Beyond Procrustes: Balancing-Free Gradient Descent for Asymmetric Low-Rank Matrix Sensing

Cong Ma∗ Yuanxin Li† Yuejie Chi†

Abstract

Low-rank matrix estimation plays a central role in various applications across science and engineering. Recently, nonconvex formulations based on matrix factorization are provably solved by simple gradient descent algorithms with strong computational and statistical guarantees. However, when the low-rank matrices are asymmetric, existing approaches rely on adding a regularization term to balance the scale of the two matrix factors which in practice can be removed safely without hurting the performance when initialized via the spectral method. In this paper, we justify this theoretically for the matrix sensing problem, which aims to recover a low-rank matrix from a small number of linear measurements. As long as the measurement ensemble satisfies the restricted isometry property, gradient descent — in conjunction with spectral initialization — converges linearly without the need of explicitly promoting balancedness of the factors; in fact, the factors stay balanced automatically throughout the execution of the algorithm. Our analysis is based on analyzing the evolution of a new distance metric that directly accounts for the ambiguity due to invertible transforms, and might be of independent interest.

1 Introduction

Low-rank matrix estimation plays a central role in many applications [2] [3] [4]. Broadly speaking, we are interested in estimating a rank-$r$ matrix $M_\ast = X_\ast Y_\ast^\top \in \mathbb{R}^{n_1 \times n_2}$ by solving a rank-constrained optimization problem:

$$\min_{M \in \mathbb{R}^{n_1 \times n_2}} \mathcal{L}(M) \quad \text{subject to} \quad \text{rank}(M) \leq r, \quad (1)$$

where $\mathcal{L}(\cdot)$ denotes a certain loss function and the rank $r$ is typically much smaller than the dimension of the matrix. To reduce computational complexity, a common approach, popularized by the work of Burer and Monteiro [5], is to factorize $M = XY^\top$ with $X \in \mathbb{R}^{n_1 \times r}$ and $Y \in \mathbb{R}^{n_2 \times r}$, and rewrite the above problem (1) into an unconstrained nonconvex optimization problem:

$$\min_{X \in \mathbb{R}^{n_1 \times r}, Y \in \mathbb{R}^{n_2 \times r}} f(X, Y) \triangleq \mathcal{L}(XY^\top). \quad (2)$$

Despite nonconvexity, one might be tempted to estimate the low-rank factors $(X, Y)$ via gradient descent, which proceeds via the following update rule: for $t \geq 0$

$$\begin{bmatrix} X_{t+1} \\ Y_{t+1} \end{bmatrix} = \begin{bmatrix} X_t \\ Y_t \end{bmatrix} - \eta_t \begin{bmatrix} \nabla_X f(X_t, Y_t) \\ \nabla_Y f(X_t, Y_t) \end{bmatrix}. \quad (3)$$

Here, $\eta_t$ is the step size and $(X_0, Y_0)$ is some proper initialization.

Significant progress has been made recently in understanding the performance of gradient descent for nonconvex matrix estimation. Somewhat surprisingly, most of the existing guarantees are not directly applicable to the vanilla gradient descent rule (3). One particular challenge is associated with the identifiability of the factors $(X, Y)$ — they...
are indistinguishable as long as their product $X Y^\top$ is the same. What is worse, if the norms of the factors become highly imbalanced, gradient descent might diverge easily. Consequently, it becomes a routine procedure to insert a regularizer $g(X, Y)$ that balances the two factors [6, 7, 8].

$$g(X, Y) \triangleq \lambda \|X^\top X - Y^\top Y\|_F^2, \quad (4)$$

where $\lambda > 0$ is some regularization parameter, and apply gradient descent to the regularized loss function instead:

$$\min_{X \in \mathbb{R}^{n_1 \times r}, Y \in \mathbb{R}^{n_2 \times r}} f_{\text{reg}}(X, Y) \triangleq f(X, Y) + g(X, Y). \quad (5)$$

For a variety of important problems such as low-rank matrix sensing and matrix completion, it has been established that gradient descent over the regularized loss function, when properly initialized, achieves compelling statistical and computational guarantees.

### 1.1 Why balancing is needed in prior work?

Before we investigate the possibility of a balancing-free procedure (i.e. vanilla gradient descent as in (3)), let us first explain using a heuristic argument why balancing is needed in the prior literatures.

To handle the asymmetric factorization, it is common to stack the two factors into one augmented factor $Z_* \triangleq \begin{bmatrix} X_* \\ Y_* \end{bmatrix} \in \mathbb{R}^{(n_1 + n_2) \times r}$ and then seek to estimate $Z_*$ directly, by rewriting the loss function with respect to the lifted low-rank matrix: $Z_*, Z_*^\top = \begin{bmatrix} X_* X_*^\top \\ Y_* Y_*^\top \end{bmatrix} \in \mathbb{R}^{(n_1 + n_2) \times (n_1 + n_2)}$. It is obvious that the loss function originally with respect to the asymmetric matrix $X_* Y_*^\top$ only constrains the off-diagonal blocks of $Z_*, Z_*^\top$ and not the diagonal ones; correspondingly, the loss function is not (restricted) strongly convex with respect to the augmented factor, unless we appropriately regularize the diagonal blocks. This gives rise to the adoption of the regularization term in (4).

To develop more intuitions regarding why this regularization term (4) may help analysis, consider a toy example of factorizing a rank-one matrix $x, y^\top$, where $f(x, y)$ and $g(x, y)$ respectively are $f(x, y) = \|x y^\top - x, y^\top\|_F^2/2$ and $g(x, y) = (\|x\|_2^2 - \|y\|_2^2)/2$. Figure [1] illustrates the landscape of the unregularized loss function $f(x, y)$ and the regularized loss function $f_{\text{reg}}(x, y)$, respectively, when the arguments are scalar-valued, i.e. $n_1 = n_2 = 1$. One can clearly appreciate the value of the regularizer: $f_{\text{reg}}(x, y)$ becomes strongly convex in the local neighborhood around the global optimum $(1, 1)$. In contrast, the Hessian of the unregularized loss function $f_{\text{reg}}(x, y)$ remains rank deficient along the ambiguity set $\{(x, y) \mid xy = 1\}$, making the analysis less tractable.
1.2 This paper: balancing-free procedure?

This goal of this paper is to understand the effectiveness of vanilla gradient descent when initialized with balanced factors. Indeed, Figure 2 plots the normalized error $\left\| X_t Y_t^\top - M_\star \right\|_F / \left\| M_\star \right\|_F$ for low-rank matrix completion, which aims to recover a low-rank matrix from a subset of its observations, with respect to the iteration count, using either a regularized loss function or an unregularized loss function when initialized by the spectral method. The two iterates converge in almost exactly the same trajectory, suggesting that gradient descent over the unregularized loss function converges almost in the same manner as its regularized counterpart, and perhaps is more natural to use in practice since it eliminates the tuning of the regularization parameters.

![Figure 2: The normalized reconstruction error with and without balancing.](image)

This paper justifies formally that even without explicit balancing in asymmetric low-rank matrix sensing, gradient descent converges linearly towards the global optimum, as long as the initialization is (nearly) balanced. As will be detailed later, our analysis is simple and built on a novel distance metric that directly accounts for the ambiguity due to invertible transformations – in contrast, the ambiguity set reduces to orthonormal transforms when the balancing regularization is present. Our key message is this:

As long as the factors are (nearly) balanced at the initialization, they will stay approximately balanced throughout the trajectory of gradient descent, and therefore no additional regularization is necessary.

1.3 Notation

We use boldface lowercase (resp. uppercase) letters to represent vectors (resp. matrices). We denote $\|x\|_2$ the $\ell_2$ norm of a vector $x$, and $X^\top$, $X^{-1}$, $\|X\|$ and $\|X\|_F$ the transpose, the inverse, the spectral norm and the Frobenius norm of a matrix $X$, respectively. Furthermore, we denote $X^{-\top} = (X^{-1})^\top = (X^\top)^{-1}$ for an invertible matrix $X$. The $k$th largest singular value of a matrix $X$ is denoted by $\sigma_k(X)$. The inner product between two matrices $X$ and $Y$ is defined as $\langle X, Y \rangle = \text{Tr} (Y^\top X)$, where $\text{Tr} (\cdot)$ denotes the trace operator. Denote by $O^{r \times r}$ the set of $r \times r$ orthonormal matrices. In addition, we use $c$ and $C$ with different subscripts to represent positive numerical constants, whose values may change from line to line.

2 Main results

Let the object of interest $M_\star \in \mathbb{R}^{n_1 \times n_2}$ be a rank-$r$ matrix whose compact Singular Value Decomposition (SVD) is given by

$$M_\star = U_\star \Sigma_\star V_\star^\top,$$
Algorithm 1 describes the gradient descent algorithm initialized by the spectral method for minimizing (9). Compared to the Procrustes Flow (PF) algorithm \[6\], which minimizes the regularized loss function in (5), the new algorithm does not include the balancing regularizer \(g\) to the Procrustes Flow (PF) algorithm. The adjoint operator \(A^* : \mathbb{R}^m \to \mathbb{R}^{n_1 \times n_2}\) is defined to be \(A^*(y) = \sum_{i=1}^m y_i A_i\).

\[ X_{t+1} = X_t - \frac{\eta_t}{\|Y_t\|^2} \cdot \sum_{i=1}^m \langle A_i, X_t Y_t^T \rangle - y_i \rangle A_i Y_t \]
\[ Y_{t+1} = Y_t - \frac{\eta_t}{\|X_t\|^2} \cdot \sum_{i=1}^m \langle A_i, X_t Y_t^T \rangle - y_i \rangle A_i^T X_t \]

Output: \(X_T\) and \(Y_T\).

where \(U \in \mathbb{R}^{n_1 \times r}, V \in \mathbb{R}^{n_2 \times r}\) and \(Σ \in \mathbb{R}^{r \times r}\) correspond to the left singular vectors, the right singular vectors and the singular values, respectively. Without loss of generality, we denote the ground truth factors as

\[ X_\star \triangleq U \Sigma_\star^{1/2}, \quad \text{and} \quad Y_\star \triangleq V \Sigma_\star^{1/2}. \] (6)

Let \(\sigma_{\text{max}} \triangleq \sigma_1(M_\star)\) (resp. \(\sigma_{\text{min}} \triangleq \sigma_r(M_\star)\)) be the largest (resp. smallest) nonzero singular value of \(M_\star\). The condition number of \(M_\star\) is therefore defined as \(κ \triangleq \sigma_{\text{max}}/\sigma_{\text{min}}\).

Since the factors are identifiable up to invertible transforms, i.e. \((X_\star, P)(Y_\star, P^{-T}) = X_\star Y_\star^T\) for any invertible matrix \(P \in \mathbb{R}^{r \times r}\), it is natural to measure the distance between two pairs of factors \(Z = [X \ Y] \in \mathbb{R}^{(n_1+n_2) \times r}\) and \(Z_\star \in [X_\star \ Y_\star] \in \mathbb{R}^{(n_1+n_2) \times r}\) via the following function:

\[ \text{dist}(Z, Z_\star) = \min_{P \in \mathbb{R}^{r \times r}, \text{invertible}} \sqrt{\|X P - X_\star\|_F^2 + \|Y P^{-T} - Y_\star\|_F^2}. \] (7)

### 2.1 Low-rank matrix sensing

Low-rank matrix sensing refers to the problem of recovering a low-rank matrix (i.e. \(M_\star\)) from a small number of linear measurements. Specifically, we are given a set of \(m\) measurements as follows

\[ y_i = \langle A_i, M_\star \rangle = \langle A_i, X_\star Y_\star^T \rangle, \quad i = 1, \ldots, m. \] (8)

where \(A_i \in \mathbb{R}^{n_1 \times n_2}\) is the \(i\)th sensing matrix. For convenience, we define \(A : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m\) as an affine transformation from \(\mathbb{R}^{n_1 \times n_2}\) to \(\mathbb{R}^m\), such that \(A(M) = \{\langle A_i, M \rangle\}_{1 \leq i \leq m}\). Consequently, one can compactly write \(y = A(M_\star)\). The adjoint operator \(A^* : \mathbb{R}^m \to \mathbb{R}^{n_1 \times n_2}\) is defined to be \(A^*(y) = \sum_{i=1}^m y_i A_i\).

To recover the low-rank matrix, a natural choice is to minimize the least-squares loss function

\[ f(X, Y) \triangleq \frac{1}{2} \|y - A(X Y^T)\|_2^2. \] (9)

Algorithm 1 describes the gradient descent algorithm initialized by the spectral method for minimizing (9). Compared to the Procrustes Flow (PF) algorithm \[6\], which minimizes the regularized loss function in (5), the new algorithm does not include the balancing regularizer \(g\) to the Procrustes Flow (PF) algorithm. To understand the performance of Algorithm 1, we adopt a standard assumption on the sensing operator \(A\), namely the Restricted Isometry Property (RIP).

1More rigorously, we should write inf instead of min in the definition of dist(·, ·). However, as we will soon see, in the cases we care about, the minimum can always be achieved by some invertible matrix \(P\)
Definition 1 (RIP) The operator $A(\cdot)$ is said to satisfy the rank-$r$ RIP with a constant $\delta_r \in [0, 1)$, if
\[(1 - \delta_r) \|M\|_F^2 \leq \|A(M)\|_2^2 \leq (1 + \delta_r) \|M\|_F^2\]
holds for all matrices $M \in \mathbb{R}^{n_1 \times n_2}$ of rank at most $r$.

It is well-known that many measurement ensembles satisfy the RIP property\textsuperscript{[11]}. 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 as long as $m$ is on the order of $(n_1 + n_2)r/\delta_r^2$.

Armed with the RIP, we have the following theoretical guarantee for the local convergence of Algorithm\textsuperscript{[1]}

Theorem 1 Suppose that $A(\cdot)$ satisfies the RIP with $\delta_{2r} \leq c$ for some sufficiently small constant $c$. Let $Z_0 \triangleq \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}$ be any initialization point that satisfies
\[
\min_{R \in O_r \times r} \|Z_0 R - Z_*\|_F \leq c_0 \frac{1}{\kappa^{3/2}} \sigma_r(X_*),
\]
for some small enough constant $c_0 > 0$. Then there exist some constant $c_1 > 0$ such that at as long as $\eta_t = \eta = c_1$, the iterates of GD (cf. (10)) satisfy
\[
\text{dist} (Z_t, Z_*) \leq \left(1 - \frac{\eta}{2\kappa c}\right)^t \text{dist} (Z_0, Z_*).
\]

In words, Theorem\textsuperscript{[1]} reveals that if the initialization $Z_0$ lands in a basin of attraction given by (11), then Algorithm\textsuperscript{[1]} converges linearly with a constant step size. To reach $\epsilon$-accuracy, i.e. $\text{dist} (Z_t, Z_*) \leq \epsilon$, it takes an order of $\kappa \log(1/\epsilon)$ iterations, which is order-wise equivalent to the regularized PF algorithm proposed in \textsuperscript{[6]}. Comparing to \textsuperscript{[6]}, which requires $\delta_{4r} \leq c$, Theorem\textsuperscript{[1]} only requires a weaker assumption $\delta_{2r} \leq c$. However, the basin of attraction allowed by Theorem\textsuperscript{[1]} is smaller than that in \textsuperscript{[6]}, which is specified by $\min_{R \in O_r \times r} \|Z_0 R - Z_*\|_F \leq c_0 \sigma_r(X_*)$.

We are still in need of finding a good initialization that obeys (11). In general, one could initialize with the balanced factors of the output after running multiple iterations of projected gradient descent (over the low-rank matrix), i.e.
\[
M_{r+1} = \mathcal{P}_r \left( M_r - \frac{1}{m} \sum_{i=1}^m (A_i, M_r) - y_i) A_i \right),
\]
where $\mathcal{P}_r(\cdot)$ is the Euclidean projection operator to set of rank-$r$ matrices. The spectral initialization specified in Algorithm\textsuperscript{[1]} can be regarded as the output at the first iteration, initialized at zero $M_0 = 0$. Based on \textsuperscript{[12, 6]}, the iterates satisfy
\[
\min_{R \in O_r \times r} \|Z_r R - Z_*\|_F \leq c_2(2\delta_{4r})^t \frac{\|M_r\|_F}{\sigma_{\min}(X_*)}
\]
for some constant $c_2$. Thus, to achieve the required initialization condition (11) using the spectral method specified in Algorithm\textsuperscript{[1]} (which corresponds to setting $\tau = 1$ in (12)), we need
\[
\delta_{4r} \leq c_2 \frac{1}{\kappa^{3/2}} \frac{\sigma_{\min}}{\|M_r\|_F}.
\]
Alternatively, if we allow multiple iterations of (12) as suggested by \textsuperscript{[6]}, we can still set $\delta_{4r} \leq \delta_\epsilon$ for a sufficiently small constant $\delta_\epsilon$, by running at least
\[
\tau \geq c_1 \log \left( \frac{\kappa^{3/2}}{\sigma_{\min}} \frac{\|M_r\|_F}{\sigma_{\min}} \right) / \log (\delta_\epsilon^{-1}) = c_2 \log (\kappa r) / \log (1/\delta_\epsilon)
\]
iters of projected gradient descent for initialization, which matches the requirement in \textsuperscript{[6]}.
3 Related Work

Low-rank matrix estimation has been extensively studied in recent years [3,4], due to its broad applicability in collaborative filtering, imaging science, and machine learning, to name a few. Convex relaxation approaches based on nuclear norm minimization are among the first set of algorithms with near-optimal statistical guarantees [2,13,14,15,16,17,18,19], however, their computational costs are often prohibitive in practice.

To cope with the computational challenges, a popular approach in practice is to invoke low-rank matrix factorization popularized by Burer and Monteiro [5] and then apply first-order methods such as gradient descent directly over the factors to recover the underlying low-rank structure. This approach is demonstrated to possess near-optimal statistical and computational guarantees in a variety of low-rank matrix recovery problems, including but not limited to [6,20,21,22,23,24,25,26,27,28]. The readers are referred to the recent overview [29] for additional references.

To the best of our knowledge, the balancing regularization term (4) was first introduced in [6] to deal with asymmetric matrix factorization, and has become a standard approach to deal with asymmetric low-rank matrix estimation [7,8,10,30,31,32]. A major benefit of adding the regularization term is to reduce the ambiguity set from invertible transforms to orthonormal transforms, so that the distance defined in (7) is minimized over $P \in O_{r \times r}$. For the special rank-one matrix recovery problem, there are some evidence in the prior literature that a balancing regularization is not needed, for example, Ma et al. [24] established that vanilla gradient descent works for blind deconvolution at a near-optimal sample complexity with spectral initialization. In [33], the trajectory of gradient descent is studied for asymmetric matrix factorization with an infinitesimal and diminishing step size; in contrast, we consider the case when the step size is constant for low-rank matrix estimation with incomplete observations. Finally, very recently, [34] also studied low-rank matrix sensing using a nonsmooth formulation without the balancing regularization via subgradient descent.

Finally, we remark that a similar regularization term (4) is also adopted when analyzing the optimization landscape of low-rank matrix estimation, e.g. [35,36,37,38]. Without such a regularization term, the landscape of matrix factorization no longer possesses the intriguing property “all saddle points are strict saddle” and therefore one cannot invoke the theory presented in [39] to argue the global convergence of gradient descent using an unregularized loss function. Our work partially bridges this gap and suggests the benign behavior of gradient descent even in the absence of local strong convexity.

4 Proof of Theorem 1

In this section, we provide the proof of Theorem 1. We first discuss some basic properties of aligning two low-rank factors via an invertible transformation. Then we prove a similar result for a warm-up case of low-rank matrix factorization. In the end, viewing matrix sensing as a perturbed version of low-rank matrix factorization helps us finish the proof of Theorem 1.

4.1 Alignment via invertible transformations

Fix a matrix $Z = \begin{bmatrix} X \\ Y \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times r}$. We define the optimal alignment matrix $Q$ between $Z$ and $Z^*$ as

$$Q \triangleq \arg \min_{P \in \mathbb{R}^{r \times r}} \|XP - X^*\|_F^2 + \|YP^\top - Y^*\|_F^2,$$

whenever the minimum is attained. As we will soon see, for the iterates $\{Z_t\}_{t \geq 0}$ generated by Algorithm 1 the optimal alignment matrix is always well defined. Furthermore, we call $Z$ and $Z^*$ aligned if the corresponding optimal alignment matrix is just the identity matrix $I_r$. Below we provide some basic understandings of this alignment matrix.

The following lemma provides a sufficient condition for the existence of the optimal alignment matrix.

**Lemma 1** Fix some matrix $Z = \begin{bmatrix} X \\ Y \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times r}$. Suppose that there exists a matrix $P \in \mathbb{R}^{r \times r}$ with $1/2 \leq \sigma_r(P) \leq \sigma_1(P) \leq 3/2$ such that

$$\max \left\{ \|XP - X^*\|_F, \|YP^\top - Y^*\|_F \right\} \leq \delta \leq \sigma_r(X^*)/80.$$

(13)
Then the optimal alignment matrix \( Q \in \mathbb{R}^{r \times r} \) between \( Z \) and \( Z^* \) exists. In addition, the matrix \( Q \) satisfies

\[
\|P - Q\| \leq \|P - Q\|_F \leq \frac{5\delta}{\sigma_r(X^*)}.
\]

Next, the lemma below presents a necessary condition for \( Q \) to be the optimal alignment matrix between \( Z \) and \( Z^* \).

**Lemma 2** Let \( Z \) and \( Z^* \) be any two matrices. Suppose that the optimal alignment matrix \( Q \) between \( Z \) and \( Z^* \) exists. Then we have

\[
\tilde{X}^\top (\tilde{X} - X^*) = (\tilde{Y} - Y^*)\top \tilde{Y},
\]

where \( \tilde{X} = XQ \) and \( \tilde{Y} = YQ^{-\top} \) are two matrices after the alignment.

Both lemmas provide basic understandings of the solution to the alignment problem with invertible transformations, which can be regarded as a generalization of the classical orthogonal Procrustes problem that only considers orthonormal transformations. Clearly, this generalized problem is more involved and our work provides some basic understandings.

### 4.2 A warm-up: low-rank matrix factorization

We consider the following minimization problem for low-rank matrix factorization

\[
f_{\text{MF}}(X, Y) = \frac{1}{2} \|XY\top - M_+\|_F^2,
\]

where \( X \in \mathbb{R}^{n_1 \times r} \) and \( Y \in \mathbb{R}^{n_2 \times r} \). The gradient descent updates with an initialization \((X_0, Y_0)\) can be written as

\[
X_{t+1} = X_t - \frac{\eta}{\sigma_{\text{max}}} \nabla_X f_{\text{MF}}(X_t, Y_t) = X_t - \frac{\eta}{\sigma_{\text{max}}} (X_t Y_t\top - M_+)Y_t;
\]

\[
Y_{t+1} = Y_t - \frac{\eta}{\sigma_{\text{max}}} \nabla_Y f_{\text{MF}}(X_t, Y_t) = Y_t - \frac{\eta}{\sigma_{\text{max}}} (X_t Y_t\top - M_+)\top X_t.
\]

Here \( \eta > 0 \) stands for the step size. We have the following theorem regarding the performance of (15), which parallels Theorem 2

**Theorem 2** Let \( Z_0 = \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} \in \mathbb{R}^{(n_1 + n_2) \times r} \) be any initialization point that satisfies

\[
\min_{R \in \mathbb{O}^{r \times r}} \|Z_0 R - Z^*\|_F \leq c_0 \frac{1}{\kappa^{1/2}} \sigma_r(X^*)
\]

for some sufficiently small constant \( c_0 > 0 \). Then setting the step size \( \eta > 0 \) to be some sufficiently some constant, the iterates of GD (cf. (15)) satisfy

\[
\text{dist}(Z_t, Z^*) \leq \left(1 - \frac{\eta}{\sigma_{\text{max}}} \right)^t \text{dist}(Z_0, Z^*).$

To prove Theorem 2 we need the following properties of the gradients of \( f_{\text{MF}}(X, Y) \); the proofs are deferred to the appendix.

**Lemma 3 (Gradient dominance)** Suppose that \( Z = \begin{bmatrix} X \\ Y \end{bmatrix} \in \mathbb{R}^{(n_1 + n_2) \times r} \) is aligned with \( Z^* \), i.e.

\[
I_r = \arg\min_{P \in \mathbb{O}^{r \times r}} \|XP - X^*_+\|_F^2 + \|YP\top - X^*_+\|_F^2.
\]

Then we have

\[
\langle X - X^*, (XY\top - M_+)Y \rangle \geq \|Y(X - X^*)\top\|_F^2 - \frac{1}{4} \|X - X^*\|_F^4,
\]

Similarly, one has

\[
\langle Y - Y^*, (XY\top - M_+)\top X \rangle \geq \|X(Y - Y^*)\top\|_F^2 - \frac{1}{4} \|Y - Y^*\|_F^4.
\]
Lemma 4 (Smoothness) Suppose that \( \|Y - Y_*\| \leq \sigma_1 (Y_*) / 4 \), then one has
\[
\| (XY^T - M_*) Y \|_F \leq \frac{3}{2} \sigma_1 (Y_*) \cdot (\| (X - X_*) Y^T \|_F + \| X (Y - Y_*)^T \|_F + \| X - X_* \|_F \| Y - Y_* \|_F).
\]
Similarly, with the proviso that \( \|X - X_*\| \leq \sigma_1 (X_*) / 4 \), one has
\[
\| (XY^T - M_*)^T X \|_F \leq \frac{3}{2} \sigma_1 (X_*) \cdot (\| (X - X_*) Y^T \|_F + \| X (Y - Y_*)^T \|_F + \| X - X_* \|_F \| Y - Y_* \|_F).
\]

4.3 Proof of Theorem 2
With the help of Lemmas 1–4, we are in position to establish Theorem 2. Denote by \( \hat{R} \in \mathbb{R}^{r \times r} \) the best rotation matrix between \( Z_0 \) and \( Z_* \), that is
\[
\hat{R} \triangleq \underset{R \in O(r)}{\text{argmin}} \| Z_0 R - Z_* \|_F.
\]
Combine the assumption of initialization (cf. (16)) and Lemma 1 to see that
\[
Q_0 \triangleq \underset{P \in \mathbb{R}^{r \times r}}{\text{argmin}} \sqrt{\| XP - X_* \|_F^2 + \| Y P^{-T} - Y_* \|_F^2}
\]
exists and in addition, one has
\[
\|Q_0 - \hat{R}\| \leq \frac{5c_0}{\kappa^{3/2}} \leq \frac{1}{400\sqrt{\kappa}}
\]
as long as \( c_0 > 0 \) is sufficiently small.

The remaining proof is inductive in nature. In particular, we aim at proving the following induction hypotheses.

1. The optimal alignment matrix \( Q_t \) between \( Z_t \) and \( Z_* \) exists.
2. The distance between \( Z_t \) and \( Z_* \) obeys
\[
dist (Z_t, Z_*) \leq \left(1 - \frac{\eta}{50\kappa}\right)^t \cdot dist (Z_0, Z_*).
\]
3. The optimal alignment matrix \( Q_t \) is nearly a rotation matrix in the sense that
\[
\|Q_t - \hat{R}\| \leq \frac{1}{400\sqrt{\kappa}}.
\]

It is straightforward to check that these three claims hold for \( t = 0 \). In what follows, we shall assume that the induction hypotheses hold for all iterations up to the \( t \)th iteration and intend to establish that they continue to hold for the \((t+1)\)th iteration.

1. We begin with demonstrating the existence of \( Q_{t+1} \). In view of the gradient update rule (15), we have
\[
X_{t+1} Q_t = X_t Q_t - \frac{\eta}{\sigma_{\max}} (X_t Y_t^T - M_*) Y_t Q_t = \tilde{X}_t - \frac{\eta}{\sigma_{\max}} (\tilde{X}_t \tilde{Y}_t^T - M_*) \tilde{Y}_t (Q_t^T Q_t),
\]
\[
Y_{t+1} Q_t^T = Y_t Q_t^T - \frac{\eta}{\sigma_{\max}} (X_t Y_t^T - M_*)^T X_t Q_t^{-T} = \tilde{Y}_t - \frac{\eta}{\sigma_{\max}} (X_t Y_t^T - M_*)^T \tilde{X}_t (Q_t^T Q_t)^{-1},
\]
where we denote
\[
\tilde{X}_t \triangleq X_t Q_t \quad \text{and} \quad \tilde{Y}_t \triangleq Y_t Q_t^{-T}.
\]
As a result, one has the following equality
\[
\|X_{t+1} Q_t - X_*\|_F^2 + \|Y_{t+1} Q_t^{-T} - Y_*\|_F^2
\]
\[
= \left\| \tilde{X}_t - X_* - \frac{\eta}{\sigma_{\max}} (\tilde{X}_t \tilde{Y}_t^T - M_*) \tilde{Y}_t A_t \right\|_F^2 + \left\| \tilde{Y}_t - Y_* - \frac{\eta}{\sigma_{\max}} (\tilde{X}_t \tilde{Y}_t^T - M_*)^T \tilde{X}_t A_t^{-1} \right\|_F^2,
\]
\[
\begin{align*}
\approx & \alpha_1 \\
\approx & \alpha_2
\end{align*}
\]
where we have denoted \( \Lambda_t \triangleq Q_t^T Q_t \). By virtue of the third induction hypothesis, namely \( \| Q_t - \hat{R} \| \leq 1/(400\sqrt{\kappa}) \), it is easy to check that \( \| \Lambda_t - I_r \| \leq 1/(180\sqrt{\kappa}) \triangleq \zeta \). Expand \( \alpha_1 \) to obtain

\[
\alpha_1 = \left\| \tilde{X}_t - X_\ast \right\|_F^2 + \left( \frac{\eta}{\sigma_{\text{max}}} \right)^2 \left\| \left( \tilde{X}_t \tilde{Y}_t^T - M_\ast \right) \tilde{Y}_t \Lambda_t \right\|_F^2 - 2 \frac{\eta}{\sigma_{\text{max}}} \left\langle \tilde{X}_t - X_\ast, \left( \tilde{X}_t \tilde{Y}_t^T - M_\ast \right) \tilde{Y}_t \Lambda_t \right\rangle.
\]

Similarly, we can decompose \( \alpha_2 \) into

\[
\alpha_2 = \left\| \tilde{Y}_t - Y_\ast \right\|_F^2 + \left( \frac{\eta}{\sigma_{\text{max}}} \right)^2 \left\| \left( \tilde{X}_t \tilde{Y}_t^T - M_\ast \right)^T \tilde{X}_t \Lambda_t^{-1} \right\|_F^2 - 2 \frac{\eta}{\sigma_{\text{max}}} \left\langle \tilde{Y}_t - Y_\ast, \left( \tilde{X}_t \tilde{Y}_t^T - M_\ast \right)^T \tilde{X}_t \Lambda_t^{-1} \right\rangle.
\]

We intend to apply Lemma 3 to lower bound the terms \( \gamma_1 \) and \( \gamma_2 \) and apply Lemma 4 to upper bound \( \beta_1 \) and \( \beta_2 \). First, since \((\tilde{X}_t, \tilde{Y}_t)\) is aligned with \((X_\ast, Y_\ast)\), we can invoke Lemma 3 to see that

\[
\gamma_1 \geq \left\langle \tilde{X}_t - X_\ast, \left( \tilde{X}_t \tilde{Y}_t^T - M_\ast \right) \tilde{Y}_t \right\rangle - \left\langle \tilde{X}_t - X_\ast, \left( \tilde{X}_t \tilde{Y}_t^T - M_\ast \right) \tilde{Y}_t (\Lambda_t - I_r) \right\rangle
\]

\[
\geq \left\| \tilde{Y}_t (\tilde{X}_t - X_\ast)^\top \right\|_F^2 - \frac{1}{4} \left\| \tilde{X}_t - X_\ast \right\|_F^4 - \left\| \Lambda_t - I_r \right\| \left\| \tilde{Y}_t \right\|_F \left\| \tilde{X}_t - X_\ast \right\|_F
\]

\[
\geq \left\| \tilde{Y}_t (\tilde{X}_t - X_\ast)^\top \right\|_F^2 - \frac{1}{4} \sigma_{\text{min}} \left\| \tilde{X}_t - X_\ast \right\|_F^2 - \zeta \left\| \tilde{X}_t \right\|_F \left\| \tilde{X}_t - X_\ast \right\|_F
\].

Here the last line follows from the bound \( \| \Lambda_t - I_r \| \leq \zeta \) and the second induction hypothesis, i.e.

\[
\left\| \tilde{X}_t - X_\ast \right\|_F^2 \leq \text{dist}^2 (Z_t, Z_\ast) \leq \text{dist}^2 (Z_0, Z_\ast) \leq \frac{1}{100} \sigma_{\text{min}}.
\]

The last term in (17) can be further bounded via Lemma 4 as

\[
\zeta \left\| \tilde{X}_t \tilde{Y}_t^T - M_\ast \right\|_F \left\| \tilde{X}_t - X_\ast \right\|_F
\]

\[
\leq \frac{3\zeta}{2} \sqrt{\sigma_{\text{max}}} \left( \left\| \left( \tilde{X}_t - X_\ast \right) \tilde{Y}_t^\top \right\|_F^2 + \left\| \tilde{X}_t \left( \tilde{Y}_t - Y_\ast \right) \right\|_F^2 + \left\| \tilde{X}_t - X_\ast \right\|_F \left\| \tilde{Y}_t - Y_\ast \right\|_F \right) \left\| \tilde{X}_t - X_\ast \right\|_F
\]

\[
\leq \frac{9\zeta \sqrt{\kappa}}{2} \left\| \tilde{X}_t \right\|_F \frac{\sqrt{\sigma_{\text{min}}}}{3} \left\| \tilde{X}_t - X_\ast \right\|_F + \frac{9\zeta \sqrt{\kappa}}{2} \left\| \tilde{X}_t \right\|_F \frac{\sqrt{\sigma_{\text{min}}}}{3} \left\| \tilde{X}_t - X_\ast \right\|_F
\]

\[
\leq \frac{81\zeta^2 \kappa}{8} \left\| \tilde{X}_t - X_\ast \right\|_F^2 + \frac{81\zeta^2 \kappa}{8} \left\| \tilde{X}_t \right\|_F \left\| \tilde{Y}_t - Y_\ast \right\|_F^2 + \sigma_{\text{min}} \frac{\sqrt{\sigma_{\text{min}}}}{8} \left\| \tilde{X}_t - X_\ast \right\|_F^2,
\]

where the last inequality arises since \( ab \leq (a^2 + b^2)/2 \) and

\[
\frac{3\zeta}{2} \sqrt{\sigma_{\text{max}}} \left\| \tilde{Y}_t - Y_\ast \right\|_F \leq \frac{3\zeta}{2} \sqrt{\sigma_{\text{max}}} \text{dist} (Z_0, Z_\ast) \leq \frac{3\zeta}{2} \sqrt{\sigma_{\text{max}}} c_0 \frac{1}{\kappa^{3/2}} \sqrt{\sigma_{\text{min}}} \leq \frac{\sigma_{\text{min}}}{72}
\]

as long as \( c_0 \) is sufficiently small. Combine the above two bounds to reach

\[
\gamma_1 \geq \left( 1 - \frac{81\zeta^2 \kappa}{8} \right) \left\| \tilde{Y}_t \left( \tilde{X}_t - X_\ast \right) \right\|_F^2 - \frac{81\zeta^2 \kappa}{8} \left\| \tilde{X}_t \left( \tilde{Y}_t - Y_\ast \right) \right\|_F^2 - \frac{\sigma_{\text{min}}}{7} \left\| \tilde{X}_t - X_\ast \right\|_F^2.
\]

Similarly \( \gamma_2 \) can be lower bounded as

\[
\gamma_2 \geq \left( 1 - \frac{81\zeta^2 \kappa}{8} \right) \left\| \tilde{X}_t \left( \tilde{Y}_t - Y_\ast \right) \right\|_F^2 - \frac{81\zeta^2 \kappa}{8} \left\| \tilde{Y}_t \left( \tilde{X}_t - X_\ast \right) \right\|_F^2 - \frac{\sigma_{\text{min}}}{7} \left\| \tilde{Y}_t - Y_\ast \right\|_F^2.
\]
which together with the bound on \( \gamma_1 \) implies

\[
\gamma_1 + \gamma_2 \geq \left( 1 - \frac{81c^2 \kappa}{4} \right) \left( \left\| \bar{Y}_t (\bar{X}_t - X_*)^\top \right\|_F^2 + \left\| \bar{X}_t (\bar{Y}_t - Y_*)^\top \right\|_F^2 \right) - \frac{\sigma_{\min}}{t} \text{dist}^2 (Z_t, Z_*)
\]

\[
\geq \frac{3}{4} \left( \left\| \bar{Y}_t (\bar{X}_t - X_*)^\top \right\|_F^2 + \left\| \bar{X}_t (\bar{Y}_t - Y_*)^\top \right\|_F^2 \right) - \frac{\sigma_{\min}}{t} \text{dist}^2 (Z_t, Z_*) ,
\]

where we plug in the definition of \( \zeta = 1/(180\sqrt{\kappa}) \).

Now we move on to controlling \( \beta_1 \) and \( \beta_2 \). Recognizing that \( \| A_t \| \leq 2 \), one has

\[
\beta_1 \leq 4 \left\| (\bar{X}_t \bar{Y}_t - M_*) \bar{Y}_t \right\|_F^2
\]

\[
\leq 9 \sigma_{\max} \left( \left\| (\bar{X}_t - X_*) \bar{Y}_t^\top \right\|_F^2 + \left\| \bar{X}_t (\bar{Y}_t - Y_*)^\top \right\|_F^2 + \left\| \bar{X}_t - X_* \right\|_F \| \bar{Y}_t - Y_* \|_F^2 \right) ,
\]

where the second line follows from Lemma 4. Apply the elementary inequality \((a + b + c)^2 \leq 3(a^2 + b^2 + c^2)\) to see that

\[
\beta_1 \leq 27 \sigma_{\max} \left( \left\| (\bar{X}_t - X_*) \bar{Y}_t^\top \right\|_F^2 + \left\| \bar{X}_t (\bar{Y}_t - Y_*)^\top \right\|_F^2 \right) + 27 \epsilon_0 \frac{\sigma_{\max} \sigma_{\min}}{\kappa^3} \| \bar{Y}_t - Y_* \|_F^2 .
\]

Here the second line relies on the fact that \( \| \bar{Y}_t - Y_* \|_F^2 \leq \text{dist}^2 (Z_0, Z_*) \leq \frac{\sigma_{\min}}{\kappa^3} \). Similarly, one can bound \( \beta_2 \) as

\[
\beta_2 \leq 27 \sigma_{\max} \left( \left\| (\bar{X}_t - X_*) \bar{Y}_t^\top \right\|_F^2 + \left\| \bar{X}_t (\bar{Y}_t - Y_*)^\top \right\|_F^2 \right) + 27 \epsilon_0 \frac{\sigma_{\max} \sigma_{\min}}{\kappa^3} \| Y_t - Y_* \|_F^2 ,
\]

which in conjunction with the bound on \( \beta_1 \) yields

\[
\beta_1 + \beta_2 \leq 54 \sigma_{\max} \left( \left\| (\bar{X}_t - X_*) \bar{Y}_t^\top \right\|_F^2 + \left\| \bar{X}_t (\bar{Y}_t - Y_*)^\top \right\|_F^2 \right) + 27 \epsilon_0 \frac{\sigma_{\max} \sigma_{\min}}{\kappa^3} \text{dist}^2 (Z_t, Z_*) .
\]

Collect all the bounds on \( \alpha_1 \) and \( \alpha_2 \) to arrive at

\[
\| X_{t+1} Q_t - X_* \|_F^2 + \| Y_{t+1} Q_t^\top - Y_* \|_F^2 \\
\leq \left( 1 + \frac{27 \epsilon_0 \eta}{\kappa^4} \right) \text{dist} (Z_t, Z_*)^2 + 27 \epsilon_0 \left( \left\| (\bar{X}_t - X_*) \bar{Y}_t^\top \right\|_F^2 + \left\| \bar{X}_t (\bar{Y}_t - Y_*)^\top \right\|_F^2 \right)
\]

\[
- 2 \frac{\eta}{\sigma_{\max}} \left\{ \left( \left\| (\bar{X}_t - X_*) \bar{Y}_t^\top \right\|_F^2 + \left\| \bar{X}_t (\bar{Y}_t - Y_*)^\top \right\|_F^2 \right) - \frac{\sigma_{\min}}{t} \text{dist} (Z_t, Z_*)^2 \right\}
\]

\[
= \left( 1 + \frac{27 \epsilon_0 \eta}{\kappa^4} + \frac{\eta}{3.5 \kappa} \right) \text{dist} (Z_t, Z_*)^2 + 27 \epsilon_0 \left( \left\| (\bar{X}_t - X_*) \bar{Y}_t^\top \right\|_F^2 + \left\| \bar{X}_t (\bar{Y}_t - Y_*)^\top \right\|_F^2 \right)
\]

\[
\leq \left( 1 + \frac{\eta}{3 \kappa} \right) \text{dist} (Z_t, Z_*)^2 - \frac{\eta}{\sigma_{\max}} \left( \left\| \bar{X}_t - X_* \right\|_F^2 + \left\| \bar{Y}_t - Y_* \right\|_F^2 \right) ,
\]

where the last line follows as long as \( \eta \leq 1/24 \). Furthermore, since \( \sigma_{\max} \bar{Y}_t \geq \sigma_{\min}/2 \) and \( \sigma_{\max} \bar{X}_t \geq \sigma_{\min}/2 \), we have

\[
\| X_{t+1} Q_t - X_* \|_F^2 + \| Y_{t+1} Q_t^\top - Y_* \|_F^2 \leq \left( 1 - \frac{\eta}{24 \kappa} \right) \text{dist} (Z_t, Z_*)^2 .
\]
2. The second induction hypothesis for the \((t + 1)\)th iteration follows immediately from the above proof. Since \(Q_{t+1}\) exists, by definition, one has

\[
\text{dist} \left( Z_{t+1}, Z_* \right) = \sqrt{\|X_{t+1}Q_{t+1} - X_*\|_F^2 + \|Y_{t+1}Q_{t+1}^\top - X_*\|_F^2} \\
\leq \sqrt{\|X_{t+1}Q_t - X_*\|_F^2 + \|Y_{t+1}Q_t^\top - X_*\|_F^2} \\
\leq \left( 1 - \frac{\eta}{50\kappa} \right) \text{dist} \left( Z_t, Z_* \right).
\]

3. It remains to show the last induction hypothesis, namely \(\|Q_{t+1} - \hat{R}\| \leq 1/(400\sqrt{\kappa})\). In view of (19), one has \(\max\{\|X_{t+1}Q_t - X_*\|_F^2, \|Y_{t+1}Q_t^\top - Y_*\|_F^2\} \leq \text{dist} \left( Z_t, Z_* \right)\). Invoke Lemma 4 again to arrive at

\[
\|Q_{t+1} - Q_t\| \leq \frac{5}{\sigma_r(X_*)} \text{dist} \left( Z_t, Z_* \right) \\
\leq \frac{5}{\sigma_r(X_*)} \left( 1 - \frac{\eta}{50\kappa} \right) c_0 \frac{1}{\kappa^{3/2}} \sigma_r(X_*) \\
\leq 5c_0 \left( 1 - \frac{\eta}{50\kappa} \right) \frac{1}{\kappa^{3/2}}.
\]

Hence, by the triangle inequality and the telescoping sum, we obtain

\[
\left\| Q_{t+1} - \hat{R} \right\| \leq \sum_{s=0}^{t} \| Q_{s+1} - Q_s \| + \| Q_0 - \hat{R} \| \\
\leq 5c_0 \sum_{s=0}^{t} \left( 1 - \frac{\eta}{50\kappa} \right) s \frac{1}{\kappa^{3/2}} + 5c_0 \frac{1}{\kappa^{3/2}} \\
< 5c_0 \sum_{s=0}^{\infty} \left( 1 - \frac{\eta}{50\kappa} \right) s \frac{1}{\kappa^{3/2}} + 5c_0 \frac{1}{\kappa^{3/2}} \\
= 5c_0 \frac{50\kappa}{\eta} \frac{1}{\kappa^{3/2}} + 5c_0 \frac{1}{\kappa^{3/2}} \\
\leq \frac{1}{400\sqrt{\kappa}}.
\]

as long as \(c_0\) is small enough and \(\eta\) is some constant.

These finish the induction step and the proof is then completed.

### 4.4 Analysis for matrix sensing

We now extend the techniques used in the proof of Theorem 2 to the matrix sensing case by leveraging the RIP. Suppose that the initialization \(Z_0\) satisfies the condition (11). By a similar argument as in [6], it is sufficient to consider the following update rule:

\[
X_{t+1} = X_t - \frac{\eta}{\sigma_{\max}} \left[ A^*A(X_tY_t^\top - M_*) \right] Y_t; \\
Y_{t+1} = Y_t - \frac{\eta}{\sigma_{\max}} \left[ A^*A(X_tY_t^\top - M_*) \right]^\top X_t.
\]

Compared with the update rule (15) for low-rank matrix factorization, the update rule for matrix sensing differs by the operation of \(A^*A\) when forming the gradient. Therefore, we expect that GD has similar behaviors as earlier as long as the operator \(A^*A\) behaves as a near isometry on low-rank matrices. This can be supplied by the following consequence of the RIP.

**Lemma 5** Suppose that \(A\) satisfies 2r-RIP with a constant \(\delta_{2r}\). Then, for all matrices \(M_1\) and \(M_2\) of rank at most \(r\), we have

\[
|\langle A(M_1), A(M_2) \rangle - \langle M_1, M_2 \rangle| \leq \delta_{2r} \|M_1\|_F \|M_2\|_F.
\]
Equivalently, we can write this as
\[
\| (A^* A - I) (M_1) M_2^T \| \leq \delta_2 r \| M_1 \|_F \| M_2 \|_F .
\]
A simple consequence is that for any \( A \in \mathbb{R}^{n \times r} \)
\[
\| (A^* A - I) \|_F \leq \delta_2 r \| A \| .
\]
Similar to before, we denote \( \tilde{X}_t = X_t Q_t \) and \( \tilde{Y}_t = Y_t Q_t^{-T} \), which are aligned with \( (X_*, Y_*) \). With this notation in place, we can rewrite the update rule as
\[
X_{t+1} Q_t = \tilde{X}_t - \frac{\eta}{\sigma_{\text{max}}} [A^* A (X_t Y_t^T - M_*)] \tilde{Y}_t \Lambda_t ,
\]
\[
Y_{t+1} Q_t^{-T} = \tilde{Y}_t - \frac{\eta}{\sigma_{\text{max}}} [A^* A (X_t Y_t^T - M_*)]^T \tilde{X}_t \Lambda_t^{-1} .
\]
where we recall \( \Lambda_t = Q_t^T Q_t \). By the definition of the distance function, we further obtain
\[
dist (Z_{t+1}, Z_t)^2 \leq \| X_{t+1} Q_t - X_* \|_F^2 + \| Y_{t+1} Q_t^{-T} - Y_* \|_F^2
\]
\[
= \| \tilde{X}_t - \frac{\eta}{\sigma_{\text{max}}} [A^* A (X_t Y_t^T - M_*)] \tilde{Y}_t \Lambda_t - X_* \|_F^2 + \| \tilde{Y}_t - \frac{\eta}{\sigma_{\text{max}}} [A^* A (X_t Y_t^T - M_*)]^T \tilde{X}_t \Lambda_t^{-1} - Y_* \|_F^2
\]
\[
= \| \tilde{X}_t - X_* \|_F^2 + \| \tilde{Y}_t - Y_* \|_F^2
\]
\[
+ \left( \frac{\eta}{\sigma_{\text{max}}} \right)^2 \left( \| [A^* A (X_t Y_t^T - M_*)] \tilde{Y}_t \Lambda_t \|_F^2 + \| [A^* A (X_t Y_t^T - M_*)]^T \tilde{X}_t \Lambda_t^{-1} \|_F^2 \right)
\]
\[
- \frac{2 \eta}{\sigma_{\text{max}}} \left( \tilde{X}_t - X_* \cdot [A^* A (X_t Y_t^T - M_*)] \tilde{Y}_t \Lambda_t \right) + \left( \tilde{Y}_t - Y_* \cdot [A^* A (X_t Y_t^T - M_*)]^T \tilde{X}_t \Lambda_t^{-1} \right) \right) .
\]
From the high level, the four terms \( \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\gamma}_1 \) and \( \tilde{\gamma}_2 \) are the perturbed versions of \( \beta_1, \beta_2, \gamma_1 \) and \( \gamma_2 \) in Section 4.3 respectively.

For the first term, we have
\[
\sqrt{\beta_1} - \sqrt{\beta_1} \leq \left( \| A^* A (X_t Y_t^T - M_*) \| \right) \tilde{Y}_t \Lambda_t - (X_t Y_t^T - M_*) \tilde{Y}_t \Lambda_t \|_F
\]
\[
\leq \left( \| (A^* A - I) (X_t Y_t^T - M_*) \| \right) \tilde{Y}_t \Lambda_t \|_F
\]
\[
\leq \left( \| (A^* A - I) \| \tilde{X}_t - X_* \|_F \right) \| \tilde{Y}_t \Lambda_t \|_F
\]
\[
\leq \delta_2 r \left( \| \tilde{X}_t - X_* \|_F \right) \| \tilde{Y}_t \|_F
\]
\[
\leq 4 \delta_2 r \| X_t - X_* \|_F \| \tilde{Y}_t \|_F + \| \tilde{X}_t \|_F \| \tilde{Y}_t - Y_* \|_F + \| \tilde{X}_t - X_* \|_F \| \tilde{Y}_t \|_F + \| \tilde{Y}_t - Y_* \|_F \|_F .
\]
Here, the first (i) and second (ii) inequalities follow from the triangle inequality. The third one (iii) uses Lemma 5 and the last relation (iv) depends on \( \| \Lambda_t \| \leq 2 \) and \( \| Y_t \| \leq 2 \sqrt{\sigma_{\text{max}}} \). Comparing (22) with (18) reveals that \( \tilde{\beta}_1 - \beta_1 \) constitutes a small perturbation to \( \beta_1 \) when \( \delta_2 r \) is small. Similar bounds hold for \( \sqrt{\beta_2} - \beta_2 \). As a result, when \( \delta_2 r \) is sufficiently small, we have
\[
\tilde{\beta}_1 + \tilde{\beta}_2 \leq 108 \sigma_{\text{max}} \left( \| \tilde{X}_t - X_* \|_F \| \tilde{Y}_t \|_F \right)^2 + \| \tilde{X}_t \|_F \| \tilde{Y}_t - Y_* \|_F \|_F^2 .
\]
\[
+ 54 c_0 \sigma_{\text{max}} \sigma_{\text{min}} \| \tilde{Z}_t - Z_\|_F^2 \| Z_\|_F^2 .
\]
We now proceed to $\tilde{\gamma}_1$, for which we have
\[
|\tilde{\gamma}_1 - \gamma_1| = \left| \left\langle \tilde{X}_t - X_*, \left[ A^* A \left( X_t Y_t^T - M_* \right) \right] \tilde{Y}_t \Lambda_t \right\rangle - \left\langle \tilde{X}_t - X_*, (X_t Y_t^T - M_*) \tilde{Y}_t \Lambda_t \right\rangle \right|
\]
\[
= \left| \left\langle \tilde{X}_t - X_*, \left[ (A^* A - I) \left( X_t Y_t^T - M_* \right) \right] \tilde{Y}_t \Lambda_t \right\rangle \right|
\]
\[
\leq \left| \text{Tr} \left( \left[ (A^* A - I) \left( X_t Y_t^T - M_* \right) \right] \tilde{Y}_t \Lambda_t \left( \tilde{X}_t - X_* \right)^T \right) \right|
\]
\[
\leq \delta_{2r} \left( \left\| \left( \tilde{X}_t - X_* \right) \tilde{Y}_t^T \right\|_F + \left\| \tilde{Y}_t \left( \tilde{Y}_t - Y_* \right)^T \right\|_F + \left\| \left( \tilde{X}_t - X_* \right) \left( \tilde{Y}_t - Y_* \right)^T \right\|_F \right) \left\| \tilde{Y}_t \Lambda_t \left( \tilde{X}_t - X_* \right)^T \right\|_F.
\]
Here once again, we utilize the triangle inequality and Lemma 5. Noticing that $\|A_t - I\|$ is small, we further have
\[
\left\| \tilde{Y}_t \Lambda_t \left( \tilde{X}_t - X_* \right)^T \right\|_F \leq \left\| \tilde{Y}_t \right\|_F + \left\| \left( \tilde{X}_t - X_* \right)^T \right\|_F + \left\| \left( \tilde{X}_t - X_* \right)^T \right\|_F + \left\| \left( \tilde{Y}_t - Y_* \right)^T \right\|_F \leq \left\| \tilde{X}_t - X_* \right\|_F
\]
where we use $\|A_t - I\| \leq \zeta$ and $\|\tilde{Y}_t\|_F \leq 2\sqrt{\sigma_{\text{max}}}$. Combine the previous two bounds and apply the basic inequality $2ab \leq a^2 + b^2$ to see
\[
|\tilde{\gamma}_1 - \gamma_1| \leq \delta_{2r} \left( \left\| \tilde{X}_t - X_* \right\|_F + 2\sigma_{\text{max}}^{1/2} \delta_{2r} \left\| \tilde{Y}_t \right\|_F + 2\sigma_{\text{max}}^{1/2} \delta_{2r} \left\| \tilde{Y}_t \right\|_F + 2\sigma_{\text{max}}^{1/2} \delta_{2r} \left\| \tilde{X}_t - X_* \right\|_F \right)
\]
\[
\leq \delta_{2r} \left( \left\| \tilde{X}_t - X_* \right\|_F + \sum_{i=1}^{r} \sigma_{\text{min}} \text{dist} \left( Z_t, Z_* \right) \right)
\]
\[
\leq \delta_{2r} \left( \left\| \tilde{X}_t - X_* \right\|_F + \sum_{i=1}^{r} \sigma_{\text{min}} \text{dist} \left( Z_t, Z_* \right) \right)
\]
as long as $\delta_{2r}$ is sufficiently small. The same bound applies to $|\tilde{\gamma}_2 - \tilde{\gamma}_1|$. As a result, as long as $\delta_{2r}$ is small enough, $\tilde{\gamma}_1 + \tilde{\gamma}_2$ is lower bounded on the same order as $\gamma_1 + \gamma_2$, say
\[
\tilde{\gamma}_1 + \tilde{\gamma}_2 \geq \frac{1}{2} \left( \left\| \tilde{Y}_t \left( \tilde{X}_t - X_* \right)^T \right\|_F^2 + \left\| \tilde{X}_t \left( \tilde{Y}_t - Y_* \right)^T \right\|_F^2 \right) - \frac{\sigma_{\text{min}}}{6} \text{dist}^2 \left( Z_t, Z_* \right),
\]
One can then repeat the same arguments for the matrix factorization case to obtain the linear convergence. For the sake of space, we omit it.

5 Conclusions

This paper establishes the local linear convergence of gradient descent for rectangular low-rank matrix sensing without explicit regularization of factor balancedness under the standard RIP assumption, as long as a balanced initialization
is provided in the basin of attraction, which can be found by the spectral method. Different from previous work, we
analyzed a new error metric that takes into account the ambiguity due to invertible transforms, and showed that it
contracts linearly even without local restricted strong convexity. We believe that our technique can be used for other
low-rank matrix estimation problems. To conclude, we outline a few exciting future research directions.

- **Low-rank matrix completion.** We believe it is possible to extend our analysis to study rectangular matrix completion
  without regularization, by combining the leave-one-out technique in [24, 32] to carefully bound the incoherence of
  the iterates for both factors even without explicit balancing.

- **Improving dependence on κ and r.** The current paper does not try to optimize the dependence with respect to κ and r
  in terms of sample complexity and the size of the basin of attraction, which are slightly worse than their regularized
counterparts. A finer analysis will likely lead to better dependencies, which we leave to the future work.

**Acknowledgment**

This work 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.

**A Proof of Lemma 1**

For notational convenience, we define the following function

\[ g(Q) \triangleq \|XQ - X^*\|_F^2 + \|YQ^{-T} - Y^*\|_F^2. \]

Clearly, the optimal alignment matrix, if exists, must be

\[ \arg\min_{Q \in \mathbb{R}^{r \times r} : Q \text{ is invertible}} g(Q). \]

With this notation in place, we consider the following constrained minimization problem:

\[ \min_{Q \in \mathbb{R}^{r \times r} : Q \text{ is invertible}} g(Q) \]

subject to \[ \|Q - P\|_F \leq \frac{5\delta}{\sigma_{\min}(X^*)}. \]

In view of Weyl’s inequality, we obtain that for any feasible \( Q \),

\[ \sigma_{\min}(Q) \geq \sigma_{\min}(P) - \frac{5\delta}{\sigma_{\min}(X^*)} \geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4}, \]

as long as \( \delta \leq \sigma_{\min}(X^*)/80 \). As a result, one sees that \( g(Q) \) is a continuous function over \{ \( Q \) : \( \|Q - P\| \leq 5\delta/\sigma_{\min}(X^*) \} \), which is a compact set over invertible matrices. Applying the Weierstrass extreme value theorem yields the claim that the minimizer of the constrained problem exists. Denote this minimizer by \( Q_1 \). In what follows, we intend to show that \( Q_1 \) is also the minimizer of the unconstrained problem. Letting \( Q \) be an arbitrary matrix with \( g(Q) \leq 2\delta^2 \) (the existence is assured since \( g(Q_1) \leq g(P) \leq 2\delta^2 \)), we have

\[ \sqrt{2}\delta \geq \|XQ - X^*\|_F \geq \|XQ - XP\|_F - \|XP - X^*\|_F, \]

which in conjunction with (13) implies

\[ (1 + \sqrt{2})\delta \geq \|X(Q - P)\|_F \geq \sigma_{\min}(X)\|Q_1 - P\|_F. \]

We now turn to investigating \( \sigma_{\min}(X) \). Weyl’s inequality tells us that

\[ |\sigma_{\min}(XP) - \sigma_{\min}(X^*)| \leq \|XP - X^*\|_F \leq \delta \leq \frac{1}{4}\sigma_{\min}(X^*), \]

which further implies

\[ \frac{3}{4}\sigma_{\min}(X^*) \leq \sigma_{\min}(XP) \leq \sigma_{\min}(X)\sigma_{\max}(P) \leq \frac{3}{2}\sigma_{\min}(X). \]
Therefore we arrive at $\sigma_{\min}(X) \geq \frac{\sigma_{\min}(X_\star)}{2}$. Putting this back to (23) yields which finally gives

$$
\|Q - P\|_F \leq 2(1 + \sqrt{2}) \frac{\delta}{\sigma_{\min}(X_\star)} < \frac{5\delta}{\sigma_{\min}(X_\star)}.
$$

In all, revisiting the proof reveals that any matrix $Q$ such that $g(Q) \leq 2\delta^2$ must obey the above bound. Therefore the minimizer of the constrained problem and that of the unconstrained one coincide with each other. This finished the proof.

**B Proof of Lemma 2**

Recall that

$$
g(P) = \|XP - X_\star\|_F^2 + \|YP^{-T} - Y_\star\|_F^2
$$

$$
= \text{Tr} \left( XPP^T X^\top \right) - 2 \text{Tr} \left( P^T X^\top X_\star \right) + \text{Tr} \left( X^\top \star X_\star \right)
$$

$$
+ \text{Tr} \left( P^{-1}Y^T YP^{-T} \right) - 2 \text{Tr} \left( P^{-1}Y^T Y_\star \right) + \text{Tr} \left( Y_\star^T Y_\star \right),
$$

whose gradient is given by

$$
\nabla g(P) = 2X^\top XP - 2X^\top X_\star - 2 \left( PP^T \right)^{-1} Y^T Y \left( PP^T \right)^{-1} P + 2P^{-1}Y_\star^T YP^{-T}.
$$

Since $Q$ minimizes $g(P)$, it must satisfy the first-order optimality condition, i.e.

$$
\nabla g(Q) = 0.
$$

Identify $\tilde{X} = XQ$ and $\tilde{Y} = YQ^{-T}$ to yield

$$
\tilde{X}^\top \tilde{X} - \tilde{X}^\top X_\star = \tilde{Y}^\top \tilde{Y} - Y^\top Y_\star.
$$

**C Proof of Lemma 3**

We prove the first part and the second part follows by symmetry. Denote $E_x = X - X_\star$ and $E_y = Y - Y_\star$. We have

$$
XY^T - M_\star = E_x Y^T + X, E_y^T.
$$

Since $Z$ is aligned with $Z_\star$, Lemma 2 tells us that $X^\top E_x = E_y^\top Y$. As a result, one has

$$
\langle X - X_\star, (XY^T - M_\star) Y \rangle = \text{Tr} \left( E_x^\top (E_x Y^T + X, E_y^T) Y \right)
$$

$$
= \text{Tr} \left( E_x^\top E_x Y^T Y \right) + \text{Tr} \left( E_x^\top X, E_y^T Y \right)
$$

$$
= \|YE_x^T\|_F^2 + \text{Tr} \left( E_y^\top Y E_x^\top X \right) - \text{Tr} \left( E_y^\top Y E_x^\top E_x \right)
$$

$$
= \|YE_x^T\|_F^2 + \|X^\top E_x\|_F^2 - \text{Tr} \left( X^\top E_x E_x^\top E_x \right). \quad (25)
$$

Complete the squares to see that

$$
\|X^\top E_x\|_F^2 - \text{Tr} \left( X^\top E_x E_x^\top E_x \right) = \left\| E_x^\top X - \frac{1}{2} E_x^\top E_x \right\|_F^2 - \frac{1}{4} \|E_x\|_F^4.
$$

Combine the previous two bounds to yield the desired result.
D Proof of Lemma 4

Again, we demonstrate the claim on $X$ and the claim on $Y$ follows by symmetry. Given the decomposition

$$XY^T - M_* = (X - X_*)Y^T + X(Y - Y_*)^T + (X_* - X)(Y - Y_*)^T,$$

we obtain

$$\| (XY^T - M_*) Y \|_F \leq \sigma_1(Y) \| XY^T - M_* \|_F$$

$$\leq \frac{3}{2} \sigma_1(Y_*) \left( \| (X - X_*)Y^T \|_F + \| X(Y - Y_*)^T \|_F + \| (X_* - X)(Y - Y_*)^T \|_F \right),$$

where the last line combines the triangle inequality and Weyl’s inequality

$$\sigma_1(Y) \leq \sigma_1(Y_*) + \| Y - Y_* \| \leq \frac{3}{2} \sigma_1(Y).$$

The proof is then finished.

References


