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

Tian Tong∗ Carnegie Mellon University
Cong Ma† Princeton University
Yuejie Chi∗ Carnegie Mellon University

May 2020; Revised June 2020

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 new algorithmic approach dubbed Scaled Gradient Descent (ScaledGD) which can be viewed as pre-conditioned or diagonally-scaled gradient descent, where the pre-conditioners are adaptive and iteration-varying with a minimal computational overhead. For low-rank matrix sensing and robust principal component analysis, 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. To the best of our knowledge, ScaledGD is the first algorithm that provably has such properties. Our analysis is also applicable to general loss functions that are restricted strongly convex and smooth over low-rank matrices. At the core of our analysis is the introduction of a new distance function that takes account of the pre-conditioners when measuring the distance between the iterates and the ground truth. As a by product of our analysis, we also remove the unnecessary regularizations that either balance the norms or maintain incoherence properties of the two low-rank factors used in previous works, thus unveiling the implicit regularization property of ScaledGD. 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, general losses

Contents

1 Introduction 2

1.1 A New Algorithm: Scaled Gradient Descent 3
1.2 Related Work 4
1.3 Paper Organization and Notations 5

∗Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Emails: {ttong1,yuejiec}@andrew.cmu.edu.
†Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544, USA; Email: congm@princeton.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:

$$\minimize_{X \in \mathbb{R}^{n_1 \times n_2}} f(X) := \frac{1}{2} \|y - A(X)\|_2^2 \text{ subject to } \text{rank}(X) \leq r,$$

(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 matrix space 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:

\[
\min_{L \in \mathbb{R}^{n_1 \times r}, R \in \mathbb{R}^{n_2 \times r}} \mathcal{L}(L, R) := f(LR^\top).
\]  

(2)

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

However, upon closer examination, existing approaches such as gradient descent and alternating minimization are still computationally expensive especially for ill-conditioned matrices. Take low-rank matrix sensing as an example: 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 \( L \), \( [TBS+16] \); 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?

1.1 A New Algorithm: Scaled Gradient Descent

In this paper, we answer this question affirmatively by proposing a new algorithm called scaled gradient descent (\texttt{ScaledGD}) to optimize (2). Given an initialization \((L_0, R_0)\), \texttt{ScaledGD} proceeds as follows

\[
\begin{align*}
L_{t+1} & = L_t - \eta \nabla_L \mathcal{L}(L_t, R_t)(R_t^\top R_t)^{-1}, \\
R_{t+1} & = R_t - \eta \nabla_R \mathcal{L}(L_t, R_t)(L_t^\top L_t)^{-1},
\end{align*}
\]

(3)

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 pre-conditioner 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 pre-conditioners that are both theoretically sound and practically cheap to compute. Such requirements are met by \texttt{ScaledGD}, where the pre-conditioners are computed by inverting two \( r \times r \) matrices, whose size is much smaller than the dimension of matrix factors. Therefore, each iteration of \texttt{ScaledGD} adds minimal overhead to the gradient computation and has the order-wise same per-iteration cost as gradient descent. Moreover, the pre-conditioners are adaptive and iteration-varying.

Theoretically, we show that \texttt{ScaledGD} achieves linear convergence at a rate \textit{independent} of the condition number of the matrix when initialized properly, e.g., using the standard spectral method, for two canonical problems: low-rank matrix sensing and robust PCA. Table 1 summarizes the performance guarantees of \texttt{ScaledGD} in terms of both statistical and computational complexities with comparisons to prior algorithms.

- \textbf{Low-rank matrix sensing.} As long as the measurement operator satisfies the standard restricted isometry property (RIP) with an RIP constant \( \delta_{2r} \lesssim 1/(\sqrt{\kappa}r) \), where \( \kappa \) is the condition number of \( X_* \), \texttt{ScaledGD} reaches \( \epsilon \)-accuracy in \( O(\log 1/\epsilon) \) iterations when initialized by the spectral method. This strictly improves the iteration complexity \( O(\kappa \log 1/\epsilon) \) of gradient descent in \([TBS+16]\) under the same sample complexity requirement.

- \textbf{Robust PCA.} Under the deterministic corruption model \([CSPW11]\), as long as the fraction \( \alpha \) of corruptions per row / column satisfies \( \alpha \lesssim 1/(\mu r^{3/2}\kappa) \), where \( \mu \) is the incoherence parameter of \( X_* \), \texttt{ScaledGD} in
| Algorithms          | Low-rank matrix sensing | Robust PCA |
|---------------------|-------------------------|------------|
| GD [TBS'16]        | $nr^2\kappa^2$          | $\kappa \log \frac{1}{\varepsilon}$ |
| AltMinSense [JNS'13]| $nr^3\kappa^4$          | $\log \frac{1}{\varepsilon}$ |
| ScaledGD (this paper)| $nr^2\kappa^2$          | $\frac{1}{\mu r^2 \kappa}$ |
| ProjGD [YPCC'16]   | $\frac{1}{\mu r^2 \kappa^2 \log \frac{1}{\varepsilon}}$ | $\kappa \log \frac{1}{\varepsilon}$ |
| AltProj [NNS'14]   | $\frac{1}{\mu r}$      | $\log \frac{1}{\varepsilon}$ |
| ScaledGD (this paper)| $\frac{1}{\mu r^2 \kappa}$ | $\log \frac{1}{\varepsilon}$ |

Table 1: Comparisons of ScaledGD with prior algorithms (with spectral initialization): ScaledGD has a comparable per-iteration cost as gradient descent (GD), while the per-iteration cost of alternating minimization is significantly higher especially for large problems. Here, we say that the output $X$ of an algorithm reaches $\varepsilon$-accuracy, if it satisfies $\|X - X^*\|_F \leq \varepsilon \sigma_r(X^*)$. Here, $n := \max\{n_1, n_2\}$, $\kappa$ and $\mu$ are the condition number and incoherence parameter of $X^*$.

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

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

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

$$F_{t+1} = F_t - \eta \nabla_F \mathcal{L}(F_t)(F_t^\top F_t)^{-1}. \quad (4)$$

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

1.2 Related Work

Our work contributes to the growing literature of design and analysis of provable nonconvex optimization procedures for high-dimensional signal estimation; see e.g. [JK17, CC18, CLC19] for an overview. A 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 do not exist 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 and completion [JNS13, ZL15, SL16, CW15, MWCC19, CLL19], 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 $\frac{1}{2}\|L^T L - R^T R\|_F^2$ [TBS+16, PKCS17] or $\frac{1}{2}\|L\|_F^2 + \frac{1}{2}\|R\|_F^2$ [ZLTW18, CCF+19, 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+19, DHL18, MLC19]. 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] (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.

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. However, this approach has no statistical and computational guarantees for its global convergence, despite its excellent empirical performance. Our analysis of ScaledGD can be viewed as providing partial justifications to the ASD method.

1.3 Paper Organization and Notations
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 and robust PCA. Section 3 provides the theoretical guarantees of ScaledGD for the two problems 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 4, we outline the proof for our main results. Section 5 illustrates the excellent empirical performance of ScaledGD in a variety of low-rank matrix estimation problems. Finally, we conclude in Section 6.

Before continuing, we introduce several notations used throughout the paper. First of all, we use boldface
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 $A_{:,j}$ denote its $i$th row and $j$th column, respectively. In addition, $\|A\|_{op}$, $\|A\|_F$, $\|A\|_{1,\infty}$, $\|A\|_{2,\infty}$, and $\|A\|_\infty$ stand for the spectral norm (i.e. the largest singular value), the Frobenius norm, the $\ell_{1,\infty}$ norm (i.e. the largest $\ell_1$ norm of the rows), the $\ell_{2,\infty}$ 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 $A$. 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. We denote

$$P_r(A) = \min_{A:\text{rank}(A) = r} \|A - \tilde{A}\|_F^2$$

(5)

as the rank-$r$ approximation of $A$, which is given by the rank-$r$ SVD of $A$ by the Eckart-Young-Mirsky theorem. We also use $\text{vec}(A)$ to denote the vectorization of a matrix $A$. Last but not least, we use the shorthand notation $a \lor b = \max\{a, b\}$ and $a \land b = \min\{a, b\}$.

## 2 Scaled Gradient Descent for Low-Rank Matrix Estimation

### 2.1 Interpretation

Before we instantiate the proposed ScaledGD algorithm on two 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

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

$$= \text{vec}(F_t) - \eta H_t^{-1} \text{vec}(\nabla_F \mathcal{L}(F_t)),$$

(6)

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. More importantly, to see why it serves as a reasonable approximation to the Hessian of $\mathcal{L}(F)$ at $F_t$, let us consider the problem of factorizing a matrix $X$ into two low-rank factors:

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

(7)

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

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

$$R_{t+1} = R_t - \eta(L_t R_t^\top - X_s)^\top L_t (L_t^\top L_t)^{-1}.$$  

(8)

The following proposition reveals that ScaledGD is equivalent to approximating the Hessian of the loss function in (7) by only keeping its diagonal blocks.

**Proposition 1.** For the matrix factorization problem (7), ScaledGD is equivalent to the following update rule:

$$\text{vec}(F_{t+1}) = \text{vec}(F_t) - \eta \begin{bmatrix} \nabla^2_{LL} \mathcal{L}(F_t) & 0 \\ 0 & \nabla^2_{RR} \mathcal{L}(F_t) \end{bmatrix}^{-1} \text{vec}(\nabla_F \mathcal{L}(F_t)).$$

Here, $\nabla^2_{LL} \mathcal{L}(F_t)$ (resp. $\nabla^2_{RR} \mathcal{L}(F_t)$) denotes the second order derivative w.r.t. $L$ (resp. $R$) at $F_t$.  
Nonetheless, it sheds light on why conditioning. However, for general loss functions, the above interpretation no longer holds in its exact form.

Therefore, (9) shows that with \( \eta \in [0, 1] \), the next iterate of ScaledGD can be interpreted as a convex combination of the current iterate and the least-squares update (10), where the latter is robust to ill-conditioning. However, for general loss functions, the above interpretation no longer holds in its exact form. Nonetheless, it sheds light on why ScaledGD is robust to ill-conditioning.

2.2 ScaledGD for Low-Rank Matrix Sensing

Assume we have collected a set of linear measurements about a rank-\( r \) matrix \( X_\star \in \mathbb{R}^{n_1 \times n_2} \), given as

\[
y = A(X_\star) \in \mathbb{R}^m,
\]

where \( A(X) = \{(A_i, X)\}_{i=1}^m : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}^m \) is the linear map modeling the measurement process. The goal of low-rank matrix sensing is to recover \( X_\star \) from \( y \), especially when the number of measurements \( m \ll n_1n_2 \), by exploiting the low-rank property. This problem has wide applications in medical imaging, signal processing, and data compression [CP11]. Writing \( X \in \mathbb{R}^{n_1 \times n_2} \) into a factored form \( X = LR^\top \), we consider the following optimization problem:

\[
\begin{align*}
\text{minimize} &\quad \mathcal{L}(F) = \frac{1}{2} \| A(LR^\top) - y \|_2^2, \\
\text{subject to} &\quad F \in \mathbb{R}^{(n_1+n_2) \times r}.
\end{align*}
\]

Here as before, \( F \) denotes the stacked factor matrix \([L^\top, R^\top]^\top\). We suggest running ScaledGD (3) with the spectral initialization to solve (12), which performs the top-\( r \) SVD on \( A^\star(y) \), where \( A^\star(\cdot) \) is the adjoint operator of \( A(\cdot) \). The full algorithm is stated in Algorithm 1. The low-rank matrix can be estimated as \( X_T = L_T R_T^\top \) after running \( T \) iterations.

\begin{algorithm}
\caption{ScaledGD for low-rank matrix sensing with spectral initialization}
\begin{algorithmic}
\Statex \textbf{Spectral initialization:} Let \( U_0 \Sigma_0 V_0^\top \) be the top-\( r \) SVD of \( A^\star(y) \), and set \( L_0 = U_0 \Sigma_0^{1/2} \) and \( R_0 = V_0 \Sigma_0^{1/2} \).
\Statex \textbf{Scaled gradient updates:} for \( t = 0, 1, 2, \ldots, T-1 \) do
\Statex \quad \quad \quad \quad L_{t+1} = L_t - \eta A^\star(A(L_t R_t^\top) - y) R_t (R_t^\top R_t)^{-1},
\Statex \quad \quad \quad \quad R_{t+1} = R_t - \eta A^\star(A(L_t R_t^\top) - y)^\top L_t (L_t^\top L_t)^{-1}.
\end{algorithmic}
\end{algorithm}

2.3 ScaledGD for Robust PCA

Assume that we have observed the data matrix \( Y = X_\star + S_\star \), which is a superposition of a rank-\( r \) matrix \( X_\star \), modeling the clean data, and a sparse matrix \( S_\star \), modeling the corruption or outliers. The goal of robust PCA [CLMW11, CSPW11] is to separate the two matrices \( X_\star \) and \( S_\star \) from their mixture \( Y \). This problem finds numerous applications in video surveillance, image processing, and so on.
**Algorithm 2 ScaledGD for robust PCA with spectral initialization**

**Spectral initialization:** Let $U_0^\top \Sigma_0 V_0^\top$ be the top-r SVD of $Y - \mathcal{T}_\alpha[Y]$, and set

$$L_0 = U_0 \Sigma_0^{1/2}, \quad \text{and} \quad R_0 = V_0 \Sigma_0^{1/2}. \quad (18)$$

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

$$S_t = \mathcal{T}_\alpha[Y - L_t R_t^\top],$$

$$L_{t+1} = L_t - \eta(\bar{L}_t R_t^\top + S_t - Y)R_t(\bar{R}_t^\top R_t)^{-1},$$

$$R_{t+1} = R_t - \eta(\bar{L}_t R_t^\top + S_t - Y)^\top \bar{L}_t(\bar{L}_t^\top \bar{L}_t)^{-1}. \quad (19)$$

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

$$S_\alpha := \{S \in \mathbb{R}^{n_1 \times n_2} : ||S||_0 \leq \alpha n_2 \text{ for all } i, \text{ and } ||S||_0 \leq \alpha n_1 \text{ for all } j\}. \quad (15)$$

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

$$\minimize_{F \in \mathbb{R}^{(n_1+n_2) \times r}, S \in S_\alpha} \mathcal{L}(F, S) = \frac{1}{2} \| LR^\top + S - Y \|_F^2. \quad (16)$$

It is thus natural to alternatively update $F = [L^\top, R^\top]^\top$ and $S$, where $F$ is updated via the proposed ScaledGD algorithm, and $S$ is updated by hard thresholding, which trims the small entries of the residual matrix $Y - LR^\top$. 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:

$$\bar{T}_\alpha[A]_{i,j} = \begin{cases} A_{i,j}, & \text{if } |A|_{i,j} \geq |A|_{\bar{\alpha} n_2}, \text{ and } |A|_{i,j} \geq |A|_{\bar{\alpha} n_1,j}, \\ 0, & \text{otherwise} \end{cases}, \quad (17)$$

where $|A|_{i,k}$ (resp. $|A|_{k,j}$) 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 $\| L^\top L - R^\top R \|_F^2$ in the loss function (16), nor the projection of the low-rank factors onto the $\ell_2, \infty$ ball in each iteration.

## 3 Theoretical Guarantees

This section is devoted to formally establishing the statistical and computational guarantees of ScaledGD for solving low-rank matrix sensing, robust PCA, and low-rank matrix estimation with general loss functions.

We start with a few necessary notations. Denote by $U_\star, \Sigma_\star, V_\star^\top$ the compact singular value decomposition (SVD) of the rank-$r$ matrix $X_\star \in \mathbb{R}^{n_1 \times n_2}$, i.e. $X_\star := U_\star \Sigma_\star V_\star^\top$. Here $U_\star \in \mathbb{R}^{n_1 \times r}$ and $V_\star \in \mathbb{R}^{n_2 \times r}$ are composed of $r$ left and right singular vectors, respectively, and $\Sigma_\star \in \mathbb{R}^{r \times r}$ is a diagonal matrix consisting of $r$ singular values of $X_\star$ organized in a non-increasing order, i.e. $\sigma_1(X_\star) \geq \cdots \geq \sigma_r(X_\star) > 0$. Define

$$\kappa := \sigma_1(X_\star)/\sigma_r(X_\star) \quad (20)$$

as the condition number of $X_\star$. Define the ground truth low-rank factors as $L_\star := U_\star \Sigma_\star^{1/2}$ and $R_\star := V_\star \Sigma_\star^{1/2}$, so that $X_\star = L_\star R_\star^\top$. Correspondingly, denote the stacked factor matrix as $F_\star := [L_\star^\top, R_\star^\top]^\top \in \mathbb{R}^{(n_1+n_2) \times r}$. Obviously, the factored representation is not unique in that for any invertible matrix $Q \in \mathbb{R}^{r \times r}$, one has $LR^\top = (LQ)(RQ^{-1})^\top$. 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 pre-conditioning. Taking both considerations together leads to the following error metric:

$$\text{dist}^2(F, F^*) := \inf_{Q \in \mathbb{R}^{r \times r}} \left\| (LQ - L^*) \Sigma_2^{1/2} \right\|_F^2 + \left\| (RQ - R^*) \Sigma_1^{1/2} \right\|_F^2. \quad (21)$$

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

$$Q := \arg\min_{Q \in \mathbb{R}^{r \times r}} \left\| (LQ - L^*) \Sigma_2^{1/2} \right\|_F^2 + \left\| (RQ - R^*) \Sigma_1^{1/2} \right\|_F^2, \quad (22)$$

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. In comparison, the previously studied distance metrics (proposed mainly for GD) either do not include the diagonal scaling [MLC19, TBS+16], or only consider the ambiguity class up to orthonormal transforms [TBS+16], which fail to unveil the benefit of the newly proposed ScaledGD method.

### 3.1 Theoretical Guarantees for Low-Rank Matrix Sensing

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. (21)) — 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 $2r$-RIP with $\delta_{2r} \leq 0.02/(\sqrt{r})$. Then for all $t \geq 0$, the iterates of the ScaledGD method in Algorithm 1 satisfy

$$\text{dist}(F_t, F^*) \leq 0.1(1 - 0.6\eta)^t \sigma_r(X^*), \quad \text{and} \quad \|L_t R_t^T - X^*\|_F \leq 0.15(1 - 0.6\eta)^t \sigma_r(X^*),$$

as long as the step size obeys $0 < \eta \leq 2/3$.

Theorem 1 establishes that the distance $\text{dist}(F_t, F^*)$ 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 $n := \max\{n_1, n_2\}$. To reach $\epsilon$-accuracy, i.e. $\|L_t R_t^T - X^*\|_F \leq \epsilon \sigma_r(X^*)$, ScaledGD takes at most $T = O(\log(1/\epsilon))$ iterations, which is independent of the condition number $\kappa$ of $X^*$. In comparison, alternating minimization with spectral initialization (AltMinSense) converges in $O(\log(1/\epsilon))$ iterations as long as $m = O(nr^3\kappa^2)$ [JNS13], where the per-iteration cost is much higher requiring solving two linear systems of size $O(mn^2)$. 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 $X_{t+1} = P_r \left( X_t - \frac{1}{m} A^*(A(X_t) - y) \right)$ 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.
3.2 Theoretical Guarantees for Robust PCA

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 $X_*$ in robust PCA [Che15].

**Definition 2.** A rank-$r$ matrix $X_* \in \mathbb{R}^{n_1 \times n_2}$ with SVD $X_* = U_* \Sigma_* V_*^\top$ is said to be $\mu$-incoherent if
\[
\|U_*\|_{2,\infty} \leq \sqrt{\mu/n_1} \|U_*\|_F = \sqrt{\mu r/n_1}, \quad \text{and} \quad \|V_*\|_{2,\infty} \leq \sqrt{\mu/n_2} \|V_*\|_F = \sqrt{\mu r/n_2}.
\]

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 $X_*$ is $\mu$-incoherent and that the corruption rate $\alpha$ obeys $\alpha \leq c/(\mu r^{3/2} \kappa)$ for some sufficiently small constant $c > 0$. Then for all $t \geq 0$, the iterates of ScaledGD in Algorithm 2 satisfy
\[
\text{dist}(F_t, F_*) \leq 0.02(1 - 0.6\eta)^t \sigma_r(X_*), \quad \text{and} \quad \|L_t R_t^\top - X_*\|_F \leq 0.03(1 - 0.6\eta)^t \sigma_r(X_*),
\]
with the proviso that $0.1 \leq \eta \leq 2/3$.

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

3.3 Theoretical Guarantees under General Loss Functions

In the sequel, we generalize our analysis of ScaledGD to minimize a general loss function in the form of (2), where the update rule of ScaledGD is given by
\[
\begin{align*}
L_{t+1} &= L_t - \eta \nabla f(L_t R_t^\top) R_t (R_t^\top R_t)^{-1}, \\
R_{t+1} &= R_t - \eta \nabla f(L_t R_t^\top)^\top L_t (L_t^\top L_t)^{-1}.
\end{align*}
\]

(23)

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

**Definition 3 (Restricted smoothness).** For a differentiable function $f : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}$, we say $f$ is rank-$r$ restricted $L$-smooth for some $L > 0$, if
\[
f(X_2) \leq f(X_1) + \langle \nabla f(X_1), X_2 - X_1 \rangle + \frac{L}{2} \|X_2 - X_1\|_F^2,
\]
for any $X_1, X_2 \in \mathbb{R}^{n_1 \times n_2}$ with rank at most $r$.

**Definition 4 (Restricted strong convexity).** For a differentiable function $f : \mathbb{R}^{n_1 \times n_2} \rightarrow \mathbb{R}$, we say $f$ is rank-$r$ restricted $\mu$-strongly convex for some $\mu \geq 0$, if
\[
f(X_2) \geq f(X_1) + \langle \nabla f(X_1), X_2 - X_1 \rangle + \frac{\mu}{2} \|X_2 - X_1\|_F^2,
\]
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 to be

\[ \kappa_f := \frac{L}{\mu}, \tag{24} \]

which plays an important role in the convergence analysis. Encouragingly, many problems can be viewed as a special case of optimizing this general loss (23), including but not limited to:

- **low-rank matrix factorization**, where the loss function \( f(X) = \frac{1}{2}\|X - X_*\|_F^2 \) in (7) satisfies \( \kappa_f = 1 \);
- **low-rank matrix sensing**, where the loss function \( f(X) = \frac{1}{2}\|\mathcal{A}(X - X_*)\|_F^2 \) in (12) satisfies \( \kappa_f \approx 1 \) when \( \mathcal{A} \) obeys the rank-\( r \) RIP with a sufficiently small RIP constant;
- **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 \( X_* \).² The following theorem establishes that as long as properly initialized, then \texttt{ScaledGD} converges linearly at a constant rate.

**Theorem 3.** Suppose that \( f(\cdot) \) is rank-\( 2r \) restricted \( L \)-smooth and \( \mu \)-strongly convex, and \( X_* \) is the minimizer of \( f \). Suppose that the initialization \( F_0 \) satisfies \( \text{dist}(F_0, F_*) \leq 0.02 \sigma_r(X_*) \). Then for all \( t \geq 0 \), the iterates of \texttt{ScaledGD} in (23) satisfy

\[ \text{dist}(F_t, F_*) \leq 0.02(1 - 0.7\mu)^t \sigma_r(X_*), \text{ and } \|L_t R_t^T - X_*\|_F \leq 0.03(1 - 0.7\mu)^t \sigma_r(X_*), \]

as long as the step size obeys \( 0 < \eta \leq 0.2/L \).

Theorem 3 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. \( \|L_t R_t^T - X_*\|_F \leq \epsilon \sigma_r(X_*) \), \texttt{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 \( X_* \). Furthermore, prior theory of vanilla gradient descent [PKCS18] requires \( O(\kappa \kappa_f \log(1/\epsilon)) \) iterations, which is worse than our rate by a factor of \( \kappa \).

### 4 Proof Sketch

In this section, we sketch the proof of the main theorems, highlighting the role of the scaled distance metric (cf. (21)) in these analyses.

#### 4.1 A Warm-Up Analysis: Matrix Factorization

To begin with, it is instrumental to examine matrix factorization with the loss (7), where the update rule of \texttt{ScaledGD} is given in (8). The following theorem, whose proof can be found in Appendix B.2, establishes that as long as \texttt{ScaledGD} is initialized close to the ground truth, then \( \text{dist}(F_t, F_*) \) will contract at a constant linear rate.

**Theorem 4.** Suppose that the step size obeys \( 0 < \eta \leq 2/3 \), and that the initialization \( F_0 \) satisfies \( \text{dist}(F_0, F_*) \leq 0.1 \sigma_r(X_*) \). Then for all \( t \geq 0 \), the iterates of the \texttt{ScaledGD} method in (8) satisfy

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

Comparing to the rate of contraction \( (1 - 1/\kappa) \) of gradient descent for matrix factorization [MLC19, CLC19]. Theorem 4 demonstrates that the pre-conditioners indeed allow better search directions in the local neighborhood of the ground truth, and hence a faster convergence rate.

²In practice, due to the presence of statistical errors due to noise, the minimizer of \( f(\cdot) \) might be only approximately low-rank, to which our analysis can be extended in a straightforward fashion.
4.2 Proof Outline for Matrix Sensing

It can be seen that the update rule (14) of ScaledGD in Algorithm 1 closely mimics (8) when \(A(\cdot)\) satisfies the RIP. Therefore, leveraging the RIP of \(A(\cdot)\) and Theorem 4, 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 4.** Suppose that \(A(\cdot)\) satisfies the 2r-RIP with \(\delta_{2r} \leq 0.02\). If the step size obeys \(0 < \eta \leq 2/3\), and \(\text{dist}(F_t, F_*^*) \leq 0.1\sigma_r(X_*)\), then \(\|L_tR_t^\top - X_*\|_F \leq 1.5\text{dist}(F_t, F_*^*)\). In addition, the \((t+1)\)th iterate \(F_{t+1}\) of the ScaledGD method in (14) 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.** The spectral initialization in (13) for low-rank matrix sensing satisfies

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

Therefore, as long as \(\delta_{2r}\) is small enough, say \(\delta_{2r} \leq 0.02/(\sqrt{r}\kappa)\) 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.

4.3 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. (25)), where we recall from (22) that \(Q_t\) is the optimal alignment matrix between \(F_t\) and \(F_*^*\).

**Lemma 3.** Suppose that \(\alpha \leq 10^{-4}/(\mu r)\). If the step size obeys \(0.1 \leq \eta \leq 2/3\), and the \(t\)th iterate satisfies \(\text{dist}(F_t, F_*^*) \leq 0.02\sigma_r(X_*^*)\), and the incoherence condition

\[
\sqrt{n_1} \left\|(L_tQ_t - L_*)\Sigma_t^{1/2} \right\|_{2,\infty} \lor \sqrt{n_2} \left\|(R_tQ_t^\top - R_*)\Sigma_*^{1/2} \right\|_{2,\infty} \leq \sqrt{\mu}r\sigma_r(X_*^*),
\]

then \(\|L_tR_t^\top - X_*\|_F \leq 1.5\text{dist}(F_t, F_*^*)\). In addition, the \((t+1)\)th iterate \(F_{t+1}\) of the ScaledGD method in (19) of Algorithm 2 satisfies

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

and the incoherence condition

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

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 (18) is close to the ground truth as long as \(\alpha\) is sufficiently small.

**Lemma 4.** The spectral initialization (18) 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 \( \alpha \leq 0.1/(\mu \kappa) \), and that \( \text{dist}(F_0, F_*) \leq 0.02\sigma_r(X_*) \). Then the spectral initialization (18) satisfies the following incoherence condition

\[
\sqrt{n_1} \left\| (L_0Q_0 - L_*)\Sigma_*^{1/2} \right\|_{2,\infty} \vee \sqrt{n_2} \left\| (R_0Q_0^\top - R_*)\Sigma_*^{1/2} \right\|_{2,\infty} \leq \sqrt{\mu \kappa} \sigma_r(X^*),
\]

where \( Q_0 \) is the optimal alignment matrix between \( F_0 \) and \( F_* \).

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

5 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.

We mainly compare ScaledGD with vanilla gradient descent (GD). The update rule of vanilla GD for solving (2) is given as

\[
\begin{align*}
L_{t+1} & = L_t - \eta_{GD} \nabla_L \mathcal{L}(L_t, R_t), \\
R_{t+1} & = R_t - \eta_{GD} \nabla_R \mathcal{L}(L_t, R_t),
\end{align*}
\]

where \( \eta_{GD} > 0 \) stands for the step size for gradient descent. To make a fair comparison, we fix the step size as \( \eta = 0.5 \) for ScaledGD, and set \( \eta_{GD} = \eta/\sigma_1(X_*) \) for vanilla GD. This choice is often recommended by the theory of vanilla GD [TBS+16, YPCC16, MWCC19] and needed for its convergence. 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:

- **Low-rank matrix sensing.** The problem formulation is detailed in Section 2.2. Here, we collect \( m = 5nr \) measurements in the form of \( y_i = \langle A_i, X_* \rangle \), in which the measurement matrices \( A_i \) are generated with i.i.d. Gaussian entries with zero mean and variance \( 1/m \).

- **Robust PCA.** The problem formulation is stated in Section 2.3. We generate the corruption with a sparse matrix \( S_* \in S_\alpha \) with \( \alpha = 0.1 \). More specifically, we generate a matrix with standard Gaussian entries and pass it through \( T_\alpha(\cdot) \) to obtain \( S_* \).

- **Low-rank matrix completion.** We assume random Bernoulli observations, where each entry of \( X_* \) is observed with probability \( p = 0.2 \) independently. The loss function is \( \mathcal{L}(L, R) = \frac{1}{2p} \sum_{(i,j) \in \Omega} (LR^\top - X_*)_{i,j}^2 \), where \( \Omega \) is the index set of observations.

- **Hankel low-rank 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 \( X_* \) is observed with probability \( p = 0.2 \) independently. The loss function is

\[
\mathcal{L}(L, R) = \frac{1}{2p} \left\| \mathcal{H}_{\Omega}(LR^\top - X_*) \right\|_F^2 + \frac{1}{2} \left\| (I - \mathcal{H})(LR^\top) \right\|_F^2,
\]

where \( I(\cdot) \) 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 \( \mathcal{H}_{\Omega}(X) := \sum_{k \in \Omega} (H_k, X)H_k \).
Figure 2: The relative errors of ScaleGD and vanilla GD with respect to 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.

For the first three problems, we generate the ground truth matrix $X_\star \in \mathbb{R}^{n \times n}$ in the following way. We first generate an $n \times r$ matrix with i.i.d. random signs, and take its $r$ left singular vectors as $U_\star$, and similarly for $V_\star$. The singular values are set to be linearly distributed from $1$ to $\kappa$. The ground truth is then defined as $X_\star = U_\star \Sigma_\star V_\star^\top$ 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

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

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

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 theses models, we can see that ScaleGD has a convergence rate independent of $\kappa$, with all curves almost overlay on each other. Under good conditioning $\kappa = 1$, ScaleGD converges at the same rate as vanilla GD; under ill conditioning,
i.e., when $\kappa$ is large, $\text{ScaleGD}$ converges much faster than vanilla GD and leads to significant computational savings.

6 Conclusions

This paper proposes scaled gradient descent ($\text{ScaledGD}$) 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 prove that for low-rank matrix sensing and robust PCA, to reach $\epsilon$-accuracy, $\text{ScaledGD}$ only takes $O(\log(1/\epsilon))$ iterations 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 $\text{ScaledGD}$ under this new metric.

This work opens up many venues for future research. On one end, in this paper we have focused on establishing the fast local convergence rate. It is interesting to study if the theory developed in this paper can be further strengthened in terms of sample complexity and the size of basin of attraction. On the other end, there are many other applications involving the recovery of an ill-conditioned low-rank matrix, such as matrix completion, robust PCA with missing data, quadratic sampling, and so on. It is of interest to establish fast convergence rates of $\text{ScaledGD}$ that are independent of the condition number for these problems as well. As it is evident from our analysis of the general loss case, $\text{ScaledGD}$ 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 addition, it is worthwhile to explore if a similar preconditioning trick can be useful to problems beyond low-rank matrix estimation.

Acknowledgements

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

References

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

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

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

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

[CCD+19] V. Charisopoulos, Y. Chen, D. Davis, M. Diaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.

[CCF+19] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *arXiv preprint arXiv:1902.07698*, 2019.
[CFMY20] Y. Chen, J. Fan, C. Ma, and Y. Yan. Bridging convex and nonconvex optimization in robust 
PCA: Noise, outliers, and missing data. arXiv preprint arXiv:2001.05484, 2020.

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

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

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

[CLL19] J. Chen, D. Liu, and X. Li. Nonconvex rectangular matrix completion via gradient descent 
without $\ell_{2,\infty}$ regularization. arXiv preprint arXiv:1901.06116, 2019.

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

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

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

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

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

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

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

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

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

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

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

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

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

[JK17] P. Jain and P. Kar. Non-convex optimization for machine learning. Foundations and Trends® 
in Machine Learning, 10(3-4):142–336, 2017.
| Reference | Authors | Title and Details |
|-----------|---------|-------------------|
| [JMD10]   | P. Jain, R. Meka, and I. S. Dhillon. | Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010. |
| [JNS13]   | P. Jain, P. Netrapalli, and S. Sanghavi. | Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013. |
| [Kaw16]   | K. Kawaguchi. | Deep learning without poor local minima. In *Advances in neural information processing systems*, pages 586–594, 2016. |
| [KC12]    | A. Kyrillidis and V. Cevher. | Matrix alps: Accelerated low rank and sparse matrix reconstruction. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 185–188. IEEE, 2012. |
| [Laf15]   | J. Lafond. | Low rank matrix completion with exponential family noise. In *Conference on Learning Theory*, pages 1224–1243, 2015. |
| [LLSW19]  | X. Li, S. Ling, T. Strohmer, and K. Wei. | Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis*, 47(3):893–934, 2019. |
| [Maz16]   | M. Mazeika. | The singular value decomposition and low rank approximation. Technical report, University of Chicago, 2016. |
| [MBM18]   | S. Mei, Y. Bai, and A. Montanari. | The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018. |
| [MLC19]   | C. Ma, Y. Li, and Y. Chi. | Beyond Procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. In *Asilomar Conference on Signals, Systems and Computers*, 2019. |
| [MWCC19]  | C. Ma, K. Wang, Y. Chi, and Y. Chen. | Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, pages 1–182, 2019. |
| [NNS*14]  | P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. | Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014. |
| [NP06]    | Y. Nesterov and B. T. Polyak. | Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006. |
| [PKCS17]  | D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi. | Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017. |
| [PKCS18]  | D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi. | Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably. *SIAM Journal on Imaging Sciences*, 11(4):2165–2204, 2018. |
| [RFP10]   | B. Recht, M. Fazel, and P. A. Parrilo. | Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010. |
| [SL16]    | R. Sun and Z.-Q. Luo. | Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016. |
| [SQW15]   | J. Sun, Q. Qu, and J. Wright. | Complete dictionary recovery using nonconvex optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2351–2360, 2015. |
| [SQW18]   | J. Sun, Q. Qu, and J. Wright. | A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018. |
| [TBS*16]  | S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. | Low-rank solutions of linear matrix equations via Procrustes flow. In *International Conference Machine Learning*, pages 964–973, 2016. |
Lemma 6. For a factor matrix \( F \), a sufficient condition guarantees the existence of the minimizer; see the lemma below.

The optimal alignment matrix between \( F \) and \( F^\ast \) involves a minimization problem over an open set (the set of invertible matrices). Hence the minimizer, the infimum, if exists, is called the optimal alignment matrix between \( F \) and \( F^\ast \).

We begin with the investigation of the new distance metric (21), where the matrix \( Q \) use

This section gathers several useful lemmas that will be used in the appendix. Throughout all lemmas, we use \( X^\ast \) to denote the ground truth low-rank matrix, with its compact SVD as \( X^\ast = U^\ast \Sigma^\ast V^\ast^\top \), and the stacked factor matrix is defined as \( F^\ast = \begin{bmatrix} L^\ast \\ R^\ast \end{bmatrix} = \begin{bmatrix} U^\ast \Sigma^\ast_1^{1/2} \\ V^\ast \Sigma^\ast_2^{1/2} \end{bmatrix} \).

A Technical Lemmas

A.1 New Distance Metric

We begin with the investigation of the new distance metric (21), where the matrix \( Q \) that attains the infimum, if exists, is called the optimal alignment matrix between \( F \) and \( F^\ast \); see (22). Notice that (21) 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 \( F^\ast \) is not guaranteed to be attained. Fortunately, a simple sufficient condition guarantees the existence of the minimizer; see the lemma below.

Lemma 6. For a factor matrix \( F := \begin{bmatrix} L \\ R \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r} \), suppose that

\[
\text{dist}(F, F^\ast) = \sqrt{\inf_{Q \in \mathbb{R}^{r \times r}} \| (LQ - L^\ast) \Sigma^\ast_1^{1/2} \|_F^2 + \| (RQ^\top - R^\ast) \Sigma^\ast_2^{1/2} \|_F^2} < \sigma_r(X^\ast),
\]

then the minimizer of the above minimization problem is attained at some \( Q \in \mathbb{R}^{r \times r} \), i.e. the optimal alignment matrix \( Q \) between \( F \) and \( F^\ast \) exists.

Proof. In view of the condition (28) and the definition of infimum, one knows that there must exist a matrix \( Q \in \mathbb{R}^{r \times r} \) such that

\[
\sqrt{\| (LQ - L^\ast) \Sigma^\ast_1^{1/2} \|_F^2 + \| (RQ^\top - R^\ast) \Sigma^\ast_2^{1/2} \|_F^2} \leq \epsilon \sigma_r(X^\ast),
\]

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

\[
\| (LQ - L^\ast) \Sigma^\ast_1^{-1/2} \|_{\text{op}} \vee \| (RQ^\top - R^\ast) \Sigma^\ast_2^{-1/2} \|_{\text{op}} \leq \epsilon.
\]

Invoke Weyl’s inequality \( |\sigma_r(A) - \sigma_r(B)| \leq \| A - B \|_{\text{op}} \), and use that \( \sigma_r(L^\ast \Sigma^\ast_1^{-1/2}) = \sigma_r(U^\ast) = 1 \) to obtain

\[
\sigma_r(LQ \Sigma^\ast_1^{-1/2}) \geq \sigma_r(L^\ast \Sigma^\ast_1^{-1/2}) - \| (LQ - L^\ast) \Sigma^\ast_1^{-1/2} \|_{\text{op}} \geq 1 - \epsilon.
\]

In addition, it is straightforward to verify that

\[
\inf_{Q \in \mathbb{R}^{r \times r}} \| (LQ - L^\ast) \Sigma^\ast_1^{1/2} \|_F^2 + \| (RQ^\top - R^\ast) \Sigma^\ast_2^{1/2} \|_F^2 = \frac{\sigma_r^2(X^\ast)}{\epsilon^2}.
\]
\[
\inf_{H \in \mathbb{R}^{r \times r}} \left\| (LQH - L_*) \Sigma^{1/2} \right\|_F^2 + \left\| (RQ^{-\top} H^{-\top} - R_*) \Sigma^{1/2} \right\|_F^2.
\]

Indeed, if the minimizer of the second optimization problem (cf. (31)) is attained at some \( H \), then \( QH \) must be the minimizer of the first problem (30). Therefore, from now on, we focus on proving that the minimizer of the second problem (31) is attained at some \( H \). In view of (30) and (31), one has

\[
\inf_{H \in \mathbb{R}^{r \times r}} \left\| (LQH - L_*) \Sigma^{1/2} \right\|_F^2 + \left\| (RQ^{-\top} H^{-\top} - R_*) \Sigma^{1/2} \right\|_F^2
\leq \left\| (LQ - L_*) \Sigma^{1/2} \right\|_F^2 + \left\| (RQ^{-\top} - R_*) \Sigma^{1/2} \right\|_F^2,
\]

Clearly, for any \( QH \) to yield a smaller distance than \( Q, H \) must obey

\[
\sqrt{\left\| (LQH - L_*) \Sigma^{1/2} \right\|_F^2 + \left\| (RQ^{-\top} H^{-\top} - R_*) \Sigma^{1/2} \right\|_F^2} \leq \sigma_r(X_*)
\]

It further implies that

\[
\left\| (L\tilde{Q}H - L_*) \Sigma^{-1/2} \right\|_{\text{op}} \vee \left\| (RQ^{-\top} H^{-\top} - R_*) \Sigma^{-1/2} \right\|_{\text{op}} \leq \epsilon.
\]

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

\[
\sigma_1(L\tilde{Q}H \Sigma^{-1/2}) \leq \sigma_1(L_*, \Sigma^{-1/2}) + \left\| (L\tilde{Q}H - L_*) \Sigma^{-1/2} \right\|_{\text{op}} \leq 1 + \epsilon.
\]

Combine (29) and (32), and use the relation \( \sigma_r(A)\sigma_1(B) \leq \sigma_1(AB) \) to obtain

\[
\sigma_r(L\tilde{Q}\Sigma^{-1/2})\sigma_1(\Sigma^{1/2} H \Sigma^{-1/2}) \leq \sigma_1(L\tilde{Q}H \Sigma^{-1/2}) \leq \frac{1 + \epsilon}{1 - \epsilon} \sigma_r(L\tilde{Q}\Sigma^{-1/2}).
\]

As a result, one has \( \sigma_1(\Sigma^{1/2} H \Sigma^{-1/2}) \leq \frac{1 + \epsilon}{1 - \epsilon}. \)

Similarly, one can show that \( \sigma_1(\Sigma^{1/2} H^{-\top} \Sigma^{-1/2}) \leq \frac{1 - \epsilon}{1 + \epsilon} \), equivalently, \( \sigma_r(\Sigma^{1/2} H \Sigma^{-1/2}) \geq \frac{1 + \epsilon}{1 - \epsilon} \). Combining the above two arguments reveals that the minimization problem (31) is equivalent to the constrained problem:

\[
\begin{align*}
\text{minimize} & \quad \left\| (L\tilde{Q}H - L_*) \Sigma^{1/2} \right\|_F^2 + \left\| (RQ^{-\top} H^{-\top} - R_*) \Sigma^{1/2} \right\|_F^2 \\
\text{subject to} & \quad \frac{1 - \epsilon}{1 + \epsilon} \leq \sigma_r(\Sigma^{1/2} H \Sigma^{-1/2}) \leq \frac{1 + \epsilon}{1 - \epsilon}.
\end{align*}
\]

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

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

**Lemma 7.** For any factor matrix \( F := \begin{bmatrix} L \\ R \end{bmatrix} \in \mathbb{R}^{(n_1 + n_2) \times r} \), suppose that the optimal alignment matrix

\[
Q = \text{argmin}_{Q \in \mathbb{R}^{r \times r}} \left\| (LQ - L_*) \Sigma^{1/2} \right\|_F^2 + \left\| (RQ^{-\top} - R_*) \Sigma^{1/2} \right\|_F^2
\]

between \( F \) and \( F_* \) exists, then \( Q \) must obey

\[
(LQ)^\top (LQ - L_*) \Sigma_* = \Sigma_* (RQ^{-\top} - R_*)^\top RQ^{-\top}.
\]
Lemma 8. For any factor matrix $F := \begin{bmatrix} L \\ R \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times r}$, the distance between $F$ and $F_\star$ satisfies

$$\text{dist}(F, F_\star) \leq \sqrt{2} + 1 \|LR^\top - X_\star\|_F.$$ 

Proof. Suppose that $X := LR^\top$ has compact SVD as $X = U \Sigma V^\top$. Without loss of generality, we can assume that $F = \begin{bmatrix} U \Sigma_1^{1/2} \\ V \Sigma_1^{1/2} \end{bmatrix}$, since any factorization of $X$ yields the same distance. Introduce two auxiliary matrices $\bar{F} := \begin{bmatrix} U \Sigma_1^{1/2} \\ -V \Sigma_1^{1/2} \end{bmatrix}$ and $\bar{F}_\star := \begin{bmatrix} U_\star \Sigma_\star_1^{1/2} \\ -V_\star \Sigma_\star_1^{1/2} \end{bmatrix}$. Apply the dilation trick to obtain

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

As a result, the squared Frobenius norm of $X - X_\star$ is given by

$$8\|X - X_\star\|_F^2 = \|FF^\top - \bar{F}_\star \bar{F}^\top + F_\star F^\top - \bar{F}_\star \bar{F}^\top\|_F^2 = \|FF^\top - \bar{F}_\star \bar{F}^\top\|_F^2 + \|F_\star F^\top - \bar{F}_\star \bar{F}^\top\|_F^2 - 2 \text{tr}((FF^\top - \bar{F}_\star \bar{F}^\top)(FF^\top - \bar{F}_\star \bar{F}^\top)) = 2\|FF^\top - \bar{F}_\star \bar{F}^\top\|_F^2 + 2\|F_\star F^\top - \bar{F}_\star \bar{F}^\top\|_F^2 \geq 2\|FF^\top - \bar{F}_\star \bar{F}^\top\|_F^2,$$

where we use the facts that $\|FF^\top - F_\star F^\top\|_F^2 = \|\bar{F} \bar{F}^\top - \bar{F}_\star \bar{F}^\top\|_F^2$ and $F_\star \bar{F} = F_\star \bar{F}_\star = 0$. Let $O := \text{sign}(F^\top F_\star)$ be the optimal orthonormal alignment matrix between $F$ and $F_\star$. Denote $\Delta := FO - F_\star$. Follow the same argument as [TBS+16, Lemma 5.14] and [GJZ17, Lemma 41] to obtain

$$4\|X - X_\star\|_F^2 \geq \|F_\star \Delta^\top + \Delta F^\top_\star + \Delta^\top \|_F^2 \geq \text{tr}(2F^\top_\star F_\star \Delta^\top \Delta + (\Delta^\top \Delta)^2 + 2(\Delta^\top \Delta)^2 + 4F^\top_\star \Delta \Delta^\top \Delta) \geq 4(\sqrt{2} - 1)\|FO - F_\star\|_F^2 \geq 4(\sqrt{2} - 1)\|FO - F_\star\Sigma_\star_1^{1/2}\|_F^2,$$

where the last inequality follows from the facts that $F^\top_\star F_\star = 2\Sigma_\star$ and that $F^\top_\star FO$ is a positive semi-definite matrix. Therefore we obtain

$$\|FO - F_\star\Sigma_\star_1^{1/2}\|_F \leq \sqrt{2} + 1\|X - X_\star\|_F.$$ 

This in conjunction with $\text{dist}(F, F_\star) \leq \|(FO - F_\star)\Sigma_\star_1^{1/2}\|_F$ yields the claimed result. 

\[\square\]
A.2 Matrix Perturbation Bounds

Lemma 9. For any $L \in \mathbb{R}^{n_1 \times r}$, $R \in \mathbb{R}^{n_2 \times r}$, denote $\Delta_L := L - L_*$ and $\Delta_R := R - R_*$. Suppose that $\|\Delta_L \Sigma_*^{-1/2}\|_{op} < 1$ and $\|\Delta_R \Sigma_*^{-1/2}\|_{op} < 1$, then one has

$$\|L(L^T L)^{-1} \Sigma_*^{1/2}\|_{op} \leq \frac{1}{1 - \|\Delta_L \Sigma_*^{-1/2}\|_{op}}, \quad \text{and} \quad \|R(R^T R)^{-1} \Sigma_*^{1/2}\|_{op} \leq \frac{1}{1 - \|\Delta_R \Sigma_*^{-1/2}\|_{op}}.$$

Proof. First, notice that

$$\|L(L^T L)^{-1} \Sigma_*^{1/2}\|_{op} = \frac{1}{\sigma_r(L \Sigma_*^{-1/2})}.$$

In addition, invoke Weyl’s inequality to obtain

$$\sigma_r(L \Sigma_*^{-1/2}) \geq \sigma_r(L, \Sigma_*^{-1/2}) - \|\Delta_L \Sigma_*^{-1/2}\|_{op} = 1 - \|\Delta_L \Sigma_*^{-1/2}\|_{op},$$

where we have used the fact that $U_* = L \Sigma_*^{-1/2}$ is an orthonormal matrix. Combine the preceding two relations to complete the proof. The claim on the factor $R$ follows from a similar argument. \hfill \Box

Lemma 10. For any $L \in \mathbb{R}^{n_1 \times r}$, $R \in \mathbb{R}^{n_2 \times r}$, denote $\Delta_L := L - L_*$ and $\Delta_R := R - R_*$. Suppose that $\|\Delta_L \Sigma_*^{-1/2}\|_{op} < 1$ and $\|\Delta_R \Sigma_*^{-1/2}\|_{op} < 1$, then one has

$$\|L(L^T L)^{-1} \Sigma_*^{1/2} - U_*\|_{op} \leq \frac{\sqrt{2}\|\Delta_L \Sigma_*^{-1/2}\|_{op}}{1 - \|\Delta_L \Sigma_*^{-1/2}\|_{op}}, \quad \text{and} \quad \|R(R^T R)^{-1} \Sigma_*^{1/2} - V_*\|_{op} \leq \frac{\sqrt{2}\|\Delta_R \Sigma_*^{-1/2}\|_{op}}{1 - \|\Delta_R \Sigma_*^{-1/2}\|_{op}}.$$

Proof. Combining $U_* = L \Sigma_*^{-1/2}$ 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

$$L(L^T L)^{-1} \Sigma_*^{1/2} - U_* = -L(L^T L)^{-1} \Delta_L \Sigma_*^{-1/2} U_* + (I_{n_1} - L(L^T L)^{-1} L^T) \Delta_L \Sigma_*^{-1/2}.$$

For any $\bar{v} \in \mathbb{R}^{n_2}$ with $\|\bar{v}\|_2 \leq 1$, the fact that $L(L^T L)^{-1} \Delta_L U_* \bar{v}$ and $(I_{n_1} - L(L^T L)^{-1} L^T) \Delta_L \Sigma_*^{-1/2} \bar{v}$ are orthogonal implies

$$\left\| \left( L(L^T L)^{-1} \Sigma_*^{1/2} - U_* \right) \bar{v} \right\|_2^2 = \left\| L(L^T L)^{-1} \Delta_L U_* \bar{v} \right\|_2^2 + \left\| (I_{n_1} - L(L^T L)^{-1} L^T) \Delta_L \Sigma_*^{-1/2} \bar{v} \right\|_2^2 \leq \left\| \frac{\|\Delta_L \Sigma_*^{-1/2}\|_{op}}{1 - \|\Delta_L \Sigma_*^{-1/2}\|_{op}} \right\|_2^2 \leq 2 \left\| \frac{\|\Delta_L \Sigma_*^{-1/2}\|_{op}}{1 - \|\Delta_L \Sigma_*^{-1/2}\|_{op}} \right\|_2^2,$$

where we have used Lemma 9 and the fact that $\|I_{n_1} - L(L^T L)^{-1} L^T\|_{op} \leq 1$ in the third line. Utilize the variational representation of the operator norm to obtain

$$\left\| L(L^T L)^{-1} \Sigma_*^{1/2} - U_* \right\|_{op} = \max_{\bar{v} \in \mathbb{R}^{n_2} : \|\bar{v}\|_2 \leq 1} \left\| \left( L(L^T L)^{-1} \Sigma_*^{1/2} - U_* \right) \bar{v} \right\|_2 \leq \frac{\sqrt{2}\|\Delta_L \Sigma_*^{-1/2}\|_{op}}{1 - \|\Delta_L \Sigma_*^{-1/2}\|_{op}}.$$

The claim on the factor $R$ follows from a similar argument. \hfill \Box
Lemma 11. For any \( L \in \mathbb{R}^{n_1 \times r}, R \in \mathbb{R}^{n_2 \times r} \), denote \( \Delta_L := L - L_\star \) and \( \Delta_R := R - R_\star \). One has

\[
\|LR^T - X_\star\|_F \leq \|L_\star \Delta_R^\top\|_F + \|\Delta_L R^\top\|_F + \|\Delta_L \Delta_R^\top\|_F \\
\leq \left( 1 + \frac{1}{2}(\|\Delta_L \Sigma^{-1/2}_\star\|_{op} + \|\Delta_R \Sigma^{-1/2}_\star\|_{op}) \right) \left( \|\Delta_L \Sigma^{1/2}_\star\|_F + \|\Delta_R \Sigma^{1/2}_\star\|_F \right).
\]

\textbf{Proof.} In light of the decomposition \( LR^T - X_\star = L \Delta_R^\top + \Delta_L R^\top + \Delta_L \Delta_R^\top \) and the triangle inequality, one obtains

\[
\|LR^T - X_\star\|_F \leq \|L_\star \Delta_R^\top\|_F + \|\Delta_L R^\top\|_F + \|\Delta_L \Delta_R^\top\|_F \\
= \|\Delta_L \Sigma^{1/2}_\star\|_F + \|\Delta_R \Sigma^{1/2}_\star\|_F + \|\Delta_L \Delta_R^\top\|_F,
\]

where we have used the facts that

\[
\|L_\star \Delta_R^\top\|_F = \|U_\star \Sigma^{1/2}_\star \Delta_R^\top\|_F = \|\Delta_R \Sigma^{1/2}_\star\|_F, \quad \text{and} \quad \|\Delta_L R^\top\|_F = \|\Delta_L \Sigma^{1/2}_\star V_\star^\top\|_F = \|\Delta_L \Sigma^{1/2}_\star\|_F.
\]

This together with the simple upper bound

\[
\|\Delta_L \Delta_R^\top\|_F = \frac{1}{2} \|\Delta_L \Sigma^{1/2}_\star (\Delta_R \Sigma^{-1/2}_\star)^\top\|_F + \frac{1}{2} \|\Delta_L \Sigma^{-1/2}_\star (\Delta_R \Sigma^{1/2}_\star)^\top\|_F \\
\leq \frac{1}{2} \|\Delta_L \Sigma^{1/2}_\star\|_{op} \|\Delta_R \Sigma^{-1/2}_\star\|_{op} + \frac{1}{2} \|\Delta_L \Sigma^{-1/2}_\star\|_{op} \|\Delta_R \Sigma^{1/2}_\star\|_{op} \\
\leq \frac{1}{2} \left( \|\Delta_L \Sigma^{1/2}_\star\|_{op} + \|\Delta_R \Sigma^{-1/2}_\star\|_{op} \right) \left( \|\Delta_L \Sigma^{1/2}_\star\|_F + \|\Delta_R \Sigma^{1/2}_\star\|_F \right)
\]

reveals that

\[
\|LR^T - X_\star\|_F \leq \left( 1 + \frac{1}{2}(\|\Delta_L \Sigma^{-1/2}_\star\|_{op} + \|\Delta_R \Sigma^{-1/2}_\star\|_{op}) \right) \left( \|\Delta_L \Sigma^{1/2}_\star\|_F + \|\Delta_R \Sigma^{1/2}_\star\|_F \right).
\]

\[\square\]

Lemma 12. For any two factor matrices \( \begin{bmatrix} L_1 & L_2 \end{bmatrix} \in \mathbb{R}^{(n_1 + n_2) \times r} \), and two invertible matrices \( Q_1, Q_2 \in \mathbb{R}^{r \times r} \), one has

\[
\|\Sigma^{1/2}_1 Q_1^{-1} Q_2 \Sigma^{1/2}_2 - \Sigma_\star\|_{op} \leq \frac{\|R_2(Q_1^{-1} - Q_2^{-1})\|_{op}}{1 - \|(R_2 Q_2^{-1} - R_\star)\Sigma^{-1/2}_\star\|_{op}}, \quad \text{and}
\]

\[
\|\Sigma^{1/2}_1 Q_1^\top Q_2^{-1} \Sigma^{1/2}_2 - \Sigma_\star\|_{op} \leq \frac{\|L_2(Q_1 - Q_2)\|_{op}}{1 - \|(L_2 Q_2 - L_\star)\Sigma^{-1/2}_\star\|_{op}}.
\]

\textbf{Proof.} Insert \( R_2^\top R_2(R_2^\top R_2)^{-1} \), and use the relation \( \|AB\|_{op} \leq \|A\|_{op} \|B\|_{op} \) to obtain

\[
\|\Sigma^{1/2}_1 Q_1^{-1} Q_2 \Sigma^{1/2}_2 - \Sigma_\star\|_{op} = \|\Sigma^{1/2}_1(Q_1^{-1} - Q_2^{-1}) R_2^\top R_2(R_2^\top R_2)^{-1} Q_2 \Sigma^{1/2}_2\|_{op} \\
\leq \frac{\|R_2(Q_1^{-1} - Q_2^{-1})\|_{op}}{1 - \|(R_2 Q_2^{-1} - R_\star)\Sigma^{-1/2}_\star\|_{op}} \|R_2(R_2^\top R_2)^{-1} Q_2 \Sigma^{1/2}_2\|_{op} \\
= \frac{\|R_2(Q_1^{-1} - Q_2^{-1})\|_{op}}{1 - \|(R_2 Q_2^{-1} - R_\star)\Sigma^{-1/2}_\star\|_{op}} \|R_2 Q_2^{-1}((R_2 Q_2^{-1})^\top R_2 Q_2^{-1})^{-1} \Sigma^{1/2}_\star\|_{op} \\
\leq \frac{\|R_2(Q_1^{-1} - Q_2^{-1})\|_{op}}{1 - \|(R_2 Q_2^{-1} - R_\star)\Sigma^{-1/2}_\star\|_{op}}.
\]

where the last inequality follows from Lemma 9.

Similarly, insert \( L_2^\top L_2(L_2^\top L_2)^{-1} \), and use the relation \( \|AB\|_{op} \leq \|A\|_{op} \|B\|_{op} \) to obtain

\[
\|\Sigma^{1/2}_1 Q_1^\top Q_2^{-1} \Sigma^{1/2}_2 - \Sigma_\star\|_{op} = \|\Sigma^{1/2}_1(Q_1^\top - Q_2^\top) L_2^\top L_2(L_2^\top L_2)^{-1} Q_2^{-1} \Sigma^{1/2}_2\|_{op}.
\]
\[
\leq \left\| L_2 (Q_1 - Q_2) \Sigma_i^{1/2} \right\|_{\text{op}} L_2 (L_2^T L_2)^{-1} Q_2^T \Sigma_i^{1/2} \right\|_{\text{op}} \\
= \left\| L_2 (Q_1 - Q_2) \Sigma_i^{1/2} \right\|_{\text{op}} \left\| L_2 Q_2 ((L_2 Q_2)^T L_2 Q_2)^{-1} \Sigma_i^{1/2} \right\|_{\text{op}} \\
\leq \frac{\left\| L_2 (Q_1 - Q_2) \Sigma_i^{1/2} \right\|_{\text{op}}}{1 - \left\| (L_2 Q_2 - L_i) \Sigma_i^{1/2} \right\|_{\text{op}}},
\]
where the last inequality follows from Lemma 9.

\section*{A.3 Partial Frobenius Norm}

We introduce the partial Frobenius norm

\[
\|X\|_{F,r} := \sqrt{\sum_{i=1}^{r} \sigma_i^2(X)}
\]

as the \(\ell_2\) norm of the vector composed of the top-\(r\) singular values of a matrix \(X\). It is straightforward to verify that \(\|\cdot\|_{r,F}\) is a norm; see also \cite{Maz16}. The following lemma provides several equivalent and useful characterizations of this partial Frobenius norm.

\begin{lemma}
For any \(X \in \mathbb{R}^{n_1 \times n_2}\), one has

\begin{align}
\|X\|_{F,r} &= \max_{\tilde{V} \in \mathbb{R}^{n_2 \times r} \colon \tilde{V}^T \tilde{V} = I_r} \|X \tilde{V}\|_F \tag{35a} \\
&= \max_{\tilde{X} \in \mathbb{R}^{n_1 \times n_2} \colon \text{rank}(\tilde{X}) \leq r} \|\tilde{X}\|_F \|X\|_F \tag{35b} \\
&= \max_{\tilde{R} \in \mathbb{R}^{n_2 \times r} \colon \|\tilde{R}\|_{op} \leq 1} \|X \tilde{R}\|_F. \tag{35c}
\end{align}

\end{lemma}

\begin{proof}
The first representation \(35a\) follows immediately from the extremal partial trace identity; see \cite[Proposition 4.4]{Maz16}, by noticing the following relation:

\[
\sum_{i=1}^{r} \sigma_i^2(X) = \max_{\tilde{V} \in \mathbb{R}^{n_2 \times \text{dim}(\tilde{V}) = r}} \text{tr}(X^T X \mid \tilde{V}) = \max_{\tilde{V} \in \mathbb{R}^{\text{rank}(X) \leq r} \colon \|\tilde{V}\|_F \leq 1} \|X \tilde{V}\|_F^2.
\]

Here the partial trace over a vector space \(\tilde{V}\) is defined as

\[
\text{tr}(X^T X \mid \tilde{V}) := \sum_{i=1}^{r} \tilde{v}_i^T X^T X \tilde{v}_i,
\]

where \(\{\tilde{v}_i\}_{1 \leq i \leq r}\) is any orthonormal basis of \(\tilde{V}\). The partial trace is invariant to the choice of orthonormal basis and therefore well-defined.

To prove the second representation \(35b\), for any \(\tilde{X} \in \mathbb{R}^{n_1 \times n_2}\) obeying \(\text{rank}(\tilde{X}) \leq r\) and \(\|\tilde{X}\|_F \leq 1\), denoting \(\tilde{X} = \tilde{U} \tilde{\Sigma} \tilde{V}^T\) as its compact SVD, one has

\[
\langle X, \tilde{X} \rangle = \langle X, \tilde{U} \tilde{\Sigma} \tilde{V}^T \rangle = \langle \tilde{X} \tilde{U} \tilde{\Sigma} \rangle \leq \|X \tilde{V}\|_F \|\tilde{U} \tilde{\Sigma}\|_F \leq \|X\|_{F,r},
\]

where the last inequality follows from \(35a\). In addition, the maximum in \(35b\) is attained at \(\tilde{X} = \mathcal{P}_r(X) \|\mathcal{P}_r(X)\|_F\), where \(\mathcal{P}_r(X)\) denotes the projection of \(X\) on to the set of rank-\(r\) matrices.

To prove the third representation \(35c\), for any \(\tilde{R} \in \mathbb{R}^{n_2 \times r}\) obeying \(\|\tilde{R}\|_{op} \leq 1\), combine the variational representation of the Frobenius norm and \(35b\) to obtain

\[
\|X \tilde{R}\|_F = \max_{L \in \mathbb{R}^{n_1 \times n_2} \colon \|L\|_F \leq 1} \langle X \tilde{R}, \tilde{L} \rangle = \max_{L \in \mathbb{R}^{n_1 \times n_2} \colon \|L\|_F \leq 1} (X, \tilde{L} \tilde{R}^T) \leq \|X\|_{F,r},
\]

where the last inequality follows from \(35b\). In addition, the maximum in \(35c\) is attained at \(\tilde{R} = \tilde{V}\), where \(\tilde{V}\) denotes the top-\(r\) right singular vectors of \(X\). \qed

\[23\]
Remark 3. For self-completeness, we also provide a detailed proof of the first representation (35a). This proof is inductive on $r$. When $r = 1$, we have

$$
\sigma_1(X) = \|Xv_1\|_2 = \max_{\tilde{v} \in \mathbb{R}^{n_2}: \|\tilde{v}\|_2 = 1} \|X\tilde{v}\|_2,
$$

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 $\tilde{V} \in \mathbb{R}^{n_2 \times r}$ such that $\tilde{V}^T\tilde{V} = I_r$, we can first pick $\tilde{v}_2, \ldots, \tilde{v}_r$ as a set of orthonormal vectors in the column space of $\tilde{V}$ that are orthogonal to $v_1$, and then pick $\tilde{v}_1$ via the Gram-Schmidt process so that $\{\tilde{v}_i\}_{i=1}^r$ provides an orthonormal basis of the column space of $\tilde{V}$. Further, by the orthogonality of $\tilde{V}$, there exists an orthonormal matrix $O \in O^{r \times r}$ such that

$$
\tilde{V} = [\tilde{v}_1, \ldots, \tilde{v}_r]O.
$$

Combining this formula with the induction hypothesis yields

$$
\|X\tilde{V}\|_2^2 = \|X[\tilde{v}_1, \ldots, \tilde{v}_r]\|_2^2
= \|X\tilde{v}_1\|_2^2 + \|X[\tilde{v}_2, \ldots, \tilde{v}_r]\|_2^2
\leq \sigma_1^2(X) + \|X - \mathcal{P}_1(X)\|_{F,r-1}^2
= \sum_{i=1}^r \sigma_i^2(X) = \|X\|_{F,r}^2,
$$

where the first line holds since $O$ is orthonormal, the third line holds since $\mathcal{P}_1(X)[\tilde{v}_2, \ldots, \tilde{v}_r] = 0$, the fourth line follows from the induction hypothesis, and the last line follows from the definition (34). In addition, the maximum in (35a) is attained at $\tilde{V} = V$, 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 14.** Fix any $X \in \mathbb{R}^{n_1 \times n_2}$ and recall the definition of $\mathcal{P}_r(X)$ in (5). One has

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

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

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

The proof is finished by observing that $\mathcal{P}_r(X)$ is rank-$r$. \hfill \square

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

**B.1 Proof of Proposition 1**

The gradients of $\mathcal{L}(F)$ in (7) with respect to $L$ and $R$ are given as

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

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

$$
\nabla^2_{L,L} \mathcal{L}(F) = (R^T R) \otimes I_{n_1}, \quad \nabla^2_{R,R} \mathcal{L}(F) = (L^T L) \otimes I_{n_2}.
$$
Viewed in the vectorized form, the ScaledGD method in (3) can be rewritten as
\[
\text{vec}(L_{t+1}) = \text{vec}(L_t) - \eta((R_t^T R_t)^{-1} \otimes I_m) \text{vec}((L_t R_t^T - X_\star) R_t) \\
= \text{vec}(L_t) - \eta(\nabla^2_{LL} L \mathcal{L}(F_t))^{-1} \text{vec}(\nabla L \mathcal{L}(F_t)), \\
\text{vec}(R_{t+1}) = \text{vec}(R_t) - \eta((L_t^T L_t)^{-1} \otimes I_n) \text{vec}((L_t R_t^T - X_\star)^T L_t) \\
= \text{vec}(R_t) - \eta(\nabla^2_{RR} R \mathcal{L}(F_t))^{-1} \text{vec}(\nabla R \mathcal{L}(F_t)).
\]

B.2 Proof of Theorem 4

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 6 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
\[
\text{dist}^2(F_{t+1}, F_\star) \leq \left\| (L_{t+1} Q_t - L_\star) \Sigma^1/2 \right\|_F^2 + \left\| (R_{t+1} Q_t^T - R_\star) \Sigma^1/2 \right\|_F^2,
\]
where we recall that \( Q_t \) is the optimal alignment matrix between \( F_t \) and \( F_\star. \) Utilize the ScaledGD update rule (8) and the decomposition \( LR^T - X_\star = \Delta_L R^T + L, \Delta_R^T \) to obtain
\[
(L_{t+1} Q_t - L_\star) \Sigma^1/2 = (L - \eta L R^T - X_\star) R (R^T R)^{-1} - L_\star \Sigma^1/2 \\
= (\Delta_L - \eta (\Delta_L R^T + L_\star \Delta_R^T) R (R^T R)^{-1}) \Sigma^1/2 \\
= (1 - \eta) \Delta_L \Sigma^1/2 - \eta L_\star \Delta_R^T R (R^T R)^{-1} \Sigma^1/2.
\]
As a result, one can expand the first square in (36) as
\[
\left\| (L_{t+1} Q_t - L_\star) \Sigma^1/2 \right\|_F^2 = (1 - \eta)^2 \text{tr}(\Delta_L \Sigma^* \Delta_L^T) - 2\eta(1 - \eta) \text{tr}(L_\star \Delta_R^T R (R^T R)^{-1} \Sigma_\star \Delta_R^T) \\
+ \eta^2 \left\| L_\star \Delta_R^T R (R^T R)^{-1} \Sigma_\star \right\|_F^2.
\]
The first term \( \text{tr}(\Delta_L \Sigma^* \Delta_L^T) \) is closely related to \( \text{dist}(F_t, F_\star), \) and hence our focus will be on relating \( \mathfrak{M}_1 \) and \( \mathfrak{M}_2 \) to \( \text{dist}(F_t, F_\star). \) We start with the term \( \mathfrak{M}_1. \) Since \( L \) and \( R \) are aligned with \( L_\star \) and \( R_\star, \) Lemma 7 tells us that \( \Sigma_\star \Delta_L^T L = R^T \Delta_R \Sigma_\star. \) This together with \( L_\star = L - \Delta_L \) allows us to rewrite \( \mathfrak{M}_1 \) as
\[
\mathfrak{M}_1 = \text{tr}(R(R^T R)^{-1} \Sigma_\star \Delta_L^T L \Delta_R^T) \\
= \text{tr}(R(R^T R)^{-1} \Sigma_\star \Delta_L^T L \Delta_R^T) - \text{tr}(R(R^T R)^{-1} \Sigma_\star \Delta_L^T \Delta_L \Delta_R^T) \\
= \text{tr}(R(R^T R)^{-1} \Delta_R \Sigma_\star \Delta_R^T) - \text{tr}(R(R^T R)^{-1} \Sigma_\star \Delta_L^T \Delta_L \Delta_R^T).
\]

Moving on to \( \mathfrak{M}_2, \) we can utilize the fact \( L_\star^T L_\star = \Sigma_\star \) and the decomposition \( \Sigma_\star = R^T R - (R^T R - \Sigma_\star) \) to obtain
\[
\mathfrak{M}_2 = \text{tr}(R(R^T R)^{-1} \Sigma_\star (R^T R)^{-1} R^T \Delta_R \Sigma_\star \Delta_R^T) \\
= \text{tr}(R(R^T R)^{-1} R^T \Delta_R \Sigma_\star \Delta_R^T) - \text{tr}(R(R^T R)^{-1} (R^T R - \Sigma_\star) (R^T R)^{-1} R^T \Delta_R \Sigma_\star \Delta_R^T).
\]
Putting $\mathcal{M}_1$ and $\mathcal{M}_2$ back to (37) yields
\[
\left\| (L_{t+1} Q_t - L_*) \Sigma_*^{1/2} \right\|_F^2 = (1 - \eta)^2 \operatorname{tr} \left( \Delta_L \Sigma_* \Delta_L^\top \right) - \eta^2 (2 - 3\eta)^2 \operatorname{tr} \left( R (R^\top R)^{-1} R^\top \Delta_R \Sigma_* \Delta_R^\top \right)
+ 2\eta (1 - \eta) \operatorname{tr} \left( R (R^\top R)^{-1} \Sigma_* \Delta_L \Delta_L^\top \right)
- \eta^2 \operatorname{tr} \left( R (R^\top R)^{-1} (R^\top R - \Sigma_*)(R^\top R)^{-1} R^\top \Delta_R \Sigma_* \Delta_R^\top \right).
\]

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

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

2. To control $\mathfrak{F}_2$, we need certain control on $\|\Delta_L \Sigma_*^{-1/2}\|_{\text{op}}$ and $\|\Delta_R \Sigma_*^{-1/2}\|_{\text{op}}$. The first induction hypothesis
\[
\operatorname{dist}(F_\ast, F_t) = \sqrt{\|\Delta_L \Sigma_*^{-1/2} \Sigma_* \|_F^2 + \|\Delta_R \Sigma_*^{-1/2} \Sigma_* \|_F^2} \leq \epsilon \sigma_r(X_*),
\]

together with the elementary inequality $\|AB\|_F \geq \sigma_r(B) \|A\|_F$ tells that
\[
\sqrt{\|\Delta_L \Sigma_*^{-1/2}\|_F^2 + \|\Delta_R \Sigma_*^{-1/2}\|_F^2} \leq \epsilon \sigma_r(X_*).
\]

In light of the relation $\|A\|_{\text{op}} \leq \|A\|_F$, this further implies
\[
\|\Delta_L \Sigma_*^{-1/2}\|_{\text{op}} \vee \|\Delta_R \Sigma_*^{-1/2}\|_{\text{op}} \leq \epsilon.
\]

Invoke Lemma 9 to see
\[
\left\| R (R^\top R)^{-1} \Sigma_*^{1/2} \right\|_{\text{op}} \leq \frac{1}{1 - \epsilon}.
\]

With these consequences, one can bound $\mathfrak{F}_2$ by
\[
|\mathfrak{F}_2| = \left| \operatorname{tr} \left( \Sigma_*^{-1/2} \Delta_R^\top R (R^\top R)^{-1} \Sigma_* \Delta_L \Sigma_*^{1/2} \right) \right|
\leq \left\| \Sigma_*^{-1/2} \Delta_R^\top R (R^\top R)^{-1} \Sigma_*^{1/2} \right\|_{\text{op}} \left| \operatorname{tr} \left( \Sigma_*^{1/2} \Delta_L \Sigma_*^{1/2} \right) \right|
\leq \|\Delta_R \Sigma_*^{-1/2}\|_{\text{op}} \left\| R (R^\top R)^{-1} \Sigma_*^{1/2} \right\|_{\text{op}} \left| \operatorname{tr} \left( \Delta_L \Sigma_* \Delta_L^\top \right) \right|
\leq \frac{\epsilon}{1 - \epsilon} \left| \operatorname{tr} \left( \Delta_L \Sigma_* \Delta_L^\top \right) \right|.
\]

3. Similarly, one can bound $|\mathfrak{F}_3|$ by
\[
|\mathfrak{F}_3| \leq \left\| R (R^\top R)^{-1} (R^\top R - \Sigma_*)(R^\top R)^{-1} R^\top \right\|_{\text{op}} \left| \operatorname{tr} \left( \Delta_R \Sigma_* \Delta_R^\top \right) \right|
\leq \left\| R (R^\top R)^{-1} \Sigma_*^{1/2} \right\|_{\text{op}}^2 \left\| \Sigma_*^{-1/2} (R^\top R - \Sigma_*) \Sigma_*^{-1/2} \right\|_{\text{op}} \left| \operatorname{tr} \left( \Delta_R \Sigma_* \Delta_R^\top \right) \right|
\leq \frac{1}{(1 - \epsilon)^2} \left\| \Sigma_*^{-1/2} (R^\top R - \Sigma_*) \Sigma_*^{-1/2} \right\|_{\text{op}} \left| \operatorname{tr} \left( \Delta_R \Sigma_* \Delta_R^\top \right) \right|.
\]

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

Take the preceding two bounds together to arrive at
\[
|\mathfrak{F}_3| \leq \frac{2\epsilon + \epsilon^2}{(1 - \epsilon)^2} \left| \operatorname{tr} \left( \Delta_R \Sigma_* \Delta_R^\top \right) \right|.
\]
Combining the bounds for $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$, one has

$$
\left\| (L_{t+1}Q_t - L_*) \Sigma_{*}^{1/2} \right\|_F^2 = \left\| (1 - \eta) \Delta_L \Sigma_{*}^{1/2} - \eta L_* \Delta_R^\top R (R^\top R)^{-1} \Sigma_{*}^{1/2} \right\|_F^2 \\
\leq \left( 1 - \eta \right)^2 + \frac{2\epsilon}{1 - \epsilon} \eta (1 - \eta) \text{tr} (\Delta_L \Sigma, \Delta_L^\top) + \frac{2\epsilon + \epsilon^2}{(1 - \epsilon)^2} \eta^2 \text{tr} (\Delta_R \Sigma, \Delta_R^\top).
$$

A similarly bound holds for the second square $\left\| (R_{t+1}Q_t - R_*) \Sigma_{*}^{1/2} \right\|_F^2$ in (36). Therefore we obtain

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

where we identify

$$
\text{dist}^2(F_t, F_*) = \text{tr}(\Delta_L \Sigma, \Delta_L^\top) + \text{tr}(\Delta_R \Sigma, \Delta_R^\top)
$$

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

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

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 6, 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 11:

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

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 (40). The proof is now completed.

### C Proof for Low-Rank Matrix Sensing

In this section, we collect the proofs for Section 4.2. We start by recording a useful lemma.

**Lemma 15** ([CP11]). Suppose that $A(\cdot)$ satisfies the $2r$-RIP with a constant $\delta_{2r}$. Then, for all matrices $X_1, X_2 \in \mathbb{R}^{n_1 \times n_2}$ 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

$$
|\text{tr} \left( (A^* A - I)(X_1X_2^\top) \right) | \leq \delta_{2r} \|X_1\|_F \|X_2\|_F.
$$

(42)
as a simple corollary, one has that for any matrix $R \in \mathbb{R}^{n_2 \times r}$:

$$
\|(A^*A - \mathcal{I})(X_1)R\|_F \leq \delta_2 \|X_1\|_F \|R\|_{op}.
$$

(43)

This is due to the fact that

$$
\|(A^*A - \mathcal{I})(X_1)R\|_F = \max_{L: \|L\|_F \leq 1} \text{tr} \left( (A^*A - \mathcal{I})(X_1)RL^\top \right)
\leq \max_{L: \|L\|_F \leq 1} \delta_2 \|X_1\|_F \|RL^\top\|_F
\leq \max_{L: \|L\|_F \leq 1} \delta_2 \|X_1\|_F \|R\|_{op} \|L^\top\|_F
= \delta_2 \|X_1\|_F \|R\|_{op}.
$$

Here, the first line follows from the definition of $\| \cdot \|_F$, the second line follows from (42), 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 $\text{dist}(F_t, F_s) \leq 0.1 \sigma_r(X_s)$ and Lemma 6, one knows that $Q_t$, the optimal alignment matrix between $F_t$ and $F_s$ exists. Therefore, for notational convenience, denote $\tilde{L} := L_t Q_t$, $R := R_t Q_t^\top$, $\Delta_L := L - L_*$, $\Delta_R := R - R_*$, and $\epsilon := 0.1$. Similar to the derivation in (38), we have

$$
\|\Delta_L \Sigma_*^{-1/2}\|_{op} \vee \|\Delta_R \Sigma_*^{-1/2}\|_{op} \leq \epsilon.
$$

(44)

The conclusion $\|L_t R_t^\top - X_*\|_F \leq 1.5 \text{dist}(F_t, F_s)$ is a simple consequence of Lemma 11; see (41) 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 $\text{dist}(F_{t+1}, F_s)$ that

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

(45)

Apply the update rule (14) and the decomposition $L R^\top - X_* = \Delta_L R^\top + L_\Delta R^\top$ to obtain

$$
(L_{t+1} Q_t - L_*) \Sigma_*^{1/2} = (L - \eta A^* A (L R^\top - X_*) R (R^\top R)^{-1} - L_*) \Sigma_*^{1/2}
= (\Delta_L - \eta (L R^\top - X_*) R (R^\top R)^{-1} - \eta (A^* A - \mathcal{I}) (L R^\top - X_*) R (R^\top R)^{-1}) \Sigma_*^{1/2}
= (1 - \eta) \Delta_L \Sigma_*^{1/2} - \eta L_\Delta R^\top R (R^\top R)^{-1} \Sigma_*^{1/2} - \eta (A^* A - \mathcal{I}) (L R^\top - X_*) R (R^\top R)^{-1} \Sigma_*^{1/2}.
$$

This allows us to expand the first square in (45) as

$$
\left\| (L_{t+1} Q_t - L_*) \Sigma_*^{1/2} \right\|_F^2 = \left\| (1 - \eta) \Delta_L \Sigma_*^{1/2} - \eta L_\Delta R^\top R (R^\top R)^{-1} \Sigma_*^{1/2} \right\|_F^2
- 2 \eta (1 - \eta) \text{tr} \left( (A^* A - \mathcal{I}) (L R^\top - X_*) R (R^\top R)^{-1} \Sigma_* \Delta_L^\top \right)
+ 2 \eta^2 \text{tr} \left( (A^* A - \mathcal{I}) (L R^\top - X_*) R (R^\top R)^{-1} \Sigma_* (R^\top R)^{-1} R^\top \Delta_R L^\top \right)
+ \eta^2 \left\| (A^* A - \mathcal{I}) (L R^\top - X_*) R (R^\top R)^{-1} \Sigma_*^{1/2} \right\|_F^2.
$$

In what follows, we shall control the four terms separately, of which $\mathcal{G}_1$ is the main term, and $\mathcal{G}_2, \mathcal{G}_3$ and $\mathcal{G}_4$ are perturbation terms.
1. Notice that the main term $\mathcal{G}_1$ has already been controlled in (39) under the condition (44). It obeys

$$\mathcal{G}_1 \leq \left( (1 - \eta)^2 + \frac{2\epsilon}{1 - \epsilon} \eta (1 - \eta) \right) \|\Delta_L \Sigma_\star^{1/2}\|^2 + \frac{2\epsilon}{(1 - \epsilon)^2} \eta^2 \|\Delta_R \Sigma_\star^{1/2}\|^2.$$ 

2. For the second term $\mathcal{G}_2$, decompose $L R^\top - X_\star = L_\star \Delta_R^\top + \Delta_L R^\top + \Delta_L \Delta_R^\top$ and apply the triangle inequality to obtain

$$|\mathcal{G}_2| = \left| \text{tr} \left( (A^* A - I) (L_\star \Delta_R^\top + \Delta_L R^\top + \Delta_L \Delta_R^\top) R (R^\top R)^{-1} \Sigma_\star \Delta_L^\top \right) \right|$$

$$\leq \left| \text{tr} \left( (A^* A - I) (L_\star \Delta_R^\top) R (R^\top R)^{-1} \Sigma_\star \Delta_L^\top \right) \right|$$

$$+ \left| \text{tr} \left( (A^* A - I) (\Delta_L R^\top) R (R^\top R)^{-1} \Sigma_\star \Delta_L^\top \right) \right|$$

$$+ \left| \text{tr} \left( (A^* A - I) (\Delta_L \Delta_R^\top) R (R^\top R)^{-1} \Sigma_\star \Delta_L^\top \right) \right|.$$ 

Invoke Lemma 15 to further obtain

$$|\mathcal{G}_2| \leq \delta_{2r} \left( \|L_\star \Delta_R^\top\|_F + \|\Delta_L R^\top\|_F + \|\Delta_L \Delta_R^\top\|_F \right) \left\| R (R^\top R)^{-1} \Sigma_\star \Delta_L^\top \right\|_F$$

$$\leq \delta_{2r} \left( \|L_\star \Delta_R^\top\|_F + \|\Delta_L R^\top\|_F + \|\Delta_L \Delta_R^\top\|_F \right) \left\| R (R^\top R)^{-1} \Sigma_\star^{1/2} \right\|_\text{op} \left\| \Delta_L \Sigma_\star^{1/2}\right\|_F,$$

where the second line follows from the relation $\|A B\|_F \leq \|A\|_\text{op} \|B\|_F$. Take the condition (44) and Lemmas 9 and 11 together to obtain

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

$$\|L_\star \Delta_R^\top\|_F + \|\Delta_L R^\top\|_F + \|\Delta_L \Delta_R^\top\|_F \leq (1 + \frac{\epsilon}{2}) \left( \|\Delta_L \Sigma_\star^{1/2}\|_F + \|\Delta_R \Sigma_\star^{1/2}\|_F \right).$$

These consequences further imply that

$$|\mathcal{G}_2| \leq \frac{\delta_{2r}(2 + \epsilon)}{2(1 - \epsilon)} \left( \|\Delta_L \Sigma_\star^{1/2}\|_F + \|\Delta_R \Sigma_\star^{1/2}\|_F \right) \|\Delta_L \Sigma_\star^{1/2}\|_F$$

$$= \frac{\delta_{2r}(2 + \epsilon)}{2(1 - \epsilon)} \left( \|\Delta_L \Sigma_\star^{1/2}\|_F^2 + \|\Delta_L \Sigma_\star^{1/2}\|_F \|\Delta_R \Sigma_\star^{1/2}\|_F \right).$$

For the term $\|\Delta_L \Sigma_\star^{1/2}\|_F \|\Delta_R \Sigma_\star^{1/2}\|_F$, we can apply the elementary inequality $2ab \leq a^2 + b^2$ to see

$$\|\Delta_L \Sigma_\star^{1/2}\|_F \|\Delta_R \Sigma_\star^{1/2}\|_F \leq \frac{1}{2} \|\Delta_L \Sigma_\star^{1/2}\|_F^2 + \frac{1}{2} \|\Delta_R \Sigma_\star^{1/2}\|_F^2.$$ 

The preceding two bounds taken collectively yield

$$|\mathcal{G}_2| \leq \frac{\delta_{2r}(2 + \epsilon)}{2(1 - \epsilon)} \left( \frac{3}{2} \|\Delta_L \Sigma_\star^{1/2}\|_F^2 + \frac{1}{2} \|\Delta_L \Sigma_\star^{1/2}\|_F^2 \right).$$

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

$$|\mathcal{G}_3| \leq \delta_{2r} \left( \|L_\star \Delta_R^\top\|_F + \|\Delta_L R^\top\|_F + \|\Delta_L \Delta_R^\top\|_F \right) \left\| R (R^\top R)^{-1} \Sigma_\star (R^\top R)^{-1} R^\top \Delta_R L_\star^\top \right\|_F$$

$$\leq \delta_{2r} \left( \|L_\star \Delta_R^\top\|_F + \|\Delta_L R^\top\|_F + \|\Delta_L \Delta_R^\top\|_F \right) \left\| R (R^\top R)^{-1} \Sigma_\star^{1/2} \right\|_\text{op} \left\| \Delta_R L_\star^\top \right\|_F$$

$$\leq \frac{\delta_{2r}(2 + \epsilon)}{2(1 - \epsilon)^2} \left( \|\Delta_L \Sigma_\star^{1/2}\|_F + \|\Delta_R \Sigma_\star^{1/2}\|_F \right) \|\Delta_R \Sigma_\star^{1/2}\|_F$$

$$\leq \frac{\delta_{2r}(2 + \epsilon)}{2(1 - \epsilon)^2} \left( \frac{3}{2} \|\Delta_L \Sigma_\star^{1/2}\|_F^2 + \frac{3}{2} \|\Delta_R \Sigma_\star^{1/2}\|_F^2 \right).$$
4. We are then left with the last term $\mathcal{S}_4$, for which we have
\[
\sqrt{\mathcal{S}_4} = \left\| (A^*A - I)(LR^T - X_*)R(R^TR)^{-1}\Sigma_\epsilon^{1/2} \right\|_F
\leq \left\| (A^*A - I)(L, \Delta R^T)R(R^TR)^{-1}\Sigma_\epsilon^{1/2} \right\|_F
+ \left\| (A^*A - I)(\Delta L \Delta R^T)R(R^TR)^{-1}\Sigma_\epsilon^{1/2} \right\|_F
+ \left\| (A^*A - I)(\Delta L \Delta R^T)R(R^TR)^{-1}\Sigma_\epsilon^{1/2} \right\|_F,
\]
where once again we use the decomposition $LR^T - X_* = \Delta L \Delta R^T + \Delta L \Delta R^T$. Use (43) and the 2r-RIP to see that
\[
\sqrt{\mathcal{S}_4} \leq \delta_{2r}\left( \|L, \Delta R^T\|_F + \|\Delta L \Delta R^T\|_F \right) \left\| R(R^TR)^{-1}\Sigma_\epsilon^{1/2} \right\|_{op}.
\]
Repeating the same argument in bounding $\mathcal{S}_2$ yields
\[
\sqrt{\mathcal{S}_4} \leq \frac{\delta_{2r}(2 + \epsilon)}{2(1 - \epsilon)} \left( \|\Delta L \Sigma_\epsilon^{1/2}\|_F + \|\Delta R \Sigma_\epsilon^{1/2}\|_F \right).
\]
We can then take the squares of both sides and use $(a + b)^2 \leq 2a^2 + 2b^2$ to reach
\[
\mathcal{S}_4 \leq \frac{\delta_{2r}^2(2 + \epsilon)^2}{2(1 - \epsilon)^2} \left( \|\Delta L \Sigma_\epsilon^{1/2}\|_F^2 + \|\Delta R \Sigma_\epsilon^{1/2}\|_F^2 \right).
\]
Taking the bounds for $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4$ collectively yields
\[
\left\| (L_{t+1}Q_t - L_*)\Sigma_\epsilon^{1/2} \right\|_F^2 \leq \left( (1 - \eta)^2 + \frac{2\epsilon}{1 - \epsilon}\eta(1 - \eta) \right) \|\Delta L \Sigma_\epsilon^{1/2}\|_F^2 + \frac{2\epsilon + \epsilon^2}{1 - \epsilon}\eta^2 \|\Delta R \Sigma_\epsilon^{1/2}\|_F^2
+ \frac{\delta_{2r}(2 + \epsilon)}{1 - \epsilon}\eta(1 - \eta) \left( \frac{3}{2}\|\Delta L \Sigma_\epsilon^{1/2}\|_F^2 + \frac{1}{2}\|\Delta R \Sigma_\epsilon^{1/2}\|_F^2 \right)
+ \frac{\delta_{2r}(2 + \epsilon)\epsilon^2}{(1 - \epsilon)^2}\eta^2 \left( \frac{1}{2}\|\Delta L \Sigma_\epsilon^{1/2}\|_F^2 + \frac{3}{2}\|\Delta R \Sigma_\epsilon^{1/2}\|_F^2 \right)
+ \frac{\delta_{2r}^2(2 + \epsilon)^2}{2(1 - \epsilon)^2}\eta^2 \left( \|\Delta L \Sigma_\epsilon^{1/2}\|_F^2 + \|\Delta R \Sigma_\epsilon^{1/2}\|_F^2 \right).
\]
Similarly, we can expand the second square in (45) and obtain a similar bound. Combine both to obtain
\[
\left\| (L_{t+1}Q_t - L_*)\Sigma_\epsilon^{1/2} \right\|_F^2 + \left\| (R_{t+1}Q_t^{-\top} - R_*)\Sigma_\epsilon^{1/2} \right\|_F^2 \leq \rho^2(\eta; \epsilon, \delta_{2r}) \text{ dist}^2(F_t, F_*),
\]
where the contraction rate is given by
\[
\rho^2(\eta; \epsilon, \delta_{2r}) := (1 - \eta)^2 + \frac{2\epsilon + \delta_{2r}(4 + 2\epsilon)}{1 - \epsilon}\eta(1 - \eta) + \frac{2\epsilon + \epsilon^2 + \delta_{2r}(4 + 2\epsilon) + \delta_{2r}^2(2 + \epsilon)^2}{(1 - \epsilon)^2}\eta^2.
\]
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
\[
\text{dist}(F_{t+1}, F_*) \leq \sqrt{\left\| (L_{t+1}Q_t - L_*)\Sigma_\epsilon^{1/2} \right\|_F^2 + \left\| (R_{t+1}Q_t^{-\top} - R_*)\Sigma_\epsilon^{1/2} \right\|_F^2}
\leq (1 - 0.6\eta) \text{ dist}(F_t, F_*).
\]

C.2 Proof of Lemma 2

With the knowledge of $\| \cdot \|_{F,r}$, we are ready to establish the claimed result. Invoke Lemma 8 to relate $\text{dist}(F_0, F_*)$ to $\| L_0 R_0^\top - X_* \|_F$, and use that $L_0 R_0^\top - X_*$ has rank at most $2r$ to obtain
\[
\text{dist}(F_0, F_*) \leq \sqrt{2 + 1} \| L_0 R_0^\top - X_* \|_F \leq \sqrt{2(\sqrt{2} + 1)} \| L_0 R_0^\top - X_* \|_{F,r}.
\]
Note that $L_0 R_0^T$ is the best rank-$r$ approximation of $\mathcal{A}^* A(X_*)$, and apply the triangle inequality combined with Lemma 14 to obtain
\[
\|L_0 R_0^T - X_*\|_{F,r} \leq \|\mathcal{A}^* A(X_*) - L_0 R_0^T\|_{F,r} + \|\mathcal{A}^* A(X_*) - X_*\|_{F,r} \\
\leq 2 \|(A^* A - T)(X_*)\|_{F,r} \leq 2\delta_{2r}\|X_*\|_F.
\]
Here, the last inequality follows from combining Lemma 13 and (43) as
\[
\|(A^* A - T)(X_*)\|_{F,r} = \max_{\tilde{R} \in \mathbb{R}^{n_2 \times r} : \|\tilde{R}\|_\infty = 1} \left\|\left((A^* A - T)(X_*)\tilde{R}\right)\right\|_F \leq \delta_{2r}\|X_*\|_F.
\]
As a result, one has
\[
\text{dist}(F_0, F_*) \leq 2\sqrt{2(\sqrt{2} + 1)}\delta_{2r}\|X_*\|_F \leq 5\delta_{2r}\sqrt{\kappa_r(X_*)}.
\]

D Proof for Robust PCA

D.1 Preliminaries

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

Lemma 16. For $S_* \in S_\alpha$ and $S = \mathcal{T}_{2a}[X_* + S_* - LR^T]$, one has
\[
\|S - S_*\|_\infty \leq 2\|LR^T - X_*\|_\infty. \tag{46}
\]
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
\[
\|S - S_* , M\| \leq \sqrt{3\nu} \left(\|L - L_*\|_{2,\infty} \Sigma_*^{1/2} \|_F + \|R - R_*\|_{2,\infty} \Sigma_*^{1/2} \|_F\right) \|M\|_F \\
+ 2\sqrt{\alpha} \left(\sqrt{n_1} \|L_M\|_{2,\infty} \|R_M\|_F \wedge \sqrt{n_2} \|L_M\|_F \|R_M\|_{2,\infty}\right) \|LR^T - X_*\|_F, \tag{47}
\]
where $\nu$ obeys
\[
\nu \geq \frac{\sqrt{n_1}}{2} \left(\|L\Sigma_*^{-1/2} \|_{2,\infty} + \|L_*\Sigma_*^{-1/2} \|_{2,\infty}\right) \vee \sqrt{\frac{n_2}{2}} \left(\|R\Sigma_*^{-1/2} \|_{2,\infty} + \|R_*\Sigma_*^{-1/2} \|_{2,\infty}\right).
\]

Proof. Denote $\Delta_L := L - L_*$, $\Delta_R := R - R_*$, and $\Delta_X := LR^T - X_*$. Let $\Omega, \Omega_*$ be the support of $S$ and $S_*$, respectively. As a result, $S - S_*$ is supported on $\Omega \cup \Omega_*$. We start with proving the first claim, i.e. (46). For $(i, j) \in \Omega$, by the definition of $\mathcal{T}_{2a}[\cdot]$, we have $(S - S_*)_{i,j} = (-\Delta_X)_{i,j}$. For $(i, j) \in \Omega_*, \Omega$, one necessarily has $S_{i,j} = 0$ and therefore $(S - S_*)_{i,j} = (-S_*)_{i,j}$. Again by the definition of the operator $\mathcal{T}_{2a}[\cdot]$, we know $|S_* - \Delta_X|_{i,j}$ is either smaller than $|S_* - \Delta_X|_{i,(2an_2)}$ or $|S_* - \Delta_X|_{(2an_1),j}$. Consequently, one has $|S_* - \Delta_X|_{i,j} \leq |\Delta_X|_{i,(2an_2)} \vee |\Delta_X|_{(an_1),j}$. Combining the two cases above, we conclude that
\[
|S - S_*|_{i,j} \leq \begin{cases} |\Delta_X|_{i,j}, & (i, j) \in \Omega \\
|\Delta_X|_{i,j} + (|\Delta_X|_{i,(2an_2)} \vee |\Delta_X|_{(an_1),j}), & (i, j) \in \Omega_* \backslash \Omega. \end{cases}
\tag{48}
\]
This immediately implies the $\ell_\infty$ norm bound (46).

Next, we prove the second claim (47). Recall that $S - S_*$ is supported on $\Omega \cup \Omega_*$. We then have
\[
\|(S - S_*, M)\| = \sum_{(i,j) \in \Omega} |S - S_*|_{i,j}|M|_{i,j} + \sum_{(i,j) \in \Omega_* \backslash \Omega} |S - S_*|_{i,j}|M|_{i,j} \\
\leq \sum_{(i,j) \in \Omega, \Omega} |\Delta_X|_{i,j}|M|_{i,j} + \sum_{(i,j) \in \Omega_* \backslash \Omega} (|\Delta_X|_{i,(2an_2)} + |\Delta_X|_{(an_1),j}) |M|_{i,j}.
\]
where the second line follows from (48). 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
\[
|\langle S - S_*, M \rangle| \leq \sum_{(i,j) \in \Omega_1, \Omega} |\Delta X|_{i,j} |M|_{i,j} + \frac{1}{2\beta} \sum_{(i,j) \in \Omega_2, \Omega} \left( |\Delta X|_{i,(\alpha_1n_2)j}^2 + |\Delta X|_{(\alpha_1n_2),j}^2 \right) + \beta \sum_{(i,j) \in \Omega, \Omega} |M|_{i,j}^2.
\]

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. For the first term $\mathfrak{A}_1$, one has
\[
\mathfrak{A}_1 \leq \sqrt{3\alpha \nu} \left( \|\Delta_L \Sigma^1/2\|_F + \|\Delta_R \Sigma^1/2\|_F \right) \|M\|_F.
\]

Claim 2. For the second term $\mathfrak{A}_2$, one has
\[
\mathfrak{A}_2 \leq 2\|\Delta X\|_F^2.
\]

Claim 3. For the third term $\mathfrak{A}_3$, one has
\[
\mathfrak{A}_3 \leq \alpha \left( n_1 \|L_M\|_{2,\infty}^2 \|R_M\|_F^2 \wedge n_2 \|L_M\|_{2,\infty}^2 \|R_M\|_F^2 \right).
\]

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

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

This finishes the proof.

Proof of Claim 1. Use the decomposition $\Delta X = \Delta_L R^T + L \Delta_R^T = L \Delta_R + \Delta_L R^T$ to obtain
\[
|\Delta X|_{i,j} \leq \|(\Delta_L \Sigma^1/2)_{i,.} \|_2 \|R \Sigma^{-1/2}_{i,.} \|_{2,\infty} + \|L \Sigma^{-1/2}_{i,.} \|_{2,\infty} \|(\Delta_R \Sigma^1/2)_{j,.} \|_2, \text{ and }
\]
\[
|\Delta X|_{i,j} \leq \|L \Sigma^{-1/2}_{i,.} \|_{2,\infty} \|(\Delta_R \Sigma^1/2)_{j,.} \|_2 + \|L \Sigma^{-1/2}_{i,.} \|_{2,\infty} \|(\Delta_L \Sigma^1/2)_{i,.} \|_2.
\]

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
\[
\mathfrak{A}_1 \leq \sum_{(i,j) \in \Omega, \Omega} \frac{\nu}{\sqrt{n_2}} \|(\Delta_L \Sigma^1/2)_{i,.} \|_2 \|M|_{i,j} + \sum_{(i,j) \in \Omega, \Omega} \frac{\nu}{\sqrt{n_1}} \|(\Delta_R \Sigma^1/2)_{j,.} \|_2 \|M|_{i,j}
\]
\[
\leq \left( \sum_{(i,j) \in \Omega, \Omega} \|(\Delta_L \Sigma^1/2)_{i,.} \|_2 / n_2 + \sum_{(i,j) \in \Omega, \Omega} \|(\Delta_R \Sigma^1/2)_{j,.} \|_2 / n_1 \right) \nu \|M\|_F.
\]
Regarding the first term, one has
\[
\sum_{(i,j) \in \Omega \cup \Omega_*} \| (\Delta_L \Sigma_1^{1/2} )_{i,j} \|_2^2 = \sum_{i: 1 \leq i \leq n_1} \sum_{j: (i,j) \in \Omega \cup \Omega_*} \| (\Delta_L \Sigma_1^{1/2} )_{i,j} \|_2^2 \\
\leq 3\alpha n_2 \sum_{i: 1 \leq i \leq n_1} \| (\Delta_L \Sigma_1^{1/2} )_{i} \|_2^2 \\
= 3\alpha n_2 \| \Delta_L \Sigma_1^{1/2} \|_F^2.
\]
Here, we utilize the fact that \( \Omega \cup \Omega_* \) has at most \( 3\alpha \)-fraction of non-zero entries per row. Similarly, we can show that
\[
\sum_{(i,j) \in \Omega \cup \Omega_*} \| (\Delta_R \Sigma_1^{1/2} )_{j,i} \|_2^2 \leq 3\alpha n_1 \| \Delta_R \Sigma_1^{1/2} \|_F^2.
\]
In all, we arrive at
\[
A_1 \leq \sqrt{3\alpha \nu} \left( \| \Delta_L \Sigma_1^{1/2} \|_F + \| \Delta_R \Sigma_1^{1/2} \|_F \right) \| M \|_F,
\]
which is the desired claim. \( \square \)

**Proof of Claim 2.** Recall that \((\Delta_X)_{i,(\alpha n_2)}\) denotes the \((\alpha n_2)\)-largest entry in \((\Delta_X)_{i,.}\). One necessarily has \( \alpha n_2 \| \Delta_X \|_{i,(\alpha n_2)} \leq \| (\Delta_X)_{i,.} \|_2^2 \).

As a result, we obtain
\[
\sum_{(i,j) \in \Omega \setminus \Omega} \| \Delta X \|_{i,(\alpha n_2)}^2 \leq \sum_{(i,j) \in \Omega_*} \| \Delta X \|_{i,(\alpha n_2)}^2 \leq \sum_{(i,j) \in \Omega_*} \frac{\| (\Delta_X)_{i,.} \|_2^2}{\alpha n_2} \\
\leq \sum_{i: 1 \leq i \leq n_1} \sum_{j: (i,j) \in \Omega_*} \frac{\| (\Delta_X)_{i,.} \|_2^2}{\alpha n_2} \\
\leq \sum_{i: 1 \leq i \leq n_1} \| (\Delta_X)_{i,.} \|_2^2 = \| \Delta X \|_F^2,
\]
where we utilize the fact that \( S_* \) contains at most \( \alpha n_2 \) nonzero entries in each row. Similarly one can show that
\[
\sum_{(i,j) \in \Omega \setminus \Omega} \| \Delta X \|_{(\alpha n_1),i}^2 \leq \| \Delta X \|_F^2.
\]
Combining the above two bounds with the definition of \( A_2 \) completes the proof. \( \square \)

**Proof of Claim 3.** By definition, \( M = L_M R_M^T \), and hence one has
\[
A_3 = \sum_{(i,j) \in \Omega \setminus \Omega} | (L_M)_{i,.} (R_M)_{j,.}^T |^2 \leq \sum_{(i,j) \in \Omega_*} | (L_M)_{i,.} (R_M)_{j,.}^T |^2.
\]
We can further upper bound \( A_3 \) as
\[
A_3 \leq \sum_{(i,j) \in \Omega_*} \| (L_M)_{i,.} \|_2^2 \| R_M \|_2^2 \| L_M \|_2^2 \| R_M \|_2^2 \| L_M \|_2^2 \| R_M \|_2^2, \\
\leq \sum_{i: 1 \leq i \leq n_1} \alpha n_2 \| (L_M)_{i,.} \|_2^2 \| R_M \|_2^2 \| L_M \|_2^2 \| R_M \|_2^2 = \alpha n_2 \| L_M \|_2^2 \| R_M \|_2^2,
\]
where we have used the fact that \( \Omega_* \) has at most \( \alpha \)-fraction of non-zero entries per row. Similarly, one can obtain
\[
A_3 \leq \alpha n_1 \| L_M \|_2^2 \| R_M \|_2^2,
\]
which completes the proof. \( \square \)
D.2 Proof of Lemma 3

D.2.1 Preliminaries

We begin with introducing several useful notation and facts. In view of the condition \( \text{dist}(F_t, F_\star) \leq 0.02\sigma_r(X_\star) \) and Lemma 6, one knows that \( Q_t \), the optimal alignment matrix between \( F_t \) and \( F_\star \) exists. Therefore, for notational convenience, denote \( L := L_tQ_t, R := R_tQ_t^{-1} \), \( \Delta_L := L - L_\star \), \( \Delta_R := R - R_\star \), \( S := S_t = T_{2\alpha}[X_\star + S_\star - LR^\top] \), and \( \epsilon := 0.02 \). The assumption \( \text{dist}(F_t, F_\star) \leq \epsilon \sigma_r(X_\star) \) together with the fact that

\[
\sqrt{\| \Delta_L \Sigma_\star^{-1/2} \|_F^2 + \| \Delta_R \Sigma_\star^{-1/2} \|_F^2} \sigma_r(X_\star) \leq \text{dist}(F_t, F_\star)
\]

implies

\[
\| \Delta_L \Sigma_\star^{-1/2} \|_{\infty} \vee \| \Delta_R \Sigma_\star^{-1/2} \|_{\infty} \leq \epsilon. \quad (49)
\]

Moreover, the incoherence condition

\[
\sqrt{n_1} \| \Delta_L \Sigma_\star^{-1/2} \|_{2, \infty} \vee \sqrt{n_2} \| \Delta_R \Sigma_\star^{-1/2} \|_{2, \infty} \leq \sqrt{\mu_r} \sigma_r(X_\star)
\]

implies

\[
\sqrt{n_1} \| \Delta_L \Sigma_\star^{-1/2} \|_{2, \infty} \vee \sqrt{n_2} \| \Delta_R \Sigma_\star^{-1/2} \|_{2, \infty} \leq \sqrt{\mu_r}, \quad (50)
\]

which combined with the triangle inequality further implies

\[
\sqrt{n_1} \| L \Sigma_\star^{-1/2} \|_{2, \infty} \vee \sqrt{n_2} \| R \Sigma_\star^{-1/2} \|_{2, \infty} \leq 2\sqrt{\mu_r} \quad (51)
\]

The conclusion \( \| L_tR_t^\top - X_\star \|_F \leq 1.5 \text{dist}(F_t, F_\star) \) is a simple consequence of Lemma 11; see (41) for a detailed argument. In what follows, we shall prove the distance contraction and the incoherence condition separately.

D.2.2 Distance Contraction

By the definition of \( \text{dist}^2(F_{t+1}, F_\star) \), one has

\[
\text{dist}^2(F_{t+1}, F_\star) \leq \|(L_{t+1}Q_t - L_\star)\Sigma_\star^{1/2}\|_F^2 + \|(R_{t+1}Q_t^{-\top} - R_\star)\Sigma_\star^{1/2}\|_F^2. \quad (52)
\]

From now on, we focus on controlling the first square \( \|(L_{t+1}Q_t - L_\star)\Sigma_\star^{1/2}\|_F^2 \). In view of the update rule, one has

\[
(L_{t+1}Q_t - L_\star)\Sigma_\star^{1/2} = (L - \eta(LR^\top + S - X_\star - S_\star)R(R^\top R)^{-1} - L_\star)\Sigma_\star^{1/2} \\
= (\Delta_L - \eta(LR^\top - X_\star)R(R^\top R)^{-1} - \eta(S - S_\star)R(R^\top R)^{-1})\Sigma_\star^{1/2} \\
= (1 - \eta)\Delta_L\Sigma_\star^{1/2} - \eta L_\star R(R^\top R)^{-1}\Sigma_\star^{1/2} - \eta(S - S_\star)R(R^\top R)^{-1}\Sigma_\star^{1/2}. \quad (53)
\]

Here, we use the notation introduced in Section D.2.1 and the decomposition \( LR^\top - X_\star = \Delta_L R^\top + L_\star \Delta_R^\top \). Take the squared Frobenius norm of both sides of (53) to obtain

\[
\|(L_{t+1}Q_t - L_\star)\Sigma_\star^{1/2}\|_F^2 = \|(1 - \eta)\Delta_L\Sigma_\star^{1/2} - \eta L_\star R(R^\top R)^{-1}\Sigma_\star^{1/2} - \eta(S - S_\star)R(R^\top R)^{-1}\Sigma_\star^{1/2}\|_F^2 \\
- 2\eta(1 - \eta)\text{tr}((S - S_\star)R(R^\top R)^{-1}\Sigma_\star \Delta_R^\top) \\
+ 2\eta^2 \text{tr}((S - S_\star)R(R^\top R)^{-1}\Sigma_\star (R^\top R)^{-1}R^\top \Delta_R L^\top) \\
+ 2\eta^2 \text{tr}((S - S_\star)R(R^\top R)^{-1}\Sigma_\star (R^\top R)^{-1}R^\top \Delta_R L^\top). \quad (54)
\]
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 (39) under the condition (49). It obeys

$$R_1 \leq \left(1 - \eta \right)^2 + \frac{2e}{1 - \epsilon} \eta \left(1 - \eta \right) \left\| \Delta_L \Sigma^1/2 \right\|_F^2 + \frac{2e + \epsilon^2}{(1 - \epsilon)^2} \eta^2 \left\| \Delta_R \Sigma^1/2 \right\|_F^2.$$

2. For the second term $R_2$, set $M := \Delta_L \Sigma^1/2 (R^\top R)^{-1} R^\top$ with $L_M := \Delta_L \Sigma^1/2 (R^\top R)^{-1} \Sigma^1/2$, $R_M := R \Sigma^1/2$, and then invoke Lemma 16 with $\nu := 3\sqrt{m}/2$ to see

$$|R_2| \leq \frac{3}{2} \sqrt{3 \alpha \mu \rho} \left( \left\| \Delta_L \Sigma^1/2 \right\|_F + \left\| \Delta_R \Sigma^1/2 \right\|_F \right) \left\| \Delta_L \Sigma^1/2 (R^\top R)^{-1} R^\top \right\|_F$$

$$+ 2\sqrt{\alpha \mu \rho} \left\| \Delta_L \Sigma^1/2 \right\|_F \left\| \Delta_L \Sigma^1/2 (R^\top R)^{-1} \Sigma^1/2 \right\|_F \left\| R \Sigma^1/2 \right\|_2, \infty \left\| LR^\top - X^*_\eta \right\|_F \leq \frac{3}{2} \sqrt{3 \alpha \mu \rho} \left( \left\| \Delta_L \Sigma^1/2 \right\|_F + \left\| \Delta_R \Sigma^1/2 \right\|_F \right) \left\| \Delta_L \Sigma^1/2 \right\|_F$$

$$+ 2\sqrt{\alpha \mu \rho} \left\| \Delta_L \Sigma^1/2 \right\|_F \left\| \Sigma^1/2 (R^\top R)^{-1} \Sigma^1/2 \right\|_F \left\| R \Sigma^1/2 \right\|_2, \infty \left\| LR^\top - X^*_\eta \right\|_F.$$

Take the condition (49) and Lemmas 9 and 11 together to obtain

$$\left\| R (R^\top R)^{-1} \Sigma^1/2 \right\|_F \leq \frac{1}{1 - \epsilon},$$

$$\left\| \Sigma^1/2 (R^\top R)^{-1} \Sigma^1/2 \right\|_F = \left\| R (R^\top R)^{-1} \Sigma^1/2 \right\|_F \leq \frac{1}{(1 - \epsilon)^2}.$$

$$\left\| LR^\top - X^*_\eta \right\|_F \leq \left(1 + \frac{\epsilon}{2}\right) \left( \left\| \Delta_L \Sigma^1/2 \right\|_F + \left\| \Delta_R \Sigma^1/2 \right\|_F \right).$$

These consequences combined with the condition (52) yield

$$|R_2| \leq \frac{3\sqrt{3 \alpha \mu \rho}}{2(1 - \epsilon)} \left( \left\| \Delta_L \Sigma^1/2 \right\|_F + \left\| \Delta_R \Sigma^1/2 \right\|_F \right) \left\| \Delta_L \Sigma^1/2 \right\|_F$$

$$+ \frac{4\sqrt{\alpha \mu \rho}}{(1 - \epsilon)^2} \left\| \Delta_L \Sigma^1/2 \right\|_F \left(1 + \frac{\epsilon}{2}\right) \left( \left\| \Delta_L \Sigma^1/2 \right\|_F + \left\| \Delta_R \Sigma^1/2 \right\|_F \right)$$

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

$$\leq \sqrt{\alpha \mu \rho} \frac{3\sqrt{3} + 4(2 + \epsilon)}{2(1 - \epsilon)} \left( \frac{3}{2} \left\| \Delta_L \Sigma^1/2 \right\|_F + \frac{1}{2} \left\| \Delta_R \Sigma^1/2 \right\|_F \right),$$

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

3. The third term $R_3$ can be controlled similarly. Set $M := L \Sigma^1/2 R (R^\top R)^{-1} \Sigma^1/2 (R^\top R)^{-1} R^\top$ with $L_M := L \Sigma^1/2$ and $R_M := R \Sigma^1/2$, and invoke Lemma 16 with $\nu := 3\sqrt{m}/2$ to arrive at

$$|R_3| \leq \frac{3}{2} \sqrt{3 \alpha \mu \rho} \left( \left\| \Delta_L \Sigma^1/2 \right\|_F + \left\| \Delta_R \Sigma^1/2 \right\|_F \right) \left\| L \Delta_L R (R^\top R)^{-1} \Sigma^1/2 (R^\top R)^{-1} R^\top \right\|_F$$

$$+ 2\sqrt{\alpha \mu \rho} \left\| L \Sigma^1/2 \right\|_2, \infty \left\| R (R^\top R)^{-1} \Sigma^1/2 (R^\top R)^{-1} R^\top \left\| R \Sigma^1/2 \right\|_F \right\| \left\| LR^\top - X^*_\eta \right\|_F \leq \frac{3}{2} \sqrt{3 \alpha \mu \rho} \left( \left\| \Delta_L \Sigma^1/2 \right\|_F + \left\| \Delta_R \Sigma^1/2 \right\|_F \right) \left\| R \Sigma^1/2 \right\|_F \left\| LR^\top - X^*_\eta \right\|_F^2.$$
For the last term $\mathcal{R}_4$, we utilize the variational representation of the Frobenius norm to see

$$\sqrt{\mathcal{R}_4} = \max_{L \in \mathbb{R}^{n \times r}} \tr \left( (S - S_\ast) R (R^T R)^{-1} \Sigma_\ast^1 \right).$$

Setting $M := \tilde{L} \Sigma_\ast^1 (R^T R)^{-1} R^T = L_M R_M^T$ with $L_M := \tilde{L} \Sigma_\ast^1 (R^T R)^{-1} \Sigma_\ast^1$ and $R_M := R \Sigma_\ast^{-1/2}$, we are ready to apply Lemma 16 again with $\nu := 3/\sqrt{\mu r}$ to reach at

$$\sqrt{\mathcal{R}_4} \leq \frac{3a_{\mu r}}{2} \left( \| \Delta_L \Sigma_\ast^1 \|_F + \| \Delta_R \Sigma_\ast^1 \|_F \right) \max_{L \in \mathbb{R}^{n \times r} \| L \|_F \leq 1} \left\| \tilde{L} \Sigma_\ast^1 (R^T R)^{-1} R^T \right\|_F$$

$$+ 2\sqrt{\alpha \mu} \max_{L \in \mathbb{R}^{n \times r} \| L \|_F \leq 1} \left\| \tilde{L} \Sigma_\ast^1 (R^T R)^{-1} \Sigma_\ast \right\|_F \| R \Sigma_\ast^{-1/2} \|_{2, \infty} \| LR^T - X_\ast \|_F$$

$$\leq \frac{3}{2} \sqrt{3/\mu r} \left( \| \Delta_L \Sigma_\ast^1 \|_F + \| \Delta_R \Sigma_\ast^1 \|_F \right) \left\| R (R^T R)^{-1} \Sigma_\ast \right\|_{op}$$

$$+ 2\sqrt{\alpha \mu} \left\| \Sigma_\ast^1 (R^T R)^{-1} \Sigma_\ast \right\|_{op} \| R \Sigma_\ast^{-1/2} \|_{2, \infty} \| LR^T - X_\ast \|_F.$$

This combined with the consequences (55) and condition (52) yields

$$\sqrt{\mathcal{R}_4} \leq \frac{3}{2} \sqrt{3/\mu r} \left( \frac{4(2+\epsilon)}{2(1-\epsilon)} \right) \left( \| \Delta_L \Sigma_\ast^1 \|_F + \| \Delta_R \Sigma_\ast^1 \|_F \right).$$

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

$$\mathcal{R}_4 \leq \alpha \mu \left( \frac{3}{2} \sqrt{3/\mu r} \left( \frac{4(2+\epsilon)}{2(1-\epsilon)} \right) \right)^2 \left( \| \Delta_L \Sigma_\ast^1 \|_F^2 + \| \Delta_R \Sigma_\ast^1 \|_F^2 \right).$$

Taking collectively the bounds for $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$ and $\mathcal{R}_4$ yields the control of $\| (L_{t+1} Q_t - L_\ast) \Sigma_\ast^{-1/2} \|_F^2$ as

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

$$+ \alpha \mu \left( \frac{3}{2} \sqrt{3/\mu r} \left( \frac{4(2+\epsilon)}{2(1-\epsilon)} \right) \right)^2 \left( \frac{3}{2} \| \Delta_L \Sigma_\ast^{-1/2} \|_F^2 + \| \Delta_R \Sigma_\ast^{-1/2} \|_F^2 \right).$$
Similarly, we can obtain the control of \( \|(R_{t+1}Q_t^\top - R_*)\Sigma_*^{1/2}\|_F^2 \). Combine them together and identify \( \text{dist}^2(F_t, F_*) = \|L_t\Sigma_*^{1/2}\|_F^2 + \|\Delta R\Sigma_*^{1/2}\|_F^2 \) to reach at

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

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

\[
\rho^2(\eta; \epsilon, \alpha \mu r) := (1 - \eta)^2 + \frac{2\epsilon}{1 - \epsilon} - \frac{2\epsilon^2 + \sqrt{\alpha \mu r}}{1 - \epsilon} \eta^2 + \frac{8(2 + \epsilon)}{1 - \epsilon} \eta^2.
\]

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

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

\[\tag{56}\]

**D.2.3 Incoherence Condition**

We start by controlling the term \( \|(L_{t+1}Q_t - L_*)\Sigma_*^{1/2}\|_{2,\infty} \). We know from (54) that

\[
(L_{t+1}Q_t - L_*)\Sigma_*^{1/2} = (1 - \eta)\Delta L\Sigma_*^{1/2} - \eta L^\top\mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\Sigma_*^{1/2} - \eta(S - S_*)\mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\Sigma_*^{1/2}.
\]

Apply the triangle inequality to obtain

\[
\left\| (L_{t+1}Q_t - L_*)\Sigma_*^{1/2} \right\|_{2,\infty} \leq (1 - \eta)\left\| \Delta L\Sigma_*^{1/2} \right\|_{2,\infty} + \eta \left\| L^\top\mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\Sigma_*^{1/2} \right\|_{2,\infty} + \eta \left\| (S - S_*)\mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\Sigma_*^{1/2} \right\|_{2,\infty}.
\]

The first term \( \|\Delta L\Sigma_*^{1/2}\|_{2,\infty} \) follows from the incoherence condition (50) as

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

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

1. For the term \( \mathfrak{T}_1 \), use the relation \( \|AB\|_{2,\infty} \leq \|A\|_{2,\infty} \|B\|_{op} \), and combine the condition (49) with the consequences (55) to obtain

\[
\mathfrak{T}_1 \leq \|L_*\Sigma_*^{-1/2}\|_{2,\infty} \left\| \Sigma_*^{1/2} \Delta R^\top\mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\Sigma_*^{1/2} \right\|_{op} \leq \|L_*\Sigma_*^{-1/2}\|_{2,\infty} \|\Delta R\Sigma_*^{1/2}\|_{op} \left\| \mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\Sigma_*^{1/2} \right\|_{op} \leq \frac{\epsilon}{1 - \epsilon} \sqrt{\frac{\mu r}{n_1}} \sigma_r(X^*),
\]

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

\[
\mathfrak{T}_2 \leq \|S - S_*\|_{2,\infty} \left\| \mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\Sigma_*^{1/2} \right\|_{op}.
\]

37
We know from Lemma 16 that $S - S_*$ has at most $3\alpha n_2$ non-zero entries in each row, and $\|S - S_*\|_\infty \leq 2\|LR^T - X_*\|_\infty$. Upper bound the $\ell_2,\infty$ norm by the $\ell_\infty$ norm as

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

Split $LR^T - X_* = \Delta_L R^T + L_* \Delta_R^T$, and use the conditions (50) and (52) to obtain

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

This combined with the consequences (55) yields

$$\Xi_2 \leq \frac{6\sqrt{3\alpha \mu r}}{1 - \epsilon} \frac{1}{\sqrt{n_1 n_2}} \sigma_r(X^*).$$

Taking collectively the bounds for $\Xi_1, \Xi_2$ yields the control

$$\left\| (L_{t+1} Q_t - L_\ast) \Sigma_{\ast}^{1/2} \right\|_{2,\infty} \leq \left( 1 - \eta + \frac{\epsilon + 6\sqrt{3\alpha \mu r}}{1 - \epsilon} \eta \right) \left( 1 - \eta \right) \sigma_r(X_\ast). \tag{57}$$

Finally, we need to change the alignment matrix from $Q_t$ to $Q_{t+1}$. (56) together with Lemma 6 demonstrates the existence of $Q_{t+1}$. Apply the triangle inequality to obtain

$$\left\| (L_{t+1} Q_{t+1} - L_\ast) \Sigma_{\ast}^{1/2} \right\|_{2,\infty} \leq \left\| (L_{t+1} Q_t - L_\ast) \Sigma_{\ast}^{1/2} \right\|_{2,\infty} + \left\| (L_{t+1} Q_t - L_{t+1} Q_{t+1}) \Sigma_{\ast}^{1/2} \right\|_{2,\infty} \leq \left( 1 - \eta \right) \sigma_r(X_\ast). \tag{57}$$

We deduce from (57) that

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

For the alignment matrix term, invoke Lemma 12 to obtain

$$\left\| \Sigma_{\ast} - \Sigma_{\ast}^{1/2} Q_t^{-1} Q_{t+1} \Sigma_{\ast}^{1/2} \right\|_{\text{op}} \leq \frac{\left\| (R_{t+1} Q_{t+1}^{-1} - Q_t^{-1}) \Sigma_{\ast}^{1/2} \right\|_{\text{op}}}{1 - \mu_t \Sigma_{\ast}^{-1/2}} \leq \frac{\left\| (R_t Q_t^{-1} - R_\ast) \Sigma_{\ast}^{1/2} \right\|_{\text{op}}}{1 - \mu_t \Sigma_{\ast}^{-1/2}} \leq \frac{2\epsilon}{1 - \epsilon} \sigma_r(X^*),$$

where we deduce from (56) that the distances using either $Q_t$ or $Q_{t+1}$ are bounded by

$$\| R_{t+1} Q_t^{-1} - R_\ast \|_{\text{op}} \leq \epsilon \sigma_r(X^*),$$

$$\| R_t Q_t^{-1} - R_\ast \|_{\text{op}} \leq \epsilon \sigma_r(X^*),$$

$$\| (R_{t+1} Q_t^{-1} - R_\ast) \Sigma_{\ast}^{-1/2} \|_{\text{op}} \leq \epsilon.$$

Combine all pieces to reach

$$\left\| (L_{t+1} Q_{t+1} - L_\ast) \Sigma_{\ast}^{1/2} \right\|_{2,\infty} \leq \left( 1 + \frac{\epsilon + 6\sqrt{3\alpha \mu r}}{1 - \epsilon} \eta \right) \left( 1 - \eta \right) \sigma_r(X_\ast) + \frac{2\epsilon}{1 - \epsilon} \left( 1 - \eta \right) \sigma_r(X_\ast).$$
With $\epsilon = 0.02$, $\alpha \mu r \leq 10^{-4}$, and $0.1 \leq \eta \leq 2/3$, we get the desired incoherence condition
\[
\left\| (L_{t+1} Q_{t+1} - L_\star) \Sigma_\star^{1/2} \right\|_{2, \infty} \leq \sqrt{\frac{\mu r}{n_1}} \sigma_r(X_\star).
\]
Similarly, we can prove the other part
\[
\left\| (R_{t+1} Q_{t+1} - R_\star) \Sigma_\star^{1/2} \right\|_{2, \infty} \leq \sqrt{\frac{\mu r}{n_2}} \sigma_r(X^\star).
\]

### D.3 Proof of Lemma 4

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

**Lemma 17 ([YPCC16, Section 6.1]).** For $S_\star \in S_\alpha$, one has $S_\star - T_\alpha[X_\star + S_\star] \in S_{2\alpha}$, and $\|S_\star - T_\alpha[X_\star + S_\star]\|_\infty \leq 2\|X_\star\|_\infty$.

**Lemma 18 ([YPCC16, Lemma 1]).** For any matrix $X \in \mathbb{R}^{n_1 \times n_2}$ that belongs to $S_\alpha$, one has
\[
\|X\|_{op} \leq \alpha \sqrt{\frac{1}{n_2}} \|X\|_\infty.
\]

With these two lemmas in place, we are ready to establish the claimed result. Invoke Lemma 8 to obtain
\[
dist(F_0, F_\star) \leq \sqrt{2 + 1} \left\| L_0 R_0^\top - X_\star \right\|_F \leq \sqrt{(\sqrt{2} + 1)2r} \left\| L_0 R_0^\top - X_\star \right\|_{op},
\]
where the last relation uses the fact that $L_0 R_0^\top - X_\star$ has rank at most $2r$. We can further apply the triangle inequality to see
\[
\left\| L_0 R_0^\top - X_\star \right\|_{op} \leq \left\| Y - T_\alpha[Y] - L_0 R_0^\top \right\|_{op} + \left\| Y - T_\alpha[Y] - X_\star \right\|_{op} \\
\leq 2\left\| Y - T_\alpha[Y] - X_\star \right\|_{op} = 2\|S_\star - T_\alpha[X_\star + S_\star]\|_{op}.
\]

Here the second inequality hinges on the fact that $L_0 R_0^\top$ is the best rank-$r$ approximation of $Y - T_\alpha[Y]$, and the last identity arises from $Y = X_\star + S_\star$. Following the same argument as in [YPCC16, Eq. (16)], we can invoke Lemma 17 and Lemma 18 to reach
\[
\|S_\star - T_\alpha[X_\star + S_\star]\|_{op} \leq 2\alpha \sqrt{\frac{1}{n_1 n_2}} \|S_\star - T_\alpha[X_\star + S_\star]\|_\infty \\
\leq 4\alpha \sqrt{\frac{1}{n_1 n_2}} \|X_\star\|_\infty \leq 4\alpha \mu r \kappa \sigma_r(X_\star),
\]
where the last step uses the incoherence assumption
\[
\|X_\star\|_\infty \leq \|U^\star\|_{2, \infty} \|\Sigma_\star\|_{op} \|V^\star\|_{2, \infty} \leq \frac{\mu r}{\sqrt{n_1 n_2}} \kappa \sigma_r(X_\star).
\]

Take the above inequalities together to arrive at
\[
dist(F_0, F_\star) \leq 8\sqrt{(\sqrt{2} + 1)\alpha \mu r^{3/2}} \kappa \sigma_r(X_\star) \leq 20\alpha \mu r^{3/2} \kappa \sigma_r(X_\star).
\]

### D.4 Proof of Lemma 5

In view of the condition $\dist(F_0, F_\star) \leq 0.02 \sigma_r(X_\star)$ and Lemma 6, one knows that $Q_0$, the optimal alignment matrix between $F_0$ and $F_\star$, exists. Therefore, for notational convenience, denote $L := L_0 Q_0$, $R := R_0 Q_0^{-1}$, $\Delta_L := L - L_\star$, $\Delta_R := R - R_\star$, and $\epsilon := 0.02$. Our objective is then translated to demonstrating
\[
\sqrt{\frac{1}{n_1}} \Delta_L \Sigma_\star^{1/2} \leq \sqrt{\frac{1}{n_2}} \Delta_R \Sigma_\star^{1/2} \leq \sqrt{\frac{\mu r}{n_1}} \sigma_r(X_\star).
\]

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 $Y = X_\star + S_\star$, one has 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
\[ \Delta_L \Sigma_*^{1/2} = (S_* - \mathcal{T}_\alpha [X_* + S_*]) R (R^\top R)^{-1} \Sigma_*^{1/2} - L \Delta_R R (R^\top R)^{-1} \Sigma_*^{1/2}. \]

Combining Claim 4 with the triangle inequality yields
\[ \| \Delta_L \Sigma_*^{1/2} \|_{2,\infty} \leq \left\| L \Delta_R R (R^\top R)^{-1} \Sigma_*^{1/2} \right\|_{2,\infty} + \left\| (S_* - \mathcal{T}_\alpha [X_* + S_*]) R (R^\top R)^{-1} \Sigma_*^{1/2} \right\|_{2,\infty}. \]

In what follows, we shall control \( \mathcal{J}_1 \) and \( \mathcal{J}_2 \) in turn.

1. For the term \( \mathcal{J}_1 \), use the relation \( \| AB \|_{2,\infty} \leq \| A \|_{2,\infty} \| B \|_{\text{op}} \) to obtain
\[ \mathcal{J}_1 \leq \| L \Sigma_*^{-1/2} \|_{2,\infty} \| \Delta_R \Sigma_*^{1/2} \|_{\text{op}} \| R (R^\top R)^{-1} \Sigma_*^{1/2} \|_{\text{op}}. \]

Clearly, by the incoherence assumption, we have \( \| L \Sigma_*^{-1/2} \|_{2,\infty} = \| U_* \|_{2,\infty} \leq \sqrt{\mu r / n_1} \). In addition, the assumption \( \text{dist} (F_0, F_*) \leq c r \sigma_r (X_*) \) entails the bound \( \| \Delta_R \Sigma_*^{1/2} \|_{\text{op}} \leq c r \sigma_r (X_*) \). Finally, repeating the argument for obtaining (49) yields \( \| \Delta_R \Sigma_*^{1/2} \|_{\text{op}} \leq \epsilon \), which together with Lemma 9 reveals \( \| R (R^\top R)^{-1} \Sigma_*^{1/2} \|_{\text{op}} \leq 1/(1 - \epsilon) \). In all, we arrive at
\[ \mathcal{J}_1 \leq \frac{\epsilon}{1 - \epsilon} \sqrt{\frac{\mu r}{n_1}} \sigma_r (X_*). \]

2. Proceeding to the term \( \mathcal{J}_2 \), use the relations \( \| AB \|_{2,\infty} \leq \| A \|_{1,\infty} \| B \|_{2,\infty} \) and \( \| AB \|_{2,\infty} \leq \| A \|_{2,\infty} \| B \|_{\text{op}} \) to obtain
\[ \mathcal{J}_2 \leq \| S_* - \mathcal{T}_\alpha [X_* + S_*] \|_{1,\infty} \| R (R^\top R)^{-1} \Sigma_*^{1/2} \|_{2,\infty} \]
\[ \leq \| S_* - \mathcal{T}_\alpha [X_* + S_*] \|_{1,\infty} \| R \Sigma_*^{1/2} \|_{2,\infty} \| \Sigma_*^{1/2} (R^\top R)^{-1} \Sigma_*^{1/2} \|_{\text{op}}. \]

Regarding \( S_* - \mathcal{T}_\alpha [X_* + S_*] \), Lemma 17 tells us that \( S_* - \mathcal{T}_\alpha [X_* + S_*] \) has at most \( 2 \alpha n_2 \) non-zero entries in each row, and also \( \| S_* - \mathcal{T}_\alpha [X_* + S_*] \|_{\infty} \leq 2 \| X_* \|_{\infty} \). Consequently, we can upper bound the \( \ell_{1,\infty} \) norm by the \( \ell_{\infty} \) norm
\[ \| S_* - \mathcal{T}_\alpha [X_* + S_*] \|_{1,\infty} \leq 2 \alpha n_2 \| S_* - \mathcal{T}_\alpha [X_* + S_*] \|_{\infty} \]
\[ \leq 4 \alpha n_2 \| X_* \|_{\infty} \]
\[ \leq 4 \alpha n_2 \sqrt{n_2} \kappa r (X_*). \]

Here the last inequality holds true due to the incoherence assumption (58). For the term \( \| R \Sigma_*^{-1/2} \|_{2,\infty} \), one can apply the triangle inequality to see
\[ \| R \Sigma_*^{-1/2} \|_{2,\infty} \leq \| R \Sigma_*^{-1/2} \|_{2,\infty} + \| \Delta_R \Sigma_*^{1/2} \|_{2,\infty} \leq \sqrt{\frac{\mu r}{n_2}} + \frac{\| \Delta_R \Sigma_*^{1/2} \|_{2,\infty}}{\sigma_r (X_*)}. \]

Last but not least, repeat the argument for (55) to obtain
\[ \left\| \Sigma_*^{1/2} (R^\top R)^{-1} \Sigma_*^{1/2} \right\|_{\text{op}} \leq \| R (R^\top R)^{-1} \Sigma_*^{1/2} \|_{\text{op}} \leq \frac{1}{(1 - \epsilon)^2}. \]

Taking together the above bounds yields
\[ \mathcal{J}_2 \leq \frac{4 \alpha \mu r \kappa}{(1 - \epsilon)^2} \sqrt{\frac{\mu r}{n_1}} \sigma_r (X_*) + \frac{4 \alpha \mu r \kappa}{(1 - \epsilon)^2} \sqrt{\frac{n_2}{n_1}} \| \Delta_R \Sigma_*^{1/2} \|_{2,\infty}. \]
Combine the bounds on $I_1$ and $I_2$ to reach
\[
\sqrt{n_1} \| \Delta_L \Sigma^1_{\ast} \|_{2, \infty} \leq \frac{\epsilon(1 - \epsilon) + 4\alpha \mu \kappa}{(1 - \epsilon)^2} \sqrt{\mu r} \sigma_r(X_\ast) + \frac{4\alpha \mu \kappa}{(1 - \epsilon)^2} \sqrt{n_2} \| \Delta_H \Sigma^1_{\ast} \|_{2, \infty}.
\]
Similarly, we have
\[
\sqrt{n_2} \| \Delta_H \Sigma^1_{\ast} \|_{2, \infty} \leq \frac{\epsilon(1 - \epsilon) + 4\alpha \mu \kappa}{(1 - \epsilon)^2} \sqrt{\mu r} \sigma_r(X_\ast) + \frac{4\alpha \mu \kappa}{(1 - \epsilon)^2} \sqrt{n_1} \| \Delta_L \Sigma^1_{\ast} \|_{2, \infty}.
\]
Taking the maximum and solving for $\sqrt{n_1} \| \Delta_L \Sigma^1_{\ast} \|_{2, \infty} \vee \sqrt{n_2} \| \Delta_L \Sigma^1_{\ast} \|_{2, \infty}$ yield the relation
\[
\sqrt{n_1} \| \Delta_L \Sigma^1_{\ast} \|_{2, \infty} \vee \sqrt{n_2} \| \Delta_L \Sigma^1_{\ast} \|_{2, \infty} \leq \frac{\epsilon(1 - \epsilon) + 4\alpha \mu \kappa}{(1 - \epsilon)^2} \sqrt{\mu r} \sigma_r(X_\ast).
\]
With $\epsilon = 0.02$ and $\alpha \mu \kappa \leq 0.1$, we get the desired conclusion
\[
\sqrt{n_1} \| \Delta_L \Sigma^1_{\ast} \|_{2, \infty} \vee \sqrt{n_2} \| \Delta_L \Sigma^1_{\ast} \|_{2, \infty} \leq \sqrt{\mu r} \sigma_r(X_\ast).
\]

**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 $$(X_\ast + S_\ast - \mathcal{T}_\alpha[X_\ast + S_\ast])R_0(R_0^\top R_0)^{-1} = L_0,$$
which is equivalent to $(X_\ast + S_\ast - \mathcal{T}_\alpha[X_\ast + S_\ast])R_0(R_0^\top R_0)^{-1} = L_0$ since $\Sigma_0 = R_0^\top R_0$. Multiply both sides by $Q_0 \Sigma_1^{1/2}$ to obtain
\[
(X_\ast + S_\ast - \mathcal{T}_\alpha[X_\ast + S_\ast])R(R^\top R)^{-1} \Sigma_1^{1/2} = L \Sigma_1^{1/2},
\]
where we recall that $L = L_0 Q_0$ and $R = R_0 Q_0^\top$. In the end, subtract $X \Sigma_1^{1/2}$ from both sides to reach
\[
(S_\ast - \mathcal{T}_\alpha[X_\ast + S_\ast])R(R^\top R)^{-1} \Sigma_1^{1/2} = L \Sigma_1^{1/2} - L_\alpha R^\top R R^\top R)^{-1} \Sigma_1^{1/2} = (L - L_\alpha) \Sigma_1^{1/2} + L_\alpha (R - R_\alpha)^\top R(R^\top R)^{-1} \Sigma_1^{1/2} = \Delta_L \Sigma_1^{1/2} + L_\alpha \Delta_H R(R^\top R)^{-1} \Sigma_1^{1/2}.
\]
This finishes the proof.

\[\square\]

**E. Proof for General Loss Functions**

**E.1 Preliminaries**

The following lemma presents a useful property of restricted smooth and convex functions.

**Lemma 19.** Suppose that $f(\cdot) : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$ is rank-$2r$ restricted $L$-smooth and rank-$2r$ restricted convex. Then the following holds for any $X_1, X_2 \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $r$

\[
\frac{1}{L} \| \nabla f(X_1) - \nabla f(X_2) \|^2_{p,r} \leq \langle \nabla f(X_1) - \nabla f(X_2), X_1 - X_2 \rangle \leq L \| X_1 - X_2 \|^2.
\]

As a simple corollary, one has

\[
\| \nabla f(X_1) - \nabla f(X_2) \|_{p,r} \leq L \| X_1 - X_2 \|.
\]

**Proof.** In view of the definition of rank-$2r$ restricted $L$-smoothness, we have for any $X_1, X_2 \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $r$

\[
f(X_1) \leq f(X_2) + \langle \nabla f(X_2), X_1 - X_2 \rangle + \frac{L}{2} \| X_1 - X_2 \|^2, \quad \text{and}
\]

\[
f(X_2) \leq f(X_1) + \langle \nabla f(X_1), X_2 - X_1 \rangle + \frac{L}{2} \| X_2 - X_1 \|^2.
\]

41
f(X_2) \leq f(X_1) + \langle \nabla f(X_1), X_2 - X_1 \rangle + \frac{L}{2} \| X_1 - X_2 \|_F^2.

Sum the above two inequalities to obtain the claimed upper bound \( \langle \nabla f(X_1) - \nabla f(X_2), X_1 - X_2 \rangle \leq L \| X_1 - X_2 \|_F^2 \).

Next, we proceed to the lower bound. Since \( f \) is rank-2r restricted \( L \)-smooth and convex, we obtain for any \( \bar{X} \in \mathbb{R}^{n_1 \times n_2} \) with rank at most 2r

\[
f(X_1) + \langle \nabla f(X_1), \bar{X} - X_1 \rangle \leq f(X_2) + \langle \nabla f(X_2), \bar{X} - X_2 \rangle + \frac{L}{2} \| \bar{X} - X_2 \|_F^2.
\]

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 \( \bar{X} = X_2 - \frac{1}{L} \mathcal{P}_r(\nabla f(X_2) - \nabla f(X_1)) \) (whose rank is at most 2r) to see

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

We can further switch the roles of \( X_1 \) and \( X_2 \) to obtain

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

Adding the above two inequalities yields the lower bound \( \langle \nabla f(X_1) - \nabla f(X_2), X_1 - X_2 \rangle \geq \frac{1}{2} \| \nabla f(X_1) - \nabla f(X_2) \|_{F,r}^2 \).

\( \square \)

E.2 Proof of Theorem 3

Suppose that the \( t \)th iterate \( F_t \) obeys the condition \( \text{dist}(F_t, F_*) \leq 0.02\sigma_r(X_*) \). In view of Lemma 6, one knows that \( Q_t \), the optimal alignment matrix between \( F_1 \) and \( F_\ast \) exists. Therefore, for notational convenience, denote \( L := L_t Q_t \), \( R := R_t Q_t^{-\top} \), \( \Delta_L := L - L_\ast \), \( \Delta_R := R - R_\ast \), and \( \epsilon := 0.02 \). Similar to the derivation in (38), we have

\[
\| \Delta_L \Sigma_*^{-1/2} \|_{op} \vee \| \Delta_R \Sigma_*^{-1/2} \|_{op} \leq \epsilon. \tag{59}
\]

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 11; see (41) for a detailed argument. From now on, we focus on proving the distance contraction.

By the definition of \( \text{dist}(F_{t+1}, F_\ast) \), one has

\[
\text{dist}^2(F_{t+1}, F_\ast) \leq \| (L_{t+1} Q_t - L_\ast) \Sigma_*^{1/2} \|_F^2 + \| (R_{t+1} Q_t^{-\top} - R_\ast) \Sigma_*^{1/2} \|_F^2. \tag{60}
\]

Introduce an auxiliary function

\[
f_\mu(X) = f(X) - \frac{\mu}{2} \| X - X_\ast \|_F^2,
\]

which is rank-2r restricted \((L - \mu)\)-smooth and rank-2r restricted convex. Using the ScaledGD update rule (23) and the decomposition \( L R_t^\top - X_\ast = \Delta_L R_t^\top + L_\ast \Delta_R^\top \), we obtain

\[
(L_{t+1} Q_t - L_\ast) \Sigma_*^{1/2} = (L - \eta \nabla f(L R_t^\top) R(R^\top R)^{-1} - L_\ast) \Sigma_*^{1/2} - \eta f_\mu(L R_t^\top) R(R^\top R)^{-1} - L_\ast \Sigma_*^{1/2} = (1 - \eta \mu) \Delta_L \Sigma_*^{1/2} - \eta \mu L_\ast \Delta_R^\top (R R^\top R)^{-1} \Sigma_*^{1/2} - \eta f_\mu(L R_t^\top) R(R^\top R)^{-1} \Sigma_*^{1/2}.
\]

As a result, one can expand the first square in (60) as

\[
\| (L_{t+1} Q_t - L_\ast) \Sigma_*^{1/2} \|_F^2 = \| (1 - \eta \mu) \Delta_L \Sigma_*^{1/2} - \eta \mu L_\ast \Delta_R^\top (R R^\top R)^{-1} \Sigma_*^{1/2} \|_F^2 \tag{61}
\]

42
In the sequel, we shall bound the four terms separately, of which $\mathfrak{S}_1$ is the dominant term, and $\mathfrak{S}_2, \mathfrak{S}_3$ and $\mathfrak{S}_4$ are small perturbation terms.

1. Notice that the main term $\mathfrak{S}_1$ has already been controlled in (39) under the condition (59). It obeys

$$\mathfrak{S}_1 \leq \left(1 - \eta \mu\right)^2 + \frac{2\epsilon}{1 - \epsilon} \eta \mu \left(1 - \eta \mu\right) \left\| \Delta_L \Sigma_r^{1/2} \right\|_F^2 + \frac{2\epsilon + \epsilon^2}{(1 - \epsilon)^2} \eta^2 \mu^2 \left\| \Delta_R \Sigma_r^{1/2} \right\|_F^2.$$

2. For the second term $\mathfrak{S}_2$, note that $\Delta_L \Sigma_r (R^T R)^{-1} R^T - \Delta_L R_*^T - \frac{1}{2} \Delta_L \Delta_R^T$ has rank at most $r$. Hence we can invoke Lemma 13 to obtain

$$|\mathfrak{S}_2| \leq \|\nabla f_\mu (LR^T)\|_{F,r} \left\| \Delta_L \Sigma_r (R^T R)^{-1} R^T - \Delta_L R_*^T - \frac{1}{2} \Delta_L \Delta_R^T \right\|_F \leq \|\nabla f_\mu (LR^T)\|_{F,r} \left\| \Delta_L \Sigma_r^{1/2} \right\|_F \left( \left\| R (R^T R)^{-1} \Sigma_r^{1/2} - V_r \right\|_{op} + \frac{1}{2} \left\| \Delta_R \Sigma_r^{1/2} \right\|_{op} \right),$$

where the second line uses $R_* = V_r \Sigma_r^{1/2}$. Notice that $\nabla f_\mu (X_*) = 0$. Invoke Lemma 19 to obtain

$$\|\nabla f_\mu (LR^T)\|_{F,r} \leq (L - \mu) \|LR^T - X_*\|_F.$$

Take the condition (59) and Lemmas 9-11 together to obtain

$$\left\| R (R^T R)^{-1} \Sigma_r^{1/2} - V_r \right\|_{op} \leq \frac{\sqrt{2\epsilon}}{1 - \epsilon},$$

$$\left\| R (R^T R)^{-1} \Sigma_r^{1/2} \right\|_{op} \leq \frac{1}{1 - \epsilon}.$$

$$\|LR^T - X_*\|_F \leq (1 + \frac{\epsilon}{2}) \left( \left\| \Delta_L \Sigma_r^{1/2} \right\|_F + \left\| \Delta_R \Sigma_r^{1/2} \right\|_F \right).$$

These consequences further imply that

$$|\mathfrak{S}_2| \leq (L - \mu) \left( 1 + \frac{\epsilon}{2} \right) \left( \left\| \Delta_L \Sigma_r^{1/2} \right\|_F + \left\| \Delta_R \Sigma_r^{1/2} \right\|_F \right) \left\| \Delta_L \Sigma_r^{1/2} \right\|_F \left( \frac{\sqrt{2\epsilon}}{1 - \epsilon} + \frac{\epsilon}{2} \right) \leq (L - \mu) \frac{\epsilon (2 + \epsilon)}{1 - \epsilon} \left( \left\| \Delta_L \Sigma_r^{1/2} \right\|_F + \left\| \Delta_L \Sigma_r^{1/2} \right\|_F \right) \left\| \Delta_R \Sigma_r^{1/2} \right\|_F \leq (L - \mu) \frac{\epsilon (2 + \epsilon)}{1 - \epsilon} \left( \frac{3}{2} \left\| \Delta_L \Sigma_r^{1/2} \right\|_F^2 + \frac{1}{2} \left\| \Delta_R \Sigma_r^{1/2} \right\|_F^2 \right),$$

where the second inequality uses the fact that $\epsilon = 0.02$ and the last line follows from $2ab \leq a^2 + b^2$.

3. As above, the third term $\mathfrak{S}_3$ can be similarly bounded as

$$|\mathfrak{S}_3| \leq \|\nabla f_\mu (LR^T)\|_{F,r} \left\| L, \Delta_R^T R (R^T R)^{-1} \Sigma_r (R^T R)^{-1} R^T \right\|_F.$$
\[ \begin{align*}
&\leq \| \nabla f_{\mu}(LR^T) \|_{F,r} \| \Delta_R \Sigma_s^{1/2} \|_F \| R(R^T R)^{-1} \Sigma_s^{1/2} \|_{\text{op}}^2 \\
&\leq (L - \mu) \left( 1 + \frac{\epsilon}{2} \right) \left( \| \Delta_L \Sigma_s^{1/2} \|_F + \| \Delta_R \Sigma_s^{1/2} \|_F \right) \| \Delta_R \Sigma_s^{1/2} \|_{\text{op}}^2 \frac{1}{(1 - \epsilon)^2} \\
&\leq (L - \mu) \frac{2 + \epsilon}{2(1 - \epsilon)^2} \left( \frac{1}{2} \| \Delta_L \Sigma_s^{1/2} \|_F^2 + \frac{3}{2} \| \Delta_R \Sigma_s^{1/2} \|_F^2 \right).
\end{align*} \]

4. For the last term \( \Phi_4 \), invoke Lemma 13 to obtain
\[ \Phi_4 \leq \| \nabla f_{\mu}(LR^T) \|_{F,r}^2 \| R(R^T R)^{-1} \Sigma_s^{1/2} \|_{\text{op}}^2 \leq \frac{1}{(1 - \epsilon)^2} \| \nabla f_{\mu}(LR^T) \|_{F,r}^2. \]

Taking collectively the bounds for \( \Phi_1, \Phi_2, \Phi_3 \) and \( \Phi_4 \) yields
\[ \left\| (L_{t+1}Q_t - L_s) \Sigma_s^{1/2} \right\|_F^2 \leq \left( 1 - \eta \mu \right)^2 + \frac{2\epsilon}{1 - \epsilon} \eta \mu (1 - \eta \mu) \left( \| \Delta_L \Sigma_s^{1/2} \|_F^2 + \frac{2\epsilon + \epsilon^2}{(1 - \epsilon)^2} \eta^2 \mu^2 \| \Delta_R \Sigma_s^{1/2} \|_F^2 \right) \\
- 2\eta (1 - \eta \mu) \left\langle \nabla f_{\mu}(LR^T), \Delta_L R_s^T + \frac{1}{2} \Delta_L \Delta_R \right\rangle \\
+ \frac{2\epsilon (2 + \epsilon)}{1 - \epsilon} \eta (L - \mu) \left( \frac{3}{2} \| \Delta_L \Sigma_s^{1/2} \|_F^2 + \frac{1}{2} \| \Delta_R \Sigma_s^{1/2} \|_F^2 \right) \\
+ \frac{2 + \epsilon}{(1 - \epsilon)^2} \eta^2 \mu (L - \mu) \left( \frac{1}{2} \| \Delta_L \Sigma_s^{1/2} \|_F^2 + \frac{3}{2} \| \Delta_R \Sigma_s^{1/2} \|_F^2 \right) \\
+ \frac{\eta^2}{(1 - \epsilon)^2} \| \nabla f_{\mu}(LR^T) \|_{F,r}^2. \]

Similarly, we can obtain the following bound for \( \left\| (R_{t+1}Q_t^{-T} - R_s) \Sigma_s^{1/2} \right\|_F^2 \)
\[ \left\| (R_{t+1}Q_t^{-T} - R_s) \Sigma_s^{1/2} \right\|_F^2 \leq \left( 1 - \eta \mu \right)^2 + \frac{2\epsilon}{1 - \epsilon} \eta \mu (1 - \eta \mu) \left( \| \Delta_R \Sigma_s^{1/2} \|_F^2 + \frac{2\epsilon + \epsilon^2}{(1 - \epsilon)^2} \eta^2 \mu^2 \| \Delta_L \Sigma_s^{1/2} \|_F^2 \right) \\
- 2\eta (1 - \eta \mu) \left\langle \nabla f_{\mu}(LR^T), L_s \Delta_L + \frac{1}{2} \Delta_L \Delta_R \right\rangle \\
+ \frac{2\epsilon (2 + \epsilon)}{1 - \epsilon} \eta (L - \mu) \left( \frac{1}{2} \| \Delta_L \Sigma_s^{1/2} \|_F^2 + \frac{3}{2} \| \Delta_R \Sigma_s^{1/2} \|_F^2 \right) \\
+ \frac{2 + \epsilon}{(1 - \epsilon)^2} \eta^2 \mu (L - \mu) \left( \frac{3}{2} \| \Delta_L \Sigma_s^{1/2} \|_F^2 + \frac{1}{2} \| \Delta_R \Sigma_s^{1/2} \|_F^2 \right) \\
+ \frac{\eta^2}{(1 - \epsilon)^2} \| \nabla f_{\mu}(LR^T) \|_{F,r}^2. \]

Further notice that \( LR^T - X_s = L_s \Delta_R + \Delta_L \Sigma_s^{1/2} + \Delta_L \Sigma_s^{1/2} \), and invoke Lemma 19 to obtain
\[ \langle \nabla f_{\mu}(LR^T), L_s \Delta_R + \Delta_L \Sigma_s^{1/2} + \Delta_L \Sigma_s^{1/2} \rangle = \langle \nabla f_{\mu}(LR^T), LR^T - X_s \rangle \geq \frac{1}{L - \mu} \| \nabla f_{\mu}(LR^T) \|_{F,r}^2. \]

With this lower bound in place, we can now combine the bounds together to arrive at
\[ \left\| (L_{t+1}Q_t - L_s) \Sigma_s^{1/2} \right\|_F^2 + \left\| (R_{t+1}Q_t^{-T} - R_s) \Sigma_s^{1/2} \right\|_F^2 \leq \rho^2 (\eta, \epsilon, L, \mu) \text{ dist}^2 (F_t, F_s) \\
- 2\eta \left( 1 - \frac{\eta \mu}{L - \mu} \right) \| \nabla f_{\mu}(LR^T) \|_{F,r}^2, \quad (61) \]
where the contraction rate is given by

\[ \rho^2(\eta; \epsilon, L, \mu) := (1 - \eta \mu)^2 + \frac{2 \epsilon}{1 - \epsilon} \eta \mu (1 - \eta \mu) + \frac{2 \epsilon + \epsilon^2}{(1 - \epsilon)^2} \eta^2 \mu^2 \]

\[ + \frac{4 \epsilon (2 + \epsilon)}{1 - \epsilon} \eta (L - \mu)(1 - \eta \mu) + \frac{4 + 2 \epsilon}{(1 - \epsilon)^2} \eta^2 \mu (L - \mu). \]

With \( \epsilon = 0.02 \) and \( 0 < \eta \leq 0.2/L \), one has \( \rho(\eta; \epsilon, L, \mu) \leq 1 - 0.7 \eta \mu \), and the second term in (61) is negative. Thus we conclude that

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

which is the desired claim.