \left\| (R_{t+1} Q_t^\top - R_\ast) \Sigma_\ast^{1/2} \right\|_F^2 \leq \rho^2(\eta; \epsilon) \text{dist}^2(F_t, F_\ast),
\]
where we identify
\[
\text{dist}^2(F_t, F_\ast) = \text{tr} \left( \Delta L \Sigma_\ast \Delta L^\top \right) + \text{tr} \left( \Delta R \Sigma_\ast \Delta R^\top \right) \tag{47}
\]
and the contraction rate $\rho^2(\eta; \epsilon)$ is given by
\[
\rho^2(\eta; \epsilon) := (1 - \eta)^2 + \frac{2 \epsilon}{1 - \epsilon} \eta (1 - \eta) + \frac{2 \epsilon + \epsilon^2}{(1 - \epsilon)^2} \eta^2.
\]
With \( \epsilon = 0.1 \) and \( 0 < \eta \leq 2/3 \), one has \( \rho(\eta; \epsilon) \leq 1 - 0.7\eta \). Thus we conclude that

\[
\text{dist}(F_{t+1}, F_*) \leq \sqrt{\left( (L_{t+1} Q_t - L_*) \Sigma_{1/2} \right)^2_F + \left( (R_{t+1} Q_t^\top - R_*) \Sigma_{1/2} \right)^2_F} \\
\leq (1 - 0.7\eta) \text{dist}(F_t, F_*) \\
\leq (1 - 0.7\eta)^{t+1} \text{dist}(F_0, F_*) \leq (1 - 0.7\eta)^t \sqrt{0.1} \sigma_r(X_*)
\]

This proves the first induction hypothesis. The existence of the optimal alignment matrix \( Q_{t+1} \) between \( F_{t+1} \) and \( F_* \) 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

\[
\|L_t R_t^\top - X_*\|_F \leq \left( 1 + \frac{\epsilon}{2} \right) \left( \|\Delta_L \Sigma_{1/2}\|_F + \|\Delta_R \Sigma_{1/2}\|_F \right) \\
\leq \left( 1 + \frac{\epsilon}{2} \right) \sqrt{2} \text{dist}(F_t, F_*) \\
\leq 1.5 \text{dist}(F_t, F_*)
\]

(48)

Here, the second line follows from the elementary inequality \( a + b \leq \sqrt{2(a^2 + b^2)} \) and the expression of \( \text{dist}(F_t, F_*) \) 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 \( A(\cdot) \) obeys the 2r-RIP with a constant \( \delta_{2r} \). Then for any \( X_1, X_2 \in \mathbb{R}^{n_1 \times n_2} \) of rank at most \( r \), one has

\[
|\langle A(X_1), A(X_2) \rangle - \langle X_1, X_2 \rangle| \leq \delta_{2r} \| X_1 \|_F \| X_2 \|_F,
\]

which can be stated equivalently as

\[
|\text{tr} \left( (A^* A - I)(X_1 X_2^\top) \right) | \leq \delta_{2r} \| X_1 \|_F \| X_2 \|_F.
\]

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

\[
\|(A^* A - I)(X_1) R\|_F \leq \delta_{2r} \| X_1 \|_F \| R \|_{op}.
\]

(50)

This is due to the fact that

\[
\|(A^* A - I)(X_1) R\|_F = \max_{L : \|L\|_{op} \leq r} \text{tr} \left( (A^* A - I)(X_1) R \tilde{L}^\top \right) \\
\leq \max_{L : \|L\|_{op} \leq r} \delta_{2r} \| X_1 \|_F \| R \|_{op} \| \tilde{L}^\top \|_F \\
= \delta_{2r} \| X_1 \|_F \| R \|_{op}.
\]

Here, the first line follows from the variational representation of the Frobenius norm, the second line follows from (49), and the third line follows from the relation \( \|AB\|_F \leq \|A\|_{op} \|B\|_F \).

C.1 Proof of Lemma 1

The proof mostly mirrors that in Section B.2. First, in view of the condition \( \text{dist}(F_t, F_*) \leq 0.1 \sigma_r(X_*) \) and Lemma 9, one knows that \( Q_t \), the optimal alignment matrix between \( F_t \) and \( F_* \), exists. Therefore, for
notational convenience, denote $L := L_{t}Q_{t}$, $R := R_{t}Q_{t}^{\top}$, $\Delta_{L} := L - L_{*}$, $\Delta_{R} := R - R_{*}$, and $\epsilon := 0.1$. Similar to the derivation in (45), we have

$$\|\Delta_{L}\Sigma_{\ast}^{1/2}\|_{\text{op}} + \|\Delta_{R}\Sigma_{\ast}^{1/2}\|_{\text{op}} \leq \epsilon. \quad (51)$$

The conclusion $\|L_{t}R_{t}^{\top} - X_{\ast}\|_{F} \leq 1.5 \text{dist}(F_{t}, F_{t})$ 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 dist$(F_{t+1}, F_{t})$ that

$$\text{dist}^{2}(F_{t+1}, F_{t}) \leq \left\|(L_{t+1}Q_{t} - L_{*})\Sigma_{\ast}^{1/2}\right\|_{F}^{2} + \left\|(R_{t+1}Q_{t}^{\top} - R_{*})\Sigma_{\ast}^{1/2}\right\|_{F}^{2}. \quad (52)$$

Apply the update rule (15) and the decomposition $LR_{t}^{\top} - X_{\ast} = \Delta_{L}R_{t}^{\top} + L_{*}\Delta_{R}^{\top}$ to obtain

$$(L_{t+1}Q_{t} - L_{*})\Sigma_{\ast}^{1/2} = (L - \eta A_{\ast}A(LR^{\top} - X_{\ast})R(R^{\top}R)^{-1} - L_{*})\Sigma_{\ast}^{1/2}$$

$$= (\Delta_{L} - \eta(LR^{\top} - X_{\ast})R(R^{\top}R)^{-1} - \eta(A_{\ast}A - I)(LR^{\top} - X_{\ast})R(R^{\top}R)^{-1})\Sigma_{\ast}^{1/2}$$

$$= (1 - \eta)\Delta_{L}\Sigma_{\ast}^{1/2} - \eta L_{*}\Delta_{R}^{\top}R(R^{\top}R)^{-1}\Sigma_{\ast}^{1/2} - \eta(A_{\ast}A - I)(LR^{\top} - X_{\ast})R(R^{\top}R)^{-1}\Sigma_{\ast}^{1/2}.$$

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

$$\left\|(L_{t+1}Q_{t} - L_{*})\Sigma_{\ast}^{1/2}\right\|_{F}^{2} = \left\{(1 - \eta)\Delta_{L}\Sigma_{\ast}^{1/2} - \eta L_{*}\Delta_{R}^{\top}R(R^{\top}R)^{-1}\Sigma_{\ast}^{1/2}\right\}_{1}^{2}$$

$$- 2\eta(1 - \eta) \text{tr} \left((A_{\ast}A - I)(LR^{\top} - X_{\ast})R(R^{\top}R)^{-1}\Sigma_{\ast}\Delta_{L}^{\top}\right) \quad \text{at}_{2}$$

$$+ 2\eta^{2} \text{tr} \left((A_{\ast}A - I)(LR^{\top} - X_{\ast})R(R^{\top}R)^{-1}\Sigma_{\ast}(R^{\top}R)^{-1}R^{\top}\Delta_{R}L_{*}^{\top}\right) \quad \text{at}_{3}$$

$$+ \eta^{2} \left\|(A_{\ast}A - I)(LR^{\top} - X_{\ast})R(R^{\top}R)^{-1}\Sigma_{\ast}^{1/2}\right\|_{F}^{2}. \quad \text{at}_{4}$$

In what follows, we shall control the four terms separately, of which $\text{at}_{1}$ is the main term, and $\text{at}_{2}, \text{at}_{3}$ and $\text{at}_{4}$ are perturbation terms.

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

$$\text{at}_{1} \leq \left((1 - \eta)^{2} + \frac{2\epsilon}{1 - \epsilon}(1 - \eta)\right)\|\Delta_{L}\Sigma_{\ast}^{1/2}\|_{F}^{2} + \frac{2\epsilon + \epsilon^{2}}{(1 - \epsilon)^{2}}\|\Delta_{R}\Sigma_{\ast}^{1/2}\|_{F}^{2}.$$

2. For the second term $\text{at}_{2}$, decompose $LR_{t}^{\top} - X_{\ast} = \Delta_{L}R_{t}^{\top} + L_{*}\Delta_{R}^{\top} + \Delta_{L}\Delta_{R}^{\top}$ and apply the triangle inequality to obtain

$$|\text{at}_{2}| = \left|\text{tr} \left((A_{\ast}A - I)(\Delta_{L}R_{t}^{\top} + L_{*}\Delta_{R}^{\top} + \Delta_{L}\Delta_{R}^{\top})R(R^{\top}R)^{-1}\Sigma_{\ast}\Delta_{L}^{\top}\right)\right|$$

$$\leq \left|\text{tr} \left((A_{\ast}A - I)(\Delta_{L}R_{t}^{\top})R(R^{\top}R)^{-1}\Sigma_{\ast}\Delta_{L}^{\top}\right)\right|$$

$$+ \left|\text{tr} \left((A_{\ast}A - I)(L_{*}\Delta_{R}^{\top})R(R^{\top}R)^{-1}\Sigma_{\ast}\Delta_{L}^{\top}\right)\right|$$

$$+ \left|\text{tr} \left((A_{\ast}A - I)(\Delta_{L}\Delta_{R}^{\top})R(R^{\top}R)^{-1}\Sigma_{\ast}\Delta_{L}^{\top}\right)\right|.$$

Invoke Lemma 17 to further obtain

$$|\text{at}_{2}| \leq \delta_{2r}(\|\Delta_{L}R_{t}^{\top}\|_{F} + \|L_{*}\Delta_{R}^{\top}\|_{F} + \|\Delta_{L}\Delta_{R}^{\top}\|_{F})\|R(R^{\top}R)^{-1}\Sigma_{\ast}\Delta_{L}^{\top}\|_{F}$$

$$\leq \delta_{2r}(\|\Delta_{L}R_{t}^{\top}\|_{F} + \|L_{*}\Delta_{R}^{\top}\|_{F} + \|\Delta_{L}\Delta_{R}^{\top}\|_{F})\|R(R^{\top}R)^{-1}\Sigma_{\ast}^{1/2}\|_{\text{op}}\|\Delta_{L}\Sigma_{\ast}^{1/2}\|_{F},$$
where the second line follows from the relation $\|AB\|_F \leq \|A\|_{op} \|B\|_F$. Take the condition (51) and Lemmas 12 and 13 together to obtain

$$\left\| R(R^TR)^{-1}\Sigma^1_{R^*}\right\|_{op} \leq \frac{1}{1 - \epsilon};$$

$$\|\Delta_L R^*\|_F + \|L^* \Delta^T_R\|_F + \|\Delta_L \Delta^T_R\|_F \leq (1 + \frac{\epsilon}{2}) \left( \|\Delta_L \Sigma^1_{R^*}\|_F + \|\Delta_R \Sigma^1_{R^*}\|_F \right).$$

These consequences further imply that

$$|\mathcal{S}_2| \leq \frac{\delta_2r(2 + \epsilon)}{2(1 - \epsilon)} \left( \|\Delta_L \Sigma^1_{R^*}\|_F + \|\Delta_R \Sigma^1_{R^*}\|_F \right) \|\Delta_L \Sigma^1_{R^*}\|_F$$

$$= \frac{\delta_2r(2 + \epsilon)}{2(1 - \epsilon)} \left( \|\Delta_L \Sigma^1_{R^*}\|_F^2 + \|\Delta_L \Sigma^1_{R^*}\|_F \|\Delta_R \Sigma^1_{R^*}\|_F \right).$$

For the term $\|\Delta_L \Sigma^1_{R^*}\|_F \|\Delta_R \Sigma^1_{R^*}\|_F$, we can apply the elementary inequality $2ab \leq a^2 + b^2$ to see

$$\|\Delta_L \Sigma^1_{R^*}\|_F \|\Delta_R \Sigma^1_{R^*}\|_F \leq \frac{1}{2} \|\Delta_L \Sigma^1_{R^*}\|_F^2 + \frac{1}{2} \|\Delta_R \Sigma^1_{R^*}\|_F^2.$$

The preceding two bounds taken collectively yield

$$|\mathcal{S}_2| \leq \frac{\delta_2r(2 + \epsilon)}{2(1 - \epsilon)} \left( \frac{3}{2} \|\Delta_L \Sigma^1_{R^*}\|_F^2 + \frac{1}{2} \|\Delta_L \Sigma^1_{R^*}\|_F^2 \right).$$

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

$$|\mathcal{S}_3| \leq \|\Delta_L R^*\|_F + \|L^* \Delta^T_R\|_F + \|\Delta_L \Delta^T_R\|_F \left\| R(R^TR)^{-1}\Sigma^1_{R^*}(R^TR)^{-1}R^* \Delta_R L^*_R \right\|_F$$

$$\leq \frac{\delta_2r(2 + \epsilon)}{2(1 - \epsilon)} \left( \|\Delta_L \Sigma^1_{R^*}\|_F + \|\Delta_R \Sigma^1_{R^*}\|_F \right) \|\Delta_R \Sigma^1_{R^*}\|_F$$

$$\leq \frac{\delta_2r(2 + \epsilon)}{2(1 - \epsilon)} \left( \|\Delta_L \Sigma^1_{R^*}\|_F^2 + \frac{1}{2} \|\Delta_R \Sigma^1_{R^*}\|_F^2 \right).$$

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

$$\sqrt{\mathcal{S}_4} = \left\| (A^*A - I)(LR^* - X^* R(R^TR)^{-1}\Sigma^1_{R^*}) \right\|_F$$

$$\leq \left\| (A^*A - I)(\Delta_L R^*) R(R^TR)^{-1}\Sigma^1_{R^*} \right\|_F$$

$$+ \left\| (A^*A - I)(L^* \Delta_R^T) R(R^TR)^{-1}\Sigma^1_{R^*} \right\|_F$$

$$+ \left\| (A^*A - I)(\Delta_L \Delta_R^T) R(R^TR)^{-1}\Sigma^1_{R^*} \right\|_F;$$

where once again we use the decomposition $LR^* - X^* = \Delta_L R^* + L^* \Delta_R^T + \Delta_L \Delta_R^T$. Use (50) to see that

$$\sqrt{\mathcal{S}_4} \leq \delta_2r \left( \|\Delta_L R^*\|_F + \|L^* \Delta_R^T\|_F + \|\Delta_L \Delta_R^T\|_F \right) \left\| R(R^TR)^{-1}\Sigma^1_{R^*} \right\|_{op}. $$

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

$$\sqrt{\mathcal{S}_4} \leq \frac{\delta_2r(2 + \epsilon)}{2(1 - \epsilon)} \left( \|\Delta_L \Sigma^1_{R^*}\|_F + \|\Delta_R \Sigma^1_{R^*}\|_F \right).$$

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

$$\mathcal{S}_4 \leq \frac{\delta_2r^2(2 + \epsilon)^2}{2(1 - \epsilon)} \left( \|\Delta_L \Sigma^1_{R^*}\|_F^2 + \|\Delta_R \Sigma^1_{R^*}\|_F^2 \right).$$
Taking the bounds for $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \mathcal{G}_4$ collectively yields

$$
\left\| (L_{t+1}Q_t - L_*)\Sigma_t^{1/2} \right\|_F^2 \leq \left( (1 - \eta)^2 + \frac{2\epsilon}{1 - \epsilon} \eta(1 - \eta) \right) \left\| \Delta_L \Sigma_1^{1/2} \right\|_F^2 + \frac{2\epsilon + \epsilon^2}{(1 - \epsilon)^2} \eta^2 \left\| \Delta_R \Sigma_1^{1/2} \right\|_F^2
+ \frac{\delta_{2r}(2 + \epsilon)}{1 - \epsilon} \eta(1 - \eta) \left( \frac{3}{2} \left\| \Delta_L \Sigma_t^{1/2} \right\|_F^2 + \frac{1}{2} \left\| \Delta_R \Sigma_t^{1/2} \right\|_F^2 \right)
+ \frac{\delta_{2r}(2 + \epsilon)^2}{(1 - \epsilon)^2} \eta^2 \left( \left\| \Delta_L \Sigma_t^{1/2} \right\|_F^2 + \left\| \Delta_R \Sigma_t^{1/2} \right\|_F^2 \right).
$$

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

$$
\left\| (L_{t+1}Q_t - L_*)\Sigma_t^{1/2} \right\|_F^2 + \left\| (R_{t+1}Q_t^{-\top} - R_*)\Sigma_t^{1/2} \right\|_F^2 \leq \rho^2(\eta; \epsilon, \delta_{2r}) \text{dist}^2(F_t, F_*),
$$

where the contraction rate is given by

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

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

$$
\text{dist}(F_{t+1}, F_*) \leq \sqrt{\left\| (L_{t+1}Q_t - L_*)\Sigma_t^{1/2} \right\|_F^2 + \left\| (R_{t+1}Q_t^{-\top} - R_*)\Sigma_t^{1/2} \right\|_F^2}
\leq (1 - 0.6\eta) \text{dist}(F_t, F_*).
$$

### C.2 Proof of Lemma 2

With the knowledge of $\| \cdot \|_{F,r}$, we are ready to establish the claimed result. Invoke Lemma 11 to relate $\text{dist}(F_0, F_*)$ to $\| L_0R_0^\top - X_* \|_F$, and use that $L_0R_0^\top - X_*$ has rank at most $2r$ to obtain

$$
\text{dist}(F_0, F_*) \leq \sqrt{\sqrt{2} + 1} \| L_0R_0^\top - X_* \|_F \leq \sqrt{2(\sqrt{2} + 1)} \| L_0R_0^\top - X_* \|_{F,r}.
$$

Note that $L_0R_0^\top$ is the best rank-$r$ approximation of $A^*A(X_*)$, and apply the triangle inequality combined with Lemma 16 to obtain

$$
\| L_0R_0^\top - X_* \|_{F,r} \leq \| A^*A(X_*) - L_0R_0^\top \|_{F,r} + \| A^*A(X_*) - X_* \|_{F,r}
\leq 2 \| (A^*A - I)(X_*) \|_{F,r} \leq 2\delta_{2r} \| X_* \|_F.
$$

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

$$
\| (A^*A - I)(X_*) \|_{F,r} = \max_{\tilde{R} \in \mathbb{R}^{n \times 2r}} \| \tilde{R} \|_{op} \leq 1 \| (A^*A - I)(X_*)\tilde{R} \|_F \leq \delta_{2r} \| X_* \|_F.
$$

As a result, one has

$$
\text{dist}(F_0, F_*) \leq 2\sqrt{2(\sqrt{2} + 1)}\delta_{2r} \| X_* \|_F \leq 5\delta_{2r} \sqrt{\kappa \sigma_r}(X_*).
$$

### D Proof for Robust PCA

We first establish a useful property regarding the truncation operator $T_{2\alpha}$. [\ldots]
Lemma 18. Given $S_\ast \in S_\alpha$ and $S = T_{2\alpha}[X_\ast + S_\ast - LR^T]$, one has

$$\|S - S_\ast\|_\infty \leq 2\|LR^T - X_\ast\|_\infty. \tag{53}$$

In addition, for any low-rank matrix $M = L_M R_M^T \in \mathbb{R}^{n_1 \times n_2}$ with $L_M \in \mathbb{R}^{n_1 \times r}, R_M \in \mathbb{R}^{n_2 \times r}$, one has

$$\|\langle S - S_\ast, M \rangle\| \leq \sqrt{3n} \nu \left( \|(L - L_\ast)\Sigma_\ast^{1/2}\|_F + \|(R - R_\ast)\Sigma_\ast^{1/2}\|_F \right) \|M\|_F$$

$$+ 2\sqrt{\alpha} \left( \frac{1}{\sqrt{n_1}} \|L_M\|_{2,\infty} \|R_M\|_F \wedge \frac{1}{\sqrt{n_2}} \|L_M\|_F \|R_M\|_{2,\infty} \right) \|LR^T - X_\ast\|_F,$$  \tag{54}

where $\nu$ obeys

$$\nu \geq \frac{\sqrt{n}}{2} \left( \|L\Sigma_{\ast}^{-1/2}\|_{2,\infty} + \|L_*\Sigma_{\ast}^{-1/2}\|_{2,\infty} \right) \sqrt{\frac{n_2}{2}} \left( \|R\Sigma_{\ast}^{-1/2}\|_{2,\infty} + \|R_*\Sigma_{\ast}^{-1/2}\|_{2,\infty} \right).$$

Proof. Denote $\Delta_L := L - L_\ast, \Delta_R := R - R_\ast$, and $\Delta_X := LR^T - X_\ast$. Let $\Omega, \Omega_\ast$ be the support of $S$ and $S_\ast$, respectively. As a result, $S - S_\ast$ is supported on $\Omega \cup \Omega_\ast$.

We start with proving the first claim, i.e. (53). For $(i, j) \in \Omega$, by the definition of $T_{2\alpha}[\cdot]$, we have $(S - S_\ast)_{i,j} = (-\Delta X)_{i,j}$. For $(i, j) \in \Omega_\ast \setminus \Omega$, one necessarily has $(S - S_\ast)_{i,j} = 0$ and therefore $(S - S_\ast)_{i,j} = -(S_\ast)_{i,j}$. Again by the definition of the operator $T_{2\alpha}[\cdot]$, we know $|S_\ast - \Delta X|_{i,j}$ is either smaller than $|S_\ast - \Delta X|_{i,|\Omega(\alpha_2)}$ or $|S_\ast - \Delta X|_{i,|\Omega(\alpha_1)}$. Furthermore, we know that $S_\ast$ contains at most $\alpha$-fraction nonzero entries per row and column. Consequently, one has $|S_\ast - \Delta X|_{i,j} \leq |\Delta X|_{i,|\Omega(\alpha_2)} \vee |\Delta X|_{i,|\Omega(\alpha_1)}$. Combining the two cases above, we conclude that

$$|S - S_\ast|_{i,j} \leq \begin{cases} |\Delta X|_{i,j}, & (i, j) \in \Omega \\ |\Delta X|_{i,j} + (|\Delta X|_{i,|\Omega(\alpha_2)} \vee |\Delta X|_{i,|\Omega(\alpha_1)}), & (i, j) \in \Omega_\ast \setminus \Omega \end{cases} \tag{55}$$

This immediately implies the $\ell_\infty$ norm bound (53).

Next, we prove the second claim (54). Recall that $S - S_\ast$ is supported on $\Omega \cup \Omega_\ast$. We then have

$$\|\langle S - S_\ast, M \rangle\| \leq \sum_{(i, j) \in \Omega} |S - S_\ast|_{i,j} |M|_{i,j} + \sum_{(i, j) \in \Omega_\ast \setminus \Omega} |S - S_\ast|_{i,j} |M|_{i,j}$$

$$\leq \sum_{(i, j) \in \Omega \cup \Omega_\ast} |\Delta X|_{i,j} |M|_{i,j} + \sum_{(i, j) \in \Omega_\ast \setminus \Omega} (|\Delta X|_{i,|\Omega(\alpha_2)} + |\Delta X|_{i,|\Omega(\alpha_1)}) |M|_{i,j},$$

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

$$\|\langle S - S_\ast, M \rangle\| \leq \sum_{(i, j) \in \Omega \cup \Omega_\ast} |\Delta X|_{i,j} |M|_{i,j} + \frac{1}{2\beta} \sum_{(i, j) \in \Omega \cup \Omega_\ast} (|\Delta X|_{i,|\Omega(\alpha_2)}^2 + |\Delta X|_{i,|\Omega(\alpha_1)}) + \beta \sum_{(i, j) \in \Omega_\ast \setminus \Omega} |M|_{i,j}^2.$$  

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

$$\mathfrak{A}_1 \leq \sqrt{3n} \nu \left( \|\Delta L\Sigma_{\ast}^{1/2}\|_F + \|\Delta R\Sigma_{\ast}^{1/2}\|_F \right) \|M\|_F.$$  

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

$$\mathfrak{A}_2 \leq 2\|\Delta X\|_F^2.$$  

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

$$\mathfrak{A}_3 \leq \alpha \left( n_1 \|L_M\|_{2,\infty}^2 \|R_M\|_F^2 \wedge n_2 \|L_M\|_F^2 \|R_M\|_{2,\infty} \right).$$  

38
Combine the pieces to reach
\[
|\langle S - S*, M \rangle| \leq \sqrt{3\alpha\nu} \left( \|\Delta_L \Sigma_*^{1/2}\|_F + \|\Delta_R \Sigma_*^{1/2}\|_F \right) \|M\|_F \\
+ \frac{\|\Delta_X\|_2^2}{\beta} + \beta \alpha \left( n_1 \|L_M\|_{2,\infty}^2 \|R_M\|_F^2 \wedge n_2 \|L_M\|_F^2 \|R_M\|_{2,\infty}^2 \right).
\]

One can then choose \( \beta \) optimally to yield
\[
|\langle S - S*, M \rangle| \leq \sqrt{3\alpha\nu} \left( \|\Delta_L \Sigma_*^{1/2}\|_F + \|\Delta_R \Sigma_*^{1/2}\|_F \right) \|M\|_F \\
+ 2\sqrt{\alpha} \left( \sqrt{n_1} \|L_M\|_{2,\infty} \|R_M\|_F \wedge \sqrt{n_2} \|L_M\|_F \|R_M\|_{2,\infty} \right) \|\Delta_X\|_F.
\]

This finishes the proof. \( \square \)

**Proof of Claim 1.** Use the decomposition \( \Delta_X = \Delta_L R^T + L_*\Delta_R = \Delta_L R^T + L\Delta_R \) to obtain
\[
|\Delta_X|_{i,j} \leq \|\Delta_L \Sigma_*^{1/2}\|_{i,2} \|R\Sigma_*^{-1/2}\|_{2,\infty} + \|L_*\Sigma_*^{-1/2}\|_{2,\infty} \|\Delta_R \Sigma_*^{1/2}\|_{j,2}, \quad \text{and}
\]
\[
|\Delta_X|_{i,j} \leq \|\Delta_L \Sigma_*^{1/2}\|_{i,2} \|R\Sigma_*^{-1/2}\|_{2,\infty} + \|L\Sigma_*^{-1/2}\|_{2,\infty} \|\Delta_R \Sigma_*^{1/2}\|_{j,2}.
\]

Take the average to yield
\[
|\Delta_X|_{i,j} \leq \frac{\nu}{\sqrt{n_2}} \|\Delta_L \Sigma_*^{1/2}\|_{i,2} + \frac{\nu}{\sqrt{n_1}} \|\Delta_R \Sigma_*^{1/2}\|_{j,2},
\]
where we have used the assumption on \( \nu \). With this upper bound on \( |\Delta_X|_{i,j} \) in place, we can further control \( \mathfrak{A}_1 \) as
\[
\mathfrak{A}_1 \leq \sum_{(i,j) \in \Omega \cup \Omega_*} \frac{\nu}{\sqrt{n_2}} \|\Delta_L \Sigma_*^{1/2}\|_{i,2} |M|_{i,j} + \sum_{(i,j) \in \Omega \cup \Omega_*} \frac{\nu}{\sqrt{n_1}} \|\Delta_R \Sigma_*^{1/2}\|_{j,2} |M|_{i,j} \\
\leq \left( \sqrt{\sum_{(i,j) \in \Omega \cup \Omega_*} \|\Delta_L \Sigma_*^{1/2}\|_{i,2}^2/n_2} + \sqrt{\sum_{(i,j) \in \Omega \cup \Omega_*} \|\Delta_R \Sigma_*^{1/2}\|_{j,2}^2/n_1} \right) \nu \|M\|_F.
\]

Regarding the first term, one has
\[
\sum_{(i,j) \in \Omega \cup \Omega_*} \|\Delta_L \Sigma_*^{1/2}\|_{i,2}^2 = \sum_{i=1}^{n_1} \sum_{j \in \Omega \cup \Omega_*} \|\Delta_L \Sigma_*^{1/2}\|_{i,2}^2 \\
\leq 3\alpha n_2 \sum_{i=1}^{n_1} \|\Delta_L \Sigma_*^{1/2}\|_{i,2}^2
\]
\[
= 3\alpha n_2 \|\Delta_L \Sigma_*^{1/2}\|_F^2,
\]
where the second line follows from the fact that \( \Omega \cup \Omega_* \) contains at most \( 3\alpha n_2 \) non-zero entries in each row. Similarly, we can show that
\[
\sum_{(i,j) \in \Omega \cup \Omega_*} \|\Delta_R \Sigma_*^{1/2}\|_{j,2}^2 \leq 3\alpha n_1 \|\Delta_R \Sigma_*^{1/2}\|_F^2.
\]

In all, we arrive at
\[
\mathfrak{A}_1 \leq \sqrt{3\alpha\nu} \left( \|\Delta_L \Sigma_*^{1/2}\|_F + \|\Delta_R \Sigma_*^{1/2}\|_F \right) \|M\|_F,
\]
which is the desired claim. \( \square \)
Proof of Claim 2. Recall that $(\Delta_X)_{i,(\alpha n_2)}$ denotes the $(\alpha n_2)$-th largest entry in the $i$-th row of $\Delta_X$. One necessarily has
\[
\alpha n_2 |(\Delta_X)_{i,(\alpha n_2)}|^2 \leq \| \Delta_X \|_2^2.
\]
As a result, we obtain
\[
\sum_{(i,j) \in \Omega_s \setminus \Omega} |(\Delta_X)_{i,(\alpha n_2)}|^2 \leq \sum_{(i,j) \in \Omega_s} |(\Delta_X)_{i,(\alpha n_2)}|^2
\]
\[
\leq \sum_{i=1}^{\alpha n_2} \sum_{j:(i,j) \in \Omega_s} \| (\Delta_X)_{i,j} \|_2^2
\]
\[
\leq \sum_{i=1}^{\alpha n_2} \| (\Delta_X)_{i,j} \|_2^2 = \| \Delta_X \|_F^2,
\]
where the last line follows from the fact that $\Omega_s$ contains at most $\alpha n_2$ nonzero entries in each row. Similarly one can show that
\[
\sum_{(i,j) \in \Omega_s \setminus \Omega} |(\Delta_X)_{i,(\alpha n_1),j}| \leq \| \Delta_X \|_F^2.
\]
Combining the above two bounds with the definition of $\mathcal{A}_2$ completes the proof.

Proof of Claim 3. By definition, $M = L_M R_M^T$, and hence one has
\[
\mathcal{A}_3 = \sum_{(i,j) \in \Omega_s \setminus \Omega} \| (L_M)_{i,(R_M)^T_j} \|^2_i \leq \sum_{(i,j) \in \Omega_s} \| (L_M)_{i,(R_M)^T_j} \|^2_i.
\]
We can further upper bound $\mathcal{A}_3$ as
\[
\mathcal{A}_3 \leq \sum_{(i,j) \in \Omega_s} \| (L_M)_{i,j} \|^2_2 \| (R_M)^T_j \|^2_2
\]
\[
\leq \sum_{i=1}^{\alpha n_2} \sum_{j:(i,j) \in \Omega_s} \| (L_M)_{i,j} \|^2_2 \| R_M \|^2_{2,\infty}
\]
\[
\leq \sum_{i=1}^{\alpha n_2} \alpha n_2 \| (L_M)_{i,j} \|^2_2 \| R_M \|^2_{2,\infty} = \alpha n_2 \| L_M \|^2_\infty \| R_M \|^2_{2,\infty},
\]
where the last line follows from the fact that $\Omega_s$ contains at most $\alpha n_2$ non-zero entries in each row. Similarly, one can obtain
\[
\mathcal{A}_3 \leq \alpha n_2 \| L_M \|^2_{2,\infty} \| R_M \|^2_\infty,
\]
which completes the proof.

D.1 Proof of Lemma 3

We begin with introducing several useful notations and facts. In view of the condition $\text{dist}(F_t, F_\ast) \leq 0.02 \sigma_r(X_\ast)$ and Lemma 9, one knows that $Q_t$, the optimal alignment matrix between $F_t$ and $F_\ast$ exists. Therefore, for notational convenience, denote $L := L_t Q_t, R := R_t Q_t^T, \Delta_L := L - L_\ast, \Delta_R := R - R_\ast, S := S_t = T_{2n}[X_\ast + S_\ast - LR^T], \epsilon := 0.02$. Similar to the derivation in (45), we have
\[
\| \Delta_L \Sigma_\ast^{-1/2} \|_{op} \vee \| \Delta_R \Sigma_\ast^{-1/2} \|_{op} \leq \epsilon.
\]
Moreover, the incoherence condition
\[
\sqrt{n_1} \| \Delta_L \Sigma_\ast^{1/2} \|_{2,\infty} \vee \sqrt{n_2} \| \Delta_R \Sigma_\ast^{1/2} \|_{2,\infty} \leq \sqrt{n_1} \sigma_r(X_\ast)
\]
implies
\[
\sqrt{n_1} \| \Delta_L \Sigma_*^{-1/2} \|_{2,\infty} \vee \sqrt{n_2} \| \Delta_R \Sigma_*^{-1/2} \|_{2,\infty} \leq \sqrt{\mu r},
\]
which combined with the triangle inequality further implies
\[
\sqrt{n_1} \| L \Sigma_*^{-1/2} \|_{2,\infty} \vee \sqrt{n_2} \| R \Sigma_*^{-1/2} \|_{2,\infty} \leq 2 \sqrt{\mu r}.
\]
The conclusion \( \| L_x R^T - X_* \|_F \leq 1.5 \text{dist}(F_t, F_*) \) 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 \( \text{dist}^2(F_{t+1}, F_*) \), one has
\[
\text{dist}^2(F_{t+1}, F_*) \leq \left\| (L_{t+1} Q_t - L_*) \Sigma_*^{-1/2} \right\|_F^2 + \left\| (R_{t+1} Q_{t}^T - R_*) \Sigma_*^{-1/2} \right\|_F^2.
\]
From now on, we focus on controlling the first square \( \| (L_{t+1} Q_t - L_*) \Sigma_*^{-1/2} \|_F^2 \). In view of the update rule (20), one has
\[
(L_{t+1} Q_t - L_*) \Sigma_*^{-1/2} = (L - \eta (L R^T + S - X_* - S_*) R (R^T R)^{-1} - L_*) \Sigma_*^{-1/2}
\]
\[
= \left( \Delta_L - \eta (L R^T - X_*) R (R^T R)^{-1} - \eta (S - S_*) R (R^T R)^{-1} \right) \Sigma_*^{-1/2}
\]
\[
= (1 - \eta) \Delta_L \Sigma_*^{-1/2} - \eta L_* \Delta_R^T R (R^T R)^{-1} \Sigma_*^{-1/2} - \eta (S - S_*) R (R^T R)^{-1} \Sigma_*^{-1/2}.
\]
Here, we use the notation introduced above and the decomposition \( L R^T - X_* = \Delta_L R^T + L_* \Delta_R^T \). Take the squared Frobenius norm of both sides of (61) to obtain
\[
\left\| (L_{t+1} Q_t - L_*) \Sigma_*^{-1/2} \right\|_F^2 = \left( \eta L_* \Delta_R^T R (R^T R)^{-1} \Sigma_*^{-1/2} \right)^2
\]
\[
- 2\eta(1 - \eta) \text{tr} \left( (S - S_*) R (R^T R)^{-1} \Sigma_* \Delta_L^T \right)
\]
\[
+ 2\eta^2 \text{tr} \left( (S - S_*) R (R^T R)^{-1} \Sigma_* (R^T R)^{-1} R^T \Delta_R L_*^T \right)
\]
\[
+ \eta^2 \left\| (S - S_*) R (R^T R)^{-1} \Sigma_*^{-1/2} \right\|_F^2.
\]
In the sequel, we shall bound the four terms separately, of which \( R_1 \) is the main term, and \( R_2, R_3 \) and \( R_4 \) are perturbation terms.

1. Notice that the main term \( R_1 \) has already been controlled in (46) under the condition (56). It obeys
\[
R_1 \leq \left( 1 - \eta \right)^2 + \frac{2\epsilon}{1 - \epsilon} \eta (1 - \eta) \left\| \Delta_L \Sigma_*^{-1/2} \right\|_F^2 + \frac{2\epsilon + \epsilon^2}{1 - \epsilon} \eta^2 \left\| \Delta_R \Sigma_*^{-1/2} \right\|_F^2.
\]
2. For the second term \( R_2 \), set \( M := \Delta_L \Sigma_* (R^T R)^{-1} R^T \) with \( L_M := \Delta_L \Sigma_* (R^T R)^{-1} \Sigma_*^{-1/2}, R_M := R \Sigma_*^{-1/2} \), and then invoke Lemma 18 with \( \nu := 3\sqrt{\mu r}/2 \) to see
\[
| R_2 | \leq \frac{3}{2} \sqrt{3\sqrt{\mu r}} \left( \left\| \Delta_L \Sigma_*^{-1/2} \right\|_F + \left\| \Delta_R \Sigma_*^{-1/2} \right\|_F \right) \left\| \Delta_L \Sigma_* (R^T R)^{-1} R^T \right\|_F
\]
\[
+ 2\sqrt{\alpha n_2} \left\| \Delta_L \Sigma_* (R^T R)^{-1} \Sigma_*^{-1/2} \right\|_F \left\| R \Sigma_*^{-1/2} \right\|_{2,\infty} \left\| LR^T - X_* \right\|_F.
\]
4. For the last term
3. The third term
arrive at
Take the condition (56) and Lemmas 12 and 13 together to obtain

\[
\|R(R^T R)^{-1} \Sigma_*^{1/2}\|_{op} \leq \frac{1}{1 - \epsilon};
\]
\[
\|\Sigma_*^{1/2}(R^T R)^{-1} \Sigma_*^{1/2}\|_{op} = \|R(R^T R)^{-1} \Sigma_*^{1/2}\|_{op} \leq \frac{1}{(1 - \epsilon)^2};
\]
\[
\|LR^T - X_*\|_{F} \leq (1 + \frac{\epsilon}{2}) \left(\|\Delta_L \Sigma_*^{1/2}\|_{F} + \|\Delta_R \Sigma_*^{1/2}\|_{F}\right).
\]

These consequences combined with the condition (59) yield

\[
|\mathcal{R}_2| \leq \frac{3\sqrt{3\alpha \mu \rho}}{2(1 - \epsilon)} \left(\|\Delta_L \Sigma_*^{1/2}\|_{F} + \|\Delta_R \Sigma_*^{1/2}\|_{F}\right) \|\Delta_L \Sigma_*^{1/2}\|_{F}
\]
\[
+ \frac{4\sqrt{\alpha \mu \rho}}{(1 - \epsilon)^{2}} \|\Delta_L \Sigma_*^{1/2}\|_{F}(1 + \frac{\epsilon}{2}) \left(\|\Delta_L \Sigma_*^{1/2}\|_{F} + \|\Delta_R \Sigma_*^{1/2}\|_{F}\right)
\]
\[
\leq \sqrt{\alpha \mu \rho} \frac{3\sqrt{3} + \frac{4(2 + \epsilon)}{2(1 - \epsilon)}}{2(1 - \epsilon)} \left(\|\Delta_L \Sigma_*^{1/2}\|_{F}^{2} + \|\Delta_L \Sigma_*^{1/2}\|_{F} \|\Delta_R \Sigma_*^{1/2}\|_{F}\right)
\]
\[
\leq \sqrt{\alpha \mu \rho} \frac{3\sqrt{3} + \frac{4(2 + \epsilon)}{2(1 - \epsilon)}}{2(1 - \epsilon)} \left(\frac{3}{2} \|\Delta_L \Sigma_*^{1/2}\|_{F}^{2} + \frac{1}{2} \|\Delta_R \Sigma_*^{1/2}\|_{F}^{2}\right),
\]

where the last inequality holds since \(2ab \leq a^{2} + b^{2}\).

3. The third term \(\mathcal{R}_3\) can be controlled similarly. Set \(M := L_* \Delta_L^T R(R^T R)^{-1} \Sigma_*(R^T R)^{-1} R^T\) with \(L_* := L_* \Sigma_*^{-1/2}\) and \(R_M := R(R^T R)^{-1} \Sigma_*(R^T R)^{-1} R^T \Delta_R \Sigma_*^{1/2}\), and invoke Lemma 18 with \(\nu := 3\sqrt{\alpha \mu / 2}\) to arrive at

\[
|\mathcal{R}_3| \leq \frac{3}{2} \sqrt{3\alpha \mu \rho} \left(\|\Delta_L \Sigma_*^{1/2}\|_{F} + \|\Delta_R \Sigma_*^{1/2}\|_{F}\right) \|L_* \Delta_L^T R(R^T R)^{-1} \Sigma_*(R^T R)^{-1} R^T\|_{F}
\]
\[
+ 2\sqrt{\alpha \mu \rho} \|L_* \Sigma_*^{-1/2}\|_{2, \infty} \|R(R^T R)^{-1} \Sigma_*(R^T R)^{-1} R^T \Delta_R \Sigma_*^{1/2}\|_{F} \|LR^T - X_*\|_{F}
\]
\[
\leq \frac{3}{2} \sqrt{3\alpha \mu \rho} \left(\|\Delta_L \Sigma_*^{1/2}\|_{F} + \|\Delta_R \Sigma_*^{1/2}\|_{F}\right) \|\Delta_R \Sigma_*^{1/2}\|_{F} \|R(R^T R)^{-1} \Sigma_*^{1/2}\|_{op} \|LR^T - X_*\|_{F}.
\]

Use the consequences (62) again to obtain

\[
|\mathcal{R}_3| \leq \frac{3}{2} \sqrt{3\alpha \mu \rho} \left(\|\Delta_L \Sigma_*^{1/2}\|_{F} + \|\Delta_R \Sigma_*^{1/2}\|_{F}\right) \|\Delta_R \Sigma_*^{1/2}\|_{F}
\]
\[
+ 2\sqrt{\alpha \mu \rho} \|\Delta_R \Sigma_*^{1/2}\|_{F}(1 + \frac{\epsilon}{2}) \left(\|\Delta_L \Sigma_*^{1/2}\|_{F} + \|\Delta_R \Sigma_*^{1/2}\|_{F}\right)
\]
\[
\leq \sqrt{\alpha \mu \rho} \frac{3\sqrt{3} + \frac{4(2 + \epsilon)}{2(1 - \epsilon)}}{2(1 - \epsilon)^{2}} \left(\|\Delta_L \Sigma_*^{1/2}\|_{F} \|\Delta_R \Sigma_*^{1/2}\|_{F} + \|\Delta_R \Sigma_*^{1/2}\|_{F}^{2}\right)
\]
\[
\leq \sqrt{\alpha \mu \rho} \frac{3\sqrt{3} + \frac{4(2 + \epsilon)}{2(1 - \epsilon)}}{2(1 - \epsilon)^{2}} \left(\frac{1}{2} \|\Delta_L \Sigma_*^{1/2}\|_{F}^{2} + \frac{3}{2} \|\Delta_R \Sigma_*^{1/2}\|_{F}^{2}\right).
\]

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

\[
\sqrt{\mathcal{R}_4} = \text{tr} \left( (S - S_*) R(R^T R)^{-1} \Sigma_*^{1/2} L^{-1} \right)
\]

42
for some $\tilde{L} \in \mathbb{R}^{n_1 \times r}$ obeying $\|\tilde{L}\|_F = 1$. Setting $M := \tilde{L} \Sigma_1^{1/2} (R^T R)^{-1} R^T = L_M R_M^T$ with $L_M := \bar{L} \Sigma_1^{1/2} (R^T R)^{-1} \Sigma_1^{1/2}$ and $R_M := \Sigma_1^{1/2}$, we are ready to apply Lemma 18 again with $\nu := 3\sqrt{\mu r}/2$ to see

$$\sqrt{R_4} \leq \frac{3}{2} \sqrt{3\alpha mr} \left( \|\Delta_L \Sigma_1^{1/2}\|_F + \|\Delta_R \Sigma_1^{1/2}\|_F \right) \left\| \tilde{L} \Sigma_1^{1/2} (R^T R)^{-1} R^T \right\|_F$$

$$+ 2\sqrt{\alpha nr} \left( \|\Delta_L \Sigma_1^{1/2}\|_F + \|\Delta_R \Sigma_1^{1/2}\|_F \right) \|R \Sigma_1^{1/2}\|_{2,\infty} \|L R^T - X_4\|_F$$

$$\leq \frac{3}{2} \sqrt{3\alpha mr} \left( \|\Delta_L \Sigma_1^{1/2}\|_F + \|\Delta_R \Sigma_1^{1/2}\|_F \right) \left\| R (R^T R)^{-1} \Sigma_1^{1/2} \right\|_{op}$$

$$+ 2\sqrt{\alpha nr} \left( \|\Sigma_1^{1/2} (R^T R)^{-1} \Sigma_1^{1/2}\|_{op} \|R \Sigma_1^{1/2}\|_{2,\infty} \|L R^T - X_4\|_F \right).$$

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

$$\sqrt{R_4} \leq \frac{\sqrt{3\alpha m r}}{2(1 - \epsilon)} \left( \|\Delta_L \Sigma_1^{1/2}\|_F + \|\Delta_R \Sigma_1^{1/2}\|_F \right).$$

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

$$R_4 \leq \alpha mr \left( \sqrt{\frac{3\sqrt{3} + 4(2 + \epsilon)}{2(1 - \epsilon)^2}} \left( \|\Delta_L \Sigma_1^{1/2}\|_F^2 + \|\Delta_R \Sigma_1^{1/2}\|_F^2 \right) \right).$$

Taking collectively the bounds for $R_1, R_2, R_3$ and $R_4$ yields the control of $\| (L_{t+1} Q_t - L_\ast) \Sigma_1^{1/2} \|_F^2$ as

$$\left\| (L_{t+1} Q_t - L_\ast) \Sigma_1^{1/2} \right\|_F^2 \leq \left( (1 - \eta)^2 + \frac{2\epsilon}{1 - \epsilon} \right) \frac{2 + \epsilon^2}{1 - \epsilon} \left( \frac{3\sqrt{3} + 4(2 + \epsilon)}{2(1 - \epsilon)^2} \right) \|\Delta_L \Sigma_1^{1/2}\|_F^2 + \|\Delta_R \Sigma_1^{1/2}\|_F^2$$

$$+ \sqrt{\alpha mr} \left( \frac{3\sqrt{3} + 4(2 + \epsilon)}{1 - \epsilon} \right) \|\Delta_L \Sigma_1^{1/2}\|_F^2 + \|\Delta_R \Sigma_1^{1/2}\|_F^2$$

$$+ \sqrt{\alpha mr} \left( \frac{3\sqrt{3} + 4(2 + \epsilon)}{2(1 - \epsilon)^2} \right) \eta^2 \left( \|\Delta_L \Sigma_1^{1/2}\|_F^2 + \|\Delta_R \Sigma_1^{1/2}\|_F^2 \right).$$

Similarly, we can obtain the control of $\| (R_{t+1} Q_t^\top - R_\ast) \Sigma_1^{1/2} \|_F^2$. Combine them together and identify $\text{dist}^2(F_t, F_\ast) = \|\Delta_L \Sigma_1^{1/2}\|_F^2 + \|\Delta_R \Sigma_1^{1/2}\|_F^2$ to reach

$$\left\| (L_{t+1} Q_t - L_\ast) \Sigma_1^{1/2} \right\|_F^2 + \left\| (R_{t+1} Q_t^\top - R_\ast) \Sigma_1^{1/2} \right\|_F^2 \leq \rho^2(\eta; \epsilon, \alpha mr) \text{dist}^2(F_t, F_\ast),$$

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

$$\rho^2(\eta; \epsilon, \alpha mr) := (1 - \eta)^2 + \frac{2\epsilon + \sqrt{\alpha mr} (6\sqrt{3} + 8(2 + \epsilon))}{1 - \epsilon} \eta^2 \left( \frac{3\sqrt{3} + 4(2 + \epsilon)}{1 - \epsilon} \right)^2 \|\Delta_L \Sigma_1^{1/2}\|_F^2 + \|\Delta_R \Sigma_1^{1/2}\|_F^2$$

$$+ 2\epsilon + \epsilon^2 + \sqrt{\alpha mr} (6\sqrt{3} + 4(2 + \epsilon)) + \alpha nm \left( \frac{3\sqrt{3} + 4(2 + \epsilon)}{1 - \epsilon} \right)^2 \eta^2.$$

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

$$\text{dist}(F_{t+1}, F_\ast) \leq \sqrt{\left\| (L_{t+1} Q_t - L_\ast) \Sigma_1^{1/2} \right\|_F^2 + \left\| (R_{t+1} Q_t^\top - R_\ast) \Sigma_1^{1/2} \right\|_F^2}$$

$$\leq (1 - 0.6\eta) \text{dist}(F_t, F_\ast). \quad (63)$$
D.1.2 Incoherence condition

We start by controlling the term \(\|(L_{t+1}Q_t - L_\ast)\Sigma_{\ast}^{1/2}\|_{2,\infty}\). We know from (61) that

\[
(L_{t+1}Q_t - L_\ast)\Sigma_{\ast}^{1/2} = (1 - \eta)\Delta_L\Sigma_{\ast}^{1/2} - \eta L_\ast\Delta_R R(R^T R)^{-1}\Sigma_{\ast}^{1/2} - \eta(S - S_\ast)R(R^T R)^{-1}\Sigma_{\ast}^{1/2}.
\]

Apply the triangle inequality to obtain

\[
\|(L_{t+1}Q_t - L_\ast)\Sigma_{\ast}^{1/2}\|_{2,\infty} \leq (1 - \eta)\|\Delta_L\Sigma_{\ast}^{1/2}\|_{2,\infty} + \eta\|L_\ast\Delta_R R(R^T R)^{-1}\Sigma_{\ast}^{1/2}\|_{2,\infty} + \eta\|(S - S_\ast)R(R^T R)^{-1}\Sigma_{\ast}^{1/2}\|_{2,\infty}.
\]

The first term \(\|\Delta_L\Sigma_{\ast}^{1/2}\|_{2,\infty}\) follows from the incoherence condition (57) as

\[
\|\Delta_L\Sigma_{\ast}^{1/2}\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n_1}}\sigma_r(X_\ast).
\]

In the sequel, we shall bound the terms \(\mathcal{T}_1\) and \(\mathcal{T}_2\).

1. For the term \(\mathcal{T}_1\), use the relation \(|A B|\|_{2,\infty} \leq |A|\|_{2,\infty} |B|_{\text{op}}\), and combine the condition (56) with the consequences (62) to obtain

\[
\mathcal{T}_1 \leq \|L_\ast\Sigma_{\ast}^{-1/2}\|_{2,\infty} \|\Sigma_{\ast}^{1/2}\Delta_R R(R^T R)^{-1}\Sigma_{\ast}^{1/2}\|_{\text{op}} \leq \|L_\ast\Sigma_{\ast}^{-1/2}\|_{2,\infty} \|\Delta_R \Sigma_{\ast}^{1/2}\|_{\text{op}} \|R(R^T R)^{-1}\Sigma_{\ast}^{1/2}\|_{\text{op}} \leq \frac{\epsilon}{1 - \epsilon} \sqrt{\frac{\mu r}{n_1}}\sigma_r(X_\ast),
\]

2. For the term \(\mathcal{T}_2\), use the relation \(|A B|\|_{2,\infty} \leq |A|\|_{2,\infty} |B|_{\text{op}}\) to obtain

\[
\mathcal{T}_2 \leq \|S - S_\ast\|_{2,\infty} \|R(R^T R)^{-1}\Sigma_{\ast}^{1/2}\|_{\text{op}}.
\]

We know from Lemma 18 that \(S - S_\ast\) has at most \(3\alpha n_2\) non-zero entries in each row, and \(|S - S_\ast|\|_{\infty} \leq 2\|LR^T - X_\ast\|_{\infty}\). Upper bound the \(\ell_{2,\infty}\) norm by the \(\ell_{\infty}\) norm as

\[
\|S - S_\ast\|_{2,\infty} \leq \sqrt{3\alpha n_2} \|S - S_\ast\|_{\infty} \leq 2\sqrt{3\alpha n_2}\|LR^T - X_\ast\|_{\infty}.
\]

Split \(LR^T - X_\ast = \Delta_L R^T + L_\ast\Delta_R^T\), and take the conditions (57) and (59) to obtain

\[
\|LR^T - X_\ast\|_{\infty} \leq \|\Delta_L R^T\|_{\infty} + \|L_\ast\Delta_R^T\|_{\infty} \leq \|\Delta_L\Sigma_{\ast}^{1/2}\|_{2,\infty} \|R\Sigma_{\ast}^{-1/2}\|_{2,\infty} + \|L_\ast\Sigma_{\ast}^{-1/2}\|_{2,\infty} \|\Delta_R \Sigma_{\ast}^{1/2}\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n_1}}\sigma_r(X_\ast) + \sqrt{\frac{\mu r}{n_2}}\sigma_r(X_\ast) + \sqrt{\frac{\mu r}{n_1}}\sigma_r(X_\ast) = \frac{3\mu r}{\sqrt{n_1 n_2}}\sigma_r(X_\ast).
\]

This combined with the consequences (62) yields

\[
\mathcal{T}_2 \leq \frac{6\sqrt{3\alpha \mu r}}{1 - \epsilon} \sqrt{\frac{\mu r}{n_1}}\sigma_r(X_\ast).
\]
Taking collectively the bounds for $\Xi_1, \Xi_2$ yields the control

$$\left\| (L_{t+1}Q_t - L_*) \Sigma_*^{1/2} \right\|_{2,\infty} \leq \left( 1 - \eta + \frac{\epsilon + 6\sqrt{3\alpha\mu\nu}}{1 - \epsilon} \eta \right) \sqrt{\frac{\mu^r}{n_1}} \sigma_r(X_*). \quad (64)$$

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

$$\left\| (L_{t+1}Q_{t+1} - L_*) \Sigma_*^{1/2} \right\|_{2,\infty} \leq \left\| (L_{t+1}Q_t - L_*) \Sigma_*^{1/2} \right\|_{2,\infty} + \left\| (L_{t+1}Q_{t+1} - L_{t+1}Q_t) \Sigma_*^{1/2} \right\|_{2,\infty}$$

$$\leq \left\| (L_{t+1}Q_t - L_*) \Sigma_*^{1/2} \right\|_{2,\infty} + \left\| L_{t+1}Q_t \Sigma_*^{-1/2} \right\|_{2,\infty} \left\| \Sigma_*^{1/2}Q_t^{-1}Q_{t+1}\Sigma_*^{1/2} - \Sigma_* \right\|_{op}.$$  

We deduce from (64) that

$$\left\| L_{t+1}Q_t \Sigma_*^{-1/2} \right\|_{2,\infty} \leq \left\| L_t \Sigma_*^{-1/2} \right\|_{2,\infty} + \left\| (L_{t+1}Q_t - L_*) \Sigma_*^{-1/2} \right\|_{2,\infty} \leq \left( 2 - \eta + \frac{\epsilon + 6\sqrt{3\alpha\mu\nu}}{1 - \epsilon} \eta \right) \sqrt{\frac{\mu^r}{n_1}}.$$  

Regarding the alignment matrix term, invoke Lemma 14 to obtain

$$\left\| \Sigma_*^{1/2}Q_t^{-1}Q_{t+1}\Sigma_*^{1/2} - \Sigma_* \right\|_{op} \leq \frac{\left\| (R_{t+1}(Q_t^{-T} - Q_{t+1}^{-T})) \Sigma_*^{1/2} \right\|_{op}}{1 - \left\| (R_{t+1}(Q_t^{-T} - R_*) \Sigma_*^{-1/2} \right\|_{op}}$$

$$\leq \frac{\left\| (R_{t+1}(Q_t^{-T} - R_*) \Sigma_*^{1/2} \right\|_{op} + \left\| (R_{t+1}(Q_{t+1}^{-T} - R_*) \Sigma_*^{1/2} \right\|_{op}}{1 - \left\| (R_{t+1}(Q_t^{-T} - R_*) \Sigma_*^{-1/2} \right\|_{op}}$$

$$\leq \frac{2\epsilon}{1 - \epsilon} \sigma_r(X_*),$$

where we deduce from (63) that the distances using either $Q_t$ or $Q_{t+1}$ are bounded by

$$\left\| (R_{t+1}Q_t^{-T} - R_*) \Sigma_*^{1/2} \right\|_{op} \leq \epsilon \sigma_r(X_*):$$

$$\left\| (R_{t+1}Q_{t+1}^{-T} - R_*) \Sigma_*^{1/2} \right\|_{op} \leq \epsilon \sigma_r(X_*);$$

$$\left\| (R_{t+1}Q_{t+1}^{-T} - R_*) \Sigma_*^{-1/2} \right\|_{op} \leq \epsilon.$$  

Combine all pieces to reach

$$\left\| (L_{t+1}Q_{t+1} - L_*) \Sigma_*^{1/2} \right\|_{2,\infty} \leq \left( 1 + \frac{\epsilon}{1 - \epsilon} \left( 1 - \eta + \frac{\epsilon + 6\sqrt{3\alpha\mu\nu}}{1 - \epsilon} \eta \right) \right) \sqrt{\frac{\mu^r}{n_1}} \sigma_r(X_*).$$

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

$$\left\| (L_{t+1}Q_{t+1} - L_*) \Sigma_*^{1/2} \right\|_{2,\infty} \leq \sqrt{\frac{\mu^r}{n_1}} \sigma_r(X_*).$$

Similarly, we can prove the other part

$$\left\| (R_{t+1}Q_{t+1}^{-T} - R_*) \Sigma_*^{1/2} \right\|_{2,\infty} \leq \sqrt{\frac{\mu^r}{n_2}} \sigma_r(X_*).$$

### 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 $S_* \in S_\alpha$, one has $\|S_* - T_\alpha |X_* + S_*\|_{\infty} \leq 2\|X_*\|_{\infty}$.

**Lemma 20** ([YPCC16, Lemma 1]). For any matrix $M \in S_\alpha$, one has $\|M\|_{op} \leq \alpha \sqrt{\mu_1 \mu_2} \|M\|_{\infty}$.
With these two lemmas in place, we are ready to establish the claimed result. Invoke Lemma 11 to obtain
\[
\text{dist}(F_0, F_\star) \leq \sqrt{2} + 1 \| L_0 R_0^T - X_\star \|_F \leq \sqrt{(\sqrt{2} + 1)2r} \| L_0 R_0^T - X_\star \|_{\text{op}},
\]
where the last relation uses the fact that \( L_0 R_0^T - X_\star \) has rank at most \( 2r \). We can further apply the triangle inequality to see
\[
\| L_0 R_0^T - X_\star \|_{\text{op}} \leq \| Y - T_\alpha [Y] - L_0 R_0^T \|_{\text{op}} + \| Y - T_\alpha [Y] - X_\star \|_{\text{op}} \\
\leq 2 \| Y - T_\alpha [Y] - X_\star \|_{\text{op}} = 2 \| S_\star - T_\alpha [X_\star + S_\star] \|_{\text{op}}.
\]
Here the second inequality hinges on the fact that \( L_0 R_0^T \) is the best rank-\( r \) approximation of \( Y - T_\alpha [Y] \), and the last identity arises from \( Y = X_\star + S_\star \). Follow the same argument as \[YPCC16, Section 6.1\], combining Lemmas 19 and 20 to reach
\[
\| S_\star - T_\alpha [X_\star + S_\star] \|_{\text{op}} \leq 2 \alpha \sqrt{n_1 n_2} \| S_\star - T_\alpha [X_\star + S_\star] \|_\infty \\
\leq 4 \alpha \sqrt{n_1 n_2} \| X_\star \|_\infty \leq 4 \alpha \mu r \kappa \sigma_r (X_\star),
\]
where the last inequality follows from the incoherence assumption
\[
\| X_\star \|_\infty \leq \| U_\star \|_{2,\infty} \| \Sigma_\star \|_{\text{op}} \| V_\star \|_{2,\infty} \leq \frac{\mu r}{\sqrt{n_1 n_2}} \kappa \sigma_r (X_\star).
\]
Take the above inequalities together to arrive at
\[
\text{dist}(F_0, F_\star) \leq 8 \sqrt{2(\sqrt{2} + 1)2r} \alpha \mu r^{3/2} \kappa \sigma_r (X_\star) \leq 20 \alpha \mu r^{3/2} \kappa \sigma_r (X_\star).
\]

**D.3 Proof of Lemma 5**

In view of the condition \( \text{dist}(F_0, F_\star) \leq 0.02 \sigma_r (X_\star) \) and Lemma 9, one knows that \( Q_0 \), the optimal alignment matrix between \( F_0 \) and \( F_\star \), exists. Therefore, for notational convenience, denote \( L := L_0 Q_0, R := R_0 Q_0^\top, \Delta_L := L - L_\star, \Delta_R := R - R_\star \), and \( \epsilon := 0.02 \). Our objective is then translated to demonstrate
\[
\sqrt{n_1} \| \Delta_L \Sigma_\star^{1/2} \|_{2,\infty} \vee \sqrt{n_2} \| \Delta_R \Sigma_\star^{1/2} \|_{2,\infty} \leq \sqrt{\mu r} \sigma_r (X_\star).
\]
From now on, we focus on bounding \( \| \Delta_L \Sigma_\star^{1/2} \|_{2,\infty} \). Since \( U_0 \Sigma_0 V_0^\top \) is the top-\( r \) SVD of \( Y - T_\alpha [Y] \), and recall that \( Y = X_\star + S_\star \), we have the relation
\[
(X_\star + S_\star - T_\alpha [X_\star + S_\star]) V_0 = U_0 \Sigma_0,
\]
which further implies the following decomposition of \( \Delta_L \Sigma_\star^{1/2} \).

**Claim 4.** One has
\[
\Delta_L \Sigma_\star^{1/2} = (S_\star - T_\alpha [X_\star + S_\star]) R(R^\top R)^{-1} \Sigma_\star^{1/2} - L_\star \Delta_R^\top R(R^\top R)^{-1} \Sigma_\star^{1/2}.
\]

Combining Claim 4 with the triangle inequality yields
\[
\| \Delta_L \Sigma_\star^{1/2} \|_{2,\infty} \leq \left\| L_\star \Delta_R^\top R(R^\top R)^{-1} \Sigma_\star^{1/2} \right\|_{2,\infty} + \left\| (S_\star - T_\alpha [X_\star + S_\star]) R(R^\top R)^{-1} \Sigma_\star^{1/2} \right\|_{2,\infty}.
\]
In what follows, we shall control \( J_1 \) and \( J_2 \) in turn.

1. For the term \( J_1 \), use the relation \( \| AB \|_{2,\infty} \leq \| A \|_{2,\infty} \| B \|_{\text{op}} \) to obtain
\[
J_1 \leq \| L_\star \Sigma_\star^{-1/2} \|_{2,\infty} \| \Delta_R \Sigma_\star^{1/2} \|_{\text{op}} \left\| R(R^\top R)^{-1} \Sigma_\star^{1/2} \right\|_{\text{op}}.
\]
The incoherence assumption tells $\|L_\Sigma^{-1/2}\|_{2,\infty} = \|U_\Sigma\|_{2,\infty} \leq \sqrt{\mu r/n_1}$. In addition, the assumption dist($F_0, F_\Sigma$) ≤ $\epsilon r(X_\Sigma)$ entails the bound $\|\Delta R_\Sigma^{-1/2}\|_{op} \leq \epsilon r(X_\Sigma)$. Finally, repeating the argument for obtaining (65) yields $\|\Delta R_\Sigma^{-1/2}\|_{op} \leq \epsilon$, which together with Lemma 12 reveals

$$\|R(R^\top R)^{-1}\Sigma_\Sigma^{-1/2}\|_{op} \leq \frac{1}{1-\epsilon}.$$ 

In all, we arrive at

$$J_1 \leq \frac{\epsilon}{1-\epsilon} \sqrt{\frac{\mu r}{n_1} \sigma_r(X_\Sigma)}.$$ 

2. Proceeding to the term $J_2$, use the relations $\|AB\|_{2,\infty} \leq \|A\|_{1,\infty} \|B\|_{2,\infty}$ and $\|AB\|_{2,\infty} \leq \|A\|_{2,\infty} \|B\|_{op}$ to obtain

$$J_2 \leq \|S_* - T_\alpha[X_* + S_*]\|_{1,\infty} \|R(R^\top R)^{-1}\Sigma_\Sigma^{-1/2}\|_{2,\infty}$$

$$\leq \|S_* - T_\alpha[X_* + S_*]\|_{1,\infty} \|R\Sigma_\Sigma^{-1/2}\|_{2,\infty} \|\Sigma_\Sigma^{-1/2}(R^\top R)^{-1}\Sigma_\Sigma^{-1/2}\|_{op}.$$ 

Regarding $S_* - T_\alpha[X_* + S_*]$, Lemma 19 tells that $S_* - T_\alpha[X_* + S_*]$ has at most $2\alpha n_2$ non-zero entries in each row, and $\|S_* - T_\alpha[X_* + S_*]\|_\infty \leq 2\|X_*\|_\infty$. Consequently, we can upper bound the $\ell_1,\infty$ norm by the $\ell_\infty$ norm as

$$\|S_* - T_\alpha[X_* + S_*]\|_{1,\infty} \leq 2\alpha n_2 \|S_* - T_\alpha[X_* + S_*]\|_\infty$$

$$\leq 4\alpha n_2 \|X_*\|_\infty$$

$$\leq 4\alpha n_2 \sqrt{\frac{\mu r}{\sqrt{n_1 n_2}}} \kappa_r(X_\Sigma).$$ 

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

$$\|R\Sigma_\Sigma^{-1/2}\|_{2,\infty} \leq \|R\Sigma_\Sigma^{-1/2}\|_{2,\infty} + \|\Delta R\Sigma_\Sigma^{-1/2}\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n_2}} + \frac{\|\Delta R\Sigma_\Sigma^{-1/2}\|_{2,\infty}}{\sigma_r(X_\Sigma)}.$$ 

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

$$\|\Sigma_\Sigma^{-1/2}(R^\top R)^{-1}\Sigma_\Sigma^{-1/2}\|_{op} = \|R(R^\top R)^{-1}\Sigma_\Sigma^{-1/2}\|_{op} \leq \frac{1}{1-\epsilon^2}.$$ 

Taking together the above bounds yields

$$J_2 \leq \frac{4\alpha \mu r k}{(1-\epsilon)^2} \sqrt{\frac{\mu r}{n_1}} \sigma_r(X_\Sigma) + \frac{4\alpha \mu r k}{(1-\epsilon)^2} \sqrt{\frac{n_2}{n_1}} \|\Delta R\Sigma_\Sigma^{-1/2}\|_{2,\infty}.$$

Combine the bounds on $J_1$ and $J_2$ to reach

$$\sqrt{n_1\|\Delta L_\Sigma^{-1/2}\|_{2,\infty}} \leq \left(\frac{\epsilon}{1-\epsilon} + \frac{4\alpha \mu r k}{(1-\epsilon)^2}\right) \sqrt{\mu r} \sigma_r(X_\Sigma) + \frac{4\alpha \mu r k}{(1-\epsilon)^2} \sqrt{n_1\|\Delta R\Sigma_\Sigma^{-1/2}\|_{2,\infty}}.$$

Similarly, we have

$$\sqrt{n_2\|\Delta R\Sigma_\Sigma^{-1/2}\|_{2,\infty}} \leq \left(\frac{\epsilon}{1-\epsilon} + \frac{4\alpha \mu r k}{(1-\epsilon)^2}\right) \sqrt{\mu r} \sigma_r(X_\Sigma) + \frac{4\alpha \mu r k}{(1-\epsilon)^2} \sqrt{n_2\|\Delta L_\Sigma^{-1/2}\|_{2,\infty}}.$$

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

$$\sqrt{n_1\|\Delta L_\Sigma^{-1/2}\|_{2,\infty}} \vee \sqrt{n_2\|\Delta L_\Sigma^{-1/2}\|_{2,\infty}} \leq \frac{\epsilon(1-\epsilon) + 4\alpha \mu r k}{(1-\epsilon)^2} \sqrt{\mu r} \sigma_r(X_\Sigma).$$

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

$$\sqrt{n_1\|\Delta L_\Sigma^{-1/2}\|_{2,\infty}} \vee \sqrt{n_2\|\Delta L_\Sigma^{-1/2}\|_{2,\infty}} \leq \sqrt{\mu r} \sigma_r(X_\Sigma).$$
Proof of Claim 4. Identify \( U_0 \) (resp. \( V_0 \)) with \( L_0 \Sigma_0^{-1/2} \) (resp. \( R_0 \Sigma_0^{-1/2} \)) to yield
\[
(X_\star + S_\star - \tau_0 [X_\star + S_\star]) R_0 \Sigma_0^{-1} = L_0,
\]
which is equivalent to \( (X_\star + S_\star - \tau_0 [X_\star + S_\star]) R_0 (R_0^\top R_0)^{-1} = L_0 \) since \( \Sigma_0 = R_0^\top R_0 \). Multiply both sides by \( Q_0 \Sigma_0^{-1/2} \) to obtain
\[
(X_\star + S_\star - \tau_0 [X_\star + S_\star]) R (R^\top R)^{-1} \Sigma_\star^{1/2} = L \Sigma_\star^{1/2},
\]
where we recall that \( L = L_0 Q_0 \) and \( R = R_0 Q_0^{-\top} \). In the end, subtract \( X \star R (R^\top R)^{-1} \Sigma_\star^{1/2} \) from both sides to reach
\[
(S_\star - \tau_0 [X_\star + S_\star]) R (R^\top R)^{-1} \Sigma_\star^{1/2} = L \Sigma_\star^{1/2} - L_\star R_\star^\top R (R^\top R)^{-1} \Sigma_\star^{1/2}
= (L - L_\star) \Sigma_\star^{1/2} + L_\star (R - R_\star)^\top R (R^\top R)^{-1} \Sigma_\star^{1/2}
= \Delta_L \Sigma_\star^{1/2} + L_\star \Delta_R R (R^\top R)^{-1} \Sigma_\star^{1/2}.
\]
This finishes the proof. \( \square \)

E Proof for Matrix Completion

E.1 New projection operator

E.1.1 Proof of Proposition 1

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

\[
L = \arg \min_{L \in \mathbb{R}^{n_1 \times r}} \left\| (L - \bar{L})(\bar{R}^\top \bar{R})^{1/2} \right\|_F^2 \quad \text{s.t.} \quad \sqrt{n_1} \left\| L (\bar{R}^\top \bar{R})^{1/2} \right\|_2,\infty \leq B.
\]

By a change of variables as \( G := L(\bar{R}^\top \bar{R})^{1/2} \) and \( \bar{G} := \bar{L}(\bar{R}^\top \bar{R})^{1/2} \), we rewrite the above problem equivalently as

\[
G = \arg \min_{G \in \mathbb{R}^{n_1 \times r}} \| G - \bar{G} \|_F^2 \quad \text{s.t.} \quad \sqrt{n_1} \| G \|_2,\infty \leq B,
\]

whose solution is given as [CW15]

\[
G_{i,} = \left( 1 \wedge \frac{B}{\sqrt{n_1} \| G_{i,} \|_2} \right) \bar{G}_{i,}, \quad 1 \leq i \leq n_1.
\]

By applying again the change of variable \( L = G(\bar{R}^\top \bar{R})^{-1/2} \) and \( \bar{L} = \bar{G}(\bar{R}^\top \bar{R})^{-1/2} \), 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 \( \bar{F} \) and \( F_\star \) as \( \bar{Q} \), whose existence is guaranteed by Lemma 9. Denoting \( \mathcal{P}_B(F) = [L^\top, R^\top]^\top \), by the definition of dist(\( \mathcal{P}_B(F), F_\star \)), we know that

\[
\begin{align*}
\text{dist}^2(\mathcal{P}_B(F), F_\star) &\leq \sum_{i=1}^{n_1} \left\| L_{i,} \bar{Q} \Sigma_\star^{1/2} - (L_{i,} \Sigma_\star^{1/2})_{i,} \right\|_2^2 + \sum_{j=1}^{n_2} \left\| R_{j,} \bar{Q}^\top \Sigma_\star^{1/2} - (R_{j,} \Sigma_\star^{1/2})_{j,} \right\|_2^2. \quad (66)
\end{align*}
\]

Recall that the condition \( \text{dist}(\bar{F}, F_\star) \leq \epsilon \sigma_r(X_\star) \) implies

\[
\left\| (L \bar{Q} - L_{\star}) \Sigma_\star^{-1/2} \right\|_\text{op} \vee \left\| (R \bar{Q}^\top - R_{\star}) \Sigma_\star^{-1/2} \right\|_\text{op} \leq \epsilon,
\]

48
which, together with $R_i \Sigma_i^{-1/2} = V_i$, further implies that
\[
\|L_i, R_i\|^2 \leq \|\tilde{L}_i, \tilde{Q}_i \Sigma_i^{1/2}\|_2 \|\tilde{R} Q^{-1} \Sigma_i^{-1/2}\|_{op} \\
\leq \|\tilde{L}_i, \tilde{Q}_i \Sigma_i^{1/2}\|_2 (\|V_i\|_{op} + \|\tilde{R} Q^{-1} - R_i \Sigma_i^{-1/2}\|_{op}) \leq (1 + \epsilon) \|\tilde{L}_i, \tilde{Q}_i \Sigma_i^{1/2}\|_2.
\]
In addition, the $\mu$-incoherence of $X$, yields
\[
\sqrt{n_1} \|\langle L_i, \Sigma_i^{1/2}\rangle_{2},\|_2 \leq \sqrt{n_1} \|U_\star\|_{2,\infty} \|\Sigma_\star\|_{op} \leq \sqrt{\mu \sigma_1(X)} \leq \frac{B}{1 + \epsilon},
\]
where the last inequality follows from the choice of $B$. Take the above two relations collectively to reach
\[
\frac{B}{\sqrt{n_1} \|\tilde{L}_i, \tilde{R}_i\|^2} \geq \frac{\|\langle L_i, \Sigma_i^{1/2}\rangle_{2},\|_2}{\|\tilde{L}_i, \tilde{Q}_i \Sigma_i^{1/2}\|_2}.
\]
We claim that performing the following projection yields a contraction on each row; see also [ZL16, Lemma 11].

**Claim 5.** For vectors $u, u_\star \in \mathbb{R}^n$ and $\lambda \geq \|u_\star\|_2/\|u\|_2$, it holds that
\[
\|(1 \land \lambda) u - u_\star\|_2 \leq \|u - u_\star\|_2.
\]

Apply Claim 5 with $u := \tilde{L}_i, \tilde{Q}_i \Sigma_i^{1/2}$, $u_\star := (\langle L_i, \Sigma_i^{1/2}\rangle_{2},\cdot)_i$, and $\lambda := B/(\sqrt{n_1} \|\tilde{L}_i, \tilde{R}_i\|^2)$ to obtain
\[
\|\tilde{L}_i, \tilde{Q}_i \Sigma_i^{1/2} - (\langle L_i, \Sigma_i^{1/2}\rangle_{2},\cdot)_i\|_2^2 = \left(1 \land \frac{B}{\sqrt{n_1} \|\tilde{L}_i, \tilde{R}_i\|^2}\right) \|\tilde{L}_i, \tilde{Q}_i \Sigma_i^{1/2} - (\langle L_i, \Sigma_i^{1/2}\rangle_{2},\cdot)_i\|_2^2 \\
\leq \|\tilde{L}_i, \tilde{Q}_i \Sigma_i^{1/2} - (\langle L_i, \Sigma_i^{1/2}\rangle_{2},\cdot)_i\|_2^2.
\]
Following a similar argument for $R_i$, and plugging them back to (66), we conclude that
\[
\text{dist}^2(P_B(\bar{F}, F_\star)) \leq \sum_{i=1}^{n_1} \|\tilde{L}_i, \tilde{Q}_i \Sigma_i^{1/2} - (\langle L_i, \Sigma_i^{1/2}\rangle_{2},\cdot)_i\|_2^2 + \sum_{j=1}^{n_2} \|\tilde{R}_j, \tilde{Q}_i^{-1} \Sigma_i^{-1/2} - (\langle R_i, \Sigma_i^{-1/2}\rangle_{2},\cdot)_j\|_2^2 = \text{dist}^2(\bar{F}, F_\star).
\]
We move on to the incoherence condition. For any $1 \leq i \leq n_1$, one has
\[
\|L_i, R_i\|^2 = \sum_{j=1}^{n_2} \langle L_i, \cdot, R_j, \cdot\rangle^2 = \sum_{j=1}^{n_2} \left(1 \land \frac{B}{\sqrt{n_1} \|\tilde{L}_i, \tilde{R}_i\|^2}\right)^2 \langle \tilde{L}_i, \tilde{R}_j, \cdot\rangle^2 = \left(1 \land \frac{B}{\sqrt{n_1} \|\tilde{L}_i, \tilde{R}_i\|^2}\right)^2 \|\tilde{L}_i, \tilde{R}_i\|^2 \\
\leq \left(1 \land \frac{B}{\sqrt{n_1} \|\tilde{L}_i, \tilde{R}_i\|^2}\right)^2 \|\tilde{L}_i, \tilde{R}_i\|^2 \\
\leq \frac{B^2}{n_1}.
\]
where (i) follows from $1 \land \frac{B}{\sqrt{n_2} \|\tilde{R}_j, \tilde{L}_i\|^2} \leq 1$, and (ii) follows from $1 \land \frac{B}{\sqrt{n_1} \|\tilde{L}_i, \tilde{R}_i\|^2} \leq \frac{B}{\sqrt{n_2} \|\tilde{R}_j, \tilde{L}_i\|^2}$. Similarly, one has $\|R_j, L_i\|^2 \leq B^2/n_2$. 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 $h(\lambda) := \|\lambda u - u_\star\|^2$. Calculate its derivative to conclude that $h(\lambda)$ is monotonically increasing when $\lambda \geq \lambda_\star := \langle u, u_\star\rangle/\|u\|^2$. Note that $\lambda \geq \|u_\star\|_2/\|u\|_2 \geq \lambda_\star$, thus $h(\lambda) \leq h(1)$, i.e. the claim holds. \qed
E.2 Proof of Lemma 7

We first record two useful lemmas regarding the projector \(P_\Omega(\cdot)\).

**Lemma 21** ([ZL16, Lemma 10]). Suppose that \(X_\ast\) is \(\mu\)-incoherent, and \(p \gtrsim \mu r \log(n_1 \lor n_2) / (n_1 \land n_2)\). With overwhelming probability, one has

\[
\left| \langle (p^{-1} P_\Omega - I)(L_A R_A^\top + L_A R_A^\top), L_A R_B^\top + L_B R_B^\top \rangle \right|
\leq C_1 \sqrt{\frac{\mu r \log(n_1 \lor n_2)}{p(n_1 \land n_2)}} \| L_A R_A^\top + L_A R_A^\top \|_F \| L_A R_B^\top + L_B R_B^\top \|_F,
\]

simultaneously for all \(L_A, L_B \in \mathbb{R}^{n_1 \times r}\) and \(R_A, R_B \in \mathbb{R}^{n_2 \times r}\), where \(C_1 > 0\) is some universal constant.

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

\[
\left| \langle (p^{-1} P_\Omega - I)(L_A R_A^\top), L_B R_B^\top \rangle \right|
\leq C_2 \sqrt{\frac{n_1 \land n_2}{p}} (\| L_A \|_F \| L_B \|_{2,\infty} \lor \| L_A \|_{2,\infty} \| L_B \|_F) (\| R_A \|_F \| R_B \|_{2,\infty} \lor \| R_A \|_{2,\infty} \| R_B \|_F),
\]

simultaneously for all \(L_A, L_B \in \mathbb{R}^{n_1 \times r}\) and \(R_A, R_B \in \mathbb{R}^{n_2 \times r}\), 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 \(\text{dist}(F_t, F_\ast) \leq 0.02 \sigma_r(X_\ast)\) and Lemma 9, one knows that \(Q_t\), the optimal alignment matrix between \(F_t\) and \(F_\ast\), exists. Therefore, for notational convenience, we denote \(L := L_t Q_t, R := R_t Q_t^\top\), \(\Delta_L := L - L_\ast, \Delta_R := R - R_\ast\), and \(\epsilon := 0.02\). In addition, denote \(\hat{F}_{t+1}\) as the update before projection as

\[
\hat{F}_{t+1} := \begin{bmatrix} \hat{L}_{t+1} \\ \hat{R}_{t+1} \end{bmatrix} = \begin{bmatrix} L_t - \eta p^{-1} P_\Omega(L_t R_t^\top - X_\ast)R_t (R_t^\top R_t)^{-1} \\ R_t - \eta p^{-1} P_\Omega(L_t R_t^\top - X_\ast)^\top L_t (L_t^\top L_t)^{-1} \end{bmatrix},
\]

and therefore \(F_{t+1} = P_B(\hat{F}_{t+1})\).

Note that in view of Lemma 6, it suffices to prove the following relation

\[
\text{dist}(\hat{F}_{t+1}, F_\ast) \leq (1 - 0.6 \eta) \text{dist}(F_t, F_\ast).
\]

The conclusion \(\| L_t R_t^\top - X_\ast \|_F \leq 1.5 \text{dist}(F_t, F_\ast)\) 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 \(\text{dist}(F_t, F_\ast) \leq \epsilon \sigma_r(X_\ast)\) and \(\sqrt{n_1} \| L R^\top \|_{2,\infty} \lor \sqrt{n_2} \| R L^\top \|_{2,\infty} \leq C_B \sqrt{\mu r} \sigma_1(X_\ast)\), one has

\[
\| \Delta_L \Sigma_\ast^{-1/2} \|_{2,\infty} \lor \| \Delta_R \Sigma_\ast^{-1/2} \|_{2,\infty} \leq \epsilon;
\]

\[
\| R (R^\top R)^{-1} \Sigma_\ast^{-1/2} \|_{2,\infty} \leq \frac{1}{1 - \epsilon};
\]

\[
\| \Sigma_\ast^{-1/2} (R^\top R)^{-1} \Sigma_\ast^{-1/2} \|_{2,\infty} \leq \frac{1}{(1 - \epsilon)^2};
\]

\[
\sqrt{n_1} \| L \Sigma_\ast^{-1/2} \|_{2,\infty} \lor \sqrt{n_2} \| R \Sigma_\ast^{-1/2} \|_{2,\infty} \leq \frac{C_B}{1 - \epsilon} \sqrt{\mu r} \sigma_1(X_\ast); \tag{68d}
\]

\[
\sqrt{n_1} \| L \Sigma_\ast^{-1/2} \|_{2,\infty} \lor \sqrt{n_2} \| R \Sigma_\ast^{-1/2} \|_{2,\infty} \leq \frac{C_B k}{1 - \epsilon} \sqrt{\mu r}; \tag{68e}
\]

\[
\sqrt{n_1} \| \Delta_L \Sigma_\ast^{-1/2} \|_{2,\infty} \lor \sqrt{n_2} \| \Delta_R \Sigma_\ast^{-1/2} \|_{2,\infty} \leq \left( 1 + \frac{C_B}{1 - \epsilon} \right) \sqrt{\mu r} \sigma_1(X_\ast); \tag{68f}
\]

\[
\sqrt{n_1} \| \Delta_L \Sigma_\ast^{-1/2} \|_{2,\infty} \lor \sqrt{n_2} \| \Delta_R \Sigma_\ast^{-1/2} \|_{2,\infty} \leq \left( 1 + \frac{C_B k}{1 - \epsilon} \right) \sqrt{\mu r}. \tag{68g}
\]
Now we are ready to embark on the proof of (67). By the definition of \( \text{dist}(F_{t+1}, F_*) \), one has
\[
\text{dist}^2(F_{t+1}, F_*) \leq \left\| (L_{t+1}Q_t - L_*) \Sigma_1 \right\|_F^2 + \left\| (R_{t+1}Q_t - R_*) \Sigma_1 \right\|_F^2, 
\]
where we recall that \( Q_t \) is the optimal alignment matrix between \( F_t \) and \( F_* \). Plug in the update rule (26) and the decomposition \( LR^T - X_* = \Delta_L R^T + L \Delta_R^T \) to obtain
\[
(\tilde{L}_{t+1}Q_t - L_*) \Sigma_1^2 = (L - \eta^{-1} \Omega(LR^T - X_*) R(R^T R)^{-1} - L_*) \Sigma_1^2 \\
= \Delta_L \Sigma_1^2 - \eta(LR^T - X_*) R(R^T R)^{-1} \Sigma_1^2 - \eta(p^{-1} \Omega - I)(LR^T - X_*) R(R^T R)^{-1} \Sigma_1^2 \\
= (1 - \eta) \Delta_L \Sigma_1^2 - \eta \Delta_R R(R^T R)^{-1} \Sigma_1^2 - \eta(p^{-1} \Omega - I)(LR^T - X_*) R(R^T R)^{-1} \Sigma_1^2.
\]
This allows us to expand the first square in (69) as
\[
\left\| (\tilde{L}_{t+1}Q_t - L_*) \Sigma_1 \right\|_F^2 = \underbrace{\left\| (1 - \eta) \Delta_L \Sigma_1^2 - \eta \Delta_R R(R^T R)^{-1} \Sigma_1^2 \right\|_F^2}_{\mathcal{P}_1} \\
- 2\eta(1 - \eta) \text{tr} \left( (p^{-1} \Omega - I)(LR^T - X_*) R(R^T R)^{-1} \Sigma_2 \right) \\
+ 2\eta^2 \text{tr} \left( (p^{-1} \Omega - I)(LR^T - X_*) R(R^T R)^{-1} \Sigma_2 \right) \text{tr} \left( R(R^T R)^{-1} \Sigma_2 \right) \\
+ \eta^2 \left\| (p^{-1} \Omega - I)(LR^T - X_*) R(R^T R)^{-1} \Sigma_1^2 \right\|_F^2.
\]
In the sequel, we shall control the four terms separately, of which \( \mathcal{P}_1 \) is the main term, and \( \mathcal{P}_2, \mathcal{P}_3 \) and \( \mathcal{P}_4 \) are perturbation terms.

1. Notice that the main term \( \mathcal{P}_1 \) has already been controlled in (46) under the condition (68a). It obeys
\[
\mathcal{P}_1 \leq \left( (1 - \eta)^2 + \frac{2\epsilon}{1 - \epsilon} \eta(1 - \eta) \right) \left\| \Delta_L \Sigma_1^2 \right\|_F^2 + \frac{2\epsilon + \epsilon^2}{(1 - \epsilon)^2} \eta^2 \left\| \Delta_R \Sigma_1^2 \right\|_F^2.
\]

2. For the second term \( \mathcal{P}_2 \), decompose \( LR^T - X_* = \Delta_L R^T + L \Delta_R^T + \Delta_L \Delta_R^T \) and apply the triangle inequality to obtain
\[
\mathcal{P}_2 = \left| \text{tr} \left( (p^{-1} \Omega - I)(\Delta_L R^T + L \Delta_R^T + \Delta_L \Delta_R^T) R(R^T R)^{-1} \Sigma, \Delta_L^T \right) \right| \\
\leq \left| \text{tr} \left( (p^{-1} \Omega - I)(\Delta_L R^T + L \Delta_R^T) R(R^T R)^{-1} \Sigma, \Delta_L^T \right) \right| \\
+ \left| \text{tr} \left( (p^{-1} \Omega - I)(\Delta_L \Delta_R^T) R(R^T R)^{-1} \Sigma, \Delta_L^T \right) \right| \\
+ \left| \text{tr} \left( (p^{-1} \Omega - I)(\Delta_L \Delta_R^T) R(R^T R)^{-1} \Sigma, \Delta_L^T \right) \right|.
\]

For the first term \( \mathcal{P}_{2,1} \), under the event \( \mathcal{E} \), we can invoke Lemma 21 to obtain
\[
\mathcal{P}_{2,1} \leq C_1 \sqrt{\frac{\pi \log(n_1 \vee n_2)}{p(n_1 \wedge n_2)}} \left\| \Delta_L R^T + L \Delta_R^T \right\|_F \left\| \Delta_L \Sigma, (R^T R)^{-1} R^T \right\|_F.
\]
where the second line follows from the relation \( \| AB \|_F \leq \| A \|_{op} \| B \|_F \). Use the condition (68c) to obtain

\[
\mathcal{P}_{2,1} \leq \frac{C_1}{(1 - \epsilon)^2} \sqrt{\frac{\mu r \log(n_1 \lor n_2)}{p(n_1 \land n_2)}} \left( \| \Delta_L \Sigma_{\ast}^{1/2} \|_F + \| \Delta_R \Sigma_{\ast}^{1/2} \|_F \right) \| \Delta_L \Sigma_{\ast}^{1/2} \|_F \| \Sigma_{\ast}^{1/2} \|_F \| \Sigma_{\ast}^{1/2} \|_F.
\]

Regarding the remaining three terms \( \mathcal{P}_{2,2}, \mathcal{P}_{2,3}, \) and \( \mathcal{P}_{2,4} \), our main hammer is Lemma 22. For instance, invoking Lemma 22 under the event \( \mathcal{E} \) with \( L_A := \Delta_L \Sigma_{\ast}^{1/2}, R_A := \Delta_R \Sigma_{\ast}^{1/2}, L_B := \Delta_L \Sigma_{\ast}^{1/2}, \) and \( R_B := \Delta_R (R^T R)^{-1} \Sigma_{\ast}^{1/2} \), we arrive at

\[
\mathcal{P}_{2,2} \leq C_2 \sqrt{\frac{n_1 \land n_2}{p}} \| \Delta_L \Sigma_{\ast}^{1/2} \|_{2, \infty} \| \Delta_L \Sigma_{\ast}^{1/2} \|_F \| \Delta_R \Sigma_{\ast}^{1/2} \|_{2, \infty} \| \Delta_R (R^T R)^{-1} \Sigma_{\ast}^{1/2} \|_F.
\]

Similarly, with the help of Lemma 22, one has

\[
\mathcal{P}_{2,3} \leq C_2 \sqrt{\frac{n_1 \land n_2}{p}} \| \Delta_L \Sigma_{\ast}^{1/2} \|_{2, \infty} \| \Delta_R \Sigma_{\ast}^{1/2} \|_F \| \Delta_R \Sigma_{\ast}^{1/2} \|_{2, \infty} \| \Sigma_{\ast}^{1/2} \|_F \| \Sigma_{\ast}^{1/2} \|_F \| \Sigma_{\ast}^{1/2} \|_F.
\]

Utilizing the consequences in Claim 6, we arrive at

\[
\mathcal{P}_{2,2} \leq \frac{C_2 \kappa}{(1 - \epsilon)^2} \left( 1 + \frac{C_B \kappa}{1 - \epsilon} \right) \sqrt{\frac{n_1 \land n_2}{p}} \| \Delta_L \Sigma_{\ast}^{1/2} \|_F \| \Delta_R \Sigma_{\ast}^{1/2} \|_F;
\]

\[
\mathcal{P}_{2,3} \leq \frac{C_2}{(1 - \epsilon)^2} \left( 1 + \frac{C_B \kappa}{1 - \epsilon} \right) \sqrt{\frac{n_1 \land n_2}{p}} \| \Delta_L \Sigma_{\ast}^{1/2} \|_F \| \Delta_R \Sigma_{\ast}^{1/2} \|_F;
\]

\[
\mathcal{P}_{2,4} \leq \frac{C_2 C_B \kappa}{(1 - \epsilon)^2} \left( 1 + \frac{C_B \kappa}{1 - \epsilon} \right) \sqrt{\frac{n_1 \land n_2}{p}} \| \Delta_L \Sigma_{\ast}^{1/2} \|_F \| \Delta_R \Sigma_{\ast}^{1/2} \|_F.
\]

We then combine the bounds for \( \mathcal{P}_{2,1}, \mathcal{P}_{2,2}, \mathcal{P}_{2,3}, \) and \( \mathcal{P}_{2,4} \) to see

\[
\mathcal{P}_2 \leq \frac{C_1}{(1 - \epsilon)^2} \sqrt{\frac{\mu r \log(n_1 \lor n_2)}{p(n_1 \land n_2)}} \left( \| \Delta_L \Sigma_{\ast}^{1/2} \|_F + \| \Delta_R \Sigma_{\ast}^{1/2} \|_F \right) \| \Delta_L \Sigma_{\ast}^{1/2} \|_F \| \Delta_R \Sigma_{\ast}^{1/2} \|_F.
\]

where we denote

\[
\delta_1 := \frac{C_1}{(1 - \epsilon)^2} \sqrt{\frac{\mu r \log(n_1 \lor n_2)}{p(n_1 \land n_2)}} , \quad \text{and} \quad \delta_2 := \frac{C_2}{(1 - \epsilon)^2} \left( \kappa + 1 + \frac{C_B \kappa}{1 - \epsilon} \right) \sqrt{\frac{n_1 \land n_2}{p}} \| \Delta_L \Sigma_{\ast}^{1/2} \|_F \| \Delta_R \Sigma_{\ast}^{1/2} \|_F.
\]

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

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

\[
\mathcal{P}_3 \leq \frac{\delta_1 + \delta_2}{2} \| \Delta_L \Sigma_{\ast}^{1/2} \|_F + \frac{3\delta_1 + \delta_2}{2} \| \Delta_R \Sigma_{\ast}^{1/2} \|_F.
\]

\[
\mathcal{P}_4 \leq \delta_1 (2\delta_1 + \delta_2) \| \Delta_L \Sigma_{\ast}^{1/2} \|_F + (\delta_1 + \delta_2)(2\delta_1 + \delta_2) \| \Delta_R \Sigma_{\ast}^{1/2} \|_F.
\]
Proof of Claim 7. This finishes the proof.

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

\[
\left\| (\tilde{L}_{t+1}Q_t - L_*) \Sigma_*^{1/2} \right\| \leq \rho^2(\eta; \epsilon, \delta_1, \delta_2) \text{dist}(F_t, F_*),
\]

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

\[
\rho^2(\eta; \epsilon, \delta_1, \delta_2) := (1 - \eta)^2 + \left( \frac{2\epsilon}{1 - \epsilon} + 4\delta_1 + 2\delta_2 \right) \eta (1 - \eta) + \left( \frac{2\epsilon + \epsilon^2}{(1 - \epsilon)^2} + 4\delta_1 + 2\delta_2 + (2\delta_1 + \delta_2)^2 \right) \eta^2.
\]

As long as \( p \geq C(\mu r c^4 \log(n_1 \vee n_2)) \mu r / (n_1 \wedge n_2) \) for some sufficiently large constant \( C \), one has \( \delta_1 \leq 0.05 \) and \( \delta_2 \leq 0.05 \) under the setting \( \epsilon = 0.02 \). When \( 0 < \eta \leq 2/3 \), one further has \( \rho(\eta; \epsilon, \delta_1, \delta_2) \leq 1 - 0.6 \eta \). Thus we conclude that

\[
\text{dist}(F_{t+1}, F_*) \leq \sqrt{\left\| (\tilde{L}_{t+1}Q_t - L_*) \Sigma_*^{1/2} \right\|^2 + \left\| (\tilde{R}_{t+1}Q_t^{\top} - R_*) \Sigma_*^{1/2} \right\|^2} \leq (1 - 0.6\eta) \text{dist}(F_t, F_*),
\]

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 \( \sqrt{mr} \| LR^\top \|_{2, \infty} \sqrt{mr} \| RL^\top \|_{2, \infty} \leq C_B \sqrt{mr} \sigma_1(X_*) \) together with the relations

\[
\| LR^\top \|_{2, \infty} \geq \sigma_r (RS_*)^{-1/2} \| LS_*^{1/2} \|_{2, \infty},
\]

\[
\| RL^\top \|_{2, \infty} \geq \sigma_r (LS_*)^{-1/2} \| RS_*^{1/2} \|_{2, \infty},
\]

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

Proof of Claim 7. We start with the term \( \Psi_3 \), for which we have

\[
\Psi_3 \leq \left\{ \text{tr} \left( (p^{-1} P - T) (\Delta_L \Delta_R^* + L_\delta \Delta_R^*) R (R^\top R)^{-1} \Sigma_* (R^\top R)^{-1} R^\top \Delta_R L_\delta^* \right) \right\}_{\Psi_{3,1}}
\]

\[
+ \left\{ \text{tr} \left( (p^{-1} P - T) (\Delta_L \Delta_R^*) R (R^\top R)^{-1} \Sigma_* (R^\top R)^{-1} R^\top \Delta_R L_\delta^* \right) \right\}_{\Psi_{3,2}}.
\]
Invoke Lemma 21 to bound $\Psi_{3,1}$ as

$$
\Psi_{3,1} \leq C_1 \sqrt[\mu r \log(n_1 \vee n_2)]{p(n_1 \wedge n_2)} \| \Delta_L R^*_r + L_\Sigma^1 R \Sigma^1 R^T \|_F \| \Delta_R \Sigma^1 R^T R^{-1} \Sigma^1 R^{-1} R^T \|_F
$$

$$
\leq C_1 \sqrt[\mu r \log(n_1 \vee n_2)]{p(n_1 \wedge n_2)} \left( \| \Delta_L \Sigma^1 \|_F + \| \Delta_R \Sigma^1 \|_F \right) \| \Delta_R \Sigma^1 \|_F \| R \Sigma^1 \|_F^{\frac{1}{2}} \| \Sigma^1 \|_F^{\frac{1}{2}} \|_{op}^2 .
$$

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

$$
\Psi_{3,1} \leq \frac{C_1}{(1-\epsilon)^2} \sqrt[\mu r \log(n_1 \vee n_2)]{p(n_1 \wedge n_2)} \left( \| \Delta_L \Sigma^1 \|_F + \| \Delta_R \Sigma^1 \|_F \right) \| \Delta_R \Sigma^1 \|_F .
$$

In regard to $\Psi_{3,2}$, we apply Lemma 22 with $L_A := \Delta_L \Sigma^1$, $R_A := \Delta_R \Sigma^{-1/2}$, $L_B := \Sigma^{-1/2}$, and $R_B := R \Sigma^{-1/2}$ to see

$$
\Psi_{3,2} \leq C_2 \sqrt[\mu r \log(n_1 \vee n_2)]{p(n_1 \wedge n_2)} \left( \| \Delta_L \Sigma^1 \|_F + \| \Delta_R \Sigma^{-1/2} \|_F \right) \| \Delta_R \Sigma^1 \|_F \| R \Sigma^1 \|_F^{\frac{1}{2}} \| \Sigma^1 \|_F^{\frac{1}{2}} \|_{op}^2 .
$$

Again, use the consequences in Claim 6 to reach

$$
\Psi_{3,2} \leq C_2 \left( \frac{3n_1 \wedge n_2}{p} \| \Delta_L \Sigma^1 \|_F \right) \sqrt[\mu r \log(n_1 \vee n_2)]{p(n_1 \wedge n_2)} \left( \frac{1+C_B}{1-\epsilon} \right) \sqrt[\mu r \log(n_1 \vee n_2)]{p(n_1 \wedge n_2)} \| \Delta_R \Sigma^1 \|_F .
$$

Combine the bounds of $\Psi_{3,1}$ and $\Psi_{3,2}$ to reach

$$
\Psi_3 \leq \frac{C_1}{(1-\epsilon)^2} \sqrt[\mu r \log(n_1 \vee n_2)]{p(n_1 \wedge n_2)} \left( \| \Delta_L \Sigma^1 \|_F + \| \Delta_R \Sigma^{-1/2} \|_F \right) \| \Delta_R \Sigma^1 \|_F
$$

$$
+ \frac{C_2}{(1-\epsilon)^2} \left( \frac{1+C_B}{1-\epsilon} \right) \sqrt[\mu r \log(n_1 \vee n_2)]{p(n_1 \wedge n_2)} \| \Delta_L \Sigma^1 \|_F \| \Delta_R \Sigma^1 \|_F
$$

$$
\leq \left( \delta_1 + \delta_2 \right) \| \Delta_L \Sigma^1 \|_F + \delta_1 \| \Delta_R \Sigma^1 \|_F
$$

$$
\leq \frac{\delta_1 + \delta_2}{2} \| \Delta_L \Sigma^1 \|_F + \frac{3\delta_1 + \delta_2}{2} \| \Delta_R \Sigma^1 \|_F .
$$

Moving on to the term $\Psi_4$, we have

$$
\sqrt[\mu r \log(n_1 \vee n_2)]{p(n_1 \wedge n_2)} \left( \| \Delta^2 \Sigma^1 \|_F \right)
$$

$$
\leq \left( \frac{(n_1 \wedge n_2)^{\gamma}}{p} \right) \left( \| \Delta_L \Sigma^1 \|_F + \| \Delta_R \Sigma^{-1/2} \|_F \right) \| \Delta_R \Sigma^1 \|_F
$$

$$
+ \left( \frac{(n_1 \wedge n_2)^{\gamma}}{p} \right) \left( \| \Delta_L \Sigma^1 \|_F + \| \Delta_R \Sigma^{-1/2} \|_F \right) \| \Delta_R \Sigma^1 \|_F
$$

$$
+ \left( \frac{(n_1 \wedge n_2)^{\gamma}}{p} \right) \left( \| \Delta_L \Sigma^1 \|_F + \| \Delta_R \Sigma^{-1/2} \|_F \right) \| \Delta_R \Sigma^1 \|_F
$$

$$
+ \left( \frac{(n_1 \wedge n_2)^{\gamma}}{p} \right) \left( \| \Delta_L \Sigma^1 \|_F + \| \Delta_R \Sigma^{-1/2} \|_F \right) \| \Delta_R \Sigma^1 \|_F
$$

$$
+ \left( \frac{(n_1 \wedge n_2)^{\gamma}}{p} \right) \left( \| \Delta_L \Sigma^1 \|_F + \| \Delta_R \Sigma^{-1/2} \|_F \right) \| \Delta_R \Sigma^1 \|_F
$$

$$
54
$$
where we have used the variational representation of the Frobenius norm for some $\tilde{L} \in \mathbb{R}^{n_1 \times r}$ obeying $\|\tilde{L}\|_F = 1$. Note that the decomposition of $\sqrt{\mathcal{P}_4}$ is extremely similar to that of $\mathcal{P}_2$. Therefore we can follow a similar argument (i.e. applying Lemma 21 and 22) to control these terms as

$$
\mathcal{P}_{4,1} \leq \frac{C_1}{(1 - \epsilon)^3} \sqrt{\frac{\mu r \log(n_1 \lor n_2)}{p(n_1 \land n_2)}} \left( \|\Delta_L \Sigma_{\star}^{1/2}\|_F + \|\Delta_R \Sigma_{\star}^{1/2}\|_F \right);
$$

$$
\mathcal{P}_{4,2} \leq \frac{C_2 \kappa}{(1 - \epsilon)^2} \left( 1 + C_{BK} (1 - \epsilon) \right) \sqrt{\frac{n_1 \land n_2}{p}} \frac{\mu r}{\sqrt{n_1 n_2}} \|\Delta_R \Sigma_{\star}^{1/2}\|_F;
$$

$$
\mathcal{P}_{4,3} \leq \frac{C_2}{(1 - \epsilon)^2} \left( 1 + C_{BK} (1 - \epsilon) \right) \sqrt{\frac{n_1 \land n_2}{p}} \frac{\mu r}{\sqrt{n_1 n_2}} \|\Delta_R \Sigma_{\star}^{1/2}\|_F;
$$

$$
\mathcal{P}_{4,4} \leq \frac{C_2 C_{BK}}{(1 - \epsilon)^3} \left( 1 + C_{BK} (1 - \epsilon) \right) \sqrt{\frac{n_1 \land n_2}{p}} \frac{\mu r}{\sqrt{n_1 n_2}} \|\Delta_R \Sigma_{\star}^{1/2}\|_F.
$$

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

$$
\sqrt{\mathcal{P}_4} \leq \delta_1 \|\Delta_L \Sigma_{\star}^{1/2}\|_F + (\delta_1 + \delta_2) \|\Delta_R \Sigma_{\star}^{1/2}\|_F.
$$

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

$$
\mathcal{P}_4 \leq \delta_1 (2\delta_1 + \delta_2) \|\Delta_L \Sigma_{\star}^{1/2}\|_F^2 + (\delta_1 + \delta_2)(2\delta_1 + \delta_2) \|\Delta_R \Sigma_{\star}^{1/2}\|_F^2.
$$

\[\square\]

### 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 $X \in \mathbb{R}^{n_1 \times n_2}$, with overwhelming probability, one has

$$
\| (p^{-1} \mathcal{P}_{\Omega} - I)(X) \|_{op} \leq C_0 \frac{\log(n_1 \lor n_2)}{p} \| X \|_\infty + C_0 \sqrt{\frac{\log(n_1 \lor n_2)}{p}} (\| X \|_{2,\infty} \lor \| X^\top \|_{2,\infty}),
$$

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

In view of Lemma 11, one has

$$
\text{dist}(\tilde{F}_0, F_{\star}) \leq \sqrt{\frac{2}{r}} + 1 \| U_0 \Sigma_0 V_0^\top - X_{\star} \|_F \leq \sqrt{(\sqrt{2} + 1)2r} \| U_0 \Sigma_0 V_0^\top - X_{\star} \|_{op},
$$

where the last relation uses the fact that $U_0 \Sigma_0 V_0^\top - X_{\star}$ has rank at most $2r$. Applying the triangle inequality, we obtain

$$
\| U_0 \Sigma_0 V_0^\top - X_{\star} \|_{op} \leq \| p^{-1} \mathcal{P}_{\Omega}(X_{\star}) - U_0 \Sigma_0 V_0^\top \|_{op} + \| p^{-1} \mathcal{P}_{\Omega}(X_{\star}) - X_{\star} \|_{op}
$$

$$
\leq 2 \| (p^{-1} \mathcal{P}_{\Omega} - I)(X_{\star}) \|_{op}.
$$

Here the second inequality hinges on the fact that $U_0 \Sigma_0 V_0^\top$ is the best rank-$r$ approximation to $p^{-1} \mathcal{P}_{\Omega}(X_{\star})$, i.e.

$$
\| p^{-1} \mathcal{P}_{\Omega}(X_{\star}) - U_0 \Sigma_0 V_0^\top \|_{op} \leq \| p^{-1} \mathcal{P}_{\Omega}(X_{\star}) - X_{\star} \|_{op}.
$$

Combining (70) and (71) yields

$$
\text{dist}(\tilde{F}_0, F_{\star}) \leq 2 \sqrt{(\sqrt{2} + 1)2r} \| (p^{-1} \mathcal{P}_{\Omega} - I)(X_{\star}) \|_{op} \leq 5 \sqrt{r} \| (p^{-1} \mathcal{P}_{\Omega} - I)(X_{\star}) \|_{op}.
$$
It then boils down to controlling $\|p^{-1}P_\Omega(X_\star) - X_\star\|_{op}$, which is readily supplied by Lemma 23 as

$$\|p^{-1}P_\Omega - I\|_{op}(X_\star) \leq C_0 \frac{\log(n_1 \lor n_2)}{p} \|X_\star\|_{\infty} + C_0 \sqrt{\frac{\log(n_1 \lor n_2)}{p}} (\|X_\star\|_{2,\infty} \lor \|X_\star^\top\|_{2,\infty}),$$

which holds with overwhelming probability. The proof is finished by plugging the following bounds from incoherence assumption of $X_\star$:

$$\|X_\star\|_{\infty} \leq \|U_\star\|_{2,\infty}\|\Sigma_\star\|_{op}\|V_\star\|_{2,\infty} \leq \frac{\mu r}{\sqrt{n_1 n_2}} \kappa \sigma_r(X_\star);$$

$$\|X_\star\|_{2,\infty} \leq \|U_\star\|_{2,\infty}\|\Sigma_\star\|_{op}\|V_\star\|_{op} \leq \sqrt{\frac{\mu r}{n_1}} \kappa \sigma_r(X_\star);$$

$$\|X_\star^\top\|_{2,\infty} \leq \|U_\star\|_{op}\|\Sigma_\star\|_{op}\|V_\star\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n_2}} \kappa \sigma_r(X_\star).$$

\section{Proof for General Loss Functions}

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

**Lemma 24.** Suppose that $f(\cdot): \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}$ is rank-$2r$ restricted $L$-smooth and rank-$2r$ restricted convex. Then for any $X_1, X_2 \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $r$, one has

$$\langle \nabla f(X_1) - \nabla f(X_2), X_1 - X_2 \rangle \geq \frac{1}{L} \|\nabla f(X_1) - \nabla f(X_2)\|_{F,r}^2.$$

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

$$f(X_1) + \langle \nabla f(X_1), \bar{X} - X_1 \rangle \leq f(\bar{X}) \leq f(X_2) + \langle \nabla f(X_2), \bar{X} - X_2 \rangle + \frac{L}{2} \|\bar{X} - X_2\|_{F}^2.$$

Reorganize the terms to yield

$$f(X_1) + \langle \nabla f(X_1), X_2 - X_1 \rangle \leq f(X_2) + \langle \nabla f(X_2) - \nabla f(X_1), \bar{X} - X_2 \rangle + \frac{L}{2} \|\bar{X} - X_2\|_{F}^2.$$

Take $\bar{X} = X_2 - \frac{1}{L} P_r(\nabla f(X_2) - \nabla f(X_1))$, whose rank is at most $2r$, to see

$$f(X_1) + \langle \nabla f(X_1), X_2 - X_1 \rangle + \frac{1}{2L} \|\nabla f(X_2) - \nabla f(X_1)\|_{F,r}^2 \leq f(X_2).$$

We can further switch the roles of $X_1$ and $X_2$ to obtain

$$f(X_2) + \langle \nabla f(X_2), X_1 - X_2 \rangle + \frac{1}{2L} \|\nabla f(X_2) - \nabla f(X_1)\|_{F,r}^2 \leq f(X_1).$$

Adding the above two inequalities yields the desired bound. \qed

\subsection{Proof of Theorem 4}

Suppose that the $t$-th iterate $F_t$ obeys the condition $\text{dist}(F_t, F_\star) \leq 0.1 \sigma_r(X_\star)/\sqrt{\kappa_f}$. In view of Lemma 9, one knows that $Q_\star$, the optimal alignment matrix between $F_t$ and $F_\star$ exists. Therefore, for notational convenience, denote $L := L_\star Q_\star, R := R_\star Q_\star^\top, \Delta_L := L - L_\star, \Delta_R := R - R_\star$, and $\epsilon := 0.1/\sqrt{\kappa_f}$. Similar to the derivation in (45), we have

$$\|\Delta_L \Sigma_\star^{-1/2}\|_{op} \lor \|\Delta_R \Sigma_\star^{-1/2}\|_{op} \leq \epsilon. \quad (72)$$

The conclusion $\|L_t R_t^\top - X_\star\|_F \leq 1.5 \text{dist}(F_t, F_\star)$ 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 dist\((F_{t+1}, F_{*})\), one has
\[
\text{dist}^2(F_{t+1}, F_{*}) \leq \left\| (L_{t+1}Q_t - L_{*})\Sigma_{*}^{1/2} \right\|_F^2 + \left\| (R_{t+1}Q_t^* - R_{*})\Sigma_{*}^{1/2} \right\|_F^2.
\] (73)

Introduce an auxiliary function
\[
f_{\mu}(X) = f(X) - \frac{\mu}{2} \| X - X_{*} \|_F^2,
\]
which is rank-2r restricted \((L - \mu)\)-smooth and rank-2r restricted convex. Using the \texttt{ScaledGD} update rule (27) and the decomposition \(LR^T - X_{*} = \Delta_L R^T + L_{*} \Delta_R^T\), we obtain
\[
(L_{t+1}Q_t - L_{*})\Sigma_{*}^{1/2} = (L - \eta \nabla f(LR^T)R(R^TR)^{-1} - L_{*})\Sigma_{*}^{1/2}
\]
\[
= (L - \eta \mu(LR^T - X_{*})R(R^TR)^{-1} - \eta \nabla f_{\mu}(LR^T)R(R^TR)^{-1} - L_{*})\Sigma_{*}^{1/2}
\]
\[
= (1 - \eta \mu)\Delta_L \Sigma_{*}^{1/2} - \eta \mu L_{*}\Delta_R^T R(R^TR)^{-1} \Sigma_{*}^{1/2} - \eta \nabla f_{\mu}(LR^T)R(R^TR)^{-1} \Sigma_{*}^{1/2}.
\]
As a result, one can expand the first square in (73) as
\[
\left\| (L_{t+1}Q_t - L_{*})\Sigma_{*}^{1/2} \right\|_F^2 = \left\| (1 - \eta \mu)\Delta_L \Sigma_{*}^{1/2} - \eta \mu L_{*}\Delta_R^T R(R^TR)^{-1} \Sigma_{*}^{1/2} \right\|_F^2
\]
\[
\Phi_1 - 2\eta(1 - \eta \mu) \left\langle \nabla f_{\mu}(LR^T), \Delta_L \Sigma_{*}^{1/2} - \eta \mu L_{*}\Delta_R^T R(R^TR)^{-1} \Sigma_{*}^{1/2} \right\rangle
\]
\[
+ 2\eta^2 \mu \left\langle \nabla f_{\mu}(LR^T), L_{*}\Delta_R^T R(R^TR)^{-1} \Sigma_{*}^{1/2} \right\rangle
\]
In the sequel, we shall bound the four terms separately.

1. Notice that the main term \(\Phi_1\) has already been controlled in (46) under the condition (72). It obeys
\[
\Phi_1 \leq \left(1 - \eta \mu\right)^2 \frac{2\epsilon}{1 - \epsilon} \eta \mu(1 - \eta \mu) \left\| \Delta_L \Sigma_{*}^{1/2} \right\|_F^2 + \frac{2\epsilon + \epsilon^2}{1 - \epsilon} \eta^2 \mu^2 \left\| \Delta_R \Sigma_{*}^{1/2} \right\|_F^2,
\]
as long as \(\eta \mu \leq 2/3\).

2. For the second term \(\Phi_2\), note that \(\Delta_L \Sigma_{*}(R^TR)^{-1} R^T - \Delta_L R_{*}^T - \frac{1}{2} \Delta_L \Delta_R^T\) has rank at most \(r\). Hence we can invoke Lemma 15 to obtain
\[
|\Phi_2| \leq \left\| \nabla f_{\mu}(LR^T) \right\|_{F,r} \left\| \Delta_L \Sigma_{*}(R^TR)^{-1} R^T - \Delta_L R_{*}^T - \frac{1}{2} \Delta_L \Delta_R^T \right\|_F
\]
\[
\leq \left\| \nabla f_{\mu}(LR^T) \right\|_{F,r} \left\| \Delta_L \Sigma_{*}^{1/2} \right\|_F \left( \left\| (R(R^TR)^{-1} \Sigma_{*}^{1/2} - V_{*}) \right\|_{op} + \frac{1}{2} \left\| \Delta_R \Sigma_{*}^{-1/2} \right\|_{op} \right),
\]
where the second line uses \(R_{*} = V_{*} \Sigma_{*}^{1/2}\). Take the condition (72) and Lemma 12 together to obtain
\[
\left\| (R(R^TR)^{-1} \Sigma_{*}^{1/2} - V_{*}) \right\|_{op} \leq \frac{1}{1 - \epsilon};
\]
\[
\left\| (R(R^TR)^{-1} \Sigma_{*}^{1/2} - V_{*}) \right\|_{op} \leq \frac{\sqrt{2\epsilon}}{1 - \epsilon}.
\]

57
These consequences further imply that
\[
|\mathfrak{E}_2| \leq (\sqrt{2}\epsilon + \frac{\epsilon}{2}) ||\nabla f_\mu(\mathbf{L}^T\mathbf{R}^T)||_{F,r} \|\Delta_L \Sigma_*^{1/2}\|_F.
\]

3. As above, the third term \(\mathfrak{E}_3\) can be similarly bounded as
\[
|\mathfrak{E}_3| \leq \||\nabla f_\mu(\mathbf{L}^T\mathbf{R}^T)||_{F,r} \|\mathbf{L}_* \Delta_\mu^\top \mathbf{R}(\mathbf{R}^\top \mathbf{R})^{-1} \Sigma_*(\mathbf{R}^\top \mathbf{R})^{-1} \mathbf{R}^\top\|_F
\]
\[
\leq \||\nabla f_\mu(\mathbf{L}^T\mathbf{R}^T)||_{F,r} \|\Delta_\mu \Sigma_*^{1/2}\|_F \|\mathbf{R}(\mathbf{R}^\top \mathbf{R})^{-1} \Sigma_*^{1/2}\|_{op}^2
\]
\[
\leq \frac{1}{(1-\epsilon)^2} \||\nabla f_\mu(\mathbf{L}^T\mathbf{R}^T)||_{F,r} \|\Delta_\mu \Sigma_*^{1/2}\|_F.
\]

4. For the last term \(\mathfrak{E}_4\), invoke Lemma 15 to obtain
\[
\mathfrak{E}_4 \leq \||\nabla f_\mu(\mathbf{L}^T\mathbf{R}^T)||_{F,r}^2 \|\mathbf{R}(\mathbf{R}^\top \mathbf{R})^{-1} \Sigma_*^{1/2}\|_{op}^2 \leq \frac{1}{(1-\epsilon)^2} \||\nabla f_\mu(\mathbf{L}^T\mathbf{R}^T)||_{F,r}^2.
\]

Taking collectively the bounds for \(\mathfrak{E}_1, \mathfrak{E}_2, \mathfrak{E}_3\) and \(\mathfrak{E}_4\) yields
\[
\left\| (\mathbf{L}_{t+1} \mathbf{Q}_t - \mathbf{L}_*) \Sigma_*^{1/2} \right\|^2_\mathbb{F} \leq \left( (1-\eta \mu)^2 + \frac{2\epsilon}{1-\epsilon} \eta \mu (1-\eta \mu) \right) \left\| \Delta_L \Sigma_*^{1/2} \right\|^2_\mathbb{F} + \frac{2\epsilon + \epsilon^2}{(1-\epsilon)^2} \eta_\mu \Sigma_* \left\| \Delta_\mu \Sigma_*^{1/2} \right\|^2_\mathbb{F}
\]
\[
+ 2\eta \sqrt{2}\epsilon + \frac{\epsilon}{2} (1-\eta \mu) \left\| \nabla f_\mu(\mathbf{L}^T\mathbf{R}^T) \right\|_{F,r} \left\| \Delta_\mu \Sigma_*^{1/2} \right\|_F
\]
\[
- 2\eta (1-\eta \mu) \left( \nabla f_\mu(\mathbf{L}^T\mathbf{R}^T), \Delta_L R^\top_* + \frac{1}{2} \Delta_\mu \Delta_R^\top \right)
\]
\[
+ \frac{2\eta^2}{(1-\epsilon)^2} \left\| \nabla f_\mu(\mathbf{L}^T\mathbf{R}^T) \right\|_{F,r} \left\| \Delta_\mu \Sigma_*^{1/2} \right\|_F + \frac{\eta^2}{(1-\epsilon)^2} \left\| \nabla f_\mu(\mathbf{L}^T\mathbf{R}^T) \right\|^2_{F,r}.
\]

Similarly, we can obtain the control of \(\left\| (\mathbf{R}_{t+1} \mathbf{Q}_t^\top - \mathbf{R}_*) \Sigma_*^{1/2} \right\|^2_\mathbb{F}. \) Combine them together to reach
\[
\left\| (\mathbf{L}_{t+1} \mathbf{Q}_t - \mathbf{L}_*) \Sigma_*^{1/2} \right\|^2_\mathbb{F} + \left\| (\mathbf{R}_{t+1} \mathbf{Q}_t^\top - \mathbf{R}_*) \Sigma_*^{1/2} \right\|^2_\mathbb{F}
\]
\[
\leq \left( (1-\eta \mu)^2 + \frac{2\epsilon}{1-\epsilon} \eta \mu (1-\eta \mu) + \frac{2\epsilon + \epsilon^2}{(1-\epsilon)^2} \eta_\mu \Sigma_* \right) \left( \left\| \Delta_L \Sigma_*^{1/2} \right\|^2_\mathbb{F} + \left\| \Delta_\mu \Sigma_*^{1/2} \right\|_F \right)
\]
\[
+ 2\eta \left( \sqrt{2}\epsilon + \frac{\epsilon}{2} (1-\eta \mu) + \frac{\eta \mu}{(1-\epsilon)^2} \right) \left\| \nabla f_\mu(\mathbf{L}^T\mathbf{R}^T) \right\|_{F,r} \left( \left\| \Delta_L \Sigma_*^{1/2} \right\|_F + \left\| \Delta_\mu \Sigma_*^{1/2} \right\|_F \right)
\]
\[
- 2\eta (1-\eta \mu) \left( \nabla f_\mu(\mathbf{L}^T\mathbf{R}^T), \Delta_L R^\top_* + \mathbf{L}_* \Delta_R^\top + \Delta_\mu \Delta_R^\top \right) + \frac{2\eta^2}{(1-\epsilon)^2} \left\| \nabla f_\mu(\mathbf{L}^T\mathbf{R}^T) \right\|^2_{F,r}
\]
\[
\leq \left( (1-\eta \mu)^2 + \frac{2\epsilon}{1-\epsilon} \eta \mu (1-\eta \mu) + \frac{2\epsilon + \epsilon^2}{(1-\epsilon)^2} \eta_\mu \Sigma_* \right) \left( \left\| \Delta_L \Sigma_*^{1/2} \right\|^2_\mathbb{F} + \left\| \Delta_\mu \Sigma_*^{1/2} \right\|_F \right)
\]
\[
+ 2\eta \left( \sqrt{2}\epsilon + \frac{\epsilon}{2} (1-\eta \mu) + \frac{\eta \mu}{(1-\epsilon)^2} \right) \left\| \nabla f_\mu(\mathbf{L}^T\mathbf{R}^T) \right\|_{F,r} \left( \left\| \Delta_L \Sigma_*^{1/2} \right\|_F + \left\| \Delta_\mu \Sigma_*^{1/2} \right\|_F \right)
\]
\[
- 2\eta \left( \frac{1-\eta \mu}{L - \mu} + \frac{\eta}{(1-\epsilon)^2} \right) \left\| \nabla f_\mu(\mathbf{L}^T\mathbf{R}^T) \right\|^2_{F,r},
\]

where the last line follows from Lemma 24 (notice that \( \nabla f_\mu(\mathbf{X}_*) = 0 \)) as
\[
(\nabla f_\mu(\mathbf{L}^T\mathbf{R}^T), \Delta_L R^\top_* + \mathbf{L}_* \Delta_R^\top + \Delta_\mu \Delta_R^\top) = (\nabla f_\mu(\mathbf{L}^T\mathbf{R}^T), \mathbf{L}^T\mathbf{R}^T - \mathbf{X}_*) \geq \frac{1}{L - \mu} ||\nabla f_\mu(\mathbf{L}^T\mathbf{R}^T)||_{F,r}^2.
\]
Notice that $c_2 > 0$ as long as $\eta \leq (1 - \epsilon)^2 / L$. Maximizing the quadratic function of $\| \nabla f_\mu(LR^T) \|_{F,r}$ yields

$$c_1 \| \nabla f_\mu(LR^T) \|_{F,r} \left( \| \Delta_L \Sigma_\star^{1/2} \|_F + \| \Delta_R \Sigma_\star^{1/2} \|_F \right) - c_2 \| \nabla f_\mu(LR^T) \|_{F,r}^2 \leq \frac{c_1^2}{4c_2} \left( \| \Delta_L \Sigma_\star^{1/2} \|_F + \| \Delta_R \Sigma_\star^{1/2} \|_F \right)^2$$

where the last inequality holds since $(a + b)^2 \leq 2(a^2 + b^2)$. Identify $\| (L_{t+1} Q_t - L_\star) \Sigma_\star^{1/2} \|_F^2 + \| (R_{t+1} Q_t^\top - R_\star) \Sigma_\star^{1/2} \|_F^2$ to obtain

$$\| (L_{t+1} Q_t - L_\star) \Sigma_\star^{1/2} \|_F^2 + \| (R_{t+1} Q_t^\top - R_\star) \Sigma_\star^{1/2} \|_F^2 \leq \rho^2 (\eta; \epsilon, \mu, L) \text{dist}^2 (F_t, F_\star),$$

where the contraction rate is given by

$$\rho^2 (\eta; \epsilon, \mu, L) := (1 - \eta \mu)^2 + \frac{2\epsilon}{1 - \epsilon} \eta \mu (1 - \eta \mu) + \frac{2\epsilon + \epsilon^2}{(1 - \epsilon)^2} \eta^2 \mu^2 + \frac{\left( \frac{\sqrt{2 \epsilon}}{1 - \epsilon} + \frac{\epsilon}{2} \right) (1 - \eta \mu) + \frac{\eta \mu}{(1 - \epsilon) \mu}}{1 - \eta \mu} \eta L.(\mu).$$

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

$$\text{dist} (F_{t+1}, F_\star) \leq \sqrt{\| (L_{t+1} Q_t - L_\star) \Sigma_\star^{1/2} \|_F^2 + \| (R_{t+1} Q_t^\top - R_\star) \Sigma_\star^{1/2} \|_F^2} \leq (1 - 0.7\eta \mu) \text{dist} (F_t, F_\star),$$

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

$$\left( \frac{\sqrt{2 \epsilon}}{1 - \epsilon} + \frac{\epsilon}{2} \right) (1 - \eta \mu) + \frac{\eta \mu}{(1 - \epsilon) \mu} \eta L.(\mu) \leq \frac{0.0016 L}{\kappa_f} + 0.078\eta \mu + 1.005\eta^2 \mu^2.$$

where the last line uses the definition (28) of $\kappa_f$. 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 $0 < \eta \mu \leq 0.4$. 

59