Fast Non-convex Matrix Sensing with Optimal Sample Complexity

Jian-Feng Cai *[1] Tong Wu †[1] Ruizhe Xia[1]

[1]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China

Abstract

We study the problem of recovering an unknown $d_1 \times d_2$ rank-r matrix from m random linear measurements. Convex methods achieve the optimal sample complexity $m = \Omega(r(d_1 + d_2))$ but are computationally expensive. Non-convex approaches, while more computationally efficient, often require suboptimal sample complexity $m = \Omega(r^2(d_1 + d_2))$. A recent advance achieves $m = \Omega(rd_1)$ for a fast non-convex approach but relies on the restrictive assumption of positive semidefinite (PSD) matrices and suffers from slow convergence in ill-conditioned settings. Bridging this gap, we show that Riemannian gradient descent (RGD) achieves both optimal sample complexity and computational efficiency without requiring the PSD assumption. Specifically, for Gaussian measurements, RGD exactly recovers the low-rank matrix with $m = \Omega(r(d_1 + d_2))$, matching the information-theoretic lower bound, and converges linearly to the global minimum with an arbitrarily small convergence rate.

1 INTRODUCTION

In this work, we study the problem of recovering an unknown matrix $\boldsymbol{X}_\star \in \mathbb{R}^{d_1 \times d_2}$ from its random linear measurements $\boldsymbol{b} := \mathcal{A}(\boldsymbol{X}_\star) \in \mathbb{R}^m$, where the linear operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ is defined as

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

Here, $\boldsymbol{A}_i \in \mathbb{R}^{d_1 \times d_2}$ are measurement matrices, $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbb{R}^{d_1 \times d_2}$, and $m \ll d_1 d_2$, making the problem inherently underdetermined. To overcome this challenge, we assume that $\boldsymbol{X}_\star$ has rank r, effectively reducing the degrees of freedom in the matrix to $r(d_1 + d_2 - r)$. Under this assumption, exact recovery of $\boldsymbol{X}_\star$ becomes theoretically feasible when the number of measurements m scales on the order of this degree of freedom. This problem, known as low-rank matrix recovery problem, lies at the intersection of theoretical and applied mathematics, with profound implications across machine learning, signal processing, and statistics. It encompasses several classical problems, such as matrix completion [Candes and Tao, 2010, Gross, 2011, Sun and Luo, 2016], phase retrieval [Candès et al., 2013], and quantum state tomography [Hsu et al., 2024], among others [Chi et al., 2019]. The core challenge lies in recovering $\boldsymbol{X}_\star$ using as few measurements m as possible, ideally matching the information-theoretic lower bound of $\Omega(r(d_1 + d_2 - r))$, while ensuring that the recovery method remains computationally efficient, operating in polynomial time as problem dimensions grow.

A prominent line of research focuses on convex relaxation methods, where the low-rank matrix is represented in $\mathbb{R}^{d_1 \times d_2}$, and the nuclear norm $\|\cdot\|_*$ is used as a convex surrogate for the rank function. For applications such as matrix sensing [Recht et al., 2010], matrix completion [Candes and Tao, 2010, Gross, 2011], and blind deconvolution and demixing [Jung et al., 2017], it has been shown that this approach can achieve exact recovery with m scaling as $\Omega(r(d_1 + d_2))$, up to logarithmic factors, matching the information-theoretically optimal sample complexity. However, these convex methods are computationally demanding, as they require optimization in the entire space $\mathbb{R}^{d_1 \times d_2}$, and the low-rank structure of the solution is not easily exploited.

To address these computational challenges, non-convex approaches have gained prominence. Factorization-based methods address this by representing the low-rank matrix as $\boldsymbol{L}\boldsymbol{R}^T$, where $\boldsymbol{L} \in \mathbb{R}^{d_1 \times r}$ and $\boldsymbol{R} \in \mathbb{R}^{d_2 \times r}$. This reduces the number of optimization variables to $r(d_1 + d_2)$, significantly fewer than the $d_1 d_2$ variables in convex approaches. Simple algorithms such as gradient descent and alternat-

[*] This work is partially supported by Hong Kong Research Grant Council GRFs 16307023 and 16306124.

[†] Main contributor.

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:497–520.

ing minimization, when initialized appropriately, have been shown to converge linearly to the global minimum under suitable assumptions on $\mathcal{A}$ and $\boldsymbol{X}_\star$ [Jain et al., 2013, Tu et al., 2016, Chen et al., 2020, Sun and Luo, 2016, Tong et al., 2021, Charisopoulos et al., 2021, Zilber and Nadler, 2022]. Another class of non-convex methods leverages manifold optimization, eliminating redundancy in the factorization parametrization either by representing factors on quotient Riemannian manifolds [Keshavan et al., 2009, Huang et al., 2017, Zheng et al., 2025] or by optimizing directly on the Riemannian manifold of rank-r matrices embedded in $\mathbb{R}^{d_1 \times d_2}$ [Wei et al., 2016, Cai and Wei, 2024, Hsu et al., 2024]. These methods are often more efficient and have also been proven to converge linearly to $\boldsymbol{X}_\star$ with the spectral initialization under appropriate conditions. However, a critical limitation of fast non-convex approaches is their suboptimal sample complexity, typically requiring $m = \Omega(r^2(d_1 + d_2))$ or higher, which scales quadratically with r. Iterative Hard Thresholding (IHT) [Tanner and Wei, 2013, Tu et al., 2016] achieves $m = \Omega(r(d_1 + d_2))$, but its computational cost is higher than the aforementioned fast non-convex methods due to repeated r-truncated singular value decompositions (SVD) on full matrices, which incur larger constant factors compared to matrix multiplication (MM) of the same computational order.

The feasibility of simultaneously achieving optimal sample complexity and low computational cost remains an open research question. Recently, Stöger and Zhu [2025] made progress in this direction for the special case of low-rank positive semidefinite (PSD) matrix sensing. By assuming Gaussian measurement matrices and representing the PSD matrix as $\boldsymbol{L}\boldsymbol{L}^T$, the authors demonstrated that factorized gradient descent can recover $\boldsymbol{X}_\star$ with sample complexity $m = \Omega(rd_1)$. However, their approach suffers from slow convergence for ill-conditioned matrices due to the dependence of the step size on the condition number of $\boldsymbol{X}_\star$. Moreover, extending these results to the more general case of non-PSD matrix recovery introduces additional challenges, particularly in balancing the factors $\boldsymbol{L}$ and $\boldsymbol{R}$ without explicit regularization [Chen et al., 2020].

In this paper, we present a theoretical result showing that Riemannian gradient descent (RGD) [Wei et al., 2016] achieves both optimal sample complexity and low computational cost for recovering rectangular low-rank matrices. Specifically, we prove that RGD can recover a rank-r matrix with optimal sample complexity $m = \Omega(r(d_1 + d_2))$ when $\mathcal{A}$ is a Gaussian measurement operator, achieving an arbitrarily small convergence rate. Unlike factorized gradient descent, our approach eliminates the need for additional regularization terms, simplifying both the theoretical analysis and the practical implementation. Furthermore, RGD is computationally efficient, as it parameterizes matrices on the Riemannian manifold with only $\Theta(r(d_1 + d_2))$ variables. By reducing the sample complexity from quadratic to linear dependence on r, our work bridges the gap between optimal sample complexity and computational efficiency, establishing RGD as a state-of-the-art method for low-rank matrix recovery. Table 1 provides a summary of the sample complexity m and computational efficiency for representative non-convex methods in low-rank matrix sensing (all quantities are stated up to order $O(\cdot)$). The per-iteration computational cost consists of two parts: (1) the common cost of applying $\mathcal{A}^*\mathcal{A}$ (dominated by matrix multiplication, MM), and (2) method-specific cost highlighted in Table 1. It may include extra MM and complex operations like QR decomposition, matrix inversion, and SVD.

The rest of the paper is organized as follows. In Section 2, we formulate the non-convex optimization problem for low-rank matrix recovery, describe the Riemannian gradient descent algorithm, and present our main theoretical result, Theorem 1. Section 3 provides the proof of the main theorem, with the Restricted Isometry Property (RIP) and the decoupling technique as key tools. Most technical details are deferred to the Appendix. Finally, we conclude with a discussion of potential directions for future research in Section 5.

2 ALGORITHMS AND RESULTS

In this section, we first formulate low-rank matrix recovery as a non-convex optimization problem on the Riemannian manifold of all rank-r matrices embedded in $\mathbb{R}^{d_1 \times d_2}$. We then describe the Riemannian gradient descent algorithm for solving this optimization problem. Finally, we present our main theoretical result.

2.1 ALGORITHMS

To recover the rank-r matrix $\boldsymbol{X}_\star \in \mathbb{R}^{d_1 \times d_2}$ from its measurement $\boldsymbol{b} = \mathcal{A}(\boldsymbol{X}_\star)$, we solve the constrained least-squares problem:

$$\begin{aligned} \min_{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}} \quad & \mathcal{L}(\boldsymbol{X}) := \frac{1}{2} \left\| \boldsymbol{b} - \mathcal{A}(\boldsymbol{X}) \right\|_2^2 \\ \text{s.t.} \quad & \operatorname{rank}(\boldsymbol{X}) = r. \end{aligned} \tag{2}$$

Solving (2) is challenging due to the non-convexity introduced by the low-rank constraint. A common approach to overcome this is to use matrix factorization, parametrizing the low-rank matrix as $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{R}^T$ with $\boldsymbol{L} \in \mathbb{R}^{d_1 \times r}, \boldsymbol{R} \in \mathbb{R}^{d_2 \times r}$. This leads to the following optimization problem:

$$\min_{\boldsymbol{L} \in \mathbb{R}^{d_1 \times r}, \boldsymbol{R} \in \mathbb{R}^{d_2 \times r}} \mathcal{L}(\boldsymbol{L}\boldsymbol{R}^T). \tag{3}$$

However, the factorization $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{R}^T$ is redundant and non-unique. Specifically, $\boldsymbol{X} = (\boldsymbol{L}\boldsymbol{Q})(\boldsymbol{R}\boldsymbol{Q}^{-T})^T$ for any

Table 1: Comparison of Non-Convex Methods for Low-Rank Matrix Sensing ($d_1 = d_2$).

Method	m	Iterations	Extra Cost/Iter
SVP [Jain et al., 2010], NIHT [Tanner and Wei, 2013]	$d_1 r$	$\log(1/\varepsilon)$	$d_1^2 r$ (SVD)
RGD [Wei et al., 2016]	$d_1 r^2 \kappa^2$	$\log(1/\varepsilon)$	$d_1^2 r$ (MM) + $d_1 r^2$(QR) + r^3 (SVD)
Scaled GD [Tong et al., 2021]	$d_1 r^2 \kappa^2$	$\log(1/\varepsilon)$	$d_1^2 r$ (MM) + r^3 (Inversion)
Factorized GD (PSD only) [Stöger and Zhu, 2025]	$d_1 r \kappa^2$	$\kappa^2 \log(1/\varepsilon)$	$d_1^2 r$ (MM)
RGD (this paper)	$d_1 r \kappa^2$	$\log(1/\varepsilon)$	$d_1^2 r$ (MM) + $d_1 r^2$(QR) + r^3 (SVD)

invertible $r \times r$ matrix $\boldsymbol{Q}$. This invariance causes the critical points of $\mathcal{L}$ to be unbounded and not isolated in parameter space, leading to potential optimization difficulties. To address this issue, some works simply assume that $\boldsymbol{L} = \boldsymbol{R}$ to recover PSD matrices [Stöger and Zhu, 2025], while others introduce an imbalance regularization term $\|\boldsymbol{L}^T\boldsymbol{L} - \boldsymbol{R}^T\boldsymbol{R}\|_F$ to the loss function in (3) [Tu et al., 2016, Ge et al., 2017]. Despite these approaches, the factorization $\boldsymbol{L}\boldsymbol{R}^T$ can still lead to an ill-conditioned Hessian. To analyze this, assume $\mathcal{A}$ is random and $\mathbb{E}[\mathcal{A}^*\mathcal{A}] = \mathcal{I}$. This assumption holds in many common low-rank matrix recovery problems, such as Gaussian matrix sensing, matrix completion, and quantum state tomography. We then consider the behavior of the expected loss function in (3), which is $\mathbb{E}[\mathcal{L}(\boldsymbol{L}\boldsymbol{R}^T)] = \frac{1}{2}\|\boldsymbol{L}\boldsymbol{R}^T - \boldsymbol{X}_\star\|_F^2$. The Hessian of $\mathbb{E}[\mathcal{L}]$ with respect to (w.r.t.) $\boldsymbol{L}$ and $\boldsymbol{R}$ is given by:

$$\nabla^2_{(\boldsymbol{L},\boldsymbol{R})}(\mathbb{E}[\mathcal{L}(\boldsymbol{L}\boldsymbol{R}^T)]) = \begin{bmatrix} (\boldsymbol{R}^T\boldsymbol{R}) \otimes \boldsymbol{I}_{d_1} & \bullet \\ \bullet^T & (\boldsymbol{L}^T\boldsymbol{L}) \otimes \boldsymbol{I}_{d_2} \end{bmatrix},$$

where $\bullet = \boldsymbol{I}_r \otimes (\boldsymbol{L}\boldsymbol{R}^T - \boldsymbol{X}_\star) + (\boldsymbol{R}^T \otimes \boldsymbol{L})\boldsymbol{K}^{(d_2,r)}$ and $\boldsymbol{K}^{(d_2,r)}$ is the commutation matrix [Von Rosen, 1988].

The condition number of the Hessian depends on those of $\boldsymbol{L}$ and $\boldsymbol{R}$, which slows convergence and ties the convergence rate to the condition number of $\boldsymbol{X}_\star$. To mitigate this, various approaches have been proposed, including preconditioning in parameter space by the inversion of the block diagonal of $\nabla^2_{(\boldsymbol{L},\boldsymbol{R})}(\mathbb{E}[\mathcal{L}(\boldsymbol{L}\boldsymbol{R}^T)])$ [Tong et al., 2021], optimization on quotient Riemannian manifolds [Keshavan et al., 2009, Huang et al., 2017, Zheng et al., 2025], and on the Riemannian manifold of rank-r matrices embedded in $\mathbb{R}^{d_1 \times d_2}$ [Wei et al., 2016, Cai and Wei, 2024, Hsu et al., 2024].

We consider the optimization over the embedded Riemannian manifold of rank-r matrices, which offers several advantages. First, the manifold representation is intrinsic, eliminating redundancy and the need for regularization in factorization-based methods. Second, the embedded manifold lies in $\mathbb{R}^{d_1 \times d_2}$, where the expected loss function simplifies to $\mathbb{E}[\mathcal{L}(\boldsymbol{X})] = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{X}_\star\|_F^2$ and the expected Hessian becomes $\mathcal{I}$, with a perfect condition number. This ensures fast convergence. Third, the operator $\mathcal{A}$ acting on matrices in $\mathbb{R}^{d_1 \times d_2}$ is well-studied, with benign properties such as RIP that can simplify analysis. In contrast, its behavior in the parameter space is less understood, requiring additional work to generalize these properties [Tong et al., 2021, Stöger and Zhu, 2025].

Let $\mathbb{M}_r = \{\boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2} : \operatorname{rank}(\boldsymbol{X}) = r\}$ be the embedded manifold of all rank-r matrices in $\mathbb{R}^{d_1 \times d_2}$. For $\boldsymbol{X} \in \mathbb{M}_r$, given its compact singular value decomposition (SVD) of $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, the tangent space at $\boldsymbol{X}$ is

$$\mathbb{T}_{\boldsymbol{X}} := \left\{\boldsymbol{U}\boldsymbol{R}^T + \boldsymbol{L}\boldsymbol{V}^T : \boldsymbol{L} \in \mathbb{R}^{d_1 \times r}, \boldsymbol{R} \in \mathbb{R}^{d_2 \times r}\right\}.$$

The orthogonal projection $\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{T}_{\boldsymbol{X}}$ has the closed-form expression

$$\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}(\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^T.$$

Then the constrained least-squares problem (2) becomes $\min_{\boldsymbol{X} \in \mathbb{M}_r} \mathcal{L}(\boldsymbol{X})$. We solve it using Riemannian gradient descent (RGD) [Absil et al., 2008, Vandereycken, 2013]:

$$\boldsymbol{X}_{t+1} = \mathcal{H}_r(\boldsymbol{X}_t - \mu \mathcal{P}_{\mathbb{T}_{\boldsymbol{X}_t}} \mathcal{A}^*(\mathcal{A}(\boldsymbol{X}_t) - \boldsymbol{b})), \forall t \in \mathbb{N}, \quad (4)$$

where:

- $\mathcal{H}_r(\cdot)$ is the hard thresholding operator and serves as a retraction, which is defined via the r-truncated SVD $\mathcal{H}_r(\boldsymbol{Z}) := \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ provided the SVD of $\boldsymbol{Z} = \sum_i \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$ with $\sigma_1 \geq \sigma_2 \geq \cdots$,
- μ is the step size, and
- $\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}_t}} \mathcal{A}^*(\mathcal{A}(\boldsymbol{X}_t) - \boldsymbol{b})$ is the Riemannian gradient of $\mathcal{L}(\boldsymbol{X})$ at $\boldsymbol{X}_t$.

The computational cost per iteration of (4) is low. Aside from applying $\mathcal{A}$ and $\mathcal{A}^*$, the most expensive operations are $\mathcal{H}_r$ and $\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}_t}}$. Since $\boldsymbol{X}_t$ can be stored in a compact SVD form as $\boldsymbol{X}_t = \boldsymbol{U}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^T$, computing $\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}_t}}$ requires only $O(r)$ matrix-vector products. Besides, in (4), $\mathcal{H}_r$ is applied to a matrix $\boldsymbol{W}_t$ in $\mathbb{T}_{\boldsymbol{X}_t}$, which has rank at most $2r$. As shown in [Wei et al., 2016], $\mathcal{H}_r(\boldsymbol{W}_t)$ can be efficiently computed using two QR decompositions of a tall matrix of width r, one SVD of a $2r \times 2r$ matrix, and a few matrix-vector products. Thus, the per-iteration computational cost of RGD is of the same order as that of gradient descent based on factorization or the quotient Riemannian manifolds. Moreover, RGD achieves a more favorable convergence rate that is independent of the condition number of the ground truth matrix and can be arbitrarily small. This results in fewer iterations to reach the target accuracy, as demonstrated in our theoretical results.

Due to the non-convexity of the problem, we also need a good initialization $\boldsymbol{X}_0$. We use the spectral initialization outlined in [Jain et al., 2013]. We initialize $\boldsymbol{X}_0$ as $\mathcal{H}_r(\mathcal{A}^*(\boldsymbol{b}))$, where $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^{d_1 \times d_2}$ is the adjoint operator of $\mathcal{A}$. Spectral initialization is a natural and common choice since $\mathbb{E}[\mathcal{A}^*(\boldsymbol{b})] = \boldsymbol{X}_\star$ and the operator $\mathcal{H}_r$ extracts the rank-r structure.

We summarize our algorithm in Algorithm 1. For simplicity, we denote $\mathbb{T}_t$ and $\mathcal{P}_{\mathbb{T}_t}$ as $\mathbb{T}_{\boldsymbol{X}_t}$ and $\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}_t}}$, respectively.

```
Algorithm 1: Riemannian Gradient Descent (RGD) for Low-Rank Matrix Recovery
Input: Measurement operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$,
       observations $\boldsymbol{b} \in \mathbb{R}^m$, step size $\mu > 0$
Stage 1 (Spectral Initialization): Define the
  initialization $\boldsymbol{X}_0 \in \mathbb{R}^{d_1 \times d_2}$ as
      $\boldsymbol{X}_0 = \mathcal{H}_r(\mathcal{A}^*(\boldsymbol{b}))$.
  Stage 2 (Iteration): for $t = 0, 1, 2, \ldots$ do
      $\boldsymbol{W}_t = \boldsymbol{X}_t - \mu \mathcal{P}_{\mathbb{T}_t} \mathcal{A}^*(\mathcal{A}(\boldsymbol{X}_t) - \boldsymbol{b})$,
      $\boldsymbol{X}_{t+1} = \mathcal{H}_r(\boldsymbol{W}_t)$.
```

2.2 MAIN RESULT

The main result of this paper provides a recovery guarantee for Algorithm 1 with optimal sample complexity. We first define the condition number of $\boldsymbol{X}_\star$ as

$$\kappa := \frac{\|\boldsymbol{X}_\star\|_2}{\sigma_{\min}(\boldsymbol{X}_\star)},$$

where $\|\cdot\|_2$ is the spectral norm (also called 2-norm) for matrices, and $\sigma_{\min}(\boldsymbol{X}_\star) := \sigma_r(\boldsymbol{X}_\star)$ is the smallest non-zero singular value of $\boldsymbol{X}_\star$. We call $\mathcal{A}$ a Gaussian measurement operator when the measurement matrices $\{\boldsymbol{A}_i\}_{i=1}^m$ in (1) have i.i.d. entries drawn from $\mathcal{N}(0,1)$. Our main theorem is stated as follows:

Theorem 1. *Let $\mathcal{A}$ be a Gaussian measurement operator. Let $\boldsymbol{X}_\star \in \mathbb{R}^{d_1 \times d_2}$ be a rank-r matrix and $\boldsymbol{b} = \mathcal{A}(\boldsymbol{X}_\star) \in \mathbb{R}^m$. Let $\{\boldsymbol{X}_t\}_{t \in \mathbb{N}}$ be the sequence generated by Algorithm 1 with step size $\mu = 1$. Then, for any $\rho \in (0,1)$, there exists a constant C depending only on ρ such that: if the number of measurements m satisfies*

$$m \geq C\kappa^2 r(d_1 + d_2),$$

with probability at least $1 - 7\exp(-(d_1 + d_2))$, it holds for all iterations $t \geq 0$ that

$$\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_F \leq \sqrt{2r}\rho^t \sigma_{\min}(\boldsymbol{X}_\star). \tag{5}$$

The proof of the theorem is deferred to Section 3. Our result attains optimal sample complexity and high computational efficiency. The key advantages of our approach are:

- The constant C in Theorem 1 depends only on the convergence rate ρ, which allows our result to achieve optimal sample complexity $m = \Omega(\kappa^2 r(d_1 + d_2))$. Importantly, this result does not require the positive semidefinite (PSD) assumption on $\boldsymbol{X}_\star$, which is a key limitation in [Stöger and Zhu, 2025]. Their work relies on the PSD structure to derive a sample complexity of $m = \Omega(\kappa^2 r d_1)$, restricting its applicability to PSD matrices. By contrast, our approach applies to general rectangular matrices, significantly broadening the scope of problems that can be addressed. This generality, combined with optimal sample complexity, underscores the versatility and strength of our method.
- The convergence rate ρ in Theorem 1 can be made arbitrarily small by choosing a sufficiently large C. Thus, our method achieves ε-accuracy for $\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_F$ in $O\left(\log\left(\sqrt{r}\sigma_{\min}(\boldsymbol{X}_\star)\varepsilon^{-1}\right)\right)$ iterations. In contrast, the step size in [Stöger and Zhu, 2025] is $O((\kappa\|\boldsymbol{X}_\star\|_2)^{-1})$, leading to a convergence rate of $1 - O(\kappa^{-2})$. This results in $O\left(\kappa^2 \log\left(\sqrt{r}\sigma_{\min}(\boldsymbol{X}_\star)\varepsilon^{-1}\right)\right)$ iterations to achieve ε-accuracy for $\|\boldsymbol{L}_t\boldsymbol{L}_t^T - \boldsymbol{X}_\star\|_F$, where $\boldsymbol{L}_t\boldsymbol{L}_t^T$ corresponds to $\boldsymbol{X}_t$ in our setting. Our method is significantly more efficient, particularly for ill-conditioned matrices.

3 THEORETICAL ANALYSIS

In this section, we prove Theorem 1. We begin by introducing the Restricted Isometry Property (RIP), which is commonly used in prior analyses. Next, we highlight the primary theoretical challenge that introduces the r^2 term in the sample complexity. To address this issue, we present the key decoupling technique, inspired by [Stöger and Zhu, 2025]. Following this, we provide the necessary supporting lemmas and conclude with the proof of the main theorem based on these results.

3.1 RESTRICTED ISOMETRY PROPERTY

The Restricted Isometry Property (RIP) is a fundamental tool in the analysis of low-rank matrix recovery problems, particularly under random Gaussian measurements. This property ensures that a measurement operator approximately preserves the geometry of low-rank matrices, which is crucial for analyzing the performance of various recovery algorithms. We introduce the definition and properties of the Restricted Isometry Property (RIP), which plays a crucial role in our analysis.

Definition 1. *The linear measurement operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ satisfies the Restricted Isometry Property*

(RIP) of rank r with RIP-constant $\delta_r \in (0,1)$ if it holds that

$$(1-\delta_r)\,\|\boldsymbol{Z}\|_F^2 \le \|\mathcal{A}(\boldsymbol{Z})\|_2^2 \le (1+\delta_r)\,\|\boldsymbol{Z}\|_F^2, \\ \forall\, \boldsymbol{Z} \in \mathbb{R}^{d_1\times d_2} \;:\; \operatorname{rank}(\boldsymbol{Z}) \le r.$$

The RIP is a uniform result, as it holds for all low-rank matrices rather than just specific matrices of interest, such as $\boldsymbol{X}_t - \boldsymbol{X}_\star$. The RIP is widely used in the theoretical analysis of matrix sensing problems. If $m = \Omega(r(d_1+d_2))$, then the measurement operator $\mathcal{A}$ satisfies the RIP of order r with high probability. The results from [Candes and Plan, 2011, Lemma 3.1] and [Stöger and Zhu, 2025, Lemma 2.2] directly extend to rectangular matrices:

Lemma 1. *Let $\mathcal{A} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ be a Gaussian measurement operator as described above. Then, $\mathcal{A}$ satisfies the RIP of rank-r with constant δ_r satisfying $\delta_r = \delta \le 1$ with probability $1-\varepsilon$ when*

$$m \ge C\delta^{-2}\left(r(d_1+d_2) + \log\left(2\varepsilon^{-1}\right)\right),$$

where $C > 0$ is a universal constant. In particular, with probability at least $1-\exp(-(d_1+d_2))$, $\mathcal{A}$ satisfies the RIP of rank r and constant δ provided $m \ge C\delta^{-2} r(d_1+d_2)$.

The following properties of the RIP will be used throughout our proofs. The mapping $\mathcal{I} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{d_1\times d_2}$ represents the identity.

Lemma 2. *Let $\mathcal{A} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ be a linear measurement operator satisfying the RIP with r_0 and RIP constant δ_{r_0} for any $r_0 \le 3r$. Then, the following statements hold:*

1. *Let $\boldsymbol{V} \in \mathbb{R}^{d_2\times r'}$ be any matrix with orthonormal columns, i.e., $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$. Then, for any matrix $\boldsymbol{Z} \in \mathbb{R}^{d_1\times d_2}$ satisfying $\operatorname{rank}(\boldsymbol{Z}) \le r$, it holds that*

   $$\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\boldsymbol{Z})\boldsymbol{V}\|_F \le \delta_{r+2r'}\|\boldsymbol{Z}\|_F. \tag{6}$$

   In particular, if we take $r' = 1$, then we have

   $$\|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\boldsymbol{Z})\|_2 \le \delta_{r+2}\|\boldsymbol{Z}\|_F. \tag{7}$$

2. *Let $\boldsymbol{x} \in \mathbb{R}^{d_1}$ be such that $\|\boldsymbol{x}\|_2 = 1$, and let $\boldsymbol{y} \in \mathbb{R}^{d_2}$ be such that $\|\boldsymbol{y}\|_2 = 1$. Define the orthogonal projection operators*

   $$\mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z}) := \left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{Z}\right\rangle \boldsymbol{x}\boldsymbol{y}^T, \\ \mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z}) := \boldsymbol{Z} - \left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{Z}\right\rangle \boldsymbol{x}\boldsymbol{y}^T.$$

   Then, for any matrix $\boldsymbol{Z} \in \mathbb{R}^{d_1\times d_2}$ satisfying $\operatorname{rank}(\boldsymbol{Z}) \le r$, we have

   $$\left|\left\langle \mathcal{A}(\boldsymbol{x}\boldsymbol{y}^T), \mathcal{A}(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z}))\right\rangle\right| \le \delta_{r+2}\|\boldsymbol{Z}\|_F. \tag{8}$$

3. *Let $\boldsymbol{X} \in \mathbb{R}^{d_1\times d_2}$ be a matrix of rank r. Then, it holds that*

   $$\sup_{\|\boldsymbol{Z}\|_F=1} \|(\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}} - \mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}})(\boldsymbol{Z})\|_F \le \delta_{2r}. \tag{9}$$

4. *Let $\boldsymbol{Z} \in \mathbb{R}^{d_1\times d_2}$ be a matrix of rank at most r. Then,*

   $$\|\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}\mathcal{A}^*\mathcal{A}(\mathcal{I} - \mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}})(\boldsymbol{Z})\|_F \\ \le \delta_{3r}\,\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}})(\boldsymbol{Z})\|_F. \tag{10}$$

The proof is in Appendix B.

3.2 LIMITATIONS OF RIP-BASED ANALYSIS

Before presenting our proof, we first highlight why uniform results based solely on the RIP are insufficient for achieving optimal sample complexity. A standard RIP-based analysis [Wei et al., 2016] typically yields a sample complexity that scales as r^2 rather than r. They show that a sufficiently small yet $O(1)$ RIP constant, requiring $m = \Omega(r(d_1+d_2))$, ensures

$$\|\mathcal{P}_{\mathbb{T}_t}(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\boldsymbol{X}_\star - \boldsymbol{X}_t)\|_F \ll \|\boldsymbol{X}_\star - \boldsymbol{X}_t\|_F. \tag{11}$$

This inequality guarantees linear convergence of $\{\boldsymbol{X}_t\}_{t\ge T}$ to $\boldsymbol{X}_\star$ in Frobenius norm for some $T \in \mathbb{N}$ whenever $\boldsymbol{X}_T$ satisfies

$$\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F \ll \sigma_{\min}(\boldsymbol{X}_\star). \tag{12}$$

To achieve this, they simply take $T = 0$ and use spectral initialization, which only achieve $\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_2 \ll \sigma_{\min}(\boldsymbol{X}_\star)$ with $m = \Omega(r(d_1+d_2)\kappa^2)$. They use $\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F \le \sqrt{2r}\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_2$ and require a RIP constant scaling as $O(1/\sqrt{r})$ to ensure (12), which in turn necessitates a sample complexity of $\Omega(r^2)$.

Alternatively, one could analyze convergence in the 2-norm, which would require a 2-norm counterpart of (11):

$$\|\mathcal{P}_{\mathbb{T}_t}(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\boldsymbol{X}_\star - \boldsymbol{X}_t)\|_2 \ll \|\boldsymbol{X}_\star - \boldsymbol{X}_t\|_2. \tag{13}$$

However, deriving (13) is challenging. Attempting to prove (13), we may consider proving a uniform result such as $\|\mathcal{P}_{\mathbb{T}_t}(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\boldsymbol{\Delta}_t)\|_2 \ll \|\boldsymbol{\Delta}_t\|_2$ for all possible $2r$-rank matrices $\boldsymbol{\Delta}_t$, but it is highly likely to fail with $\Omega(r)$ in sample complexity. Indeed, [Stöger and Zhu, 2025] provides a related negative result:

$$\sup_{\operatorname{rank}(\boldsymbol{Z})\le r} \|(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\boldsymbol{Z})\|_2 \ge \frac{1}{16}\|\boldsymbol{Z}\|_2\sqrt{\frac{r^2 d_1}{m}}.$$

Although their setting differs slightly from ours, this result underscores the difficulty of establishing uniform 2-norm bounds analogous to RIP.

Instead of relying on the uniform results, we leverage the fact that $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}}$ is a discrete sequence and approach (13) directly. However, since $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}}$ is generated by $\mathcal{A}$ and is thus dependent on it, the absence of a uniform result necessitates techniques to decouple them. One common way is resampling [Candès et al., 2015], but it increases the sample complexity. Inspired by [Stöger and Zhu, 2025], we used a delicate decoupling technique, which will be elaborated in the following section.

3.3 KEY DECOUPLING TECHNIQUE

Define $\boldsymbol{\Delta}_t := \boldsymbol{X}_\star - \boldsymbol{X}_t$. As illustrated in the previous section, the key is to control $\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})\,(\boldsymbol{\Delta}_t)\|_2$. We first recall a typical method to control the 2-norm of a general random matrix $\boldsymbol{M} \in \mathbb{R}^{d_1 \times d_2}$ [Vershynin, 2018]. Define $\mathbb{S}^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$ and $\mathbb{S} := \mathbb{S}^{d_1-1} \times \mathbb{S}^{d_2-1}$. We can construct an ε-net $\mathcal{N}_1 \in \mathbb{S}^{d_1-1}$ and an ε-net $\mathcal{N}_2 \in \mathbb{S}^{d_2-1}$ with $\varepsilon = \frac{1}{4}$, and let

$$\mathcal{N} := \mathcal{N}_1 \times \mathcal{N}_2 \in \mathbb{S}. \tag{14}$$

It is well known that the size of ε-net for $\mathbb{S}^{d-1}$ can be smaller than $(\frac{3}{\varepsilon})^d$, so $|\mathcal{N}| \le 12^{d_1+d_2}$. Then we have:

$$\begin{aligned}\|\boldsymbol{M}\|_2 &= \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{S}} \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{M}\right\rangle\right| \\ &\le \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}} \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{M}\right\rangle\right| + \frac{2}{4}\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{S}} \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{M}\right\rangle\right|,\end{aligned}$$

which imples

$$\|\boldsymbol{M}\|_2 \le 2\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}} \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{M}\right\rangle\right|.$$

Substituting $\boldsymbol{M} = (\mathcal{A}^*\mathcal{A} - \mathcal{I})\,(\boldsymbol{\Delta}_t)$, we turn to estimate $\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}} \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\,(\boldsymbol{\Delta}_t)\right\rangle\right|$.

For any $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}$, we have

$$\begin{aligned}&|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\,(\boldsymbol{\Delta}_t)\right\rangle| \\ &\le |\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}\boldsymbol{\Delta}_t\right)\right\rangle| \\ &+ \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}\boldsymbol{\Delta}_t\right)\right\rangle\right|.\end{aligned}$$

The first term on the right-hand side is smaller than $O(\sqrt{\frac{r(d_1+d_2)}{m}})\|\boldsymbol{\Delta}_t\|_2$ by (7) in Lemma 1 if RIP is satisfied, and the second one equals to

$$I := \left|\frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T\rangle\langle \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{A}_i), \boldsymbol{\Delta}_t\rangle\right|. \tag{15}$$

We define

$$\boldsymbol{A}_i^{(\boldsymbol{x},\boldsymbol{y})} := \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{A}_i) = \boldsymbol{A}_i - \left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{A}_i\right\rangle \boldsymbol{x}\boldsymbol{y}^T. \tag{16}$$

Using the rotation invariance property of Gaussian random variables, $\{\boldsymbol{A}_i^{(\boldsymbol{x},\boldsymbol{y})}\}_{i=1}^m$ are stochastically independent of $\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T\rangle\}_{i=1}^m$. If $\boldsymbol{\Delta}_t$ is independent of $\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T\rangle\}_{i=1}^m$, it is not difficult to deal with it.

Lemma 3. *For any $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}$ and $\boldsymbol{Z}$ independent of $\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T\rangle\}_{i=1}^m$, it holds with probability at least $1 - 2\exp(-8(d_1+d_2))$ that*

$$\begin{aligned}&\left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right)\right\rangle\right| \\ &\le 4\sqrt{\frac{d_1+d_2}{m}}\left\|\mathcal{A}(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z}))\right\|_2\end{aligned} \tag{17}$$

Proof. Under the assumption, $\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T\rangle\}_{i=1}^m$ are independent of $\left\{\left\langle \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{A}_i), \boldsymbol{Z}\right\rangle\right\}_{i=1}^m$. Then, for all $x > 0$, with probability at least $1 - 2\exp\left(-x^2/2\right)$,

$$\begin{aligned}&\left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right)\right\rangle\right| \\ &= \left|\frac{1}{m}\sum_{i=1}^m \left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{A}_i\right\rangle\left\langle \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{A}_i), \boldsymbol{Z}\right\rangle\right| \\ &\le \frac{x}{m}\sqrt{\sum_{i=1}^m \left\langle \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{A}_i), \boldsymbol{Z}\right\rangle^2} \\ &= \frac{x}{\sqrt{m}}\left\|\mathcal{A}(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z}))\right\|_2.\end{aligned}$$

The inequality follows from the fact that, conditioning on $\{\langle \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{A}_i), \boldsymbol{Z}\rangle\}_{i=1}^m$, $\sum_{i=1}^m \left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{A}_i\right\rangle\left\langle \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{A}_i), \boldsymbol{Z}\right\rangle$ is a Gaussian variable with mean 0 and variance $\sum_{i=1}^m \left\langle \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{A}_i), \boldsymbol{Z}\right\rangle^2$. Then it holds directly from the tail probability of Gaussian random variables. Choose $x = 4\sqrt{d_1+d_2}$, and the failure probability is at most $2\exp\left(-8(d_1+d_2)\right)$. □

We assume that $\boldsymbol{Z}$ has a rank less than $2r$ here, since all the matrices we care about in this section have rank less than $2r$. If we rely solely on RIP, we can bound this term as $O(\sqrt{\frac{r^2(d_1+d_2)}{m}})\|\boldsymbol{Z}\|_2$ using (8). In contrast, this lemma converts it into the right-hand side of (17), and we can eliminate the factor r and bound the term as $O(\sqrt{\frac{r(d_1+d_2)}{m}})\|\boldsymbol{Z}\|_2$ using (7).

However, we can not take $\boldsymbol{Z} = \boldsymbol{\Delta}_t$ since the $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}}$ is generated by $\mathcal{A}$ and thus dependent on $\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T\rangle\}_{i=1}^m$. To relieve the statistical dependence between $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}}$ and $\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T\rangle\}_{i=1}^m$, the central idea is to introduce a virtual sequence $\{\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\}_{t\in\mathbb{N}}$ that is independent of $\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T\rangle\}_{i=1}^m$ to approximate the real sequence $\{\boldsymbol{X}_t\}_{t\in\mathbb{N}}$.

To this end, we construct a modified measurement operator $\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})} : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^{m+1}$ that is statistically independent of $\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T\rangle\}$ to approximate $\mathcal{A}$ as follows:

$$\left[\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}(\boldsymbol{Z})\right]_i := \begin{cases}\frac{1}{\sqrt{m}}\langle \boldsymbol{A}_i^{(\boldsymbol{x},\boldsymbol{y})}, \boldsymbol{Z}\rangle, & \text{for } i \in [m], \\ \langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{Z}\rangle, & \text{for } i = m+1.\end{cases}$$

The first m terms are Gaussian random measurements of $\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})$ and independent of $\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T\rangle\}_{i=1}^m$ by (16), and the $m+1$-th term is introduced to collect the information of $\mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})$ deterministically. From this construction, $\mathbb{E}\mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})} = \mathbb{E}(\mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T} + \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}\mathcal{A}^*\mathcal{A}\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}) = \mathcal{I} = \mathbb{E}\mathcal{A}^*\mathcal{A}$, which means $\mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}$ approximates $\mathcal{A}^*\mathcal{A}$ well in terms of expectation. For more properties of $\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}$ and its relationship with $\mathcal{A}$, see Lemma 13 in the Appendix.

Finally we define the virtual sequence $\{\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\}_{t\in\mathbb{N}}$ to be the sequence generated by Algorithm 1 with input data

$\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}$ and $\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}(\boldsymbol{X}_\star)$ as follows: for $t = 0$,

$$\boldsymbol{X}_0^{(\boldsymbol{x},\boldsymbol{y})} = \mathcal{H}_r\left(\mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\left(\boldsymbol{X}_\star\right)\right),$$

and, for $t \geq 0$,

$$\begin{aligned} \boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})} &= \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} \\ &\quad - \mu \mathcal{P}_{\mathbb{T}_t^{(\boldsymbol{x},\boldsymbol{y})}} \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\left(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{X}_\star\right), \\ \boldsymbol{X}_{t+1}^{(\boldsymbol{x},\boldsymbol{y})} &= \mathcal{H}_r(\boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})}), \end{aligned}$$

where $\mathbb{T}_t^{(\boldsymbol{x},\boldsymbol{y})}$ is the tangent space of the manifold $\mathbb{M}_r$ at $\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}$.

Consequently, $\{\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\}_{t \in \mathbb{N}}$ is independent of $\left\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T \rangle\right\}_{i=1}^m$ and approximates $\{\boldsymbol{X}_t\}_{t \in \mathbb{N}}$. The stochastic independence properties and approximation properties inherent in the construction of the virtual sequence significantly benefit the analysis. A straightforward analysis yields a corollary of Lemma 3 specified for our virtual sequence $\{\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\}_{t \in \mathbb{N}}$. For simplicity, we denote $[m] = \{1, \ldots, m\}$ and $[m] - 1 = \{0, \ldots, m-1\}$.

Lemma 4. *With probability at least* $1 - 2\exp(-2(d_1 + d_2))$, *it holds that*

$$\begin{aligned} &\left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{X}_\star - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right)\right\rangle\right| \\ &\leq 4\sqrt{\frac{d_1 + d_2}{m}} \left\|\mathcal{A}\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{X}_\star - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right)\right\|_2, \\ &\qquad \forall\, (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{N} \text{ in } (14) \text{ and } t \in [12^{d_1+d_2}] - 1. \end{aligned} \tag{18}$$

Proof. Notice that $\boldsymbol{X}_\star - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}$ is independent of $\{\langle \boldsymbol{A}_i, \boldsymbol{x}\boldsymbol{y}^T \rangle\}_{i=1}^m$, so we can take $\boldsymbol{Z} = \boldsymbol{X}_\star - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}$ for any t and $(\boldsymbol{x}, \boldsymbol{y})$ in Lemma 3. We simply take a union bound, and then (18) is satisfied with probability at least $1 - 2|\mathcal{N}|T\exp(-8(d_1 + d_2)) \geq 1 - 2\exp(-2(d_1 + d_2))$. □

Using Lemma 4, we can finally get an estimation of $\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{X}_\star - \boldsymbol{X}_t)\|_2$:

Lemma 5. *Let* $\mathcal{N}$ *be in* (14). *Let* $\{\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\}_{t \in \mathbb{N}}$ *be the virtual sequence constructed for* $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{N}$. *Assume that* $\mathcal{A}$ *satisfies RIP of rank* $6r$, *and let* $\delta = \delta_{6r} \leq 1$. *Assume that* (18) *holds. Then we have*

$$\begin{aligned} &\forall t \in [12^{d_1+d_2}] - 1, \quad \|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{X}_\star - \boldsymbol{X}_t)\|_2 \\ &\leq \sigma_1 \|\boldsymbol{X}_\star - \boldsymbol{X}_t\|_2 + \sigma_2 \sup_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}} \left\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F, \end{aligned} \tag{19}$$

where $\sigma_1 = 16\sqrt{\frac{2r(d_1+d_2)}{m}} + 2\delta$ *and* $\sigma_2 = 4\delta + 16\sqrt{\frac{d_1+d_2}{m}}$.

Its proof is deferred to Appendix C. When $\mathcal{A}$ is Gaussian measurement operator, $\delta = O(\frac{r(d_1+d_2)}{m})$ with high probability from Lemma 1, so σ_1 and σ_2 can become arbitrarily close to 0 as m increases. This result approaches (13), with an additional error term arising from the distance between the real and virtual sequences. Consequently, we are going to control both the distances from $\boldsymbol{X}_t$ to $\boldsymbol{X}_\star$ and from $\boldsymbol{X}_t$ to $\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}$ at initialization and demonstrate that these distances contract during the iterations. Although (19) is not uniform and holds for at most T steps, it enables (12) with $T = O(\ln r)$, allowing convergence analysis in [Wei et al., 2016] available with $m = \Omega(\kappa^2 r(d_1 + d_2))$.

3.4 PROOF OF THE MAIN THEOREM

In this section, we provide of proof of Theorem 1. The proof is divided into three phases: the initialization, the first T steps to meet $\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F \ll \sigma_{\min}(\boldsymbol{X}_\star)$ in (12), and the subsequence steps where the linear convergence in Frobenius norm is guaranteed [Wei et al., 2016]. For simplicity, we denote for $t \in \mathbb{N}$:

$$E_t := \|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_2 + \sup_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}} \left\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F. \tag{20}$$

Phase I: Initialization. We show in Lemma 6 (whose proof is in Appendix D) that E_0 can be small with high probability provided $m = \Omega(\kappa^2 r(d_1 + d_2))$. This is a non-PSD version of [Stöger and Zhu, 2025, Lemma 4.1].

Lemma 6. *Let* $c_1 > 0$ *be arbitrarily given. Then there exists a constant* $C > 0$ *such that when* $m \geq C\kappa^2 r(d_1 + d_2)$, *with probability at least* $1 - 4\exp(-(d_1 + d_2))$, *it holds that:*

$$E_0 \leq c_1 \sigma_{\min}(\boldsymbol{X}_\star), \tag{21}$$

where E_0 *is defined in* (20).

Phase II: Contraction in 2-norm in the first T steps. Using (18) and (19), we estimate E_t by induction starting from (21).

Lemma 7. *Let* c_1 *be an absolute constant such that* $c_1 \in (0, 0.001)$. *Assume that* $\mathcal{A}$ *satisfies RIP of rank* $6r$, *and let* $\delta = \delta_{6r} < \frac{1}{24}c_1$. *Assume that* (18) *and* (21) *hold. Then there exists a constant* $C > 0$ *depending on* c_1 *only such that when* $m \geq C\kappa^2 r(d_1 + d_2)$,

$$E_t \leq (1000c_1)^t c_1 \sigma_{\min}(\boldsymbol{X}_\star), \qquad \forall\ t \in [12^{d_1+d_2}]. \tag{22}$$

This lemma is critical, and its proof differs significantly from the parallel one for factorized gradient descent in [Stöger and Zhu, 2025]. Specifically, the gradient is projected onto the tangent space of $\boldsymbol{X}_t$, requiring careful analysis of the projection operator, as detailed in Lemma 11 in the Appendix. Additionally, our algorithm incorporates a hard-thresholding operator after the gradient descent step, for which Lemma 10 is necessary to bound the error introduced by thresholding. The detailed proof is provided in Appendix E.

By choosing c_1 sufficiently small and $T = O(\ln r)$, Lemma 7 implies $\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F \leq \sqrt{2r}\,\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_2 \leq \sqrt{2r}E_T \ll \sigma_{\min}(\boldsymbol{X}_\star)$.

Phase III: Contraction in Frobenius norm in the subsequent steps. With $\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F \ll \sigma_{\min}(\boldsymbol{X}_\star)$, we can directly apply the result from [Wei et al., 2016] to establish the convergence of $\boldsymbol{X}_t$ to $\boldsymbol{X}_\star$ in Frobenius norm with $m = \Omega(\kappa^2 r(d_1 + d_2))$. For completeness, we introduce the following lemma (whose proof is in Appendix E).

Lemma 8 ([Wei et al., 2016])**.** *Let c_2 be an arbitrary constant that satisfies $0 < 6c_2 < 1$. Assume that the measurement operator $\mathcal{A}$ satisfies the RIP of rank $6r$ with constant $\delta_{6r} < c_2$. Assume that*

$$\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F \le c_2 \sigma_{\min}(\boldsymbol{X}_\star) \tag{23}$$

for some $T \in \mathbb{N}$. Then it holds for all $t \ge T$ that

$$\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_F \le (6c_2)^{t-T} \|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F.$$

Combining these three phases, we can give the proof of Theorem 1.

Proof of Theorem 1. Recall that $\rho \in (0, 1)$ is the target convergence rate, and we have denoted $E_0 = \|\boldsymbol{X}_0 - \boldsymbol{X}_\star\|_2 + \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}} \|\boldsymbol{X}_0 - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F$. We define the constants $c_2 = \rho/6$, $c_1 = \min\{\frac{\rho}{1000}, \frac{1}{2000}, e^{\frac{\ln 2 - \frac{1}{2}}{2}} c_2\} < 1$, and $\delta = \min\{c_2, \frac{1}{24}c_1\}$.

The proof relies on the following events:

- $\mathcal{A}$ satisfies RIP of rank $6r$ with $\delta_{6r} < \delta$. By Lemma 1, this event holds with probability at least $1 - \exp(-(d_1 + d_2))$ provided that $m \ge C'\kappa^2 r(d_1 + d_2)$.
- The inequality (18) holds. By Lemma 4, this occurs with probability at least $1 - 2\exp(-2(d_1 + d_2))$.
- The initial error satisfies $E_0 \le c_1 \sigma_{\min}(\boldsymbol{X}_\star)$, i.e., (21) holds. By Lemma 6, this is true with probability at least $1 - 4\exp(-(d_1 + d_2))$ when $m \ge C''\kappa^2 r(d_1 + d_2)$.

Applying a union bound, the probability that all these three events occur simultaneously is at least $1 - 7\exp(-(d_1 + d_2))$.

Assuming these events hold, we proceed with the proof. Combining the RIP, (18), and (21), it follows from Lemma 7 that for all $t \in [12^{d_1+d_2}]$,

$$\begin{aligned} \|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_F &\le \sqrt{2r}\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_2 \le \sqrt{2r}\rho_1^t E_0 \\ &\le \sqrt{2r}\rho_1^t c_1 \sigma_{\min}(\boldsymbol{X}_\star) \le \sqrt{2r}\rho^t \sigma_{\min}(\boldsymbol{X}_\star), \end{aligned} \tag{24}$$

where $\rho_1 = 1000c_1 \le \rho < 1$, and the number of measurements satisfies $m \ge C'''\kappa^2 r(d_1 + d_2)$.

Let $T = \ln(2r) \le 12^{d_1+d_2}$. A straightforward calculation shows that

$$\frac{\frac{1}{2}\ln 2r + \ln\frac{c_1}{c_2}}{\ln 2r} \le \frac{1}{2} + 2\ln\frac{c_1}{c_2} \overset{c_1 < e^{\frac{\ln 2 - \frac{1}{2}}{2}} c_2}{<} \ln 2 < \ln\frac{1}{1000c_1}.$$

This implies $\sqrt{2r}\rho_1^T c_1 \sigma_{\min}(\boldsymbol{X}_\star) < c_2\sigma_{\min}(\boldsymbol{X}_\star)$, which ensures that (23) holds. Using this result and the RIP, Lemma 8 guarantees that for $t \ge T$,

$$\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_F \le \rho^{t-T}\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F. \tag{25}$$

Combining (24) for $t \in [T]$ and (25) for $t \ge T$, we obtain the convergence result (5).

To conclude, we determine the number of measurements required by taking the maximum of the conditions on m throughout the proof:

$$m \ge C\kappa^2 r(d_1 + d_2),$$

where $C = \max\{C', C'', C'''\}$. □

4 EXPERIMENT

In this section, we evaluate the performance of the Riemannian Gradient Descent (RGD) algorithm, as described in Algorithm 1, on Gaussian matrix sensing problems. We present phase transition diagrams to illustrate the relationship between sample complexity m and the rank r or condition number κ of $\boldsymbol{X}_\star$. Furthermore, we compare the efficiency of RGD with factorized gradient descent (GD) methods in ill-conditioned settings.

Phase Transition Diagram We study the phase transition behavior of RGD by systematically varying the rank r and the number of measurements m in Gaussian matrix sensing problems, with fixed dimensions ($d_1 = 60$, $d_2 = 80$) and condition number $\kappa = 2$. For each (r, m) pair, we perform 20 independent trials. A trial is considered successful if $\frac{\|\boldsymbol{X}_N - \boldsymbol{X}_\star\|_F}{\|\boldsymbol{X}_\star\|_F} \le 10^{-2}$ after $N = 100$ iterations. This setup allows us to empirically estimate the success rate as a function of m and r. Figure 1 (left) reveals a sharp phase transition, where the minimal sample complexity m required for successful recovery increases linearly with the rank r. We further examine how the condition number κ affects the sample complexity m. Keeping the dimensions fixed as before and setting the rank $r = 10$, we vary κ from 1 to 280. The nearly horizontal boundary in Fig. 1 (right) indicates that increasing the condition number κ has little effect on the sample complexity m required for successful recovery. Explaining this empirical insensitivity may require new theoretical insights.

Comparison with Factorized GD We also compare the convergence speed of RGD and factorized GD in ill-conditioned settings. We use square matrices ($d_1 = d_2 = 80$), rank $r = 15$, $m = 13200$, and condition number $\kappa = 20$. Stepsizes are set to $\mu = 1$ for RGD, and $\mu = 0.9$ (empirically optimal) and $\mu = 1$ for GD. The ground truth $\boldsymbol{X}_\star$ is PSD, following [Stöger and Zhu, 2025]. As shown in Fig. 2, RGD is stable and converges rapidly, while GD becomes unstable at larger stepsizes.

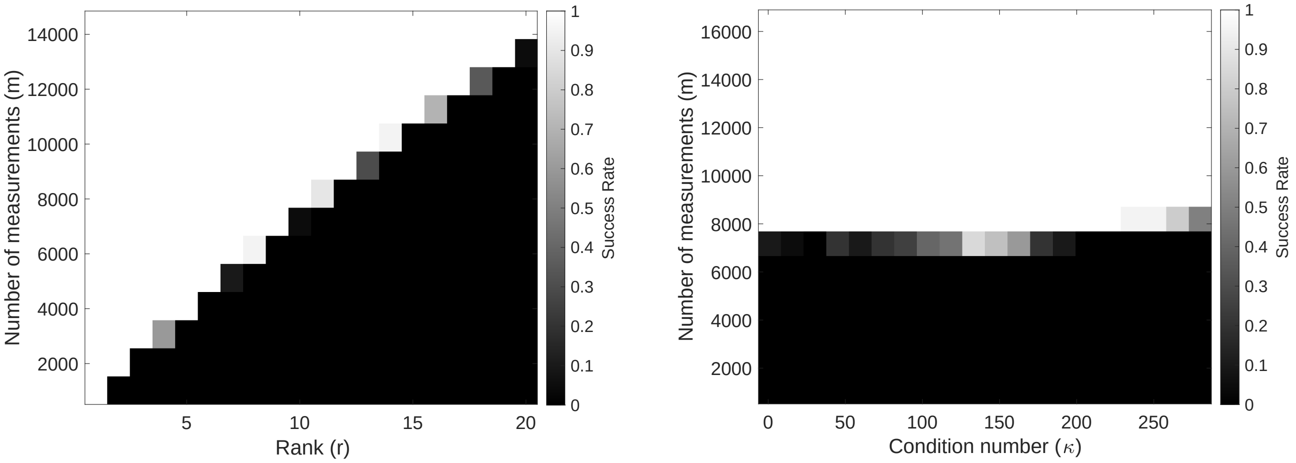

Figure 1: Phase transition diagrams for Gaussian matrix sensing: (left) m vs. r; (right) m vs. κ. Black indicates failure; white indicates success.

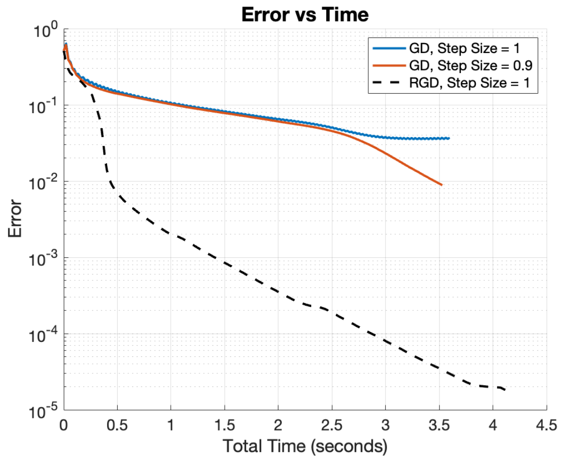

Figure 2: Error versus time for RGD and factorized GD. GD with a large stepsize ($\mu = 1$, blue) oscillates, while RGD (dashed) is stable and efficient. GD with the empirically optimal stepsize ($\mu = 0.9$, red) is also shown.

5 CONCLUSION AND OPEN PROBLEMS

In this work, we proved that the Riemannian gradient descent algorithm with spectral initialization can recover a rank-r matrix $\boldsymbol{X}_\star$ of size $d_1 \times d_2$ using $O(r(d_1+d_2)\kappa^2)$ Gaussian measurements, which is optimal among fast non-convex methods. Furthermore, its convergence rate is independent of κ, making it computationally efficient even when $\boldsymbol{X}_\star$ is ill-conditioned.

Convex approaches based on nuclear norm minimization need only $\Omega(r(d_1+d_2))$ samples in the matrix sensing scenario, while our result is suboptimal by a factor of κ^2. As a local search algorithm operating on the rank-r matrix manifold, our RGD method's performance naturally depends on the geometric properties at the solution point $\boldsymbol{X}_\star$. Existing analyses of the embedded manifold's local geometry (e.g., Lemma 5 in [Luo and Trillos, 2022]) demonstrate that the curvature at $\boldsymbol{X}_\star$ scales with the condition number κ. This relationship is further evidenced in our two lemmas in Appendix A.3, which show κ-dependence in tangent space perturbations. This dependence on κ is a common feature of fast non-convex methods, as shown in Table 1. Interestingly, our experiment result suggests that m might decouple from κ, opening pathways for future research into improved initialization strategies or refined geometric analyses.

Moreover, the proof relies on a decoupling technique that critically depends on the rotational invariance of Gaussian random variables, posing an interesting and challenging direction for future research to establish optimal sample complexity in other settings, such as matrix completion and quantum state tomography.

References

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Jianfeng Cai and Ke Wei. Solving systems of phaseless equations via riemannian optimization with optimal sampling complexity. *Journal of Computational Mathematics*, 42(3):755–783, 2024.

E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theor.*, 57 (4):2342–2359, April 2011. ISSN 0018-9448. doi: 10.1109/TIT.2011.2111771. URL https://doi.org/10.1109/TIT.2011.2111771.

Emmanuel J. Candes and Terence Tao. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Transactions on Information Theory*, 56 (5):2053–2080, May 2010. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2010.2044061. URL http://ieeexplore.ieee.org/document/5452187/.

Emmanuel J. Candès, Thomas Strohmer, and Vladislav Voroninski. PhaseLift: Exact and Stable Signal Recovery from Magnitude Measurements via Convex Programming. *Communications on Pure and Applied Mathematics*, 66 (8):1241–1274, August 2013. ISSN 00103640. doi: 10.1002/cpa.21432. URL https://onlinelibrary.wiley.com/doi/10.1002/cpa.21432.

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, September 2015. ISSN 1063-5203. doi: 10.1016/j.acha.2014.09.004. URL https://www.sciencedirect.com/science/article/pii/S1063520314001201.

Vasileios Charisopoulos, Yudong Chen, Damek Davis, Mateo Díaz, Lijun Ding, and Dmitriy Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *Foundations of Computational Mathematics*, 21(6):1505–1593, 2021.

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

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5): 566–806, 2021.

Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, October 2019. ISSN 1941-0476. doi: 10.1109/TSP.2019.2937282. URL https://ieeexplore.ieee.org/abstract/document/8811622. Conference Name: IEEE Transactions on Signal Processing.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

David Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, March 2011. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2011.2104999. URL http://arxiv.org/abs/0910.1879. arXiv:0910.1879 [cs].

Ming-Chien Hsu, En-Jui Kuo, Wei-Hsuan Yu, Jian-Feng Cai, and Min-Hsiu Hsieh. Quantum State Tomography via Nonconvex Riemannian Gradient Descent. *Physical Review Letters*, 132(24):240804, June 2024. doi: 10.1103/PhysRevLett.132.240804. URL https://link.aps.org/doi/10.1103/PhysRevLett.132.240804. Publisher: American Physical Society.

Wen Huang, Kyle A Gallivan, and Xiangxiong Zhang. Solving phaselift by low-rank riemannian optimization methods for complex semidefinite constraints. *SIAM Journal on Scientific Computing*, 39(5):B840–B859, 2017.

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

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

Peter Jung, Felix Krahmer, and Dominik Stöger. Blind demixing and deconvolution at near-optimal rate. *IEEE Transactions on Information Theory*, 64(2):704–727, 2017.

Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Advances in neural information processing systems*, 22, 2009.

Yuetian Luo and Nicolas Garcia Trillos. Nonconvex matrix factorization is geodesically convex: Global landscape analysis for fixed-rank matrix optimization from a riemannian perspective. *arXiv preprint arXiv:2209.15130*, 2022.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, January 2010. ISSN 0036-1445, 1095-7200. doi: 10.1137/070697835. URL http://epubs.siam.org/doi/10.1137/070697835.

Dominik Stöger and Yizhe Zhu. Non-convex matrix sensing: Breaking the quadratic rank barrier in the sample complexity. In *Proceedings of the 38th Annual Conference on Learning Theory (COLT 2025)*. PMLR, September 2025. URL http://arxiv.org/abs/2408.13276. arXiv:2408.13276 [stat].

Ruoyu Sun and Zhi-Quan Luo. Guaranteed Matrix Completion via Non-convex Factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, November 2016. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2016.2598574. URL http://arxiv.org/abs/1411.8003. arXiv:1411.8003 [cs].

Jared Tanner and Ke Wei. Normalized iterative hard thresholding for matrix completion. *SIAM Journal on Scientific Computing*, 35(5):S104–S125, 2013.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22 (150):1–63, 2021.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International conference on machine learning*, pages 964–973. PMLR, 2016.

Bart Vandereycken. Low-Rank Matrix Completion by Riemannian Optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. doi: 10.1137/110845768. URL https://doi.org/10.1137/110845768. _eprint: https://doi.org/10.1137/110845768.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Dietrich Von Rosen. Moments for the inverted wishart distribution. *Scandinavian Journal of Statistics*, pages 97–109, 1988.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, March 1972. ISSN 1572-9125. doi: 10.1007/BF01932678. URL https://doi.org/10.1007/BF01932678.

Ke Wei, Jian-Feng Cai, Tony F Chan, and Shingyu Leung. Guarantees of riemannian optimization for low rank matrix recovery. *SIAM Journal on Matrix Analysis and Applications*, 37(3):1198–1222, 2016.

Shixin Zheng, Wen Huang, Bart Vandereycken, and Xiangxiong Zhang. Riemannian optimization using three different metrics for hermitian psd fixed-rank constraints. *Computational Optimization and Applications*, pages 1–50, 2025.

Pini Zilber and Boaz Nadler. Gnmr: A provable one-line algorithm for low rank matrix recovery. *SIAM journal on mathematics of data science*, 4(2):909–934, 2022.

Fast Non-convex Matrix Sensing with Optimal Sample Complexity (Supplementary Material)

Jian-Feng Cai [1] **Tong Wu** [1] **Ruizhe Xia**[1]

[1]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China

A PRELIMINARY THEOREMS AND LEMMAS

In this section, we present some preliminary theorems and lemmas, which are fundamental and will be frequently used in our proofs.

A.1 SUPPORTING THEOREMS

We begin with Weyl's inequality, which is useful for estimating the singular values of a perturbed matrix.

Theorem 2 (Weyl's inequality). *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2}$ be two matrices with singular values $\sigma_1(\boldsymbol{A}) \geq \sigma_2(\boldsymbol{A}) \geq \cdots \geq \sigma_{\min\{d_1,d_2\}}(\boldsymbol{A})$ and $\sigma_1(\boldsymbol{B}) \geq \sigma_2(\boldsymbol{B}) \geq \cdots \geq \sigma_{\min\{d_1,d_2\}}(\boldsymbol{B})$. Then for any $i \in [\min\{d_1, d_2\}]$ it holds that:*

$$|\sigma_i(\boldsymbol{A}) - \sigma_i(\boldsymbol{B})| \leq \|\boldsymbol{A} - \boldsymbol{B}\|_2.$$

The following Bernstein inequality helps control the tail probabilities of certain random events.

Theorem 3 ([Vershynin, 2018, Theorem 2.8.1], Bernstein's inequality). *Let $X_1, \ldots, X_N$ be independent, mean-zero, sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{\sum_{i=1}^{N}\|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right)\right], \tag{26}$$

where $\|\cdot\|_{\psi_1}$ is the sub-exponential norm and $c > 0$ is an absolute constant.

A.2 PERTURBATION BOUNDS FOR EIGENSPACE

For a matrix $\boldsymbol{Z} \in \mathbb{R}^{d_1 \times d_2}$ with SVD $\boldsymbol{Z} = \boldsymbol{U_Z \Sigma_Z V_Z^\top}$, we let $\boldsymbol{U}_{\boldsymbol{Z},r} \in \mathbb{R}^{d_1 \times r}$ be the matrix consisting of the first r columns of $\boldsymbol{U_Z}$, and $\boldsymbol{U}_{\boldsymbol{Z},r,\perp} \in \mathbb{R}^{d_1 \times (d_1 - r)}$ be the matrix consisting of the remaining $d_1 - r$ columns. The matrices $\boldsymbol{V}_{\boldsymbol{Z},r}$ and $\boldsymbol{V}_{\boldsymbol{Z},r,\perp}$ are defined similarly. The matrix $\boldsymbol{\Sigma}_{\boldsymbol{Z},r}$ is an $r \times r$ diagonal matrix consisting of the first r singular values of $\boldsymbol{\Sigma_Z}$. The singular values of $\boldsymbol{Z}$ are ordered such that their magnitudes are decreasing, i.e., $\sigma_1(\boldsymbol{Z}) \geq \sigma_2(\boldsymbol{Z}) \geq \ldots \geq \sigma_{\min\{d_1,d_2\}}(\boldsymbol{Z})$. For simplicity, we use $\boldsymbol{U}_1$ to denote $\boldsymbol{U}_{\boldsymbol{Z}_1}$, $\boldsymbol{U}_{1,r}$ to denote $\boldsymbol{U}_{\boldsymbol{Z}_1,r}$, and $\boldsymbol{U}_{2,r}$ to denote $\boldsymbol{U}_{\boldsymbol{Z}_2,r}$. Other notations are simplified similarly.

The following lemma bounds the perturbation of the subspace spanned by the first r singular vectors of $\boldsymbol{Z}_1$ in terms of the spectral gap of $\boldsymbol{Z}_1$ and the perturbation on the matrix itself:

[*]This work is partially supported by Hong Kong Research Grant Council GRFs 16307023 and 16306124.

[†]Main contributor.

Lemma 9 ([Wedin, 1972], Non-symmetric version of Davis-Kahan inequality). *Let $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2 \in \mathbb{R}^{d_1\times d_2}$ be two matrices with singular value decompositions*

$$\boldsymbol{Z}_1 = \begin{bmatrix}\boldsymbol{U}_{1,r} & \boldsymbol{U}_{1,r,\perp}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{1,r} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{1,r,\perp}\end{bmatrix}\begin{bmatrix}\boldsymbol{V}_{1,r}^T\\ \boldsymbol{V}_{1,r,\perp}^T\end{bmatrix},$$

and

$$\boldsymbol{Z}_2 = \boldsymbol{Z}_1 + \boldsymbol{\Delta} = \begin{bmatrix}\boldsymbol{U}_{2,r} & \boldsymbol{U}_{2,r,\perp}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{2,r} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{2,r,\perp}\end{bmatrix}\begin{bmatrix}\boldsymbol{V}_{2,r}^T\\ \boldsymbol{V}_{2,r,\perp}^T\end{bmatrix},$$

respectively. If $\sigma_r(\boldsymbol{Z}_1) > \sigma_{r+1}(\boldsymbol{Z}_1)$ and

$$\|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\|_2 \le \left(1 - \frac{1}{\sqrt{2}}\right)(\sigma_r(\boldsymbol{Z}_1) - \sigma_{r+1}(\boldsymbol{Z}_1)), \tag{27}$$

then

$$\max\left\{\left\|\boldsymbol{U}_{2,r,\perp}^\top \boldsymbol{U}_{1,r}\right\|_F, \left\|\boldsymbol{V}_{2,r,\perp}^\top \boldsymbol{V}_{1,r}\right\|_F\right\} \le \frac{\sqrt{2}\left(\left\|\boldsymbol{U}_1^T\boldsymbol{\Delta}\right\|_F + \|\boldsymbol{\Delta}\boldsymbol{V}_1\|_F\right)}{\sigma_r(\boldsymbol{Z}_1) - \sigma_{r+1}(\boldsymbol{Z}_1)}.$$

The following lemma bounds the distance between two matrices after applying the thresholding operator $\mathcal{H}_r$, assuming they are sufficiently close. To use this result, we first provide a lower bound on the spectral gap of $\boldsymbol{Z}_1$ and show that it is large enough compared to both $\|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\|_2$ and $\sigma_r(\boldsymbol{Z}_1)$.

Lemma 10. *Let $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ satisfy the same conditions as in Lemma 9. Assume further that the spectral gap is large enough such that*

$$s := \sigma_r(\boldsymbol{Z}_1) - \sigma_{r+1}(\boldsymbol{Z}_1) \ge \frac{1}{c_0}\sigma_{r+1}(\boldsymbol{Z}_1)$$

for some constant $c_0 > 0$. Then, there exist constants C_1 and C_2 depending only on c_0 and satisfying $C_1 \le C_2 \le 6c_0 + 10$ such that

$$\|\mathcal{H}_r(\boldsymbol{Z}_1) - \mathcal{H}_r(\boldsymbol{Z}_2)\|_2 \le C_1(\sigma_r(\boldsymbol{Z}_1) - \sigma_{r+1}(\boldsymbol{Z}_1)),$$

and

$$\begin{aligned}\|\mathcal{H}_r(\boldsymbol{Z}_1) - \mathcal{H}_r(\boldsymbol{Z}_2)\|_F &\le C_2\left(\|(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\boldsymbol{V}_{1,r}\|_F + \left\|\boldsymbol{U}_{1,r}^T(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\right\|_F\right)\\ &\le 2C_2\|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\|_F.\end{aligned}$$

Proof. Recall that we have defined

$$s = \sigma_r(\boldsymbol{Z}_1) - \sigma_{r+1}(\boldsymbol{Z}_1).$$

By Weyl's inequality (see Theorem 2) and (27), it follows that

$$\sigma_r(\boldsymbol{Z}_2) - \sigma_{r+1}(\boldsymbol{Z}_2) \ge \sigma_r(\boldsymbol{Z}_1) - \sigma_{r+1}(\boldsymbol{Z}_1) - 2\|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\|_2 \ge (\sqrt{2} - 1)s.$$

Therefore, for $i = 1, 2$, the rank-r approximation $\boldsymbol{Z}_{i,r} = \mathcal{H}_r(\boldsymbol{Z}_i)$ is uniquely defined, as $\sigma_r(\boldsymbol{Z}_i) > \sigma_{r+1}(\boldsymbol{Z}_i)$. Moreover, by Weyl's inequality and (27), we have

$$|\sigma_{r+1}(\boldsymbol{Z}_2)| \le |\sigma_{r+1}(\boldsymbol{Z}_1)| + (1 - 1/\sqrt{2})s \le (c_0 + 1 - 1/\sqrt{2})s.$$

Let $c := c_0 + 1 - 1/\sqrt{2}$, noting that $c > c_0$. We then derive the following estimate:

$$\begin{aligned}\|\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r}\|_2 &\le \|\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_1\|_2 + \|\boldsymbol{Z}_1 - \boldsymbol{Z}_2\|_2 + \|\boldsymbol{Z}_2 - \boldsymbol{Z}_{2,r}\|_2\\ &\le |\sigma_{r+1}(\boldsymbol{Z}_1)| + |\sigma_{r+1}(\boldsymbol{Z}_2)| + (1 - 1/\sqrt{2})(|\sigma_r(\boldsymbol{Z}_1)| - |\sigma_{r+1}(\boldsymbol{Z}_1)|)\\ &\le \underbrace{(2c + 1 - 1/\sqrt{2})}_{C_1}s,\end{aligned} \tag{28}$$

where the constant C_1 satisfies $C_1 \le 2c_0 + 3$.

Let $\boldsymbol{Z}_{1,r} = \boldsymbol{U}_{1,r}\boldsymbol{\Sigma}_{1,r}\boldsymbol{V}_{1,r}^T$ and $\boldsymbol{Z}_{2,r} = \boldsymbol{U}_{2,r}\boldsymbol{\Sigma}_{2,r}\boldsymbol{V}_{2,r}^T$ be the SVDs of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, respectively. Since

$$\|\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r}\|_F^2 \le \|(\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r})\boldsymbol{V}_1\|_F^2 = \|(\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r})\boldsymbol{V}_{1,r}\|_F^2 + \|(\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r})\boldsymbol{V}_{1,r,\perp}\|_F^2,$$

taking the square root of both sides and using $\sqrt{a^2+b^2} \leq |a| + |b|$, we obtain

$$\|\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r}\|_F \leq \|(\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r})\boldsymbol{V}_{1,r}\|_F + \|(\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r})\boldsymbol{V}_{1,r,\perp}\|_F. \tag{29}$$

We now estimate the two terms on the right-hand side separately.

- For the first term, we have:

$$\begin{aligned}
\|(\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r})\boldsymbol{V}_{1,r}\|_F &= \|(\boldsymbol{Z}_1 - \boldsymbol{Z}_{2,r})\boldsymbol{V}_{1,r}\|_F \\
&\leq \|(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\boldsymbol{V}_{1,r}\|_F + \|(\boldsymbol{Z}_2 - \boldsymbol{Z}_{2,r})\boldsymbol{V}_{1,r}\|_F \\
&= \|(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\boldsymbol{V}_{1,r}\|_F + \|\boldsymbol{U}_{2,r,\perp}\boldsymbol{\Sigma}_{2,r,\perp}\boldsymbol{V}_{2,r,\perp}^T\boldsymbol{V}_{1,r}\|_F \\
&\overset{\text{(a)}}{\leq} \left(1 + \frac{\sqrt{2}|\sigma_{r+1}(\boldsymbol{Z}_2)|}{|\sigma_r(\boldsymbol{Z}_1)| - |\sigma_{r+1}(\boldsymbol{Z}_1)|}\right) \|(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\boldsymbol{V}_{1,r}\|_F \\
&\quad + \frac{\sqrt{2}|\sigma_{r+1}(\boldsymbol{Z}_2)|}{|\sigma_r(\boldsymbol{Z}_1)| - |\sigma_{r+1}(\boldsymbol{Z}_1)|} \left\|\boldsymbol{U}_{1,r}^T(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\right\|_F \\
&\leq (1 + \sqrt{2}c)\|(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\boldsymbol{V}_{1,r}\|_F + \sqrt{2}c\left\|\boldsymbol{U}_{1,r}^T(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\right\|_F,
\end{aligned}$$

  where step (a) follows from

$$\|\boldsymbol{U}_{2,r,\perp}\boldsymbol{\Sigma}_{2,r,\perp}\boldsymbol{V}_{2,r,\perp}^T\boldsymbol{V}_{1,r}\|_F \leq |\sigma_{r+1}(\boldsymbol{Z}_2)|\|\boldsymbol{V}_{2,r,\perp}^T\boldsymbol{V}_{1,r}\|_F$$

  and Lemma 9.

- For the second term, we further split it into two parts:

$$\begin{aligned}
\|(\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r})\boldsymbol{V}_{1,r,\perp}\|_F &\leq \|\boldsymbol{U}_{1,r}^T(\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r})\boldsymbol{V}_{1,r,\perp}\|_F + \|\boldsymbol{U}_{1,r,\perp}^T(\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r})\boldsymbol{V}_{1,r,\perp}\|_F \\
&\leq \|\boldsymbol{U}_{1,r}^T(\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r})\|_F + \|\boldsymbol{U}_{1,r,\perp}^T\boldsymbol{Z}_{2,r}\boldsymbol{V}_{1,r,\perp}\|_F.
\end{aligned}$$

  The last term is estimated as:

$$\begin{aligned}
\|\boldsymbol{U}_{1,r,\perp}^T\boldsymbol{Z}_{2,r}\boldsymbol{V}_{1,r,\perp}\|_F &= \|\boldsymbol{U}_{1,r,\perp}^T\boldsymbol{U}_{2,r}\boldsymbol{\Sigma}_{2,r}\boldsymbol{V}_{2,r}^T\boldsymbol{V}_{1,r,\perp}\|_F \\
&\leq \|\boldsymbol{U}_{1,r,\perp}^T\boldsymbol{U}_{2,r}\boldsymbol{\Sigma}_{2,r}\|_2\|\boldsymbol{V}_{2,r}^T\boldsymbol{V}_{1,r,\perp}\|_F \\
&= \|\boldsymbol{U}_{1,r,\perp}^T\boldsymbol{U}_{2,r}\boldsymbol{\Sigma}_{2,r}\boldsymbol{V}_{2,r}^T\|_2\|\boldsymbol{V}_{2,r}^T\boldsymbol{V}_{1,r,\perp}\|_F \\
&\overset{\text{(a)}}{=} \|\boldsymbol{U}_{1,r,\perp}^T\boldsymbol{Z}_{2,r}\|_2\|\boldsymbol{V}_{2,r,\perp}^T\boldsymbol{V}_{1,r}\|_F \\
&= \|\boldsymbol{U}_{1,r,\perp}^T(\boldsymbol{Z}_{2,r} - \boldsymbol{Z}_{1,r})\|_2\|\boldsymbol{V}_{2,r,\perp}^T\boldsymbol{V}_{1,r}\|_F \\
&\leq \|\boldsymbol{Z}_{2,r} - \boldsymbol{Z}_{1,r}\|_2\|\boldsymbol{V}_{2,r,\perp}^T\boldsymbol{V}_{1,r}\|_F \\
&\overset{\text{(b)}}{\leq} \frac{\sqrt{2}(2c + 1 - 1/\sqrt{2})s}{s}\left(\|(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\boldsymbol{V}_{1,r}\|_F + \left\|\boldsymbol{U}_{1,r}^T(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\right\|_F\right) \\
&= \sqrt{2}(2c + 1 - 1/\sqrt{2})\left(\|(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\boldsymbol{V}_{1,r}\|_F + \left\|\boldsymbol{U}_{1,r}^T(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\right\|_F\right),
\end{aligned}$$

  where step (a) uses $\|\boldsymbol{V}_{2,r}^T\boldsymbol{V}_{1,r,\perp}\|_F = \|\boldsymbol{V}_{2,r,\perp}^T\boldsymbol{V}_{1,r}\|_F$ [Chen et al., 2021, Lemma 2.5], and step (b) follows from (28).

Combining these estimates, we obtain

$$\|\boldsymbol{Z}_{1,r} - \boldsymbol{Z}_{2,r}\|_F \leq C_2\left(\|(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\boldsymbol{V}_{1,r}\|_F + \left\|\boldsymbol{U}_{1,r}^T(\boldsymbol{Z}_1 - \boldsymbol{Z}_2)\right\|_F\right),$$

where $C_2 = \sqrt{2}(2c + 1 - 1/\sqrt{2}) + (1 + \sqrt{2}c) \leq 6c_0 + 10$ is a constant. □

A.3 BOUNDS ON THE DISTANCE BETWEEN PROJECTIONS

We introduce key lemmas used in the convergence analysis of the RGD algorithm, which have been stated and proved in [Wei et al., 2016]. The following result bounds the projection distance between the singular vector subspaces of two matrices:

Lemma 11. *Let $\boldsymbol{X}_t$ and $\boldsymbol{X}$ be two rank-r matrices with compact SVDs $\boldsymbol{X}_t = \boldsymbol{U}_t \boldsymbol{\Sigma}_t \boldsymbol{V}_t^T$ and $\boldsymbol{X} = \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^T$, respectively.*

1. *The distance between the projection matrices of their singular vector subspaces satisfies the following bounds:*

$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^T - \boldsymbol{U}\boldsymbol{U}^T\right\|_2 \leq \frac{\|\boldsymbol{X}_t - \boldsymbol{X}\|_2}{\sigma_{\min}(\boldsymbol{X})}, \quad \left\|\boldsymbol{V}_t\boldsymbol{V}_t^T - \boldsymbol{V}\boldsymbol{V}^T\right\|_2 \leq \frac{\|\boldsymbol{X}_t - \boldsymbol{X}\|_2}{\sigma_{\min}(\boldsymbol{X})};$$
$$\left\|\boldsymbol{U}_t\boldsymbol{U}_t^T - \boldsymbol{U}\boldsymbol{U}^T\right\|_F \leq \frac{\sqrt{2}\,\|\boldsymbol{X}_t - \boldsymbol{X}\|_F}{\sigma_{\min}(\boldsymbol{X})}, \quad \left\|\boldsymbol{V}_t\boldsymbol{V}_t^T - \boldsymbol{V}\boldsymbol{V}^T\right\|_F \leq \frac{\sqrt{2}\,\|\boldsymbol{X}_t - \boldsymbol{X}\|_F}{\sigma_{\min}(\boldsymbol{X})}.$$

2. *Let $\mathcal{P}_{\mathbb{T}_t}$ and $\mathcal{P}_{\mathbb{T}}$ be the projection operators onto the tangent spaces of the rank-r matrix manifold at $\boldsymbol{X}_t$ and $\boldsymbol{X}$, respectively. Then, the following bounds hold:*

$$\sup_{\|\boldsymbol{Z}\|_2=1} \|(\mathcal{P}_{\mathbb{T}_t} - \mathcal{P}_{\mathbb{T}})\,\boldsymbol{Z}\|_2 \leq \frac{2\,\|\boldsymbol{X}_t - \boldsymbol{X}\|_2}{\sigma_{\min}(\boldsymbol{X})} \quad \textit{and} \quad \sup_{\|\boldsymbol{Z}\|_2=1} \|(\mathcal{P}_{\mathbb{T}_t} - \mathcal{P}_{\mathbb{T}})\,\boldsymbol{Z}\|_F \leq \frac{2\sqrt{2}\,\|\boldsymbol{X}_t - \boldsymbol{X}\|_F}{\sigma_{\min}(\boldsymbol{X})}.$$

Proof. We prove only the second assertion, as the first assertion is identical to [Wei et al., 2016, Lemma 4.2].

By the definition of $\mathcal{P}_{\mathbb{T}_t}$ and $\mathcal{P}_{\mathbb{T}}$, we have

$$\begin{aligned}(\mathcal{P}_{\mathbb{T}_t} - \mathcal{P}_{\mathbb{T}})\,\boldsymbol{Z} &= \left(\boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{V}_t\boldsymbol{V}_t^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\boldsymbol{Z}\boldsymbol{V}_t\boldsymbol{V}_t^T\right) - \left(\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{Z}\boldsymbol{V}\boldsymbol{V}^T\right)\\ &= \left(\boldsymbol{U}_t\boldsymbol{U}_t^T - \boldsymbol{U}\boldsymbol{U}^T\right)\boldsymbol{Z}\left(\boldsymbol{I} - \boldsymbol{V}_t\boldsymbol{V}_t^T\right) + \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^T\right)\boldsymbol{Z}\left(\boldsymbol{V}_t\boldsymbol{V}_t^T - \boldsymbol{V}\boldsymbol{V}^T\right).\end{aligned}$$

Taking the spectral norm on both sides yields:

$$\sup_{\|\boldsymbol{Z}\|_2=1} \|(\mathcal{P}_{\mathbb{T}_t} - \mathcal{P}_{\mathbb{T}})\,\boldsymbol{Z}\|_2 \leq \|\boldsymbol{U}_t\boldsymbol{U}_t^T - \boldsymbol{U}\boldsymbol{U}^T\|_2 + \|\boldsymbol{V}_t\boldsymbol{V}_t^T - \boldsymbol{V}\boldsymbol{V}^T\|_2 \leq \frac{2\,\|\boldsymbol{X}_t - \boldsymbol{X}\|_2}{\sigma_{\min}(\boldsymbol{X})}.$$

Similarly, taking the Frobenius norm on both sides gives:

$$\sup_{\|\boldsymbol{Z}\|_2=1} \|(\mathcal{P}_{\mathbb{T}_t} - \mathcal{P}_{\mathbb{T}})\,\boldsymbol{Z}\|_F \leq \|\boldsymbol{U}_t\boldsymbol{U}_t^T - \boldsymbol{U}\boldsymbol{U}^T\|_F + \|\boldsymbol{V}_t\boldsymbol{V}_t^T - \boldsymbol{V}\boldsymbol{V}^T\|_F \leq \frac{2\sqrt{2}\,\|\boldsymbol{X}_t - \boldsymbol{X}\|_F}{\sigma_{\min}(\boldsymbol{X})}.$$

□

The following lemma provides second-order information about $\mathbb{M}_r$, the smooth manifold of all rank-r matrices.

Lemma 12 ([Wei et al., 2016], Lemma 4.1)**.** *Let $\boldsymbol{X}_t \in \mathbb{M}_r$ with compact SVD $\boldsymbol{X}_t = \boldsymbol{U}_t\boldsymbol{\Sigma}_t\boldsymbol{V}_t^T$, and let $\mathbb{T}_t$ denote the tangent space of $\mathbb{M}_r$ at $\boldsymbol{X}_t$. Let $\boldsymbol{X} \in \mathbb{M}_r$ be another rank-r matrix. Then, the following inequalities hold:*

$$\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_t})\,\boldsymbol{X}\|_F \leq \frac{1}{\sigma_{\min}(\boldsymbol{X})}\,\|\boldsymbol{X}_t - \boldsymbol{X}\|_2\,\|\boldsymbol{X}_t - \boldsymbol{X}\|_F \leq \frac{1}{\sigma_{\min}(\boldsymbol{X})}\,\|\boldsymbol{X}_t - \boldsymbol{X}\|_F^2\,,$$
$$\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_t})\,\boldsymbol{X}\|_2 \leq \frac{1}{\sigma_{\min}(\boldsymbol{X})}\,\|\boldsymbol{X}_t - \boldsymbol{X}\|_2^2\,.$$

Proof. By the definition of the projection operators $\mathcal{P}_{\mathbb{T}_t}$ and $\mathcal{P}_{\mathbb{T}}$, we have:

$$\begin{aligned}(\mathcal{I} - \mathcal{P}_{\mathbb{T}_t})\,\boldsymbol{X} &= (\mathcal{P}_{\mathbb{T}} - \mathcal{P}_{\mathbb{T}_t})\,\boldsymbol{X}\\ &= \left(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\right)\boldsymbol{X}\left(\boldsymbol{I} - \boldsymbol{V}_t\boldsymbol{V}_t^T\right) + \left(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^T\right)\boldsymbol{X}\left(\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{V}_t\boldsymbol{V}_t^T\right)\\ &= \left(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\right)\boldsymbol{X}\left(\boldsymbol{I} - \boldsymbol{V}_t\boldsymbol{V}_t^T\right)\\ &= \left(\boldsymbol{U}\boldsymbol{U}^T - \boldsymbol{U}_t\boldsymbol{U}_t^T\right)(\boldsymbol{X} - \boldsymbol{X}_t)\left(\boldsymbol{I} - \boldsymbol{V}_t\boldsymbol{V}_t^T\right).\end{aligned}$$

Taking the spectral and Frobenius norms on both sides and applying Lemma 11 completes the proof. □

B PROOF IN RESTRICTED ISOMETRY PROPERTY

For completeness, we provide the proof and relevant references regarding the properties of the Restricted Isometry Property (RIP) in this section.

Proof of Lemma 2. Assertions 1, 2, and 4 follow directly from a non-symmetric version of [Stöger and Zhu, 2025, Lemma 2.4] and [Wei et al., 2016, Lemma 4.4].

We now prove Assertion 3. Consider the following chain of inequalities:

$$\begin{aligned}\sup_{\|\boldsymbol{Z}\|_F=1} \|(\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}} - \mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}})(\boldsymbol{Z})\|_F &\overset{\text{(a)}}{=} \sup_{\|\boldsymbol{Z}\|_F=1} |\langle(\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}} - \mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}})(\boldsymbol{Z}), \boldsymbol{Z}\rangle| \\ &= \sup_{\|\boldsymbol{Z}\|_F=1} \left| \|\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}(\boldsymbol{Z})\|_F^2 - \|\mathcal{A}\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}(\boldsymbol{Z})\|_2^2 \right| \\ &\overset{\text{(b)}}{\leq} \sup_{\|\boldsymbol{Z}\|_F=1} \delta_{2r} \|\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}(\boldsymbol{Z})\|_F^2 \leq \delta_{2r},\end{aligned}$$

where:

- Step (a) follows because $\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}} - \mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}\mathcal{A}^*\mathcal{A}\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}$ is a self-adjoint operator, and the operator norm is expressed in its variational form.
- Step (b) holds because $\mathcal{P}_{\mathbb{T}_{\boldsymbol{X}}}(\boldsymbol{Z})$ has rank at most $2r$, and RIP applies.

This completes the proof of Assertion 3. □

C PROOFS IN DECOUPLING TECHNIQUE

The following lemma describes the properties of the operator $\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}$ and its relationship with $\mathcal{A}$. It follows directly from the definition of $\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}$.

Lemma 13. *For any matrix $\boldsymbol{Z} \in \mathbb{R}^{d_1 \times d_2}$, the following properties hold:*

$$\begin{aligned}&\left(\mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\right)\left(\mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right) = \mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z}), \\ &\left(\mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\right)\left(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right) = (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right) - \left\langle \mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right), \mathcal{A}\left(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right)\right\rangle \boldsymbol{x}\boldsymbol{y}^T, \\ &\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\right)(\boldsymbol{Z}) = (\mathcal{A}^*\mathcal{A} - I)\mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z}) + \left\langle \boldsymbol{x}\boldsymbol{y}^T, \mathcal{A}^*\mathcal{A}\left(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right)\right\rangle \boldsymbol{x}\boldsymbol{y}^T.\end{aligned} \tag{30}$$

Proof. We prove each assertion separately.

First assertion: By the definition of $\boldsymbol{A}_{i,(\boldsymbol{x},\boldsymbol{y})}$, we have $\langle \boldsymbol{A}_{i,(\boldsymbol{x},\boldsymbol{y})}, \mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\rangle = 0$. Consequently,

$$\left(\mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\right)\left(\mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right) = \frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{A}_{i,(\boldsymbol{x},\boldsymbol{y})}, \mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\rangle \boldsymbol{A}_{i,(\boldsymbol{x},\boldsymbol{y})} + \langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{Z}\rangle \boldsymbol{x}\boldsymbol{y}^T = \langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{Z}\rangle \boldsymbol{x}\boldsymbol{y}^T.$$

This establishes the first assertion.

Second assertion: For the orthogonal projection $\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})$, we observe that

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

This proves the second assertion.

Third assertion: For the difference $\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}$, we decompose $\boldsymbol{Z}$ as $\boldsymbol{Z} = \mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}\boldsymbol{Z} + \mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}\boldsymbol{Z}$. Then,

$$
\begin{aligned}
&(\mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})} - \mathcal{I})\boldsymbol{Z} = (\mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})} - \mathcal{I})(\mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}\boldsymbol{Z} + \mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}\boldsymbol{Z}) \\
&\overset{\text{(a)}}{=} (\mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})} - \mathcal{I})\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}\boldsymbol{Z} \\
&\overset{\text{(b)}}{=} (\mathcal{A}^*\mathcal{A} - \mathcal{I})(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}\boldsymbol{Z}) - \left\langle \mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right), \mathcal{A}\left(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right)\right\rangle \boldsymbol{x}\boldsymbol{y}^T \\
&= (\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{Z} - \mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}\boldsymbol{Z}) - \left\langle \mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right), \mathcal{A}\left(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right)\right\rangle \boldsymbol{x}\boldsymbol{y}^T,
\end{aligned}
$$

where (a) follows from the first assertion and (b) follows from the second assertion. Rearranging terms completes the proof of the third assertion. □

Now we can prove Lemma 5, which bounds $\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{X}_\star - \boldsymbol{X}_t\right)\|_2$ using the virtual sequence $\{\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\}_{t\in\mathbb{N}}$.

Proof of Lemma 5. Let $\boldsymbol{\Delta}_t := \boldsymbol{X}_\star - \boldsymbol{X}_t$ and $\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})} := \boldsymbol{X}_\star - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}$. From the construction of the net $\mathcal{N}$, we have

$$
\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{\Delta}_t\right)\|_2 = \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathbb{S}^{d_1-1}\times\mathbb{S}^{d_2-1}} \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{\Delta}_t\right)\right\rangle\right| \le 2 \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}} \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{\Delta}_t\right)\right\rangle\right|.
$$

For every $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}$, applying the triangle inequality yields

$$
\begin{aligned}
&\left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{\Delta}_t\right)\right\rangle\right| \\
&\le \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right\rangle\right| + \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{\Delta}_t\right)\right\rangle\right| \\
&\le \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right\rangle\right| + \left\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{\Delta}_t\right)\right\|_2 \\
&\overset{\text{(a)}}{\le} \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right\rangle\right| + \delta\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F \\
&\le \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right)\right\rangle\right| + \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right)\right\rangle\right| + \delta\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F,
\end{aligned}
$$

where (a) follows from (7) in Lemma 2, which is a consequence of RIP of $\mathcal{A}$. We now estimate the first two terms in the last line.

- **Second term:** The second term can be bounded as

$$
\begin{aligned}
&\left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\mathcal{P}_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right)\right\rangle\right| \\
&= \left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A} - \mathcal{I})\,\boldsymbol{x}\boldsymbol{y}^T\right\rangle \left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\rangle\right| = \left|\left(\left\|\mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right)\right\|_2^2 - 1\right)\left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\rangle\right| \\
&\overset{\text{(a)}}{\le} \delta\left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\rangle\right| \le \delta\|\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_2 \le \delta\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F + \delta\left\|\boldsymbol{\Delta}_t\right\|_2,
\end{aligned}
$$

  where (a) follows from the definition of the RIP property.

- **First term:** Under the assumption that (18) holds, the first term can be estimated as

$$
\begin{aligned}
&\left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, (\mathcal{A}^*\mathcal{A})\left(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right)\right\rangle\right| \\
&\le 4\sqrt{\frac{d_1+d_2}{m}}\left\|\mathcal{A}\left(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right)\right\|_2 \overset{\text{(a)}}{\le} 8\sqrt{\frac{d_1+d_2}{m}}\left\|\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right)\right\|_F \le 8\sqrt{\frac{d_1+d_2}{m}}\left\|\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F \\
&\le 8\sqrt{\frac{d_1+d_2}{m}}\left\|\boldsymbol{\Delta}_t\right\|_F + 8\sqrt{\frac{d_1+d_2}{m}}\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F \\
&\overset{\text{(b)}}{\le} 8\sqrt{\frac{2r(d_1+d_2)}{m}}\left\|\boldsymbol{\Delta}_t\right\|_2 + 8\sqrt{\frac{d_1+d_2}{m}}\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F,
\end{aligned}
$$

where (a) follows from the RIP property of $\mathcal{A}$, $\operatorname{rank}(\mathcal{P}^{\perp}_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})})) \le 2r+2$, and $1+\delta_{2r+2} \le 2$, and (b) follows from $\operatorname{rank}(\boldsymbol{\Delta}_t) \le 2r$.

Combining all the estimated terms and taking the supreme over $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}$, we obtain the final bound:

$$\left\|\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)(\boldsymbol{\Delta}_t)\right\|_2 \le \left(2\delta + 16\sqrt{\frac{2r(d_1+d_2)}{m}}\right)\|\boldsymbol{\Delta}_t\|_2 + \left(2\delta+2\delta+16\sqrt{\frac{d_1+d_2}{m}}\right)\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F .$$

This completes the proof. □

D PROOF OF INITIALIZATION

The proof of Lemma 6 follows a structure similar to that of [Stöger and Zhu, 2025, Lemma 4.1].

Proof of Lemma 6. To prove this lemma, we establish the following two inequalities:

$$\|\boldsymbol{X}_\star - \boldsymbol{X}_0\|_2 \le \frac{1}{2}c_1\sigma_{\min}(\boldsymbol{X}_\star), \tag{31}$$

and

$$\left\|\boldsymbol{X}_0 - \boldsymbol{X}_0^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F \le \frac{1}{2}c_1\sigma_{\min}(\boldsymbol{X}_\star), \quad \forall\,(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}. \tag{32}$$

First, from Lemma 1, with probability at least $1-\exp(-(d_1+d_2))$, the operator $\mathcal{A}$ satisfies RIP of rank $6r$ with $\delta_{6r}=\delta$ when $m \ge c\delta^{-2}r(d_1+d_2)$, where c is a universal constant. This implies that, with the same probability, $\mathcal{A}$ satisfies RIP of rank $6r$ with constant $\delta = \sqrt{\frac{cr(d_1+d_2)}{m}}$. We choose $m > cr(d_1+d_2)$ to ensure that $\delta < 1$.

Then, we have

$$\begin{aligned}\left\|\left(\mathcal{A}^*\mathcal{A}\right)(\boldsymbol{X}_\star) - \boldsymbol{X}_\star\right\|_2 &\le 2\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\frac{1}{m}\sum_{i=1}^m \boldsymbol{x}^T\left(\langle\boldsymbol{A}_i,\boldsymbol{X}_\star\rangle\boldsymbol{A}_i - \boldsymbol{X}_\star\right)\boldsymbol{y}\\ &= 2\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\frac{1}{m}\sum_{i=1}^m\left(\langle\boldsymbol{A}_i,\boldsymbol{X}_\star\rangle\boldsymbol{x}^T\boldsymbol{A}_i\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{X}_\star\boldsymbol{y}\right).\end{aligned}$$

The expectation can be computed as $\mathbb{E}\langle\boldsymbol{A}_i,\boldsymbol{X}_\star\rangle\boldsymbol{x}^T\boldsymbol{A}_i\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{X}_\star\boldsymbol{y}$. From [Vershynin, 2018], we have

$$\|\langle\boldsymbol{A}_i,\boldsymbol{X}_\star\rangle\boldsymbol{x}^T\boldsymbol{A}_i\boldsymbol{y}\|_{\psi_1} \le \|\langle\boldsymbol{A}_i,\boldsymbol{X}_\star\rangle\|_{\psi_2}\|\boldsymbol{x}^T\boldsymbol{A}_i\boldsymbol{y}\|_{\psi_2} \le K\|\boldsymbol{X}_\star\|_F,$$

where K is a universal constant, and therefore the centered version satisfies

$$\|\langle\boldsymbol{A}_i,\boldsymbol{X}_\star\rangle\boldsymbol{x}^T\boldsymbol{A}_i\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{X}_\star\boldsymbol{y}\|_{\psi_1} \le K\|\boldsymbol{X}_\star\|_F.$$

Applying Bernstein's inequality, we obtain:

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m\left(\langle\boldsymbol{A}_i,\boldsymbol{X}_\star\rangle\boldsymbol{x}^T\boldsymbol{A}_i\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{X}_\star\boldsymbol{y}\right)\right| \ge t\right) \le 2\exp\left(-C'\min\left\{\frac{mt^2}{\|\boldsymbol{X}_\star\|_F^2},\frac{mt}{\|\boldsymbol{X}_\star\|_F}\right\}\right).$$

Setting $t = C''(\sqrt{\frac{d_1+d_2}{m}} + \frac{d_1+d_2}{m})\|\boldsymbol{X}_\star\|_F$, the probability is less than $2\exp(-C''C'(d_1+d_2))$ for a fixed pair $(\boldsymbol{x},\boldsymbol{y})$. Taking a union bound over all $(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}$, we obtain

$$\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left|\frac{1}{m}\sum_{i=1}^m\left(\langle\boldsymbol{A}_i,\boldsymbol{X}_\star\rangle\boldsymbol{x}^T\boldsymbol{A}_i\boldsymbol{y} - \boldsymbol{x}^T\boldsymbol{X}_\star\boldsymbol{y}\right)\right| \le C''(\sqrt{\frac{d_1+d_2}{m}} + \frac{d_1+d_2}{m})\|\boldsymbol{X}_\star\|_F$$

with probability at least $1-2\exp((\ln 12 - C''C')(d_1+d_2))$. We choose C'' sufficiently large such that $\ln 12 - C''C' < -4$, ensuring a high success probability. Consequently, with probability at least $1-2\exp\left(-4(d_1+d_2)\right)$,

$$\left\|\left(\mathcal{A}^*\mathcal{A}\right)(\boldsymbol{X}_\star) - \boldsymbol{X}_\star\right\|_2 \le 2C''(\sqrt{\frac{d_1+d_2}{m}} + \frac{d_1+d_2}{m})\|\boldsymbol{X}_\star\|_F \le 2C''(\sqrt{\frac{d_1+d_2}{m}} + \frac{d_1+d_2}{m})\sqrt{r}\kappa\sigma_{\min}(\boldsymbol{X}_\star).$$

We choose a proper constant C_1 and let $m \geq C_1\kappa^2 r(d_1+d_2)$ to make the constant before $\sigma_{\min}(\boldsymbol{X}_\star)$ less than or equal to $\min\{\frac{1}{4}c_1, \frac{1}{10}\}$, and then we obtain

$$\|(\mathcal{A}^*\mathcal{A})(\boldsymbol{X}_\star) - \boldsymbol{X}_\star\|_2 \leq \min\{\frac{1}{4}c_1, \frac{1}{10}\}\sigma_{\min}(\boldsymbol{X}_\star). \tag{33}$$

This, together with Weyl's inequality, implies that the spectral gap for $(\mathcal{A}^*\mathcal{A})(\boldsymbol{X}_\star)$ satisfies:

$$s_1 := \sigma_r\left((\mathcal{A}^*\mathcal{A})(\boldsymbol{X}_\star)\right) - \sigma_{r+1}\left((\mathcal{A}^*\mathcal{A})(\boldsymbol{X}_\star)\right) \geq \frac{4}{5}\sigma_{\min}(\boldsymbol{X}_\star) > 0. \tag{34}$$

As a result, $\boldsymbol{X}_0 = \mathcal{H}_r(\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_\star))$ is uniquely defined. Using the best rank-r approximation property of $\boldsymbol{X}_0$, we obtain

$$\begin{aligned}\|\boldsymbol{X}_\star - \boldsymbol{X}_0\|_2 &\leq \|\boldsymbol{X}_\star - (\mathcal{A}^*\mathcal{A})(\boldsymbol{X}_\star)\|_2 + \|(\mathcal{A}^*\mathcal{A})(\boldsymbol{X}_\star) - \boldsymbol{X}_0\|_2 \\ &\leq 2\,\|\boldsymbol{X}_\star - (\mathcal{A}^*\mathcal{A})(\boldsymbol{X}_\star)\|_2\,.\end{aligned}$$

Thus, combining it with (33), we obtain (31).

From Lemma 13, we have

$$\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\right)(\boldsymbol{X}_\star) = \langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{X}_\star\rangle\,(\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{x}\boldsymbol{y}^T\right) + \left\langle \boldsymbol{x}\boldsymbol{y}^T, \mathcal{A}^*\mathcal{A}\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{Z})\right)\right\rangle \boldsymbol{x}\boldsymbol{y}^T. \tag{35}$$

Therefore,

$$\begin{aligned}\left\|\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\right)(\boldsymbol{X}_\star)\right\|_2 &\leq \|\boldsymbol{X}_\star\|_2\|\,(\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{x}\boldsymbol{y}^T\right)\|_2 + \left|\left\langle \mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right), \mathcal{A}\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{X}_\star\right)\right)\right\rangle\right| \\ &:= I_1 + I_2.\end{aligned} \tag{36}$$

From (7) in Lemma 2, it follows that

$$I_1 \leq \|\boldsymbol{X}_\star\|_2 \cdot \delta \leq \kappa\sigma_{\min}(\boldsymbol{X}_\star)\sqrt{\frac{cr(d_1+d_2)}{m}}.$$

To estimate I_2, we use

$$\left\langle \mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right), \mathcal{A}\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{X}_\star\right)\right)\right\rangle = \frac{1}{m}\sum_{i=1}^m \langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{A}_i\rangle \left\langle \boldsymbol{A}_i, \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{X}_\star\right)\right\rangle.$$

Here, $\sum_{i=1}^m \langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{A}_i\rangle \left\langle \boldsymbol{A}_i, \mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{X}_\star\right)\right\rangle$ is a sum of m independent sub-exponential random variables with mean zero due to the rotation invariance of the Gaussian measure. Each term has a sub-exponential norm $K\left\|\boldsymbol{X}_\star\right\|_F$ with constant K. Applying Bernstein's inequality, we obtain that for each fixed $(\boldsymbol{x},\boldsymbol{y})$, with probability at least $1 - \exp(-4(d_1+d_2))$,

$$I_2 = \left|\left\langle \mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right), \mathcal{A}\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}\left(\boldsymbol{X}_\star\right)\right)\right\rangle\right| \leq c_2\kappa\sigma_{\min}\left(\boldsymbol{X}_\star\right)\sqrt{r}\left(\sqrt{\frac{d_1+d_2}{m}} + \frac{d_1+d_2}{m}\right),$$

where c_2 is a constant depending only on K. Taking a union bound over all $(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}$ and combining I_1 and I_2, we obtain that, with probability at least $1 - \exp(-(d_1+d_2))$,

$$\left\|\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\right)(\boldsymbol{X}_\star)\right\|_2 \leq c_3\kappa\sigma_{\min}\left(\boldsymbol{X}_\star\right)\sqrt{r}\left(\sqrt{\frac{d_1+d_2}{m}} + \frac{d_1+d_2}{m}\right), \quad \forall\,(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{N}. \tag{37}$$

By choosing a proper C_2 and letting $m \geq C_2\kappa^2 r(d_1+d_2)$, (37) implies

$$\left\|\left(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\right)(\boldsymbol{X}_\star)\right\|_2 \leq \frac{4(1-1/\sqrt{2})}{5}\sigma_{\min}(\boldsymbol{X}_\star) \leq (1-1/\sqrt{2})s_1, \tag{38}$$

where in the last inequality we have used (34). Furthermore, by using (34) and (33), we obtain

$$c_0 := \frac{\sigma_{r+1}(\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_\star))}{s_1} \leq \frac{\frac{1}{10}}{\frac{4}{5}} \leq 1.$$

Applying Lemma 10 to $\boldsymbol{Z}_1 := \mathcal{A}^*\mathcal{A}(\boldsymbol{X}_\star)$ and $\boldsymbol{Z}_2 := \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}(\boldsymbol{X}_\star)$ and noticing that $\boldsymbol{X}_0 = \mathcal{H}_r(\boldsymbol{Z}_1)$ and $\boldsymbol{X}_0^{(\boldsymbol{x},\boldsymbol{y})} = \mathcal{H}_r(\boldsymbol{Z}_2)$, we obtain

$$\begin{aligned}\|\boldsymbol{X}_0 - \boldsymbol{X}_0^{(\boldsymbol{x},\boldsymbol{y})}\|_2 &\leq 16 \left\| \left(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}\right)(\boldsymbol{X}_\star)\right\|_2 \\ &\leq c_3'\kappa\sigma_{\min}(\boldsymbol{X}_\star)\sqrt{r}\left(\sqrt{\frac{d_1+d_2}{m}} + \frac{d_1+d_2}{m}\right),\end{aligned} \tag{39}$$

where we have used (37) in the last inequality, and

$$\begin{aligned}&\|\boldsymbol{X}_0 - \boldsymbol{X}_0^{(\boldsymbol{x},\boldsymbol{y})}\|_F \\ &\leq 16\left(\|(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})})(\boldsymbol{X}_\star)\boldsymbol{V}_{1,r}\|_F + \|\boldsymbol{U}_{1,r}^T(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})})(\boldsymbol{X}_\star)\|_F\right) \\ &\overset{(a)}{\leq} 16\left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{X}_\star\right\rangle\right| \left\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{x}\boldsymbol{y}^T\right)\boldsymbol{V}_{1,r}\right\|_F + 16\left|\left\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{X}_\star\right\rangle\right| \left\|\boldsymbol{U}_{1,r}^T(\mathcal{A}^*\mathcal{A} - \mathcal{I})\left(\boldsymbol{x}\boldsymbol{y}^T\right)\right\|_F \\ &\quad + 16\left|\left\langle \mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right), \mathcal{A}\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{X}_\star)\right)\right\rangle\right| \|\boldsymbol{x}\boldsymbol{y}^T\boldsymbol{V}_{1,r}\|_F + 16\left|\left\langle \mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right), \mathcal{A}\left(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{X}_\star)\right)\right\rangle\right| \|\boldsymbol{U}_{1,r}^T\boldsymbol{x}\boldsymbol{y}^T\|_F \\ &\overset{(b)}{\leq} 64c\kappa\sigma_{\min}(\boldsymbol{X}_\star)\sqrt{r}\left(\sqrt{\frac{d_1+d_2}{m}} + \frac{d_1+d_2}{m}\right),\end{aligned} \tag{40}$$

where (a) follows from (35), and (b) follows from (6) and (8) in Lemma 2, $\|\boldsymbol{U}_{1,r}\|_2 \leq 1$, and $\|\boldsymbol{V}_{1,r}\|_2 \leq 1$. We choose a proper constant $C_2' > C_2$ and let $m \geq C_2'\kappa^2 r(d_1+d_2)$ to ensure that the last term in (40) is not greater than $\frac{1}{2}\sigma_{\min}(\boldsymbol{X}_\star)$ and thus (32).

Throughout the proof, we have imposed several lower bounds on m. We then take their maximum, i.e., $m \geq C\kappa^2 r(d_1+d_2)$ with $C = \max\{c, C_1, C_2'\}$, to complete the proof. □

E PROOFS IN CONVERGENCE ANALYSIS

This section presents the proof of Lemma 7, a key result in our analysis. Unlike the corresponding argument for factorized gradient descent in [Stöger and Zhu, 2025], our proof requires analyzing the projection of the gradient onto the tangent space of $\boldsymbol{X}_t$, which relies on Lemma 11 and Lemma 12. Additionally, the use of a hard-thresholding operator after the gradient step introduces errors that are bounded using Lemma 10.

Proof of Lemma 7. From the assumption of this lemma,

$$c_1 < \frac{1}{1000}, \tag{41}$$

and we have $\mathcal{A}$ satisfies RIP of rank $6r$ with

$$\delta = \delta_{6r} \leq \frac{1}{24}c_1 < 1. \tag{42}$$

Besides, (19) holds with this δ for $t \leq T \leq 12^{d_1+d_2}$.

We prove this theorem by induction. The assumption (21) of this lemma gives $E_0 \leq c_1\sigma_{\min}(\boldsymbol{X}_\star)$. Assume that

$$E_0 \leq c_1\sigma_{\min}(\boldsymbol{X}_\star), \quad E_1 \leq (1000c_1)c_1\sigma_{\min}(\boldsymbol{X}_\star), \quad \cdots, \quad E_t \leq (1000c_1)^t c_1\sigma_{\min}(\boldsymbol{X}_\star).$$

We will need to show that $E_{t+1} \leq (1000c_1)^{t+1}c_1\sigma_{\min}(\boldsymbol{X}_\star)$, i.e.,

$$\|\boldsymbol{X}_{t+1} - \boldsymbol{X}_\star\|_2 + \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}} \|\boldsymbol{X}_{t+1} - \boldsymbol{X}_{t+1}^{(\boldsymbol{x},\boldsymbol{y})}\|_F \leq c_1(1000c_1)^{t+1}\sigma_{\min}(\boldsymbol{X}_\star).$$

For this purpose, we estimate $\|\boldsymbol{X}_{t+1} - \boldsymbol{X}_\star\|_2$ and $\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}} \|\boldsymbol{X}_{t+1} - \boldsymbol{X}_{t+1}^{(\boldsymbol{x},\boldsymbol{y})}\|_F$, respectively. Notice that the inductive assumption $E_t \leq (1000c_1)^t c_1\sigma_{\min}(\boldsymbol{X}_\star)$ implies

$$\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_2 \leq c_1\sigma_{\min}(\boldsymbol{X}_\star) \qquad \text{and} \qquad \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}} \|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F \leq c_1\sigma_{\min}(\boldsymbol{X}_\star). \tag{43}$$

Estimate $\|\boldsymbol{X}_{t+1}-\boldsymbol{X}_*\|_2$. We first compute $\|\boldsymbol{W}_t-\boldsymbol{X}_\star\|_2$. By decomposing $\boldsymbol{X}_\star-\boldsymbol{X}_t$ onto $\mathbb{T}_t$ and $\mathbb{T}_t^\perp$, we obtain

$$\begin{aligned}
&\|\boldsymbol{W}_t-\boldsymbol{X}_\star\|_2\\
&=\|(\mathcal{I}-\mathcal{P}_{\mathbb{T}_t}\mathcal{A}^*\mathcal{A})\left(\boldsymbol{X}_\star-\boldsymbol{X}_t\right)\|_2\\
&\le\|(\mathcal{I}-\mathcal{P}_{\mathbb{T}_t})\left(\boldsymbol{X}_\star-\boldsymbol{X}_t\right)\|_2+\|\mathcal{P}_{\mathbb{T}_t}\left(I-\mathcal{A}^*\mathcal{A}\right)\left(\boldsymbol{X}_\star-\boldsymbol{X}_t\right)\|_2\\
&\overset{(a)}{\le}\frac{1}{\sigma_{\min}(\boldsymbol{X}_\star)}\|\boldsymbol{X}_t-\boldsymbol{X}_\star\|_2^2+\|\mathcal{P}_{\mathbb{T}_t}\left(I-\mathcal{A}^*\mathcal{A}\right)\left(\boldsymbol{X}_\star-\boldsymbol{X}_t\right)\|_2\\
&\overset{(b)}{\le}\left(c_1+3\left(16\sqrt{\frac{2r(d_1+d_2)}{m}}+2\delta\right)\right)\|\boldsymbol{X}_t-\boldsymbol{X}_\star\|_2+12\left(\delta+4\sqrt{\frac{d_1+d_2}{m}}\right)\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{X}_t-\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F\\
&\le\left(\frac{3}{2}c_1+48\sqrt{\frac{2r(d_1+d_2)}{m}}\right)\|\boldsymbol{X}_t-\boldsymbol{X}_\star\|_2+\left(\frac{1}{2}c_1+48\sqrt{\frac{d_1+d_2}{m}}\right)\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{X}_t-\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F,
\end{aligned}$$

where (a) follows from Lemma 12, and (b) from (19) in Lemma 5, the first equation in (43), and $\sup_{\|\boldsymbol{Z}\|_2=1}\|\mathcal{P}_{\mathbb{T}_t}\boldsymbol{Z}\|_2\le 3$. We choose a proper constant C' and let $m\ge C'\kappa^2r(d_1+d_2)$ to make the coefficients before $\|\boldsymbol{X}_t-\boldsymbol{X}_\star\|_2$ and $\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{X}_t-\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F$ above are both smaller than $2c_1$. Then we have:

$$\|\boldsymbol{X}_\star-\boldsymbol{W}_t\|_2\le 2c_1\left(\|\boldsymbol{X}_t-\boldsymbol{X}_\star\|_2+\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{X}_t-\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F\right)\le c_1\sigma_{\min}(\boldsymbol{X}_\star),\tag{44}$$

where in the last inequality we used the fact that $2c_1<1$ and the inductive assumption. This, together with Weyl's inequality, implies that $\sigma_r(\boldsymbol{W}_t)\ge(1-c_1)\sigma_{\min}(\boldsymbol{X}_\star)>c_1\sigma_{\min}(\boldsymbol{X}_\star)\ge\sigma_{r+1}(\boldsymbol{W}_t)$ and

$$s:=\sigma_r(\boldsymbol{W}_t)-\sigma_{r+1}(\boldsymbol{W}_t)\ge(1-2c_1)\sigma_{\min}(\boldsymbol{X}_\star)>0,\tag{45}$$

i.e., the spectral gap of $\boldsymbol{W}_t$ is positive. Then, $\boldsymbol{X}_{t+1}=\mathcal{H}_r(\boldsymbol{W}_t)$ is uniquely defined, which is the best rank-r approximation to $\boldsymbol{W}_t$. Therefore,

$$\begin{aligned}
\|\boldsymbol{X}_{t+1}-\boldsymbol{X}_\star\|_2&\le\|\boldsymbol{X}_{t+1}-\boldsymbol{W}_t\|_2+\|\boldsymbol{W}_t-\boldsymbol{X}_\star\|_2\le 2\|\boldsymbol{W}_t-\boldsymbol{X}_\star\|_2\\
&\le 4c_1\left(\|\boldsymbol{X}_t-\boldsymbol{X}_\star\|_2+\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{X}_t-\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F\right),
\end{aligned}\tag{46}$$

where in the last inequality we have used (44).

Estimate $\|\boldsymbol{X}_{t+1}-\boldsymbol{X}_{t+1}^{(\boldsymbol{x},\boldsymbol{y})}\|_F$. Since $\boldsymbol{X}_{t+1}=\mathcal{H}_r(\boldsymbol{W}_t)$ and $\boldsymbol{X}_{t+1}^{(\boldsymbol{x},\boldsymbol{y})}=\mathcal{H}_r(\boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})})$, applying Lemma 10, we can upper bound $\|\boldsymbol{X}_{t+1}-\boldsymbol{X}_{t+1}^{(\boldsymbol{x},\boldsymbol{y})}\|_F$ by $\|\boldsymbol{W}_t-\boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F$. We first bound $\|\boldsymbol{W}_t-\boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F$ by

$$\begin{aligned}
&\|\boldsymbol{W}_t-\boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F\\
&=\|(\boldsymbol{X}_t-\mathcal{P}_{\mathbb{T}_t}\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t-\boldsymbol{X}_\star))-(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}-\mathcal{P}_{\mathbb{T}_t^{(\boldsymbol{x},\boldsymbol{y})}}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}^*\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}-\boldsymbol{X}_\star))\|_F\\
&\le\|(\mathcal{I}-\mathcal{P}_{\mathbb{T}_t})\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F+\|\mathcal{P}_{\mathbb{T}_t}\left(\boldsymbol{X}_t-\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t-\boldsymbol{X}_\star)-\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}+\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}-\boldsymbol{X}_\star)\right)\|_F\\
&\qquad+\|\mathcal{P}_{\mathbb{T}_t}(\mathcal{A}^*\mathcal{A}-\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}^*\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})})(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}-\boldsymbol{X}_\star)\|_F+\|(\mathcal{P}_{\mathbb{T}_t}-\mathcal{P}_{\mathbb{T}_t^{(\boldsymbol{x},\boldsymbol{y})}})\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}^*\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})}(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}-\boldsymbol{X}_\star)\|_F\\
&:=I_1+I_2+I_3+I_4.
\end{aligned}\tag{47}$$

We estimate the four terms respectively.

- Bounding I_1. I_1 is a second-order term about $\|\boldsymbol{X}_t-\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F$. Indeed, Lemma 12 implies

$$I_1\le\frac{1}{\sigma_{\min}(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})})}\|\boldsymbol{X}_t-\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F^2.\tag{48}$$

We need to derive a lower bound for $\sigma_{\min}(\boldsymbol{X}_t)$ and $\sigma_{\min}(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})})$ respectively. From Weyl's inequality and the inductive assumption (43), we have

$$\sigma_{\min}(\boldsymbol{X}_t)\ge\sigma_{\min}(\boldsymbol{X}_\star)-\|\boldsymbol{X}_t-\boldsymbol{X}_\star\|_2\ge\sigma_{\min}(\boldsymbol{X}_\star)-c_1\sigma_{\min}(\boldsymbol{X}_\star)\ge(1-c_1)\sigma_{\min}(\boldsymbol{X}_\star)$$

and

$$\sigma_{\min}(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}) \geq \sigma_{\min}(\boldsymbol{X}_t) - \left\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F \geq (1-2c_1)\sigma_{\min}(\boldsymbol{X}_\star). \tag{49}$$

Plugging it in (48) gives

$$I_1 \leq \frac{1}{(1-2c_1)\sigma_{\min}(\boldsymbol{X}_\star)}\left\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F^2 \leq \frac{c_1}{1-2c_1}\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F \leq 2c_1\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F, \tag{50}$$

where we have used the inductive assumption (43) in the second inequality and (41) in the last inequality.

- Bounding I_2. We estimate I_2 by projecting $\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}$ onto $\mathbb{T}_t$ and $\mathbb{T}_t^\perp$ respectively as follows:

$$\begin{aligned}
I_2 &= \|\mathcal{P}_{\mathbb{T}_t}(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})})\|_F \\
&\leq \|\mathcal{P}_{\mathbb{T}_t}(\mathcal{I} - \mathcal{A}^*\mathcal{A})\mathcal{P}_{\mathbb{T}_t}(\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})})\|_F + \|\mathcal{P}_{\mathbb{T}_t}(\mathcal{I} - \mathcal{A}^*\mathcal{A})(\mathcal{I} - \mathcal{P}_{\mathbb{T}_t})(\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})})\|_F \\
&\overset{\text{(a)}}{\leq} \delta_{2r}\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F + \delta_{3r}\|(\mathcal{I} - \mathcal{P}_{\mathbb{T}_t})(\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})})\|_F \\
&\overset{\text{(b)}}{\leq} \delta_{2r}\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F + \delta_{3r}\frac{1}{\sigma_{\min}(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})})}\left\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F^2 \\
&\overset{\text{(c)}}{\leq} \left(\delta_{2r} + \frac{c_1\delta_{3r}}{1-2c_1}\right)\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F \\
&\overset{\text{(d)}}{\leq} 2c_1\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F,
\end{aligned} \tag{51}$$

where:

  - step (a) follows from the properties 3 and 4 of RIP in Lemma 2,
  - step (b) follows from Lemma 11,
  - step (c) follows from the inductive assumption (43) and (49),
  - step (d) follows from $c_1 < \frac{1}{1000}$ and $\delta < \frac{1}{24}c_1$.

- Bounding I_3. For I_3, we denote $\boldsymbol{\Delta}_t := \boldsymbol{X}_t - \boldsymbol{X}_\star$ and $\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})} := \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{X}_\star$. Then, I_3 is estimated as follows:

$$\begin{aligned}
I_3 &= \|\mathcal{P}_{\mathbb{T}_t}(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})})(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{X}_\star)\|_F \\
&\overset{\text{(a)}}{\leq} \|\mathcal{P}_{\mathbb{T}_t}(\mathcal{A}^*\mathcal{A} - \mathcal{I})\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\rangle \boldsymbol{x}\boldsymbol{y}^T\|_F + \|\langle \mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right), \mathcal{A}(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}))\rangle \mathcal{P}_{\mathbb{T}_t}(\boldsymbol{x}\boldsymbol{y}^T)\|_F \\
&\leq \|\mathcal{P}_{\mathbb{T}_t}(\mathcal{A}^*\mathcal{A} - \mathcal{I})\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\rangle \boldsymbol{x}\boldsymbol{y}^T\|_F + \left|\langle \mathcal{A}\left(\boldsymbol{x}\boldsymbol{y}^T\right), \mathcal{A}(\mathcal{P}^\perp_{\boldsymbol{x}\boldsymbol{y}^T}(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}))\rangle\right| \\
&\overset{\text{(b)}}{\leq} \|\mathcal{P}_{\mathbb{T}_t}(\mathcal{A}^*\mathcal{A} - \mathcal{I})\langle \boldsymbol{x}\boldsymbol{y}^T, \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\rangle \boldsymbol{x}\boldsymbol{y}^T\|_F \\
&\quad + \left(8\sqrt{\frac{2r(d_1+d_2)}{m}}\|\boldsymbol{\Delta}_t\|_2 + 8\sqrt{\frac{d_1+d_2}{m}}\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F\right) \\
&\leq \|\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_2\left(\|\mathcal{P}_{\mathbb{T}_t}(\mathcal{A}^*\mathcal{A} - \mathcal{I})\mathcal{P}_{\mathbb{T}_t}\boldsymbol{x}\boldsymbol{y}^T\|_F + \|\mathcal{P}_{\mathbb{T}_t}(\mathcal{A}^*\mathcal{A})(\mathcal{I} - \mathcal{P}_{\mathbb{T}_t})\boldsymbol{x}\boldsymbol{y}^T\|_F\right) \\
&\quad + \left(8\sqrt{\frac{2r(d_1+d_2)}{m}}\|\boldsymbol{\Delta}_t\|_2 + 8\sqrt{\frac{d_1+d_2}{m}}\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F\right) \\
&\overset{\text{(c)}}{\leq} \left(8\sqrt{\frac{2r(d_1+d_2)}{m}} + \delta_{2r} + \delta_{3r}\right)\|\boldsymbol{\Delta}_t\|_2 \\
&\quad + \left(8\sqrt{\frac{d_1+d_2}{m}} + \delta_{2r} + \delta_{3r}\right)\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F \\
&\overset{\text{(d)}}{\leq} \left(8\sqrt{\frac{2r(d_1+d_2)}{m}} + \frac{1}{12}c_1\right)\|\boldsymbol{\Delta}_t\|_2 \\
&\quad + \left(8\sqrt{\frac{d_1+d_2}{m}} + \frac{1}{12}c_1\right)\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F,
\end{aligned} \tag{52}$$

where:

- step (a) follows form (30) in Lemma 13,
- step (b) follows from (18) in Lemma 4,
- step (c) from $\|\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_2 \leq \|\boldsymbol{\Delta}_t\|_2 + \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F$ and Lemma 1,
- step (d) from $\delta_{2r} \leq \delta_{3r} \leq \delta \leq \frac{1}{24}c_1$ by assumption (42).

We further denote the upper bound for I_3 in the last inequality as I_3', that is

$$I_3 \leq I_3' := \left(8\sqrt{\frac{2r(d_1+d_2)}{m}} + \frac{1}{12}c_1\right)\|\boldsymbol{\Delta}_t\|_2 + \left(8\sqrt{\frac{d_1+d_2}{m}} + \frac{1}{12}c_1\right)\sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F.$$

- Bounding I_4. We estimate I_4 as in the following:

$$\begin{aligned}
I_4 &\leq \|(\mathcal{P}_{\mathbb{T}_t} - \mathcal{P}_{\mathbb{T}_t^{(\boldsymbol{x},\boldsymbol{y})}})\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{X}_\star)\|_F + \|\mathcal{P}_{\mathbb{T}_t}(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})})(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{X}_\star)\|_F \\
&\quad + \|\mathcal{P}_{\mathbb{T}_t^{(\boldsymbol{x},\boldsymbol{y})}}(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})})(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{X}_\star)\|_F \\
&\overset{(a)}{\leq} \|(\mathcal{P}_{\mathbb{T}_t} - \mathcal{P}_{\mathbb{T}_t^{(\boldsymbol{x},\boldsymbol{y})}})\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{X}_\star)\|_F + 2I_3' \\
&\leq \|(\mathcal{P}_{\mathbb{T}_t} - \mathcal{P}_{\mathbb{T}_t^{(\boldsymbol{x},\boldsymbol{y})}})(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{X}_\star)\|_F + \|(\mathcal{P}_{\mathbb{T}_t} - \mathcal{P}_{\mathbb{T}_t^{(\boldsymbol{x},\boldsymbol{y})}})(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{X}_\star)\|_F + 2I_3' \\
&\overset{(b)}{\leq} \frac{4\sqrt{2}}{\sigma_{\min}(\boldsymbol{X}_t)}\left\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F\left(\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})})\|_2 + \|\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_2\right) + 2I_3' \\
&\overset{(c)}{\leq} \frac{4\sqrt{2}c_1}{1-2c_1}\left(\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})})\|_2 + \|\boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_2\right) + 2I_3' \\
&\leq \frac{4\sqrt{2}c_1}{1-2c_1}\left(\|\boldsymbol{\Delta}_t\|_2 + \|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F + \|(\mathcal{A}^*\mathcal{A} - \mathcal{I})\boldsymbol{\Delta}_t\|_2 + \|(\mathcal{A}^*\mathcal{A} - I)(\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})})\|_2\right) + 2I_3' \qquad (53)\\
&\overset{(d)}{\leq} \frac{4\sqrt{2}c_1}{1-2c_1}\left(1 + 16\sqrt{\frac{2r(d_1+d_2)}{m}} + 2\delta\right)\|\boldsymbol{\Delta}_t\|_2 \\
&\quad + \frac{4\sqrt{2}c_1}{1-2c_1}\left(1 + 4\delta + 16\sqrt{\frac{d_1+d_2}{m}} + \delta_{r+2}\right)\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F + 2I_3' \\
&\overset{(e)}{\leq} 4\sqrt{2} \times \frac{500}{499} \times \left(1 + 16\sqrt{\frac{2r(d_1+d_2)}{m}} + \frac{1}{500}\right)c_1\|\boldsymbol{\Delta}_t\|_2 \\
&\quad + 4\sqrt{2} \times \frac{500}{499} \times \left(1 + 16\sqrt{\frac{d_1+d_2}{m}} + \frac{1}{200}\right)c_1\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F + 2I_3',
\end{aligned}$$

where:

- step (a) follows from $\|\mathcal{P}_{\mathbb{T}_t^{(\boldsymbol{x},\boldsymbol{y})}}(\mathcal{A}^*\mathcal{A} - \mathcal{A}^*_{(\boldsymbol{x},\boldsymbol{y})}\mathcal{A}_{(\boldsymbol{x},\boldsymbol{y})})(\boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})} - \boldsymbol{X}_\star)\|_F$ and be estimated similarly as in (52),
- step (b) follows from Lemma 11,
- step (c) follows from (43) and (49),
- step (d) follows from Lemma 5 and $\|(\mathcal{A}^*\mathcal{A} - \mathcal{I})(\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})})\|_2 \leq \delta_{r+2}\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F$,
- step (e) follows from $\delta < c_1 < \frac{1}{1000}$.

We choose a proper constant C'' and let $m \geq C''\kappa^2 r(d_1+d_2)$ to make the coefficients before $\|\boldsymbol{\Delta}_t - \boldsymbol{\Delta}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F$ and $\|\boldsymbol{\Delta}_t\|_2$ in the last term of (52) are both smaller than $2c_1$, and those in the last term (excluding $2I_3'$) of (53) are smaller than $8c_1$. As a result, $I_3 \leq I_3' \leq 2c_1E_t$, and $I_4 \leq 8c_1E_t + 2I_3' \leq 12c_1E_t$. Besides, we have $I_1 \leq 2c_1E_t$ by (50) and $I_2 \leq 2c_1E_t$ by (51). Substituting all these in (47) gives:

$$\|\boldsymbol{W}_t - \boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F \leq (2\cdot 3 + 12)c_1\left(\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_2 + \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F\right) \leq c_1\sigma_{\min}(\boldsymbol{X}_\star), \qquad (54)$$

where we use $18c_1 < 1$ and the inductive assumption (43) in the last inequality.

To estimate $\|\boldsymbol{X}_{t+1} - \boldsymbol{X}_{t+1}^{(\boldsymbol{x},\boldsymbol{y})}\|_F$, we check the validity of Lemma 10. First, from (45) and (54),

$$\|\boldsymbol{W}_t - \boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_2 \leq \|\boldsymbol{W}_t - \boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F \leq \frac{c_1}{1-2c_1}\left(\sigma_r(\boldsymbol{W}_t) - \sigma_{r+1}(\boldsymbol{W}_t)\right) \overset{(41)}{<} (1 - 1/\sqrt{2})s.$$

Second, we define $c_0 := \frac{c_1}{1-2c_1}$, and we have $\sigma_{r+1}(\boldsymbol{W}_t) \leq c_0 s$ by (45). Then all conditions in Lemma 10 are met, and therefore it implies that: there exists a constant C_2 that is only related to c_1 such that

$$\begin{aligned}\|\boldsymbol{X}_{t+1} - \boldsymbol{X}_{t+1}^{(\boldsymbol{x},\boldsymbol{y})}\|_F &\leq 2C_2\|\boldsymbol{W}_t - \boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F \leq 2(\frac{6c_1}{1-2c_1} + 10)\|\boldsymbol{W}_t - \boldsymbol{W}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F \\ &\leq 996c_1\left(\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_2 + \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\left\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\right\|_F\right),\end{aligned} \tag{55}$$

where we use (54) in the last inequality. Summing up (55) and (46) gives

$$\begin{aligned}\|\boldsymbol{X}_{t+1} - \boldsymbol{X}_\star\|_2 + \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\|\boldsymbol{X}_{t+1} - \boldsymbol{X}_{t+1}^{(\boldsymbol{x},\boldsymbol{y})}\|_F &\leq 1000c_1(\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_2 + \sup_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{N}}\|\boldsymbol{X}_t - \boldsymbol{X}_t^{(\boldsymbol{x},\boldsymbol{y})}\|_F) \\ &\leq c_1(1000c_1)^{t+1}\sigma_{\min}(\boldsymbol{X}_\star),\end{aligned} \tag{56}$$

where in the last inequality we have used (43) and (41).

Throughout the proof, we have imposed two lower bounds on m. We then take their maximum, i.e., $m \geq C\kappa^2 r(d_1 + d_2)$ with $C = \max\{C', C''\}$, to complete the proof. □

For completeness, we also include the proof of Lemma 8, which was established in prior work given the initialization $\mathcal{H}_r(\mathcal{A}^*(\boldsymbol{b}))$ [Wei et al., 2016, Theorem 2.2]. We slightly modify the proof and show that whenever $\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F$ is sufficiently small, RGD will converge linearly to $\boldsymbol{X}_\star$.

Proof of Lemma 8. The proof follows the same structure as [Wei et al., 2016, Theorem 2.2]. Since $\boldsymbol{X}_{t+1} = \mathcal{H}_r(\boldsymbol{W}_t)$ is the best rank-r approximation to $\boldsymbol{W}_t$, we have

$$\|\boldsymbol{X}_{t+1} - \boldsymbol{X}_\star\|_F \leq \|\boldsymbol{X}_{t+1} - \boldsymbol{W}_t\|_F + \|\boldsymbol{W}_t - \boldsymbol{X}_\star\|_F \leq 2\left\|\boldsymbol{W}_t - \boldsymbol{X}_\star\right\|_F.$$

Substituting $\boldsymbol{W}_t = \boldsymbol{X}_t + \mathcal{P}_{\mathbb{T}_t}\mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}_\star - \boldsymbol{X}_t\right)$ into the above inequality yields:

$$\begin{aligned}\|\boldsymbol{X}_{t+1} - \boldsymbol{X}_\star\|_F &\leq 2\left\|\left(\mathcal{I} - \mathcal{P}_{\mathbb{T}_t}\mathcal{A}^*\mathcal{A}\right)\left(\boldsymbol{X}_\star - \boldsymbol{X}_t\right)\right\|_F \\ &\leq 2\left\|\left(\mathcal{I} - \mathcal{P}_{\mathbb{T}_t}\right)\left(\boldsymbol{X}_\star - \boldsymbol{X}_t\right)\right\|_F + 2\left\|\mathcal{P}_{\mathbb{T}_t}\left(I - \mathcal{A}^*\mathcal{A}\right)\left(\boldsymbol{X}_\star - \boldsymbol{X}_t\right)\right\|_F \\ &\overset{(a)}{\leq} \frac{2}{\sigma_{\min}(\boldsymbol{X}_\star)}\left\|\boldsymbol{X}_t - \boldsymbol{X}_\star\right\|_F^2 + 2\left\|\mathcal{P}_{\mathbb{T}_t}\left(I - \mathcal{A}^*\mathcal{A}\right)\mathcal{P}_{\mathbb{T}_t}\left(\boldsymbol{X}_\star - \boldsymbol{X}_t\right)\right\|_F \\ &\quad + 2\left\|\mathcal{P}_{\mathbb{T}_t}\left(I - \mathcal{A}^*\mathcal{A}\right)\left(I - \mathcal{P}_{\mathbb{T}_t}\right)\left(\boldsymbol{X}_\star - \boldsymbol{X}_t\right)\right\|_F \\ &\overset{(b)}{\leq} \frac{2}{\sigma_{\min}(\boldsymbol{X}_\star)}\left\|\boldsymbol{X}_t - \boldsymbol{X}_\star\right\|_F^2 + 2\delta_{2r}\|\boldsymbol{X}_\star - \boldsymbol{X}_t\|_F + 2\delta_{3r}\left\|\boldsymbol{X}_\star - \boldsymbol{X}_t\right\|_F \\ &\leq \left(2\frac{\|\boldsymbol{X}_\star - \boldsymbol{X}_t\|_F}{\sigma_{\min}(\boldsymbol{X}_\star)} + 4c_2\right)\left\|\boldsymbol{X}_\star - \boldsymbol{X}_t\right\|_F,\end{aligned}$$

where (a) follows from Lemma 12, and (b) follows from Lemma 2 and the inequalities $\delta_{2r} \leq \delta_{3r} \leq \delta_{6r} \leq c_2$.

Define

$$\gamma_t = 2\frac{\|\boldsymbol{X}_\star - \boldsymbol{X}_t\|_F}{\sigma_{\min}(\boldsymbol{X}_\star)} + 4c_2.$$

By the condition (23), we have $\gamma_T \leq 6c_2 < 1$. The remainder of the proof proceeds by induction. Assume $\gamma_k < 6c_2$ for $k = T, T+1, \ldots, t$. Then, we have

$$\|\boldsymbol{X}_t - \boldsymbol{X}_\star\|_F \leq (6c_2)^{t-T}\|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F \leq \|\boldsymbol{X}_T - \boldsymbol{X}_\star\|_F \leq c_2\sigma_{\min}(\boldsymbol{X}_\star).$$

Therefore, $\gamma_{t+1} \leq 6c_2$. By induction, we conclude that $\gamma_t < 6c_2$ for all $t \geq T$. □