Leave-One-Out Approach for Matrix Completion: Primal and Dual Analysis

Lijun Ding\textsuperscript{\textcopyright} and Yudong Chen\textsuperscript{\textcopyright}

Abstract—In this paper, we introduce a powerful technique based on Leave-One-Out analysis to the study of low-rank matrix completion problems. Using this technique, we develop a general approach for obtaining fine-grained, entrywise bounds for iterative stochastic procedures in the presence of probabilistic dependency. We demonstrate the power of this approach in analyzing two of the most important algorithms for matrix completion: (i) the non-convex approach based on Projected Gradient Descent (PGD) for a rank-constrained formulation, also known as the Singular Value Projection algorithm, and (ii) the convex relaxation approach based on nuclear norm minimization (NNM). Using this approach, we establish the first convergence guarantee for the original form of PGD without regularization or sample splitting, and in particular shows that it converges linearly in the infinity norm. For NNM, we use this approach to study a fictitious iterative procedure that arises in the dual analysis. Our results show that NNM recovers an $d$-by-$d$ rank-$r$ matrix with $O(\mu r \log(\mu r) d \log d)$ observed entries. This bound has optimal dependence on the matrix dimension and is independent of the condition number. To the best of our knowledge, none of previous sample complexity results for tractable matrix completion algorithms satisfies these two properties simultaneously.

Index Terms—Matrix completion, leave-one-out, statistical learning.

I. INTRODUCTION

THE matrix completion problem concerns recovering a low-rank matrix given a (typically random) subset of its entries. To study the sample complexity and algorithmic behaviors of this problem, one often needs to analyze an iterative procedure in the presence of dependency across the iterations and the entries of the iterates. Such dependency creates significant difficulties in both the design and analysis of algorithms, often leading to sub-optimal bounds as well as complicated and unrealistic algorithms that are not used in practice.

To overcome these challenges, in this paper we introduce a powerful technique, based on the Leave-One-Out argument, to the study of matrix completion problems. Leave-One-Out, as an analytical technique, allows one to isolate the effect of dependency on individual entries, and establish entrywise bounds for the iterative procedures. We use this technique to obtain new theoretical guarantees for two archetypal algorithms for matrix completion: (i) the non-convex approach based on applying Projected Gradient Descent (PGD) to a rank-constrained formulation, also known as the Singular Value Projection (SVP) algorithm [21]; (ii) the convex relaxation method based on Nuclear Norm Minimization (NNM) [4]. We use leave-one-out in two different ways. For PGD, we employ this technique to study the primal solution path of the algorithm. For NNM, we analyze an iterative procedure that arises in the analysis, particularly for constructing a dual solution that certifies the optimality of the desired primal solution.

Concretely, consider the problem of recovering a rank-$r$ matrix $M^* \in \mathbb{R}^{d_1 \times d_2}$ given a subset of its entries, $\{M^*_{ij} : (i, j) \in \Omega\}$. As this problem is ill-posed for general $\Omega$, it is standard to assume that $\Omega$ is generated according to some probabilistic model. To recover $M^*$, a natural idea is to seek a low-rank matrix that is consistent with the observations in $\Omega$. Based on this idea, two most representative algorithms for matrix completion are the following:

a) Projected gradient descent for rank-constrained formulation: This approach is based on solving a natural rank-constrained, least-squares formulation for matrix completion:

\[
\begin{align*}
\text{minimize}_{X \in \mathbb{R}^{d_1 \times d_2}} & \quad \frac{1}{2} \| \Pi_\Omega(X) - \Pi_\Omega(M^*) \|^2_F \\
\text{subject to} & \quad \text{rank}(X) \leq r;
\end{align*}
\]

where $\Pi_\Omega(\cdot) : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R}^{d_1 \times d_2}$ is the linear operator that zeros out the entries outside $\Omega$. PGD applied to the above optimization problem takes the form

\[
\begin{align*}
M^0 &= 0; \text{ for } t = 0, 1, 2, \ldots, \\
M^{t+1} &= \mathcal{P}_r \left( M^t - \eta_t (\Pi_\Omega(M^t) - \Pi_\Omega(M^*)) \right),
\end{align*}
\]

where $\eta_t$ is the step size and $\mathcal{P}_r$ is the projection operator onto the set of rank-$r$ matrices. Note that while this set is non-convex, the projection $\mathcal{P}_r$ can be efficiently computed by the rank-$r$ singular value decomposition (SVD). This approach is also known as Singular Value Projection (SVP) or iterative hard thresholding [21].

b) Nuclear norm minimization: Since the rank function is non-convex, another popular approach for matrix completion is based on replacing the rank with a convex surrogate, namely the nuclear norm. This relaxation leads to the following convex
nuclear norm minimization (NNM) problem [4]:

\[
\text{minimize}_{X \in \mathbb{R}^{d_1 \times d_2}} \|X\|_{\text{nuc}} \\
\text{subject to} \quad \Pi_{\Omega}(X) = \Pi_{\Omega}(M^*),
\]

(3)

where \(\|X\|_{\text{nuc}}\) denotes the nuclear norm of \(X\), defined as the sum of its singular values.

Both PGD and NNM can be efficiently computed/solved. The key statistical question here is when these two approaches recover the true low-rank matrix \(M^*\) under natural probabilistic models for the observed data \(\Omega\). Perhaps surprisingly, while matrix completion has been extensively studied, a complete answer to the above question remains elusive. As we elaborate below, existing techniques are fundamentally insufficient in this regard, either relying on assumptions that are difficult (sometimes impossible) to verify, or resulting in performance bounds that are inherently sub-optimal.

**A. Our Contributions**

Our key insight to the answer of the above question, is as follows. While the PGD and NNM approaches appear completely different, their analysis can both be reduced to studying a stochastic iterative procedure of the form

\[
\theta^{t+1} = F(\theta^t; \delta), \quad \text{for } t = 0, 1, \ldots.
\]

(4)

Here \(F(\cdot; \delta)\) is a possibly nonlinear and implicit mapping with a fixed point \(\theta^*\), and \(\delta\) represents a random data vector. Note that the same \(\delta\) is used in all iterations. Therefore, a major challenge here is that the iterates \(\{\theta^t\}\) are dependent through the common data \(\delta\). The Leave-One-Out analysis allows us to isolate such dependency and provide fine-grained, entrywise convergence guarantees, namely, bounds on \(\|\theta^t - \theta^*\|_{\infty} := \max_i |\theta^t_i - \theta^*_i|\). As discussed below, these bounds are crucial to the analysis of PGD and NNM. We now elaborate.

1) **Projected Gradient Descent**: The PGD algorithm (2) can be recognized as a special case of the iteration (4), where the nonlinear map \(F\) is given implicitly by the projection \(P_M\) (i.e., an SVD), and the random data \(\delta\) corresponds to the observed indices \(\Omega\). A major roadblock in analyzing this iterative procedure involves showing that for all \(t\), the differences of iterates \(M^{t+1} - M^t\) remain incoherent, which roughly means that they are entrywise well-bounded. Such bounds are challenging to obtain due to the probabilistic dependency across the iterations. The seminal work [21] only provides partial results, imposing as an assumption that \(M^{t+1} - M^t\) is incoherent. Another set of work [20], [22], [23] resorts to a sample splitting trick, that is, assuming that a fresh set of independent observations \(\Omega\) is used in each iteration. As we comment on in greater details in Section III-B, this trick is artificial, difficult to implement, and unnecessary in practice; moreover it leads to sample complexity bounds that are either non-rigorous or inherently suboptimal.

Using leave-one-out, we are able to study the original form (4) of PGD—without sample splitting—and rigorously prove that the iterates indeed remain entrywise small. In fact, a stronger conclusion is established: the iterates converge geometrically in entrywise norm to \(M^*\); see Theorem 1 for details.

2) **Nuclear Norm Minimization**: To prove that the convex NNM program (3) recovers the underlying matrix \(M^*\) as the optimal solution, it suffices to show that \(M^*\) satisfies the first-order optimality condition, which stipulates the existence of a corresponding dual optimal solution (often called a “dual certificate”). Such a dual certificate can be constructed using an iterative procedure, akin to dual ascent, in the form of the iteration (4). The celebrated Golfing Scheme argument [18] implements such a procedure, but it crucially relies on the sample splitting trick to circumvent the dependency across iterations. While this argument has proved to be fruitful and led to the best sampling complexity bounds to date [3], [7], [30], it is well-recognized that sample splitting is a workaround and results in fundamentally sub-optimal bounds.

Using leave-one-out, we are able to analyze the dual ascent procedure with correlated iterations and establish entrywise bounds, which ensure dual feasibility. Our results imply that NNM recovers a \(d\)-by-\(d\) rank-\(r\) matrix \(M^*\) given a number of \(C\mu r \log(\mu r) d \log d\) observed entries, where \(\mu\) is the incoherence parameter of \(M^*\) and \(C\) is a universal constant; see Theorem 2 for details. To the best of our knowledge, this is the first sample complexity result, for a tractable algorithm, that has the optimal scaling with the dimension \(d\) while at the same time carries absolutely no dependence on the condition number of \(M^*\). We believe that our result paves the way to finally matching the information-theoretic lower bound \(C\mu r d \log d\) [5].

We emphasize that in both settings above, the mapping \(F(\theta; \delta)\) cannot satisfy a contraction property over all \(\theta\)'s, even when restricted to those that are low-rank and incoherent; see Sections III and IV for detailed discussion. Therefore, standard techniques from stochastic approximation [26] are insufficient for our problem. The key step in our analysis is to show that with high probability, a type of contraction is satisfied by the sequence of iterates \(\{\theta^t\}\) generated by the procedure (4); that is, the iterative procedure avoids the bad regions of \(\theta\) in which contraction fails to hold. The leave-one-out argument plays a key role in establishing this probabilistic, sequence-specific convergence result.

**B. Discussion**

In this paper we focus on the PGD and NNM approaches. While algorithms for matrix completion abound, PGD and NNM are of fundamental importance. In particular, NNM, and more broadly convex relaxation methods, remains one of the most versatile, robust and statistically efficient approaches to high-dimensional statistical problems. Similarly, PGD plays a unique role in a growing line of work on non-convex methods. It is recognized as a particularly natural and simple algorithm, does not require a two-step procedure of “initialization + local refinement”, and in fact is often used as an initialization procedure for other algorithms [9], [23], [32], [35]. Moreover, PGD involves one of the most important numerical procedure: computing the best low-rank approximation using SVD. Many other algorithms can be either viewed as approximate or perturbed versions of PGD/SVD, or as computationally efficient procedures for solving NNM.
In this sense, while we apply Leave-One-Out to PGD and NNM specifically, we believe that this technique is useful more broadly in studying other iterative procedures for statistical problems with complex probabilistic structures. Indeed, when preparing an early version of this manuscript [13], we became aware of the independent work in [10], [29], which uses a related technique to analyze other iterative methods for non-convex formulations of matrix completion and phase retrieval problems. We discuss this work in more details in Section II.

Our results can also be viewed as establishing a form of implicit regularization. In particular, note that neither the rank-constrained formulation (1), nor the iterative procedures we consider for PGD and NNM, has an explicit mechanism for regularizing their solutions to have small entry-wise ($\ell_\infty$) norms. The goal of the leave-one-out analysis is precisely to show that this property is satisfied automatically, with high probability, by the solution sequence generated by the iterative procedures. From this perspective, our results are complementary to a very recent line of work on $\ell_1/\ell_2$ implicit regularization of (stochastic) gradient descent methods [19], [28].

**Paper Organization:** In Section II, we review and compare with existing work in the literature. In Section III, we present our main results for PGD and NNM. In Section IV, we outline the main ingredients of our Leave-One-Out based technique. Using this technique, we prove our results for PGD for NNM in Sections V and VI, respectively. The proofs of some technical lemmas are deferred to the appendix.

## II. Related Work and Comparison

Leave-One-Out has many incarnations, and is often used as an algorithmic technique, e.g., for cross-validation. As an analytical technique, leave-one-out has been employed to study robust M-estimation [14], de-biased Lasso estimators [24], and spectral and MLE methods for ranking problems [11].

Most related to our work are several recent papers that use leave-one-out in problems involving low-rank matrix estimation. The work in [36] studies the generalized power method for phase synchronization problems. The work in [1] derives general entrywise perturbation bounds for spectral decomposition. The contemporary work in [6], [10], [29] also studies nonconvex formulations of matrix completion, but focuses on a different algorithm, namely, gradient descent applied to an unconstrained and factorized objective function.

Besides the differences in problem settings, our use of leave-one-out differs from the above work in the following three main aspects:

1) The work in [1], [11] consider “one-shot” spectral methods, which only involve a single SVD operation. In contrast, the PGD method we study is an iterative procedure with multiple sequential SVD operations. As will become clear in our proofs, even studying the second iteration of PGD involves a very different analysis than that for one-shot algorithms. In particular, we need to track the propagation of errors and dependency through many (potentially infinite) iterations, which requires careful induction and probabilistic arguments.

2) The work in [6], [10], [29], [36] study gradient descent and power methods, which are iterative procedures in the form of (4). Both methods correspond to an explicit and relatively simple mapping $F$. The PGD algorithm is much more complicated, as it involves computing the SVD, a highly nonlinear operation that is defined implicitly and variationally. The analysis of PGD is hence significantly harder, requiring quite delicate use of matrix perturbation and concentration bounds.

3) In our analysis of NNM, leave-one-out is used in a different context: instead of studying an actual algorithmic procedure, we use leave-one-out to study an (unimplementable) iterative procedure that arises in the dual analysis of the convex program.

The exact low-rank matrix completion problem is studied in the seminar work [4], which initialized the use of the NNM approach. Follow-up work on NNM includes [5], [18], [30], with the best existing sample complexity result given in [7]. The PGD algorithm is proposed in [21] under the name SVP, although no rigorous guarantees are provided for matrix completion. Other iterative algorithms based on non-convex formulations of matrix completion have been proposed; a partial list of work in this line includes [9], [20], [25], [32], [35]. These algorithms are different from PGD: they are typically based on a factorized formulation (rather than the rank-constrained formulation (1)), and often require explicit $\ell_\infty$ projection/regularization and a careful initialization procedure via SVD. With the $\ell_\infty$ regularization, a remarkable recent result shows that this factorized formulation in fact has no spurious local minima [16], [17], though the resulting iteration complexity bounds therein are quite pessimistic. After presenting our main theorem, we provide a more quantitative comparison with the above results.

PGD and NNM have been well studied for the related problem of matrix sensing [21], [31], whose standard formulation can be viewed as a simpler version of matrix completion with the projection $\Pi_{\Omega}$ replaced by a linear operator $A$ that satisfies certain restricted isometry property (RIP) over all low-rank matrices. The lack of such a global RIP/contraction property in matrix completion makes it a more challenging problem and is precisely the reason why leave-one-out is needed.

## III. Problem Setup and Main Theorems

In this section, we describe the formal setup of the matrix completion problem and present our main results, namely convergence and sample complexity guarantees for PGD and NNM.

**Notations:** For each integer $d > 0$, define the set $[d] := \{1, 2, \ldots, d\}$. We write $f = O(g)$ or $f \preceq g$ if $f \leq C \cdot g$ for a universal numerical constant $C$. Denote by $e_i$ the $i$-th standard basis vector, $I$ the all-one vector, and $I$ the identity matrix, in appropriate dimensions. The set of symmetric matrices in $\mathbb{R}^{d \times d}$ is denoted by $S^{d \times d}$. For a matrix $Z$, let $Z_i$ be its $i$-th row and $Z_j$ its $j$-th column. The Frobenius norm, operator norm (maximum singular value) and nuclear norm (sum of singular values) of a matrix are denoted by $\| \cdot \|_F$, $\| \cdot \|_2$, and $\| \cdot \|_{\text{op}}$, respectively. Two other matrix norms are used:
\[ \|Z\|_{\infty} := \max_{i,j} |Z_{ij}| \] for the entrywise \( \ell_\infty \) norm, and
\[ \|Z\|_{2,\infty} := \max_i \|Z_i\|_2 \] for the maximum row \( \ell_2 \) norm. The \( i \)-th largest singular value of a matrix \( Z \) is \( \sigma_i(Z) \), and the best rank-\( k \) approximation of \( Z \) in Frobenius norm is \( P_k(Z) \).

For two matrices \( A \) and \( B \), we write \( A \otimes B := AB^T \) for their outer product. We denote by \( I \) the identity operator on matrices.

A. Matrix Completion Setup

In matrix completion, the goal is to recover an unknown rank-\( r \) matrix \( M^* \in \mathbb{R}^{d_1 \times d_2} \) given partial observations of its entries indexed by the set \( \Omega \subset [d_1] \times [d_2] \). Define the observation indicator \( \delta_{ij} := 1 \{ (i,j) \in \Omega \} \), as well as the sampling operator \( \Pi_\Omega : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2} \) via
\[ (\Pi_\Omega(Z))_{ij} = Z_{ij}\delta_{ij} = \begin{cases} Z_{ij}, & (i,j) \in \Omega, \\ 0, & (i,j) \notin \Omega. \end{cases} \]

It is well known that when most of the entries of \( M^* \) equal zero, it is impossible to recover \( M^* \) unless all of its entries are observed [4]. To avoid such pathological situations, we impose the standard assumption that \( M^* \) is incoherent in the following sense:

**Definition 1 (Incoherence):** A matrix \( M \in \mathbb{R}^{d_1 \times d_2} \) with rank-\( r \) SVD \( M = USV^T \) is \( \mu \)-incoherent if
\[ \|U\|_{2,\infty} \leq \sqrt{\frac{\mu r}{d_1}} \quad \text{and} \quad \|V\|_{2,\infty} \leq \sqrt{\frac{\mu r}{d_2}}. \]

In the sequel, we assume that \( M^* \) has rank \( r \) and is \( \mu \)-incoherent. If another matrix \( Z \in \mathbb{R}^{d_1 \times d_2} \) is \( \mathcal{O}(\mu) \)-incoherent, we simply say that \( Z \) is incoherent. Denote by \( \kappa := \sigma_1(M^*)/\sigma_r(M^*) \) the condition number of \( M^* \). Throughout this paper, by *with high probability* (w.h.p.), we mean with probability at least \( 1 - c_1(d_1 + d_2)^{-c_2} \) for some universal constants \( c_1, c_2 > 0 \).\(^1\)

B. Analysis of Projected Gradient Descent

To study the PGD algorithm (2), we consider a standard probabilistic setting where the observation indices \( \Omega \) are randomly generated. We focus on the following symmetric and positive semidefinite setting.\(^2\)

**Model 1:** Under the model SMC\((M^*, p)\), the matrix \( M^* \in \mathcal{S}^{d_1 \times d_2} \) is symmetric positive semidefinite, and the observation indices \( \Omega \) is such that \( \mathbb{P}( (i,j) \in \Omega ) = p \) independently for all \( i \geq j \), where \( p \in (0, 1] \), and that \( (j,i) \in \Omega \) if and only if \( (i,j) \in \Omega \).

The PGD algorithm (2) is first proposed by Jain et al. in [21]. They observe empirically that the objective value

\[ \|\Pi_\Omega(M^t - M^*)\|_2^2 \] of the PGD iterate decreases quickly to zero; accordingly, they conjecture that the iterate \( M^t \) is guaranteed to converge to \( M^* \) [21, Conjecture 4.3]. Below we reproduce their conjecture, which is rephrased under our symmetric setting:

**Conjecture 1:** For some numbers \( C, C' > 0 \) depending on \( r \) and \( \mu \), the following holds under the model SMC\((M^*, p)\). If \( p \geq C \log^2 d \), then with high probability, the PGD algorithm (2) with fixed step size \( \eta_t = \mathcal{O}(\frac{1}{\sqrt{T}}) \) outputs a matrix \( M^t \) of rank at most \( r \) such that \( ||\Pi_\Omega(M^t) - \Pi_\Omega(M^*)||_2^2 \leq \epsilon \) after \( C' \log \frac{1}{\epsilon} \) iterations; moreover, \( M^t \) converges to \( M^* \).

This conjecture remains open since the proposal of PGD. The original PGD paper [21] argues that PGD would converge if the operator \( \Pi_\Omega \) is assumed to satisfy a form of Restricted Isometry Property (RIP), i.e., \( \Pi_\Omega(M^t - M^*) \) preserves the Frobenius norm of the error matrix \( M^t - M^* \). However, RIP cannot hold uniformly for all low-rank matrices—just consider matrices with only one non-zero entry. Even when one restricts to incoherent iterates \( M^t \), the error matrix \( M^t - M^* \), being the difference of two incoherent matrices, need not be incoherent itself.

Instead of relying on RIP and the Frobenius (Euclidean) norm geometry, we directly control the entries of the error matrix using Leave-One-Out. In particular, we show that *every entry* of the eigenvectors of \( M^t \) converges to that of \( M^* \) simultaneously and geometrically, hence \( M^t \) converges entrywise to \( M^* \) as well. This result formally proves Conjecture 1.

**Theorem 1:** Under the model SMC\((M^*, p)\), if \( p \geq C_0 \mu^2 \log^2 d \) for some universal constant \( C_0 \), then with high probability the PGD algorithm (2) with fixed step size \( \eta_t = \frac{1}{p} \) satisfies the bound
\[ \|M^t - M^*\|_{\infty} \leq \left( \frac{1}{2} \right)^t \|\sigma_1(M^*)\|_1, \quad \text{for all } t \geq 1. \] (5)

Moreover, the first few iterations satisfy the tighter bound
\[ \|M^t - M^*\|_{\infty} \leq \left( \frac{1}{2} \right)^t \frac{\|\sigma_1(M^*)\|_1}{d}, \quad \text{for all } t \in [\log d]. \]

We prove Theorem 1 in Section V, by casting PGD into a stochastic iterative procedure in the form of (4), namely \( \theta^{t+1} = F(\theta^t; \delta) \). In particular, the random data \( \delta \) consists of the observation indicators \( \delta_{ij} := 1 \{ (i,j) \in \Omega \} \) generated according to Model 1, and the PGD iteration (2) acts on the primal variable \( \theta^t \equiv M^t \) with the map \( F \) given by
\[ F(\cdot; \delta) := P_r \left( (I - p^{-1}\Pi_\Omega) \cdot + p^{-1}\Pi_\Omega(M^*) \right). \] (6)

We establish entrywise geometric convergence of this procedure using the leave-one-out technique; the main ideas of the analysis is outlined in Section IV.

1) Discussion and Comparison: Theorem 1 establishes, for the first time, the convergence of the original form of PGD. Our result holds when the same set of observed entries \( \Omega \) is used in all iterations, without any sampling splitting.

In comparison, existing work in [20], [22], [23] has considered certain modified versions of PGD. These algorithms are significantly more complicated than the original PGD: they typically proceed in a stagewise fashion and rely on sampling splitting, i.e., using an independent set of observations \( \Omega \) in each iteration. It is well recognized that sample splitting has

\[^1\] In all our proofs, the value of \( c_2 \) can be made arbitrarily large as long as the constant \( C_0 \) in Theorems 1 and 2 is sufficiently large. In this case, if each of a polynomial (in \( d_1 \) and \( d_2 \)) number of events holds w.h.p. (with probability \( \geq 1 - c_1(d_1 + d_2)^{-c_2} \)), then by union bound the intersection of these events also holds w.h.p. (with probability \( \geq 1 - c_1(d_1 + d_2)^{-c_2} \) for a constant \( c_2 < c_2 \).

\[^2\] We consider this setting for the sake of a streamlined presentation of our main techniques. Our results can be extended to the general asymmetric case either via a direct analysis, or by using an appropriate form of the standard dilation argument (see, e.g., [35]), though the proofs will become more tedious.
several major drawbacks [20], [32]. Firstly, it is a wasteful way of using the data, and leads to sample complexity bounds that grow (unnecessarily) with the number of iterations. Secondly, the use of sampling splitting is artificial, resulting in algorithmic complications that are not needed in practice. Finally, as observed in [20], [32], naive sample splitting (i.e., partitioning \( \Omega \) into disjoint subsets) in fact does not ensure the required independence; rigorously addressing this technical subtlety (as done in [20]) leads to even more complicated algorithms that are sensitive to the generative model of \( \Omega \) and hence hardly practical.

A consequence of Theorem 1 is that the PGD iterates \( M_t \) remains incoherent throughout the iterations (though incoherence and RIP are no longer needed explicitly in our convergence proof). Note that PGD, and the nonconvex program (1) it aims to solve, have no explicit regularization mechanism to ensure incoherence. Therefore, while natural and simple, the PGD algorithm is effective for quite delicate probabilistic reasons, which are tied to the specific algorithmic procedure and cannot be simply explained by the geometry of the optimization problem (1).

C. Analysis of Nuclear Norm Minimization

To present our results on NNM, we consider the following setting in which the ground-truth \( M^* \) is allowed to be a general rectangular and asymmetric matrix.

Model 2: Under the model \( MC(M^*, p) \), \( M^* \) is a \( d_1 \)-by-\( d_2 \) matrix, and the observation indices \( \Omega \) is such that \( \mathbb{P}( (i, j) \in \Omega ) = p \) independently for all \( i, j \), where \( p \in (0, 1] \).

Starting with the seminar papers [4], [25], a long line of work has been devoted to proving sample complexity results for the above model, that is, sufficient conditions for recovering \( M^* \) using NNM and other algorithms. We summarize the state-of-the-art in Table I, omitting other existing results that are strictly dominated by those in the table.

Using Leave-One-Out, we are able to improve upon this long line of work and establish the following new sample complexity result for NNM.

**Theorem 2:** Under the model \( MC(M^*, p) \), if \( p \geq C_0 \mu r \log((\mu r) \log(\max(d_1, d_2))) \) for some universal constant \( C_0 \), then with high probability \( M^* \) is the unique minimizer of the NNM program (3).

We prove this theorem in Section VI, by connecting NNM to the stochastic iterative procedure (4) in the form \( \theta^{t+1} = \mathcal{F}(\theta^t, \delta) \). In particular, we consider an iterative procedure acting on the dual variable \( \theta^t \) of NNM, with the map \( \mathcal{F} \) given by

\[
\mathcal{F}(\cdot; \delta) = (\mathcal{P}_\delta - p^{-1} \mathcal{P}_\delta \Pi_{\Theta}) (\cdot); \tag{7}
\]

here \( \mathcal{P}_\delta \) is the projection onto the tangent space at \( M^* \) with respect to the set of low-rank matrices (the explicit expression of \( \mathcal{P}_\delta \) is given in Section VI), and the data \( \delta \) consists of the observation indicators \( \delta_{ij} := 1 \{ (i, j) \in \Omega \} \) under Model 2.

We show that with high probability, the above procedure converges to an optimal dual solution that certifies the primal optimality of \( M^* \) to the NNM program (3). A key step is the proof is to show that the iterates are dual feasible, which in turn requires bounding their \( \ell_\infty \) norm. We do so using the leave-one-out technique, with the main ideas of the analysis outlined in Section IV.

1) Discussion and Comparison: In the setting with \( d_1 = d_2 = d \), Theorem 2 shows that NNM recovers \( M^* \) w.h.p. provided that the expected number of observed entries satisfies \( pd^2 \gtrsim \mu r \log(\mu r) d \log d \). Note that this bound is independent of the condition number \( \kappa \) of \( M^* \). An information-theoretic lower bound on the sample complexity is established in [5], which shows that \( pd \gtrsim \mu r d \log d \) is necessary for any algorithm. Our bound hence has the optimal dependence on \( d \) and \( \kappa \), and is sub-optimal by a logarithmic term of the incoherence parameter \( \mu \) and rank \( r \).

Let us compare Theorem 2 with the sample complexity results in Table I. The best existing result for NNM appears in [7], which establishes a bound that scales sub-optimally with \( \mu^2 d \). This gap is a fundamental consequence of their proof techniques, as they rely on the Golfing Scheme [18] that splits \( \Omega \) into \( \log d \) disjoint subsets to ensure independence. All other previous results in the table have non-trivial dependence on the condition number \( \kappa \). While it is common to see dependency on \( \kappa \) in the time complexity, the appearance of \( \kappa \) in the sample complexity is unnecessary. To the best of our knowledge, our result is the only one that achieves optimal dependence on both the condition number and the dimension for tractable algorithms; in particular, our result is not dominated by any existing results.

We note that the very recent work in [2] obtains a sample complexity result that matches the lower bound; their bound, however, is achieved by an algorithm with running time exponential in \( d \).

IV. LEAVE-ONE-OUT ANALYSIS OF STOCHASTIC ITERATIVE PROCEDURES

As mentioned, we prove our main results for PGD and NNM by using Leave-One-Out to analyze certain stochastic iterative procedures and obtain entrywise bounds. In this section, we present the main ingredients of this approach. We first describe the general ideas of Leave-One-Out in the context of the abstract stochastic iteration (4), and then discuss the additional steps needed for the concrete settings of PGD and NNM. The complete proofs are given in Sections V and VI to follow.
A. Stochastic Iterative Procedures

Consider the stochastic iterative procedure in (4), namely, \( \theta^{t+1} = F(\theta^t; \delta) \). Here \( \delta = (\delta_1, \ldots, \delta_N) \in \mathbb{R}^N \) is a random data vector with independent coordinates, and \( F(\cdot; \delta) : \mathbb{R}^N \rightarrow \mathbb{R}^N \) is a nonlinear map with a fixed point \( \theta^* \). For simplicity, we assume that \( \theta^* = 0 \). Our goal is to study the convergence behavior of the iterates \( \theta^t \) to the fixed point \( \theta^* \).

If \( F \) is a contraction in the \( \ell_2 \) norm in the sense that with high probability, uniformly for all \( \theta, \theta' \),
\[
\| F(\theta; \delta) - F(\theta'; \delta) \|_2 \leq \alpha \| \theta - \theta' \|_2,
\]
for some \( \alpha < 1 \), then it is straightforward to show that the \( \ell_2 \) distance to the fixed point, \( \| \theta^t \|_2 \), decreases geometrically to zero. This contraction argument is classical, but often insufficient.

- In some cases, one is interested in controlling the entrywise behaviors of the iterates, i.e., bounding its \( \ell_\infty \) norm \( \| \theta^t \|_\infty \). Using the worst-case inequality \( \| \theta^t \|_\infty \leq \| \theta^t \|_2 \), together with the above \( \ell_2 \) distance bound, is often far too loose. This is the situation we will encounter in the analysis of NNM.

- Worse yet, there are settings where the \( \ell_2 \) contraction does not hold uniformly for all \( \theta \) and \( \theta' \); instead, a restricted version holds:
\[
\| F(\theta; \delta) - F(\theta'; \delta) \|_2 \leq \alpha \| \theta - \theta' \|_2, \\
\forall \theta, \theta' : \| \theta \|_\infty, \| \theta' \|_\infty \leq b
\]
for some small number \( b \). In this case, establishing convergence of \( \theta^t \) requires one to first control the \( \ell_\infty \) norm of \( \theta^t \). This is the situation we will encounter in the analysis of PGD.

In both situations, one needs to control the individual coordinates of the iterates. The Leave-One-Out argument allows us to do so by exploiting the independence of the coordinates of the data vector \( \delta \), and by exploiting the fine-grained structures of the map \( F \).

For illustration, we assume that iteration (4) is separable w.r.t. the data vector \( \delta \), in the sense that
\[
\theta^{t+1}_i = F_i(\theta^t; \delta_i).
\]
That is, the \( i \)-th coordinate of the iterate has explicit dependence only on the \( i \)-th coordinate of the data \( \delta \). Note that \( \theta^{t+1}_i \) also depends on all coordinates of \( \theta^t \), which in turn depends on the entire vector \( \delta \). Consequently, all coordinates of \( \theta^{t+1} \), for all iterations \( t \), are still correlated with each other.

Our crucial observation is that the map \( F(\theta; \delta) \) is often not too sensitive to individual coordinates of \( \theta \). In this case, we expect that the randomness of \( \delta \) propagates slowly across the coordinates, so the correlation between \( \theta^{t+1}_i \) and \( \{ \delta_j, j \neq i \} \) is relatively weak even though they are not independent. To formalize this insensitivity property, we assume that \( F \) satisfies, in addition to the restricted \( \ell_2 \)-contraction bound (8), the following \( \ell_\infty/\ell_2 \) Lipschitz condition
\[
\| F(\theta; \delta) - F(\theta'; \delta) \|_\infty \leq \beta \| \theta - \theta' \|_2, \quad \forall \theta, \theta'.
\]
The value of \( \beta \) is often small, since we are comparing \( \ell_\infty \) norm with \( \ell_2 \) norm. However, one should expect that \( \beta < \frac{1}{\sqrt{N}} \), as otherwise we would have \( \ell_\infty \) contraction/non-expansion, which is what we try to prove in the first place.

B. Leave-One-Out Analysis

We are now ready to describe the leave-one-out argument, which allows us to exploit the properties (8)-(10) and isolate the behavior of individual coordinates. For each \( i \in [N] \), let \( \delta_i^- := (\delta_1, \ldots, \delta_{i-1}, 0, \delta_{i+1}, \ldots, \delta_N) \) be the vector obtained from the original data vector \( \delta \) by zeroing out its \( i \)-th coordinate. Consider the fictitious iteration (used only in the analysis) \( c \)
\[
\theta^{0,i} = \theta^0; \theta^{t+1,i} = F(\theta^{t,i}; \delta_i^-), t = 1, 2, \ldots
\]
Crucially, \( \theta^{t+1,i} \) is independent of \( \delta_i \) by construction. Our strategy is to show that the leave-one-out iterates \( \theta^{t,i} \) closely approximate the original iterates \( \theta^t \), thereby leveraging the independence in \( \theta^{t,i} \) to bound the coordinates of \( \theta^t \).

To this end, we use induction on \( t \), with the hypothesis that \( \| \theta^t - \theta^{t,i} \|_2 \) (proximity) and \( \| \theta^t \|_\infty \) (\( \ell_\infty \) bound) are small in an appropriate sense. For the next iteration \( t+1 \), it is intuitive that \( \| \theta^{t+1} - \theta^{t+1,i} \|_2 \) should remain small, since \( \theta^{t+1} \) and \( \theta^{t+1,i} \) are computed using two data vectors different at only one coordinate. More quantitatively, we can establish the proximity property by
\[
\| \theta^{t+1} - \theta^{t+1,i} \|_2 = \| F(\theta^t; \delta) - F(\theta^{t,i}; \delta_i^-) \|_2 \\
\leq \| F(\theta^t; \delta) - F(\theta^{t,i}; \delta) \|_2 + \| F(\theta^{t,i}; \delta) - F(\theta^{t,i}; \delta_i^-) \|_2 \\
= \| F(\theta^t; \delta) - F(\theta^{t,i}; \delta) \|_2 + \| F_i(\theta^{t,i}; \delta_i^-) - F_i(\theta^{t,i}; \delta_i) \|_2
\]
where the last step follows from the separability assumption (9). The discrepancy term above involves two independent quantities \( \theta^t \) and \( \delta_i \), and can typically be controlled by standard concentration arguments. To bound the \( \ell_2 \)-Lipschitz term above, we invoke the restricted \( \ell_2 \)-contraction condition (8) under the \( \ell_\infty \) induction hypothesis, thus obtaining \( \| F(\theta^t; \delta) - F(\theta^{t,i}; \delta) \|_2 \leq \alpha \| \theta^t - \theta^{t,i} \|_2 \). The above bounds combined with the proximity hypothesis on \( \| \theta^t - \theta^{t,i} \|_2 \) yield an (often contracting) upper bound on \( \| \theta^{t+1} - \theta^{t+1,i} \|_2 \), so the proximity bound holds for the next iteration.

Turning to the coordinates of the original iterate \( \theta^{t+1} \), we use separability (9) to compute
\[
| \theta^{t+1}_i - F_i(\theta^{t+1}; \delta_i) | \\
= | F_i(\theta^{t+1}; \delta_i) - F_i(\theta^{t+1,i}; \delta_i) + F_i(\theta^{t+1,i}; \delta_i) | \\
\leq | F_i(\theta^t; \delta) - F_i(\theta^{t,i}; \delta) | + | F_i(\theta^{t,i}; \delta_i) | \\
\leq \beta \| \theta^t - \theta^{t,i} \|_2 + | F_i(\theta^{t,i}; \delta_i) |,
\]
where the last step follows from the Lipschitz conditions (10). The first term \( \| \theta^t - \theta^{t,i} \|_2 \) above is bounded under the proximity hypothesis; the second term \( F_i(\theta^{t,i}; \delta_i) \) again involves two independent quantities and can be handled as before. Putting together, we have established an upper bound on each coordinate \( | \theta^{t+1}_i | \) of the original iterate (4), as desired.

To sum up, by using the above arguments, we reduce the challenging problem of controlling the individual coordinates of \( \theta^t \) to two easier tasks:
1) Control the quantity $\mathcal{F}_t(\theta; \delta_t)$ when $\theta$ and $\delta_t$ are independent. This quantity measures the sensitivity of $\mathcal{F}$ under an independent random perturbation to one coordinate of the data vector $\delta$.

2) Control the quantity $F(\theta; \delta) - F(\theta'; \delta)$ in various norms when $\theta, \theta'$ are small entrywise. This quantity measures the sensitivity of $F$ with respect to the iterate $\theta$. This task can be accomplished using the (restricted) Lipschitz properties (8) and (10) of $F$, which can often be established even when $\ell_2$- or $\ell_\infty$-contraction fails to hold uniformly.

C. Analysis of PGD and NNM Using Leave-One-Out

To study PGD and NNM, we instantiate the abstract procedure (4) as in equations (6) and (7), respectively. The analysis of these two procedures follows the general strategy outlined above, though the proof involves several technical complications:

- In the matrix completion setting, the iterates $\theta^t$ and the random data $\delta$ are both matrices, so the separability property (9), and accordingly the leave-one-out sequences $\{\theta^{t,i}\}$, take a more complicated form involving the rows and columns of these matrices.

- Consequently, in addition to bounding the entrywise norm of the iterates, we often need to bound their row-wise norm as well. In the case of PGD, we in fact do so for the eigenvectors of the iterates.

- We need to establish the Lipschitz properties (8) and (10) with small enough $\alpha$ and $\beta$, which requires the use of appropriate concentration and matrix perturbation bounds.

In addition, PGD involves an unbounded number of iterations, yet the high probability bounds obtained by leave-one-out are only valid for poly($d$) iterations, due to the use of union bounds. Fortunately, after this many iterations, PGD already enters a small neighborhood of the fixed point $\theta^*$, within which one can establish certain uniform concentration bounds that are valid for an arbitrary number of iterations.

For NNM, a direct application of leave-one-out as in the last subsection would establish a sample complexity result of the form $p \gtrsim \text{poly}(\mu r) \log d$. This bound has the right dependence on $d$, but it is vastly sub-optimal in terms $\mu$ and $r$. To remove these superfluous $\mu r$ factors and establish a tighter bound $p \gtrsim \mu r \log(\mu r) \log d$ in Theorem 2, we take a hybrid approach that “warm-starts” the iterative procedure by running a small number (in particular, $O(\log \mu r)$) of iterations with sample splitting. As mentioned in Section III-C, this approach gives the best sample complexity upper bound to date, but we have not been able to remove the extra $\log \mu r$ factor that is absent in the information-theoretic lower bound (cf. Table 1).

V. PROOF OF THEOREM 1

In this section, we prove our convergence guarantee for PGD in Theorem 1. The proof makes use of the auxiliary lemmas given in Appendix C. Let $\sigma_1$ denote the largest eigenvalue of $M^*$ and $\sigma_r$ the smallest nonzero eigenvalue. Recall that $\kappa := \frac{\sigma_2}{\sigma_r}$ is the condition number of $M^*$.

Proof outline: After recording some preliminary steps in Section V-A, we present the main steps of the proof in two parts, following the strategy given in Section IV. Let $t_0 = 5 \log_2 d + 12$.

• Part 1: We prove that w.h.p., for all $t = 1, \ldots, t_0$, there holds the infinity norm bound

$$\|M^t - M^*\|_\infty \leq \frac{1}{d} \left( \frac{1}{2} \right)^t \sigma_1.$$  \hspace{1cm} (14)

This bound is proved using an induction argument and the leave-one-out technique. The proof proceeds in two sub-steps.

- Part 1(a): We first establish the base case, that is, the bound (14) holds for $t = 1$. This step is itself non-trivial, and is presented in Section V-B.

- Part 1(b): We next perform the induction step, in which we assume that the induction hypothesis holds for $t$ and show that it is also valid for $t + 1$. This step is presented in Section V-C.

• Part 2: We show in Section V-D that w.h.p., for all $t \geq t_0$, there holds the Frobenius norm bound

$$\|M^t - M^*\|_F \leq \left( \frac{1}{2} \right)^{t-t_0} \|M^{t_0} - M^*\|_F.$$  \hspace{1cm} (15)

thereby controlling the error of PGD for an infinite number of iterations.

Combining the above two bounds (14) and (15), we conclude that w.h.p. $\|M^t - M^*\|_{\infty} \leq \left( \frac{1}{2} \right)^t \sigma_1$ for all $t \geq 1$, which establishes the first part of Theorem 1. The second part of the theorem is exactly the bound (14).

A. Preliminaries

Throughout the proof, we use $c$ and $C$ to denote sufficiently large universal constants that may differ from line to line. Recall the assumption $p \geq C^2 \mu r \mu r \log d$.

Recall that a constant step size $\eta_t = \frac{1}{t}$ is used in the PGD iteration (2); accordingly, we define the operator $\mathcal{H}_t := I - \frac{1}{\eta_t^2} \Pi_\Omega$. With this notation, the PGD iteration (2) can be written compactly as

$$M^{t+1} = \mathcal{P}_r(M^t + \mathcal{H}_t(M^t - M^*)),$$  \hspace{1cm} (16)

We write the rank-$r$ eigenvalue decompositions of $M^t$ and $M^*$ as $M^t = (F^t \Lambda^t) \otimes F^t$ and $M^* = (F^* \Lambda^*) \otimes F^*$ respectively. Both $F^t$ and $F^*$ are in $\mathbb{R}^{d \times r}$ and have orthonormal columns. The matrices $\Lambda^t$ and $\Lambda^*$ are in $\mathbb{R}^{r \times r}$ and are diagonal matrices. Note that $\Lambda = \text{diag}(\sigma_1, \ldots, \sigma_r)$.

For the purpose of analysis, we consider a leave-one-out version of PGD. Let $\mathcal{H}^{(m)}_t$ be the operator derived from $\mathcal{H}_t$ with the $m$-th row and column observed; that is,

$$\left( \mathcal{H}_t^{(m)} Z \right)_{ij} = \begin{cases} (1 - \frac{1}{p} \delta_{ij}) Z_{ij}, & i \neq m, j \neq m, \\ 0, & i = m \text{ or } j = m. \end{cases}$$
For each $m \in [d]$, define the following leave-one-out sequence:

$$M^0,m = 0;$$
$$M^{t+1,m} = \mathcal{P}_r \left[ M^* + \mathcal{H}_t^{(-m)} \left( M^{t,m} - M^* \right) \right], \quad (17)$$

for $t = 0, 1, 2, \ldots$. We write the rank-$r$ eigen decomposition of $M^{t,m}$ as $M^{t,m} = (F^{t,m} \Lambda^{t,m} F^{t,m})$. Here $F^{t,m} \in \mathbb{R}^{d \times r}$ has orthonormal columns and $\Lambda^{t,m} \in \mathbb{R}^{r \times r}$ is diagonal. By construction, the sequence $(F^{t,m}, \Lambda^{t,m})_{t=0,1,...}$ is independent of $\delta_m$ and $\delta_m$, i.e., the $m$-th row and column of $\Omega$. It is convenient to let $m = 0$ correspond the original PGD iteration, e.g., $F^{t,0} = F^t$ and $\mathcal{H}_t^{(-0)} = \mathcal{H}_\Omega$.

A few notations are needed for measuring the distance between the column spaces of two matrices $F^t, F^t \in \mathbb{R}^{d \times r}$. For each $m \in \{0\} \cup [d]$, set $H^{t,m} := (F^t)^T F^{t,m}$ and its rank-$r$ SVD be $H^{t,m} = U \Sigma V^T$. It is known that the orthonormal matrix $G^{t,m} := U \Sigma V^T$ is the minimizer of the problem $\min_{Q \in \mathbb{R}^{d \times r}} \|Q^t F^t - F^t Q\|_F$ [16, Lemma 6]. Similarly, for each pair of leave-one-out iterates $F^{t,i}$ and $F^{t,m}$, we set $H^{t,i,m} := (F^{t,i})^T F^{t,i,m}$ and define the orthogonal matrix $G^{t,i,m}$ accordingly. We again use the convention that $H^t = H_t^0, G^t = G_t^0$.

The following notations are defined for each step $t = 0, 1, \ldots$. Denote the residual matrix of the original PGD (16) by $E^t \equiv E^{t,0} := \mathcal{H}_t^0 (M^t - M^*)$, and the residual matrix of the $m$-th leave-one-out sequence (17) by $E^{t,m} := \mathcal{H}_t^{(-m)} (M^{t,m} - M^*)$. For each $i, m \in \{0\} \cup [d]$, denote the difference of the iterates from the true $E^* := 0$ by $\Delta^t,i,m := E^{t,i} - F^t \Lambda^{t,i,m} G^{t,i,m}$. The distance between a pair of iterates by $D^t,i,m := \|D^{t,i} - F^t \Lambda^{t,i,m} G^{t,i,m}\|_F$ and the non-commutativity measure by $S^t,i,m := \|D^{t,i} - G^{t,i,m}\|_F$. Finally, define the shorthands $\Delta^{t,\infty} := \arg\max_{\Delta^t,i,m} \|\Delta^t,i,m\|_2, D^{t,\infty} := \arg\max_{D^t,i,m} \|D^t,i,m\|_F$, $E^{t,\infty} := \arg\max_{E^t,i,m} \|E^t,i,m\|_F$ and $S^{t,\infty} := \arg\max_{S^t,i,m} \|S^t,i,m\|_F$.

Unequal eigenvalues and non-commutativity: Since the eigenvalues of $M^{t,m}$ are unequal in general, the proof is complicated by the fact that the diagonal eigenvalue matrix $\Lambda^{t,m}$ does not commute with the matrices $H^{t,m}$ and $G^{t,i,m}$. We record two technical lemmas for handling this issue. The first lemma is proved in Section A-A using techniques from [1], [15].

**Lemma 1:** Suppose that $W := M^* + E$, where $E \in \mathcal{S}^{d \times d}$. Denote the eigenvalue decomposition of $M^*$ as $M^* = (F^* \Lambda^*) \in \mathbb{R}^{d \times r}$ having orthonormal columns and $\Lambda^* \in \mathbb{R}^{d \times d}$, where $F^* \in \mathbb{R}^{d \times d}$ is the matrix whose columns are the top-$r$ orthonormal eigenvectors of $W$. Let the SVD of the matrix $H := (F^*)^T F$ be $H = U \Sigma V^T$. Let $G := UV^T$. If $\|E\|_F < \frac{1}{2} \sigma_r$, then we have the bounds

$$\| \Lambda G - GA \|_F \leq \left( 2 + 2 \sigma_1 \frac{1}{\sigma_r - \|E\|_F} \right) \|E\|_F,$$

$$\| \Lambda H - GA \|_F \leq \left( 2 + \sigma_1 \frac{1}{\sigma_r - \|E\|_F} \right) \|E\|_F,$$

$$\| \Lambda G - HA \|_F \leq \left( 2 + \sigma_1 \frac{1}{\sigma_r - \|E\|_F} \right) \|E\|_F.$$
2) \( \ell_{2,\infty} \): Proximity and Non-Commutativity Bounds: We claim that the following three intermediate inequalities hold \( w.h.p. \) for all \( i,m \):

\[
\|\Delta_{m}^{1,i}\|_2 \leq \frac{1}{2} \sqrt{\frac{\mu r}{d}} + \frac{2}{15} \|\Delta_{1,m}\|_{2,\infty}
\]

\[
|D_{1,m}^{1,i}| \leq \frac{1}{8} \|\Delta_{1,m}\|_{2,\infty} + \frac{1}{2} \sqrt{\frac{\mu r}{d}}.
\]

\[
|\left[(\mathcal{H}_{\Omega}^{(-0)} - \mathcal{H}_{\Omega}^{(-m)})M^*\right]F_{1,m}\|_2 \leq \frac{\sigma_r}{2^2} \|\Delta_{1,m}\|_{2,\infty} + \frac{\sigma_r}{32} \left(\frac{1}{2}\right) \sqrt{\frac{\mu r}{d}}.
\]

We postpone the proofs of (20a), (20b) and (20c) to Sections V-B3, V-B4 and V-B5, respectively. With these three inequalities, the last three bounds in the induction hypothesis follow easily, as we show below.

First, plugging (20b) into (20a), we obtain that \( w.h.p. \)

\[
\|\Delta_{m}^{1,i}\|_2 \leq \frac{1}{2} \sqrt{\frac{\mu r}{d}} + \frac{1}{6} \|\Delta_{1,m}\|_{2,\infty}.
\]

The maximum of the last LHS over \( i \in \{0\} \cup [d] \) and \( m \in [d] \) is \( \|\Delta_{1,m}\|_{2,\infty} \). Taking this maximum and rearranging terms, we obtain the desired \( \ell_{2,\infty} \) bound (18b) in the induction hypothesis for \( t = 1 \).

Next, plugging the \( \ell_{2,\infty} \) bound (18b) we just proved into (20b), we obtain \( w.h.p. \)

\[
|D_{1,m}^{1,i}| \leq \frac{1}{8} \|\Delta_{1,m}\|_{2,\infty} + \frac{1}{2} \sqrt{\frac{\mu r}{d}}.
\]

The maximum of the last LHS over \( i \) and \( m \) is \( \|D_{1,m}\|_2 \). Taking this maximum proves the desired proximity bound (18c) in the induction hypothesis for \( t = 1 \).

Finally, we have \( w.h.p. \)

\[
\|S_{1,m}^{1,0}\|_2 = \|S_{1,m}^{1,0}\|_2 \\
\leq 4 \cdot \left(\frac{1}{2}\right) \sqrt{\frac{\mu r}{d}}.
\]

where step (a) follows from Lemma 2 whose premise is satisfied because of the bound (19), and step (b) follows from the inequalities (20c) and (18b). The maximum of the last LHS over \( m \) is \( \|S_{1,m}^{1,0}\|_2 \). Taking this maximum proves the desired non-commutativity bound (18d) in the induction hypothesis for \( t = 1 \). We have thus completed the proof of the base case of the hypothesis.

3) Proof of Intermediate Inequality (20a): We focus on the \( m \)-th row of the difference matrix \( \Delta_{m}^{1,i} := F_{m}^{1,i} - F_{m}^{1,i} \). Using the facts \( F_{m}^{1,i} \Lambda_{i}^{1,i} = (M^* + \mathcal{E}_{m}^{0,i})F_{m}^{1,i} \), we have the expression

\[
\Delta_{m}^{1,i} = F_{m}^{1,i} - F_{m}^{1,i} \cdot G_{m}^{1,i}.
\]

Rearranging the last RHS yields

\[
\Delta_{m}^{1,i} = e_{m}^{T} F_{m}^{1,i} \Lambda_{m}^{1,i} \left( (F_{m}^{1,i})^{T} (\Lambda_{m}^{1,i})^{-1} - (\Lambda_{m}^{1,i})^{-1} \right)
\]

\[
eq e_{m}^{T} F_{m}^{1,i} \Lambda_{m}^{1,i} \left( (F_{m}^{1,i})^{T} (\Lambda_{m}^{1,i})^{-1} - (\Lambda_{m}^{1,i})^{-1} \right)
\]

\[
= e_{m}^{T} F_{m}^{1,i} \Lambda_{m}^{1,i} \left( (F_{m}^{1,i})^{T} (\Lambda_{m}^{1,i})^{-1} - (\Lambda_{m}^{1,i})^{-1} \right)
\]

\[
\leq \frac{\sqrt{\frac{\mu r}{d}}}{\sigma_r} \frac{1}{C} + \frac{1}{C} \sqrt{\frac{\mu r}{d}}.
\]

Below we bound each of the three terms \( T_{1}, T_{2} \) and \( T_{3} \).

a) Bounding \( T_{1} \): Note that \( T_{1} = e_{m}^{T} F_{m}^{*} R \), where

\[
R := \Lambda_{m}^{1,i} \left( (F_{m}^{1,i})^{T} (\Lambda_{m}^{1,i})^{-1} - (\Lambda_{m}^{1,i})^{-1} \right)
\]

Recall that by definition \( H_{1,1} := (F_{m}^{1,i})^{T} (\Lambda_{m}^{1,i})^{-1} - (\Lambda_{m}^{1,i})^{-1} \) has SVD \( \hat{U} \Sigma \hat{V}^{T} \) and \( G_{1,1} := \hat{U} \Sigma \). Therefore, Lemma 1 is applicable, which gives that \( w.h.p. \)

\[
\|R\|_{op} \leq \left(2 + 2 \sigma_r \sqrt{\frac{1}{\sigma_r - \|E_{m}^{0,i}\|_2}}\right) \|E_{m}^{0,i}\|_2 \cdot \frac{1}{\sigma_r} \leq \frac{1}{C} \left(\frac{1}{\sigma_r} \sqrt{\frac{\mu r}{d}} \right),
\]

where the last step is due to the bound (19) on \( \|E_{m}^{0,i}\|_2 \). Thus \( T_{1} \) can be bounded \( w.h.p. \) as

\[
\|T_{1}\|_2 \leq \|e_{m}^{T} F_{m}^{*} R\|_2 \leq \|e_{m}^{T} F_{m}^{*}\|_2 \|R\|_{op} \leq \frac{1}{C} \left(\frac{1}{\sigma_r} \sqrt{\frac{\mu r}{d}} \right),
\]

b) Bounding \( T_{2} \): Using Weyl’s inequality \( \|\Lambda_{m}^{1,i} - \Lambda_{m}^{1,i}\|_{\infty} \leq \|E_{m}^{0,i}\|_2 \) and the bound (19) on \( \|E_{m}^{0,i}\|_2 \), we have \( w.h.p. \)

\[
\|\Lambda_{m}^{1,i} - \Lambda_{m}^{1,i}\|_{\infty} \leq \left(\frac{1}{C} \sqrt{\frac{\mu r}{d}} \right),
\]

It then follows that \( w.h.p. \)

\[
\|T_{2}\|_2 \leq \|e_{m}^{T} F_{m}^{0,i} F_{m}^{1,i}\|_2 \|R\|_{op} \|\Lambda_{m}^{1,i} - \Lambda_{m}^{1,i}\|_{\infty} \leq \left(\frac{1}{C} \sqrt{\frac{\mu r}{d}} \right) \|e_{m}^{T} F_{m}^{0,i} F_{m}^{1,i}\|_2 \|R\|_{op} \|\Lambda_{m}^{1,i} - \Lambda_{m}^{1,i}\|_{\infty} \leq \left(\frac{1}{C} \sqrt{\frac{\mu r}{d}} \right),
\]

Combining the above bounds for \( T_{1}, T_{2}, \) and \( T_{3} \), we obtain that \( w.h.p. \)

\[
\|\Delta_{m}^{1,i}\|_2 \leq \frac{1}{C} \frac{1}{\sigma_r} \sqrt{\frac{\mu r}{d}} \left(\frac{1}{2}\right) + \frac{1}{C} \left(\frac{1}{\sigma_r} \sqrt{\frac{\mu r}{d}} \right).
\]
To proceed, we further control \( \|e_m^T E^{0,i} F^{1,i}\|_2 \). If \( i = m \), then \( e_m^T E^{0,i} = 0 \) and we are done. Below we assume \( i \neq m \). Let \( Q \in \mathbb{R}^r \times r \) be an orthogonal matrix whose value is determined later. We write

\[
e_m^T E^{0,i} F^{1,i} = e_m^T E^{0,i} F^{1,m} Q + e_m^T E^{0,i} (F^{1,i} - F^{1,m} Q)
\]

\[
= e_m^T H_1^{(-i)} (-M^*) F^{1,m} Q
\]

\[
+ e_m^T E^{0,i} (F^{1,i} - F^{1,m} Q).
\]

(24)

We bound \( \hat{T}_1 \) and \( \hat{T}_2 \) below.

d) Bounding \( \hat{T}_1 \): Denote the shorthand \( R_{mj} \) in this paragraph as

\[
R_{mj} := \sum_{k=1}^{d} \left( 1 - \frac{1}{p} \delta_{mk} \right) (-M^*) F^{1,m}_{kj}.
\]

We have the bound

\[
\|\hat{T}_1\|_2 = \|e_m^T H_1^{(-i)} (-M^*) F^{1,m}\|_2
\]

\[
\leq \sqrt{r} \max_{j \in [r]} |R_{mj}|.
\]

(25)

Note that \( F^{1,m} \) is independent of \( \{\delta_{mk}, k \in [d]\} \) by construction. Therefore, Bernstein’s inequality (Lemma 10) ensures that for each \( j \in [r] \), with probability at least \( 1 - d^{-12} \), there holds the inequality

\[
|R_{mj}| = \sum_{k=1}^{d} \left( 1 - \frac{1}{p} \delta_{mk} \right) (-M^*) F^{1,m}_{kj}
\]

\[
\leq \frac{C \log d}{p} \|F^{1,m}\|_{2,\infty} \|M^*_m\|_2
\]

\[
+ \frac{C \log d}{p} \|F^{1,m}\|_{2,\infty} \|M^*\|_{\infty}.
\]

(26)

For the first term \( \hat{T}_{1a} \), we have

\[
\hat{T}_{1a}^{(a)} \leq \sqrt{\frac{C \log d}{p} \|\Delta^{1,m}\|_{2,\infty} + \sqrt{\frac{\mu}{d}}} \sigma_1 \sqrt{\frac{\mu}{d}}
\]

\[
\leq \sigma_r \left[ \frac{1}{\sqrt{16\sqrt{r}}} \|\Delta^{1,m}\|_{2,\infty} + \frac{1}{32 \sqrt{r} 2^{16 \kappa_2 \mu r}} \frac{9}{\mu} \right].
\]

where step (a) is due to the facts that \( \|F^{1,m}\|_{2,\infty} \leq \|\Delta^{1,m}\|_{2,\infty} + \|F^*\|_{2,\infty} \), that \( \|M^*_m\|_2 \leq \|F^*_m\|_2 \), \( \|F^*_m\|_2 \leq \|F^*\|_2 \leq \sqrt{\frac{\mu}{d}} \), and step (b) is due to the assumption \( p \geq \frac{e^{3 \log d} \mu r^4}{d} \). For the second term \( \hat{T}_{1b} \), we follow a similar argument as above to obtain that w.h.p.

\[
\hat{T}_{1b} \leq \sigma_r \left[ \frac{1}{\sqrt{16\sqrt{r}}} \|\Delta^{1,m}\|_{2,\infty} + \frac{1}{32 \sqrt{r} 2^{16 \kappa_2 \mu r}} \frac{9}{\mu} \right].
\]

We plug the bounds for \( \hat{T}_{1a} \) and \( \hat{T}_{1b} \) into the inequality (26), and take a union bound over all \( j \in [r] \) and \( m \in [d] \). Combining with the inequality (25), we obtain that w.h.p.

\[
\|\hat{T}_1\|_2 \leq \frac{\sigma_r}{8} \|\Delta^{1,m}\|_{2,\infty} + \frac{\sigma_r}{2^{20 \kappa_2 \mu^2}} \sqrt{\frac{\mu}{d}}.
\]

(27)

e) Bounding \( \hat{T}_2 \): Recalling the definition of \( G^{1,i,m} \) in Section V-A, we choose \( Q \equiv \tilde{G}^{1,i,m} \) so that \( F^{1,i} - F^{1,m} Q = D^{1,i,m} \). We thus have \( \hat{T}_2 = e_m^T H_{1i}^{(-i)} (M^*) D^{1,i,m} \) by definition of \( \hat{T}_2 \). It follows that w.h.p.

\[
\|\hat{T}_2\|_2 \leq \|H_{1i}^{(-i)} (M^*)\|_\infty \|D^{1,i,m}\|_f
\]

\[
\leq \frac{2c}{\sqrt{d \log d}} \|M^*\|_{\infty} \|D^{1,i,m}\|_f,
\]

where step (a) is due to Cauchy-Schwarz, and step (b) is due to Lemma 21. Using the assumptions that \( \|M^*\|_{\infty} \leq \frac{2c}{\sqrt{d \log d}} \) as \( M^* \) is \( \mu \)-incoherent and \( p \geq \frac{e^{3 \log d} \mu r^4}{d} \), we obtain that w.h.p.

\[
\|\hat{T}_2\|_2 \leq \frac{\sigma_r}{8} \|D^{1,i,m}\|_f.
\]

(28)

Plugging the bounds (27) and (28) for \( \hat{T}_1 \) and \( \hat{T}_2 \) into (24), we get that w.h.p.

\[
\|e_m^T E^{0,i} F^{1,i}\|_2 \leq \sigma_r \left[ \frac{1}{8} \|\Delta^{1,i,m}\|_{2,\infty} + \frac{1}{8} \|D^{1,i,m}\|_f,
\]

\[
+ \frac{1}{8} \frac{1}{2^{16 \kappa_2 \mu r}} \sqrt{\frac{\mu r}{d}}.
\]

Further plugging this bound into (23), we obtain the first intermediate inequality (20a).

4) Proof of Intermediate Inequality (20b): To bound \( D^{1,i,m} \), we begin by recalling that by definition,

\[
M^{1-j} = (F^{1,i} A^{1,j}) \otimes F^{1-j} = \mathcal{P}_r [M^* + \mathcal{H}_{1i}^{(-i)} (-M^*)],
\]

for \( j = i \) and \( m \). Weyl’s inequality ensures that the eigen gap \( \delta \) between the \( r \)-th and \( (r+1) \)-th eigenvalues of the matrix \( M^* + \mathcal{H}_{1i}^{(-i)} (-M^*) \) is at least \( \delta \geq \sigma_r - 2 \|E^{0,i}\|_\infty \geq \frac{15}{8} \sigma_r \) w.h.p., where the last inequality follows from the bound (19) on \( \|E^{0,i}\|_\infty \). Let \( W^{1,i} := (\mathcal{H}_{1i}^{(-i)} - \mathcal{H}_{1i}^{(-m)})(M^*) \), for which we have w.h.p. \( \|W^{1,i}\|_\infty \leq \|E^{0,i}\|_\infty + \|E^{0,m}\|_\infty \leq \frac{15}{8} \sigma_r \) again thanks to (19). Recalling the definition of \( D^{1,i,m} \) and applying the Davis-Kahan Theorem (Lemma 14), we obtain that w.h.p.

\[
\|D^{1,i,m}\|_f \leq \sqrt{2} \|W^{1,i} F^{1,i,m}\|_f - \|W^{1,i} F^{1,i,m}\|_f,
\]

\[
\leq \frac{1}{7/8} \sigma_r,
\]

(29)

To proceed, we consider two cases: \( i = 0 \) and \( i \neq 0 \).

a) The case \( i = 0 \): In this case, the quantity \( \|W^{0,i} F^{1,i,m}\|_f \) is bounded in the inequality (20c) to be proved below in Section V-B5. Plugging (20c) into the inequality (29), we have w.h.p.

\[
\|D^{1,0,m}\|_f \leq \frac{1}{16} \|\Delta^{1,m}\|_{2,\infty} + \frac{1}{2^{20 \kappa_2 \mu^2}} \frac{\mu r}{d}.
\]

(30)

b) The case \( i \neq 0 \): For any orthonormal matrices \( Q_1, Q_2 \in \mathbb{R}^{r \times r} \), the optimality of \( D^{1,i,m} \) implies that

\[
\|D^{1,i,m}\|_f \leq \|F^{1,i} - F^{1,m} Q_1\|_f
\]

\[
\leq \|F^{1,i} - F^{1,0} Q_2\|_f
\]

\[
+ \|F^{1,0} Q_2 Q_1^{-1} - F^{1,m} Q_1\|_f.
\]

(31)
Choosing $Q_2$ to minimize $\|F^{1,i} - F^{1,0}Q_2\|_\|$, and then $Q_1$ to minimize $\|F^{1,0}Q_2Q_1^{-1} - F^{1,m}\|_\|$, we have $\|F^{1,i} - F^{1,0}Q_2\|_\| = \|D_1^{1,0,i}\|_\|$, and $\|F^{1,0}Q_2Q_1^{-1} - F^{1,m}\|_\| = \|D_1^{1,0,m}\|_\|$. It then follows from (31) that w.h.p.
\[
\|D_1^{1,i,m}\|_\| \leq \|D_1^{1,0,i}\|_\| + \|D_1^{1,0,m}\|_\|.
\]
where step (a) is due to the inequality (30) proved above.

In view of the bounds (30) and (32) for two cases, we have established the second intermediate inequality (20b).

5) Proof of Intermediate Inequality (20c): We begin by observing that the operator $\mathcal{H}_\Omega^{(-i)} - \mathcal{H}_\Omega^{(-m)}$ is supported only on the $m$-th row and column. This observation gives
\[
\|[(\mathcal{H}_\Omega^{(-i)} - \mathcal{H}_\Omega^{(-m)})(M^*)]\|_\|
= \sum_{j \leq r} \sum_{k \leq d, k \neq m} \left(1 - \frac{\delta_{mk}}{p}\right)M_{mk}F^{1,m}_{mj}\]
+ \sum_{j \leq r} \sum_{k \leq d} \left(1 - \frac{\delta_{mk}}{p}\right)M_{mk}F^{1,m}_{kj}\].

Decomposing the matrix $(\mathcal{H}_\Omega^{(-i)} - \mathcal{H}_\Omega^{(-m)})(M^*)$ into two terms accordingly, we have

\[
\|[(\mathcal{H}_\Omega^{(-i)} - \mathcal{H}_\Omega^{(-m)})(M^*)]\|_\|
\leq \sqrt{\sum_{j \leq r} \sum_{k \leq d, k \neq m} \left(1 - \frac{\delta_{mk}}{p}\right)^2 M_{km}F^{1,m}_{mj}}
+ \sqrt{\sum_{j \leq r} \sum_{k \leq d} \left(1 - \frac{\delta_{mk}}{p}\right)^2 M_{mk}F^{1,m}_{kj}}\].

For the first term $B_1$, note that
\[
\sqrt{\sum_{k \leq d, k \neq m} \left(1 - \frac{\delta_{mk}}{p}\right)^2 M_{km}}^2 = \|\mathcal{H}_\Omega(M^*)e_m\|_\|
\leq (1 + \|F^{1,m}\|_\|).\|\mathcal{H}_\Omega(M^*)\|_\|
\]

By the union bound, these inequalities hold w.h.p.

\[
\|M^t - M^*\|_\|
\leq \kappa \sigma_1 \left(\frac{1}{2}\right)^{t} \frac{1}{d}.\]

Using Lemma 22 in the following step (a), we obtain that w.h.p.

\[
\|E^{t,i}\|_\|
\leq \frac{r \log d}{\sigma_1} \left(\frac{1}{2}\right)^{t} \frac{1}{d}.\]

We have proved that the operator norm bound (18a) holds for the next iteration.

2) $\ell_2, \infty$ Norm Bound (18b): We focus on the $m$-th row of the difference matrix $\Delta^{t+1, i} := F^{t+1, i} - F^{t+1, i}G^{t+1, i}$. By definition, the matrices $\Lambda^{t+1, i}$ and $F^{t+1, i}$ correspond to the top $r$ eigenvalues and eigenvectors of $M^* + E^{t, i}$, whence

\[
(M^* + E^{t, i})F^{t+1, i} = F^{t+1, i}G^{t+1, i}.
\]

Recalling $M^* = (F^* \Lambda^*) \odot F^*$, we have

\[
\Delta^{t+1, i} = F^{t+1, i} - F^* G^{t+1, i}.
\]

C. Part 1(b): Induction Step

Suppose that the induction hypothesis (18b) holds for the $t$-th iteration, where $t \geq 1$. We shall prove that it also holds for the $(t + 1)$-th iteration. Again, in the proof we shall show that various inequalities hold w.h.p. for each fixed indices $i \in \{0, 1, \ldots, d\}$ and $m \in [d]$. By the union bound, these inequalities hold simultaneously for all indices w.h.p.

1) Operator Norm Bound (18a): To bound $E^{t,i} := \mathcal{H}_\Omega^{(-i)}(M^{t,i} - M^*)$, we shall apply Lemma 22, which requires an $\ell_\infty$ norm bound on $M^{t,i} - M^*$. To this end, let us record several useful bounds. The $\ell_2, \infty$ bound (18b) in the induction hypothesis implies that $\|\Delta^{t,i}\|_\\infty \leq \frac{1}{\sqrt{\mu r}} \left(\frac{1}{2}\right)^{t}$. w.h.p.; consequently, $\|E^{t,i}\|_\\infty \leq 2 \sqrt{\mu r}$. Wely’s inequality together with the operator norm bound (18a) in the induction hypothesis implies that $\|\Lambda^{t,i} - \Lambda^*\|_\\infty \leq \|E^{t-1,i}\|_\\infty \leq \frac{1}{\sqrt{\mu r}} \left(\frac{1}{2}\right)^{t} \sigma_r$ w.h.p.; consequently, $\|\Lambda^{t,i}\|_\\infty \leq \sigma_1$ w.h.p. With these bounds, we may apply Lemma 19 to obtain that w.h.p.

\[
\|M^t - M^*\|_\\infty \leq \kappa \sigma_1 \left(\frac{1}{2}\right)^{t} \frac{1}{d}.\]

We can bound the $\ell_2$ norms of $T_1, T_2, T_3$ by following the same arguments used in Section V-B3 for bounding $T_1, T_2, T_3$.
therein. Doing so yields that w.h.p.
\[
\|\Delta_{t+1}^{i,j}\|_2 \leq \frac{1}{C} \frac{1}{2^{6\mu r + 2}} \frac{\sqrt{\mu r}}{d} \left( \frac{1}{2} \right)^{t+1} C \|1 + \frac{1}{C^2} \|_{\infty} \|\epsilon_{t+1, i}^{T} F_{t+1, i}\|_2.
\]
To proceed, we control \(\epsilon_{t+1, i}^{T} F_{t+1, i}\) and thereby establish the \(\epsilon_{t+1}^{\infty}\) error bound (18b) for \(t+1\). Note that if \(i = m\), then \(\epsilon_{t+1, i}^{T} F_{t+1, i} = 0\) by construction and we are done. In the following, we assume \(i \neq m\).

3) Bounding \(\|\epsilon_{t+1, i}^{T} F_{t+1, i}\|_2\) and Establishing the \(\epsilon_{t+1}^{\infty}\) Bound: Let \(Q := (G^{t+1, m})^T G^{t+1, m} \in \mathbb{R}^{r \times r}\), which satisfies \(QQ^T = I\). The reason for this choice shall become clear later.

We use the decomposition
\[
\epsilon_{t+1, i}^{T} F_{t+1, i} = \epsilon_{t+1, m}^{T} F_{t+1, i} Q + \epsilon_{t+1, i}^{T} (F_{t+1, i} - F_{t+1, i} Q) = \epsilon_{t+1, m}^{T} F_{t+1, i} Q + \epsilon_{t+1, i}^{T} (F_{t+1, i} - F_{t+1, i} Q)
\]

\[= \epsilon_{t+1, i}^{T} (M_{t+1}^{i} - M^*) F_{t+1, i} Q + \epsilon_{t+1, i}^{T} (F_{t+1, i} - F_{t+1, i} Q)
\]

\[= \epsilon_{t+1, i}^{T} (M_{t+1}^{i} - M^*) F_{t+1, i} Q + \epsilon_{t+1, i}^{T} (F_{t+1, i} - F_{t+1, i} Q)
\]

Below we control each of the terms \(\hat{T}_1, \hat{T}_2, \hat{T}_3\).

a) Controlling \(\hat{T}_1\): Denote the shorthand \(R_{t+1}^i\) in this paragraph as
\[
R_{t+1}^i = \sum_{k=1}^{d} \left( 1 - \frac{1}{p} \delta_{mk} \right) (M_{mk} - M_{mk}^*) F_{t+1, i}.
\]
We begin with the inequality
\[
\|\hat{T}_1\|_2 = \|\epsilon_{t+1, i}^{T} (M_{t+1}^{i} - M^*) F_{t+1, i}\|_2 \leq \sqrt{\max_{j \leq r} \|R_{t+1}^i\|_2}.
\]
Note that \(F_{t+1, i}\) is independent of \(\{\delta_{mk}, k \in [d]\}\) by construction. Therefore, Bernstein’s inequality (Lemma 10) ensures that for each \(j \in [r]\), with probability at least \(1 - d^{-1/2}\), there holds the inequality
\[
\|R_{t+1}^i\|_2 \leq \sqrt{\frac{\log d}{p} \|F_{t+1, i}\|_2 \|M_{t+1}^{i} - M^*\|_2}.
\]
To further bound \(\hat{T}_{1a}\) and \(\hat{T}_{1b}\), we shall apply Lemma 19 to control the \(\ell_2\) and \(\ell_\infty\) norms of the vector \(M_{t+1}^{i} - M_{t+1}^*\). To this end, we recall the bounds proved before (33): \(\|\Delta_{t+1}^{i}\|_2 \leq \frac{1}{C \cdot 2^{6\mu r + 2}} \frac{\sqrt{\mu r}}{d} \left( \frac{1}{2} \right)^{t+1}\)

\[\|\Delta_{t+1}^{i}\|_2 \leq \frac{1}{C \cdot 2^{6\mu r + 2}} \frac{\sqrt{\mu r}}{d} \left( \frac{1}{2} \right)^{t+1}\]

\[\|\Delta_{t+1}^{i}\|_2 \leq \frac{1}{C \cdot 2^{6\mu r + 2}} \frac{\sqrt{\mu r}}{d} \left( \frac{1}{2} \right)^{t+1}\]

Thus, we have
\[
\|\hat{T}_{1a}\|_2 \leq 2\sigma_1 \|\Delta_{t+1}^{i}\|_2 \leq 4\sigma_1 \|\Delta_{t+1}^{i}\|_2 \leq 4\sigma_1 \sqrt{\frac{\mu r}{d}} \|F_{t+1, i}\|_2.
\]
Plugging the above bound into (37) and (38), we conclude that w.h.p.
\[
\|\hat{T}_{1a}\|_2 \leq 2\sigma_1 \|\Delta_{t+1}^{i}\|_2 \leq 4\sigma_1 \sqrt{\frac{\mu r}{d}} \|F_{t+1, i}\|_2.
\]

b) The \(\hat{T}_2\) term: To bound the \(\hat{T}_2\) in inequality (36), we first note that the matrix \(M_{t+1}^{i} - M_{t+1}^*\) can be decomposed into three terms as
\[
M_{t+1}^{i} - M_{t+1}^* = (F_{t+1, i} M_{t+1}^{i}) \otimes (F_{t+1, i} M_{t+1}^*) \otimes (F_{t+1, i} M_{t+1}^*)
\]

\[= (F_{t+1, i} M_{t+1}^{i}) \otimes (F_{t+1, i} M_{t+1}^*) \otimes (F_{t+1, i} M_{t+1}^*)
\]

\[+ (F_{t+1, i} G_{t+1, i} M_{t+1}^{i}) \otimes (F_{t+1, i} G_{t+1, i} M_{t+1}^*) \otimes (F_{t+1, i} G_{t+1, i} M_{t+1}^*)
\]

\[+ (F_{t+1, i} G_{t+1, i} M_{t+1}^{i}) \otimes (F_{t+1, i} G_{t+1, i} M_{t+1}^*) \otimes (F_{t+1, i} G_{t+1, i} M_{t+1}^*)
\]

\[+ (F_{t+1, i} G_{t+1, i} M_{t+1}^{i}) \otimes (F_{t+1, i} G_{t+1, i} M_{t+1}^*) \otimes (F_{t+1, i} G_{t+1, i} M_{t+1}^*)
\]

\[+ (F_{t+1, i} G_{t+1, i} M_{t+1}^{i}) \otimes (F_{t+1, i} G_{t+1, i} M_{t+1}^*) \otimes (F_{t+1, i} G_{t+1, i} M_{t+1}^*)
\]

Therefore, we can bound \(\hat{T}_2\) by splitting it into three terms accordingly:
\[
\|\hat{T}_2\|_2 \leq \|\hat{T}_{2a}\|_2 \leq 2\sigma_1 \|\Delta_{t+1}^{i}\|_2 \leq 8\sigma_1 \sqrt{\frac{\mu r}{d}} \|F_{t+1, i}\|_2.
\]
The term \( B_1 \) can be written explicitly as
\[
B_1 = \sum_{l \leq r} \left( \sum_{j \leq r} (D_{t,m,l}^t)^2 \right) \leq \sum_{l \leq r} \left( \sum_{j \leq r} \|D_{t,m,l}^t\|^2 \right) \leq \sum_{l \leq r} \left( \sum_{j \leq r} \|D_{t,m,l}^t\|^2 \right),
\]
where we use Cauchy-Schwarz in step (a). It follows that
\[
B_1 \leq \sum_{l \leq r} \left( \sum_{j \leq r} \|D_{t,m,l}^t\|^2 \right) \leq \sum_{l \leq r} \left( \sum_{j \leq r} \|D_{t,m,l}^t\|^2 \right).
\]
Recalling that \( \delta_{mk} \) and \( F_{kj}^t \) are independent by construction, we apply Bernstein inequality (Lemma 10) to obtain w.h.p.
\[
\|R_{m,j}^t\|_p \leq \sum_{k \leq d} F_{k,j}^t \left( 1 - \frac{\delta_{mk}}{p} \right) F_{kl}^t.
\]
Then we have w.h.p.
\[
B_2 := \|e_m H_{t}^{(i)} ((D_{t,m}^t)^t \otimes F_{i,m}) \Delta_{t+1,m}^t\|_2
\leq \|D_{t,m}^t \Delta_{t,m}^t \|_2 \cdot \max_{l,i} \|R_{m,j}^t\|_2,\]
where step (a) follows from Lemma 2, and step (b) follows from the triangle inequality and the induction hypothesis (18a) on \( \|E_{t+1}^i\|_p \). Combining the above bound with the bound on \( \|E_{t+1}^i\|_p \), we get
\[
\|F_{i+1}^t S_{t+1,m}^t\|_p \leq \|D_{t+1,m}^t \|_2 \leq \sqrt{\frac{\mu r}{d}} \quad \text{w.h.p.,}
\]
and (46). In particular, we have
\[
B_2 \leq \|e_m H_{t}^{(i)} ((D_{t,m}^t)^t \otimes F_{i,m}) \Delta_{t+1,m}^t\|_2
\leq \|D_{t,m}^t \Delta_{t,m}^t \|_2 \cdot \max_{l,i} \|R_{m,j}^t\|_2,\]
where step (a) follows from Lemma 2, and step (b) follows from the triangle inequality and the induction hypothesis (18a) on \( \|E_{t+1}^i\|_p \). Combining the above bound with the bound on \( \|E_{t+1}^i\|_p \), we obtain that w.h.p.
\[
\|F_{i+1}^t S_{t+1,m}^t\|_p \leq \|D_{t+1,m}^t \|_2 \leq \sqrt{\frac{\mu r}{d}} \quad \text{w.h.p.,}
\]
and (46). In particular, we have
\[
B_2 \leq \|e_m H_{t}^{(i)} ((D_{t,m}^t)^t \otimes F_{i,m}) \Delta_{t+1,m}^t\|_2
\leq \|D_{t,m}^t \Delta_{t,m}^t \|_2 \cdot \max_{l,i} \|R_{m,j}^t\|_2,\]
where step (a) follows from Lemma 2, and step (b) follows from the triangle inequality and the induction hypothesis (18a) on \( \|E_{t+1}^i\|_p \). Combining the above bound with the bound on \( \|E_{t+1}^i\|_p \), we get
\[
\|F_{i+1}^t S_{t+1,m}^t\|_p \leq \|D_{t+1,m}^t \|_2 \leq \sqrt{\frac{\mu r}{d}} \quad \text{w.h.p.,}
\]
and (46). In particular, we have
\[
B_2 \leq \|e_m H_{t}^{(i)} ((D_{t,m}^t)^t \otimes F_{i,m}) \Delta_{t+1,m}^t\|_2
\leq \|D_{t,m}^t \Delta_{t,m}^t \|_2 \cdot \max_{l,i} \|R_{m,j}^t\|_2,\]
where step (a) follows from Lemma 2, and step (b) follows from the triangle inequality and the induction hypothesis (18a) on \( \|E_{t+1}^i\|_p \). Combining the above bound with the bound on \( \|E_{t+1}^i\|_p \), we get
\[
\|F_{i+1}^t S_{t+1,m}^t\|_p \leq \|D_{t+1,m}^t \|_2 \leq \sqrt{\frac{\mu r}{d}} \quad \text{w.h.p.,}
\]
and (46). In particular, we have
\[
B_2 \leq \|e_m H_{t}^{(i)} ((D_{t,m}^t)^t \otimes F_{i,m}) \Delta_{t+1,m}^t\|_2
\leq \|D_{t,m}^t \Delta_{t,m}^t \|_2 \cdot \max_{l,i} \|R_{m,j}^t\|_2,\]
where step (a) follows from Lemma 2, and step (b) follows from the triangle inequality and the induction hypothesis (18a) on \( \|E_{t+1}^i\|_p \). Combining the above bound with the bound on \( \|E_{t+1}^i\|_p \), we get
\[
\|F_{i+1}^t S_{t+1,m}^t\|_p \leq \|D_{t+1,m}^t \|_2 \leq \sqrt{\frac{\mu r}{d}} \quad \text{w.h.p.,}
\]
and (46). In particular, we have
\[
B_2 \leq \|e_m H_{t}^{(i)} ((D_{t,m}^t)^t \otimes F_{i,m}) \Delta_{t+1,m}^t\|_2
\leq \|D_{t,m}^t \Delta_{t,m}^t \|_2 \cdot \max_{l,i} \|R_{m,j}^t\|_2,\]
where step (a) follows from Lemma 2, and step (b) follows from the triangle inequality and the induction hypothesis (18a) on \( \|E_{t+1}^i\|_p \). Combining the above bound with the bound on \( \|E_{t+1}^i\|_p \), we get
\[
\|F_{i+1}^t S_{t+1,m}^t\|_p \leq \|D_{t+1,m}^t \|_2 \leq \sqrt{\frac{\mu r}{d}} \quad \text{w.h.p.,}
\]
and (46). In particular, we have
\[
B_2 \leq \|e_m H_{t}^{(i)} ((D_{t,m}^t)^t \otimes F_{i,m}) \Delta_{t+1,m}^t\|_2
\leq \|D_{t,m}^t \Delta_{t,m}^t \|_2 \cdot \max_{l,i} \|R_{m,j}^t\|_2,\]
where step (a) follows from Lemma 2, and step (b) follows from the triangle inequality and the induction hypothesis (18a) on \( \|E_{t+1}^i\|_p \). Combining the above bound with the bound on \( \|E_{t+1}^i\|_p \), we get
\[
\|F_{i+1}^t S_{t+1,m}^t\|_p \leq \|D_{t+1,m}^t \|_2 \leq \sqrt{\frac{\mu r}{d}} \quad \text{w.h.p.,}
\]
and (46). In particular, we have
\[
B_2 \leq \|e_m H_{t}^{(i)} ((D_{t,m}^t)^t \otimes F_{i,m}) \Delta_{t+1,m}^t\|_2
\leq \|D_{t,m}^t \Delta_{t,m}^t \|_2 \cdot \max_{l,i} \|R_{m,j}^t\|_2,\]
where the quantity $R_{mij}^{t,*}$ is defined as

$$R_{mij}^{t,*} := \sum_{k \leq d} D_{k}^{m,i}(1 - \frac{\delta_{mk}}{p}) F_{kl}^*.$$  

To bound $R_{mij}^{t,*}$, denoting by 1 the all one vector, we find that

$$|R_{mij}^{t,*}| \leq \sum_{k \leq d} D_{k}^{m,i}(1 - \frac{\delta_{mk}}{p}) \|F_{kl}^*\|_2.$$  

Applying Lemma 21 with $\|F\|_2, \leq \frac{\mu_r}{d}$, we have w.h.p. $\|H_{\Omega}(1 \otimes F_{l}^*)\|_p \leq c \sqrt{\frac{\log d}{p}} \sqrt{\frac{\mu_r}{d}}$. Moreover, the proximity condition (18c) in the induction hypothesis implies $\|D_{t,i}^{m,i}\|_2 \leq \frac{1}{2^\frac{\mu_r}{d}} (\frac{1}{2})^t \sqrt{\frac{\mu_r}{d}}$, and the $\ell_2, \infty$ bound (18b) in the hypotheses implies $\|F_{t}^{i} \Lambda_t^{i} G_{t,l}^{m,i}\|_2, \leq \frac{1}{2} \sigma_t \sqrt{\frac{\mu_r}{d}}$. Plugging these bounds into (52) and (53) and recalling the assumption $p \geq \frac{\log d}{d}$, we obtain that w.h.p.

$$B_1 \leq \frac{\sigma_t}{2^t} \frac{1}{2^{16 \mu_r^2}} \left( \frac{1}{2} \right)^t \frac{1}{\sqrt{d}} \sqrt{\frac{\mu_r}{d}}.$$  

By a similar argument, we can bound the term $B_2 := \|e_{m}^{T}H_{\Omega}^{(-1)}((F_{t}^{i} \Lambda_t^{i} G_{t,l}^{m,i}) \otimes D_{t,i}^{m,i}) \Delta_{t+1,m}\|_2$ as

$$B_2 \leq \|F_{t}^{i} \Lambda_t^{i} G_{t,l}^{m,i}\|_2, \|r_{l,j} \leq \frac{\Delta_{t+1,m}}{2^\frac{\mu_r}{d}} \otimes D_{t,i}^{m,i}\|_2.$$  

$$\leq \frac{1}{32} \Delta_{t+1,m} \Delta_{t+1,m}, 2^\infty.$$  

where $R_{mij}^{t,*} := e_{m}^{T}H_{\Omega}^{(-1)}(1 \otimes \Delta_{t+1,m}) D_{t,i}^{m,i}$. Combining the above bounds on $B_1$ and $B_2$, we obtain that w.h.p.,

$$\frac{\bar{T}_{3c}}{32} \bar{T}_{2c} \leq \frac{1}{2^t} \frac{1}{2^{16 \mu_r^2}} \left( \frac{1}{2} \right)^t \frac{1}{\sqrt{d}} \sqrt{\frac{\mu_r}{d}}.$$  

Plugging the above bounds (50), (51) and (54) on $\bar{T}_{2a}, \bar{T}_{2b}, \bar{T}_{2c}$ into inequality (43), we obtain that w.h.p.

$$||T_2||_2 \leq \frac{\sigma_t}{10} ||\Delta_{t+1,m}||_2, \infty + \frac{3\sigma_t}{32} \frac{1}{2^{16 \mu_r^2}} \left( \frac{1}{2} \right)^t \frac{1}{\sqrt{d}} \sqrt{\frac{\mu_r}{d}}$$.

$$+ \frac{\bar{T}_{3a}}{\bar{T}_{3b}} \frac{1}{\Delta_{t+1,m} \Delta_{t+1,m}}, 2^\infty.$$  

$$c) The \bar{T}_3 term: To bound the third term \bar{T}_3 in inequality (36), we observe that \begin{align} \bar{T}_3 &= \|e_{m}^{T}E_{t}^{i}((F_{t}^{i+1} - \bar{F}_{t}^{i+1,m} Q)\|_2 \leq \|e_{m}^{T}E_{t}^{i} \Delta_{t+1,m}\|_2 \bar{T}_{3a} + \|e_{m}^{T}E_{t}^{i} \Delta_{t+1,m} \Delta_{t+1,m}\|_2 \bar{T}_{3b}, \end{align} $$

where the last step is due to the choice of the orthogonal matrix $Q = (G_{t}^{i+1,m})^T \Delta_{t+1,i}$. Recalling $E_{t}^{i} := H_{\Omega}^{(-1)}(M^{t,i} - M^*)$, we write $\bar{T}_{3a}$ explicitly as

$$\bar{T}_{3a} = \|e_{m}^{T}H_{\Omega}^{(-1)}(M^{t,i} - M^*) \Delta_{t+1,i}\|_2.$$  

$$Since p \geq \frac{\log d}{d}$, we have $\sum_{j} \delta_{ij} \leq 2p d$ w.h.p. uniformly for all $j$. It follows that

$$\bar{T}_{3a}^2 \leq (2d) \|M_{m}^{t,i} - M^*_m\|_2, \infty \|\Delta_{t+1,i}\|_\infty^2 + d \|M_{m}^{t,i} - M^*_m\|_2, \infty \|\Delta_{t+1,i}\|_\infty^2.$$  

The term $\|\Delta_{t+1,i}\|_\infty$ can be bounded using the (33) proved in Section V-B, w.h.p.

$$\|\Delta_{t+1,i}\|_\infty \leq 20 \sigma_t \frac{1}{2^{16 \mu_r^2}} \left( \frac{1}{2} \right) \frac{1}{\sqrt{d}}.$$  

Putting together, we obtain that w.h.p.

$$\bar{T}_3 \leq 3\sigma_t \sqrt{d} \times 20 \frac{1}{64 \mu_r^2} \left( \frac{1}{2} \right) \frac{1}{\sqrt{d}} \|\Delta_{t+1,i}\|_\infty \|\Delta_{t+1,i}\|_\infty \leq \frac{\sigma_t}{128} \|\Delta_{t+1,i}\|_\infty.$$  

$$The term $\bar{T}_{3b}$, in (56) can be bounded using the same argument as above, which gives that w.h.p. $\bar{T}_{3b} \leq \frac{\sigma_t}{128} \|\Delta_{t+1,m}\|_2, \infty$.

Plugging the above bounds on $\bar{T}_{3a}$ and $\bar{T}_{3b}$ into (56), we have w.h.p.

$$||T_3||_2 \leq \frac{\sigma_t}{64} \|\Delta_{t+1,i}\|_\infty.$$  

Plugging the bounds (41), (55) and (57) on $\{\bar{T}_i, i = 1, 2, 3\}$ into inequality (36), we obtain that w.h.p.

$$\|\Delta_{t+1,i}\|_2 \leq \frac{1}{4} \frac{1}{2^{16 \mu_r^2}} \left( \frac{1}{2} \right)^t \frac{1}{\sqrt{d}} \sqrt{\frac{\mu_r}{d}} + \frac{100}{99} \sigma_t \left( \frac{3}{32} \|\Delta_{t+1,m}\|_2, \infty \right) + \frac{1}{32} \|\Delta_{t+1,m}\|_2, \infty \leq \frac{1}{2} \frac{1}{2^{16 \mu_r^2}} \left( \frac{1}{2} \right)^t \frac{1}{\sqrt{d}} \sqrt{\frac{\mu_r}{d}}$$.

$$d) Completing proof of $\ell_2, \infty$ bound in the induction hypothesis: We now plug the bound (58) on $\|e_{m}^{T}E_{t}^{i}((F_{t}^{i+1} - \bar{F}_{t}^{i+1,m} Q)\|_2$ into the inequality (35), thereby obtaining that w.h.p.

$$\|\Delta_{t+1,i}\|_2 \leq \frac{1}{4} \frac{1}{2^{16 \mu_r^2}} \left( \frac{1}{2} \right)^t \frac{1}{\sqrt{d}} \sqrt{\frac{\mu_r}{d}} + \frac{25}{99} \|\Delta_{t+1,i}\|_2, \infty.$$  

Taking the maximum of both sides of (59) over $i, m$ and rearranging terms, we obtain that w.h.p.

$$\|\Delta_{t+1,i}\|_\infty \leq \frac{1}{2^{16 \mu_r^2}} \left( \frac{1}{2} \right)^t \frac{1}{\sqrt{d}} \sqrt{\frac{\mu_r}{d}},$$

thereby proving the $\ell_2, \infty$ norm bound (18b) in the induction hypothesis for $t + 1$. 
4) Proximity Bound (18c) and Non-Commutativity Bound (18d) in the Induction Hypothesis: Recall that by definition,
\[
M^{t+1,i} = (F^{t+1,i} + M^{t,i}) \odot F^{t+1,i},
\]
\[
= P_r [M^* + \mathcal{H}_{t+1}^{-1}(M^{t,i} - M^*)],
\]
\[
M^{t+1,m} = (F^{t+1,m} \Lambda^{t+1,m}) \odot F^{t+1,m},
\]
\[
= P_r [M^* + \mathcal{H}_{t+1}^{-1}(M^{t,m} - M^*)].
\]

By Weyl's inequality, the eigen gap $\delta$ between the $r$-th and $(r+1)$-th eigenvalues of $M^* + \mathcal{H}_{t+1}^{-1}(M^{t,i} - M^*)$ is at least $\delta \geq \sigma_r - 2\|E^{t,i}\|_w \geq \sigma_r - \frac{\sigma_r}{10}$ w.h.p., where we use the bound (34) on $\|E^{t,i}\|_w$.

We consider $M^* + \mathcal{H}_{t+1}^{-1}(M^{t,m} - M^*)$ as a perturbed version of $M^* + \mathcal{H}_{t+1}^{-1}(M^{t,i} - M^*)$. Let
\[
W := \begin{cases}
\mathcal{H}_{t+1}^{-1}(M^{t,i} - M^*) : \text{discrepancy term} \\
\mathcal{H}_{t+1}^{-1}(M^{t,m} - M^*) : \text{contracting term}
\end{cases}
\]

be the corresponding perturbation matrix, decomposed into two terms following the strategy outlined in equation (12) in Section IV. Using the bound (34) on $\|E^{t,i}\|_w$ again, we have w.h.p.
\[
\|W\|_w \leq \|E^{t,i}\|_w + \|E^{t,m}\|_w \leq \frac{\sigma_r}{10}.
\]

Consequently, Davis-Kahan's inequality (Lemma 14) ensures that w.h.p.
\[
\|D^{t+1,i,m}\|_w \leq \sqrt{2} \|W\|_w \leq \frac{\sigma_r}{10} + \frac{\sigma_r}{10} = \frac{\sigma_r}{5}.
\]

To proceed, we control the two RHS terms to obtain a bound on $\|D^{t+1,i,m}\|_w$. We first consider the case $i = 0$, deferring the case $i \neq 0$ to later.

a) The $W^{(a)}$ term: Introduce the shorthand $M^{t,m}_{sk} := (1 - \frac{\delta_{mk}}{p})(M^{t,m}_{mk} - M^{t,m}_{mk})$. Noting that $W^{(a)}$ is only nonzero at its $m$-th row and column, we have the following explicit expression:
\[
\|W^{(a)} F^{t+1,m}\|_w = \left( \sum_{j < r} \sum_{k < d, k \neq m} (M^{t,m}_{sk} F^{t+1,m}_{mj})^2 \right) + \sum_{j < r} \sum_{k < d} (M^{t,m}_{sk} F^{t+1,m}_{kj})^2 + \sum_{j < r} \sum_{k < d, k \neq m} (M^{t,m}_{sk} F^{t+1,m}_{kj}) \cdot \frac{\sigma_r}{10}.
\]

For the term $T_1$, recalling that $F^{t+1,m} = \Delta^{t+1,m} + F^* G^{t+1,m}$ and $E^{t,m} := \mathcal{H}_t(M^{t,m} - M^*)$, we find that
\[
T_1 \leq \|F^{t+1,m}\|_w \|\mathcal{H}_t(M^{t,m} - M^*)e_m\|_2 \cdot \frac{\sigma_r}{10}.
\]

Combining this inequality with the assumption $\|F^*\|_w \leq \sqrt{\frac{\mu r}{d}}$ and the bounds (34) and (60) on $\|E^{t,m}\|_w$ and $\|\Delta^{t+1,m}\|_w$, we obtain that w.h.p.
\[
T_1 \leq \frac{\sigma_r}{64} \frac{1}{2^{2m} \sqrt{\mu r}} \left( \frac{1}{2} \right)^{t+1} \sqrt{\frac{\mu r}{d}}.
\]

For the term $T_2$, we begin with the bound
\[
T_2 \leq \sqrt{r} \max_{j \leq r} \left| \sum_{k \leq d} M^{t,m}_{sk} F^{t+1,m}_{kj} \right|.
\]

Bounding the last RHS using the same argument as in the derivation of (41), we find that w.h.p.
\[
T_2 \leq \frac{\sigma_r}{64} \frac{1}{2^{2m} \sqrt{\mu r}} \left( \frac{1}{2} \right)^{t+1} \sqrt{\frac{\mu r}{d}}.
\]

Plugging the above bounds on $T_1$ and $T_2$ into (62), we conclude that w.h.p.
\[
\|W^{(a)} F^{t+1,m}\|_w \leq \frac{\sigma_r}{32} \frac{1}{2^{2m} \sqrt{\mu r}} \left( \frac{1}{2} \right)^{t+1} \sqrt{\frac{\mu r}{d}}.
\]

b) The $W^{(b)}$ term: Specializing the decomposition in (42) to $M^{t,0} := M^t$, we have
\[
M^{t,m} - M^t = (D^{t,m,0} \Lambda^{t,0}) \odot F^{t,m} - (F^t S^{t,m,0}) \odot F^{t,m} + (F^t \Lambda G^{t,m,0}) \odot D^{t,m,0}.
\]

Accordingly, we split $\|W^{(b)} F^{t+1,m}\|_w$ into three terms as
\[
\|W^{(b)} F^{t+1,m}\|_w = \|\mathcal{H}_t((D^{t,m,0} \Lambda^{t,0}) \odot F^{t,m}) F^{t+1,m}\|_w + \|\mathcal{H}_t((F^t S^{t,m,0}) \odot F^{t,m}) F^{t+1,m}\|_w + \|\mathcal{H}_t((F^t \Lambda G^{t,m,0}) \odot D^{t,m,0}) F^{t+1,m}\|_w.
\]

For the term $R_1$, we introduce the shorthand $D^{t,m,0} := D^{t,m,0} \Lambda^{t,m}$ and further split $R_1$ into three terms:
\[
R_1 = \|\mathcal{H}_t(D^{t,m,0} \odot F^{t,m}) F^{t+1,m}\|_w \leq \|\mathcal{H}_t(D^{t,m,0} \odot F^{t,m}) \Delta^{t+1,m}\|_w + \|\mathcal{H}_t(D^{t,m,0} \odot F^{t,m}) F^*\|_w.
\]

For the term $R_2$, we introduce the shorthand $G^{t,m,0} := G^{t,m,0} \odot F^*$ and further split $R_2$ into three terms:
\[
R_2 = \|\mathcal{H}_t(D^{t,m,0} \odot \Delta^{t+1,m}) F^*\|_w + \|\mathcal{H}_t(D^{t,m,0} \odot \Delta^{t+1,m}) F^*\|_w.
\]
where we use triangle inequality in steps (a) and (b), and the unitary invariance of \( \| \cdot \| \) in step (a).

Denote the shorthand \( R_{m_l j, l, 2}^{t+1} \) as
\[
R_{m_l j, l, 2}^{t+1} := \sum_{k \leq d} F_{k, l, 2}^t \Delta_{k_l, t}^{t+1, m} \left( 1 - \frac{\delta_{j k}}{p} \right).
\]
We first write the term \( R_{1a} \) in (66) explicitly as
\[
R_{1a} = \left( \sum_{j \leq d, l \leq r} \left( \sum_{l \leq r} (D_{A, m, 0}^t)_{j, l} R_{m_l j, l, 2}^{t+1} \right) \right)^2
\]
with Cauchy-Schwarz in step (a) and that
\[
\sum_{j \leq d, l \leq r} \left( (D_{A, m, 0}^t)_{j, l} \right)^2 = \| D_{A, m, 0}^t \|_2^2 \text{ in step (b)}.
\]
Note that
\[
\max_{l \leq r} \sum_{l \leq r} (R_{m_l j, l, 2}^{t+1})^2 \leq \sqrt{r} \max_{l \leq r} \sum_{l \leq r} (R_{m_l j, l, 2}^{t+1})^2.
\]
Therefore, we may continue from equation (67) to obtain that w.h.p.
\[
R_{1a} \leq \sum_{j \leq d, l \leq r} \left( \sum_{l \leq r} (D_{A, m, 0}^t)_{j, l} R_{m_l j, l, 2}^{t+1} \right)^2
\]
with Cauchy-Schwarz in step (a) and that
\[
\sum_{j \leq d, l \leq r} \left( (D_{A, m, 0}^t)_{j, l} \right)^2 = \| D_{A, m, 0}^t \|_2^2 \text{ in step (b)}.
\]
Note that \( R_{1a} \) can be written compactly as \( \sum_{l \leq r} \mathcal{H}_\Omega (F_{i, l, 2}^t \otimes (S_{F, m, 0}^t)^{t+1, m}) \leq 2 \sqrt{\frac{\log d}{p}} \times \| F_{t+1, m}^t \|_{2, \infty} \) w.h.p. by Lemma 22, and
\[
\| S_{F, m, 0}^t \|_{\infty} \leq \| (S_{F, m, 0}^t)^{t+1, m} \|_{\infty} \leq \| S_{F, m, 0}^t \|_{\infty} \text{ by construction.}
\]
Plugging these bounds into (73), we obtain that w.h.p.
\[
R_{2} \leq \max_{l, l \leq r} \sum_{l \leq r} \left( \sum_{l \leq r} (R_{m_l j, l, 2}^{t+1})^2 \right) \geq \mathcal{H}_\Omega (F_{i, l, 2}^t \otimes (S_{F, m, 0}^t)^{t+1, m}) \leq 2 \sqrt{\frac{\log d}{p}} \times \| F_{t+1, m}^t \|_{2, \infty} \text{ w.h.p. by Lemma 22, and}
\]
\[
\| S_{F, m, 0}^t \|_{\infty} \leq \| (S_{F, m, 0}^t)^{t+1, m} \|_{\infty} \leq \| S_{F, m, 0}^t \|_{\infty} \text{ by construction.}
\]
Plugging these bounds into (74) and (72), we obtain that w.h.p.
\[
R_{2} \leq \max_{l, l \leq r} \sum_{l \leq r} \left( \sum_{l \leq r} (R_{m_l j, l, 2}^{t+1})^2 \right) \geq \mathcal{H}_\Omega (F_{i, l, 2}^t \otimes (S_{F, m, 0}^t)^{t+1, m}) \leq 2 \sqrt{\frac{\log d}{p}} \times \| F_{t+1, m}^t \|_{2, \infty} \text{ w.h.p. by Lemma 22, and}
\]
\[
\| S_{F, m, 0}^t \|_{\infty} \leq \| (S_{F, m, 0}^t)^{t+1, m} \|_{\infty} \leq \| S_{F, m, 0}^t \|_{\infty} \text{ by construction.}
\]
Plugging these bounds into (74) and (72), we obtain that w.h.p.
\[
R_{2} \leq \max_{l, l \leq r} \sum_{l \leq r} \left( \sum_{l \leq r} (R_{m_l j, l, 2}^{t+1})^2 \right) \geq \mathcal{H}_\Omega (F_{i, l, 2}^t \otimes (S_{F, m, 0}^t)^{t+1, m}) \leq 2 \sqrt{\frac{\log d}{p}} \times \| F_{t+1, m}^t \|_{2, \infty} \text{ w.h.p. by Lemma 22, and}
\]
\[
\| S_{F, m, 0}^t \|_{\infty} \leq \| (S_{F, m, 0}^t)^{t+1, m} \|_{\infty} \leq \| S_{F, m, 0}^t \|_{\infty} \text{ by construction.}
\]
Plugging these bounds into (74) and (72), we obtain that w.h.p.
\[
R_{2} \leq \max_{l, l \leq r} \sum_{l \leq r} \left( \sum_{l \leq r} (R_{m_l j, l, 2}^{t+1})^2 \right) \geq \mathcal{H}_\Omega (F_{i, l, 2}^t \otimes (S_{F, m, 0}^t)^{t+1, m}) \leq 2 \sqrt{\frac{\log d}{p}} \times \| F_{t+1, m}^t \|_{2, \infty} \text{ w.h.p. by Lemma 22, and}
\]
\[
\| S_{F, m, 0}^t \|_{\infty} \leq \| (S_{F, m, 0}^t)^{t+1, m} \|_{\infty} \leq \| S_{F, m, 0}^t \|_{\infty} \text{ by construction.}
\]
Plugging these bounds into (74) and (72), we obtain that w.h.p.
\[
R_{2} \leq \max_{l, l \leq r} \sum_{l \leq r} \left( \sum_{l \leq r} (R_{m_l j, l, 2}^{t+1})^2 \right) \geq \mathcal{H}_\Omega (F_{i, l, 2}^t \otimes (S_{F, m, 0}^t)^{t+1, m}) \leq 2 \sqrt{\frac{\log d}{p}} \times \| F_{t+1, m}^t \|_{2, \infty} \text{ w.h.p. by Lemma 22, and}
\]
\[
\| S_{F, m, 0}^t \|_{\infty} \leq \| (S_{F, m, 0}^t)^{t+1, m} \|_{\infty} \leq \| S_{F, m, 0}^t \|_{\infty} \text{ by construction.}
\]
hypothesis, the bound \( \| \Lambda^{t,i} \|_{\infty} \leq 2\sigma_1 \) derived before (33), and the assumption \( p \geq \frac{c_0 \mu r}{d \log d} \). Doing so yields that \( \text{w.h.p.} \)

\[
R_3 \leq \frac{\sigma_1}{64 \lambda^2 \mu r^2} \left( \frac{1}{2} \right)^{t+1} \sqrt{\frac{\mu r}{d}}.
\]

(77)

Plugging the above bounds (71), (75) and (77) for \( \{ R_i, i \in [3] \} \) into (65), we obtain that \( \text{w.h.p.} \)

\[
\| W^{(b)} F^{t+1,m} \|_p \leq \frac{\sigma_1}{16 \lambda^2 \mu r^2} \left( \frac{1}{2} \right)^{t+1} \sqrt{\frac{\mu r}{d}}.
\]

(78)

c) Completing proof of proximity and non-commutativity bounds in induction hypothesis: Plugging the bounds (64) and (78) on \( \| W^{(a)} F^{t+1,m} \|_p \) and \( \| W^{(b)} F^{t+1,m} \|_p \) into inequality (61), we get \( \text{w.h.p.} \)

\[
\| D^{t+1,0,m} \|_p \leq \frac{2}{\sigma_1} \left( \| W^{(a)} F^{t+1,m} \|_p + \| W^{(b)} F^{t+1,m} \|_p \right).
\]

(79)

To bound \( \| D^{t+1,i.m} \|_p \) for \( i \neq 0 \), we follow the same argument used in deriving the inequality (32). This argument yields that \( \text{w.h.p.} \)

\[
\| D^{t+1,i.m} \|_p \leq \| D^{t+1,0,m} \|_p + \| D^{t+1,0,m} \|_p \leq \frac{1}{2} \left( \frac{1}{2} \right)^{t+1} \sqrt{\frac{\mu r}{d}}.
\]

Taking the maximum of both sides of the last two equations over \( i, m \), we establish the proximity bound (18c) in the induction hypothesis for \( t+1 \).

Finally, to establish the non-commutativity bound (18d) in the induction hypothesis for \( t+1 \), we apply Lemma 2 to obtain that \( \text{w.h.p.} \)

\[
\| S^{t+1,0,m} \|_p \leq \| S^{t+1,0,m} \|_p \leq 4k \| W F^{t+1,m} \|_p \leq \sigma_1 \left( \frac{1}{2} \right)^{t+1} \sqrt{\frac{\mu r}{d}}.
\]

Taking maximum over \( m \) on both sides proves the non-commutativity bound.

Recall we proved the operator norm and \( \ell_2, \infty \) norm bounds of the induction hypothesis for \( t+1 \) in (34) and (60), respectively. Therefore, we have completed the induction step and concluded that the induction hypothesis (18) holds \( \text{w.h.p.} \) for each \( t \in [t_0] \), where \( t_0 := 5 \log_2 d + 12 \). Invoking Lemma 19 and taking a union bound over \( t \in [t_0] \), we deduce from the hypothesis (18) that the desired error bound (14) on the original matrix \( M^t \) holds \( \text{w.h.p.} \) for the first \( t_0 \) iterations; that is, \( \| M^t - M^* \|_\infty \leq \frac{1}{2} \left( \frac{1}{2} \right)^{t_0} \sigma_1, \forall t \in [t_0] \).

D. Part 2: Bounds for an Infinite Number of Iterations

The previous induction argument is insufficient for controlling all iterations \( t > t_0 \), as the union bound would fail for an unbounded number of \( t \)'s. To control the error for an infinite number of PGD iterations, we employ a different argument and establish a uniform \textit{Frobenius} norm error bound as in (15), i.e., \( \text{w.h.p.} \) there holds \( \| M^t - M^* \|_p \leq \left( \frac{1}{2} \right)^{t-t_0} \| M^{t_0} - M^* \|_p, \forall t \geq t_0 \).

We prove the Frobenius bound (15) by induction. The base case \( t = t_0 \) trivially holds. Below we assume that (15) holds for all iterations \( t_0, t_0+1, \ldots, t \). For each \( t \), define the shorthand \( \tilde{D}^t := M^t - M^* \).

The proof relies on the following \textit{uniform} bound, which is proved in Appendix A-C.

\textbf{Lemma 3:} In the setting of SMC \((M^*, p)\), suppose \( p \geq \frac{c_0 \mu r^2}{d \log d} \). Then with probability at least \( 1 - 4d^{-2} \), the bound

\[
\left| \left( M^* - M^* \right) - \left( \Pi_{\Omega} (M^* - M^*) \right) \right| \leq \frac{1}{d} \left( \left\| M^* - M^* \right\|_p \left\| M - M^* \right\|_p \right)
\]

holds simultaneously for all rank-\( r \) matrices \( M, M^* \in \mathcal{S}^{d \times d} \) satisfying \( \max\{ \| M \|_\infty, \| M^* \|_\infty \} \leq \frac{\mu r}{n} \sigma_1 (M^*) \) and \( \max\{ \| M - M^* \|_p, \left\| M^* - M^* \right\|_p \} \leq \frac{\mu r}{n} \times \frac{1}{\sqrt{d}} \sigma_1 (M^*) \).

We record several facts that are useful in verifying the premise of Lemma 3. First note that

\[
\| \tilde{D}^t \|_p \leq \| \tilde{D}^{t_0} \|_p \leq d \| \tilde{D}^{t_0} \|_\infty \leq \frac{1}{2} \left( \frac{1}{2} \right)^{t_0} \frac{1}{d} \left( \frac{1}{2} \right)^{t_0} \mu r \sqrt{d}.
\]

(80)

We apply the uniform bound in Lemma 22, and combining with the above bound on \( \| \tilde{D}^t \|_\infty \) and the assumption \( p \geq \frac{c_0 \mu r^2}{d \log d} \), we have

\[
\| E^t \|_p \leq \| \tilde{H}_{\Omega} (\tilde{D}^t) \|_p \leq 2c \sqrt{\frac{2d \log d}{p}} \| \tilde{D}^t \|_\infty \leq \frac{1}{64^2 \sqrt{d^2} \sigma_1}.
\]

By definition, \( M^{t+1} \) is the best rank-\( r \) approximation of \( M^* + E^t \) in Frobenius norm, whence

\[
\| M^{t+1} - (M^* + E^t) \|_p^2 \leq \| E^t \|_p^2.
\]

(82)

Consequently, we have

\[
\| \tilde{D}^{t+1} \|_p \leq \| M^{t+1} - (M^* + E^t) \|_p + \| E^t \|_p \leq 2 \| E^t \|_p \leq 2 \times \frac{1}{64^2 \sqrt{d^2} \sigma_1},
\]

(83)

where step \( (a) \) follows from (82), and step \( (b) \) follows from

\[
\| E^t \|_p \leq \sqrt{d} \| E^t \|_p \text{ and the previous bound on } \| E^t \|_p.
\]

We also have

\[
\| M^{t+1} \|_\infty \leq \| M^* \|_\infty + \| \tilde{D}^{t+1} \|_\infty \leq \frac{\mu r}{d} + \frac{\mu r}{d} \leq \frac{2}{d} \mu r \frac{\mu r}{d} \sigma_1.
\]

(84)
In view of the inequalities (81), (80), (83) and (84), we see that the premise of Lemma 3 is satisfied by letting $M^* = M^{t+1}, M' = M^*$. Expanding the square on the LHS of inequality (82), we have

$$\|\tilde{D}^{t+1} \|_r^p \leq -2(\tilde{D}^{t+1}, E^t) \leq -2(\tilde{D}^{t+1} \ (I - p^{-1}\Pi_\Omega)(\tilde{D}^t)) \leq 2\left| \frac{1}{p}(\Pi_\Omega \tilde{D}^{t+1}, \Pi_\Omega \tilde{D}^t) - (\tilde{D}^{t+1}, \tilde{D}^t) \right| \leq 2 \times \frac{1}{4} \|\tilde{D}^{t+1}\|_r \|\tilde{D}^t\|_r,$$

where in last step we apply Lemma 3. The above inequality implies the contraction $\|\tilde{D}^{t+1}\|_r \leq \frac{1}{2}\|\tilde{D}^t\|_r$, hence the Frobenius norm bound (15) also holds for $t + 1$. We have thus completed the induction step and established (15) for all the $t \geq t_0$.

We can now complete the proof of Theorem 1. We have w.h.p.

$$\|M^t - M^*\|_\infty \leq \|M^t - M^*\|_r \leq \left(\frac{1}{2}\right)^{t-t_0} \|M_{t_0} - M^*\|_r \leq \left(\frac{1}{2}\right)^{t-t_0} \cdot d \cdot \|M_{t_0} - M^*\|_\infty \leq \left(\frac{1}{2}\right)^{t-t_0} \cdot d \cdot \left(\frac{1}{2}\right)^{t_0} \sigma_1 = \left(\frac{1}{2}\right)^t \sigma_1, \forall t \geq t_0,$$

where step (a) follows from (14) and step (b) follows from (15). With the above bound, as well as the bound (14) for $1 \leq t \leq t_0$, we have established the claim in Theorem 1.

VI. PROOF OF THEOREM 2

In this section we prove Theorem 2 for NNM using a combination of our leave-one-out framework and the Golfing Scheme introduced in [18], [30]. We assume that $d = d_1 = d_2$ for simplicity; the proof of the general case follows the same lines. We shall make use of the auxiliary lemmas given in Appendix C.

A. Preliminaries

Let the singular value decomposition of $M^*$ be $M^* = U \Sigma V^T$. For a matrix $Z \in \mathbb{R}^{d \times d}$, we define the projections $\mathcal{P}_T(Z) := UU^T Z + ZV V^T - UU^T Z V V^T$ and $\mathcal{P}_{T^c}(Z) := Z - \mathcal{P}_T(Z) = (I - UU^T) Z (I - V V^T)$. Introduce the shorthand $\mathcal{R}_\Omega := \frac{1}{p} \Pi_\Omega$, which has the explicit expression $[\mathcal{R}_\Omega(Z)]_{ij} = \frac{1}{p} \delta_{ij} Z_{ij}$. We also define the operator $\mathcal{H}_{\Omega} := \mathcal{I} - \mathcal{R}_\Omega = \mathcal{I} - \frac{1}{p} \Pi_\Omega$ and the linear subspace $T := \{ \mathcal{P}_T(Z) \mid Z \in \mathbb{R}^{d \times d} \}$. For a linear map $A$ on matrices, its operator norm is defined as $\|A\|_\infty := \max_{Z : \|Z\|_r = 1} \|A(Z)\|_r$.

We make use of the following standard result, which provides a deterministic sufficient condition for the optimality of $M^*$ to the nuclear norm minimization problem.

**Proposition 1** ([7, Proposition 2]): Suppose that $p \geq \frac{1}{r}$. The matrix $M^*$ is the unique optimal solution to the NNM problem (3) if the following conditions hold:

1. $\|\mathcal{P}_T \mathcal{R}_{\Omega} \mathcal{P}_T - \mathcal{P}_T\|_\infty \leq \frac{1}{2}$
2. There exists a dual certificate $Y \in \mathbb{R}^{d \times d}$ that satisfies $\Pi_\Omega(Y) = Y$ and
   a) $\|\mathcal{P}_{T^c}(Y)\|_\infty \leq \frac{1}{2}$
   b) $\|\mathcal{P}_T(Y) - U V^T\|_\infty \leq \frac{1}{\mu r}$

The first condition in Proposition 1 can be verified using the following well-known result from the matrix completion literature.

**Proposition 2** ([4, Theorem 4.1], [8, Lemma 11]): If $p \geq \frac{\mu r \log d}{\delta^2}$, then with high probability

$$\|\mathcal{P}_T \mathcal{R}_{\Omega} \mathcal{P}_T - \mathcal{P}_T\|_\infty \leq \frac{1}{64}.$$

We are left to construct a dual certificate $Y$ such that Condition 2 in Proposition 1 is satisfied. Recalling the definition of the row-wise $\ell_2$ norm in Section III-A, we further define the doubly $\ell_{2,\infty}$ norm $\|Z\|_{2,\infty} := \max\{\|Z\|_{2,\infty} \|Z^T\|_{2,\infty}\}$, which plays a crucial role in the dual certificate construction.

**d) Constructing the Dual Certificate:** Our strategy is to construct the desired certificate $Y$ by running an iterate procedure that uses the same set of random samples, and then apply leave-one-out to analyze these correlated iterations. This procedure is warm-started by first employing the Golfing Scheme for $O(\log \mu_\Omega)$ iterations, each using an independent set of samples; as mentioned, doing so allows us to achieve tighter dependence on $\mu_\Omega$.

Now for the details. Set $k_0 := C_0 \max\{1, \log(\mu_\Omega)\}$ for some large enough numerical constant $C_0$. Suppose that the set $\Omega$ of observed entries is generated from $\Omega = \cup_{t=1}^{k_0} \Omega_t$, where for each $t$ and matrix index $(i, j)$ we have $P[(i, j) \in \Omega_t] = q : = 1 - (1 - p)^{\delta_{ij}}$ independently of all others. Clearly this $\Omega$ has the same distribution as the original model MC($M^*, p$). Denote the projection $\mathcal{R}_{\Omega_t} := [\Pi_{\Omega_t}(Z)]_{ij} = Z_{ij} \mathbb{I}((i, j) \in \Omega_t)$ and $\mathcal{R}_{\Omega_0} = \frac{1}{q} \Pi_\Omega$. Following our strategy, we use independent samples in the first $k_0$ iterations: set $W^{0} := U V^T$ and

$$W^t := \mathcal{P}_T \mathcal{H}_{\Omega_t}(W^{t-1}), \quad t = 1, 2, \ldots, k_0 - 1, \quad (85)$$

where $\mathcal{H}_{\Omega_t} = I - \frac{1}{q} \Pi_\Omega$. We then use the same sample set $\Omega_{k_0}$ in the next $t_0 := 2 \log d + 2$ iterations: set $Z^0 = W^{k_0 - 1}$ and for $t = 1, 2, \ldots, t_0 - 1$ define

$$Z^t := \mathcal{P}_T \mathcal{H}_{\Omega_{k_0}}(Z^{t-1}) = (\mathcal{P}_T \mathcal{H}_{\Omega_{k_0}})^t(W^{k_0 - 1}). \quad (86)$$

The final dual certificate $Y$ is constructed by summing up the above iterates: set

$$Y_1 := \sum_{t=1}^{k_0 - 1} \mathcal{R}_{\Omega_t} \mathcal{P}_T(W^{t-1}),$$

$$Y_2 := \sum_{t=1}^{k_0 - 1} \mathcal{R}_{\Omega_{k_0}} \mathcal{P}_T(Z^{t-1}),$$

$$Y := Y_1 + Y_2. \quad (87)$$
Below we show that the matrix \(Y\) satisfies the conditions in Proposition 1.

e) Validating Condition 2(b): Note \(q \geq \frac{p}{k_0} \gtrsim \frac{\mu r \log d}{d}\) as \(p \gtrsim \frac{\mu r \log d \log(\mu r)}{d}\). Applying Proposition 2 with \(\Omega\) replaced by \(\Omega_i\), we obtain that w.h.p., for all \(t = 1, 2, \ldots, k_0 - 1,\)
\[
\|W^t\|_F \leq \|P_T H_{\Omega_i} P_T\|_F \|W^{t-1}\|_F \\
\leq \frac{1}{16} \|W^{t-1}\|_F,
\]  
(88)
and for all \(t = 1, 2, \ldots, t_0,\)
\[
\|Z^t\|_F \leq \|P_T H_{\Omega_i} P_T\|_F \|Z^{t-1}\|_F \\
\leq \frac{1}{16} \|Z^{t-1}\|_F.
\]  
(89)
Using the last two inequalities, we obtain that w.h.p.,
\[
\|P_T(Y) - U V^T\|_F \overset{(a)}{\leq} \|Z^{t_0}\|_F
\leq \left(\frac{1}{2}\right)^{t_0} \|Z_0\|_F = \left(\frac{1}{2}\right)^{t_0} \|W_{k_0-1}\|_F
\leq \left(\frac{1}{2}\right)^{t_0} \|U V^T\|_F
\leq \frac{\sqrt{\mu r}}{4d^2},
\]  
(90)
where step (a) follows from the fact that \(\|P_T(Y) - U V^T = -Z_0\|_F\), which can be verified by definition and direct computation. Therefore, Condition 2(b) in Proposition 1 is satisfied.

f) Validating Condition 2(a): From the definitions of \(Y_1, Y_2,\) and \(Y\) in (87), we have
\[
\|P_{\perp}(Y)\|_F \overset{(a)}{\leq} \sum_{i=1}^{k_0-1} \|P_{\perp}(R_{\Omega_i} P_T - P_T)(W^{t-1})\|_F
+ \sum_{i=1}^{t_0} \|P_{\perp}(R_{\Omega_i} P_T - P_T)(Z^{t-1})\|_F
\leq \sum_{i=1}^{k_0-1} \|H_{\Omega_i} W^{t-1}\|_F + \sum_{i=1}^{t_0} \|H_{\Omega_i} Z^{t-1}\|_F
\leq \frac{1}{2^{k_0-1}} \left(\frac{1}{\mu r}\right)^{10},
\]  
(91)
where in step (a) we use the facts that \(\|P_{\perp}(A)\|_F = \|(I - U U^T) A (I - V V^T)\|_F \leq \|A\|_F\) for any \(A \in \mathbb{R}^{d \times d}\) and that \(Z^{t-1}, W^{t-1} \in T\). To bound the term \(T_1\), we follow exactly the same arguments in [7, “Validating Condition 2(a)”, pp 12-13], which gives that \(T_1 \leq \frac{1}{4}\) w.h.p. Introduce the shorthand
\[
G : = \frac{1}{2^{k_0-1}} \left(\frac{1}{\mu r}\right)^{10},
\]
which will be used throughout the rest of the proof. Turning to the term \(T_2\), we claim that w.h.p.,
\[
\|Z^t\|_F \leq \frac{1}{2^t} \mu r \sqrt{\mu r} G, \quad t = 0, 1, \ldots, t_0.
\]  
(92)
We prove this bound later. Taking it as given for now, we apply the second inequality in Lemma 22 with \(\Omega\) replaced by \(\Omega_{k_0}\), which gives that w.h.p.
\[
\|H_{\Omega_{k_0}} (Z^t)\|_F \leq \left(\frac{1}{\sqrt{2}}\right)^{t+3}, \quad t = 0, 1, \ldots, t_0.
\]
Plugging into the expression of \(T_2\), we obtain \(T_2 \leq \frac{1}{4}\). Combining the bounds \(T_1\) and \(T_2\), we see that Condition 2(a) in Proposition 1 is satisfied, thereby establishing Theorem 2.

The rest of this section is devoted to proving the bound (91). We first show that the bound holds for \(t = 0\). Recall the definition of \(W^t\) in (85). Note that \(\|U V^T\|_F \leq \frac{\mu r}{\sqrt{d}}\), and that \(\Omega_i\) is independent of \(W^{t-1}\) for each \(t \leq k_0\). Applying Lemma 21 gives that w.h.p.,
\[
\|W^t\|_F \leq \frac{1}{16} \frac{\mu r}{d}, \quad t = 1, 2, \ldots, k_0.
\]  
(93)
Using the above inequality and recalling the definition \(Z^0 : = W^{k_0-1}\), we obtain that w.h.p.,
\[
\|Z^0\|_F \leq G \frac{\mu r}{d} \frac{1}{(\mu r)^5} \leq \frac{\mu r}{d} G,
\]
\[
\|Z^0\|_F \leq d \cdot G \cdot \frac{\mu r}{d} \frac{1}{(\mu r)^5} \leq G,
\]
\[
\|Z^0\|_{(2,\infty)^2} \leq \sqrt{d} \cdot G \frac{\mu r}{d} \frac{1}{(\mu r)^5} \leq \sqrt{\frac{\mu r}{d}} G,
\]  
(93)
provided that the constant \(C_0\) in \(k = C_0 \log(\mu r)\) is sufficiently large. Therefore, the bound (91) holds for \(t = 0\). Below we prove the bound for \(1 \leq t \leq t_0\) using the leave-one-out technique.

B. Leave-One-Out Analysis of the \(Z^t\) Sequence

In this subsection, we abuse the notation and write \(\Omega_{k_0}\) as \(\Omega\), whose observation probability is \(q : = 1 - (1 - p) \frac{1}{k_0} \gtrsim \frac{\mu r \log d}{d}\) since \(p \gtrsim \frac{\mu r \log d \log(\mu r)}{d}\). For each \(w \in [d] \times [d]\), define the operator \(\mathcal{H}_{\Omega_{(w)}} : \mathbb{R}^{d \times d} \to \mathbb{R}^{d \times d}\) by
\[
(\mathcal{H}_{\Omega_{(w)}} Z)_{ij} = \begin{cases} \left(1 - \frac{1}{q} \delta_{ij}\right) Z_{ij}, & i \neq w_1 \text{ and } j \neq w_2, \\ 0, & i = w_1 \text{ or } j = w_2. \end{cases}
\]
Let \(\mathcal{H}_{\Omega_{(w)}} : = \mathcal{H}_{\Omega} - \mathcal{H}_{\Omega_{(w)}}\). For each \(w \in [d] \times [d]\), we introduce the leave-one-out sequence
\[
Z^0_{(w)} = Z^0; \quad Z^{t,w} = (P_T \mathcal{H}_{\Omega_{(w)}} (Z^{t-1,w}),
\]
for \(t = 1, 2, \ldots\) By construction, this sequence is independent of \(\delta_{w_1,w_2}\) and \(\mathcal{H}_{\Omega_{(w)}}\), a property we crucially rely on below.

We first record a few technical lemmas that provide concentration bounds for the operators \(\mathcal{H}_{\Omega}\) and \(\mathcal{H}_{\Omega_{(w)}}\). The first lemma is proved in Appendix B-A.

Lemma 4: If \(q \gtrsim \frac{\mu r \log d}{d}\), then we have w.h.p.
\[
\|P_T \mathcal{H}_{\Omega_{(w)}} (Z)\|_F \leq \frac{1}{8} \sqrt{\frac{\mu r}{d}} \|Z\|_F
\]
uniformly for all \(Z \in T\). The same statement holds with \(\mathcal{H}_{\Omega}\) replaced by \(\mathcal{H}_{\Omega_{(w)}}\) for each \(w \in [d] \times [d]\).

The lemma below is proved in Appendix B-B.
Lemma 5: If \( q \gtrsim \frac{\mu r \log d}{d} \), then for each \( t = 1, 2, \ldots, t_0 \), we have w.h.p.

\[
\max\left\{ \| (P_T \mathcal{H}_1(Z^{t-1}))_{v_1}, (P_T \mathcal{H}_1(Z^{t-1}))_{v_2} \|_2 \right\} \leq \frac{1}{32} \left[ \| Z^{t-1,v} \|_{(2, \infty)^2} + \frac{\log d}{q} \| Z^{t-1,v} \|_{\infty} \right] + \frac{\mu r}{d} \| Z^{t-1,v} \|_F
\]

The same statement holds with \( \mathcal{H}_1 \) replaced by \( \mathcal{H}_1^{(-w)} \) for each \( w \in [d] \times [d] \).

The lemma below is proved in Appendix B-C.

Lemma 6: If \( q \gtrsim \frac{\mu r \log d}{d} \), then for each \( t = 1, 2, \ldots, t_0 \), we have w.h.p.

\[
\left( P_T \mathcal{H}_1(Z^{t-1}) \right)_{v_1, v_2} \leq \frac{1}{32} \| Z^{t-1,v} \|_{\infty} + \frac{1}{32} \frac{\mu r}{d} \| Z^{t-1,v} \|_F
\]

The same statement holds with \( \mathcal{H}_1 \) replaced by \( \mathcal{H}_1^{(-w)} \) for each \( w \in [d] \times [d] \).

The lemma below is proved in Appendix B-D.

Lemma 7: If \( q \gtrsim \frac{\mu r \log d}{d} \), then we have w.h.p., uniformly for all \( Z \in \mathcal{T} \)

\[
\| P_T \mathcal{H}_1(Z) \|_{(2, \infty)^2} \leq \frac{1}{32} \| Z \|_F.
\]

The same statement holds with \( \mathcal{H}_1 \) replaced by \( \mathcal{H}_1^{(-w)} \) for each \( w \in [d] \times [d] \).

The lemma below is proved in Appendix B-E.

Lemma 8: If \( q \gtrsim \frac{\mu r \log d}{d} \), then for each \( w \in [d] \times [d] \) and each fixed \( Z \in \mathcal{T} \), we have with probability at least \( 1 - d^{-q} \),

\[
\| P_T \mathcal{H}_1^{(w)}(Z) \|_F \leq \frac{1}{32} \| Z \|_{(2, \infty)^2} + \frac{\log d}{q} \| Z \|_{\infty}.
\]

Finally, we note that the inequalities (89) and (93) imply that w.h.p.

\[
\| Z^t \|_F \leq \frac{1}{2} G, \quad \forall t = 1, 2, \ldots, t_0.
\]

We are now ready to prove the inequality (91) by induction on \( t \). The induction hypothesis is

\[
\| Z^{t-1} \|_{(2, \infty)^2} \leq \left( \frac{1}{2} \right)^{t-1} \sqrt{\frac{\mu r}{d} G}, \quad (95a)
\]

\[
\| Z^{t-1,w} \|_{(2, \infty)^2} \leq \left( \frac{1}{2} \right)^{t-1} \sqrt{\frac{\mu r}{d} G}, \quad \forall w \in [d]^2, \quad (95b)
\]

\[
\| Z^{t-1} \|_\infty \leq \left( \frac{1}{2} \right)^{t-1} \sqrt{\frac{\mu r}{d} G}, \quad (95c)
\]

\[
\| Z^{t-1,w} \|_\infty \leq \left( \frac{1}{2} \right)^{t-1} \sqrt{\frac{\mu r}{d} G}, \quad \forall w \in [d]^2, \quad (95d)
\]

\[
\| Z^{t-1} - Z^{t-1,w} \|_F \leq \left( \frac{1}{2} \right)^{t-1} \sqrt{\frac{\mu r}{d} G}, \quad \forall w \in [d]^2. \quad (95e)
\]

Here \( [d]^2 = [d] \times [d] \). We have proved the base case in equation (93), noting that \( Z^0,w = Z^0 \). Assuming that the bounds in (95) hold for \( t - 1 \), we show below that each of them also holds for \( t \) w.h.p.

\[
\| Z^{t-1,w} \|_{(2, \infty)^2} \leq \left( \frac{1}{2} \right)^{t-1} \sqrt{\frac{\mu r}{d} G}. \quad (96)
\]

We bound the term \( T_3 \) using the induction hypothesis (95b) and (95d), and bound \( T_2 \) using inequality (94). For the term \( T_3 \), we apply Lemma 7 and the induction hypothesis (95e) to obtain that w.h.p.

\[
T_3 \leq \frac{1}{32} \| Z^{t-1} - Z^{t-1,w} \|_F \leq \frac{1}{32} \left( \frac{1}{2} \right)^{t-1} \sqrt{\frac{\mu r}{d} G}.
\]

Combining the above bounds, we obtain that \( \| Z^{t_1,w} \|_2 \leq \left( \frac{1}{2} \right)^t \sqrt{\frac{\mu r}{d} G} \) w.h.p. We can bound \( \| Z^{t_1,w}_1 \|_2 \) in a similar way. Taking a union bound over all \( w \in [d] \times [d] \) proves the inequality (95a) for \( t \).

h) The \| \cdot \|_{(2, \infty)^2} \) bound (95b): Fix \( v, w \in [d] \times [d] \). We have

\[
\| Z^{t_1,w}_1 \|_2 = \| (P_T \mathcal{H}_1^{(w)}(Z^{t-1,w}))_{v_1, v_2} \|_2 \leq \| (P_T \mathcal{H}_1^{(w)}(Z^{t-1,w}))_{v_1, v_2} \|_2 + \| (P_T \mathcal{H}_1^{(w)}(Z^{t-1,w}) - Z^{t-1,w}_1)_{v_1, v_2} \|_2.
\]

The first RHS term can be bounded in a similar way as in the above proof of (95a). The second term can be bounded using Lemma 7 and the induction hypothesis (95e). Combining the two bounds gives \( \| Z^{t_1,w}_1 \|_2 \leq \left( \frac{1}{2} \right)^t \sqrt{\frac{\mu r}{d} G} \) w.h.p. Taking a union bound over \( v \) and \( w \) proves the inequality (95b) for \( t \).

i) The \( \| \cdot \|_{\infty} \) bound (95c): Fix \( w \in [d] \times [d] \). Lemma 6 ensures that w.h.p.

\[
\| Z^{t_1,w}_1 \| = \| (P_T \mathcal{H}_1(Z^{t-1}))_{w_1, w_2} \|_2 \leq \frac{1}{32} \| Z^{t-1,w} \|_\infty + \frac{\mu r}{d} \| Z^{t-1,w} \|_F
\]

We bound \( T_1 \) and \( T_3 \) using the induction hypothesis (95c) and (95e), respectively, and bound \( T_2 \) using inequality (94). Doing so gives \( \| Z^t \|_\infty \leq \left( \frac{1}{2} \right)^t \sqrt{\frac{\mu r}{d} G} \) w.h.p., and taking a union over \( w \) proves the inequality (95c) for \( t \).
j) The \( \| \cdot \|_\infty \) bound (95d): Fix \( v, w \in [d] \times [d] \). We have
\[
\| Z^{t,w}_v \|_D = \left( \left( P_TH_{\Omega}^{(w)}(Z_{t-1}^{t-1,w}) \right)_v \right)
\leq \left( P_TH_{\Omega}^{(w)}(Z_{t-1}^{t-1,w}) \right)_v
\leq \left( P_TH_{\Omega}^{(w)}(Z_{t-1}^{t-1,w}) \right)_v
+ \left( P_TH_{\Omega}^{(w)}(Z_{t-1}^{t-1,w} - Z_{t-1}^{t-1}) \right)_v
\].

The first RHS term can be bounded in a similar way as in
the above proof of (95c). To bound the second RHS term,
we apply Lemma 4 to obtain that \( w.h.p. \)
\[
\left( P_TH_{\Omega}^{(w)}(Z_{t-1}^{t-1,w} - Z_{t-1}^{t-1}) \right)_v
\leq \frac{1}{8} \sqrt{\frac{\mu_r \nu}{d}} \| Z_{t-1}^{t-1,w} - Z_{t-1} \|_D.
\]

Combining the above bounds and the induction hypothe-
sis (95e), we get that \( \| Z^{t,w}_v \| \leq \left( \frac{3}{4} \right)^t \sqrt{\mu \nu} G w.h.p. \) Taking
a union bound over \( v \) and \( w \) proves the inequality (95d) for \( t \).

k) The proximity condition (95e): Fix \( w \in [d] \times [d] \).
We have \( w.h.p. \)
\[
\| Z^t - Z^{t,w}_v \|_D
= \left( P_TH_{\Omega}^{(w)}(Z_{t-1}^{t-1,w}) \right)_v
\leq \left( P_TH_{\Omega}^{(w)}(Z_{t-1}^{t-1,w}) \right)_v
+ \left( P_TH_{\Omega}^{(w)}(Z_{t-1}^{t-1,w}) \right)_v
\leq \frac{1}{8} \| Z_{t-1}^{t-1,w} - Z_{t-1}^{t-1,w} \|_D + \| H_{\Omega}^{(w)}(Z_{t-1}^{t-1,w}) \|_D
\leq \frac{1}{8} \| Z_{t-1}^{t-1,w} - Z_{t-1}^{t-1,w} \|_D
+ \left( \frac{1}{4} \| Z_{t-1}^{t-1,w} \|_{(2,\infty)^2} + \frac{1}{q} \log \frac{d}{q} \| Z_{t-1}^{t-1,w} \|_{\infty} \right)
\]

where we use Proposition 2 in step (a) and Lemma 8 in step
(b). Bounding the last RHS using the induction hypothe-
sis (95f) and the fact that \( q \geq \frac{2 \log d}{\sqrt{\mu \nu}} \), we obtain
\( \| Z^t - Z^{t,w}_v \|_D \leq \left( \frac{3}{4} \right)^t \sqrt{\mu \nu} G w.h.p. \). Taking a union bound over \( v \) and \( w \) proves
the inequality (95e) for \( t \).

We have completed the induction step. Running this argument
for \( t_0 := 2 \log_2 d + 2 \) steps and taking a union bound,
we establish the claimed inequality (91).

APPENDIX A
PROOF OF LEMMAS IN SECTION V
In this section, we prove the technical Lemmas 1, 2 and 3
used in the proof of PGD in Section V.

A. Proof of Lemma 1
Proof: We only prove the first inequality in the lemma.
The other two inequalities can be proved similarly.

We make use of the following known result.

Lemma 9 (1, Lemma 3): Under the setting of Lemma 1,
we have the bounds
\[
\| H \|_{\infty} \leq 1 \quad \text{and} \quad \| H - G \|_{\infty} \leq \frac{1}{\sigma_r \| D \|_{\infty}} \| E \|_{\infty}
\]
and
\[
\| \Lambda^* H - H \Lambda^* \|_{\infty} \leq 2 \| E \|_{\infty}
\]

Returning to the proof of the first inequality in Lemma 1,
we have
\[
\| \Theta = \| \Lambda^* G - G \Lambda^* \|_{\infty}
\]\( (a) \)
\[
\leq \| \Lambda^* G - \Lambda^* H \|_{\infty} + \| H \Lambda^* - G \Lambda^* \|_{\infty}
+ \| H \Lambda^* - \Lambda^* H \|_{\infty}
\]
\[
\leq \| H \Lambda^* \|_{\infty} + \| H + G \Lambda^* \|_{\infty}
+ \| H \Lambda^* - \Lambda^* H \|_{\infty}
\]
\[
\leq \left( 2 + 2 \sigma_1 \frac{1}{\sigma_r \| D \|_{\infty}} \right) \| E \|_{\infty}
\]

where we use the triangle inequality in step (a), the
sub-multiplicative property of the operator norm \( \| \cdot \|_{\infty} \) in step
(b), and Lemma 9 and the assumption \( \| E \|_{\infty} < \frac{1}{2} \sigma_r \) in
step (c).

B. Proof of Lemma 2
Proof: Using the definition of \( \tilde{F} \), we have
\[
\tilde{A} \tilde{F} = \tilde{F} \Lambda \quad \Rightarrow \quad (A + E) \tilde{F} = \tilde{F} \Lambda
\]
\[
\Rightarrow \quad \Lambda_1 F_1^T \tilde{F} + F_2 \Lambda_2 F_2^T \tilde{F} + E \tilde{F} = \tilde{F} \Lambda
\]

Right multiplying the last equation by \( F_1^T \) on both sides,
we get
\[
\Lambda_1 F_1^T \tilde{F} + F_1^T E \tilde{F} = \tilde{F} \tilde{F} \Lambda
\]
\[
\Rightarrow \quad \Lambda_1 H - H \Lambda = - F_1^T E \tilde{F}
\]
\[
\Rightarrow \quad \Lambda_1 H - H \Lambda \|_{\infty} = \| F_1^T E \tilde{F} \|_{\infty}
\]

Consequently, with \( \Theta = \text{diag}(\cos \theta_1, \ldots, \cos \theta_d) \) denoting
the matrix of the principal angles between the column spaces
of \( F \) and \( F_1 \), we obtain
\[
\| \Lambda_1 G - G \Lambda \|_{\infty} \leq \| \Lambda_1 \|_{\infty} \leq \| \Lambda_1 H - H \Lambda \|_{\infty}
\]
\[
+ \| H \Lambda - G \|_{\infty} + \| H \Lambda - H \Lambda \|_{\infty}
\]
\[
\leq \Lambda_1 (A) \| G - H \|_{\infty} + \| F_1^T E \tilde{F} \|_{\infty}
\]
\[
+ \left( \frac{1}{\sigma_r (A) \| D \|_{\infty}} \right) \| G - H \|_{\infty}
\]
\[
\leq \left( 2 \Lambda_1 (A) + \| E \|_{\infty} \right) \| A \|_{\infty} \| \tilde{F} \|_{\infty}
\]
\[
\leq \left( \frac{2 \Lambda_1 (A) + \| E \|_{\infty}}{\sigma_r (A) \| D \|_{\infty}} \right) \| \tilde{F} \|_{\infty}
\]

where step (a) holds because \( \| H - G \|_{\infty} = \| I \cos \Theta \|_{\infty} \leq \| \sin \Theta \| \| \sin \Theta \|_{\infty} \). This proves the Frobenius
norm bound in the lemma. The operator norm bound can be
proved in a similar way.

C. Proof of Lemma 3
Proof: In this proof, we make use of the auxiliary lemmas
given in Appendix C. Here, we define the matrix \( F^* \in \mathbb{R}^{d \times r} \)
to be the one satisfying \( M^* = F^* \otimes F^* \). Note that \( F^* \) may
not have orthonormal columns in this definition.
Let \( M = F \otimes F \) and \( M^+ = F^+ \otimes F^+ \), where \( F, F^+ \in \mathbb{R}^{n \times r} \). Set \( W := M - M^* \) and \( W^+ := M^+ - M^* \).
Also let \( Q := \inf_{Q \in \mathbb{R}^{r \times r}, Q_0^T = 1} \| F^+ - F^* \|_{\infty} \)
and \( Q^+ := \inf_{Q \in \mathbb{R}^{r \times r}, Q_0^T = 1} \| F^+ - F^* \|_{\infty} \). Define \( \Delta := F - F^* \) and
\[ \Delta^+ := F^+ Q - F^* . \] We first record two useful inequalities.

Lemma 15 ensures that
\[
\| \Delta \|_p \leq \frac{3}{\sqrt{\sigma_T}} \| W \|_p \leq 3 \times \frac{1}{64^2} \sqrt{\frac{\sigma_T(M^*)}{\kappa^2}}
\]
and
\[
\| \Delta^+ \|_p \leq 3 \times \frac{1}{\sqrt{\sigma_T}} \| W^+ \|_p \leq 3 \times \frac{1}{64^2} \sqrt{\frac{\sigma_T(M^*)}{\kappa^2}} .
\]

(97)

The differences \( W \) and \( W^+ \) can be expressed as
\[
W = F^* \otimes \Delta + \Delta^+ \otimes F^* + \Delta \otimes \Delta^+ \text{ and } W^+ = F^* \otimes \Delta^+ + \Delta^+ \otimes F^* + \Delta \otimes \Delta^+. 
\]

With these expressions, we decompose the inner product of interest as
\[
\langle W^+, W \rangle = \frac{1}{p} \langle \Omega (W^+), \Omega (W) \rangle
\]
\[
= \left( \langle R_{\Delta, F^*}, R_{\Delta^+, F^*} \rangle - \frac{1}{p} \langle \Omega (R_{\Delta, F^*}), \Omega (R_{\Delta^+, F^*}) \rangle \right)_{(T_1)}
\]
\[
+ \left( \langle R_{\Delta, F^*}, \Delta^+_\perp \rangle - \frac{1}{p} \langle \Omega (R_{\Delta, F^*}), \Omega (\Delta^+_\perp) \rangle \right)_{(T_2)}
\]
\[
+ \left( \langle R_{\Delta^+, F^*}, \Delta_2 \rangle - \frac{1}{p} \langle \Omega (R_{\Delta^+, F^*}), \Omega (\Delta_2) \rangle \right)_{(T_3)}
\]
\[
+ \langle \Delta^+_\perp, \Delta_2 \rangle - \frac{1}{p} \langle \Omega (\Delta^+_\perp), \Omega (\Delta_2) \rangle . \]

(98)

For \( T_1 \), we apply Lemma 17 with some \( \epsilon \leq \frac{1}{\sigma_T} \), which ensures \( w.h.p. \),
\[
|T_1| \leq 4\epsilon \| F^\ast \otimes \Delta \|_p \| F^\ast \otimes \Delta^+ \|_p
\]
\[
\leq 4\epsilon \sigma_1(M^*) \| \Delta \|_p \| \Delta^+ \|_p
\]
\[
\leq \frac{1}{8} \| W^+ \|_p \| W \|_p ,
\]
where the last step follows from the bounds in (97) and letting \( \epsilon = \frac{1}{\sigma_T} \) for some large constant \( c \).

To bound \( T_2, T_3 \), and \( T_4 \), we recall the premise of the proposition that \( \max \{ \| F \|_2, \| F^\ast \|_2 \| F^\ast \otimes F^\ast \|_2, \leq 2\sqrt{\frac{D}{d}} \sigma_1(M^*), \) which implies that \( \max \{ \| F \|_2, \| F^\ast \|_2 \| F^\ast \otimes F^\ast \|_2 \leq \sqrt{\frac{2D \sigma_1(M^* \| d}{d}} . \)

It follows that
\[
\| \Delta \|_2, \| \Delta^+ \|_2 \leq \sqrt{\frac{2D \sigma_1(M^* \| d}{d}} \text{ and similarly } \| \Delta^+ \|_2 \leq \sqrt{\frac{2D \sigma_1(M^* \| d}{d}} .
\]

These bounds allow us to use Lemmas 17 and 18. In particular, for \( T_2, \) letting \( \epsilon = \frac{1}{\sigma_T} \) for a sufficiently large constant \( c \), we have \( w.h.p. \),
\[
|T_2| \leq \| F^\ast \otimes \Delta \|_p \| F^\ast \otimes \Delta^+ \|_p
\]
\[
+ \| (1/\sqrt{p}) \Omega (F^\ast \otimes \Delta \otimes F^\ast) \|_p
\]
\[
\times \| (1/\sqrt{p}) \Omega (\Delta \otimes \Delta^+) \|_p \]
\[
\leq 2 \| F^\ast \|_p \| \Delta \|_p^2 + 4 \| F^\ast \otimes \Delta \|_p
\]
\[
\leq \frac{1}{24} \| W^+ \|_p \| W \|_p ,
\]
where in step (a) we use Lemma 17 for the term \( \| (1/\sqrt{p}) \Omega (F^\ast \otimes \Delta \otimes F^\ast) \|_p \) and Lemma 18 with the above \( \epsilon \) for \( \| (1/\sqrt{p}) \Omega (\Delta \otimes \Delta^+) \|_p \), and in step (b) we use (97) and the above choice of \( \epsilon \). Note that applying Lemma 18 with the above \( \epsilon \) requires \( p \geq \frac{\mu^2 \epsilon^2 (\log d)c^8}{d} \), which is satisfied under the premise of the proposition. A similar argument shows that
\[
|T_3| \leq \frac{1}{24} \| W^+ \|_p \| W \|_p w.h.p.
\]

For \( T_4 \), with the same \( \epsilon \) as above and applying Lemma 18, we have \( w.h.p. \),
\[
|T_4| \leq \frac{1}{24} \| W^+ \|_p \| W \|_p
\]
\[
\leq \frac{1}{24} \| W^+ \|_p \| W \|_p ,
\]

where the last step follows from the above choice of \( \epsilon \) and the bounds (97).

Combing the above bounds on \( T_1, T_4 \), we obtain that
\[
\left| \langle W^+, W \rangle - \frac{1}{p} \langle \Omega (W^+), \Omega (W) \rangle \right|
\]
\[
\leq \sum_{i=1}^{4} |T_i| \leq \frac{1}{4} \| W^+ \|_p \| W \|_p ,
\]
thereby completing the proof of Lemma 3. \( \square \)

APPENDIX B
PROOF OF LEMMAS IN SECTION VI

In this section, we prove the technical Lemmas 4–8 used in the proof of NN in Section VI. For the first four lemmas, we prove the bounds for \( H_{\Omega} \) only; the bounds for \( H_{\Omega}^{(-w)} \) can be proved similarly.

A. Proof of Lemma 4

Proof: For each fixed \((i, j) \in [d] \times [d] \), we have
\[
e_{ij}^T [P_T H_{\Omega}(Z)] e_j = (P_T H_{\Omega}(Z), e_{ij}^T)
\]
\[
\stackrel{(a)}{=} (P_T H_{\Omega} P_T (Z), e_{ij}^T)
\]
\[
= (Z, P_T H_{\Omega} P_T (e_{ij}^T)) ,
\]

where \( P_T \) is the projection onto the subspace spanned by the columns of \( T \).

Proof of Lemma 5

Proof: For each fixed \((i, j) \in [d] \times [d] \), we have
\[
e_{ij}^T [P_T H_{\Omega}(Z)] e_j = (P_T H_{\Omega}(Z), e_{ij}^T)
\]
\[
\stackrel{(a)}{=} (P_T H_{\Omega} P_T (Z), e_{ij}^T)
\]
\[
= (Z, P_T H_{\Omega} P_T (e_{ij}^T)) ,
\]

where \( P_T \) is the projection onto the subspace spanned by the columns of \( T \).

Proof of Lemma 6

Proof: For each fixed \((i, j) \in [d] \times [d] \), we have
\[
e_{ij}^T [P_T H_{\Omega}(Z)] e_j = (P_T H_{\Omega}(Z), e_{ij}^T)
\]
\[
\stackrel{(a)}{=} (P_T H_{\Omega} P_T (Z), e_{ij}^T)
\]
\[
= (Z, P_T H_{\Omega} P_T (e_{ij}^T)) ,
\]

where \( P_T \) is the projection onto the subspace spanned by the columns of \( T \).

Proof of Lemma 7

Proof: For each fixed \((i, j) \in [d] \times [d] \), we have
\[
e_{ij}^T [P_T H_{\Omega}(Z)] e_j = (P_T H_{\Omega}(Z), e_{ij}^T)
\]
\[
\stackrel{(a)}{=} (P_T H_{\Omega} P_T (Z), e_{ij}^T)
\]
\[
= (Z, P_T H_{\Omega} P_T (e_{ij}^T)) ,
\]

where \( P_T \) is the projection onto the subspace spanned by the columns of \( T \).

Proof of Lemma 8

Proof: For each fixed \((i, j) \in [d] \times [d] \), we have
\[
e_{ij}^T [P_T H_{\Omega}(Z)] e_j = (P_T H_{\Omega}(Z), e_{ij}^T)
\]
\[
\stackrel{(a)}{=} (P_T H_{\Omega} P_T (Z), e_{ij}^T)
\]
\[
= (Z, P_T H_{\Omega} P_T (e_{ij}^T)) ,
\]

where \( P_T \) is the projection onto the subspace spanned by the columns of \( T \).
where the step (a) is due to $Z \in \mathcal{T}$. Applying the Cauchy-Schwarz inequality, we obtain that w.h.p.

$$\langle Z, \mathcal{P}_T \mathcal{H}_T (e_{v_j}^T) \rangle \leq \|Z\|_2 \|\mathcal{P}_T \mathcal{H}_T (e_{v_j}^T)\|_2 \leq \frac{1}{16} \|Z\|_2 \|\mathcal{P}_T (e_{v_j})\|_2 \leq \frac{1}{8} \sqrt{\frac{\mu r}{d}} \|Z\|_2,$$

where the inequality (b) holds because $\|\mathcal{P}_T \mathcal{H}_T \mathcal{P}_T \|_w \leq \frac{1}{16}$ w.h.p. by Proposition 2, and the last inequality follows from direct computation using the definition $\mathcal{P}_T$.

**B. Proof of Lemma 5**

**Proof:** We first recorded two useful identities:

$$UU^T \mathcal{P}_T (Z) = UU^T (UU^T Z + ZVV^T - UU^T ZVV^T) = UU^T Z,$$

$$\mathcal{P}_T (Z) VV^T = (UU^T Z + ZVV^T - UU^T ZVV^T) VV^T = ZVV^T.$$

(99a)

Now fix $v \in [d] \times [d]$. By definition of $\mathcal{P}_T$ we have

$$\| (\mathcal{P}_T \mathcal{H}_T (Z^{-1} v), \|_2 \leq \| e_{v_1}^T UU^T \mathcal{H}_T (Z^{-1} v) \|_2 + \| e_{v_1}^T \mathcal{H}_T (Z^{-1} v) VV^T \|_2

\leq 2 \| e_{v_1}^T UU^T \mathcal{H}_T (Z^{-1} v) \|_2 + \| e_{v_1}^T \mathcal{H}_T (Z^{-1} v) VV^T \|_2

\leq 2 \| e_{v_1}^T UU^T \mathcal{H}_T (Z^{-1} v) \|_2 + \| e_{v_1}^T \mathcal{H}_T (Z^{-1} v) VV^T \|_2$$

(99b)

For $T_1$, we have w.h.p.

$$\frac{1}{2} \leq \| e_{v_1}^T UU^T \mathcal{H}_T (Z^{-1} v) \|_2 + \| e_{v_1}^T \mathcal{H}_T (Z^{-1} v) VV^T \|_2 \leq \frac{1}{64} \sqrt{\frac{\mu r}{d}} \|Z^{-1}\|_2,$$

where we use $Z^{-1} \in \mathcal{T}$ in step (a), the identity (99a) in step (b), and Proposition 2 in step (c).

For $T_2$, note that $Z^{1,v}$ is independent of $\{\delta_{v,j}, j \in [d]\}$. Conditioning on $Z^{1,v}$, we write $T_2$ as sum of independent vectors:

$$T_2 = \| e_{v_1}^T [\mathcal{H}_T (Z^{-1} v)] V \|_2 \leq \sum_{1 \leq j \leq d} \| (1 - q^{-1} \delta_{v,j}) Z_{v,j}^{-1,v} V \|_2,$$

(100)

We compute the bounds

$$\| (1 - q^{-1} \delta_{v,j}) Z_{v,j}^{-1,v} V \|_2 \leq \frac{1}{q} \| Z^{-1,v} \|_\infty \|V\|_2 \leq \frac{1}{q} \sqrt{\frac{\mu r}{d}} \| Z^{-1,v} \|_\infty =: B,$$

$$\| \sum_j \mathbb{E} \left[ (1 - q^{-1} \delta_{v,j}) Z_{v,j}^{-1,v} V_j \right] \|_2 \leq \frac{1}{1} \| Z^{-1,v} \|_\infty \|V_j\|_2 \leq \frac{\mu r}{qd} \| Z^{-1,v} \|_\infty^2 =: \sigma^2.$$
\[ \leq |e_{v_1}^T UUT \mathcal{H}_\Omega(Z^{t-1,v})e_{v_2}| \]
\[ + \frac{1}{64} \sqrt{\frac{\mu r}{d}} \|Z^{t-1} - Z^{t-1,v}\|_F, \]

where we use the equality (99a) in step (a), \( Z^{t-1}, Z^{t-1,v} \in \mathcal{T} \) in step (b), and Proposition 2 in step (c). To proceed, note that \( Z^{t-1,v} \) is independent of \( \{\delta_{kv_2}, k \in [d]\} \) by construction. Therefore, we have w.h.p.,

\[ |e_{v_1}^T UUT \mathcal{H}_\Omega(Z^{t-1,v})e_{v_2}| \]
\[ = \sum_k (UUT)_{v_1,k} |1 - \frac{1}{q} \delta_{kv_2}| Z^{t-1,v} \]
\[ \leq c \sqrt{\frac{\log d}{d}} \sqrt{\frac{\mu r}{d}} \|Z^{t-1,v}\|_\infty + \frac{\log d \mu r}{q} \|Z^{t-1,v}\|_\infty \]
\[ \leq \frac{1}{64} \|Z^{t-1,v}\|_\infty, \]

where we use Bernstein’s inequality (Lemma 10) in step (a), and \( q \geq \frac{\mu r \log d}{d} \) in step (b). Thus, \( T_1 \) satisfies

\[ T_1 \leq \frac{1}{64} \|Z^{t-1,v}\|_\infty + \frac{1}{64} \sqrt{\frac{\mu r}{d}} \|Z^{t-1} - Z^{t-1,v}\|_F \text{ w.h.p.} \]

By a similar argument, the same bound holds for \( T_2 \). For \( T_3 \), we have w.h.p.,

\[ T_3 \leq \|e_{v_1}^T UUT \mathcal{P}_\mathcal{T} \mathcal{H}_\Omega \mathcal{P}_\mathcal{T}(Z^{t-1})VVT e_{v_2}| \]
\[ \leq \|e_{v_1}^T(UUT)\|_2 \|\mathcal{P}_\mathcal{T} \mathcal{H}_\Omega \mathcal{P}_\mathcal{T}(Z^{t-1})\|_F \|VVT e_{v_2}\|_2 \]
\[ \leq \sqrt{\frac{\mu r}{d}} \cdot \frac{1}{64} \|Z^{t-1}\|_F \sqrt{\frac{\mu r}{d}}, \]

where we use the equality (99a) in step (a), and Proposition 2 in step (b). Combining the above bounds over all \( v \in [d] \times [d] \), proves the lemma.

**D. Proof of Lemma 7**

**Proof:** Fix \( j \in [d] \). Since \( Z \in \mathcal{T} \), we have

\[ \|\mathcal{P}_\mathcal{T} \mathcal{H}_\Omega(Z)e_j\|_2 = \sup_{\|x\|_2=1} \langle |\mathcal{P}_\mathcal{T} \mathcal{H}_\Omega(Z)e_j|, x \rangle \]
\[ = \sup_{\|x\|_2=1} \langle \mathcal{P}_\mathcal{T} \mathcal{H}_\Omega \mathcal{P}_\mathcal{T}(Z), xe_j^T \rangle \]
\[ \leq \|Z\|_F \sup_{\|x\|_2=1} \|\mathcal{P}_\mathcal{T} \mathcal{H}_\Omega \mathcal{P}_\mathcal{T}(xe_j^T)\|_F, \]

Bounding the last RHS using Proposition 2, we obtain that w.h.p.,

\[ \|\mathcal{P}_\mathcal{T} \mathcal{H}_\Omega(Z)e_j\|_2 \leq \frac{1}{64} \|Z\|_F \sup_{\|x\|_2=1} \|xe_j^T\|_F \]
\[ = \frac{1}{64} \|Z\|_F. \]

The same bound holds for \( \|e_i^T \mathcal{P}_\mathcal{T} \mathcal{H}_\Omega(Z)\|_2 \) for each \( i \in [d] \) by a similar argument. The lemma then follows from a union bound over \( i \in [d] \) and \( j \in [d] \).

**E. Proof of Lemma 8**

**Proof:** Fix \( w \in [d] \times [d] \) and \( Z \in \mathcal{T} \). By definition of \( \mathcal{P}_\mathcal{T} \) and the fact that \( \|U\|_w \leq 1, \|V\|_w \leq 1 \), we have

\[ \|\mathcal{P}_\mathcal{T} \mathcal{H}_\Omega^{(w)}(Z)\|_F \]
\[ = \|U^T \mathcal{H}_\Omega^{(w)}(Z)(I + VVT) + \mathcal{H}_\Omega^{(w)}(Z)VVT\|_F \]
\[ \leq 2\|U^T \mathcal{H}_\Omega^{(w)}(Z)\|_F + \|\mathcal{H}_\Omega^{(w)}(Z)V\|_F. \]

Below we bound the first RHS term; the second term can bounded similarly. Since only the \( w_1 \)-th row and \( w_2 \)-th column of \( \mathcal{H}_\Omega^{(w)}(Z) \) are non-zero, we have

\[ \|U^T \mathcal{H}_\Omega^{(w)}(Z)\|_F \leq \|U^T \mathcal{H}_\Omega^{(w)}(Z)\|_{w_2} 2 \]
\[ + \|U^T \mathcal{H}_\Omega^{(w)}(Z)\|_{w_1} 2 \]
\[ \leq \left( \sum_{i=1}^d U_{i}(1 - q^{-1} \delta_{iv_2})Z_{iv_2} \right) T_1 \]
\[ + \sqrt{\frac{\mu r}{d}} \|\mathcal{H}_\Omega^{(w)}(Z)\|_{w_1} 2. \]

Note that \( T_1 \) is the sum of independent vectors and has the same form as the term \( T_2 \) in equation (100) in the proof of Lemma 5. Following the same arguments therein, we obtain that w.h.p.,

\[ T_1 \leq \frac{1}{256} \sqrt{\frac{\log d}{q}} \|Z\|_\infty + \frac{1}{256} \|Z\|_{(2,\infty)^2}. \]

To bound \( T_2 \), define the operator \( \mathcal{P}_w : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}^{d \times d} \) by \( \mathcal{P}_w(Z) = Z_{ij} 1 \{i = w_1 \text{ or } j = w_2\} \). We have w.h.p.,

\[ T_2 \leq \left( \sqrt{\frac{\mu r}{d}} \|\mathcal{H}_\Omega^{(w)}(Z)\|_w \right) \]
\[ + \left( \frac{\mu r}{d} \|\mathcal{H}_\Omega^{(w)}(Z)\|_{w_2} \right) \]
\[ + \left( \frac{\mu r}{d} \right) \|\mathcal{H}_\Omega^{(w)}(Z)\|_{w_1} \]
\[ \leq \frac{1}{256} \sqrt{\frac{\log d}{q}} \|Z\|_\infty + \frac{1}{256} \|Z\|_{(2,\infty)^2}, \]

where step (a) follows from the inequality \( \|A\|_{(2,\infty)^2} \leq \|A\|_w \), \( \forall A \), step (b) follows from \( \mathcal{H}_\Omega^{(w)} = \mathcal{H}_\Omega \mathcal{P}_w \), and step (c) follows from Lemma 16. Combining the bounds for \( T_1 \) and \( T_2 \), we get that w.h.p.,

\[ \|U^T \mathcal{H}_\Omega^{(w)}(Z)\|_F \leq \frac{1}{128} \sqrt{\frac{\log d}{q}} \|Z\|_\infty + \frac{1}{128} \|Z\|_{(2,\infty)^2}. \]

Plugging this inequality into the bound for \( \|\mathcal{P}_\mathcal{T} \mathcal{H}_\Omega^{(w)}(Z)\|_F \), we prove the lemma.
APPENDIX C

AUXILIARY LEMMAS

In this section, we record several technical lemmas that are used in the proofs of our main theorem.

A. Standard Concentration and Perturbation Bounds

The lemmas in this subsection are standard concentration and matrix perturbation inequalities.

Lemma 10 (Bernstein): Let $X_1, \ldots, X_n$ be $n$ independent random variables with $|X_i| \leq B$ with mean 0. For each $t > 0$ we have

$$\Pr(\sum_{i=1}^{n} X_i \geq t) \leq 2 \exp\left(-\frac{t^2}{2nB^2}\right).$$

Lemma 11 (Vector Bernstein [18, Theorem 11]): Let $\{v_k\}$ be a finite sequence of independent $d$ dimensional random vectors. Suppose that $E[v_k] = 0$ and $\|v_k\|_2 \leq B$, a.s., and put $\sigma \geq \sqrt{\sum_k \|v_k\|_2^2}$. Then for all $t \geq 0$,

$$\Pr(\sum_k v_k \geq t) \leq (n+1) \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Lemma 12 (Matrix Bernstein [33]): Consider a finite sequence $\{Z_k\}$ of independent $n_1 \times n_2$ random matrices that satisfy $E[Z_k] = 0$ and $\|Z_k\|_F \leq D$ a.s. Let $\sigma^2$ be the maximum of $\sqrt{\sum_k \|Z_kZ_k^T\|_F}$ and $\sqrt{\sum_k \|Z_k^2\|_F}$. Then for all $t \geq 0$ we have

$$\Pr(\sum_k Z_k \geq t) \leq (n_1 + n_2) \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Lemma 13 (Subspace Distance Equivalence [34, Proposition 2.2]): If the matrices $V_1, V_2 \in \mathbb{R}^{d \times r}$ have orthonormal columns, then

$$\frac{1}{2} \inf_{Q \in \mathbb{R}^{d \times r}, QQ^T = I} \|V_1 - V_2Q\|_2^2$$

$$\leq \|\sin(V_1, V_2)\|_2^2$$

$$\leq \inf_{Q \in \mathbb{R}^{d \times r}, QQ^T = I} \|V_1 - V_2Q\|_2^2,$$

where $(V_1, V_2)$ denotes the principal angles between the column spaces of $V_1$ and $V_2$.

Lemma 14 (Davis-Kahan sin $\Theta$ Theorem [12, 27]): Suppose that $A, W \in \mathbb{R}^{d \times d}$ are symmetric matrices, and $A = A + W$. Let $\delta = \lambda_r(A) - \lambda_{r+1}(A)$ be the gap between the top $r$-th and $(r+1)$-th eigenvalues of $A$, and $U, \bar{U}$ be matrices whose columns are the $r$ leading orthonormal eigenvectors of $A$ and $A$ respectively. If $\delta > \|W\|_\infty$, then for any unitarily invariant norm $\|\cdot\|$, we have

$$\|\sin(U, \bar{U})\| \leq \|WU\|_\infty \frac{\delta}{\delta - \|W\|_\infty}.$$  

Consequently, by Lemma 13 there exists a matrix $O \in \mathbb{R}^{r \times r}$ satisfying $OO^T = I$ and

$$\|U - \bar{O}U\|_r \leq \sqrt{\frac{2\|WU\|_\infty}{\delta - \|W\|_\infty}},$$

Lemma 15 (16, Lemma 6]): Given matrices $F, F^+ \in \mathbb{R}^{d \times r}$, let $M = F \otimes F$ and $M^+ = F^+ \otimes F^+$. Also let

$$\Delta = F - F^+O^*$$

where $O^* = \arg \min_{O \in \mathbb{R}^{d \times r}, O \otimes O^T = I} \|F - F^+O\|_r$. We have

$$\|\Delta \otimes \Delta\|_r \leq 2\|M - M^+\|_r^2,$$

and

$$\sigma_r(M^+\|\Delta\|_r \leq \frac{1}{2(\sqrt{3} - 1)} \|M - M^+\|_r^2.$$
Let us first bound the matrices $R_i$, $i = 1, 2, 3$ in terms of their infinity norms. We have
\[
\begin{align*}
&\|R_1\|_\infty \leq \|\Delta\|_{2,\infty} \|F\|_{2,\infty}\|A\|_\infty, \\
&\|R_2\|_\infty \leq \|\Delta\|_{2,\infty} \|F^*\|_{2,\infty}\|\Lambda\|_\infty, \\
&\|R_3\|_\infty \leq \|F^*\|_{2,\infty}\|\Lambda - \Lambda^*\|_\infty \|F\|_{2,\infty}, \\
&\quad \text{where we use Wely's inequality in step (a). (a) also bound}
\end{align*}
\]

where step (a) follows from Lemma 17 and step (c) holds since $F^*$ is orthonormal with $\|F\|_\infty = 1$.

Lemma 21: In the setting of SMC($M^*, p$) or MC($M^*, p$), there exists a numerical constant $c > 0$ such that if $p \geq \frac{\log d}{d}$, then for each fixed matrix $Z \in \mathbb{R}^{d \times d}$, the following inequalities hold w.h.p.:
\[
\|\mathcal{H}_\Omega^{(-i)}(Z)\|_\infty \leq \|\mathcal{H}_\Omega(Z)\|_\infty \leq 2c \sqrt{d} \log d \|Z\|_\infty.
\]

Proof: Since the $i$-th column and $i$-th row of $\mathcal{H}_\Omega^{(-i)}(Z)$ is zero, we have $\|\mathcal{H}_\Omega^{(-i)}(Z)\|_\infty \leq \|\mathcal{H}_\Omega(Z)\|_\infty$. Thus it remains to bound $\|\mathcal{H}_\Omega(Z)\|_\infty$. Under MC($M^*, p$), such a bound has been established in the literature using the matrix Bernstein inequality (Lemma 12); see, e.g., [3, Lemma 3.1] and [8, Lemma 12] for the proof. The proof under the symmetric setting SMC($M^*, p$) follows the same lines; we omit the details here.

Lemma 22 (Uniform version of Lemma 21): In the setting of SMC($M^*, p$) or MC($M^*, p$), there exists a numerical constant $c > 0$ such that if $p \geq \frac{\log d}{d}$, then w.h.p. the following bounds hold:
\[
\begin{align*}
&\|\mathcal{H}_\Omega(U \otimes V)\|_\infty \leq 2c \sqrt{d} \log d \|U\|_{2,\infty} \|V\|_{2,\infty}, \\
&\forall U, V \in \mathbb{R}^{d \times r}, \\
&\|\mathcal{H}_\Omega(A)\|_\infty \leq 2c \sqrt{d} \log d \|A\|_\infty, \\
&\forall A \in \mathbb{R}^{d \times d} : \text{rank}(A) \leq r, \\
&\|\mathcal{H}_\Omega^{(-i)}(A)\|_\infty \leq 2c \sqrt{d} \log d \|A\|_\infty, \\
&\forall A \in \mathbb{R}^{d \times d} : \text{rank}(A) \leq r, \forall i \in [d].
\end{align*}
\]

Proof: The first and last inequalities in the lemma are immediate consequence of the second inequality. In particular,
the second inequality follows from noting that $\sqrt{r} \leq r$ and $\|U \otimes V\|_\infty \leq \|U\|_2 \|V\|_2$. The last inequality follows from the fact $\mathcal{H}_\Omega^{(i-1)}$ sets the $i$-th row and column of $\mathcal{H}_\Omega(A)$ to 0, hence $\|\mathcal{H}_\Omega^{(i-1)}(A)\|_\infty \leq \|\mathcal{H}_\Omega(A)\|_\infty$. It remains to prove the second inequality in the lemma.

Since $\text{rank}(A) = r$, we have the decomposition $A = U \otimes V$ where $U, V \in \mathbb{R}^{d \times r}$. Let $u_i = (u_i^{(1)}, \ldots, u_i^{(r)})$ and $v_j = (v_j^{(1)}, \ldots, v_j^{(r)})$ be the $i$-th row and $j$-th row of $U$ and $V$, respectively. We make use of the variational representation of the spectral norm:

$$\|\mathcal{H}_\Omega(U \otimes V)\|_\sigma = \sup_{\|a\|_2 = \|b\|_2 = 1} \langle \mathcal{H}_\Omega(U \otimes V), a \otimes b \rangle.$$ 

Recalling the definition $\mathcal{H}_\Omega := I - \frac{1}{p}\Pi_\Omega$, we have

$$\langle \mathcal{H}_\Omega(U \otimes V), a \otimes b \rangle = \langle U \otimes V, a \otimes b \rangle - \frac{1}{p} \langle \Pi_\Omega(U \otimes V), a \otimes b \rangle = \sum_{i,j} (u_i, v_j) a_i b_j - \frac{1}{p} \sum_{i,j \in \Omega} (u_i, v_j) a_i b_j = \langle (1 \otimes 1, (U \otimes V) \circ (a \otimes b)) \rangle - \frac{1}{p} \langle \Pi_\Omega(1 \otimes 1), (U \otimes V) \circ (a \otimes b) \rangle,$$

where $\circ$ denotes the Hadamard product. It follows that

$$\|\mathcal{H}_\Omega(U \otimes V)\|_\sigma \leq \|\mathcal{H}_\Omega(1 \otimes 1)\|_\sigma \sup_{\|a\|_2 = \|b\|_2 = 1} \|(U \otimes V) \circ (a \otimes b)\|_\infty.$$ 

On the one hand, Lemma 21 applied to the fixed matrix $Z = 1 \otimes 1$ guarantees that $\|\mathcal{H}_\Omega(1 \otimes 1)\|_\sigma \leq 2\sqrt{d \log(d)/p}$ w.h.p. On the other hand, note that $\|(U \otimes V) \circ (a \otimes b)\|_{i,j} = (a_i u_i, b_j v_j)$, so the matrix $(U \otimes V) \circ (a \otimes b)$ has rank at most $r$. It follows that

$$\|(U \otimes V) \circ (a \otimes b)\|_\infty \leq \sqrt{r} - \|(U \otimes V) \circ (a \otimes b)\|_F = \sqrt{r} \sum_{i,j} (a_i b_j)^2 (u_i, v_j)^2 \leq \sqrt{r} \sum_{i,j} (a_i b_j)^2 \max_{i,j} |u_i, v_j| = \sqrt{r} \cdot 1 \cdot \|U \otimes V\|_\infty.$$ 

Combining pieces, we establish the second inequality in the lemma.

\section*{Acknowledgment}

Y. Chen would like to thank Y. Chen for inspiring discussion. The authors thank the anonymous reviewer for his/her constructive comments.

\section*{References}

[1] E. Abbe, J. Fan, K. Wang, and Y. Zhong, “Entrywise eigenvector analysis of random matrices with low expected rank,” 2017, arXiv:1709.09565. [Online]. Available: http://arxiv.org/abs/1709.09565.

[2] M.-F. Balcan, Y. Liang, D. P. Woodruff, and H. Zhang, “Matrix completion and related problems via strong duality,” in LIPICS-Leibniz International Proceedings in Informatics, vol. 94, Wadern, Germany: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[3] E. J. Candès, X. Li, Y. Ma, and J. Wright, “Robust principal component analysis?” J. ACM, vol. 58, no. 3, p. 11, May 2011.

[4] E. J. Candès and B. Recht, “Exact matrix completion via convex optimization,” Found. Comput. Math., vol. 9, no. 6, p. 717, 2009.

[5] E. J. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” IEEE Trans. Inf. Theory, vol. 56, no. 5, pp. 2053–2080, May 2010.

[6] J. Chen, D. Liu, and X. Li, “Nonconvex rectangular matrix completion via gradient descent without $\ell_2,\infty$ regularization,” 2019, arXiv:1901.06116. [Online]. Available: http://arxiv.org/abs/1901.06116.

[7] Y. Chen, “Incoherence-optimal matrix completion,” IEEE Trans. Inf. Theory, vol. 61, no. 5, pp. 2909–2923, May 2015.

[8] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis, “Low-rank matrix recovery from errors and erasures,” IEEE Trans. Inf. Theory, vol. 59, no. 7, pp. 4324–4337, Jul. 2013.

[9] Y. Chen and M. J. Wainwright, “Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees,” 2015, arXiv:1509.03025. [Online]. Available: http://arxiv.org/abs/1509.03025.

[10] Y. Chen, Y. Chi, J. Fan, and C. Ma, “Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval,” Math. Program., vol. 176, nos. 1–2, pp. 5–37, Jul. 2019.

[11] Y. Chen, J. Fan, C. Ma, and K. Wang, “Spectral method and regularized MLE are both optimal for top-$K$ ranking,” 2017, arXiv:1707.09971. [Online]. Available: http://arxiv.org/abs/1707.09971.

[12] C. Davis and W. M. Kahan, “The rotation of eigenvectors by a perturbation. III,” SIAM J. Numer. Anal., vol. 7, no. 1, pp. 1–46, Mar. 1970.

[13] L. Ding and Y. Chen, “Leave-one-out approach for matrix completion: Primal and dual analysis,” 2018, arXiv:1803.07554. [Online]. Available: http://arxiv.org/abs/1803.07554.

[14] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu, “On robust regression with high-dimensional predictors,” Proc. Natl. Acad.Sci. USA, vol. 110, no. 36, pp. 14557–14562, Sep. 2013.

[15] J. Fan, D. Wang, K. Wang, and Z. Zhu, “Distributed estimation of principal eigenspaces,” 2017, arXiv:1702.06488. [Online]. Available: http://arxiv.org/abs/1702.06488.

[16] R. Ge, C. Jin, and Y. Zheng, “No spurious local minima in nonconvex low rank problems: A unified geometric analysis,” 2017, arXiv:1704.00708. [Online]. Available: http://arxiv.org/abs/1704.00708.

[17] R. Ge, J. D. Lee, and T. Ma, “Matrix completion has no spurious local minimum,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 2973–2981.

[18] D. Gross, “Recovering low-rank matrices from few coefficients in any basis,” IEEE Trans. Inf. Theory, vol. 57, no. 3, pp. 1548–1566, Mar. 2011.

[19] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, “Implicit regularization in matrix factorization,” in Proc. Inf. Theory Appl. Workshop (ITA), Feb. 2018, pp. 6151–6159.

[20] M. Hardt and M. Wooters, “Fast matrix completion without the condition number,” in Proc. Conf. Learn. Theory, 2014, pp. 638–678.

[21] P. Jain, R. Meka, and I. S. Dhillon, “Guaranteed rank minimization via singular value projection,” in Proc. Adv. Neural Inf. Process. Syst., 2010, pp. 937–945.

[22] P. Jain and P. Netrapalli, “Fast exact matrix completion with finite samples,” in Proc. Conf. Learn. Theory, 2015, pp. 1007–1034.

[23] P. Jain, P. Netrapalli, and S. Sanghavi, “Low-rank matrix completion using alternating minimization,” in Proc. 45th Annu. ACM Symp. Symp. Theory Comput. STOC, 2013, pp. 665–674.

[24] A. Javanmard and A. Montanari, “De-biasing the lasso: Optimal sample size for Gaussian designs,” 2015, arXiv:1508.02757. [Online]. Available: http://arxiv.org/abs/1508.02757.

[25] R. H. Keshavan, A. Montanari, and S. Oh, “Matrix completion from a few entries,” IEEE Trans. Inf. Theory, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[26] H. Kushner and G. G. Yin, Stochastic Approximation and Recursive Algorithms and Applications, vol. 35. Springer, 2003.

[27] R.-C. Li, “Relative perturbation theory: II. Eigenspace and singular subspace variations,” SIAM J. Matrix Anal. Appl., vol. 20, no. 2, pp. 471–492, Jan. 1998.
[28] Y. Li, T. Ma, and H. Zhang, “Algorithmic regularization in overparameterized matrix sensing and neural networks with quadratic activations,” in Proc. Conf. Learn. Theory, 2018, pp. 2–47.

[29] C. Ma, K. Wang, Y. Chi, and Y. Chen, “Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution,” 2017, arXiv:1711.10467. [Online]. Available: http://arxiv.org/abs/1711.10467

[30] B. Recht, “A simpler approach to matrix completion,” J. Mach. Learn. Res., vol. 12, pp. 3413–3430, Jan. 2011.

[31] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” SIAM Rev., vol. 52, no. 3, pp. 471–501, Jan. 2010.

[32] R. Sun and Z.-Q. Luo, “Guaranteed matrix completion via non-convex factorization,” IEEE Trans. Inf. Theory, vol. 62, no. 11, pp. 6535–6579, Nov. 2016.

[33] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” Found. Comput. Math., vol. 12, no. 4, pp. 389–434, Aug. 2012.

[34] V. Q. Yu and J. Lei, “Minimax sparse principal subspace estimation in high dimensions,” Ann. Statist., vol. 41, no. 6, pp. 2905–2947, Dec. 2013.

[35] Q. Zheng and J. Lafferty, “Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent,” 2016, arXiv:1605.07051. [Online]. Available: http://arxiv.org/abs/1605.07051

[36] Y. Zhong and N. Boumal, “Near-optimal bounds for phase synchronization,” 2017, arXiv:1703.06605. [Online]. Available: http://arxiv.org/abs/1703.06605

Lijun Ding received the B.S. degree in mathematics and economics from the Hong Kong University of Science and Technology in 2014 and the M.S. degree in statistics from The University of Chicago in 2016. He is currently pursuing the Ph.D. degree with the School of Operations Research and Information Engineering, Cornell University. His research interests include large-scale optimization and statistical learning.

Yudong Chen received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, and the Ph.D. degree in electrical and computer engineering from The University of Texas at Austin in 2013. From 2013 to 2015, he was a Post-Doctoral Scholar with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley. He is currently an Assistant Professor with the School of Operations Research and Information Engineering, Cornell University. His research interests include machine learning, high-dimensional and robust statistics, convex and nonconvex optimization, and applications in networks and financial systems.