Subspace Estimation from Unbalanced and Incomplete Data Matrices: $\ell_{2,\infty}$ Statistical Guarantees

Changxiao Cai\(^*\)  Gen Li\(^\dagger\)  Yuejie Chi\(^\ddagger\)  H. Vincent Poor\(^*\)  Yuxin Chen\(^*\)

October 11, 2019

Abstract
This paper is concerned with estimating the column space of an unknown low-rank matrix $\mathbf{A}^* \in \mathbb{R}^{d_1 \times d_2}$, given noisy and partial observations of its entries. There is no shortage of scenarios where the observations — while being too noisy to support faithful recovery of the entire matrix — still convey sufficient information to enable reliable estimation of the column space of interest. This is particularly evident and crucial for the highly unbalanced case where the column dimension $d_2$ far exceeds the row dimension $d_1$, which is the focal point of the current paper.

We investigate an efficient spectral method, which operates upon the sample Gram matrix with diagonal deletion. We establish statistical guarantees for this method in terms of both $\ell_2$ and $\ell_{2,\infty}$ estimation accuracy, which improve upon prior results if $d_2$ is substantially larger than $d_1$. To illustrate the effectiveness of our findings, we develop consequences of our general theory for three applications of practical importance: (1) tensor completion from noisy data, (2) covariance estimation with missing data, and (3) community recovery in bipartite graphs. Our theory leads to improved performance guarantees for all three cases.

Keywords: spectral method, principal component analysis, tensor completion, covariance estimation, missing data, spectral clustering, leave-one-out analysis

Contents

1 Introduction .......................... 2

2 Problem formulation .............. 4
  2.1 Models ................................ 4
  2.2 Notations ............................ 6

3 Main results ......................... 6
  3.1 A spectral method with diagonal deletion .......... 6
  3.2 Theoretical guarantees .............. 7
  3.3 Consequences for specific models ............... 9
    3.3.1 Symmetric tensor completion .............. 10
    3.3.2 Covariance estimation with missing data ........ 12
    3.3.3 Community recovery in bipartite stochastic block models .......... 15
  3.4 Numerical experiments ............ 16

\begin{itemize}
  \item Corresponding author: Yuxin Chen.
  \item \(^*\)Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA; email: \{ccai,poor,yuxin.chen\}@princeton.edu.
  \item \(^\dagger\)Department of Electronic Engineering, Tsinghua University, Beijing, China 100086, USA; email: g-li16@mails.tsinghua.edu.cn.
  \item \(^\ddagger\)Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; email: yuejiechi@cmu.edu.
\end{itemize}
Consider the problem of estimating the column space of a low-rank matrix $A^* = [A_{i,j}^*]_{1 \leq i \leq d_1, 1 \leq j \leq d_2}$, based on noisy and highly incomplete observations of its entries. To set the stage, suppose that we observe

$$A_{i,j} = A_{i,j}^* + N_{i,j}, \quad \forall (i,j) \in \Omega,$$

where $\Omega \subseteq \{1, \cdots, d_1\} \times \{1, \cdots, d_2\}$ is the sampling set, and $A_{i,j}$ denotes the observed entry at location $(i,j)$, which is corrupted by noise $N_{i,j}$. In contrast to the classical matrix completion problem that aims
to fill in all missing entries [CR09, KMO10, CLC19, CC18b], the current paper focuses solely on estimating the column space of $A^*$, which is oftentimes a less stringent requirement. A problem of this kind arises in numerous applications. We immediately give several representative examples as follows.

- **Symmetric tensor completion and estimation.** This problem, which seeks to estimate a low-rank symmetric tensor from partial observations of its entries [BM16, XYZ17, MS18], spans various applications like visual data inpainting [LMWY13] and medical imaging [SHKM14]. Consider, for example, a symmetric order-3 tensor $T^*$ of canonical polyadic (CP) rank $r$:

$$
T^* = \sum_{s=1}^{r} w_s^* \otimes w_s^* \otimes w_s^* := \sum_{s=1}^{r} (w_s^*)^{\otimes 3} \in \mathbb{R}^{d \times d \times d},
$$

where $\{w_s^* \in \mathbb{R}^d\}_{s=1}^{r}$ represents a collection of tensor factors.\(^1\) An alternative representation of $T^*$ can be obtained by unfolding the tensor of interest into the following $d \times d^2$ matrix

$$
A^* = \sum_{s=1}^{r} w_s^* (w_s^* \otimes w_s^*)^\top \in \mathbb{R}^{d \times d^2}. \quad (2)
$$

Consequently, estimation of the subspace spanned by $\{w_s^* \}_{s=1}^{r}$ from partial noisy entries of $T^*$ — which serves as a common and crucial step for tensor completion [MS18] — is equivalent to estimating the column space of $A^*$ from incomplete data. Notably, the unfolded matrix $A^* \in \mathbb{R}^{d \times d^2}$ becomes extremely fat as the dimension $d$ grows.

- **Covariance estimation with missing data.** Suppose we have available a sequence of $n$ independent sample vectors $\{x_i \in \mathbb{R}^d\}_{i=1}^{n}$, drawn from a certain low-dimensional subspace. To be more specific, suppose that

$$
x_i = B^* f_i^* + \eta_i, \quad 1 \leq i \leq n,
$$

where $B^* \in \mathbb{R}^{d \times r}$ encodes the $r$-dimensional principal subspace underlying the data (sometimes referred to as the factor loading matrix in factor models [LM62, FWZZ18]), $f_i^* \sim \mathcal{N}(0, I_r)$ represents some random coefficients, and $\eta_i$ stands for some independent random noise. Several fundamental statistical models fall in the same vein of this model, including, for example, the generalized spiked model [BY12]. An important task amounts to estimating the subspace spanned by $B^*$, or equivalently, the $r$-dimensional principal subspace of the covariance matrix of $x_i$. If we write $F^* = [f_1^*, \cdots, f_n^*] \in \mathbb{R}^{r \times n}$ and $N = [\eta_1, \cdots, \eta_n] \in \mathbb{R}^{d \times n}$, then it boils down to estimating the column space of the data matrix

$$
\mathbf{X} := B^* F^* + N \in \mathbb{R}^{d \times n}.
$$

Further, this task becomes more challenging in the presence of missing data (where we only get to see highly incomplete entries of $\mathbf{X}$), which is clearly subsumed by the problem described in (1). Of particular interest is a regime where a substantial amount of data are missing (which may yield a fat data matrix $\mathbf{X}$), such that the individual sample vectors cannot possibly be recovered; this effectively allows one to protect the privacy of individual samples. Fortunately, one might still hope to faithfully estimate the principal subspace of the covariance matrix, provided that a sufficiently large number of sample vectors are queried.

- **Community recovery in bipartite graphs.** Community recovery is often concerned with clustering a collection of individuals or nodes into different communities, on the basis of the interactions between pairs of nodes. In many complex networks, such pairwise interactions might only occur when the two nodes involved belong to two disjoint groups. This requires community recovery in bipartite networks, which is sometimes referred to as biclustering [Dhi01, LCJ14, AHB14, ZA18]. To illuminate the connection between

\(^1\)Here and throughout, for any vectors $a, b, c \in \mathbb{R}^d$, we denote by $a \otimes b \otimes c$ a $d \times d \times d$ array whose $(i, j, k)$-th entry is given by $a_{ij} b_{jk}$. In addition, we let $a \otimes b := \begin{bmatrix} a_{11} b & \cdots & a_{1d} b \\ \vdots & \ddots & \vdots \\ a_{d1} b & \cdots & a_{dd} b \end{bmatrix}$ represent a $d^2$-dimensional vector.
biclustering and the problem considered herein, consider a bipartite stochastic block model (BSBM) with two disjoint groups of nodes $\mathcal{U}$ and $\mathcal{V}$. Suppose that the nodes in $\mathcal{U}$ (resp. $\mathcal{V}$) form two clusters. For each pair of nodes $(i, j) \in (\mathcal{U}, \mathcal{V})$, there is an edge connecting them with probability depending only on the community memberships of $i$ and $j$. As it turns out, if we denote by $A \in \mathbb{R}^{d_1 \times d_2}$ the bi-adjacency matrix of the observed random bipartite graph or its centered version, then $A^* := \mathbb{E}[A]$ exhibits a low-rank structure (as we shall elaborate momentarily). Perhaps more importantly, the column subspace of $A^*$ reveals the community memberships of all nodes in $\mathcal{U}$. As a result, this biclustering problem is tightly connected to subspace estimation given noisy observations of a low-rank matrix. In particular, when the size of $\mathcal{V}$ is substantially larger than that of $\mathcal{U}$, one might encounter a situation where only the nodes in $\mathcal{U}$ (rather than those in $\mathcal{V}$) can be reliably clustered. This calls for development of “one-sided” community recovery algorithms, that is, the type of algorithms that guarantee reliable clustering of $\mathcal{U}$ without worrying about the clustering accuracy in $\mathcal{V}$.

Since we concentrate primarily on estimating the column space of $A^*$, it is natural to expect a lower sample complexity as well as a weaker requirement on the noise, in comparison to the conditions required for reliable estimation/completion of the whole matrix—particularly for those highly unbalanced problems with drastically different dimensions $d_1$ and $d_2$. Focusing on a spectral method applied to the Gram matrix $AA^\top$ with diagonal deletion, the present paper establishes statistical guarantees in terms of the sample complexity and the estimation accuracy, both of which improve significantly upon the results supported by prior matrix completion theory. Our theory allows for appealing estimation error bounds measured by the $\ell_{2,\infty}$ norm (i.e. the maximum error among all rows of the matrix), which were previously rarely available. Finally, we develop concrete consequences of our general theory for all three applications mentioned above, leading to improved performance guarantees compared to prior literature.

The rest of this paper is organized as follows. Section 2 formulates the problem and introduces basic definitions and notations. In Section 3, we present our main theory and apply it to the aforementioned applications. Numerical experiments are also provided at the end of this section. Section 4 provides an overview of related prior works. The proof of our main theory is outlined in Section 5, with the proofs of many auxiliary lemmas postponed to the appendix. We conclude the paper with a discussion of future directions in Section 6.

## 2 Problem formulation

### 2.1 Models

**Low-rank matrix.** Suppose that the unknown matrix $A^* \in \mathbb{R}^{d_1 \times d_2}$ is rank-$r$, where the row dimension $d_1$ and the column dimension $d_2$ are allowed to be drastically different. Assume that the (compact) singular value decomposition (SVD) of $A^*$ is given by

$$A^* = U^* \Sigma^* V^* \top = \sum_{i=1}^r \sigma_i^* u_i^* v_i^* \top. \quad (3)$$

Here, $\sigma_1^* \geq \sigma_2^* \geq \cdots \geq \sigma_r^* > 0$ represent the $r$ nonzero singular values of $A^*$, and $\Sigma^* \in \mathbb{R}^{r \times r}$ is a diagonal matrix whose diagonal entries are given by $\{\sigma_1^*, \cdots, \sigma_r^*\}$. The columns of $U^* = [u_1^*, \cdots, u_r^*] \in \mathbb{R}^{d_1 \times r}$ (resp. $V^* = [v_1^*, \cdots, v_r^*] \in \mathbb{R}^{d_2 \times r}$) are orthonormal, which are the top-$r$ left (resp. right) singular vectors of $A^*$. We define and denote the condition number of $A^*$ as follows

$$\kappa := \frac{\sigma_1^*}{\sigma_r^*}, \quad (4)$$

and take

$$d := \max \{d_1, d_2\}. \quad (5)$$

**Incoherence.** Further, we impose certain incoherence conditions on the unknown matrix $A^*$, which are commonly adopted in the matrix completion literature (e.g. [CR09, KMO10, CLC19]).
Definition 1 (Incoherence parameters). Define the incoherence parameters $\mu_0$, $\mu_1$ and $\mu_2$ as follows

$$\mu_0 := \max_{1 \leq i \leq d_1, 1 \leq j \leq d_2} \frac{|A_{i,j}^*|^2}{\|A^*\|_F^2},$$

$$\mu_1 := \max_{r \leq i \leq d_1} \frac{\|U^*e_i\|_2^2}{d_1} \quad \text{and} \quad \mu_2 := \max_{r \leq i \leq d_2} \frac{\|V^*e_i\|_2^2}{d_2},$$

where $e_i$ is the $i$-th standard basis vector of compatible dimensionality.

Intuitively, when $\mu_0$, $\mu_1$ and $\mu_2$ are all small, the energies of the matrices $A^*$, $U^*$ and $V^*$ are (nearly) evenly spread out across all entries, rows, and columns. For notational simplicity, we shall set

$$\mu := \max \{\mu_0, \mu_1, \mu_2\}.$$  

Random sampling and random noise. Suppose that we have only collected noisy observations of the entries of $A^*$ over a sampling set $\Omega \subseteq \{1, \ldots, d_1\} \times \{1, \ldots, d_2\}$. Specifically, we observe

$$A_{i,j} = \begin{cases} A_{i,j}^* + N_{i,j}, & (i,j) \in \Omega, \\ 0, & \text{else}, \end{cases}$$

where $N_{i,j}$ denotes the noise at location $(i,j)$. For notational simplicity, we shall write

$$A = \mathcal{P}_\Omega(A) = \mathcal{P}_\Omega(A^*) + \mathcal{P}_\Omega(N),$$

where $\mathcal{P}_\Omega$ represents the Euclidean projection onto the subspace of matrices supported on $\Omega$. In addition, this paper concentrates on random sampling and random noise as follows.

Assumption 1 (Random sampling). Each $(i,j)$ is included in the sampling set $\Omega$ independently with probability $p$.

Assumption 2 (Random noise). The noise $N_{i,j}$'s are independent random variables and satisfy the following conditions: for each $1 \leq i \leq d_1, 1 \leq j \leq d_2$,

1. (Zero mean) $\mathbb{E}[N_{i,j}] = 0$;
2. (Variance) $\text{Var}(N_{i,j}) \leq \sigma^2$;
3. (Magnitude) Each $N_{i,j}$ satisfies either of the following condition:
   
   (a) $|N_{i,j}| \leq R$;
   
   (b) $N_{i,j}$ has a symmetric distribution satisfying $\mathbb{P}\{|N_{i,j}| > R\} \leq c_ie^{-\frac{R^2}{\sigma^2}}$ for some universal constant $c_i > 0$.

Here, $R$ is some quantity obeying

$$\frac{R^2}{\sigma^2} \leq C_i \min \{p\sqrt{d_1d_2}, pd_2\} \log d$$

for some universal constant $C_i > 0$.

As a remark, Assumption 2 allows the largest possible size $R$ of each noise component to be substantially larger than its typical size $\sigma$. For example, if $p \approx 1$, then $R$ can be $\min \{(d_1d_2)^{1/4}, \sqrt{d_2}\}$ times larger than $\sigma$ (ignoring any logarithmic factor). In addition, the noise $N_{i,j}$'s do not necessarily have identical variance; in fact, our formulation allows us to accommodate the heteroscedasticity of noise (i.e. the scenario where the noise has location-varying variance).
Goal. Given incomplete and noisy observations about \( A^* \in \mathbb{R}^{d_1 \times d_2} \) (cf. (9)), we seek to estimate \( U^* \in \mathbb{R}^{d_1 \times r} \) modulo some global rotation. We emphasize once again that the aim here is not to estimate the entire matrix. In truth, there are many unbalanced cases with \( d_2 \gg d_1 \) such that (1) reliable estimation of \( U^* \) is feasible, but (2) faithful estimation of the whole matrix \( A^* \) is information theoretically impossible.

2.2 Notations
We denote \([n] := \{1, \cdots, n\}\). For any matrix \( A \in \mathbb{R}^{d_1 \times d_2}\), we use \( \sigma_i(A) \) and \( \lambda_i(A) \) to represent the i-th largest singular value and the i-th largest eigenvalue of \( A \), respectively. Let \( A_{i,:} \) and \( A_{:,j} \) denote respectively the i-th row and the j-th column of \( A \). Let \( \|A\| \) (resp. \( \|A\|_p \)) represent the spectral norm (resp. the Frobenius norm) of \( A \). We also denote by \( \|A\|_{2,\infty} := \max_{i \in [d_1]} \|A_{i,:}\|_2 \) and \( \|A\|_\infty := \max_{i \in [d_1], j \in [d_2]} |A_{i,j}| \) the \( \ell_{2,\infty} \) norm and the entrywise \( \ell_\infty \) norm of \( A \), respectively. Similarly, for any tensor \( T \), we use \( \|T\|_\infty \) to represent the largest magnitude of the entries of \( T \). Moreover, we denote by \( \mathcal{P}_{\text{diag}} \) the projection onto the subspace that vanish outside the diagonal, and define \( \mathcal{P}_{\text{off-diag}} \) such that \( \mathcal{P}_{\text{off-diag}}(A) := A - \mathcal{P}_{\text{diag}}(A) \). Let \( O^{r \times r} \) stand for the set of \( r \times r \) orthonormal matrices. In addition, we use \( \text{diag}(a) \) to represent a diagonal matrix whose \((i,i)\)-th entry is equal to \( a_i \). Throughout this paper, the notations \( C,C_1,\cdots,c,c_1,\cdots \) denote absolute positive constants whose values may change from line to line.

Furthermore, for any real-valued functions \( f(d_1,d_2) \) and \( g(d_1,d_2) \), \( f(d_1,d_2) \lesssim g(d_1,d_2) \) or \( f(d_1,d_2) = O(g(d_1,d_2)) \) means that \( \frac{f(d_1,d_2)}{g(d_1,d_2)} \leq C_1 \) for some constant \( C_1 > 0 \); \( f(d_1,d_2) \gtrsim g(d_1,d_2) \) means that \( \frac{f(d_1,d_2)}{g(d_1,d_2)} \geq C_2 \) for some universal constant \( C_2 > 0 \); \( f(d_1,d_2) \asymp g(d_1,d_2) \) means that \( C_1 \leq \frac{f(d_1,d_2)}{g(d_1,d_2)} \leq C_2 \) for some universal constants \( C_1, C_2 > 0 \); \( f(d_1,d_2) \approx o(g(d_1,d_2)) \) means that \( f(d_1,d_2) = \mathcal{O}(g(d_1,d_2)) \) as \( \min\{d_1,d_2\} \to \infty \). In addition, \( f(d_1,d_2) \ll g(d_1,d_2) \) (resp. \( f(d_1,d_2) \gg g(d_1,d_2) \)) means that there exists some sufficiently small (resp. large) constant \( c_1 > 0 \) (resp. \( c_2 > 0 \)) such that \( f(d_1,d_2) \leq c_1 g(d_1,d_2) \) (resp. \( f(d_1,d_2) \geq c_2 g(d_1,d_2) \)) holds true for all sufficiently large \( d_1 \) and \( d_2 \).

3 Main results
3.1 A spectral method with diagonal deletion
Recall that \( A = [A_{i,j}]_{1 \leq i \leq d_1, 1 \leq j \leq d_2} \) is the zero-padded data matrix (see (9)). It is easily seen that, under our random sampling model (i.e. Assumption 1), \( p^{-1} A \) serves as an unbiased estimator of \( A^* \). One might thus expect the left singular subspace of \( A \) to form a reasonably good estimator of the subspace spanned by \( U^* \). As it turns out, when \( A^* \) is a very fat matrix (namely, \( d_2 \gg d_1 \)), this approach might fail to work when the sample complexity is not sufficiently large or when the noise size is not sufficiently small.

This paper adopts an alternative route by resorting to the sample Gram matrix \( AA^\top \) (properly rescaled). Straightforward calculation reveals that
\[
\frac{1}{p^2} \mathbb{E} [AA^\top] = A^* A^* + \left( \frac{1}{p} - 1 \right) \mathcal{P}_{\text{diag}} (A^* A^* \top) + \frac{1}{p} \text{diag} \left[ \sum_{j=1}^{d_2} \text{Var}(N_{i,j}) \right]_{1 \leq i \leq d_1},
\]
where \( \text{diag}(a) \) with \( a \in \mathbb{R}^{d_1} \) represents a diagonal matrix whose \((i,i)\)-th entry equals \( a_i \). The identity (12) implies that the diagonal components of \( p^{-2} \mathbb{E} [AA^\top] \) are significantly inflated, which might need to be properly suppressed.

In order to remedy the above-mentioned diagonal inflation issue, we adopt a simple strategy that zeros out all diagonal entries; that is, performing the spectral method on the following matrix
\[
G = \frac{1}{p^2} \mathcal{P}_{\text{off-diag}} (AA^\top)
\]
with \( \mathcal{P}_{\text{off-diag}}(M) := M - \mathcal{P}_{\text{diag}}(M) \) denoting projection onto the set of zero-diagonal matrices. This clearly satisfies
\[
\mathbb{E} [G] = \mathcal{P}_{\text{off-diag}} (A^* A^* \top) = \mathcal{P}_{\text{off-diag}} (U^* \Sigma^* U^* \top).
\]
If the diagonal entries of $A^* A^{*\top}$ are not too large, then one has $A^* A^{*\top} \approx P_{\text{off-diag}}(A^* A^{*\top})$ and, as a result, the rank-$r$ eigen-subspace of $G$ might form a reliable estimate of the subspace spanned by $U^*$. The procedure is summarized in Algorithm 1.

\begin{algorithm}
\textbf{Algorithm 1} The spectral method on the diagonal-deleted Gram matrix

1: \textbf{Input}: sampling set $\Omega$, observed entries $\{A_{i,j}\mid (i,j) \in \Omega\}$, sampling rate $p$, rank $r$. \\
2: \textbf{Compute} the (truncated) rank-$r$ eigen-decomposition $U\Lambda U^\top$ of $G$, where $U \in \mathbb{R}^{d_1 \times r}$, $\Lambda \in \mathbb{R}^{r \times r}$, and $G := P_{\text{off-diag}}(\frac{1}{p^2}AA^\top)$. \hspace{1cm} (14)

Here, $A$ is defined in (10) and $P_{\text{off-diag}}(M)$ zeros out the diagonal entries of a matrix $M$.

3: \textbf{Output} $U$ as the subspace estimate, and $\Sigma = \Lambda^{1/2}$ as the spectrum estimate.
\end{algorithm}

We note that this is clearly not a new algorithmic idea. In fact, proper handling of the diagonal entries (e.g., diagonal deletion, diagonal reweighting) has already been recommended in multiple different contexts, including bipartite stochastic block models [FP16], covariance estimation [Lou13, Lou14, LW12, EvdG19], tensor completion [MS18], to name just a few.

### 3.2 Theoretical guarantees

In general, one can only hope to estimate $U^*$ up to global rotation. With this in mind, we introduce the following rotation matrix $R := \arg \min_{Q \in \mathcal{O}^{r \times r}} \|UQ - U^*\|_F$, \hspace{1cm} (15)

where $\mathcal{O}^{r \times r}$ stands for the set of $r \times r$ orthonormal matrices. In words, $R$ is the global rotation matrix that best aligns $U$ and $U^*$. Equipped with this notation, the following theorem delivers upper bounds on the difference between the obtained estimate $U$ and the ground truth $U^*$. The proof is postponed to Section 5.

\begin{theorem}
Assume that the following conditions hold

$$p \geq c_0 \max \left\{ \frac{\mu_1 r \log d}{\sqrt{d_1 d_2}}, \frac{\mu_2 \sqrt{r} \log d}{d_2} \right\}, \quad \sigma_r \leq c_1 \min \left\{ \frac{\sqrt{p}}{r \sqrt{d_1 d_2} \sqrt{\log d}}, \frac{1}{\kappa}, \frac{1}{\kappa^2 \log d} \right\} \quad \text{and} \quad r \leq \frac{c_2 d_1}{\mu_1 \kappa^2}, \hspace{1cm} (16)$$

where $c_0 > 0$ is some sufficiently large constant and $c_1, c_2 > 0$ are some sufficiently small constants. Then with probability at least $1 - O\left(d^{-10}\right)$, the matrices $U$ and $\Sigma$ returned by Algorithm 1 satisfy

$$\|UR - U^*\| \lesssim \mathcal{E}_\text{gen}, \hspace{1cm} (17a)$$

$$\|UR - U^*\|_{2,\infty} \lesssim \kappa^2 \sqrt{\frac{\mu r d_1}{d_2}} \cdot \mathcal{E}_\text{gen}, \hspace{1cm} (17b)$$

$$\|\Sigma - \Sigma^*\| \lesssim \sigma_r \cdot \mathcal{E}_\text{gen}, \hspace{1cm} (17c)$$

where $R$ is defined in (15), and

$$\mathcal{E}_\text{gen} := \frac{\mu_2 \sqrt{r} \log d}{\sqrt{d_1 d_2} p} + \frac{\mu_2 \sqrt{r} \log d}{d_2 p} + \frac{\sigma_r^2 \sqrt{d_1 d_2} \log d}{p} + \frac{\sigma_r \sqrt{d_2 \log d}}{p} + \frac{\mu_1 \kappa^2 r}{d_1}. \hspace{1cm} (18)$$

\end{theorem}

\begin{remark}
If there is no missing data, namely, $p = 1$, then Theorem 1 continues to hold if the first two terms on the right-hand side of (18) are removed.

In a nutshell, Theorem 1 asserts that Algorithm 1 produces reliable estimates of the column subspace of $A^*$ — with respect to both the spectral norm and the $\|\cdot\|_{2,\infty}$ norm — under certain conditions imposed on
Table 1: The dominant term of the noise effect in $\sigma^2 \sqrt{d_1 d_2} \frac{1}{p}$ if $d_2 \geq d_1$ (omitting logarithmic factors and assuming $r, \kappa, \mu \approx 1$).

| | large-noise regime (i.e. $\sigma/\sigma^* \gtrsim \sqrt{p/d_2}$) | small-noise regime (i.e. $\sigma/\sigma^* \lesssim \sqrt{p/d_2}$) |
|---|---|---|
| dominant term | $\frac{\sigma^2}{\sigma^*} \frac{\sqrt{d_1 d_2}}{p}$ | $\frac{\sigma^2}{\sigma^*} \frac{1}{\sqrt{d_2 p}}$ |

the sample size and the noise size. For instance, consider the settings where $\mu, \kappa \approx 1$ and $r \ll d_1 \leq d_2$: as long as the following condition holds:

$$p \gg \frac{r \log^2 d}{d_1 d_2} \quad \text{and} \quad \frac{\sigma^2}{\sigma^*^2} \ll \frac{p}{d_1 d_2 \log d};$$  

(19)

the proposed spectral method achieves consistent estimation with high probability, namely,

$$\min_{Q \in \mathbb{O}^{r \times r}} \frac{\|UQ - U^*\|}{\|U^*\|} \ll 1, \quad \min_{Q \in \mathbb{O}^{r \times r}} \frac{\|UQ - U^*\|_{2,\infty}}{\|U^*\|_{2,\infty}} \ll 1 \quad \text{and} \quad \frac{\|\Sigma - \Sigma^*\|}{\|\Sigma^*\|} \ll 1. \quad \text{(20)}$$

Our upper bound (18) on the spectral norm error contains five terms. The first two terms of (18) represent the influence of observation noise; the third and the fourth terms of (18) represent the noise effect; and the last term of (18) arises due to the bias caused by diagonal deletion. In particular, the last term is expected to be vanishingly small in the low-rank and incoherent case. Interestingly, both the missing data effect and the noise effect are captured by two different terms, which we shall interpret in what follows. Note that a primary focus of this paper is to demonstrate the feasibility of obtaining a tight control of the $\ell_{2,\infty}$ statistical error. This is particularly evident for the low-rank, incoherent, and well-conditioned case with $r, \mu, \kappa = O(1)$, in which our theory (cf. (17a) and (17b)) reveals that the $\ell_{2,\infty}$ error can be a factor of $\sqrt{d_1}$ smaller than the spectral norm error. The discussion below focuses on this case (namely, $r, \mu, \kappa = O(1)$), with all logarithmic factors omitted for simplicity of presentation.

- Let us first examine the influence of observation noise, which reads

$$\frac{\sigma^2}{\sigma^*^2} \frac{\sqrt{d_1 d_2}}{p} + \frac{\sigma}{\sigma^*^2} \sqrt{\frac{d_1}{p}}.$$

This contains a quadratic term as well as a linear term w.r.t. $\sigma/\sigma^*$. To interpret this, consider, for example, the case without missing data (i.e. $p = 1$) and decompose

$$AA^T = A^*A^{*\top} + A^*N^{*\top} + N^T A^* + NN^T,$$

(omitting logarithmic factors)

which clearly explains why eigenspace perturbation bounds depend both linearly and quadratically on the noise magnitudes. In general, the quadratic term $\frac{\sigma^2}{\sigma^*^2} \sqrt{d_1 d_2} \frac{1}{p}$ is dominant when the signal-to-noise ratio (SNR) is not large enough; as the noise decreases to a sufficiently low level, the linear term starts to enter the picture. See Table 1 for a more precise summary.

- Next, we examine the influence of missing data and assume $\sigma = 0$ to simplify the discussion. If we view $N_{\text{missing}} = \frac{1}{p} A - A^*$ as a zero-mean perturbation matrix, then one can write

$$\frac{1}{p^2} AA^T = A^*A^{*\top} + A^*N_{\text{missing}}^{*\top} + N_{\text{missing}} A^{*\top} + N_{\text{missing}}N_{\text{missing}}^{*\top}.$$

Similar to the above noisy case with $p = 1$, this decomposition explains why the influence of missing data contains two terms as well (see Table 2)

$$\frac{1}{\sqrt{d_1 d_2 p}} + \frac{1}{\sqrt{d_2 p}}.$$
Table 2: The dominant term of the missing data effect in $\frac{1}{\sqrt{d_1 d_2 p}} + \frac{1}{\sqrt{d_2 p}}$ if $d_2 \geq d_1$ (omitting logarithmic factors and assuming $r, \kappa, \mu \approx 1$).

### Comparison with prior results

To demonstrate the effectiveness of our bounds, we take a moment to compare them with several prior results. Once again, the discussion below focuses on the case with $\max\{\mu, \kappa, r\} \approx 1$.

- To begin with, we compare our spectral norm bound with that required for matrix completion [KMO10, CT10, CCF19, CLC19] in the noise-free case (i.e. $\sigma = 0$). Suppose that $d_2 \geq d_1$. As is well known, for both spectral algorithms and optimization-based methods, the sample complexities required for faithful matrix completion need to satisfy $pd_1 d_2 \gtrsim d_2 \text{poly log } d$. In comparison, our algorithm enables faithful estimation of the column subspace under the sample size $pd_1 d_2 \gtrsim \sqrt{d_1 d_2} \text{ poly log } d$. This confirms that the sample complexity required for subspace estimation might be much lower than that required for matrix completion (i.e. by a factor of $\sqrt{d_2/d_1}$). In other words, to ensure faithful recovery of the column space, the number of observations per column only needs to satisfy $pd_1 \gtrsim \sqrt{d_1 / d_2} \text{ poly log } d$, which decreases as the number of columns $d_2$ grows.

- Next, we compare our $\| \cdot \|_{2,\infty}$ bound with the theoretical guarantees derived in [AFWZ17] for the case obeying $d_2 \gtrsim d_1 \text{ log }^2 d$. The theory in [AFWZ17, Theorem 3.4] requires the sample size and the noise to satisfy $p \gtrsim d_1^{-1} \text{ log } d$ and $\sigma / \sigma^* \lesssim \sqrt{\frac{p}{d_2 \text{ log } d}}$, both of which are more stringent requirements than ours (namely, $p \gtrsim \frac{\text{log}^2 d}{d_1 d_2}$ and $\sigma / \sigma^* \lesssim \frac{\text{log } d}{d_2 \text{ log } d}$). To be more precise, our sample size requirement improves upon [AFWZ17] by a factor of $d_1 \text{ log }^2 d$ and our noise condition improves upon [AFWZ17] by a factor of $(d_2/d_1)^{1/4}$. Again, this arises because [AFWZ17] seeks to estimate the whole matrix as opposed to its column subspace.

- We then compare our results with [MS18], which studies a diagonal-reshcaling algorithm for the noise-free case (i.e. $\sigma = 0$). Combining [MS18, Theorem 6.2] with the standard Davis-Kahan matrix perturbation theory, we can easily see that their spectral norm bound for subspace estimation reads

$$\frac{\text{poly log } d}{\sqrt{d_1 d_2 p}} + \frac{\text{poly log } d}{\sqrt{d_2 p}}.$$

This coincides with our bound except for the last term of (18) (due to the bias incurred by diagonal deletion). In comparison, our theory offers additional $\ell_2,\infty$ statistical guarantees and covers the noisy case, thus strengthening the theory presented in [MS18].

- Additionally, we compare our spectral norm bound with the results derived in [ZCW18]. Consider the noiseless case where $\sigma = 0$. It is proven in [ZCW18, Theorem 6] (see also the remark that follows) that: if the sample size satisfies $pd_1 d_2 \gtrsim \max\{d_1^{1/3} d_2^{2/3}, d_1\} \text{ poly log } d$, then the HeteroPCA estimator is consistent in estimating the column subspace (namely, achieving a relative $\ell_2$ estimation error not exceeding $o(1)$). In comparison, our theory claims that Algorithm 1 is guaranteed to yield consistent column subspace estimation as long as the sample size obeys $pd_1 d_2 \gtrsim \sqrt{d_1 d_2} \text{ poly log } d$. Consequently, if we omit logarithmic terms, then our sample complexity improves upon the theoretical support of HeteroPCA by a factor of $(d_2/d_1)^{1/6}$ if $d_2 \geq d_1$.

### 3.3 Consequences for specific models

We showcase the consequence of Theorem 1 in three concrete applications previously introduced in Section 1 in relatively simple settings, with the purpose of highlighting the broad applicability of our main results, rather than striving for full generality.
3.3.1 Symmetric tensor completion

We begin by considering the problem of symmetric tensor completion. Suppose that the unknown order-3 tensor $T^* = \sum_{s=1}^r (w_s^*)^{\otimes 3} \in \mathbb{R}^{d \times d \times d}$ is symmetric with CP rank $r$. The goal is to estimate the subspace spanned by $\{w_s^*\}_{s=1}^r$, on the basis of the noisy tensor $T = [T_{i,j,k}]_{1 \leq i,j,k \leq d}$ obeying

$$T_{i,j,k} = \begin{cases} T_{i,j,k}^* + N_{i,j,k}, & (i,j,k) \in \Omega, \\ 0, & (i,j,k) \notin \Omega, \end{cases}$$

(21)

where $T_{i,j,k}$ is the observed entry in location $(i,j,k)$, $N_{i,j,k}$ is the associated independent random noise satisfying Assumption 2, and $\Omega \subseteq [d]^3$ stands for a sampling set obtained via uniform random sampling with sampling rate $p$ (namely, each entry is observed independently with probability $p$). Algorithm 2 summarizes the spectral method for tensor completion based on our general algorithm.

**Algorithm 2** The spectral method for tensor completion

1: **Input:** sampling set $\Omega$, observed entries $\{T_{i,j,k} \mid (i,j,k) \in \Omega\}$, sampling rate $p$, CP-rank $r$.
2: Let $A \in \mathbb{R}^{d \times d^2}$ be the mode-1 matricization of the observed tensor $T$ (see (21)), namely, set $A_{i,(j-1)d+k} = T_{i,j,k}$ for each $(i,j,k) \in [d]^3$, and employ $A$ as the input of Algorithm 1.
3: **Output** $U \in \mathbb{R}^{d \times r}$ returned by Algorithm 1 as the subspace estimate.

In order to provide theoretical support for this algorithm, we introduce a few notations. First, we introduce the following quantities

$$\kappa_{tc} := \frac{\lambda_{21}^{\max}}{\lambda_{21}^{\min}}, \quad \lambda_{21}^{\min} := \min_{1 \leq i \leq r} \|w_i^*\|_2^3, \quad \lambda_{21}^{\max} := \max_{1 \leq i \leq r} \|w_i^*\|_2^3.$$  

(22)

Informally, $\kappa_{tc}$ captures the condition number of the unknown tensor. Additionally, similar to matrix completion, we impose the following incoherence assumptions that enable efficient tensor completion:

**Assumption 3** (Incoherence). Suppose that the tensor $T^*$ and its tensor factors $\{w_s^*\}_{s=1}^r$ satisfy

$$\|T^*\|_\infty \leq \sqrt{\frac{\mu_4}{d^3}} \|T^*\|_F, \quad \text{(23a)}$$

$$\|w_i^*\|_\infty \leq \sqrt{\frac{\mu_4}{d}} \|w_i^*\|_2, \quad 1 \leq i \leq r \quad \text{(23b)},$$

$$\langle w_i^*, w_j^* \rangle \leq \sqrt{\frac{\mu_5}{d}} \|w_i^*\|_2 \|w_j^*\|_2, \quad 1 \leq i \neq j \leq r \quad \text{(23c)}$$

with incoherence parameters $\mu_3, \mu_4, \mu_5 > 0$.

For notational convenience, we also set

$$\mu_{tc} := \max \{\mu_3, \mu_4^2\}. \quad \text{(24)}$$

Given that the tensor factors $\{w_s^*\}_{1 \leq s \leq r}$ are in general not orthogonal to each other, we introduce the following orthonormal matrix $U^* \in \mathbb{R}^{d \times r}$ to represent the subspace spanned by $\{w_s^*\}_{1 \leq s \leq r}$:

$$U^* := W^*(W^*^TW^*)^{-1/2}, \quad W^* := [w_1^*, \cdots, w_r^*] \in \mathbb{R}^{d \times r}. \quad \text{(25)}$$

With these in place, we are now ready to quantify the estimation error of this spectral algorithm. The proof is deferred to Appendix A.1.

**Corollary 1** (Symmetric tensor completion). Consider the above tensor completion model. There exist some universal constants $c_0, c_1, c_2 > 0$ such that if

$$p \geq c_0 \max \left\{ \frac{\mu_{tc}\kappa_{tc}^2R\log^2 d}{d^{3/2}}, \frac{\mu_{tc}\kappa_{tc}^2R\log^2 d}{d^2} \right\}, \quad \frac{\sigma}{\lambda_{21}^{\min}} \leq c_1 \min \left\{ \frac{\sqrt{p}}{\kappa_{tc}\sqrt{\log d}} \frac{1}{\kappa_{tc}} \right\}$$

$$\frac{1}{\kappa_{tc}} \sqrt{d} \log d \right\}$$

$$10$$
and

\[ r \leq c_2 \min \left\{ \frac{d}{\kappa_{tc}^2 \mu_4}, \frac{1}{\kappa_{tc}^2} \sqrt{\frac{d}{\mu_5}} \right\}, \]

then with probability exceeding \( 1 - O(d^{-10}) \), Algorithm 2 yields

\[ \|UR - U^*\| \lesssim \mathcal{E}_{tc}, \quad (26a) \]
\[ \|UR - U^*\|_{2,\infty} \lesssim \kappa_{tc}^2 \sqrt{\frac{\mu_{tc}r}{d}} \mathcal{E}_{tc}, \quad (26b) \]

where \( R := \arg \min_{Q \in \mathcal{O}^{n \times r}} \|UQ - U^*\|_F \) and

\[ \mathcal{E}_{tc} := \frac{\mu_{tc} \kappa_{tc}^2 r \log d}{d^{3/2} p} + \sqrt{\frac{\mu_{tc} \kappa_{tc}^2 r \log d}{d^2 p}} + \frac{\sigma^2}{\lambda_{\min}^2} \frac{d^{3/2} \log d}{p} + \frac{\sigma \kappa_{tc}}{\lambda_{\min}^2} \sqrt{\frac{d \log d}{p}} + \frac{\mu_4 \kappa_{tc}^2 r}{d}. \quad (27) \]

As discussed in several related work (e.g. [XYZ17, MS18, CLPC19]), once we obtain reliable estimates of the subspace spanned by the tensor factors, we can further exploit the tensor structure to estimate the unknown tensor. Indeed, in many tensor completion algorithms, subspace estimation serves as a crucial initial step for tensor completion. Moreover, while prior works only provide \( \ell_2 \) estimation error bounds, Corollary 1 further delivers \( \ell_{2,\infty} \) statistical guarantees, which reflect a stronger sense of statistical accuracy. Further, Corollary 1 also reveals that the estimation errors are spread out across all rows of the subspace estimate.

It is widely recognized that spectral methods are often unable to attain optimal estimation accuracy; for instance, when \( \sigma = 0 \) and \( p \neq 1 \), the spectral method fails to achieve perfect recovery. More often than not, spectral methods are employed to produce a rough initial estimate that outperforms the random guess, which can then be refined via other algorithms (e.g. nonconvex optimization algorithms like gradient descent and alternating minimization [JO14, XY17, CLPC19]). Consequently, our discussion below focuses on understanding the sample size and the signal-to-noise (SNR) required for achieving consistent estimation (namely, obtaining an \( o(1) \) relative estimation error). For convenience of presentation, we again focus on the low-rank, incoherent, and well-conditioned case with \( r, \mu, \kappa \ll 1 \). In this case, our results in Corollary 1 indicate that

\[ \min_{Q \in \mathcal{O}^{n \times r}} \|UQ - U^*\| = o(1) \quad \text{and} \quad \min_{Q \in \mathcal{O}^{n \times r}} \|UQ - U^*\|_{2,\infty} = o(1/\sqrt{d}) \quad (28) \]

with high probability, provided that the sample size and the noise satisfy

\[ p \gtrsim \frac{\log^2 d}{d^{3/2}} \quad \text{and} \quad \frac{\sigma}{\lambda_{\min}^2} \lesssim \sqrt{\frac{p}{d^{3/2} \log d}}. \quad (29) \]

Several remarks are in order.

- **Sample complexity.** It is widely conjectured that the sample complexity \( pd^3 \) required to reconstruct a order-3 tensor in polynomial time — even in the noiseless case — is at least \( d^{3/2} \) (or equivalently, \( p \gtrsim 1/d^{3/2} \) [BM15, XYZ17, MS18]. Therefore, our theory reveals that spectral methods achieve consistent estimation (w.r.t. both \( \| \cdot \| \) and \( \| \cdot \|_{2,\infty} \), as long as the sample size is slightly above the (conjectured) computational limit. Moreover, it is easily seen that the bias incurred by deleting the diagonal is much smaller than the error due to missing data, which justifies the rationale that diagonal deletion does not harm the performance by much.

- **Noise size.** We now comment on the noise size requirement. It is easily seen that the maximum magnitude of the entries of \( T^* \) in this case is \( \|T^*\|_{\infty} \asymp \lambda_{\max}^{\star} / d^{3/2} \). As a result, the noise size condition in (29) is equivalent to

\[ \frac{\sigma}{\|T^*\|_{\infty}} \lesssim \sqrt{pd^{3/2}}. \]
Taken together with our sample size requirement \( p \geq \frac{\log^2 d}{\eta^2} \), this condition allows the noise magnitude in each observed entry to significantly exceed the size of the corresponding entry, which covers a broad range of scenarios of practical interest. In addition, in the fully-observed case (i.e. \( p = 1 \)) with i.i.d. Gaussian noise, the authors in [ZX18] showed that the noise size condition (29) — up to some log factor — is necessary for any polynomial-time algorithm to achieve consistent estimation, provided that a certain hypergraph planted clique conjecture holds.

- **\( \ell_{2,\infty} \) estimation error control.** Our results reveal that the estimation errors of \( U \) are fairly de-localized; there is no row of \( U \) that suffers from an estimation error larger than \( 1/\sqrt{d} \), provided that \( U \) is properly rotated to account for the global rotational ambiguity. As far as we know, this is a sort of performance guarantees previously unavailable. In addition, we note that [XZ19, Theorem 4] derived an appealing \( \ell_{2,\infty} \) statistical error bound for an algorithm called HOSVD, under the tensor de-noising setting. In comparison to the Gaussian noise considered therein, our results accommodate the case with missing data and possibly spiky noise.

Finally, we remark that in the fully-observed case (i.e. \( p = 1 \)) with i.i.d. Gaussian noise, it can be seen from [ZX18, Theorem 1] that (26a) is suboptimal; in fact, the minimax risk consists only of the linear term in \( \sigma \) (namely, \( \frac{1}{\sqrt{d}} \), if we omit log factors and assume \( r, \mu_2, \mu_5, \kappa > 1 \)). This is a typical drawback of the spectral method. However, the spectral estimate offers a reasonably good initial estimate for this problem, and one can often employ optimization-based iterative refinement paradigms (like gradient descent [CLPC19]) to obtain minimax optimal estimates.

### 3.3.2 Covariance estimation with missing data

Next, we develop concrete consequences of our main theorem for covariance estimation with missing data. Recall the model introduced in Section 1, and assume that

\[
x_i = B^* f_i^* + \eta_i \in \mathbb{R}^d, \quad f_i^* \overset{i.i.d.}{\sim} \mathcal{N}(0, I_r), \quad 1 \leq i \leq n,
\]

where the noise vector \( \eta_i = [\eta_{i,j}]_{1 \leq j \leq d} \) consists of independent Gaussian components obeying

\[
\mathbb{E}[\eta_{i,j}] = 0 \quad \text{ and } \quad \text{Var}[\eta_{i,j}] \leq \sigma^2.
\]

Note that the noise components might have different variance — a scenario often referred to as heteroscedecity. What we observe is a partial set of entries of \( x_i = [x_{i,j}]_{1 \leq j \leq d} \), namely, we only observe

\[
x_{i,j}, \quad \forall (i,j) \in \Omega,
\]

where \( \Omega \) is an index set obtained by random sampling with rate \( p \) (i.e. each \( i, j \) is included in \( \Omega \) independently with probability \( p \)). The goal is to estimate the subspace spanned by \( B^* \), or even \( B^* B^{*\top} \), from this set of partial and noisy observations. Our spectral method for covariance estimation is summarized in Algorithm 3.

**Algorithm 3** The spectral method for covariance estimation

1. **Input:** sampling set \( \Omega \), observed entries \( \{X_{i,j} \mid (i,j) \in \Omega \} \), sampling rate \( p \), rank \( r \).
2. Let \( A = P_\Omega(X) \in \mathbb{R}^{d \times n} \) with \( X = [x_1, \ldots, x_n] \), and use \( A \) as the input of Algorithm 1. Let \( U \in \mathbb{R}^{d \times r} \) and \( \Sigma \in \mathbb{R}^{r \times r} \) be the estimates returned by Algorithm 1, and set \( B := \frac{1}{\sqrt{n}} U \Sigma \).
3. **Output** \( U \) as the subspace estimate and \( S := BB^\top \) as the covariance estimate.

In order to present our performance guarantees, we make a few more assumptions. Without loss of generality, we shall assume

\[
B^* = U^* \Lambda^{1/2} \quad \text{ and } \quad S^* := B^* B^{*\top} = U^* \Lambda^* U^{*\top},
\]

where \( U^* \in \mathbb{R}^{d \times r} \) consists of orthonormal columns, and \( \Lambda^* = \text{diag}(\lambda_1^*, \ldots, \lambda_r^*) \in \mathbb{R}^{r \times r} \) is a diagonal matrix with \( \lambda_1^* \geq \cdots \geq \lambda_r^* \geq 0 \). We denote

\[
\kappa_{ee} := \frac{\lambda_1^*}{\lambda_r^*}.
\]

In order to allow for reliable estimation, we further impose the following incoherence assumption:
Assumption 4 (Incoherence). Suppose that $U^*$ satisfies

$$\|U^*\|_{2,\infty} = \sqrt{\frac{\mu_{ce}}{d}} \|U^*\|_F = \sqrt{\frac{\mu_{ce} r}{d}}$$

with an incoherence parameter $\mu_{ce} > 0$.

We are now positioned to derive statistical estimation guarantees using our general theorem. The following result is a consequence of Theorem 1; see Appendix A.2.

Corollary 2 (Covariance estimation). Consider the above covariance estimation model with missing data. There exist universal constants $c_0, c_1 > 0$ such that if

$$n \geq c_0 \max \left\{ \frac{\mu_{ce} r^2 \log^2 (n + d)}{dp^2}, \frac{\mu_{ce} r^2 \log^2 (n + d)}{p}, \frac{\mu_{ce} r^2 \log^2 (n + d)}{\lambda_r^2 p^2}, \frac{\mu_{ce} r^2 \log^2 (n + d)}{\lambda_r^2 p^2} \right\},$$

and

$$r \leq c_1 \frac{d}{\mu_{ce} r^2},$$

then with probability exceeding $1 - O((n + d)^{-10})$, Algorithm 3 yields

$$\|UR - U^*\|_F \lesssim \mathcal{E}_{ce},$$

$$\|UR - U^*\|_{2,\infty} \lesssim \frac{\mu_{ce} r \log (n + d)}{d} \cdot \mathcal{E}_{ce},$$

$$\|S - S^*\|_F \lesssim \frac{\mu_{ce} \lambda_r \cdot \mathcal{E}_{ce},}{d},$$

$$\|S - S^*\|_{\infty} \lesssim \frac{\mu_{ce} r \log (n + d)}{d} \lambda_r \cdot \mathcal{E}_{ce},$$

where $R := \arg \min_{Q \in \Omega^{-x}} \|UR - U^*\|_F$ and

$$\mathcal{E}_{ce} := \frac{\mu_{ce} r^2 \log^2 (n + d)}{\sqrt{dn} p} + \frac{\mu_{ce} r^2 \log^2 (n + d)}{np} + \frac{\sigma^2}{\lambda_r^2} \sqrt{d \frac{\log (n + d)}{n}} + \frac{\sigma}{\lambda_r} \sqrt{d \frac{\log (n + d)}{n}} + \frac{\mu_{ce} r \log (n + d)}{d}.$$

Remark 2. We note that in the fully observed case (i.e. $p = 1$), the first two terms on the right-hand side of (33) can be removed.

To facilitate interpretation, let us again focus on the case where $\mu_{ce}, \kappa_{ce} \asymp 1$. Corollary 2 demonstrates that for any given sampling rate $p$, we can achieve

$$\min_{Q \in \Omega^{-x}} \|UR - U^*\| = o(1), \quad \min_{Q \in \Omega^{-x}} \|UR - U^*\|_{2,\infty} = o(\sqrt{r/d}),$$

$$\|S - S^*\| = o(\lambda_r^*), \quad \|S - S^*\|_{\infty} = o(\lambda_r^* r/d),$$

as long as the number $n$ of samples satisfies

$$n \gg \max \left\{ \frac{r^2}{dp^2}, \frac{r \sigma^2 d}{\lambda_r^2 p^2}, \frac{\sigma^2 d}{\lambda_r^2 p^2} \right\} \text{poly log } d.$$

Throughout this subsection, the sample size refers to $n$ — the number of sample vectors $\{x_i\}_{1 \leq i \leq n}$ we have available.

• Interpretation via the effective rank. We first interpret Corollary 2 for a simple case, namely, the spiked covariance matrix model with homoscedastic noise. More specifically, each sample vector $x_i$ is independently drawn from $N(0, S^* + \sigma^2 I_d)$. The standard notion of effective rank is thus given by [Ver12]

$$r_{eff} := \frac{\text{tr}(S^* + \sigma^2 I_d)}{\|S^* + \sigma^2 I_d\|} = \frac{\sum r \lambda_r^* + \sigma^2 d}{\lambda_1^* + \sigma^2} \asymp \begin{cases} d, & \text{if } \frac{\sigma}{\lambda_1^*} \gg 1, \\ r, & \text{if } \frac{\sigma}{\lambda_1^*} \lesssim \sqrt{r}, \\ \frac{\sigma^2 d}{\lambda_1^*}, & \text{otherwise.} \end{cases}$$
In the regime $\sigma \lesssim \sqrt{\lambda^*_r}$, the error bound in (33) can alternatively be expressed as

$$\mathcal{E}_{ce} \approx \left( \frac{r_{eff}}{\sqrt{dp}} + \sqrt{\frac{r_{eff}}{np}} \right) \text{poly log}(n + d) + \frac{r}{d},$$

and the sample complexity requirement for consistent estimation of $U^*$ and $S^*$ (i.e. (34)) becomes

$$n \gg \max \left\{ \frac{r_{eff}^2}{dp^2}, \frac{r_{eff}}{p} \right\} \text{poly log } d.$$  \hspace{1cm} (37)

This matches the condition for covariance matrix estimation in [Ver12, Remark 5.53] in the fully observed case ($p = 1$) (neglecting logarithmic terms). More importantly, we have obtained stronger entrywise statistical guarantees for estimating $S$ under the same conditions.

- **Near minimaxity.** Imagine that all entries are observed (i.e. $p = 1$) and that $\min_{i,j} \text{Var}(\eta_{i,j}) \approx \max_{i,j} \text{Var}(\eta_{i,j}) \approx \sigma^2$. In this scenario, our $\| \cdot \|$ subspace estimation guarantees (see Corollary 2 and Remark 2) simplify to

$$\mathcal{E}_{ce} \approx \sqrt{\frac{d}{n}} \left( \frac{\sigma^2}{\lambda^*_r} + \frac{\sigma}{\sqrt{\lambda^*_r}} \right) \text{poly log}(n + d) + \frac{r}{d}. $$

The first term above matches the minimax lower bound derived in [ZCW18, Theorem 2] up to some logarithmic factor, whereas the second term (which is due to diagonal deletion) is fairly small in the low rank and incoherent case.

- **$\ell_{2,\infty}$ estimation error.** Our result (32b) reveals that the estimation error of the subspace estimate $U$ is evenly spread across all rows (modulo some global rotation). Now we compare our results with prior work. We shall focus attention on the case with $r \approx 1$ and ignore all log factors.

  - Suppose that $\sigma = 1$. In this setting, [ZWS19, Theorem 4] demonstrates that if

  $$n \gtrsim \max \left\{ \frac{1}{p^2}, \frac{d^2}{\lambda^*_r + 2}, \frac{d}{\lambda^*_r p^2} \right\} \text{poly log } d,$$

  then one has

  $$\min_{Q \in O^{d \times r}} \| UQ - U^* \|_{2,\infty} \lesssim \frac{1}{p \sqrt{n}} \left( \frac{1}{\sqrt{\lambda^*_r}} + \frac{1}{\lambda^*_r} \right) \left( 1 + \frac{\sqrt{d}}{\lambda^*_r} \right) \text{poly log } d$$

  with high probability. In comparison, our sample size requirement for consistent estimation improves upon [ZWS19, Theorem 4] by a factor of $\min \left\{ d, p^{-1} \right\}$. Moreover, our estimation error bound improves upon [ZWS19, Theorem 4] by a factor of $\min \left\{ \sqrt{\lambda^*_r}, \frac{1}{\sqrt{d}} \right\}$ if $\sqrt{d} \ll \lambda^*_r \lesssim d$, by a factor of $\frac{\sqrt{d}}{\lambda^*_r}$ when $\lambda^*_r \lesssim 1$, and by a factor of $\min \left\{ \frac{\sqrt{d}}{\lambda^*_r \sqrt{n}}, \sqrt{\frac{d}{\lambda^*_r}} \right\}$ if $1 \ll \lambda^*_r \lesssim \sqrt{d}$.

  - In the absence of missing data, the $\ell_{2,\infty}$ error bound presented in [CTP19b, Theorem 1.1] reads (ignoring logarithmic terms)

  $$\min_{Q \in O^{d \times r}} \| UQ - U^* \|_{2,\infty} \lesssim \begin{cases} \sqrt{\frac{1}{nd}} & \text{for } \frac{\sigma}{\sqrt{\lambda^*_r}} \lesssim \frac{1}{\sqrt{d}} \\ \frac{\sigma^2}{\lambda^*_r \sqrt{\frac{d}{n}}} & \text{for } \frac{1}{\sqrt{n}} \ll \frac{\sigma}{\sqrt{\lambda^*_r}} \lesssim 1 \end{cases}$$

  with high probability. Consequently, our result improves upon the error bound by a factor of $\frac{\sqrt{d}}{\sqrt{\lambda^*_r}}$ if $\frac{1}{\sqrt{d}} \ll \frac{\sigma}{\sqrt{\lambda^*_r}} \lesssim 1$, while being able to handle the case with larger noise (namely, $\frac{\sigma}{\sqrt{\lambda^*_r}} \gg 1$).

Finally, we remark that Algorithm 3 can be applied straightforwardly to the streaming data setting — namely, when the data samples $\{x_i\}$ arrive sequentially over time — by recursively updating the Gram matrix $G$. This certifies Algorithm 3 as a memory-efficient and guaranteed algorithm for streaming principal component analysis (PCA) with missing data [BCL18].
3.3.3 Community recovery in bipartite stochastic block models

In this subsection, we apply our general theorem to the problem of community recovery in the following bipartite stochastic block model (BSBM).

- **Biclustering structure.** Consider two disjoint collections of nodes $\mathcal{U}$ and $\mathcal{V}$, which are of size $n_u$ and $n_v$, respectively. Suppose that each collection of nodes can be clustered into two communities. To be more precise, let $\mathcal{I}_1 \subseteq \mathcal{U}$ and $\mathcal{I}_2 = \mathcal{U} \setminus \mathcal{I}_1$ (resp. $\mathcal{J}_1 \subseteq \mathcal{V}$ and $\mathcal{J}_2 = \mathcal{V} \setminus \mathcal{J}_1$) be two non-overlapping communities in $\mathcal{U}$ (resp. $\mathcal{V}$) that contain $n_u/2$ (resp. $n_v/2$) nodes each. Without loss of generality, we assume that $\mathcal{I}_1$ contains the first $n_u/2$ nodes of $\mathcal{U}$, and $\mathcal{J}_1$ contains the first $n_v/2$ nodes of $\mathcal{V}$; these are of course *a priori* unknown.

- **Measurement model.** What we observe is a random bipartite graph generated based on the community memberships of the nodes. In the simplest version of BSBMs, a pair of nodes $(i,j) \in (\mathcal{U},\mathcal{V})$ is connected by an edge independently with probability $q_{in}$ if either $(i,j) \in (\mathcal{I}_1,\mathcal{J}_1)$ or $(i,j) \in (\mathcal{I}_2,\mathcal{J}_2)$ holds, and with probability $q_{out}$ otherwise. Here, $0 \leq q_{out} \leq q_{in} \leq 1$ represent the edge densities. If we denote by $C \in \{0,1\}^{n_u \times n_v}$ the bi-adjacency matrix of this random bipartite graph, then one has

$$\mathbb{P}\{C_{i,j} = 1\} = \begin{cases} q_{in}, & \text{if } (i,j) \in (\mathcal{I}_1,\mathcal{J}_1) \text{ or } (i,j) \in (\mathcal{I}_2,\mathcal{J}_2), \\ q_{out}, & \text{otherwise.} \end{cases}$$

While our theory covers a broad range of $n_u$ and $n_v$, we emphasize the case where $n_v \gg n_u$ (namely, $\mathcal{V}$ contains far more nodes than $\mathcal{U}$). In such a case, it is not uncommon to encounter a situation where one can only hope to recover the community memberships of the nodes in $\mathcal{U}$ but not those in $\mathcal{V}$.

To attempt community recovery, we look at a centered version of the bi-adjacency matrix\(^2\)

$$A := C - \frac{q_{in} + q_{out}}{2} 1_{n_u} 1_{n_v}^\top. \quad (38)$$

Recognizing that

$$A^* := \mathbb{E}[A] = \frac{q_{in} - q_{out}}{2} \begin{bmatrix} 1_{n_u/2} 1_{n_v/2}^\top, & -1_{n_u/2} 1_{n_v/2}^\top \\ -1_{n_u/2} 1_{n_v/2}^\top, & 1_{n_u/2} 1_{n_v/2}^\top \end{bmatrix} = \frac{q_{in} - q_{out}}{2} \begin{bmatrix} 1_{n_u/2} & 1_{n_v/2}^\top \\ -1_{n_u/2} & -1_{n_v/2}^\top \end{bmatrix}, \quad (39)$$

we see that the leading eigenvector of $A^*$ reveals the community memberships of all nodes. Motivated by this observation, our algorithm for recovering the community memberships in $\mathcal{U}$ proceeds as follows:

**Algorithm 4** The spectral method for BSBM

1. **Input:** observed bi-adjacency matrix $C$, edge probabilities $q_{in}, q_{out}$.
2. Employ $A$ (cf. (38)) as the input of Algorithm 1, and let $u = [u_i] \in \mathbb{R}^{n_u}$ be the output returned by Algorithm 1 (which serves as the estimate of the leading left singular subspace of $A^*$).
3. **Output:** for any $i \in \mathcal{U}$, we claim that $i$ belongs to the first community if $u_i > 0$, and the second community otherwise.

In what follows, we define

$$n := n_u + n_v, \quad (40)$$

and declare exact community recovery of $\mathcal{U}$ if the partition of the nodes returned by our algorithm coincides precisely with the true partition $(\mathcal{I}_1, \mathcal{I}_2)$.

With these in place, we are ready to invoke our general theory to demonstrate the effectiveness of the above algorithm, as asserted by the following corollary.

**Corollary 3** (Bipartite stochastic block model). Consider the above bipartite stochastic block model. There exists some universal constant $c_0 > 0$ such that if

$$\frac{(q_{in} - q_{out})^2}{q_{in}} \geq c_0 \max \left\{ \frac{\log n}{\sqrt{n_u n_v}}, \frac{\log n}{n_v} \right\}, \quad (41)$$

\(^2\)Here, we assume prior knowledge about $q_{in}$ and $q_{out}$. Otherwise, the quantity $\frac{q_{in} + q_{out}}{2}$ can also be easily estimated.
then Algorithm 4 achieves exact community recovery of $\mathcal{U}$ with probability exceeding $1 - O(n^{-10})$.

We then take a moment to discuss the implications of Corollary 3. For simplicity of presentation, we shall focus on the scenario with $q_n \gg q_{out} = o(1)$ and $n_u \leq n_v$.  

- **Exact recovery via the spectral method alone.** Consider the following sparse regime, where

$$q_n = \frac{a \log n}{\sqrt{n_u n_v}} \quad \text{and} \quad q_{out} = \frac{b \log n}{\sqrt{n_u n_v}}$$

for some absolute positive constants $a \geq b$. Corollary 3 demonstrates that we can achieve exact recovery when $\frac{(a-b)^2}{a} \geq 1$. This improves upon prior results presented in [FP16]. More specifically, the results in [FP16] only guaranteed almost exact recovery of community memberships (namely, obtaining correct community memberships for a fraction $1 - o(1)$ of the nodes). In comparison, our results assert that the spectral estimates alone are sufficient to reveal exact community memberships for all nodes in $\mathcal{U}$; there is no need to invoke further refinement procedures to clean up the remaining errors.

- **Near optimality.** In the balanced case where $n_u \asymp n_v$, the condition $\frac{(a-b)^2}{a} \geq 1$ above is known to be information-theoretically optimal up to a constant factor. In the unbalanced case with $n_v \geq n_u$, prior work has identified a sharp threshold for detection — the problem of recovering a fraction $1/2 + \epsilon$ of the community memberships for an arbitrarily small fixed constant $\epsilon > 0$. Specifically, such results reveal a fundamental lower limit that requires $\frac{(d_1 - d_2)^2}{d_1} \geq \frac{1}{\sqrt{n_u n_v}}$ [FPV15, FP16], thus implying the information-theoretic optimality of the spectral method (up to a logarithmic factor).

### 3.4 Numerical experiments

To confirm the applicability of our algorithm and the theoretical findings, we conduct a series of numerical experiments. All results reported in this subsection are averaged over 100 independent Monte Carlo trials. For the sake of comparisons, we also report the numerical performance of the vanilla spectral method (namely, returning the $r$-dimensional principal column subspace of $A$ directly without proper diagonal deletion).

#### Subspace estimation for random low-rank data matrices.

We start with subspace estimation for a randomly generated matrix $A^\ast$. Specifically, generate $A^\ast = Z_1 Z_2^\top$, where $Z_1 \in \mathbb{R}^{d_1 \times r}$, $Z_2 \in \mathbb{R}^{d_2 \times r}$ consist of i.i.d. standard Gaussian entries. The noise matrix contains i.i.d. Gaussian entries, namely, $N_{i,j,k} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ for each $(i,j) \in [d_1] \times [d_2]$. Figure 1(a) - Figure 1(c) plot respectively the numerical estimation errors of the estimate $U$ vs. the sampling rate $p$, the column dimension $d_2$, and the standard deviation $\sigma$ of noise. Two types of estimation errors are reported: (1) the relative spectral norm error $\|UR - U^\ast\| / \|U^\ast\|$; (2) the relative $\ell_{2,\infty}$ norm error $\|UR - U^\ast\|_{2,\infty} / \|U^\ast\|_{2,\infty}$, where $R := \arg\min_{Q \in \mathbb{S}_d} \|UQ - U^\ast\|_F$. As can be seen from the plots, Algorithm 1 yields reasonably good estimates in terms of both the spectral norm and the $\ell_{2,\infty}$ norm, outperforming the vanilla spectral method in all experiments.

#### Tensor completion from noise data.

Next, we consider numerically the problem of tensor completion from noisy observations of its entries. Recall the notations in Section 3.3.1. We generate $W^\ast \in \mathbb{R}^{d \times r}$ with i.i.d. standard Gaussian entries, and generate $N_{i,j,k} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ independently for each $(i,j,k) \in [d]^3$. Figure 2(a) and Figure 2(b) illustrate the relative estimation errors of the subspace estimate $U$ vs. the sampling rate $p$ and noise standard deviation $\sigma$, respectively. Encouragingly, Figure 2 shows that Algorithm 2 accurately recovers the subspace spanned by the tensor factors of interest (with respect to both the spectral norm and the $\ell_{2,\infty}$ norm); in particular, it is capable of producing faithful subspace estimates even when the vanilla spectral method fails.

#### Covariance estimation with missing data.

The next series of experiments is concerned with covariance estimation with missing data. Recall the notations in Section 3.3.2. We draw $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma^\ast)$ independently with $\Sigma^\ast = U^\ast U^\ast^\top$, where $U^\ast \in \mathbb{R}^{d \times r}$ is a i.i.d. standard Gaussian random matrix in $\mathbb{R}^{d \times r}$, and $e_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2 I_d)$ for each $1 \leq i \leq n$. We first consider the estimation error of the subspace. The numerical
Figure 1: Relative estimation errors of the subspace estimate \(U\) for both Algorithm 1 and the vanilla spectral method. The results are reported for (a) relative error vs. sampling rate \(p\) (where \(d_1 = 100, d_2 = 1000, r = 4, \sigma = 1\)), (b) relative error vs. column dimension \(d_2\) (where \(d_1 = 100, r = 4, \sigma = 1, p = \frac{2r\log(d_1+d_2)}{\sqrt{d_1d_2}}\)), and (c) relative error vs. noise standard deviation \(\sigma\) (where \(d_1 = 100, d_2 = 1000, r = 4, p = 0.1\)).

Figure 2: Relative estimation errors of the subspace \(U\) spanned by tensor components in tensor completion. The results are plotted for (a) relative error vs. sampling rate \(p\) (where \(d = 100, r = 4, \sigma = 2\)), and (b) relative error vs. noise standard deviation \(\sigma\) (where \(d = 100, r = 4, p = 0.1\)).

estimation errors of the estimate \(U\) vs. the sampling rate \(p\), the sample size \(n\) and the noise standard deviation \(\sigma\) are plotted in Figure 3(a) – Figure 3(c), respectively. We then turn to the estimation accuracy of the covariance matrix. The numerical estimation errors of the estimate \(S\) of Algorithm 3 and the vanilla spectral method vs. the sampling rate \(p\), the sample size \(n\) and the noise standard deviation \(\sigma\) are plotted in Figure 4, respectively. Similar to previous experiments, Algorithm 3 produces reliable estimates both in terms of the spectral norm, the \(\ell_2,\infty\) norm and the \(\ell_\infty\) norm accuracy.

Community recovery in bipartite stochastic block model. Finally, we conduct numerical experiments for community recovery in bipartite stochastic block models. The parameters are chosen to be \(q_{\text{in}} = \frac{a\log(n_u+n_v)}{\sqrt{n_u+n_v}}\) and \(q_{\text{out}} = \frac{b\log(n_u+n_v)}{\sqrt{n_u+n_v}}\) for some constants \(a > b > 0\). Figure 5(a) reveals a phase transition phenomenon concerned with exact community recovery. As can be seen, Algorithm 4 always succeeds in achieving exact recovery once \(a\) — or equivalently \(q_{\text{in}}\) — exceeds a certain threshold, which outperforms the vanilla spectral method. In Figure 5(b), we vary the number \(n_u\) nodes in \(V\) and plot the empirical success rates for exact recovery. The advantage of Algorithm 4 compared to the vanilla spectral method can be clearly seen from the plot.

4 Further related work

A natural class of spectral algorithms to estimate the leading singular subspace of a matrix — when given a noisy and sub-sampled copy of the true matrix — is to compute the leading left singular subspace of the
observed data matrix. Despite the simplicity of this idea, this type of spectral methods provably achieves appealing performances for multiple statistical problems when the true matrix is (nearly) squared. A partial list of examples includes low-rank matrix estimation and completion [KMO10, JNS13, Cha15, CW15, MWCC17], community detection [RCY11, PF14, AFWZ17, Lei19], and synchronization and alignment [Sin11, AFWZ17, CC18a, SHSS16]. The $\ell_2$ statistical analysis of such algorithms relies heavily on the matrix perturbation theory such as the Davis-Kahan sin $\Theta$ theorem [DK70], the Wedin theorem [Wed72], and their extensions [Vu11, YWS14, OVV18, CZ18, CLC19, ZCW18].

However, the above-mentioned approach might lead to sub-optimal performance when the row dimension and the column dimension of the matrix differ dramatically. This issue has already been recognized in multiple contexts, but not limited to unfolding-based spectral methods for tensor estimation [HSS16, XY17, MS18, ZX19, ZX18] and spectral methods for biclustering [FP16]. Motivated by this sub-optimality issue, an alternative is to look at the “sample Gram matrix” which, as one expects, shares the same leading left singular space as the original observed data matrix. However, in the highly noisy or highly subsampled regime, the diagonal entries of the sample Gram matrix are highly biased, thus requiring special care. Several different treatments of diagonal components have been adopted for different contexts, including proper rescaling [MS18, Lou14, GRESS16], deletion [FP16], and iterative updates [ZCW18]. The deletion strategy is perhaps the simplest of this kind, as it does not require estimation of noise parameters. We note, however, that performing more careful iterative updates might be beneficial for certain heteroskedastic noise scenarios; see [ZCW18] for detailed discussions.

An important application of our work is the problem of tensor completion and estimation [GRY11, LMWY13, RPP13, KOKC13, MHWG14, RM14, GSSB17, SRK09, ZY16, YZ17, HZC18]. Despite its similarity to matrix completion, tensor completion is considerably more challenging; for concreteness and simplicity, we shall only discuss order-3 symmetric rank-$r$ tensors in $\mathbb{R}^{d^3}$. Motivated by the success of matrix completion, a simple strategy for tensor completion/denoising is to unfold the observed tensor into a $d \times d^2$ matrix and to apply standard matrix completion methods for completion. However, existing statistical guarantees derived in the matrix completion literature [CR09, KMO10, Gro11] do not lead to useful bounds unless the sample size exceeds the order of $rd^2$, which far exceeds the requirement for other methods such as the sum-of-squares (SOS) hierarchy [BM15, BM16, PS17]. The work by [MS18] demonstrates that unfolding-based spectral algorithms can also lead to useful estimates under minimal sample complexity, as long as we look at the “Gram matrix” instead. In addition, such spectral algorithms also play an important role in initializing other nonconvex optimization methods [XY17, XYZ17, CLPC19].

In addition, there is an enormous literature on covariance estimation and principal component analysis (PCA). For instance, the classical spiked covariance model [Job01] has been extensively studied; in particular, the high-dimensional setting has inspired much investigation from both algorithmic and analysis perspectives [JL09, Pau07, BL08, Nad08, CY12, Ma13, CMW13, CMW15]. More recently, the result for the classical spiked covariance model has been extended to accommodate heteroskedastic noise [ZCW18]. When it comes to incomplete data, a variety of methods have been introduced [Kie97, EvdG19, JH12]. For instance, Lounici

![Figure 3](https://example.com/figure3.png)
considered estimating the top eigenvector in the setting of sparse PCA in [Lou13], and further proposed an estimator for the covariance matrix in [Lou14]. In [CZ16], bandable and sparse covariance matrices are considered. In addition, most of the prior work considered uniform random subsampling, and the recent work [ZWS19, PO19] began to account for heterogeneous missingness patterns.

Turning to the problem of community recovery or graph clustering, we note that extensive research has been carried out on stochastic block models or censored block models, which can be viewed as special cases of uni-partite networks [MNS14, Mas14, ABH15, MNS15, HWX16, JCSX11, CKST16, CRV15, GV16, CSG16, JMRT16, CLX18, CL15, GMZZ17]. The algorithms that enable exact community recovery in these block models include two-stage approaches [ABH15, MNS14] and semidefinite programming [HWX16, AL18, ABKK17, Ban18, GV16]. In addition, spectral clustering algorithms have been extensively studied as well [CO06, CO10, RCY11, YP14, LR15, YP16, AFWZ17, GMZZ17, Vu18, OVW18, STFP12, LMX15]. While this class of algorithms was originally developed to yield almost exact recovery (e.g. [ABH15]), the recent work by [AFWZ17, Lei19] uncovered that spectral methods alone are sufficient to achieve optimal exact community recovery (a.k.a. achieving strong consistency) for stochastic block models. The interested reader is referred to [Abb18] for an in-depth overview. Various extensions of the SBMs have been introduced and studied in the last few years. Our work contributes to this growing literature by justifying the optimality of spectral methods in bipartite stochastic block models [FPV15, FP16, GLMZ16].

Further, entrywise statistical analysis has recently received significant attention for various statistical problems [FWZ16, AFWZ17, MSC17, CFMW19, CCF18, ZB18, ASSS19, Leib19, EBW18, CTP19b, CTP19a, XZ19, RV15, PW19]. For instance, entrywise guarantees for spectral methods are obtained in [CCF18, EBW18] based on an algebraic Neumann trick, while the results in [ZB18, AFWZ17, CFMW19] were established based on a leave-one-out analysis. The work by [KL16, KX16, CCF18] went one step further by controlling an arbitrary linear form of the eigenvectors or singular vectors of interest. These results, however, typically lead to suboptimal performance guarantees when the row dimension and the column dimension of the matrix are substantially different.

Finally, we recently became aware of an unpublished work by Abbe, Fan and Wang [AFW19], which
also considers statistical guarantees of PCA beyond the usual $\ell_2$ analysis; in particular, they develop an analysis framework that delivers tight $\ell_p$ perturbation bounds. Note, however, that their results are very different from the ones presented here. For instance, the results presented herein emphasize the scenarios with drastically different $d_1$ and $d_2$, which are not the main focus of [AFW19].

5 Analysis

In this section, we discuss in detail the analysis techniques employed to establish Theorem 1. This is built upon a leave-one-out (as well as a leave-two-out) analysis strategy that is particularly effective in controlling entrywise and $\ell_{2,\infty}$ estimation errors [EKBB13, EK15, ZB18, CFMW19, AFWZ17, SCC17, CCFM19, CFMY19, CLL19, LBEK18, PW19].

5.1 Leave-one-out and leave-two-out estimates

In order to facilitate the analysis when bounding $\|UR - U^*\|_{2,\infty}$, we introduce a set of auxiliary leave-one-out matrices — a powerful analysis technique that has been employed to decouple complicated statistical dependency. It is worth emphasizing that these procedures are never executed in practice. Specifically, for each $1 \leq m \leq d_1$, we introduce an auxiliary matrix

$$A^{(m)} = P_{-m,:} (A) + p P_{m,:} (A^*) ,$$

where $P_{-m,:}$ (resp. $P_{m,:}$) represents the projection onto the subspace of matrices supported on the index subset $\{d_1 \setminus \{m\}\} \times [d_2]$ (resp. $\{m\} \times [d_2]$). In other words, $A^{(m)}$ is obtained by replacing all entries in the $m$-th row by their expected values (taking into account the sampling rate). By construction, (1) $A^{(m)}$ is statistically independent of the data in the $m$-th row of $A$, and (2) $A^{(m)}$ is expected to be quite close to $A$, as we only discard a small fraction of data when constructing $A^{(m)}$. These two observations taken together allow for optimal control of the estimation error in the $m$-th row of $U$.

Armed with the leave-one-out matrices, we are ready to introduce auxiliary leave-one-out procedures for subspace estimation. Similar to the matrix $G$ in Algorithm 1 (whose eigenspace serves as an estimate of the column space of $U^*$), we define an auxiliary matrix $G^{(m)} \in \mathbb{R}^{d_1 \times d_1}$ as follows:

$$G^{(m)} = P_{\text{off-diag}} \left( \frac{1}{p^2} A^{(m)} (A^{(m)})^\top \right) ,$$

where $P_{\text{off-diag}} (\cdot)$ (as already defined in Section 2.2) extracts out all off-diagonal entries from a matrix. The auxiliary procedure, which is summarized in Algorithm 5, is very similar to Algorithm 1 except that it operates upon $G^{(m)}$. 

Figure 5: Empirical success rates for exact community recovery in bipartite stochastic block models, where $q_{in} = \frac{a \log(n_u + n_v)}{\sqrt{n_un_v}}$ and $q_{out} = \frac{b \log(n_u + n_v)}{\sqrt{n_un_v}}$. The results are shown for (a) empirical success rate vs. the number $n_v$ of nodes in $V$ (where $n_u = 100$, $a = 0.8$, $b = 0.01$), and (b) empirical exact recovery rate vs. $a$ (where $n_u = 100$, $n_v = 10000$, $b = 0.01$).
Algorithm 5 The $m$-th leave-one-out sequence

1. **Input:** sampling set $\Omega$, observed entries $\{A_{i,j} \mid (i,j) \in \Omega\}$, true entries $\{A_{m,j}^* \mid j \in [d_2]\}$, sampling rate $p$, rank $r$.
2. Let $U^{(m)} A^{(m)} U^{(m)\top}$ be the (truncated) rank-$r$ eigen-decomposition of $G^{(m)}$. Here, $G^{(m)}$ and $A^{(m)}$ are defined respectively in (43) and (42).
3. **Output** $U^{(m)}$ as the subspace estimate and $\Sigma^{(m)} = (A^{(m)})^{1/2}$ as the spectrum estimate.

Given that $A^{(m)}$ (resp. $G^{(m)}$) is very close to $A$ (resp. $G$), one would naturally expect $U^{(m)}$ — the $r$-dimensional principal eigenspace of $G^{(m)}$ — to stay extremely close to the original estimate $U$. This fact will be formalized shortly.

As it turns out, given that the spectral method is applied to the Gram matrix (which is a quadratic form of the original data matrix), introducing the leave-one-out sequences alone is not yet sufficient for our purpose; we still need to introduce an additional set of “leave-two-out” matrices, in the hope of simultaneously handling the row-wise and the column-wise statistical dependency. Specifically, for each $1 \leq m \leq d_1$ and each $1 \leq l \leq d_2$, define the following auxiliary matrices:

$$A^{(m,l)} := P_{-m,l}(A) + p P_{m,l}(A^*), \quad (44a)$$

$$G^{(m,l)} := P_{\text{off-diag}}(p P_{m,l}(A^{(m,l)} \top)), \quad (44b)$$

where $P_{-m,l}$ (resp. $P_{m,l}$) denotes the projection onto the subspace of matrices supported on $\{(d_1 \setminus \{m\}) \times \{d_2\} \setminus \{l\}\}$ (resp. $\{m\} \times \{l\}$). Similar to $A^{(m)}$, $A^{(m,l)}$ is generated by replacing all data lying on the $m$-th row and the $l$-th column of $A$ by their expected values (taking into account the sampling rate). The precise procedure is summarized in Algorithm 6.

Similar to the leave-one-out estimates, one expects the new leave-two-out estimates $U^{(m,l)}$ to be extremely close to $U^{(m)}$ (and hence $U$).

Algorithm 6 The $(m,l)$-th leave-two-out sequence

1. **Input:** sampling set $\Omega$, observed entries $\{A_{i,j} \mid (i,j) \in \Omega\}$, true entries $\{A_{m,j}^* \mid j \in [d_2]\} \cup \{A_{i,l}^* \mid i \in [d_1]\}$, sampling rate $p$, rank $r$.
2. Let $U^{(m,l)} A^{(m,l)} U^{(m,l)\top}$ be the (truncated) rank-$r$ eigen-decomposition of $G^{(m,l)}$. Here, $G^{(m,l)}$ and $A^{(m,l)}$ are defined respectively in (44b) and (44a).
3. **Output** $U^{(m,l)}$ as the subspace estimate and $\Sigma^{(m,l)} = (A^{(m,l)})^{1/2}$ as the spectrum estimate.

5.2 Key lemmas

In this subsection, we provide several lemmas that play a crucial role in establishing our main theorem. These lemmas are primarily concerned with the proximity between the original estimate, the leave-one-out estimates, and the ground truth. Throughout this section, we let

$$G^* := A^* A^{\star\top} = U^* \Sigma^2 U^{\star\top}. \quad (45)$$

To begin with, we demonstrate that $G$ is sufficiently close to $G^*$ when the difference is measured by the spectral norm. In view of standard matrix perturbation theory (which we shall make precise later), the proximity of $G$ and $G^*$ is crucial in bounding the difference between $U$ and $U^*$. The proof is deferred to Appendix B.2.

**Lemma 1.** Instate the assumptions of Theorem 1. With probability at least $1 - O(d^{-10})$, one has

$$\|G - G^*\| \lesssim \frac{\mu \sqrt{\sigma_1^2 \log d}}{\sqrt{d_1 d_2} p} + \sqrt{\frac{\mu r \sigma_1^2 \log d}{d_2 p}} + \frac{\sigma^2 \sqrt{d_1 d_2 \log d}}{p} + \sigma_1 \sqrt{\frac{d_1 \log d}{p}} + \|A^*\|_{2,\infty}^2. \quad (46)$$

$$= \approx \omega$$
In order to get a better sense of the term $\zeta_{op}$ appearing above, we make note of a straightforward yet useful fact, which reveals that $\zeta_{op}$ is much smaller than any nonzero eigenvalue of $G^{*}$.

**Fact 1.** Instantiate the assumptions of Theorem 1. Then the quantity $\zeta_{op}$ as defined in (46) obeys

$$
\zeta_{op} + \|A^*\|_{2,\infty}^2 \leq \frac{\mu r \sigma_r^2 \log d}{\sqrt{d_1 d_2 p}} + \sqrt{\frac{\mu r \sigma_r^4 \log d}{d_2 p}} + \sigma_r^2 \sqrt{d_1 d_2 \log d} + \sigma_r^2 \frac{d_1 \log d}{p} + \frac{\mu_1 r \sigma_r^2}{d_1}
$$

where $\|A^*\|_{2,\infty}^2 \leq \frac{\mu_1 r \sigma_r^2}{d_1}$ (cf. Lemma 11).

Further, the following lemma upper bounds the difference between $G$ and $G^{*}$ in the $m$-th row, when projected onto the subspace represented by $U^*$; the proof is postponed to Appendix B.3. This result gives a more refined control of the difference between $G$ and $G^{*}$.

**Lemma 2.** Instantiate the assumptions of Theorem 1. With probability at least $1 - O(d^{-10})$, the following holds simultaneously for all $1 \leq m \leq d_1$:

$$
\| (G - G^{*})_{m,:} U^* \|_2 \lesssim (\zeta_{op} + \|A^*\|_{2,\infty}^2) \sqrt{\frac{\mu r}{d_1}},
$$

where $\zeta_{op}$ is defined in (46).

The next step, which is also the most challenging and crucial step, lies in showing that: every row of $U$, under certain global linear transformation, serves as a good approximation of the corresponding row of $U^*$. Towards this end, we begin with the following preparations:

- We first introduce the following matrix $H$ to represent the linear transformation we have in mind:

$$
H := U^{T} U^*.
$$

(47)

While this is not a rotation matrix, it is quite close to the rotation matrix $R$ defined in (15).

- In addition, we find it convenient to express

$$
U^* = G^{*} U^* (\Sigma^{*})^{-2}.
$$

Combining this with Lemma 2, one would expect $U^*$ and $G U^{*} (\Sigma^{*})^{-2}$ to be reasonably close, namely,

$$
U^* \approx G U^{*} (\Sigma^{*})^{-2}.
$$

(48)

With these in hand, the following lemma (together with Lemma 2) asserts that

$$
UH \approx GU^{*} (\Sigma^{*})^{-2} \approx U^*
$$

in an $\ell_{2,\infty}$ sense.

**Lemma 3.** Instantiate the assumptions of Theorem 1, and recall the definition of $\zeta_{op}$ in (46). With probability at least $1 - O(d^{-10})$, one has

$$
\|UH - GU^{*} (\Sigma^{*})^{-2}\|_{2,\infty} \lesssim \frac{\kappa^2 (\zeta_{op} + \|A^*\|_{2,\infty}^2)}{\sigma_r^2} \left( \|UH\|_{2,\infty} + \sqrt{\frac{\mu r}{d_1}} \right).
$$

The proof of this lemma, however, goes far beyond conventional matrix perturbation theory, and requires delicate decoupling of statistical dependencies. This is accomplished via leave-one-out and leave-two-out analysis arguments. In what follows, we take a moment to explain the high-level idea.

To establish Lemma 3, we first learn from standard matrix perturbation theory [AFWZ17, Lemma 2] that: for each $1 \leq m \leq d_1$,

$$
\| (UH - GU^{*} (\Sigma^{*})^{-2})_{m,:}\|_2 \lesssim \frac{1}{\lambda_r (G^{*})} \|G - G^{*}\| \|G_{m,:} U^*\|_2 + \frac{1}{\lambda_r (G^{*})} \|G_{m,:} (UH - U^*)\|_2
$$

holds, provided that $G$ and $G^{*}$ are sufficiently close.
• The first term on the right-hand side of (49) can already be controlled by Lemma 1 and Lemma 2.

• The second term on the right-hand side of (49), however, is considerably more difficult to analyze, due to the complicated statistical dependence between $G_m$ and $UH$. In order to decouple statistical dependency, we resort to the leave-one-out sequence $U^{(m)}$ introduced in Algorithm 5 and use the triangle inequality to bound

$$
\|G_m; (UH - U^*)\|_2 \leq \|G_m; (UH - U^{(m)}H^{(m)})\|_2 + \|G_m; (U^{(m)}H^{(m)} - U^*)\|_2,
$$

where $H^{(m)} := U^{(m)\top}U^*$. As mentioned before, the leave-one-out estimate $U^{(m)}$ enjoys two nice properties.

(1) The true estimate $U$ and the leave-one-out estimate $U^{(m)}$ are exceedingly close, as asserted by the following lemma (to be established in Appendix B.5).

**Lemma 4.** Instate the assumptions of Theorem 1, and recall the definition of $H$ in (47). With probability at least $1 - O(d^{-10})$, the following holds simultaneously for all $1 \leq m \leq d_1$:

$$
\left\|U^{(m)}U^{(m)\top} - UU^\top\right\|_F \lesssim \frac{\zeta_{\text{op}}}{\sigma^2} \left(\|UH\|_{2,\infty} + \sqrt{\frac{\mu r}{d_1}}\right),
$$

where $\zeta_{\text{op}}$ is defined in (46).

This result in turn allows us to control the first term on the right-hand side of (50).

(2) Due to the statistical independence between $A_m$ and $U^{(m)}$, the matrices $G_m$ and $U^{(m)}$ turn out to be nearly independent. This allows one to invoke simple concentration inequalities to develop tight bounds for the second term on the right-hand side of (50). The detailed proof can be found in Appendix B.4.

Finally, we make a remark on a technical issue encountered in the proof of Lemma 4. Recall that $U^{(m)}$ is obtained by simply replacing the $m$-th row of $A$ with its population version, which indicates the statistical dependency between $U^{(m)}$ and the $m$-th row of $A$. However, there is still some delicate statistical dependency between $U^{(m)}$ and the columns of $A$ that need to be carefully coped with. Fortunately, the leave-two-out estimate $U^{(m,l)}$ — which is obtained by dropping not only the $m$-th row of $A$ but also the $l$-th of its columns — allows us to decouple the dependency between $U^{(m,l)}$ (and hence $U$ and $U^{(m)}$) and the $l$-th column of $A$. This is precisely the main reason why we introduce additional leave-two-out estimates.

### 5.3 Proof of Theorem 1

We are now positioned to establish our main theorem. The proof is split into two parts.

#### 5.3.1 Statistical accuracy measured by $\|\cdot\|$ 

We begin by establishing the spectral norm bound (17c). Let $\lambda_i$ and $\lambda_i^*$ be the $i$-th largest eigenvalue of $A$ and $A^*$, respectively. From Lemma 1 and Weyl’s inequality, one finds that

$$
\max_{1 \leq i \leq r} |\lambda_i - \lambda_i^*| = \|A - A^*\| \leq \|G - G^*\| \lesssim \zeta_{\text{op}} + \|A^*\|_{2,\infty} \leq \sigma_i^2 \cdot \mathcal{E}_{\text{gen}},
$$

where $\zeta_{\text{op}}$ and $\mathcal{E}_{\text{gen}}$ are defined in (46) and (18), respectively. Here, the last inequality arises from the simple fact that $\|A^*\|_{2,\infty} \leq \frac{\mu r + \sigma_i^2}{\sigma_i^2}$ (cf. Lemma 11). By virtue of Fact 1, we know that $\|A - A^*\| \ll \sigma_i^2$. Given that $A^* = \Sigma^2$ and $A = \Sigma^2$, this implies that for each $1 \leq i \leq r$,

$$
\frac{1}{4} \sigma_i^2 = \frac{1}{4} \lambda_i^* \leq |\lambda_i - \lambda_i^*| \leq \lambda_i \leq \lambda_i^* + |\lambda_i - \lambda_i^*| \leq 4 \lambda_i^* = 4 \sigma_i^2,
$$

thus indicating that

$$
\frac{1}{2} \sigma_i^2 \leq \sigma_i \leq 2 \sigma_i^2.
$$
In conclusion,

\[ \| \Sigma - \Sigma^\bullet \| = \max_{1 \leq i \leq r} | \sigma_i - \sigma_i^\bullet | = \max_{1 \leq i \leq r} \left| \frac{\sigma_i^2 - \sigma_i^{2\bullet}}{\sigma_i + \sigma_i^\bullet} \right| \leq \max_{1 \leq i \leq r} \left\| \frac{\Lambda - \Lambda^\bullet}{2\sigma_i^\bullet} \right\| \lesssim \sigma_i^\bullet \cdot \epsilon_{\text{gen}} \]

as claimed. Here, (a) comes from (52), whereas (b) follows from (51).

Next, we turn attention to (17a). In view of the celebrated Davis-Kahan sin \Theta Theorem [DK70], one can upper bound

\[
\| U R - U^\bullet \| \leq \sqrt{2} \left\| U U^\top - U^\bullet U^{\bullet\top} \right\| \leq \frac{\sqrt{2} \| G - G^\bullet \|}{\lambda_r(G^\bullet) - \lambda_{r+1}(G)} \leq \frac{\sqrt{2} \| G - G^\bullet \|}{\sigma_r^2 - \| G - G^\bullet \|},
\]

where \( R \) is defined in (15). Here, (i) is a well-known inequality connecting two different subspace distance metrics, (ii) follows from the Davis-Kahan sin \Theta theorem, (iii) arises from the Weyl inequality, and the last identity follows since \( \lambda_r(G^\bullet) = \sigma_r^2 \) and \( \lambda_{r+1}(G^\bullet) = 0 \) (recall the definition in (45)). In addition, Lemma 1 and Fact 1 tell us that

\[ \| G - G^\bullet \| \lesssim \| A^\bullet \|_{2,\infty} \ll \sigma_r^2 \]

with probability at least \( 1 - O(d^{-10}) \). This taken collectively with (53) implies that, with probability at least \( 1 - O(d^{-10}) \),

\[ \| U R - U^\bullet \| \leq \frac{2 \sqrt{2} \| G - G^\bullet \|}{\sigma_r^2} \lesssim \epsilon_{\text{gen}}. \]

**5.3.2 Statistical accuracy measured by \( \| \cdot \|_{2,\infty} \)**

Before continuing to the proof, we find it convenient to introduce a few more notations. In addition to the rotation matrix \( R \) defined in (15) and the linear transformation \( H \) defined in (47), we define

\[ \text{sgn}(H) := \bar{U} \bar{V}^\top \in \mathbb{R}^{d_1 \times d_1}, \]

where the columns of \( \bar{U} \in \mathbb{R}^{d_1 \times d_1} \) (resp. \( \bar{V} \in \mathbb{R}^{d_1 \times d_1} \)) are the left (resp. right) singular vectors of \( H \). It is well-known that [TB77, Theorem 2]

\[ R = \text{sgn}(H). \]

We now move on to establishing the advertised bound (17b).

1. To begin with, we claim that \( U H \) is extremely close to \( U R \), provided that \( \| G - G^\bullet \| \) is sufficiently small. To this end, recognizing that \( \| G - G^\bullet \| \lesssim \zeta_{\text{op}} \ll \sigma_r^2 \) (according to Lemma 1 and Fact 1), we can apply [AFWZ17, Lemma 3] to show that

\[ \| H^{-1} \| \leq 1 \quad \text{and} \quad \sqrt{\| H - \text{sgn}(H) \|} \leq \| U U^\top - U^{\bullet\top} U^\bullet \| \lesssim \frac{\zeta_{\text{op}}}{\sigma_r^2}, \]

where the last inequality follows from (55). Thus, invoke the identity (57) to arrive at

\[
\| U H - U R \|_{2,\infty} = \| U H - U \text{sgn}(H) \|_{2,\infty} = \| U H H^{-1} (H - \text{sgn}(H)) \|_{2,\infty} \leq \| U H \|_{2,\infty} \| H^{-1} \| \| H - \text{sgn}(H) \| \lesssim \left( \frac{\zeta_{\text{op}}}{\sigma_r^2} \right)^2 \| U H \|_{2,\infty} \ll \frac{\zeta_{\text{op}}}{\sigma_r^2} \| U H \|_{2,\infty}.
\]

This in turn allows us to focus attention on bounding \( \| U H - U^\bullet \|_{2,\infty} \) (instead of \( \| U R - U^\bullet \|_{2,\infty} \)).
2. Next, recall that \( G^* = U^* \Sigma^* U^* \) and hence \( G^* U^* (\Sigma^*)^{-2} = U^* \). Invoke the triangle inequality to reach
\[
\| UH - U^* \|_{2,\infty} \leq \| UH - GU^* (\Sigma^*)^{-2} + GU^* (\Sigma^*)^{-2} - G^* U^* (\Sigma^*)^{-2} \|_{2,\infty} + \| UH - GU^* (\Sigma^*)^{-2} \|_{2,\infty}
\]
\[
\leq \| (G - G^*) U^* (\Sigma^*)^{-2} \|_{2,\infty} + \| UH - GU^* (\Sigma^*)^{-2} \|_{2,\infty}
\]
\[
\leq \frac{1}{\sigma_r^2} \| (G - G^*) U^* \|_{2,\infty} + \| UH - GU^* (\Sigma^*)^{-2} \|_{2,\infty}.
\]
(59)

Regarding the first term of (59), Lemma 2 reveals that with probability at least \( 1 - O(d^{-10}) \),
\[
\frac{1}{\sigma_r^2} \| (G - G^*) U^* \|_{2,\infty} \leq \frac{\zeta_{\text{op}} + \| A^* \|_{2,\infty}^2}{\sigma_r^2} \sqrt{\frac{\mu r}{d_1}}.
\]
(60)

With regards to the second term of (59), Lemma 3 demonstrates that
\[
\| UH - GU^* (\Sigma^*)^{-2} \|_{2,\infty} \leq \kappa^2 \left( \frac{\zeta_{\text{op}} + \| A^* \|_{2,\infty}^2}{\sigma_r^2} \right) \left( \| UH \|_{2,\infty} + \sqrt{\frac{\mu r}{d_1}} \right)
\]
(61)

with probability at least \( 1 - O(d^{-10}) \). Combine (60) and (61) to arrive at
\[
\| UH - U^* \|_{2,\infty} \leq \kappa^2 \left( \frac{\zeta_{\text{op}} + \| A^* \|_{2,\infty}^2}{\sigma_r^2} \right) \left( \| UH \|_{2,\infty} + \sqrt{\frac{\mu r}{d_1}} \right).
\]
(62)

3. As a byproduct of (62) and Fact 1, we see that
\[
\| UH - U^* \|_{2,\infty} \ll \| UH \|_{2,\infty} + \sqrt{\frac{\mu r}{d_1}}.
\]

It then follows from the triangle inequality that
\[
\| UH \|_{2,\infty} \leq \| UH - U^* \|_{2,\infty} + \| U^* \|_{2,\infty} \ll o(1) \| UH \|_{2,\infty} + \sqrt{\frac{\mu r}{d_1}},
\]
thus indicating that
\[
\| UH \|_{2,\infty} \leq 2 \sqrt{\frac{\mu r}{d_1}}.
\]
(63)

Substitution into (58) and (62) gives
\[
\| UH - UR \|_{2,\infty} \ll \frac{\zeta_{\text{op}}}{\sigma_r^2} \sqrt{\frac{\mu r}{d_1}} \text{ and } \| UH - U^* \|_{2,\infty} \leq \kappa^2 \left( \frac{\zeta_{\text{op}} + \| A^* \|_{2,\infty}^2}{\sigma_r^2} \right) \sqrt{\frac{\mu r}{d_1}}.
\]

Combining the above results yields
\[
\| UR - U^* \|_{2,\infty} \leq \| UH - U^* \|_{2,\infty} + \| UH - UR \|_{2,\infty}
\]
\[
\leq \kappa^2 \left( \frac{\zeta_{\text{op}} + \| A^* \|_{2,\infty}^2}{\sigma_r^2} \right) \sqrt{\frac{\mu r}{d_1}}.
\]

Substituting the value of \( \zeta_{\text{op}} \) into the above inequality and using the upper bound \( \| A^* \|_{2,\infty} \leq \frac{\mu r \sigma_r^2}{d_1} \) (cf. Lemma 11), we conclude the proof.
6 Discussion

This paper explores spectral methods tailored to subspace estimation for low-rank matrices with missing entries. In comparison to prior literature, our findings are particularly interesting when the column dimension $d_2$ far exceeds the row dimension $d_1$. In many scenarios, even though the observed data are either too noisy or too incomplete to support reliable recovery of the entire matrix (so that prior matrix completion results often become inapplicable), they might still be informative enough if the purpose is merely to estimate the column subspace of the unknown matrix. In fact, this suggests a potentially useful paradigm for privacy-preserving estimation or learning: the inability to recover the entire matrix facilitates the protection of personal data, yet it is still possible to retrieve useful subspace information for inference and learning. Our main contribution lies in establishing $\ell_{2,\infty}$ statistical guarantees for subspace estimation, therefore providing a stronger form of performance guarantees compared to the usual $\ell_2$ perturbation bounds.

Moving forward, there are many directions that are worth pursuing. For example, our current theory is likely suboptimal with respect to the dependence on the rank, the condition number, and the incoherence parameters. How to further refine our analysis to improve these aspects? In addition, can we extend our algorithm and theory to accommodate more general sampling patterns? Going beyond estimation, an important direction lies in statistical inference and uncertainty quantification for subspace estimation, namely, how to construct valid and hopefully optimal confidence regions that are likely to contain the unknown column subspace? It would also be interesting to investigate how to incorporate other structural prior (e.g., sparsity) to further reduce the sample complexity and/or improve the estimation accuracy. Finally, another interesting avenue for future exploration is the extension to distributed or decentralized settings.

Acknowledgements

Y. Chen is supported in part by the grants AFOSR YIP award FA9550-19-1-0030, ONR N00014-19-1-2120, ARO W911NF-18-1-0303, NSF CCF-1907661 and IIS-1900140, and the Princeton SEAS Innovation Award. Y. Chi is supported in part by the grants ONR N00014-18-1-2142 and N00014-19-1-2404, ARO W911NF-18-1-0303, and NSF CCF-1806154 and ECCS-1818571. H. V. Poor is supported in part by the NSF grant DMS-1736417. C. Cai is supported in part by Gordon Y. S. Wu Fellowships in Engineering.

A Proofs for corollaries

A.1 Proof of Corollary 1

Recall that the spectral algorithm considered herein (cf. Section 3.3.1) operates upon the noisy copy of the mode-1 matricization of the tensor $T^*$, namely,
\[
A^* = \sum_{s=1}^{r} w^*_s (w^*_s \otimes w^*_s)^\top.
\]

Consequently, in order to apply Theorem 1, the main step boils down to estimating the spectrum and the incoherence parameters of $A^*$. Specifically, we need to upper bound the condition number $\kappa$, as well as the incoherence parameters $\mu_0, \mu_1$ and $\mu_2$ as introduced in Definition 1.

Before proceeding, we introduce a few notations that simplify the presentation. Define
\[
\lambda^*_i := \|w^*_i\|_2^3, \quad 1 \leq i \leq r,
\]
and let $\lambda^*_{(i)}$ denote the $i$-th largest value in $\{\lambda^*_i\}_{i=1}^{r}$. We also recall that
\[
\lambda^*_{\min} := \min_{1 \leq i \leq r} \|w^*_i\|_2^3 \quad \text{and} \quad \lambda^*_{\max} := \max_{1 \leq i \leq r} \|w^*_i\|_2^3.
\]

In addition, we define two matrices of interest
\[
\begin{align*}
\tilde{W}^* := [w^*_1, \cdots, w^*_r] \in \mathbb{R}^{d \times r}, \quad \tilde{W}^* := [w^*_1 \otimes w^*_1, \cdots, w^*_r \otimes w^*_r] \in \mathbb{R}^{d^2 \times r},
\end{align*}
\]
where \( \overline{w}_s^* := w_s^* / \| w_s^* \|_2 \), and \( a \otimes b := \begin{bmatrix} a_1 b \\ \vdots \\ a_d b \end{bmatrix} \). In addition, let \( D^* \in \mathbb{R}^{r \times r} \) be a diagonal matrix with diagonal entries

\[
D^*_{s,s} = \| w_s^* \|_2, \quad 1 \leq s \leq r.
\]

These allow us to express

\[
G^* = A^* A^{\ast \top} = \overline{W}^\top D^* \overline{W}^* + \overline{W}^\top \tilde{W}^* D^* \tilde{W}^*. \tag{65}
\]

In the sequel, we begin by quantifying the spectrum of \( G^* \), which in turn allows us to understand the spectrum of \( A^* \).

- We first look at the eigenvalues of the matrices \( \overline{W}^\top \overline{W}^* \) and \( \tilde{W}^* \tilde{W}^* \). Towards this, let us write

\[
\overline{W}^\top \overline{W}^* = I_r + C, \quad \text{and} \quad \tilde{W}^* \tilde{W}^* = I_r + \tilde{C}
\]

for some matrices \( C, \tilde{C} \in \mathbb{R}^{r \times r} \). It follows immediately from the incoherence assumption (23) that

\[
\| C \|_\infty \leq \sqrt{\mu_5 / d} \quad \text{and} \quad \| \tilde{C} \|_\infty \leq \mu_5 / d,
\]

thus leading to the simple bounds

\[
\| C \| \leq r \| C \|_\infty \leq r \sqrt{\mu_5 / d}, \quad \| \tilde{C} \| \leq r \| \tilde{C} \|_\infty \leq \mu_5 r / d. \tag{66}
\]

These taken collectively with (65) and Weyl’s inequality yield

\[
\max_{i \in [r]} \lambda_i (\overline{W}^\top \overline{W}^*) - 1 \leq \| C \| \leq r \sqrt{\mu_5 / d} \quad \text{and} \quad \max_{i \in [r]} \lambda_i (\tilde{W}^* \tilde{W}^*) - 1 \leq \| \tilde{C} \| \leq \mu_5 r / d,
\]

which essentially tell us that

\[
\| \overline{W} \| = \sqrt{\lambda_1 (\overline{W}^\top \overline{W}^*)} \leq \sqrt{1 + r \sqrt{\mu_5 / d}} \quad \text{and} \quad \| \tilde{W}^* \| = \sqrt{\lambda_1 (\tilde{W}^* \tilde{W}^*)} \leq \sqrt{1 + \mu_5 r / d}. \tag{67}
\]

- Returning to \( G^* \), one invoke the definition (65) to deduce that

\[
G^* = \overline{W}^\top D^* \overline{W}^* + \overline{W}^\top \tilde{D}^3 \tilde{C} D^* \tilde{W}^*.
\]

Observe that the eigenvalues of \( \overline{W}^\top D^3 \overline{W}^* \) are identical to those of \( \overline{W}^\top D^3 \overline{W}^* \), where the latter can be further decomposed as (in view of (65))

\[
(\overline{W}^\top D^3) \overline{W}^* D^3 = D^3 \overline{W}^\top \overline{W}^* D^3 = D^6 + D^3 \tilde{C} D^3.
\]

This taken together with Weyl’s inequality, (66) and (67) shows that

\[
\left| \lambda_i (G^*) - \lambda_i (D^6) \right| \leq \left\| D^3 \tilde{C} D^3 \right\| \leq \left\| D^* \right\|^6 \| C \| \leq r \sqrt{\frac{\mu_5}{d}} \lambda_{\max}^2
\]

for each \( 1 \leq i \leq r \). In addition,

\[
\left| \lambda_i (G^*) - \lambda_i (D^6) \right| \leq \left\| \overline{W}^\top D^3 \tilde{C} D^3 \overline{W}^* \right\| \leq \left\| \overline{W} \right\|^2 \left\| D^* \right\|^6 \| \tilde{C} \| \leq \frac{\mu_5 r}{d} \left( 1 + r \sqrt{\frac{\mu_5}{d}} \right) \lambda_{\max}^2.
\]

As a result, invoke the triangle inequality to see that

\[
\left| \lambda_i (G^*) - \lambda_i (D^6) \right| \leq \left| \lambda_i (G^*) - \lambda_i (\overline{W}^\top D^6 \overline{W}^*) \right| + \left| \lambda_i (\overline{W}^\top D^6 \overline{W}^*) - \lambda_i (D^6) \right| \leq \frac{\mu_5 r}{d} \left( 1 + r \sqrt{\frac{\mu_5}{d}} \right) \lambda_{\max}^2 + r \sqrt{\frac{\mu_5}{d}} \lambda_{\max}^2 \leq 3r \sqrt{\frac{\mu_5}{d}} \lambda_{\max}^2,
\]

for each \( 1 \leq i \leq r \), where the last inequality holds under the assumption that \( r \sqrt{\mu_5 / d} \leq 1 \). This means

\[
\left| \lambda_i (G^*) - \lambda_{(i)}^2 \right| \leq 3r \sqrt{\frac{\mu_5}{d}} \lambda_{\max}^2,
\]

where \( \lambda_{(i)}^* \) denotes the \( i \)-th largest value in \( \{ \lambda_i^* \}_{i=1}^r \).
Recalling that $\mu_{tc} := \max \{\mu_3, \mu_4^2\}$, $\kappa_{tc} := \lambda_{max}^*/\lambda_{min}^*$ and the rank assumption $r \ll \kappa_{tc}^{-2} \sqrt{d/\mu_5}$, we find that
\[
\lambda_i(G^*) = \lambda_{max}^* + o \left( r \sqrt{\frac{\mu_5}{d}} \right) \lambda_{max}^2 \\
\text{and } \sigma_i(A^*) = \lambda_{(i)}^*(1 + o(1)). \tag{68}
\]

As a result, we immediately arrive at
\[
\sigma_1(A^*) = \lambda_{max}^*(1 + o(1)), \quad \sigma_r(A^*) = \lambda_{min}^*(1 + o(1)), \quad \text{and } \kappa = \frac{\sigma_1(A^*)}{\sigma_r(A^*)} \lesssim \kappa_{tc}.
\]

Next, we turn attention to bounding the incoherence parameters of $A^*$. Let $A^* = U^* \Sigma^* V^* \top$ be the (compact) SVD of $A^*$. It is seen from (64) that the column space of $U^*$ (resp. $V^*$) coincides with the column space of $\tilde{W}$ (resp. $\tilde{W}^*$). Therefore, there exist orthonormal matrices $H_1$ and $H_2$ such that
\[
U^* H_1 = \tilde{W}(\tilde{W}^\top \tilde{W})^{-1/2} \quad \text{and} \quad V^* H_2 = \tilde{W}^*(\tilde{W}^{*\top} \tilde{W}^*)^{-1/2}.
\]

These allow us to bound
\[
\begin{align*}
\|U^*\|_{2, \infty} & = \|U^* H_1\|_{2, \infty} \leq \|\tilde{W}\|_{2, \infty} \| (\tilde{W}^{*\top} \tilde{W})^{-1/2} \| \leq \sqrt{\frac{\mu_4 r}{d}} \sqrt{\frac{1}{\lambda_r(\tilde{W}^{*\top} \tilde{W})}} \lesssim \sqrt{\frac{\mu_4 r}{d}} \sqrt{\frac{1}{1 - 1/3}} \leq \sqrt{\frac{2 \mu_4 r}{d}}, \\
\|V^*\|_{2, \infty} & = \|V^* H_2\|_{2, \infty} \leq \|\tilde{W}^*\|_{2, \infty} \| (\tilde{W}^{*\top} \tilde{W}^*)^{-1/2} \| \leq \sqrt{\frac{\mu_4^2 r}{d^2}} \sqrt{\frac{1}{\lambda_r(\tilde{W}^{*\top} \tilde{W}^*)}} \lesssim \frac{\mu_4 r}{d} \sqrt{\frac{1}{1 - 1/3}} \leq \sqrt{\frac{2 \mu_4^2 r}{d^2}},
\end{align*}
\]

which follow from (67) and the assumption that $r \ll \sqrt{d/\mu_5}$. Moreover, the incoherence assumption (23) gives that
\[
\mu_0 = \frac{d^3 \|A^*\|_{\infty}^2}{\|A^*\|_F^2} = \frac{d^3 \|T^*\|_{2, \infty}^2}{\|T^*\|_F^2} \leq \mu_3.
\]

To conclude, the above analysis reveals that
\[
\mu_0 \leq \mu_3, \quad \mu_1 \lesssim \mu_4, \quad \mu_2 \lesssim \mu_4^2, \quad \mu \lesssim \max \{\mu_3, \mu_4^2\} = \mu_{tc} \quad \text{and} \quad \kappa \lesssim \kappa_{tc},
\]

where $\mu = \max \{\mu_0, \mu_1, \mu_2\}$ and $\kappa = \sigma_1(A^*)/\sigma_r(A^*)$. With these estimates in place, Corollary 1 follows immediately from Theorem 1.

### A.2 Proof of Corollary 2

In the problem of covariance estimation with missing data, the ground truth $A^*$ is effectively given by $B^* F^*$, which obeys
\[
A^* = B^* F^* = U^* \Lambda^{1/2} F^* \in \mathbb{R}^{d \times n}, \quad F^* = [f_1^*, \cdots, f_n^*] \in \mathbb{R}^{r \times n}
\]

with $f_i^* \overset{i.i.d.}{\sim} \mathcal{N}(0, I_r)$. We note that by our assumption on the sample size, one has $n \gg \kappa_{ce}^2 (r + \log d)$, where $\kappa_{ce} = \lambda_1^*/\lambda_r^*$. In addition, we note that under the assumption of Corollary 2, one has
\[
\varepsilon_{ce} \ll \kappa_{ce}^{-1} \lesssim 1, \tag{69}
\]

where $\varepsilon_{ce}$ and $\kappa_{ce}$ are defined in (33) and (31), respectively.

#### A.2.1 Estimation error of the principle subspace

In this section, we will prove (32a) and (32b). To begin the proof, we verify the condition of the random noise (cf. (11)). From standard Gaussian concentration results, one is allowed to choose $R \asymp \sigma \sqrt{\log(n + d)}$, 

28
so that $|\eta_{i,j}| \leq R$ for all $i$ and $j$ with probability $1 - O((n + d)^{-12})$. Under our sample size condition that $n \gg \max\left\{ \frac{\log(n + d)}{dp^2}, \frac{\log(n + d)}{p} \right\}$, the requirement (11) is satisfied, namely,

$$\frac{R^2}{\sigma^2} \approx \log(n + d) \lesssim \min\left\{ \frac{p\sqrt{dn}}{\log(n + d)}, \frac{pn}{\log(n + d)} \right\}.$$ 

Next, we turn to the properties of $B^* F^*$ and start by looking at its spectrum. Define

$$C := F^* F^* - nI_r,$$

which allows us to write

$$G^* = B^* F^* (B^* F^*)^T = U^* \Lambda^*^{1/2} F^* F^* \Lambda^*^{1/2} U^* = \underbrace{U^* \Lambda^* U^*}_{=: S^*} + \underbrace{U^* \Lambda^*^{1/2} C \Lambda^*^{1/2} U^*}_{=: \Delta}.$$ (70)

Using standard results on Gaussian random matrices [Ver12], one obtains

$$\|C\| \lesssim \max\left\{ \sqrt{n} \left( \sqrt{r} + \sqrt{\log(n + d)} \right), r + \log(n + d) \right\} \approx \sqrt{n} \left( \sqrt{r} + \sqrt{\log(n + d)} \right),$$

$$|\sigma_i(F^*) - \nu| \lesssim \sqrt{r} + \sqrt{\log(n + d)}$$ (71)

with probability at least $1 - O((n + d)^{-10})$, provided that $n \gg r + \log(n + d)$. It then follows from Weyl's inequality that

$$|\lambda_i(G^*) - \lambda_i(nS^*)| = |\lambda_i(G^*) - \lambda_i^* n| \leq \|\Delta\| \leq \|C\| \|U^*\|^2 \|\Lambda^*\| \lesssim \lambda_i^* \sqrt{n} \left( \sqrt{r} + \sqrt{\log(n + d)} \right).$$ (72)

Under the sample size assumption $n \gg \kappa_{ce}^2 (r + \log(n + d))$, we conclude that

$$\lambda_i(G^*) = \lambda_i^* n (1 + o(1)) \quad \text{and} \quad \sigma_i(B^* F^*) = \sqrt{\lambda_i^* n (1 + o(1))},$$ (73)

and hence

$$\kappa(G^*) = \frac{\lambda_1(G^*)}{\lambda_r(G^*)} \approx \kappa_{ce} \quad \text{and} \quad \kappa(B^* F^*) = \frac{\sigma_1(B^* F^*)}{\sigma_r(B^* F^*)} \approx \sqrt{\kappa_{ce}}.$$ (74)

Further, we look at the entrywise infinity norm of $B^* F^*$. From standard Gaussian concentration inequalities,

$$\|B^* F^*\|_{\infty} = \max_{i,j} \left| \langle (U^* \Lambda^*^{1/2})_i, f^*_j \rangle \right| \lesssim \|U^* \Lambda^*^{1/2}\|_{2,\infty} \sqrt{\log(n + d)}$$

$$\leq \|U^*\|_{2,\infty} \|\Lambda^*\|^{1/2} \sqrt{\log(n + d)} \leq \sqrt{\frac{\lambda_1^* \mu_{ce} r \log(n + d)}{d}}$$

holds with probability at least $1 - O((n + d)^{-10})$. Meanwhile, one has

$$\|B^* F^*\|_F \geq \|B^*\|_F \sigma_r(F^*) \geq \|\Lambda^*^{1/2}\|_F \sigma_r(F^*) \geq \sqrt{\lambda_i^* r n} \geq \frac{1}{\kappa_{ce}} \sqrt{\lambda_i^* r n},$$

where the last step follows from (73) and (74). As a result,

$$\|B^* F^*\|_{\infty} \leq \frac{\mu_{ce} \kappa_{ce} \log(n + d)}{nd} \|B^* F^*\|_F$$ (75)

Recalling the definition of $\mu_0$ in (6), one obtains

$$\mu_0 \lesssim \frac{\mu_{ce} \kappa_{ce} \log(n + d)}{nd}.$$ (76)

When it comes to the incoherence parameters $\mu_1$ and $\mu_2$ (cf. (7)), it can be easily verified that

$$\mu_1 = \frac{d}{r} \|U^*\|_{2,\infty}^2 = \mu_{ce}.$$
In addition, recognizing the existence of an orthonormal matrix $H_2$ such that $V^*H_2 = F^*(F^*F^*)^{-1/2}$, we can bound

\[
\|V^*\|_{2,\infty} = \|V^*H_2\|_{2,\infty} \leq \|F^*T\|_{2,\infty} \left(\|F^*F^*\|\right)^{-1/2} \leq \left(\sqrt{r} + \sqrt{\log(n + d)}\right) \frac{\sqrt{\log(n + d)}}{\sigma_r(F^*)},
\]

where (i) follows from the standard Gaussian concentration result that $\|F^*\|_{2,\infty} - \sqrt{r}$ is of order $\sqrt{\log(n + d)}$ with probability $1 - O((n + d)^{-20})$, (ii) arises from (71), and (iii) holds true under our sample size assumption. Consequently, we obtain

\[
\mu_2 = \frac{n}{r} \|V^*\|_{2,\infty}^2 \leq \log(n + d).
\]

Thus far, we have shown that

\[
\mu_0 \lesssim \mu_{ce}\kappa_{ce} \log(n + d), \quad \mu_1 = \mu_{ce}, \quad \mu_2 \lesssim \log(n + d), \quad \mu \lesssim \mu_{ce}\kappa_{ce} \log(n + d) \quad \text{and} \quad \kappa \lesssim \sqrt{\kappa_{ce}},
\]

where $\mu = \max\{\mu_0, \mu_1, \mu_2\}$ and $\kappa = \kappa(B^*F^*)$. Applying Theorem 1 immediately establishes the claims (32a) and (32b) in Corollary 2. Along the way, we have also established the following upper bound (see Lemma 1), which will be useful in the sequel:

\[
\|G - G^*\| \lesssim \lambda_r(G^*) \cdot \mathcal{E}_{ce} \approx \lambda^*_r n \cdot \mathcal{E}_{ce}. \tag{77}
\]

Here, we recall that $G = \frac{1}{p}P_{\text{off-diag}}(P_{\Omega}(X)P_{\Omega}(X)^\top)$.

### A.2.2 Estimation error of the covariance matrix

It remains to prove (32c) and (32d). Before proceeding, we first recall that $U\Lambda U^\top$ is the top-$r$ eigendecomposition of $G$,

\[
\Sigma = \Lambda^{1/2}, \quad B = \frac{1}{\sqrt{n}}U\Sigma, \quad B^* = U^*\Lambda^{1/2} \quad \text{and} \quad R = \arg\min_{Q \in \mathcal{O}_{r \times r}} \|UQ - U^*\|_F. \tag{78}
\]

Let us also define

\[
K := \arg\min_{Q \in \mathcal{O}_{r \times r}} \|BK - B^*\|_F.
\]

It is well known that the minimizer $K$ is given by $[TB77]$

\[
K = \text{sgn}(B^\top B^*),
\]

where the $\text{sgn}(\cdot)$ function is defined in (56). Since $K$ is an orthonormal matrix, one can express

\[
S - S^* = (BK)(BK)^\top - B^*B^* = (BK - B^*)(BK)^\top + B^*(BK - B^*)^\top. \tag{79}
\]

As a result, everything boils down to controlling $\|BK - B^*\|$ and $\|BK - B^*\|_{2,\infty}$. To this end, we use (78) to reach the following useful decomposition

\[
BK - B^* = \frac{1}{\sqrt{n}}U\Lambda^{1/2} (K - R) + U \left(\frac{1}{\sqrt{n}}\Lambda^{1/2}R - RA^{1/2}\right) + (UR - U^*) \Lambda^{1/2}. \tag{80}
\]

Given that $U\frac{1}{n}\Lambda U^\top$ is the top-$r$ eigendecomposition of $\frac{1}{n}G$, an important step lies in controlling the difference between $\frac{1}{n}G$ and $S^*$. Recalling the matrix $\Delta$ as defined in (70), one can use (72), (77) as well as the definition of $\mathcal{E}_{ce}$ (cf. (33)) to obtain

\[
\left\|\frac{1}{n}G - S^*\right\| \leq \frac{1}{n} \|G - G^*\| + \left\|\frac{1}{n}G^* - S^*\right\| = \frac{1}{n} \|G - G^*\| + \frac{1}{n} \|\Delta\|
\]

30
\[
\frac{1}{n} \lambda_r (G^*) \cdot \mathcal{E}_{ce} + \frac{\lambda_1^*}{\sqrt{n}} (\sqrt{r} + \sqrt{\log (n + d)}) \approx \lambda_r^* \cdot \mathcal{E}_{ce},
\]

where the last inequality makes use of the identity \( \lambda_r (G^*) \approx n \lambda_r^* \). Hence, apply [MWCC17, Lemma 46, Lemma 47] (with slight modification on \( \kappa \)) and Weyl’s inequality to show that

\[
\left\| \frac{1}{\sqrt{n}} A^{1/2} R - RA^* \right\|_{1/2} \lesssim \frac{\kappa (S^*)}{\lambda_r (S^*)} \left\| \frac{1}{n} G - S^* \right\| \lesssim \kappa ce \sqrt{\lambda_r^*} \cdot \mathcal{E}_{ce}; \tag{81}
\]

\[
\left\| K - R \right\| \lesssim \frac{\sqrt{\kappa (S^*)}}{\lambda_r (S^*)} \left\| \frac{1}{n} G - S^* \right\| \lesssim \sqrt{\kappa ce} \cdot \mathcal{E}_{ce}. \tag{82}
\]

In addition, it follows from Weyl’s inequality that

\[
\left\| \frac{1}{n} \Lambda - \Lambda^* \right\| \lesssim \left\| \frac{1}{n} G - S^* \right\| \lesssim \lambda_r^* \cdot \mathcal{E}_{ce},
\]

which combined with (69) gives

\[
\frac{1}{n} \left\| \Lambda \right\| \lesssim \left\| \frac{1}{n} \Lambda - \Lambda^* \right\| + \left\| \Lambda^* \right\| \lesssim \lambda_r^* \cdot \mathcal{E}_{ce} + \sqrt{\lambda_1^*} \approx \lambda_1^*
\]  

under our assumptions.

We are ready to upper bound the difference between \( BK - B^* \). Plugging (32a), (81) (82) and (83) into (80) shows that

\[
\left\| BK - B^* \right\| \leq \frac{1}{\sqrt{n}} \left\| \Lambda \right\|_{1/2} \left\| K - R \right\| + \left\| \frac{1}{\sqrt{n}} A^{1/2} R - RA^* \right\|_{1/2} + \left\| UR - U^* \right\| \left\| \Lambda^* \right\|_{1/2}
\]

\[
\lesssim \sqrt{\kappa ce \lambda_1^*} \cdot \mathcal{E}_{ce} + \kappa ce \sqrt{\lambda_r^*} \cdot \mathcal{E}_{ce} + \sqrt{\lambda_1^*} \cdot \mathcal{E}_{ce}
\]

\[
\lesssim \sqrt{\kappa ce} \cdot \mathcal{E}_{ce}. \tag{84}
\]

Since \( K \in \mathcal{O}^{r \times r} \), this also implies that

\[
\left\| B \right\| = \left\| BK \right\| \leq \left\| BK - B^* \right\| + \left\| B^* \right\| \lesssim \kappa ce \sqrt{\lambda_1^*} \cdot \mathcal{E}_{ce} + \sqrt{\lambda_1^*} \approx \lambda_1^*, \tag{85}
\]

where the last step results from (69). In addition, (32b), (69) and the fact that \( R \in \mathcal{O}^{r \times r} \) guarantees that

\[
\left\| U \right\|_{2, \infty} = \left\| UR \right\|_{2, \infty} \leq \left\| UR - U^* \right\|_{2, \infty} + \left\| U^* \right\|_{2, \infty} \lesssim \kappa^{3/2} ce \sqrt{\frac{\mu ce^r \log (n + d)}{d}} + \sqrt{\frac{\mu ce^r}{d}} \lesssim \sqrt{\kappa ce \mu ce^r \log (n + d)}. \tag{86}
\]

Consequently, it follows from the decomposition (80) that

\[
\left\| BK - B^* \right\|_{2, \infty} \leq \left\| U \right\|_{2, \infty} \frac{1}{\sqrt{n}} \left\| \Lambda \right\|_{1/2} \left\| K - R \right\| + \left\| U \right\|_{2, \infty} \left\| \frac{1}{\sqrt{n}} A^{1/2} R - RA^* \right\|_{1/2} + \left\| UR - U^* \right\|_{2, \infty} \left\| \Lambda^* \right\|_{1/2}
\]

\[
\lesssim \sqrt{\kappa^2 ce^{r} \mu ce^r \lambda_1^* \log (n + d)} \cdot \mathcal{E}_{ce} + \sqrt{\frac{\kappa^3 ce^{r} \mu ce^r \lambda_1^* \log (n + d)}{d}} \cdot \mathcal{E}_{ce} + \sqrt{\frac{\kappa^2 ce^{r} \mu ce^r \lambda_1^* \log (n + d)}{d}} \cdot \mathcal{E}_{ce}
\]

\[
\lesssim \sqrt{\frac{\kappa^2 ce^{r} \mu ce^r \lambda_1^* \log (n + d)}{d}} \cdot \mathcal{E}_{ce}. \tag{86}
\]

Combining (86) and (69) gives that

\[
\left\| B \right\|_{2, \infty} = \left\| BK \right\|_{2, \infty} \leq \left\| BK - B^* \right\|_{2, \infty} + \left\| B^* \right\|_{2, \infty} \lesssim \sqrt{\frac{\kappa^2 ce^{r} \mu ce^r \lambda_1^* \log (n + d)}{d}} \cdot \mathcal{E}_{ce} + \sqrt{\frac{\mu ce^r \lambda_1^*}{d}}
\]

\[
\lesssim \sqrt{\frac{\mu ce^r \lambda_1^* \log (n + d)}{d}}, \tag{87}
\]

31
where we use the fact that \( \|B^*\|_{2,\infty} = \|U^*A^*1/2\|_{2,\infty} \lesssim \|U^*\|_{2,\infty} \|A^*\|^{1/2} \lesssim \sqrt{\mu_{\text{ce}} \lambda_1^*/d} \) and \( \epsilon_{\text{ce}} \ll \kappa_{\text{ce}}^{-1} \).

To finish up, we substitute (84) and (85) into (79) to find that
\[
\|S - S^*\| \lesssim \|BK - B^*\| (\|B^*\| + \|BK\|) \leq \|BK - B^*\| (\|B^*\| + \|B\|) \lesssim \kappa_{\text{ce}} \lambda_1^* \epsilon_{\text{ce}}.
\]

Combining (86) and (87) reveals that
\[
\|S - S^*\|_{\infty} \leq \|BK - B^*\|_{2,\infty} (\|BK\|_{2,\infty} + \|B^*\|_{2,\infty}) \leq \|BK - B^*\|_{2,\infty} (\|B\|_{2,\infty} + \|B^*\|_{2,\infty}) \lesssim \frac{\kappa_{\text{ce}}^2 \mu_{\text{ce}} \lambda_1^* \log (n + d)}{d} \epsilon_{\text{ce}}.
\]

We have therefore established all claims.

A.3 Proof of Corollary 3

Recall from our calculation (39) that
\[
A^* := \mathbb{E}[A] = \frac{(q_{\text{in}} - q_{\text{out}}) \sqrt{n_u n_v}}{2} u^* v^\top
\]
is a rank-1 matrix, where
\[
u^* := \frac{1}{\sqrt{n_u}} \begin{bmatrix} 1_{n_u/2} \n -1_{n_u/2} \end{bmatrix} \quad \text{and} \quad v^* := \frac{1}{\sqrt{n_v}} \begin{bmatrix} 1_{n_v/2} \n -1_{n_v/2} \end{bmatrix}.
\]

Let \( u \in \mathbb{R}^{n_u} \) be the leading eigenvector of \( G \) (cf. (13) and Algorithm 4). To establish Corollary 3, the main step boils down to showing that, under the conditions of Corollary 3,
\[
\min \{ \|u - u^*\|_{\infty}, \|u + u^*\|_{\infty} \} \lesssim \frac{1}{\sqrt{n_u}} \epsilon_{\text{bsbm}}, \tag{88}
\]
holds with probability exceeding \( 1 - O(n^{-10}) \), where
\[
\epsilon_{\text{bsbm}} := \frac{q_{\text{in}}}{(q_{\text{in}} - q_{\text{out}}) \sqrt{n_u n_v}} \log n + \frac{\sqrt{q_{\text{in}}}}{q_{\text{in}} - q_{\text{out}}} \sqrt{\log n} + \frac{1}{\sqrt{n_u}}. \tag{89}
\]

If this claim (88) holds, then under our condition (41) one has \( \epsilon_{\text{bsbm}} \ll 1 \), and hence
\[
\min \{ \|u - u^*\|_{\infty}, \|u + u^*\|_{\infty} \} \lesssim \frac{1}{\sqrt{n_u}} \epsilon_{\text{bsbm}} < \frac{1}{\sqrt{n_u}}.
\]

In other words, one has either \( \text{sign}(u_i) = \text{sign}(u^*_i) \) for all \( 1 \leq i \leq n_u \), or \( \text{sign}(u_i) = -\text{sign}(u^*_i) \) for all \( 1 \leq i \leq n_u \). This tells us that the entrywise rounding operation applied to \( u \) is sufficient to recover exactly the community memberships of all nodes in \( \mathcal{U} \).

The rest of the proof is devoted to establishing the claim (88). In order to apply Theorem 1, it suffices to estimate the spectrum and the incoherence parameters of \( A^* \), as well as some simple statistical properties of \( N := A - A^* \).

- We begin by looking at \( A^* \), which has rank 1 and satisfies
\[
\sigma_1(A^*) = \frac{(q_{\text{in}} - q_{\text{out}}) \sqrt{n_u n_v}}{2}, \quad \|A^*\|_{\infty} = \frac{q_{\text{in}} - q_{\text{out}}}{2}, \quad \|u^*\|_{\infty} = \frac{1}{\sqrt{n_u}}, \quad \|v^*\|_{\infty} = \frac{1}{\sqrt{n_v}}.
\]

Recalling the definition of \( \mu_0, \mu_1, \mu_2 \) in (6) and (7), we obtain
\[
\mu_0 = \frac{n_u n_v}{\|A^*\|_F^2} \|A^*\|_{\infty}^2 = 1, \quad \mu_1 = n_u \|u^*\|_{2,\infty}^2 = 1, \quad \mu_2 = n_v \|v^*\|_{2,\infty}^2 = 1, \quad \kappa = 1.
\]
• Next, we consider the maximum magnitude $R$ and the maximum variance $\sigma^2$ of all entries of $N$ (see Assumption 2). Clearly, one has

$$R = \max_{i,j} |N_{i,j}| \leq 1,$$

$$\sigma^2 = \max_{i,j} \text{Var}(N_{i,j}) = \max \{ q_{in}(1 - q_{in}), q_{out}(1 - q_{out}) \} \leq \max \{ q_{in}, q_{out} \} < q_{in},$$

which follows since $N_{i,j}$ is a centered Bernoulli random variable with parameter either $q_{in}$ or $q_{out}$. From the assumption (41) and the fact $q_{in}^2 \geq (q_{in} - q_{out})^2$, we know that

$$q_{in} \geq \frac{(q_{in} - q_{out})^2}{q_{in}} \geq \frac{\log n}{\sqrt{nn_v}} + \frac{\log n}{n_v}.$$

Putting the above estimates together, we can straightforwardly verify the random noise requirement (11), namely,

$$\frac{R^2}{\sigma^2} \leq \frac{1}{q_{in}} \lesssim \min \left\{ \frac{\sqrt{nn_v}, n_v}{\log n} \right\}.$$

With the preceding bounds in place, Corollary 3 is an immediate consequence of Theorem 1.

B Proofs for key lemmas

This section aims to establish the key lemmas listed in Section 5.2.

B.1 Auxiliary quantities, notation, and preliminary facts

To simplify our treatment, the proofs shall consider the influence of missing data and that of noise altogether. Specifically, throughout this section, we shall define a rescaled version of $A$ as follows

$$A^2 := \frac{1}{p} A = A^* + E \in \mathbb{R}^{d_1 \times d_2},$$

where the matrix $E$ represents the aggregate perturbation

$$E := \frac{1}{p} \mathbb{P}_{\Omega} (A^*) - A^* + \frac{1}{p} \mathbb{P}_{\Omega} (N).$$

Clearly, $E \in \mathbb{R}^{d_1 \times d_2}$ is a random matrix with independent zero-mean entries and $\mathbb{E}[A^*] = A^*$. In addition, we define the corresponding leave-one-out and leave-two-out versions

$$A_{*,(m)} := \frac{1}{p} A^{(m)},$$

$$A_{*,(m,l)} := \frac{1}{p} A^{(m,l)},$$

for each $1 \leq m \leq d_1$, $1 \leq l \leq d_2$.

As we shall see momentarily, it is convenient to introduce the following quantities regarding the above perturbation matrix $E$: (1) $\max_{i \in [d_1], j \in [d_2]} |E_{i,j}|$; (2) $\max_{i \in [d_1], j \in [d_2]} \sqrt{\mathbb{E}[E_{i,j}^2]}$; (3) $\max_{i \in [d_1]} \sqrt{\sum_{j \in [d_2]} \mathbb{E}[E_{i,j}^2]}$; (4) $\max_{j \in [d_2]} \sqrt{\sum_{i \in [d_1]} \mathbb{E}[E_{i,j}^2]}$. In our settings, it is easy to verify — using the definition of incoherence parameters (cf. Definition 1), the assumptions of the random noise (cf. Assumption 2), and Lemma 11 — that the quantities defined above admit the following upper bounds

$$\max_{i \in [d_1], j \in [d_2]} |E_{i,j}| \leq \frac{\|A^*\|_p + R}{p} \lesssim \frac{\mu r \sigma^2}{d_1 d_2 p} + \sigma \min \left\{ \frac{d_1}{d_2}, \frac{d_2}{d_1} \right\} =: B,$$

$$\max_{i \in [d_1], j \in [d_2]} \sqrt{\mathbb{E}[E_{i,j}^2]} \lesssim \frac{\|A^*\|_p + \sigma}{\sqrt{p}} \leq \sigma \sqrt{\frac{d_1 d_2}{d_1 d_2 p}} + \sigma \sqrt{\frac{d_2}{p}} =: \sigma_{\%},$$

$$\max_{i \in [d_1]} \sqrt{\sum_{j \in [d_2]} \mathbb{E}[E_{i,j}^2]} \leq \frac{\|A^*\|_{2,\infty} + \sigma \sqrt{d_2}}{\sqrt{p}} \leq \sigma \sqrt{\frac{\mu r \sigma^2}{d_1 d_2 p}} + \sigma \sqrt{\frac{d_2}{p}} =: \sigma_{\text{row}}.$$
\[
\max_{j \in [d_2]} \sum_{i \in [d_1]} \mathbb{E} [E_{i,j}^2] \leq \frac{\|A^*\|_{2,\infty}^2 + \sigma \sqrt{d_1}}{\sqrt{p}} \leq \sigma_1^2 \frac{\mu_r}{d_2 p} + \sigma \frac{d_1}{p} =: \sigma_{\text{col}},
\]

with probability exceeding \(1 - O(d^{-12})\). Further, the following lemma singles out a few other useful properties about these quantities (to be established in Appendix B.6), which will be useful throughout the proof.

**Lemma 5.** Instate the assumptions of Theorem 1. Then with probability at least \(1 - O(d^{-12})\), we have

\[
\begin{align*}
B &\lesssim \min \left\{ \sqrt{\sigma_{\text{row}} \sigma_{\text{col}}}, \sigma_{\text{row}} \right\}; \\
\sigma^2 &\lesssim B \log d \|A^*\| \sqrt{\frac{\mu_r}{d_2}} \lesssim \sigma_{\text{col}} \sqrt{\log d} \|A^*\| \sqrt{\frac{\mu_r}{d_1}}; \\
\sqrt{\frac{\mu_r}{d_1}} &\gtrsim B \log^{3/2} d \|A^*\|_\infty; \\
\sigma_r^2 &\gtrsim \max \left\{ \kappa^2 \sigma_{\text{col}} \sigma_{\text{row}} \log d, \kappa^2 \sigma_{\text{col}} \sqrt{\log d} \|A^*\|, \kappa^2 \|A^*\|_{2,\infty}^2, \sigma_{\text{row}} \sqrt{\frac{d_1 \log d}{\mu_r}} \|A^*\|_{2,\infty}, B \log d \|A^*\|_\infty \right\}.
\end{align*}
\]

**B.2 Proof of Lemma 1**

The main component of the proof is to demonstrate that

\[
\|G - G^*\| \lesssim (\sigma_{\text{row}} + \sigma_{\text{col}}) (\sigma_{\text{col}} + \|A^*\|_{2,\infty}) \log d + \sigma_{\text{col}} \sqrt{\log d} \|A^*\| + \|A^*\|_{2,\infty}^2,
\]

By substituting the values of \(\sigma_{\text{row}}\) and \(\sigma_{\text{col}}\) (cf. (94)) into the above expression, one derives

\[
\delta_{\text{op}} \lesssim \zeta_{\text{op}} + \|A^*\|_{2,\infty}^2,
\]

where \(\zeta_{\text{op}}\) is defined in (46). Therefore, Lemma 1 is an immediate consequence of (96) and (97). The remainder of the proof amounts to justifying (96).

Recall the definitions of \(G\) and \(G^*\) in (14) and (45), respectively. Given that \(A^* = A^* + E\), we can expand

\[
G - G^* = \mathcal{P}_{\text{off-diag}} (A^* A^* - A^* A^*) = \mathcal{P}_{\text{off-diag}} (A^* A^* - A^* A^*) - \mathcal{P}_{\text{diag}} (A^* A^*)
\]

\[
= \mathcal{P}_{\text{off-diag}} (EE^T) + \mathcal{P}_{\text{off-diag}} (EA^* + A^* E^T) - \mathcal{P}_{\text{diag}} (A^* A^*),
\]

where \(\mathcal{P}_{\text{off-diag}}\) and \(\mathcal{P}_{\text{diag}}\) are defined in Section 2.2. In what follows, we control these three terms separately.

**B.2.1 Step 1: bounding the term \(\mathcal{P}_{\text{off-diag}} (EE^T)\)**

We first consider the term \(\mathcal{P}_{\text{off-diag}} (EE^T)\). Since \(\{E_{i,j}\}_{i \in [d_1], j \in [d_2]}\) are independent zero-mean random variables, we can express

\[
\mathcal{P}_{\text{off-diag}} (EE^T) = \sum_{1 \leq i \leq d_1} (E_{i,-} E_{i}^T - D_i)
\]

as a sum of independent zero-mean random matrices, where \(D_i\) is a random diagonal matrix in \(\mathbb{R}^{d_1 \times d_1}\) with entries

\[
(D_i)_{i,i} = E_{i,i}^2.
\]

We intend to invoke the truncated matrix Bernstein inequality [HSSS16, Proposition A.7] to control the spectral norm of (99). To this end, we need to look at a few quantities.
• We first bound the spectral norm of the following covariance matrix

\[ \Sigma_{ns} := \sum_{1 \leq i \leq d_2} E \left[ (E_{i,l} E_{i,l}^T - D_i)^2 \right] \in \mathbb{R}^{d_1 \times d_1}. \]

Straightforward computation reveals that \( \Sigma_{ns} \) is a diagonal matrix with entries

\[ (\Sigma_{ns})_{i,i} = \sum_{i \leq i \leq d_2} E [E_{i,l}^2] \sum_{i \neq i} E [E_{i,l}^2] \leq \sigma^2_{row} \sigma^2_{col} \]

for each \( 1 \leq i \leq d_1 \). This immediately reveals that

\[ V_{ns} := \| \Sigma_{ns} \| \leq \sigma^2_{row} \sigma^2_{col}. \] (101)

• Next, we turn to upper bounding the spectral norm of each summand \( E_{i,l} E_{i,l}^T - D_i \). As shown in the proof of Lemma 12, one has

\[ \mathbb{P}\left\{ \left\| E_{i,l} \right\|_2^2 - M_1 \geq t \right\} \leq 2 \exp \left( -\frac{3}{8} \min \left\{ \frac{t^2}{V_1}, \frac{t}{L_1} \right\} \right), \quad t > 0, \]

where \( M_1, L_1 \) and \( V_1 \) are given respectively by

\[ M_1 := \mathbb{E} \left[ \| E_{i,l} \|_2^2 \right] \leq \sigma^2_{col}, \]
\[ L_1 := \max_{1 \leq i \leq d_1} \left| E_{i,l}^2 - \mathbb{E} [E_{i,l}^2] \right| \leq 2 B^2, \]
\[ V_1 := \sum_{1 \leq i \leq d_1} \text{Var} (E_{i,l}^2) \leq B^2 \sigma^2_{col}. \]

In addition, with probability exceeding \( 1 - O(d^{-20}) \),

\[ \| E_{i,l} \|_2^2 \leq M_1 + L_1 \log d + \sqrt{V_1} \log d \leq \sigma^2_{col} + B^2 \log d + \sqrt{B^2 \sigma^2_{col} \log d} \leq B^2 \log d + \sigma^2_{col}, \]

where the last line comes from the AM-GM inequality \( 2\sqrt{B^2 \sigma^2_{col} \log d} \leq B^2 \log d + \sigma^2_{col} \). This together with the definition \( D_i := \text{diag}(E_{i,1}^2, \ldots, E_{i,d_1}^2) \) gives

\[ \| E_{i,l} E_{i,l}^T - D_i \| \leq \| E_{i,l} \|_2^2 + \| D_i \| \leq 2 \| E_{i,l} \|_2^2 \leq B^2 \log d + \sigma^2_{col}. \]

Therefore, if we set

\[ L_{ns} := C \left( B^2 \log d + \sigma^2_{col} \right) \] (103)

for some sufficiently large constant \( C > 0 \), then the above argument reveals that

\[ L_{ns} \geq \frac{C}{3} \left( M_1 + L_1 \log d + \sqrt{V_1} \log d \right) \geq \frac{C}{3} \max \left\{ \sqrt{V_1} \log d, L_1 \log d \right\}. \]

• In addition, one can easily bound that

\[ \mathbb{E} \left[ \| E_{i,l} \|_2^2 \mathbb{I}\left\{ \| E_{i,l} \|_2^2 \geq L_{ns} \right\} \right] \leq L_{ns} \mathbb{P}\left\{ \| E_{i,l} \|_2^2 \geq L_{ns} \right\} + \int_{L_{ns}}^{\infty} \mathbb{P}\left\{ \| E_{i,l} \|_2^2 \geq t \right\} dt \leq O(d^{-20}) L_{ns} + \int_{L_{ns}}^{\infty} \mathbb{P}\left\{ \| E_{i,l} \|_2^2 \geq t \right\} dt. \]

Moreover, we know that \( \min \left\{ t^2 / V_1, t / L_1 \right\} \geq t / \max \left\{ \sqrt{V_1} \log d, L_1 \right\} \) for any \( t \geq L_{ns} / 2 \). As a result, for sufficiently large \( d \), we have

\[ \int_{L_{ns}}^{\infty} \mathbb{P}\left\{ \| E_{i,l} \|_2^2 \geq t \right\} dt \leq \int_{L_{ns}}^{\infty} \mathbb{P}\left\{ \| E_{i,l} \|_2^2 - M_1 \geq t / 2 \right\} dt. \]
\[ \leq 4 \int_{\frac{L_{ns}}{L}}^{\infty} \exp \left( -\frac{3}{8} \min \left\{ \frac{t^2}{V_{1}} \frac{t}{L_{t}} \right\} \right) dt \]
\[ \leq 4 \int_{\frac{L_{ns}}{L}}^{\infty} \exp \left( -\frac{3}{8} \max \left\{ \sqrt{V_{1}/\log d}, L_{1} \right\} \right) dt \]
\[ \leq \max \left\{ \sqrt{V_{1}/\log d}, L_{1} \right\} \exp \left( -\frac{3}{16} \max \left\{ \sqrt{V_{1}/\log d}, L_{1} \right\} \right) \]
\[ \leq \max \left\{ \sqrt{V_{1}/\log d}, L_{1} \right\} \exp \left( -\frac{3C}{32} \log d \right) \]
\[ \ll \frac{L_{ns}}{d^2}, \]
provided that \( C > 0 \) is sufficiently large. Consequently, we have
\[
R_{ns} := \mathbb{E} \left[ ||E_{i,t} E_{i,t}^T - D_i||^2 \mathbbm{1} \left\{ ||E_{i,t} E_{i,t}^T - D_i|| \geq L_{ns} \right\} \right] \leq \mathbb{E} \left[ 2 ||E_{i,t}||^2 \mathbbm{1} \left\{ 2 ||E_{i,t}||^2 \geq L_{ns} \right\} \right] \ll \frac{L_{ns}}{d^2}. \tag{104}
\]

With estimates (101), (103) and (104) in place, we are ready to apply the truncated matrix Bernstein inequality [HSSS16, Proposition A.7] to obtain that, with probability at least \( 1 - O \left( d^{-10} \right) \),
\[
\| P_{\text{off-diag}} (EE^T) \| = \left\| \sum_{1 \leq l \leq d_2} E_{i,l} E_{i,l}^T - D_i \right\| \lesssim d_2 R_{ns} + L_{ns} \log d + \sqrt{V_{ns} \log d}
\]
\[ \asymp L_{ns} \log d + \sqrt{V_{ns} \log d} \]
\[ \lesssim B^2 \log^2 d + \sigma_{col}^2 \log d + \sigma_{row} \sigma_{col} \sqrt{\log d} \]
\[ \lesssim \sigma_{col} (\sigma_{row} + \sigma_{col}) \log d, \tag{105} \]
where the last line results from the identity \( B^2 \log d \lesssim \sigma_{row} \sigma_{col} \) (See (95a)).

### B.2.2 Step 2: bounding the term \( P_{\text{off-diag}} (A^* E^T + E A^*) \)

Next, we turn attention to \( P_{\text{off-diag}} (A^* E^T + E A^*) \). By symmetry, it suffices to control to the spectral norm of \( P_{\text{off-diag}} (A^* E^T) \). To this end, we first express
\[
P_{\text{off-diag}} (A^* E^T) = \sum_{1 \leq l \leq d_2} (A^*_{i,l} E_{i,l}^T - \bar{D}_i)
\]
as a sum of independent zero-mean random matrices, where \( \bar{D}_i \) is a diagonal matrix obeying
\[
(\bar{D}_i)_{i,i} = A^*_{i,i} E_{i,i}. \tag{107}
\]

To control (106), we need to first look at two matrices defined as follows
\[
\bar{\Sigma}_{\text{crs}} := \sum_{1 \leq l \leq d_2} \mathbb{E} \left[ (A^*_{i,l} E_{i,l}^T - \bar{D}_i) (A^*_{i,l} E_{i,l}^T - \bar{D}_i)^T \right]
\]
\[
\bar{\Sigma}_{\text{crs}} := \sum_{1 \leq l \leq d_2} \mathbb{E} \left[ (A^*_{i,l} E_{i,l}^T - \bar{D}_i)^T (A^*_{i,l} E_{i,l}^T - \bar{D}_i) \right].
\]

Straightforward computation shows that
\[
(\bar{\Sigma}_{\text{crs}})_{i,i} = \sum_{1 \leq l \leq d_2} A^*_{i,l} \mathbb{E} \left[ ||E_{i,l}||^2 - E_{l,i}^2 \right], \quad i \in [d_2],
\]
\[
(\bar{\Sigma}_{\text{crs}})_{i,j} = \sum_{1 \leq l \leq d_2} A^*_{i,l} A^*_{j,l} \mathbb{E} \left[ ||E_{l,i}||^2 - E_{l,i}^2 - E_{l,j}^2 \right], \quad i \neq j,
\]
and \( \bar{\Sigma}_{\text{crs}} \in \mathbb{R}^{d_1 \times d_1} \) is a diagonal matrix with entries
\[
(\bar{\Sigma}_{\text{crs}})_{i,i} = \sum_{1 \leq l \leq d_2} \left( ||A^*_{l,i}||^2 - A^*_{l,i}^2 \right) \mathbb{E} \left[ E_{l,i}^2 \right], \quad i \in [d_1].
\]
Hence we have

\[
\|\hat{\Sigma}_{\text{crs}}\| \leq \max_{1 \leq i \leq d_1} |(\hat{\Sigma}_{\text{crs}})_{i,i}| \lesssim \sigma_{\text{row}}^2 \|A^*\|^2_{2,\infty}.
\] (108)

To control the spectral norm of \(\hat{\Sigma}_{\text{crs}}\), we further decompose it as \(\hat{\Sigma}_{\text{crs}} = \hat{\Sigma}_{\text{crs}}' - \hat{\Sigma}_{\text{crs}}''\), where

\[
(\hat{\Sigma}_{\text{crs}}')_{i,i} = \sum_{1 \leq l \leq d_2} A_{i,l}^* E \left[|E_{i,l}|^2\right], \quad i \in [d_1],
\]

\[
(\hat{\Sigma}_{\text{crs}}')_{i,j} = \sum_{1 \leq l \leq d_2} A_{i,l}^* A_{j,l}^* E \left[|E_{i,l}|^2\right], \quad i \neq j,
\]

and

\[
(\hat{\Sigma}_{\text{crs}}'')_{i,i} = \sum_{1 \leq l \leq d_2} A_{i,l}^2 E \left[|E_{i,l}|^2\right], \quad i \in [d_1],
\]

\[
(\hat{\Sigma}_{\text{crs}}'')_{i,j} = \sum_{1 \leq l \leq d_2} A_{i,l}^* A_{j,l}^* E \left[|E_{i,l}|^2 + |E_{j,l}|^2\right], \quad i \neq j.
\]

- The spectral norm of \(\hat{\Sigma}_{\text{crs}}'\) can be easily upper bounded by

\[
\|\hat{\Sigma}_{\text{crs}}'\| \leq \max_{1 \leq i \leq d_2} E \left[|E_{i,i}|^2\right] \|A^* A^\top\| \leq \sigma_{\text{col}}^2 \|A^*\|^2.
\] (109)

- Regarding \(\hat{\Sigma}_{\text{crs}}''\), we first decompose \(\hat{\Sigma}_{\text{crs}}'' + \mathcal{P}_{\text{diag}}(\hat{\Sigma}_{\text{crs}}'') = B_1 + B_2\), where the diagonal entries of \(B_1\) and \(B_2\) are identical and equal to \(\sum_{1 \leq l \leq d_2} A_{i,l}^2 E \left[|E_{i,l}|^2\right], (1 \leq i \leq d_2)\) while their off-diagonal parts are given by

\[
(B_1)_{i,j} = \sum_{1 \leq l \leq d_2} A_{i,l}^2 E \left[|E_{i,l}|^2\right] A_{j,l}^* \quad \text{and} \quad (B_2)_{i,j} = \sum_{1 \leq l \leq d_2} A_{i,l}^* A_{j,l}^* E \left[|E_{i,l}|^2 + |E_{j,l}|^2\right], \quad i \neq j.
\]

Let \(C\) be a matrix in \(\mathbb{R}^{d_1 \times d_2}\) with entries \(C_{i,j} = A_{i,j}^* E \left[|E_{i,j}|^2\right]\). One can easily check that \(B_1 = \sum_{1 \leq l \leq d_2} C_{i,l} A_{i,l}^\top = CA^\top\) and develop an upper bound

\[
\|B_1\| \leq \|C\| \|A^*\| \leq \|C\| \|A^*\| \|A^*\| \leq \sigma_{\text{col}}^2 \|A^*\| \|A^*\|.
\]

Note that the same bound also holds for \(B_2\). Therefore, we arrive at

\[
\|\hat{\Sigma}_{\text{crs}}''\| \leq \|\mathcal{P}_{\text{diag}}(\hat{\Sigma}_{\text{crs}}'')\| + \|\mathcal{P}_{\text{diag}}(\hat{\Sigma}_{\text{crs}}'') + \hat{\Sigma}_{\text{crs}}''\| \leq \|\mathcal{P}_{\text{diag}}(\hat{\Sigma}_{\text{crs}}'')\| + \|B_1\| + \|B_2\|
\]

\[
\leq \sigma_{\infty}^2 \|A^*\|_{2,\infty}^2 + \sigma_{\text{col}}^2 \|A^*\| \|A^*\| + \sigma_{\infty}^2 \|A^*\|_{2,\infty}^2 + \sigma_{\text{col}}^2 \|A^*\| \|A^*\|
\]

\[
\leq \sigma_{\infty}^2 \sqrt{r} \|A^*\|_{2,\infty}^2,
\]

where we have used the facts that \(\|A^*\|_{2,\infty} \leq \|A^*\|\) and \(\|A^*\| \leq \sqrt{r} \|A^*\|\). Consequently, the above bounds taken collectively yield

\[
\|\hat{\Sigma}_{\text{crs}}\| \leq \|\hat{\Sigma}_{\text{crs}}'\| + \|\hat{\Sigma}_{\text{crs}}''\| \lesssim (\sigma_{\text{col}}^2 + \sigma_{\infty}^2 \sqrt{r}) \|A^*\|_{2,\infty}^2 \lesssim \sigma_{\text{col}}^2 \|A^*\|_{2,\infty}^2,
\] (110)

where the last step uses (94b) and (94d).

Putting (108), (109) and (110) together yields

\[
V_{\text{crs}} := \max \left\{ \|\hat{\Sigma}_{\text{crs}}\|, \|\hat{\Sigma}_{\text{crs}}'\| \right\} \lesssim \sigma_{\text{col}}^2 \|A^*\|_{2,\infty}^2 + \sigma_{\text{row}}^2 \|A^\top\|_{2,\infty}^2.
\] (111)

Second, we turn to the spectral norm of each summand \(A_{i,l}^* E_{i,l}^\top - \hat{D}_l\). Recalling the definition that \(\hat{D}_l = \text{diag}(A_{1,l}^* E_{1,l}, \ldots, A_{d_2,l}^* E_{d_2,l})\), we can obtain

\[
\|A_{i,l}^* E_{i,l}^\top - \hat{D}_l\| \leq \|A_{i,l}^*\|_{2,\infty} \|E_{i,l}\|_{2,\infty} + \|\hat{D}_l\| \leq 2 \|A_{i,l}^*\|_{2,\infty} \|E_{i,l}\|_{2,\infty} \leq 2 \|A^\top\|_{2,\infty} \|E_{i,l}\|_{2,\infty}.
\]
Set
\[ L_{crs} := C \sqrt{L_{ns}} \| A^* \|_{2,\infty} \lesssim (\sigma_{col} + B \sqrt{\log d}) \| A^* \|_{2,\infty}, \tag{112} \]
where \( L_{ns} \) is defined in (103) and \( C > 0 \) is some sufficiently large universal constant. Then with probability at least \( 1 - O \left( d^{-20} \right) \), one has
\[
\| A^* E_{i,t} - \tilde{D}_t \| \leq 2 \| A^* \|_{2,\infty} \| E_{i,t} \|_2 \lesssim L_{crs},
\]
where the last inequality comes from (102).

Third, we need to control
\[
R_{crs} := \mathbb{E} \left[ \| A^* E_{i,t} - \tilde{D}_t \| \right] \{ \| A^* E_{i,t} - \tilde{D}_t \| \geq L_{crs} \}.\]

From Jensen’s inequality and (104), we know that
\[
\mathbb{E} \left[ \| E_{i,t} \|_2 \right] \{ \| E_{i,t} \|_2 \geq \sqrt{L_{ns}} \} \leq \sqrt{ \mathbb{E} \left[ \| E_{i,t} \|_2^2 \right] \{ \| E_{i,t} \|_2 \geq \sqrt{L_{ns}} \} } \ll \sqrt{\frac{L_{ns}}{d}}.
\]

By the definition of \( L_{crs} \) in (112) and the fact that
\[
\left\{ \| A^* E_{i,t} \|_2 \| E_{i,t} \|_2 \geq L_{crs} \right\} \subset \left\{ \| A^* \|_{2,\infty} \| E_{i,t} \|_2 \geq L_{crs} \right\} = \left\{ \| E_{i,t} \|_2 \geq C \sqrt{L_{ns}} \right\},
\]
one has
\[
R_{crs} \leq \mathbb{E} \left[ 2 \| A^* E_{i,t} \|_2 \| E_{i,t} \|_2 \right] \{ 2 \| A^* E_{i,t} \|_2 \| E_{i,t} \|_2 \geq L_{crs} \} \ll \| A^* \|_{2,\infty} \mathbb{E} \left[ \| E_{i,t} \|_2 \right] \{ \| E_{i,t} \|_2 \geq \frac{C}{2} \sqrt{L_{ns}} \} \ll \frac{1}{d} \| A^* \|_{2,\infty} \sqrt{L_{ns}} \approx \frac{L_{crs}}{d}. \tag{113} \]

With (111), (112) and (113) in place, we can apply the truncated matrix Bernstein inequality to obtain that, with probability at least \( 1 - O \left( d^{-10} \right) \),
\[
\| \mathcal{P}_{\text{off-diag}} \left( A^* E^T \right) \| \leq \| \sum_{1 \leq i \leq d_2} (A^* E_{i,t} - \tilde{D}_i) \| \lesssim d_2 R_{crs} + L_{crs} \log d + \sqrt{V_{crs} \log d}
\]
\[
\approx L_{crs} \log d + \sqrt{V_{crs} \log d}
\]
\[
\lesssim (\sigma_{col} \log d + B \log^{3/2} d) \| A^* \|_{2,\infty} + \sigma_{row} \log d \| A^* \|_{2,\infty} + \sigma_{col} \log d \| A^* \|
\]
\[
\lesssim (\sigma_{row} + \sigma_{col}) \log d \| A^* \|_{2,\infty} + \sigma_{col} \log d \| A^* \| \tag{114}
\]

under the condition (95a) that \( B \sqrt{\log d} \lesssim \sqrt{\sigma_{row} \sigma_{col}} \leq \max \{ \sigma_{row}, \sigma_{col} \} \).

### B.2.3 Step 3: combining Step 1 and Step 2

Taking together (105), (114) and (98), we conclude that
\[
\| G - G^* \| \leq (\sigma_{row} + \sigma_{col}) (\sigma_{col} + \| A^* \|_{2,\infty}) \log d + \sigma_{col} \log d \| A^* \| + \| \mathcal{P}_{\text{diag}} \left( A^* A^T \right) \|
\]
\[
\lesssim (\sigma_{row} + \sigma_{col}) (\sigma_{col} + \| A^* \|_{2,\infty}) \log d + \sigma_{col} \log d \| A^* \| + \| A^* \|_{2,\infty}^2,
\]
where we have used the basic property \( \| \mathcal{P}_{\text{diag}} \left( A^* A^T \right) \| = \| A^* \|_{2,\infty}^2 \).
B.3 Proof of Lemma 2

We first claim that, for any fixed matrix $W$, with probability at least $1 - O(d^{-10})$, the following holds for any $1 \leq m \leq d$:

$$
\| (G - G^*)_{m,:} \|_2 \lesssim \left( \sigma_{\text{col}} (\sigma_{\text{row}} + \|A^*\|_{2,\infty}) \sqrt{\log d} + B \log d \| A^* \|_{\infty} + \| A^* \|_{2,\infty} \right) \| W \|_{2,\infty}
+ \sigma_{\text{row}} \sqrt{\log d} \| A^* \|_{2,\infty} \| W \|. 
$$

(115)

In particular, taking $W = U^*$ gives

$$
\| (G - G^*)_{m,:} U^* \|_2 \lesssim \delta_{\text{row}} \sqrt{\frac{\mu_r \| W \|}{d_1}},
$$

where

$$
\delta_{\text{row}} := \sigma_{\text{col}} (\sigma_{\text{row}} + \|A^*\|_{2,\infty}) \sqrt{\log d} + B \log d \| A^* \|_{\infty} + \sqrt{\frac{d_1 \log d \mu_r}{\| W \|}} \sigma_{\text{row}} \| A^* \|_{2,\infty} + \| A^* \|_{2,\infty}^2. 
$$

(116)

Using the values of $B, \sigma_{\infty}, \sigma_{\text{row}}$ and $\sigma_{\text{col}}$ specified in (94), one can easily verify that

$$
\delta_{\text{row}} \lesssim \zeta_{\text{op}} + \| A^* \|_{2,\infty}^2,
$$

(117)

where $\zeta_{\text{op}}$ is defined in (46). This leads to the advertised bound.

The rest of the proof is thus devoted to proving the claim (115). Recall the definitions of $G$ and $G^*$ in (14) and (45). For any $m, i \in [d_1]$, we can expand

$$
(G - G^*)_{m,i} = \langle A^*_m, A^*_i \rangle - \langle A^*_m, A^*_i \rangle = \langle E_{m,:}, E_{i,:} \rangle + \langle A^*_m, A^*_i \rangle, \quad i \neq m;
$$

$$
(G - G^*)_{m,m} = -G^*_{m,m} = -\| A^* \|^2. 
$$

(119)

This allows us to derive

$$
\| (G - G^*)_{m,:} W \|_2 \leq \left\| \sum_{i:i \neq m} \langle E_{m,:}, E_{i,:} \rangle W_{i,:} \right\|_2 + \left\| \sum_{i:i \neq m} \langle A^*_m, A^*_i \rangle W_{i,:} \right\|_2 + \left\| G^*_{m,m} W_{m,:} \right\|_2.
$$

(118)

We shall control each of these four terms separately.

- For the first term on the right-hand side of (118), we know that

$$
\sum_{i:i \neq m} \langle E_{m,:}, E_{i,:} \rangle W_{i,:} = \sum_{(i,j):i \neq m} E_{m,j} E_{i,j} W_{i,:}
$$

is a sum of independent zero-mean random vectors conditional on $\{E_{m,j}\}_{j \in [d_2]}$. In view of the matrix Bernstein inequality, it suffices to control the following two quantities

$$
L_1 := \max_{(i,j):i \neq m} \| E_{m,j} E_{i,j} W_{i,:} \|_2 \leq \max_{i,j} |E_{i,j}|^2 \| W \|_{2,\infty} \leq B^2 \| W \|_{2,\infty},
$$

$$
V_1 := \sum_{(i,j):i \neq m} E^2_{m,j} \mathbb{E} \left[ |E_{i,j}|^2 \right] \| W \|_{2,\infty}^2 \leq \| W \|_{2,\infty}^2 \sum_{i,j} E^2_{m,i} \sum_{i,j} \mathbb{E} \left[ |E_{i,j}|^2 \right] \leq \| W \|_{2,\infty}^2 \sigma_{\text{col}}^2 \| E_{m,:} \|_{2,\infty}^2, 
$$

where $B$ and $\sigma_{\text{col}}$ are defined in (94). According to Lemma 12, the following holds with probability at least $1 - O(d^{-12})$,

$$
V_1 \lesssim \sigma_{\text{col}}^2 (\sigma_{\text{row}}^2 + B^2 \log d) \| W \|_{2,\infty}^2 \lesssim \sigma_{\text{col}}^2 \sigma_{\text{row}}^2 \| W \|_{2,\infty}^2, 
$$

where we use the condition (95a) (namely, $B \lesssim \sigma_{\text{row}} / \sqrt{\log d}$). Apply the matrix Bernstein inequality to demonstrate that with probability exceeding $1 - O(d^{-12})$,

$$
\left\| \sum_{i:i \neq m} \langle E_{m,:}, E_{i,:} \rangle W_{i,:} \right\|_2 \lesssim L_1 \log d + \sqrt{V_1 \log d} \lesssim (B^2 \log d + \sigma_{\text{col}} \sigma_{\text{row}} \log d) \| W \|_{2,\infty}
\lesssim \sigma_{\text{col}} \sigma_{\text{row}} \sqrt{\log d} \| W \|_{2,\infty}, 
$$

(119)

where the last line follows from (95a) (i.e. $B \lesssim \sqrt{\sigma_{\text{row}} / \sigma_{\text{col}} \log d}$).
• Regarding the second term on the right-hand side of (118), apply the same argument as above to show that
\[
\left\| \sum_{i:i\neq m} \langle A^*_m, E_{i,:} \rangle W_{i,:} \right\|_2 \lesssim (\sigma_{\text{col}} \sqrt{\log d} \| A^* \|_{2,\infty} + B \log d \| A^* \|_{\infty}) \| W \|_{2,\infty}
\]
holds with probability at least \(1 - O(d^{-12})\).

• Turning to the third term on the right-hand side of (118), we have
\[
\left\| \sum_{i:i\neq m} \langle E_{m,:}, A^*_i \rangle W_{i,:} \right\|_2 \leq \left\| \sum_{1\leq j\leq d} E_{m,j} (A^* W)_{j,:} \right\|_2,
\]
where the summands are independent zero-mean random vectors. Let us compute that
\[
L_2 := \max_{j \in [d_2]} \left\| E_{m,j} (A^* W)_{j,:} \right\|_2 \leq B \| A^* W \|_{2,\infty} \leq B \| A^T \|_{2,\infty} \| W \|;
\]
\[
V_2 = \sum_j E \left[ E_{m,j} \right] \left( A^* W \right)_{j,:} \right\|_2 \leq \sigma_{\text{row}}^2 \| A^* W \|_{2,\infty} \leq \sigma_{\text{row}}^2 \| A^T \|_{2,\infty} \| W \|^2.
\]
Then the matrix Bernstein inequality reveals that with probability exceeding \(1 - O(d^{-12})\),
\[
\left\| \sum_j E_{m,j} (A^* W)_{j,:} \right\|_2 \lesssim L_2 \log d + \sqrt{V_2} \log d \lesssim (B \log d + \sigma_{\text{row}} \sqrt{\log d}) \| A^* \|_{2,\infty} \| W \|
\]
\[
\approx \sigma_{\text{row}} \sqrt{\log d} \| A^T \|_{2,\infty} \| W \|,
\]
where the last line follows from the condition (95a) (i.e. \(B \lesssim \sigma_{\text{row}} / \sqrt{\log d}\).

• The last term on the right-hand side of (118) can simply be upper bounded by
\[
\| G_{m,m} W_{m,:} \|_2 \leq \| A^* \|_{2,\infty} \| W \|_{2,\infty}.
\]
Putting (119), (120), (121) and (122) together yields
\[
\left\| (G - G^*)_{m,:} \right\|_2 \lesssim \left( \sigma_{\text{col}} (\sigma_{\text{row}} + \| A^* \|_{2,\infty}) \sqrt{\log d} + B \log d \| A^* \|_{\infty} + \| A^* \|_{2,\infty}^2 \right) \| W \|_{2,\infty}
\]
as claimed.

B.4 Proof of Lemma 3

We claim for the moment that
\[
\left\| UH - GU^* (\Sigma^*)^{-2} \right\|_{2,\infty} \lesssim \left( \frac{\delta_{\text{op}} + \delta_{\text{loo}}}{\sigma^2} \right)^2 \left( \| UH \|_{2,\infty} + \sqrt{\frac{\mu r}{d_1}} \right),\]
where \(\delta_{\text{op}}\) is defined in (96), and \(\delta_{\text{loo}}\) is defined as follows:
\[
\delta_{\text{loo}} := \sigma_{\text{col}} \sigma_{\text{row}} \log d + \sigma_{\text{col}} \| A^* \| \sqrt{\log d}.
\]
Using the values of \(\sigma_{\text{row}}\) and \(\sigma_{\text{col}}\) specified in (94), one can easily see that \(\delta_{\text{loo}} \lesssim \zeta_{\text{op}}\), where \(\zeta_{\text{op}}\) is defined in (46). In addition, recall that we have already shown that \(\delta_{\text{op}} \lesssim \zeta_{\text{op}} + \| A^* \|_{2,\infty}^2\). Putting these together establishes the lemma.

We now start to prove the claim (123). To this end, consider an arbitrary \(m \in [d_1]\). In view of [AFWZ17, Lemma 2], we can decompose
\[
\left\| (UH - GU^* (\Sigma^*)^{-2})_{m,:} \right\|_2 \lesssim \left( \frac{1}{\sigma^2} \right) \left| G - G^* \right| \| G_{m,:} U^* \|_2^2 + \frac{1}{\sigma^2} \| G_{m,:} (UH - U^*) \|_2.
\]
• To bound the first term of (125), we apply Lemma 2, (117) and Fact 1 to reach

$$\| (G - G^*) U^* \|_{2, \infty} \lesssim \delta_{\text{row}} \sqrt{\frac{\mu r}{d_1}} \ll \sigma_r^2 \sqrt{\frac{\mu r}{d_1}}.$$  

The triangle inequality then gives

$$\| GU^* \|_{2, \infty} \leq \| (G - G^*) U^* \|_{2, \infty} + \| G^* \| \| U^* \|_{2, \infty} \lesssim \sigma_r^2 \sqrt{\frac{\mu r}{d_1}}.$$  

(126)

This taken collectively with the upper bound on $\| G - G^* \|$ (cf. (96)) gives

$$\frac{1}{\sigma_r^2} \| G - G^* \| \| G_m U^* \|_2 \leq \frac{1}{\sigma_r^2} \| G - G^* \| \| GU^* \|_{2, \infty} \lesssim \frac{\delta_{\text{op}} \kappa_2^2}{\sigma_r^2} \sqrt{\frac{\mu r}{d_1}}.$$  

(127)

• Turning to the second term of (125), we start with the following bound

$$\| G_m \ (U H - U^*) \|_2 \leq \| G_m (U H - U (m) H (m)) \|_2 + \| G_m (U (m) H (m) - U^*) \|_2.$$  

Lemma 2 tells us that

$$\| G_m \|_2 \leq \| (G - G^*)_m \|_2 + \| G^* \|_{2, \infty} \lesssim \sigma_r^2 \| U^* \|_{2, \infty} \leq \sigma_r^2 \| U^* \| = \sigma_r^2,$$

which makes use of the fact that $\| G^* \|_{2, \infty} \leq \| U^* \|_{2, \infty} \| \Sigma^* \| \| U^* \| = \sigma_r^2 \| U^* \|_{2, \infty}$. This combined with (130) (to be established shortly in the proof of Lemma 4) and the definitions of $H$ and $H (m)$ gives

$$\| G_m (U (m) H (m) - U^*) \|_2 \leq \| G_m \|_2 \| U (m) H (m) - U^* \|$$

$$= \| G_m \|_2 \| (U (m) U (m)^T - U U^T) U^* \|$$

$$\lesssim \| G_m \|_2 \| (U (m) U (m)^T - U U^T) \|$$

$$\lesssim \frac{\sigma_r^2}{\sigma_r^2} \delta_{\text{row}} \left( \| U H \|_{2, \infty} + \frac{\| \mu r \|}{\| d_1 \|} \right)$$

$$\lesssim \delta_{\text{row}} \kappa_2 \left( \| U H \|_{2, \infty} + \frac{\| \mu r \|}{\| d_1 \|} \right).$$  

(128)

In addition, Lemma 10 shows that with probability at least $1 - O(d^{-11})$,

$$\| G_m (U (m) H (m) - U^*) \|_2 \lesssim \delta_{\text{row}} \| U (m) H (m) - U^* \|_{2, \infty} + \delta_{\text{op}} \kappa_2^2 \sqrt{\frac{\mu r}{d_1}}$$

$$\lesssim \delta_{\text{row}} \| U (m) H (m) \|_{2, \infty} + \delta_{\text{row}} \| U^* \|_{2, \infty} + \delta_{\text{op}} \kappa_2^2 \sqrt{\frac{\mu r}{d_1}}$$

$$\lesssim \delta_{\text{row}} \| U H \|_{2, \infty} + (\delta_{\text{row}} + \delta_{\text{op}} \kappa_2^2) \sqrt{\frac{\mu r}{d_1}},$$  

(129)

where the inequality (129) results from (140) (also established shortly in the proof of Lemma 4).

Then claim immediately follows from (127), (128), (129) and the union bound.

B.5 Proof of Lemma 4

To begin with, recalling the definition of $\delta_{\text{row}}$ (cf. (124)), we claim that

$$\| U (m) U (m)^T - U U^T \|_F = \frac{1}{\sigma_r^2} \left( \sigma_{\text{col}} \sigma_{\text{row}} \log d + \sigma_{\text{col}} \| A^* \| \sqrt{\log d} \right) \left( \| U H \|_{2, \infty} + \sqrt{\frac{\mu r}{d_1}} \right).$$  

(130)
As mentioned before, one has \( \delta_{\text{op}} \lesssim \zeta_{\text{op}} \), from which the lemma follows immediately. The rest of the proof thus boils down to proving the claim (130).

We shall apply the Davis-Kahan sin \( \Theta \) theorem [DK70] to derive

\[
\|UU^T - U^{(m)}(m)^T\|_F \leq \frac{\|G - G^{(m)}\|U^{(m)}\|_F}{\lambda_r(G^{(m)}) - \lambda_{r+1}(G)} \leq \frac{2\|G - G^{(m)}\|U^{(m)}\|_F}{\sigma^2_r},
\]

(131)

Here, the last inequality follows since, by Weyl’s inequality,

\[
\lambda_r(G^{(m)}) - \lambda_{r+1}(G) \geq \lambda_r(G^*) - \|G^{(m)} - G^*\| - \|G - G^*\| = \sigma^2_r - \|G^{(m)} - G^*\| - \|G - G^*\| \\
\geq \frac{\sigma^2_r}{2},
\]

(132)

where the last line follows since \( \|G^{(m)} - G^*\| \lesssim \delta_{\text{op}} \leq \sigma^2_r \) — an immediate consequence of Lemma 6 and Condition (95d). As a side note, the fact \( \|G^{(m)} - G^*\| \leq \sigma^2_r \) also implies (according to [AFW17, Lemma 3])

\[
\|(H^{(m)})^{-1}\| \lesssim 1,
\]

(133)

which will be useful later.

It remains to control the term \( \|(G - G^{(m)})U^{(m)}\|_F \) in (131). Recall the definitions of \( G \) and \( G^{(m)} \) in (14) and (43), respectively. It is straightforward to see that \( G - G^{(m)} \) is a rank-2 symmetric matrix with nonzero entries located only in the \( m \)-th row and the \( m \)-th column. Simple calculation reveals that

\[
(G - G^{(m)})_{m,i} = \langle E_{m,:}, A^*_{i,:} \rangle, \quad i \neq m;
\]

(134)

\[
(G - G^{(m)})_{m,m} = 0.
\]

(135)

We can then derive

\[
\|(G - G^{(m)})U^{(m)}\|_F = \|(G - G^{(m)})U^{(m)}H^{(m)}(H^{(m)})^{-1}\|_F \leq \|(G - G^{(m)})U^{(m)}H^{(m)}\|_F\|(H^{(m)})^{-1}\| \\
\lesssim \|(G - G^{(m)})U^{(m)}H^{(m)}\|_F \\
\leq \|\mathcal{P}_{m,:}(G - G^{(m)})U^{(m)}H^{(m)}\|_F + \|\mathcal{P}_{:,m}(G - G^{(m)})U^{(m)}H^{(m)}\|_F,
\]

(136)

where the second line arises due to (133), and \( \mathcal{P}_{m,:} \) (resp. \( \mathcal{P}_{:,m} \)) is the projection onto the subspace of matrix supported on \( \{m\} \times [d_2] \) (resp. \( [d_1] \times \{m\} \)).

To bound the first term of (136), we make the observation (using (134) and (135)) that

\[
\|\mathcal{P}_{m,:}(G - G^{(m)})U^{(m)}H^{(m)}\|_F = \|(A^* - A^*)_{m,:} \mathcal{P}_{-,m,:}(A^*)^T U^{(m)}H^{(m)}\|_F.
\]

Controlling this quantity requires the assistance of leave-two-out matrices. Here, we only state our bound: with probability at least \( 1 - O(d^{-11}) \), one has

\[
\|\mathcal{P}_{m,:}(G - G^{(m)})U^{(m)}H^{(m)}\|_F \lesssim \sigma_{\text{col}}(\sigma_{\text{raw}} \log d + \|A^*\| \sqrt{\log d}) \left( \|U^{(m)}H^{(m)}\|_{2,\infty} + \sqrt{\frac{\mu}{d_1}} \right).
\]

(137)

This bound will be restated in Lemma 7 and established in Appendix C. Turning to the second term of (136), we apply Lemma 6 (also established in Appendix C) to obtain

\[
\|\mathcal{P}_{m,:}(G - G^{(m)})U^{(m)}H^{(m)}\|_F \leq \left( \sum_{1 \leq i \leq d_1} (G - G^{(m)})_{i,m} \right)^{1/2} \|U^{(m)}H^{(m)}\|_{m,:} \\
\leq \|G - G^{(m)}\|\|U^{(m)}H^{(m)}\|_{2,\infty} \\
\leq \left( \sigma_{\text{col}}(\sigma_{\text{raw}} + \|A^*\|_{2,\infty}) \sqrt{\log d} \right) \|U^{(m)}H^{(m)}\|_{2,\infty}
\]

(138)
with probability at least $1 - O\left(d^{-11}\right)$. Hence, we can combine (137), (138) and (131) to yield

$$
\|U^{(m)}U^{(m)\top} - UU\top\|_F \\
\leq \left(\sigma_{\text{col}}\sigma_{\text{row}}\log d + \sigma_{\text{col}}\|A^*\|\sqrt{\log d}\right)\left(\|U^{(m)}H^{(m)}\|_{2,\infty} + \sqrt{\frac{\mu\|r\|}{d_1^*}}\right)
$$

$$
= \frac{\delta_{\text{loq}}}{\sigma_r^2}\left(\|U^{(m)}H^{(m)}\|_{2,\infty} + \sqrt{\frac{\mu\|r\|}{d_1^*}}\right),
$$

(139)

where $\delta_{\text{loq}}$ is defined in (124). As a result, the proof is complete as long as we can show that

$$
\|U^{(m)}H^{(m)}\|_{2,\infty} \lesssim \|UH\|_{2,\infty} + \sqrt{\frac{\mu\|r\|}{d_1}}.
$$

(140)

To finish up, it remains to justify this inequality (140). To this end, from the definitions of $H^{(m)}$ and $H$ we have

$$
\|U^{(m)}H^{(m)}\|_{2,\infty} \leq \|U^{(m)}H^{(m)} - UH\|_{2,\infty} + \|UH\|_{2,\infty}
$$

$$
= \|\left(U^{(m)}U^{(m)\top} - UU\top\right)U^*\|_{2,\infty} + \|UH\|_{2,\infty}
$$

$$
\leq \|U^{(m)}U^{(m)\top} - UU\top\|_F\|U^*\| + \|UH\|_{2,\infty}
$$

$$
= \|U^{(m)}U^{(m)\top} - UU\top\|_F + \|UH\|_{2,\infty}.
$$

(141)

Under the condition (95d), it is easily seen that $\delta_{\text{loq}} \ll \sigma_r^2$. This together with (139) gives

$$
\|U^{(m)}U^{(m)\top} - UU\top\|_F \leq 0.5\|U^{(m)}H^{(m)}\|_{2,\infty} + 0.5\sqrt{\frac{\mu\|r\|}{d_1}}.
$$

(142)

which combined with (141) yields

$$
\|U^{(m)}H^{(m)}\|_{2,\infty} \leq 2\|UH\|_{2,\infty} + \sqrt{\frac{\mu\|r\|}{d_1}}
$$

(143)

as claimed.

### B.6 Proof of Lemma 5

We start with (95a). In view of the definitions of $B, \sigma_{\infty}, \sigma_{\text{row}}$ and $\sigma_{\text{col}}$ in (94), we have with probability at least $1 - O\left(d^{-12}\right)$,

$$
B^2 = \frac{\mu\|r\|^2}{d_1 d_2 p^2} + \frac{\sigma^2 \min\{\sqrt{d_1 d_2}, d_2\}}{p \log d},
$$

$$
\sigma^2_{\text{row}} = \frac{\mu\|r\|^2}{d_1 p} + \frac{\sigma^2 d_2}{p},
$$

$$
\sigma_{\text{row}}\sigma_{\text{col}} = \frac{\mu\|r\|^2}{\sqrt{d_1 d_2} p} + \frac{2\sigma_1^* \sqrt{\|r\|}}{p} + \frac{\sigma^2 \sqrt{d_1 d_2}}{p} \approx \frac{\mu\|r\|^2}{\sqrt{d_1 d_2} p} + \frac{\sigma^2 \sqrt{d_1 d_2}}{p},
$$

where we have used the AM-GM inequality (i.e. $2\sigma_1^* \sqrt{\|r\|} \leq \frac{\mu\|r\|^2}{\sqrt{d_1 d_2}} + \sigma^2 \sqrt{d_1 d_2}$) in the last line. Therefore,

$$
B^2 \log d \lesssim \sigma_{\text{row}}\sigma_{\text{col}} \quad \text{and} \quad B^2 \log d \lesssim \sigma_{\text{row}}^2.
$$

hold as long as $p \gtrsim (d_1 d_2)^{-1/2} \log d$ and $p \gtrsim d_2^{-1} \log d$.

The next step is to establish (95b). Let us consider the first inequality. By (94), it is easily seen that

$$
\sigma^2_{\infty} = \frac{\mu\|r\|^2}{d_1 d_2 p} + \frac{\sigma^2}{p};
$$
\[
B \log d \|A^*\| \sqrt{\frac{\mu r}{d_2}} = \frac{\mu r \sigma^2 \log d}{\sqrt{\log d}} + \sigma r \min \left\{ \sqrt{d_1 d_2}, \sqrt{d_2} \right\} \sqrt{\frac{\mu r \log d}{d_2 p}}.
\]

As a consequence, the first inequality holds as long as \( \frac{\mu r}{\sigma^2} \lesssim \min \left\{ \sqrt{d_1 d_2}, \sqrt{d_2} \right\} \sqrt{\frac{\mu r \log d}{d_2}} \), which is satisfied by our noise assumption that \( \frac{\mu r}{\sigma^2} \ll \frac{\sqrt{\mu r \log d}}{\sqrt{d_1 d_2}} \). To show the second inequality, we note that

\[
B \sqrt{\log d} \|A^*\| \sqrt{\frac{\mu r}{d_2}} \leq \left\{ \frac{\sigma r \sqrt{\mu r \log d}}{d_2 p} + \sigma r \sqrt{d_1} \right\} \sqrt{\frac{\mu r}{d_1}}.
\]

Recognizing that \( \sigma_{\text{col}} \|A^*\| = \sigma r \sqrt{\mu r \log d} \), we prove the second inequality provided that \( p \gtrsim d_2^{-1} \log d \).

When it comes to (95c): by virtue of Lemma 11 and (94), one has

\[
\frac{B \log^{3/2} d \|A^*\|}{\|A^*\|^2} \leq \frac{\mu r \log^{3/2} d}{d_1 d_2 p} + \frac{\sigma \sqrt{\mu r \log d}}{\sigma_1 r \sqrt{d_1 d_2} \sqrt{\log d}}.
\]

Consequently, (95c) holds provided \( p \gtrsim \frac{\sqrt{\mu r \log^{3/2} d}}{\sqrt{d_1 d_2}} \) and \( \sigma_1 \lesssim \frac{\sqrt{d_1 \log d}}{\sqrt{d_1 d_2} \sqrt{\log d}} \), which holds under our assumptions that \( p \gg \frac{\mu r \log^{3/2} d}{d_1 d_2} \) and \( \sigma_1 \ll \frac{\sqrt{d_1 \log d}}{\sqrt{d_1 d_2} \sqrt{\log d}} \).

Finally, using the definitions in (94) and Lemma 11, one obtains the following bounds:

\[
\sigma_{\text{row}} \sigma_{\text{col}} \log d \geq \frac{\mu r \sigma^2 \log d}{\sqrt{d_1 d_2} p} + \frac{\sigma^2 \sqrt{d_1 d_2} \log d}{p},
\]

\[
\sigma_{\text{col}} \sqrt{\log d} \|A^*\| = \sigma^2 \sqrt{\mu r \log d} \sqrt{d_1} \log d ;
\]

\[
\|A^*\|^2 \leq \frac{\mu r \sigma^2 \log d}{d_1} ;
\]

\[
\sigma_{\text{row}} \sqrt{\frac{d_1 \log d}{\mu r}} \|A^*\| \leq \sigma^2 \sqrt{\mu r \log d} \sqrt{d_1} \log d ;
\]

\[
B \|A^*\| \log d \leq \frac{\mu r \sigma^2 \log d}{d_1 d_2 p} + \frac{\sigma^2 \sqrt{\mu r \log d}}{\sqrt{d_1 d_2} \sqrt{\log d}}.
\]

Therefore, it is easy to verify (95d) under the assumptions of Theorem 1.

C Proofs for auxiliary lemmas

This section establishes several useful technical lemmas useful for proving our main theorems. Throughout this section, we shall frequently use the quantities \( B, \sigma_\infty, \sigma_{\text{row}} \) and \( \sigma_{\text{col}} \) defined in (94). In fact, it suffices to bear in mind the following bounds

\[
B \geq \max_{i,j} |E_{i,j}|; \quad \sigma_\infty \geq \max_{i,j} \sqrt{\frac{\text{Var}(E_{i,j})}}{\text{Var}(E_{i,j})};
\]

\[
\sigma_{\text{row}} \geq \max_i \sqrt{\sum_j \text{Var}(E_{i,j})}; \quad \sigma_{\text{col}} \geq \max_j \sqrt{\sum_i \text{Var}(E_{i,j})}.
\]

C.1 Auxiliary technical lemmas

We first gather all technical lemmas to be established in this section, and begin with the following lemma, which shows that the leave-one-out sequence \( G^{(m)} \) is close to \( G \) and \( G^* \) when measured by the spectral norm.
Lemma 6. Instate the assumptions of Theorem 1. With probability at least $1 - O(d^{-11})$, one has
\begin{align*}
\|G^{(m)} - G\| &\lesssim \sigma_{\text{row}} (\sigma_{\text{col}} + \|A^*\|_{2,\infty}) \sqrt{\log d}, \\
\|G^{(m)} - G^*\| &\lesssim \delta_{\text{op}} = (\sigma_{\text{row}} + \sigma_{\text{col}}) (\sigma_{\text{col}} + \|A^*\|_{2,\infty}) \log d + \sigma_{\text{col}} \sqrt{\log d} \|A^*\| + \|A^*\|_{2,\infty}^2,
\end{align*}
where $\sigma_{\text{row}}$ and $\sigma_{\text{col}}$ are defined in (94).

Proof. See Appendix C.2.

Similar to $H$ (defined in (47)), we also introduce the following matrices for each $(m, l) \in [d_1] \times [d_2]$:
\begin{align*}
H^{(m)} &:= U^{(m)\top} A^*, \\
H^{(m,l)} &:= U^{(m,l)\top} A^*,
\end{align*}
where $U^{(m)}$ and $U^{(m,l)}$ are defined in Algorithm 5 and Algorithm 6, respectively.

Lemma 7 serves a crucial step towards proving Lemma 4.

Lemma 7. Instate the assumptions of Theorem 1. For any fixed $1 \leq m \leq d_1$, with probability at least $1 - O(d^{-11})$, one has
\begin{align*}
\| (A^* - A^*)_{m, \mathcal{P}_{-m,:}} (A^*)_{m,:} U^{(m)\top} H^{(m)} \|_2 \\
\lesssim \sigma_{\text{col}} (\sigma_{\text{row}} \log d + \|A^*\| \sqrt{\log d}) \left( \|U^{(m)\top} H^{(m)}\|_{2,\infty} + \sqrt{\frac{\mu r}{d_1}} \right),
\end{align*}
where $A^*$ is defined in (90), and $\sigma_{\text{row}}$ and $\sigma_{\text{col}}$ are both defined in (94).

Proof. See Appendix C.3.

The proof of Lemma 7 relies on an upper bound on the $\ell_{2,\infty}$ norm of $\mathcal{P}_{-m,:} (A^*)_{m,:} U^{(m)\top} H^{(m)}$, which is formalized below in Lemma 8. This is built upon a leave-two-out argument.

Lemma 8. Instate the assumptions of Theorem 1. With probability at least $1 - O(d^{-10})$, the following holds simultaneously for all $m \in [d_1]$,
\begin{align*}
\| \mathcal{P}_{-m,:} (A^*)_{m,:} U^{(m)\top} H^{(m)} \|_{2,\infty} \lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|U^{(m)\top} H^{(m)}\|_{2,\infty} + \|A^*\| \sqrt{\frac{\mu r}{d_2}},
\end{align*}
where $A^*$ is defined in (90), and $B$ and $\sigma_{\text{col}}$ are defined in (94).

Proof. See Appendix C.4.

The proof of Lemma 8 requires the proximity between $U^{(m)}$ and $U^{(m,l)}$, which is demonstrated below in Lemma 9.

Lemma 9. Instate the assumptions of Theorem 1. With probability at least $1 - O(d^{-10})$, the following holds simultaneously for any $m \in [d_1]$ and $l \in [d_2]$,
\begin{align*}
\|U^{(m)} U^{(m)\top} - U^{(m,l)} U^{(m,l)\top}\| \lesssim \frac{1}{\sigma_r^2} \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right)^2 \|U^{(m)} H^{(m)}\|_{2,\infty} \\
+ \frac{\sigma_{\infty}^2}{\sigma_r^2} + \frac{1}{\sigma_r^2} \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right) \|A^*\|_{2,\infty},
\end{align*}
where $B, \sigma_{\infty}$ and $\sigma_{\text{col}}$ are defined in (94).

Proof. See Appendix C.5.

Finally, Lemma 10 stated below constitutes the main part of Lemma 3 (recalling the decomposition in (49) and (50)).
Lemma 10. Instate the assumptions of Theorem 1. For each fixed \( m \in [d_1] \), the following holds with probability exceeding \( 1 - O\left(d^{-11}\right) \),
\[
\left\| G_{m,:}(U^{(m)}H^{(m)} - U^*) \right\|_2 \lesssim \delta_{\text{col}} \left\| U^{(m)}H^{(m)} - U^* \right\|_{2,\infty} + \delta_{\text{op}}k^2 \sqrt{\frac{mp}{d_1}},
\]
where \( \delta_{\text{col}} \) and \( \delta_{\text{op}} \) are defined in (46) and (124), respectively.

Proof. See Appendix C.6.

C.2 Proof of Lemma 6

Recall the definitions of \( G \) and \( G^{(m)} \) in (14) and (43). As shown in (134) in the proof of Lemma 4 in Appendix B.5, we know that \( G - G^{(m)} \) is a rank-2 symmetric matrix with nonzero entries located only in the \( m \)-th row and the \( m \)-th column. In particular, one has
\[
(G - G^{(m)})_{m,i} = \langle E_{m,:}, A^*_i \rangle, \quad i \neq m,
\]
\[
(G - G^{(m)})_{m,m} = 0,
\]
thus indicating that
\[
(G - G^{(m)})_{m,:} = E_{m,:}[P_{-m,:}(A^*)^T.
\]
This allows us to upper bound
\[
\left\| G - G^{(m)} \right\| \leq \left\| G - G^{(m)} \right\|_F \lesssim \left\| (G - G^{(m)})_{m,:} \right\|_2
\]
\[
= \left\| E_{m,:}[P_{-m,:}(A^*)^T \right\|_2
\]
\[
\lesssim \sigma_{\text{row}}(\sigma_{\text{col}} + \|A^*\|_{2,\infty})\sqrt{\log d}.
\]
Here, the last line follows the following. First, notice \( A^* = A^* + E \) and
\[
E_{m,:}[P_{-m,:}(A^* + E)]^T = \sum_{i:i\neq m} \langle E_{m,:), E_i \rangle e_i^T + \sum_{i:i\neq m} \langle E_{m,:}, A^* i \rangle e_i^T,
\]
where \( e_i \) is the \( i \)-th standard basis in \( \mathbb{R}^{d_1} \). It follows from (119) and (121) shown in the proof of Lemma 2 (cf. Appendix B.3) that with probability at least \( 1 - O(d^{-11}) \),
\[
\left\| E_{m,:}[P_{-m,:}(A^* + E)]^T \right\|_2 \leq \left\| \sum_{i:i\neq m} \langle E_{m,:}, E_i \rangle e_i^T \right\|_2 + \left\| \sum_{i:i\neq m} \langle E_{m,:}, A^* i \rangle e_i^T \right\|_2
\]
\[
\lesssim \sigma_{\text{col}}\sigma_{\text{row}}\sqrt{\log d} + \sigma_{\text{row}}\sqrt{\log d}\|A^*\|_{2,\infty}.
\]
In addition, the above bound combined with Lemma 1 immediately yields (146). The proof is complete by taking the union bound over \( 1 \leq m \leq d_1 \).

C.3 Proof of Lemma 7

By construction, the \( m \)-th row of \( A^* - A^* \) is independent of \( [P_{-m,:}(A^*)]^T U^{(m)}H^{(m)} \). As a result,
\[
(A^* - A^*)_{m,:} [P_{-m,:}(A^*)]^T U^{(m)}H^{(m)} = \sum_{j \in [d_2]} E_{m,j} \left( [P_{-m,:}(A^*)]^T U^{(m)}H^{(m)} \right)_j,
\]
can be viewed as a sum of independent zero-mean random vectors (where the randomness comes from \( \{E_{m,j}\}_{j \in [d_2]} \)). It is straightforward to calculate that
\[
L := \max_{j \in [d_2]} \left\| E_{m,j} \left( [P_{-m,:}(A^*)]^T U^{(m)}H^{(m)} \right)_j \right\|_2 \leq B \left\| P_{-m,:}(A^*)^T U^{(m)}H^{(m)} \right\|_{2,\infty},
\]
\[ V := \sum_{j \in [d_2]} \mathbb{E} \left[ E_{m,j}^2 \right] \left\| \left[ \mathcal{P}_{m,:} (A^\ast) \right]^\top U^{(m)} H^{(m)} \right\|_2^2 \leq \sigma_2^\ast \left\| \mathcal{P}_{m,:} (A^\ast) \right\|_2^2 \left\| U^{(m)} H^{(m)} \right\|_F^2. \]

In view of the matrix Bernstein inequality, it boils down to controlling \( L \) and \( V \). To this end, let us first bound \( L \). From Lemma 8, one has that with probability at least \( 1 - O \left( d^{-11} \right) \),

\[ L \lesssim B (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \left\| U^{(m)} H^{(m)} \right\|_{2,\infty} + B \left\| A^\ast \right\| \sqrt{\frac{\mu r}{d_2}}. \quad (149) \]

Regarding \( V \), Lemma 13 guarantees the following upper bound with probability exceeding \( 1 - O \left( d^{-11} \right) \),

\[
\left\| \left[ \mathcal{P}_{m,:} (A^\ast) \right]^\top U^{(m)} H^{(m)} \right\|_F \leq \left\| \mathcal{P}_{m,:} (A^\ast) \right\| \left\| U^{(m)} H^{(m)} \right\|_F
\leq \sqrt{d_1} \left\| A^\ast \right\| \left\| U^{(m)} H^{(m)} \right\|_{2,\infty}
\lesssim \sqrt{d_1} \left( B \log d + (\sigma_{\text{row}} + \sigma_{\text{col}}) \sqrt{\log d} + \left\| A^\ast \right\| \right) \left\| U^{(m)} H^{(m)} \right\|_{2,\infty}
\lesssim \sqrt{d_1} \left( \sigma_{\text{row}} \sqrt{\log d} + \left\| A^\ast \right\| \right) \left\| U^{(m)} H^{(m)} \right\|_{2,\infty},
\]

where the last inequality follows from the condition (95d) that \( B \log d + \sigma_{\text{col}} \sqrt{\log d} \ll \sigma_2^\ast \). Applying the matrix Bernstein inequality yields that with probability at least \( 1 - O \left( d^{-11} \right) \): one has

\[
\left\| (A^\ast - A^\ast)_{m,:} [\mathcal{P}_{m,:} (A^\ast)]^\top U^{(m)} H^{(m)} \right\|_2 \lesssim L \log d + \sqrt{\log d}
\lesssim (B^2 \log^2 d + B \sigma_{\text{col}} \log^3 2 d) \left\| U^{(m)} H^{(m)} \right\|_{2,\infty} + B \log d \left\| A^\ast \right\| \sqrt{\frac{\mu r}{d_2}}
\lesssim \sqrt{d_1} \sigma_2 \left( \sigma_{\text{row}} \log d + \left\| A^\ast \right\| \sqrt{\log d} \right) \left\| U^{(m)} H^{(m)} \right\|_{2,\infty} + \sigma_{\text{col}} \left\| A^\ast \right\| \sqrt{\log d} \sqrt{\frac{\mu r}{d_1}}
\lesssim \sigma_{\text{col}} \left( \sigma_{\text{col}} \log d + \left\| A^\ast \right\| \sqrt{\log d} \right) \left( \left\| U^{(m)} H^{(m)} \right\|_{2,\infty} + \sqrt{\frac{\mu r}{d_1}} \right),
\]

where (i) uses (94) that \( \sigma_{2,\text{col}} \simeq d_1 \sigma_2 \), as well as the conditions (95a), (95d) and (95b) (namely, \( B \lesssim \sqrt{\sigma_{\text{row}} \sigma_{\text{col}} \log d} \), \( B \log d \ll \left\| A^\ast \right\| \) and \( B \log d \sqrt{\mu r/d_2} \ll \sigma_{\text{col}} \sqrt{\log d} \sqrt{\mu r/d_1} \).

### C.4 Proof of Lemma 8

For notational convenience, we denote

\[ A_{s,(m),0} := \mathcal{P}_{m,:} (A^\ast). \]

Fix an arbitrary \( l \in [d_2] \), and we would like to upper bound \( \left\| A_{s,(m),0}^\top U^{(m)} H^{(m)} \right\|_2 \). The main difficulty here lies in the complicated statistical dependence between \( A_{s,(m),0}^\top U^{(m)} H^{(m)} \) and \( U^{(m)} H^{(m)} \). We recall the definitions of the auxiliary matrices \( U^{(m)} \) and \( H^{(m)} \) in Algorithm 6 and (147b), respectively. By construction, \( A_{s,(m),0} \) is independent of \( U^{(m)} \) and \( H^{(m)} \). Moreover, Lemma 9 guarantees that \( U^{(m)} H^{(m)} \) is extremely close to \( U^{(m)} H^{(m)} \). Thus, invoke the triangle inequality to upper bound

\[
\left\| A_{s,(m),0}^\top U^{(m)} H^{(m)} \right\|_2 \leq \left\| A_{s,(m),0}^\top U^{(m)} H^{(m)} - E \left[ A_{s,(m),0}^\top U^{(m)} H^{(m)} \right] \right\|_2 + \left\| E \left[ A_{s,(m),0}^\top U^{(m)} H^{(m)} \right] \right\|_2
\leq \alpha_1 + \alpha_2
\]

\[
+ \left\| A_{s,(m),0} \right\|_2 \left\| U^{(m)} H^{(m)} - U^{(m)} H^{(m)} \right\|_2
\leq \alpha_3.
\]
Before moving on, we make note of the following two useful upper bounds on $\|U^{(m)}U^{(m)\top} - U^{(m,l)}U^{(m,l)\top}\|$ based on Lemma 9. On the one hand, one has with probability at least $1 - O(d^{-13})$,
\[
\|U^{(m)}U^{(m)\top} - U^{(m,l)}U^{(m,l)\top}\| \leq \frac{1}{\sigma^2} \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right)^2 \|U^{(m)}H^{(m)}\|_{2,\infty} + \sigma_{\text{col}}^2 \|A^*\|_{2,\infty} \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right) \|A^*\|_{2,\infty} + \sigma_{\text{col}} \sqrt{\log d} \|A^*\| \\
\leq \frac{\delta_{\text{op}}}{\sigma^2} \ll 1. \tag{152}
\]
where (i) follows from the facts that $\sigma_{\text{col}} \leq \|U^{(m)}H^{(m)}\|_{2,\infty} \leq \|U^{(m)}H^{(m)}\| \leq 1$ and $\|A^*\|_{2,\infty} \leq \|A^*\|$; (152) is due to the definition of $\delta_{\text{op}}$ in (94d), the definition of $\delta_{\text{op}}$ (cf. (46)) as well as conditions (95a) and (95d). On the other hand, we can also bound
\[
\|U^{(m)}U^{(m)\top} - U^{(m,l)}U^{(m,l)\top}\| \leq \frac{1}{\sigma^2} \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right)^2 \|U^{(m)}H^{(m)}\|_{2,\infty} + \sigma_{\text{col}}^2 \|A^*\|_{2,\infty} \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right) \|A^*\|_{2,\infty} \leq \frac{1}{\sigma^2} \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right) \|A^*\|_{2,\infty} \leq \frac{\delta_{\text{op}}}{\sigma^2} \ll 1. \tag{153}
\]
where (i) arises from (95b) and the inequality that $\|A^*\|_{2,\infty} \leq \|A^*\| \|V^*\|_{2,\infty} \leq \|A^*\| \sqrt{\mu L / d_2}$, and (153) is due to conditions (95a) and (95d). In the sequel, we control the $\alpha_i$'s separately.

- For $\alpha_1$, it is easy to see that
\[
\left( A^{(m),0}_{:,l} - E\left[ A^{(m),0}_{:,l} \right] \right) U^{(m,l)}H^{(m,l)} = \sum_{i:j \neq m} E_{i,j}(U^{(m,l)}H^{(m,l)})_{i,:}
\]
is a sum of independent zero-mean random vectors conditional on $\{E_{i,j}\}_{i,j \in [d], j \in [d] \setminus \{l\}}$. Straightforward calculation gives that
\[
L := \max_{i:j \neq m} \|E_{i,j}(U^{(m,l)}H^{(m,l)})_{i,:}\|_2 \leq B \|U^{(m,l)}H^{(m,l)}\|_{2,\infty},
\]
\[
V := \sum_{i:j \neq m} E_{i,j}^2 \|U^{(m,l)}H^{(m,l)}_{i,:}\|_2^2 \leq \sigma_{\text{col}}^2 \|U^{(m,l)}H^{(m,l)}\|_{2,\infty}^2.
\]
Then we apply the matrix Bernstein inequality to obtain that with probability at least $1 - O(d^{-13})$,
\[
\left\| \left( A^{(m),0}_{:,l} - E\left[ A^{(m),0}_{:,l} \right] \right) U^{(m,l)}H^{(m,l)} \right\|_2 \leq \log d + \sqrt{V \log d} \\
\leq \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right) \|U^{(m,l)}H^{(m,l)}\|_{2,\infty} + \|U^{(m,l)}U^{(m,l)\top} - U^{(m,l)}U^{(m,l)\top}\|, \tag{154}
\]
where the last line results from the following observation:
\[
\|U^{(m,l)}H^{(m,l)}\|_{2,\infty} \leq \|U^{(m,l)}H^{(m,l)}\|_{2,\infty} + \|U^{(m,l)U^{(m,l)\top} - U^{(m,l)}U^{(m,l)\top}\|_{2,\infty}
\leq \|U^{(m,l)}H^{(m,l)}\|_{2,\infty} + \|U^{(m,l)U^{(m,l)\top} - U^{(m,l)}U^{(m,l)\top}\|.
\tag{155}
\]
• Turning to $\alpha_2$, we obtain the simple upper bound
\[
\| \mathbb{E} \left[ A_{i,l}^{(m,0)} \right]^T U^{(m,1)} H^{(m,1)} \|_2 \leq \| A_{i,l}^* \|_2 \| U^{(m,1)} H^{(m,1)} \| \leq \| A^* \|_{2,\infty},
\] (156)

• With regards to $\alpha_3$, Lemma 12 reveals that with probability at least $1 - O \left( d^{-13} \right)$,
\[
\| A_{i,l}^{(m,0)} \|_2 \| U^{(m)} H^{(m)} - U^{(m,1)} H^{(m,1)} \| \leq \| A_{i,l} \|_2 \| U^{(m)} U^{(m,1)^T} - U^{(m,1)} U^{(m,1)^T} \|
\lesssim \left( \| A^* \|_{2,\infty} + B \sqrt{\log d + \sigma_{\text{col}}} \right) \| U^{(m)} U^{(m,1)^T} - U^{(m,1)} U^{(m,1)^T} \|
\lesssim \| A^* \|_{2,\infty} + \left( B \sqrt{\log d + \sigma_{\text{col}}} \right) \| U^{(m)} U^{(m,1)^T} - U^{(m,1)} U^{(m,1)^T} \|,
\] (157)

where we use (152) in the last step.

Combining (154), (156), (157) implies that with probability greater than $1 - O \left( d^{-13} \right)$,
\[
\| A_{i,l}^{(m,0)^T} U^{(m)} H^{(m)} \|_2 \lesssim \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right) \left( \| U^{(m)} H^{(m)} \|_{2,\infty} + \| U^{(m)} U^{(m,1)^T} - U^{(m,1)} U^{(m,1)^T} \| \right) + \| A^* \|_{2,\infty}
\lesssim \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right) \| U^{(m)} H^{(m)} \|_{2,\infty} + \frac{1}{\sigma_r^2} \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right)^2 \| A^* \| \sqrt{\frac{\mu r}{d_2}} + \| A^* \| \sqrt{\frac{\mu r}{d_2}}
\lesssim \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right) \| U^{(m)} H^{(m)} \|_{2,\infty} + \| A^* \| \sqrt{\frac{\mu r}{d_2}},
\]

where (i) is by (153) and $\| A^* \|_{2,\infty} \leq \| A^* \| \sqrt{\mu r/d_2}$, and (ii) follows from conditions (95a) and (95d). The proof is complete by taking the union bound over $1 \leq l \leq d_2$.

C.5 Proof of Lemma 9

Fix arbitrary $m \in [d_1]$ and $l \in [d_2]$. Recalling the definitions of $G^{(m)}$ and $G^{(m,1)}$ in (43) and (44b), we see that $G^{(m)} - G^{(m,1)}$ is symmetric with entries
\[
(G^{(m)} - G^{(m,1)})_{i,j} = E_{i,l} E_{j,l} + A_{i,l}^* E_{j,l} + E_{i,l} A_{j,l}^*, \quad i \neq m, j \neq m, i \neq j,
\]
\[
(G^{(m)} - G^{(m,1)})_{i,m} = A_{i,m}^* E_{i,l}, \quad i \neq m,
\]
\[
(G^{(m)} - G^{(m,1)})_{i,i} = 0, \quad 1 \leq i \leq d_1.
\]

Note that $G^{(m)} - G^{(m,1)}$ depends only on $\{E_{i,l}\}_{i \in [d_1] \setminus \{m\}}$ and is hence statistically independent of $U^{(m,1)}$ and $H^{(m,1)}$. In particular, we can express
\[
\mathcal{P}_m (G^{(m)} - G^{(m,1)}) = \mathcal{P}_{\text{off-diag}} \mathcal{P}_m (E_{i,l} E_{j,l}^T + E_{i,l} A_{j,l}^* + A_{i,l}^* E_{j,l}^T),
\]
where $\mathcal{P}_m$ is the projection onto the subspace of matrices supported on $\{(i, j) \in [d_1] \times [d_2] : i \neq m \text{ and } j \neq m\}$ and $\mathcal{P}_{\text{off-diag}}$ extracts the off-diagonal part. In addition,
\[
(G^{(m)} - G^{(m,1)})_{m,:} = A_{m,l}^* E_{l,l} - A_{m,l}^* E_{m,l} e_m^T,
\]
where $e_m$ stands for the $m$-th standard basis in $\mathbb{R}^{d_1}$.

In the sequel, we shall apply the Davis-Kahan $\sin \Theta$ theorem to prove the claim. Towards this end, we need to control $\| G^{(m)} - G^{(m,1)} \|$ and $\| (G^{(m)} - G^{(m,1)}) U^{(m,1)} \|$. 

49
C.5.1 Step 1: controlling $\|G^{(m)} - G^{(m,l)}\|

Recall that we have already dealt with $\mathcal{P}_m(G^{(m)} - G^{(m,l)})$ in the proof of Lemma 1 in Appendix B.2. Straightforward computation gives

$$
\|\mathcal{P}_m(G^{(m)} - G^{(m,l)})\| \leq \|E_{i,l}\|_2^2 + \|E_{:,l}\|_2 \|A^\top\|_{2,\infty} \leq \|E_{i,l}\|_2^2 + \|E_{:,l}\|_2 \|A^\star\|,
$$

$$
\|\mathcal{P}_m(G^{(m)} - G^{(m,l)})\| \leq \bigg(\|G^{(m)} - G^{(m,l)}\|_{m,:}\bigg)_{2,\infty} \leq \|E_{i,l}\|_2 \|A^\star\|_\infty,
$$

where $\mathcal{P}_m$ is the projection onto the subspace of matrices supported on $\{(i,j) \in [d_1] \times [d_2] : i = m \text{ or } j = m\}$. In view of (102) (shown in the proof of Lemma 1 in Appendix B.2), we know that

$$
\|G^{(m)} - G^{(m,l)}\| \leq \|\mathcal{P}_m(G^{(m)} - G^{(m,l)})\| + \|\mathcal{P}_m(G^{(m)} - G^{(m,l)})\|
\lesssim \|E_{i,l}\|_2^2 + \|E_{:,l}\|_2 \|A^\star\|,
\lesssim B^2 \log d + \sigma_{\text{col}}^2 + (B \sqrt{\log d} + \sigma_{\text{col}}) \|A^\star\|
\lesssim \sigma_r^2,
$$

where the last step results from the conditions (95a) and (95d). Since $\|G^{(m)} - G^\star\| \lesssim \delta_{\text{op}} \ll \sigma_r^2$ by Lemma 6 and the condition (95d), this also implies

$$
\|G^{(m,l)} - G^\star\| \lesssim \sigma_r^2 \quad \text{and} \quad \|H^{(m,l)}\|^{-1} \lesssim 1,
$$

(158)

according to [AFWZ17, Lemma 3]. Moreover, it follows from Weyl’s inequality that

$$
\lambda_r(G^{(m)}) - \lambda_{r+1}(G^{(m)}) - \|G^{(m)} - G^{(m,l)}\| \geq \lambda_r(G^\star) - \lambda_{r+1}(G^\star) - 2 \|G^{(m)} - G^\star\| - \|G^{(m)} - G^{(m,l)}\|
\gtrsim \sigma_r^2.
$$

(159)

C.5.2 Step 2: controlling $\|(G^{(m)} - G^{(m,l)})U^{(m,l)}\|

In view of (158), we can obtain

$$
\|(G^{(m)} - G^{(m,l)})U^{(m,l)}\| \leq \|(G^{(m)} - G^{(m,l)})U^{(m,l)}H^{(m,l)}\| (H^{(m,l)})^{-1}
\lesssim \|(G^{(m)} - G^{(m,l)})U^{(m,l)}H^{(m,l)}\|
\leq \|\mathcal{P}_m(G^{(m)} - G^{(m,l)})U^{(m,l)}H^{(m,l)}\|_{F,\alpha_1}
\lesssim \|\mathcal{P}_m(G^{(m)} - G^{(m,l)})U^{(m,l)}H^{(m,l)}\|_{F,\alpha_2}.
$$

(160)

Therefore, it suffices to control $\alpha_1$ and $\alpha_2$ separately.

- Regarding $\alpha_1$, Lemma 12 reveals that, with probability at least $1 - O(d^{-13})$,

$$
\|\mathcal{P}_m(G^{(m)} - G^{(m,l)})U^{(m,l)}H^{(m,l)}\|_F
\leq \|(G^{(m)} - G^{(m,l)})U^{(m,l)}H^{(m,l)}\|_2 + \|G^{(m)} - G^{(m,l)}\|_{m,:} \|U^{(m,l)}H^{(m,l)}\|_{2,\infty}
\leq \|A^\star\|_\infty \|E_{i,l}\|_2 \|U^{(m,l)}H^{(m,l)}\|_2 + \|A^\star\|_\infty \|E_{:,l}\|_2 \|U^{(m,l)}H^{(m,l)}\|_{2,\infty}
\lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^\star\|_\infty \|U^{(m,l)}H^{(m,l)}\|_{2,\infty}.
$$

(161)

- When it comes to $\alpha_2$, since the spectral norm of a submatrix is always less than that of its original matrix, we can further upper bound

$$
\|\mathcal{P}_m(G^{(m)} - G^{(m,l)})U^{(m,l)}H^{(m,l)}\| \leq \|(E_{i,l}E_{:,l}^\top - D_{i,l}) U^{(m,l)}H^{(m,l)}\|
=: \beta_1
$$
where $D_l$ and $\tilde{D}_l$ are defined in (100) and (107) in Appendix B.2. In what follows, let us bound $\beta_1$ and $\beta_2$.

- To bound $\beta_1$, we have

$$\| (E_t, t E_t^\top - D_l) U^{(m,l)} H^{(m,l)} \| \leq \| E_t \|_2 \| (E_t, t)^\top U^{(m,l)} H^{(m,l)} \|_2 + \| D_l U^{(m,l)} H^{(m,l)} \|.$$ 

By Lemma 12, one has with probability at least $1 - O(d^{-13}),$

$$ \| E_t \|_2 \| (E_t, t)^\top U^{(m,l)} H^{(m,l)} \|_2 \lesssim (B \sqrt{ \log d + \sigma_{\text{col}}} ) ( B \log d + \sigma_{\text{col}} \sqrt{ \log d } ) \| U^{(m,l)} H^{(m,l)} \|_{2, \infty} \leq ( B \log d + \sigma_{\text{col}} \sqrt{ \log d } )^2 \| U^{(m,l)} H^{(m,l)} \|_{2, \infty}. $$

As for the second term $\| D_l U^{(m,l)} H^{(m,l)} \|$, we first observe that

$$ \| \mathbb{E}[D_l] U^{(m,l)} H^{(m,l)} \| \leq \| \mathbb{E}[D_l] \| \| U^{(m,l)} H^{(m,l)} \| \leq \max_{i \in [d]} \mathbb{E} \| E_{i,t} \| \| U^{(m,l)} H^{(m,l)} \| \leq \sigma_{\text{col}}^2. $$

Additionally, the deviation $(D_l - \mathbb{E}[D_l]) U^{(m,l)} H^{(m,l)} = \sum_{i \in [d]} (E_{i,t}^2 - \mathbb{E}[E_{i,t}^2]) e_i e_i^\top U^{(m,l)} H^{(m,l)}$ is a sum of independent zero-mean random matrices. By the matrix Bernstein inequality, we have with probability at least $1 - O(d^{-13}),$

$$ \| (D_l - \mathbb{E}[D_l]) U^{(m,l)} H^{(m,l)} \| \lesssim \left( \max_{i \in [d]} | E_{i,t}^2 - \mathbb{E}[E_{i,t}^2] | \log d + \sqrt{ \sum_{i \in [d]} \mathbb{E}[E_{i,t}^2] \log d } \right) \| e_i e_i^\top U^{(m,l)} H^{(m,l)} \|_{2, \infty} \lesssim \left( B^2 \log d + \sigma_{\text{col}}^2 \log d \right) \| U^{(m,l)} H^{(m,l)} \|_{2, \infty} \lesssim \left( B^2 \log d + \sigma_{\text{col}}^2 \sqrt{ \log d } \right) \| U^{(m,l)} H^{(m,l)} \|_{2, \infty},$$

where we have used the AM-GM inequality in (i). Combining the estimates above yields

$$ \beta_1 = \left\| (E_t, t E_t^\top - D_l) U^{(m,l)} H^{(m,l)} \right\| \lesssim ( B \log d + \sigma_{\text{col}} \sqrt{ \log d } )^2 \| U^{(m,l)} H^{(m,l)} \|_{2, \infty} + \sigma_{\text{col}}^2. $$

- Turning to $\beta_2$, we see from Lemma 12 that with probability at least $1 - O(d^{-13})$, one has

$$ \| A_{i,t}^* E_{i,t}^\top U^{(m,l)} H^{(m,l)} \| \leq \| A_{i,t}^* \|_{2, \infty} \| E_{i,t}^\top U^{(m,l)} H^{(m,l)} \| \lesssim \left( B \log d + \sigma_{\text{col}} \sqrt{ \log d } \right) \| A_{i,t}^* \|_{2, \infty} \| U^{(m,l)} H^{(m,l)} \|_{2, \infty} $$

and

$$ \| E_{i,t} A_{i,t}^* U^{(m,l)} H^{(m,l)} \| \leq \| E_{i,t} \|_2 \| A_{i,t}^* \|_{2, \infty} \| U^{(m,l)} H^{(m,l)} \| \lesssim \left( B \log d + \sigma_{\text{col}} \right) \| A_{i,t}^* \|_{2, \infty}. $$

In addition, $\tilde{D}_l U^{(m,l)} H^{(m,l)} = \sum_{i \in [d]} A_{i,t}^* E_{i,t} e_i e_i^\top U^{(m,l)} H^{(m,l)}$ is a sum of independent zero-mean random matrices. It then follows from the matrix Bernstein inequality that with probability at least $1 - O(d^{-13}),$

$$ \| \tilde{D}_l U^{(m,l)} H^{(m,l)} \| \lesssim \left( \max_{i \in [d]} | A_{i,t}^* E_{i,t} | \log d + \sqrt{ \sum_{i \in [d]} | A_{i,t}^* E_{i,t}^2 | \log d } \right) \| U^{(m,l)} H^{(m,l)} \|_{2, \infty} $$
\[ \lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\|_\infty \|U^{(m,l)}H^{(m,l)}\|_{2,\infty}. \]

Hence, we know that
\[
\beta_2 = \left\| (A^*_1 E_{j,l}^T + E_{j,l} A_1^* - 2 \hat{D}_j) U^{(m,l)} H^{(m,l)} \right\| \\
\lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\|_\infty \|U^{(m,l)}H^{(m,l)}\|_{2,\infty} + (B \sqrt{\log d} + \sigma_{\text{col}}) \|A^*\|_{2,\infty} \\
\lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\|_{2,\infty}, \quad (163)
\]

which results from the facts that \( \|A^*\|_\infty \leq \|A^*\|_{2,\infty} \) and \( \|U^{(m,l)}H^{(m,l)}\|_{2,\infty} \leq \|U^{(m,l)}H^{(m,l)}\|_1 \leq 1. \)

Putting (162) and (163) together yields that
\[
\|P_{-\lambda}(G^{(m)} - G^{(m,l)})U^{(m,l)}H^{(m,l)}\| \lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d})^2 \|U^{(m,l)}H^{(m,l)}\|_{2,\infty} + \sigma_{\infty}^2 + (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\|_{2,\infty}.
\]

This combined with (161) and (160) implies
\[
\| (G^{(m)} - G^{(m,l)})U^{(m,l)} \| \lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d})^2 \|U^{(m,l)}H^{(m,l)}\|_{2,\infty} + \sigma_{\infty}^2 + \sigma_{\infty}^2 \\
+ (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\|_{2,\infty} + (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\|_{2,\infty} \\
\overset{(i)}{\lesssim} (B \log d + \sigma_{\text{col}} \sqrt{\log d})^2 \|U^{(m,l)}H^{(m,l)}\|_{2,\infty} + \sigma_{\infty}^2 + (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\|_{2,\infty} \\
\overset{(ii)}{\lesssim} (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|U^{(m,l)}H^{(m,l)}\|_{2,\infty} + \sigma_{\infty}^2 + (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\|_{2,\infty}.
\]

where (i) is due to the facts that \( \|U^{(m,l)}H^{(m,l)}\|_{2,\infty} \leq \|U^{(m,l)}H^{(m,l)}\|_1 \leq 1, \) and (165) arises from the inequality \( \|A^*\|_\infty \leq \|A^*\|_{2,\infty}. \)

C.5.3 Step 3: combining Step 1 and Step 2

From (159) and (165), we apply the Davis-Kahan sin\( \Theta \) theorem to obtain that with probability exceeding \( 1 - O(d^{-12}), \)
\[
\|U^{(m)}U^{(m)^T} - U^{(m,l)}U^{(m,l)^T}\| \leq \frac{1}{\lambda_r (G^{(m)}) - \lambda_{r+1} (G^{(m)})} \| (G^{(m)} - G^{(m,l)})U^{(m,l)}\| \\
\lesssim \frac{1}{\sigma_r^{-2}} \| (G^{(m)} - G^{(m,l)})U^{(m,l)}\| \\
\overset{(i)}{\lesssim} \frac{1}{\sigma_r^{-2}} (B \log d + \sigma_{\text{col}} \sqrt{\log d})^2 \left( \|U^{(m)}H^{(m)}\|_{2,\infty} + \|U^{(m)}U^{(m)^T} - U^{(m,l)}U^{(m,l)^T}\| \right) \\
+ \frac{\sigma_{\infty}^2}{\sigma_r^{-2}} + \frac{1}{\sigma_r^{-2}} (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\|_{2,\infty} \\
\overset{(ii)}{\lesssim} \frac{1}{\sigma_r^{-2}} (B \log d + \sigma_{\text{col}} \sqrt{\log d})^2 \|U^{(m)}H^{(m)}\|_{2,\infty} + \sigma_1 (1) \|U^{(m)}U^{(m)^T} - U^{(m,l)}U^{(m,l)^T}\| \\
+ \frac{\sigma_{\infty}^2}{\sigma_r^{-2}} + \frac{1}{\sigma_r^{-2}} (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\|_{2,\infty}.
\]

Here, we have used (155) in (i), and the condition (95d) (i.e. \( \max \{B \log d, \sigma_{\text{col}} \sqrt{\log d}\} \ll \sigma_r^*) \) in (ii). Rearrange the inequalities and taking the union bound over \( m \in [d_1] \) and \( l \in [d_2] \) complete the proof.
C.6 Proof of Lemma 10

Recall the definition of $G$ in (13). We can express

$$G_{m,:}(U^{(m)}H^{(m)} - U^*) = A^*_{m,:}[P_{-m,:}(A^*)]^T(U^{(m)}H^{(m)} - U^*).$$

Consequently, one can upper bound

$$\|G_{m,:}(U^{(m)}H^{(m)} - U^*)\|_2 \leq \left\|A^*_{m,:}[P_{-m,:}(A^*)]^T(U^{(m)}H^{(m)} - U^*)\right\|_2 =: \beta_1$$

$$+ \left\|(A^* - A^*)[P_{-m,:}(A^*)]^T(U^{(m)}H^{(m)} - U^*)\right\|_2 =: \beta_2.$$

In what follows, we shall control $\beta_1$ and $\beta_2$ separately.

- To upper bound $\beta_1$, we have

$$\left\|A^*_{m,:}[P_{-m,:}(A^*)]^T(U^{(m)}H^{(m)} - U^*)\right\|_2 \leq \left\|A^*_{m,:}[P_{-m,:}(A^*)]^T\right\|_2 \|U^{(m)}H^{(m)} - U^*\|.$$

It is straightforward to derive

$$\left\|A^*_{m,:}[P_{-m,:}(A^*)]^T\right\|_2 \leq \left\|A^*_{m,:}A^T\right\|_2 \leq \|A^*_{m,:}\|_2 \|A^*\| \leq \sigma_1^* \|A^*\|_{2,\infty}.$$

In addition, Lemma 14 indicates that

$$\|A^*_{m,:}E^T\|_2^2 = \sum_i \left(\sum_j A^*_{m,i}E_{i,j}\right)^2 \lesssim (\sigma_{col}^2 + \sigma_{dol}^2 \log^2 d) \|A^*\|_{2,\infty}^2 + B^2 \|A^*\|_{\infty}^2 \log^3 d$$

$$\leq (\sigma_{col}^2 + B^2 \log^2 d) \|A^*\|_{2,\infty}^2 + B^2 \|A^*\|_{\infty}^2 \log^3 d$$

holds with probability at least $1 - O(d^{-11})$. Hence, we have

$$\left\|A^*_{m,:}[P_{-m,:}(A^*)]^T\right\|_2 \leq \|A^*_{m,:}A^T\|_2 + \|A^*_{m,:}E^T\|_2$$

$$\lesssim \sigma_1^* \|A^*\|_{2,\infty} + (\sigma_{col} + B \log d) \|A^*\|_{2,\infty} + B \log^3 d \|A^*\|_{\infty}$$

$$\lesssim \sigma_1^* \|A^*\|_{2,\infty} + \sigma_1^2 \sqrt{\frac{\mu}{d_1}} \lesssim \delta_1 \sqrt{\frac{\mu}{d_1}},$$

using conditions (95a), (95c) and (95d). Moreover, from Lemma 1 and Lemma 6, we know that

$$\left\|U^{(m)}H^{(m)} - U^*\right\| \lesssim \left\|U^{(m)}U^{(m)^T} - U^*U^*\right\| \leq \frac{\|G^{(m)} - G^*\|}{\lambda_r(G^*) - \lambda_{r+1}(G^{(m)})}$$

$$\leq \frac{\|G^{(m)} - G^*\|}{\lambda_r(G^*) - \lambda_{r+1}(G^*) - \|G^{(m)} - G^*\|}$$

$$\lesssim \frac{1}{\sigma_r^2} \|G^{(m)} - G^*\| \lesssim \frac{\delta_{op}^2}{\sigma_r^2},$$

where $\delta_{op}$ is defined in (46). Combining (166) and (167) yields

$$\left\|A^*_{m,:}[P_{-m,:}(A^*)]^T(U^{(m)}H^{(m)} - U^*)\right\|_2 \lesssim \|A^*_{m,:}[P_{-m,:}(A^*)]^T\|_2 \|U^{(m)}H^{(m)} - U^*\| \lesssim \delta_{op} \sqrt{\frac{\mu}{d_1}},$$

(168)
Next, we look at $\beta_2$. Before we start, we pause to note that by (153), one has
\[
\|U^{(m)}U^{(m)\top} - U^{(m,l)}U^{(m,l)\top}\| \lesssim \frac{1}{\sigma_r^2} (B \log d + \sigma_{\text{col}} \sqrt{\log d})^2 (\|U^{(m)}H^{(m)} - U^*\|_{2,\infty} + \|U^*\|_{2,\infty})
+ \frac{1}{\sigma_r^2} (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\| \sqrt{\frac{\mu r}{d_2}}.
\] (169)

We now ready to control $(A^* - A^*)_m, [P_{-m,:}(A^*)]^\top (U^{(m)}H^{(m)} - U^*)$, which can be accomplished in the same way as in the proof of Lemma 7 in Appendix C.3. We omit the proof details for conciseness here and only give the proof sketch. First, we can use $U^{(m,l)}H^{(m,l)} - U^*$ as the surrogate for $U^{(m)}H^{(m)} - U^*$ to deal with the statistical dependence issue, and apply the Bernstein inequality to show that with probability at least $1 - O(d^{-11})$,
\[
\| [P_{-m,:}(A^*)]^\top (U^{(m)}H^{(m)} - U^*) \|_{2,\infty}
\lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|U^{(m)}H^{(m)} - U^*\|_{2,\infty}
+ \|A^*\|_{2,\infty} \|U^{(m)}H^{(m)} - U^*\|_{2,\infty}
+ \left( \|A^*\|_{2,\infty} + B \sqrt{\log d} + \sigma_{\text{col}} \right) \|U^{(m)}U^{(m)\top} - U^{(m,l)}U^{(m,l)\top}\|
\lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|U^{(m)}H^{(m)} - U^*\|_{2,\infty}
+ \frac{1}{\sigma_r^2} (B \log d + \sigma_{\text{col}} \sqrt{\log d})^2 (\|U^{(m)}H^{(m)} - U^*\|_{2,\infty} + \|U^*\|_{2,\infty})
+ \frac{1}{\sigma_r^2} (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|A^*\| \sqrt{\frac{\mu r}{d_2}}
\lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|U^{(m)}H^{(m)} - U^*\|_{2,\infty}
+ \frac{\delta_{\text{op}}}{\sigma_r^2} \left( \|U^*\|_{2,\infty} + \kappa \sqrt{\frac{\mu r}{d_2}} \right),
\] (170)

where (i) follows from (152), (167) and (169) and the inequality $\|A^*\|_{2,\infty} \leq \|A^*\| \sqrt{\mu r/d_2}$; (170) arises from the definition of $\delta_{\text{op}}$ in (96) and conditions (95a) and (95d) (namely, $B \log d + \sigma_{\text{col}} \sqrt{\log d} \ll \sigma_r^2 / \kappa$ and $(B \log d + \sigma_{\text{col}} \sqrt{\log d})^2 \ll \delta_{\text{op}} \ll \sigma_r^2$). Applying the matrix Bernstein inequality yields that with probability at least $1 - O(d^{-11})$,
\[
\| (A^* - A^*)_m, [P_{-m,:}(A^*)]^\top (U^{(m)}H^{(m)} - U^*) \|_2
\lesssim B \log d \| [P_{-m,:}(A^*)]^\top (U^{(m)}H^{(m)} - U^*) \|_{2,\infty}
+ \sqrt{\frac{1}{d_1}} \sigma_{\text{row}} \log d + \|A^*\| \sqrt{\log d} \|U^{(m)}H^{(m)} - U^*\|_{2,\infty}
\lesssim B \log d \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right) \|U^{(m)}H^{(m)} - U^*\|_{2,\infty}
+ \delta_{\text{op}} \frac{B \log d}{\sigma_r^2} \left( \|U^*\|_{2,\infty} + \kappa \sqrt{\frac{\mu r}{d_2}} \right)
+ \sigma_{\text{col}} \left( \sigma_{\text{row}} \log d + \|A^*\| \sqrt{\log d} \right) \|U^{(m)}H^{(m)} - U^*\|_{2,\infty}
\lesssim \sigma_{\text{col}} \left( \sigma_{\text{row}} \log d + \|A^*\| \sqrt{\log d} \right) \|U^{(m)}H^{(m)} - U^*\|_{2,\infty}
+ \delta_{\text{op}} \frac{B \log d}{\sigma_r^2} \left( \|U^*\|_{2,\infty} + \kappa \sqrt{\frac{\mu r}{d_2}} \right)
\lesssim \sigma_{\text{col}} \left( \sigma_{\text{row}} \log d + \|A^*\| \sqrt{\log d} \right) \|U^{(m)}H^{(m)} - U^*\|_{2,\infty}
+ \delta_{\text{op}} \frac{\kappa \sigma_{\text{col}} \sqrt{\log d}}{\sigma_r^2} \left( \|U^*\|_{2,\infty} + \delta_{\text{op}} \frac{\mu r}{d_1} \right)
\lesssim \sigma_{\text{col}} \left( \sigma_{\text{row}} \log d + \|A^*\| \sqrt{\log d} \right) \|U^{(m)}H^{(m)} - U^*\|_{2,\infty}
+ \delta_{\text{op}} \frac{\mu r}{d_1},
\]
Here, (i) follows from (94) and (170); (ii) is due to conditions (95a) and (95d) that \( B^2 \log d \lesssim \sigma \text{col} \sigma \text{row} \), \( B \log d \ll \sigma \text{col}^* \) and \( B \log d \left( B \log d + \sigma \text{col} \sqrt{\log d} \right) \lesssim B \log^2 d + \sigma \text{col}^2 \log d \leq \delta \text{op} \); (iii) holds true because of (95b) and (95d) that \( B \log d \ll \sigma \text{col}^* \); and (iv) arises from (95d) that \( \sigma \text{col} \sqrt{\log d} \ll \sigma \text{col}^*/\kappa \). Recalling the definition of \( \delta \text{lo} \) in (124), we obtain that

\[
\left\| (A^* - A^*)_{m,:} \left[ P_{-m,:} (A^*) \right]^\top (U^{(m)} H^{(m)} - U^*) \right\|_2 \lesssim \delta \text{lo} \left\| U^{(m)} H^{(m)} - U^* \right\|_2, \infty + o(1) \delta \text{op} \sqrt{\frac{\mu r}{d}}. \tag{171}
\]

Putting (168) and (171) together, we arrive at the advertised bound.

### D A few more auxiliary lemmas

In this section, we establish a few auxiliary facts that are useful throughout the proof of the main theorem. We begin with some basic properties about the truth \( A^* \) and \( G^* \).

**Lemma 11.** Recall the definition of the incoherence parameters in Definition 1. Then one has

\[
\| A^* \|_{2, \infty} \leq \sqrt{\frac{\mu_1 \sigma_1^2}{d_1}}, \quad \| A^\top \|_{2, \infty} \leq \sqrt{\frac{\mu_2 \sigma_1^2}{d_2}},
\]

\[
\| G^* \|_{2, \infty} \leq \sqrt{\frac{\mu_1 \sigma_2^4}{d_1}}, \quad \| A^* \|_\infty \leq \min \left\{ \sqrt{\frac{\mu_1 \mu_2 r^2}{d_1 d_2}}, \sigma_1^* \| U^* \|_{2, \infty}, \sigma_1^* \| V^* \|_{2, \infty} \right\}.
\]

Next, we summarize several facts related to the matrix \( E \) defined in (91), which contains independent zero-mean entries.

**Lemma 12.** Fix any matrices \( W_1 \) and \( W_2 \). With probability greater than \( 1 - O \left( d^{-20} \right) \), the following holds

\[
\max_{i \in [d_1]} \sum_{j \in [d_2]} E_{i,j}^2 \lesssim B^2 \log d + \sigma_{\text{row}}^2,
\]

\[
\max_{i \in [d_1]} \left\| E_{i,:} W_1 \right\|_2 \lesssim \left( B \log d + \sigma_{\text{row}} \sqrt{\log d} \right) \left\| W_1 \right\|_{2, \infty},
\]

\[
\max_{j \in [d_2]} \sum_{i \in [d_1]} E_{i,j}^2 \lesssim B^2 \log d + \sigma_{\text{col}}^2,
\]

\[
\max_{j \in [d_2]} \left\| E_{:,j} \right\|_{2, \infty} \lesssim \left( B \log d + \sigma_{\text{col}} \sqrt{\log d} \right) \left\| W_2 \right\|_{2, \infty},
\]

where \( \sigma_{\text{row}}, \sigma_{\text{col}}, \) and \( B \) are respectively upper bounds on \( \max_{i \in [d_1]} \sqrt{\sum_{j \in [d_2]} E_{i,j}^2} ; \max_{j \in [d_2]} \sqrt{\sum_{i \in [d_1]} E_{i,j}^2} \) and \( \max_{i \in [d_1], j \in [d_2]} |E_{i,j}| \); see (94) for precise definitions. As a result, one has

\[
\| E \|_{2, \infty} \lesssim B \sqrt{\log d + \sigma_{\text{row}}},
\]

\[
\| A^* \|_{2, \infty} \lesssim \| A^\top \|_{2, \infty} + B \sqrt{\log d + \sigma_{\text{row}}},
\]

\[
\| E^\top \|_{2, \infty} \lesssim B \sqrt{\log d + \sigma_{\text{col}}},
\]

\[
\| A^\top \|_{2, \infty} \lesssim \| A^\top \|_{2, \infty} + B \sqrt{\log d + \sigma_{\text{col}}},
\]

where \( A^* = A^* + E \) is defined in (90).

**Lemma 13.** With probability greater than \( 1 - O \left( d^{-20} \right) \), one has

\[
\| E \| \lesssim B \log d + (\sigma_{\text{row}} + \sigma_{\text{col}}) \sqrt{\log d},
\]

where \( B, \sigma_{\text{row}} \) and \( \sigma_{\text{col}} \) are defined in (94).

55
Lemma 14. Fix any vector \( w \in \mathbb{R}^{d_2} \). With probability at least \( 1 - O\left(d^{-20}\right) \), one has
\[
\sum_{i \in [d_1]} \left( \sum_{j \in [d_2]} w_j E_{i,j} \right)^2 \lesssim \|w\|_2^2 \left( \sigma_{\text{col}}^2 + \sigma_\infty^2 \log^2 d \right) + \|w\|_\infty^2 B^2 \log^3 d,
\]
where \( B, \sigma_\infty \) and \( \sigma_{\text{col}} \) are defined in (94).

D.1 Proof of Lemma 11
Given the SVD of \( A^* = U^* \Sigma^* V^* \top \), one has \( G^* = A^* A^\top = U^* \Sigma^2 U^\top \). Using the definition of the incoherence parameters, one can derive
\[
\|A^*\|_{2, \infty} = \max_{i \in [d_1]} \|U^*_i \Sigma^* V^* \top\|_2 \leq \max_{i \in [d_1]} \|U^*_i \|_2 \|\Sigma^*\| \|V^*\| \leq \sigma_1^* \|U^*\|_{2, \infty} \leq \frac{\mu_1 \sigma_1^2}{d_1}.
\]
\[
\|A^\top\|_{2, \infty} = \max_{j \in [d_2]} \|V^*_j \Sigma^* U^* \top\|_2 \leq \max_{j \in [d_2]} \|V^*_j \|_2 \|\Sigma^*\| \|U^*\| \leq \sigma_1^* \|V^*\|_{2, \infty} \leq \frac{\mu_2 \sigma_1^2}{d_2}.
\]
\[
\|G^*\|_{2, \infty} = \max_{i \in [d_1]} \|U^*_i \Sigma^2 U^\top\|_2 \leq \max_{i \in [d_1]} \|U^*_i \|_2 \|\Sigma^2\| \|U^*\| \leq \sigma_1^2 \|U^*\|_{2, \infty} \leq \frac{\mu_1 \sigma_1^4}{d_1}.
\]
Moreover, the Cauchy-Schwartz inequality allows one to upper bound
\[
\|A^*\|_\infty = \max_{(i,j) \in [d_1] \times [d_2]} \|U^*_i \Sigma^* (V^*_j) \top\| \leq \|U^*\|_{2, \infty} \|\Sigma^*\| \|V^*\|_{2, \infty} \leq \sigma_1^* \|U^*\|_{2, \infty} \|V^*\|_{2, \infty}.
\]
In view of the simple bounds \( \|U^*\|_{2, \infty} \leq \|U^*\| \leq 1 \) and \( \|V^*\|_{2, \infty} \leq \|V^*\| \leq 1 \), we conclude that
\[
\|A^*\|_\infty \leq \sigma_1^* \|U^*\|_{2, \infty} \quad \text{and} \quad \|A^*\|_{2, \infty} \leq \sigma_1^* \|V^*\|_{2, \infty}.
\]

D.2 Proof of Lemma 12
We shall only prove the results concerning \( \sigma_{\text{col}} \); the results concerning \( \sigma_{\text{row}} \) follow immediately via nearly identical arguments.

In view of the Bernstein inequality, we have
\[
P \left\{ \sum_{i \in [d_1]} E_{i,j}^2 - M_1 \geq t \right\} \leq 2 \exp \left( -\frac{3}{8} \min \left\{ \frac{t^2}{\Var \left\{ E_{i,j}^2 \right\}} \right\} \right), \quad t > 0,
\]
where \( M_1, L_1 \) and \( S_1 \) are given respectively by
\[
M_1 := \sum_{i \in [d_1]} \mathbb{E} \left[ E_{i,t}^2 \right] \leq \sigma_{\text{col}}^2,
\]
\[
L_1 := \max_{i \in [d_1]} E_{i,t}^2 - \mathbb{E} \left[ E_{i,t}^2 \right] \leq B^2 + \sigma_\infty^2 \leq 2B^2,
\]
\[
V_1 := \sum_{i \in [d_1]} \Var \left( E_{i,t}^2 \right) \leq \sum_{i \in [d_1]} \mathbb{E} \left[ E_{i,t}^2 \right] \leq B^2 \sigma_{\text{col}}^2.
\]
Here, we have made use of the fact that \( \sigma_\infty \leq B \). As a result, one has
\[
\sum_{i \in [d_1]} E_{i,j}^2 \lesssim M_1 + L_1 \log d + \sqrt{V_1 \log d} \lesssim \sigma_{\text{col}}^2 + B^2 \log d + B \sigma_{\text{col}} \sqrt{\log d}
\]
\[
\lesssim \sigma_{\text{col}}^2 + B^2 \log d
\]
with probability exceeding \( 1 - O\left(d^{-20}\right) \), where the last line arises from the AM-GM inequality (namely, \( 2B \sigma_{\text{col}} \sqrt{\log d} \leq \sigma_{\text{col}}^2 + B^2 \log d \)). As an immediate consequence, with probability at least \( 1 - O\left(d^{-20}\right) \),
\[
\|E_{i,j}\|_2 = \sqrt{\sum_{i \in [d_1]} E_{i,j}^2} \lesssim \sigma_{\text{col}} + B \sqrt{\log d},
\]
where \( B, \sigma_\infty \) and \( \sigma_{\text{col}} \) are defined in (94).
\[ \|A_{i,j}\|_2 \leq \|A_{i,j}'\|_2 + \|E_{i,j}\|_2 \lesssim \|A^*\|_{2,\infty} + \sigma_{\text{col}} + B\sqrt{\log d}. \]

Next, we turn to the claim concerning a fixed matrix \( W_2 \). Observe that \((E_{i,j})^\top W_2 = \sum_{i \in [d_1]} E_{i,j}(W_2)_{i,:}\) is a sum of independent zero-mean random vectors. In order to invoke standard concentration inequalities, we compute

\[
L_2 := \max_{i \in [d_1]} \|E_{i,:}(W_2)_{i,:}\|_2 \leq B \|W_2\|_{2,\infty},
\]

\[
V_2 := \sum_{i \in [d_1]} \mathbb{E} \left[ E_{i,j}^2 \right] \| (W_2)_{i,:}\|_2^2 \leq \sigma_{\text{col}}^2 \|W_2\|_{2,\infty}^2.
\]

Invoking the matrix Bernstein inequality yields that with probability exceeding \( 1 - O\left(d^{-20}\right) \),

\[
\| (E_{i,:})^\top W_2 \|_2 \lesssim L_2 \log d + \sqrt{V_2 \log d} \lesssim (B \log d + \sigma_{\text{col}} \sqrt{\log d}) \|W_2\|_{2,\infty}.
\]

**D.3 Proof of Lemma 13**

First, we can write

\[ E = \sum_{i \in [d_1], j \in [d_2]} E_{i,j} e_i e_j^\top \]

as a sum of independent zero-mean random matrices (since \( \mathbb{E}[E_{i,j}] = 0 \)). We make the observation that

\[
L := \max_{i \in [d_1], j \in [d_2]} \| E_{i,j} e_i e_j^\top \| \leq B;
\]

\[
V := \max \left\{ \left\| \sum_{i \in [d_1], j \in [d_2]} \mathbb{E} \left[ E_{i,j}^2 \right] e_i e_j^\top \right\|, \left\| \sum_{i \in [d_1], j \in [d_2]} \mathbb{E} \left[ E_{i,j}^2 \right] e_j e_i^\top \right\| \right\} \leq \sigma_{\text{row}}^2 + \sigma_{\text{col}}^2.
\]

It then follows from the matrix Bernstein inequality that, with probability at least \( 1 - O\left(d^{-10}\right) \),

\[
\| E \| \lesssim L \log d + \sqrt{V \log d} \lesssim B \log d + (\sigma_{\text{row}} + \sigma_{\text{col}}) \sqrt{\log d}.
\]

**D.4 Proof of Lemma 14**

Let us define a sequence of independent zero-mean random variables \( \{X_i\}_{1 \leq i \leq d_1} \) as follows

\[ X_i := \sum_{j \in [d_2]} w_j E_{i,j}. \]

It is easily seen that

\[
\max_{j \in [d_2]} |w_j E_{i,j}| \leq \|w\|_{\infty} B; \quad \mathbb{E} \left[ X_i^2 \right] = \sum_{j \in [d_2]} w_j^2 \sigma_{i,j}^2 \leq \|w\|_2^2 \sigma_{\infty}^2.
\]

We can therefore apply the Bernstein inequality to show that, with probability at least \( 1 - O\left(d^{-11}\right) \),

\[
|X_i| \lesssim (\|w\|_{\infty} B) \log d + \sqrt{\left( \|w\|_2^2 \sigma_{\infty}^2 \right) \log d} :=: R. \tag{172}
\]

Next, let us introduce a sequence of independent random variables \( \{Y_i\}_{1 \leq i \leq d_1} \), obtained by truncating \( X_i \)

\[ Y_i \triangleq X_i 1 \{|X_i| \leq CR\} \]

for some sufficiently large absolute constant \( C > 0 \). From (172) and the union bound, we know that \( Y_i = X_i \) holds simultaneously for all \( 1 \leq i \leq d_1 \) with probability at least \( 1 - O\left(d^{-10}\right) \).
Further, it is straightforward to compute that

\[ M_2 := \sum_{i \in [d_1]} \mathbb{E} [Y_i^2] \leq \sum_{i \in [d_1]} \mathbb{E} [X_i^2] \leq \sum_{i \in [d_1], j \in [d_2]} w_{ij}^2 \sigma_{ij}^2 \leq \|w\|_2^2 \sigma_{col}^2; \]

\[ L_2 := \max_{i \in [d_1]} |Y_i^2 - \mathbb{E} [Y_i^2]| \lesssim R^2 \lesssim \|w\|_2^2 B^2 \log^2 d + \|w\|_2^2 \sigma_{col}^2 \log d; \]

\[ V_2 := \sum_{i \in [d_1]} \text{Var}(Y_i^2) \leq \sum_{i \in [d_1]} \mathbb{E} [Y_i^2] \leq \sum_{i \in [d_1], j \in [d_2]} \sum_{i \in [d_1], j \neq j_2} w_{ij}^2 \mathbb{E} [E_{i,j_1}^2] + \sum_{i \in [d_1], j \neq j_2} w_{ij}^2 \mathbb{E} [E_{i,j_1}^2] \mathbb{E} [E_{i,j_2}^2] \]

\[ \lesssim \|w\|_\infty^2 \|w\|_2^2 B^2 \sigma_{col}^2 + \|w\|_2^4 \sigma_{\infty}^2 \sigma_{col}^2. \]

We then apply the Bernstein inequality to conclude that with probability at least 1 – \( O(d^{-10}) \):

\[ \sum_{i \in [d_1]} Y_i^2 \lesssim M_2 + L_2 \log d + \sqrt{V_2 \log d} \]

\[ \lesssim \|w\|_2^2 \sigma_{col}^2 + \|w\|_\infty^2 B^2 \log^2 d + \|w\|_2^2 \sigma_{\infty}^2 \log d + \left( \|w\|_\infty \mathbb{E} [w]_2 B \sigma_{col} + \|w\|_2^2 \sigma_{\infty} \sigma_{col} \right) \sqrt{\log d} \]

\[ \asy \|w\|_2^2 \left( \sigma_{col}^2 + \sigma_{\infty}^2 \log^2 d \right) + \|w\|_2^2 B^2 \log^3 d, \]

where the last line arises from the AM-GM inequality (namely, 2 \( \|w\|_\infty \mathbb{E} [w]_2 B \sigma_{col} \sqrt{\log d} \leq \|w\|_\infty^2 B^2 \log^2 d + \|w\|_\infty^2 \sigma_{col} \sqrt{\log d} \leq \|w\|_2^2 \sigma_{\infty} \sigma_{col} \sqrt{\log d} \leq \|w\|_2^2 \sigma_{\infty}^2 \log d + \|w\|_2^2 \sigma_{col}^2 \)).

References

[Abb18] E. Abbe. Community detection and stochastic block models. Foundations and Trends in Communications and Information Theory, 14(1-2):1–162, 2018.

[ABH15] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. IEEE Transactions on Information Theory, 62(1):471–487, 2015.

[ABKK17] N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla. Multisection in the stochastic block model using semidefinite programming. In Compressed Sensing and its Applications, pages 125–162. Springer, 2017.

[AFW19] E. Abbe, J. Fan, and K. Wang. An \( \ell_p \) analysis of kernel principal component analysis with application to spectral clustering. working paper, 2019.

[AFWZ17] E. Abbe, J. Fan, K. Wang, and Y. Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. arXiv:1709.09565, 2017.

[AHB14] T. Alzahrani, K. J. Horadam, and S. Boztas. Community detection in bipartite networks using random walks. In Complex Networks V, pages 157–165. Springer, 2014.

[AL18] A. A. Amini and E. Levina. On semidefinite relaxations for the block model. The Annals of Statistics, 46(1):149–179, 2018.

[ASSS19] A. Agarwal, D. Shah, D. Shen, and D. Song. On robustness of principal component regression. accepted to Neural Information Processing Systems (NeurIPS), 2019.

[Ban18] A. S. Bandeira. Random laplacian matrices and convex relaxations. Foundations of Computational Mathematics, 18(2):345–379, 2018.

[BCL18] L. Balzano, Y. Chi, and Y. M. Lu. Streaming PCA and subspace tracking: The missing data case. Proceedings of the IEEE, 106(8):1293–1310, 2018.

[BL08] P. J. Bickel and E. Levina. Covariance regularization by thresholding. The Annals of Statistics, 36(6):2577–2604, 2008.
[BM15] B. Barak and A. Moitra. Tensor prediction, rademacher complexity and random 3-xor. arXiv preprint arXiv:1501.06521, 2015.

[BM16] B. Barak and A. Moitra. Noisy tensor completion via the sum-of-squares hierarchy. In Conference on Learning Theory, pages 417–445, 2016.

[BY12] Z. Bai and J. Yao. On sample eigenvalues in a generalized spiked population model. Journal of Multivariate Analysis, 106:167–177, 2012.

[CC18a] Y. Chen and E. Candès. The projected power method: An efficient algorithm for joint alignment from pairwise differences. Communications on Pure and Applied Mathematics, 71(8):1648–1714, 2018.

[CC18b] Y. Chen and Y. Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. IEEE Signal Processing Magazine, 35(4):14–31, July 2018.

[CCFM19] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. Mathematical Programming, 176(1-2):5–37, July 2019.

[CFMW19] Y. Chen, J. Fan, C. Ma, and K. Wang. Spectral method and regularized MLE are both optimal for top-K ranking. Annals of Statistics, 47(4):2204–2235, August 2019.

[CFMY19] Y. Chen, J. Fan, C. Ma, and Y. Yan. Inference and uncertainty quantification for noisy matrix completion. arXiv:1906.04159, accepted to Proceedings of the National Academy of Sciences (PNAS), 2019.

[Cha15] S. Chatterjee. Matrix estimation by universal singular value thresholding. The Annals of Statistics, 43(1):177–214, 2015.

[CKST16] Y. Chen, G. Kamath, C. Suh, and D. Tse. Community recovery in graphs with locality. In International Conference on Machine Learning, pages 689–698, 2016.

[CL15] T. T. Cai and X. Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. The Annals of Statistics, 43(3):1027–1059, 2015.

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

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

[CLPC19] C. Cai, G. Li, V. Poor, and Y. Chen. Nonconvex low-rank symmetric tensor completion from noisy data. accepted to Neural Information Processing Systems (NeurIPS), 2019.

[CLX18] Y. Chen, X. Li, and J. Xu. Convexified modularity maximization for degree-corrected stochastic block models. The Annals of Statistics, 46(4):1573–1602, 2018.

[CMW13] T. T. Cai, Z. Ma, and Y. Wu. Sparse PCA: Optimal rates and adaptive estimation. The Annals of Statistics, 41(6):3074–3110, 2013.

[CMW15] T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. Probability theory and related fields, 161(3-4):781–815, 2015.
[CO06] A. Coja-Oghlan. A spectral heuristic for bisecting random graphs. *Random Structures & Algorithms*, 29(3):351–398, 2006.

[CO10] A. Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

[CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[CRV15] P. Chin, A. Rao, and V. Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*, pages 391–423, 2015.

[CSG16] Y. Chen, C. Suh, and A. J. Goldsmith. Information recovery from pairwise measurements. *IEEE Transactions on Information Theory*, 62(10):5881–5905, 2016.

[CT10] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, May 2010.

[CTP19a] J. Cape, M. Tang, and C. E. Priebe. Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *Biometrika*, 106(1):243–250, 2019.

[CTP19b] J. Cape, M. Tang, and C. E. Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5):2405–2439, 2019.

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

[CY12] T. T. Cai and M. Yuan. Adaptive covariance matrix estimation through block thresholding. *The Annals of Statistics*, 40(4):2014–2042, 2012.

[CZ16] T. T. Cai and A. Zhang. Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of multivariate analysis*, 150:55–74, 2016.

[CZ18] T. T. Cai and A. Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics*, 46(1):60–89, 2018.

[Dhi01] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.

[DK70] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[EBW18] J. Eldridge, M. Belkin, and Y. Wang. Unperturbed: spectral analysis beyond Davis-Kahan. In *Algorithmic Learning Theory*, pages 321–358, 2018.

[EK15] N. El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, pages 1–81, 2015.

[EKBB+13] N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, 2013.

[EvdG19] A. Elsener and S. van de Geer. Sparse spectral estimation with missing and corrupted measurements. *Stat*, 8(1):e229, 2019.

[FP16] L. Florencu and W. Perkins. Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959, 2016.
[FPV15] V. Feldman, W. Perkins, and S. Vempala. Subsampled power iteration: a unified algorithm for block models and planted csp’s. In Advances in Neural Information Processing Systems, pages 2836–2844, 2015.

[FWZ16] J. Fan, W. Wang, and Y. Zhong. An $\ell_\infty$ eigenvector perturbation bound and its application to robust covariance estimation. arXiv preprint arXiv:1603.03516, 2016.

[FWZZ18] J. Fan, K. Wang, Y. Zhong, and Z. Zhu. Robust high dimensional factor models with applications to statistical machine learning. arXiv preprint arXiv:1808.03889, 2018.

[GLMZ16] C. Gao, Y. Lu, Z. Ma, and H. H. Zhou. Optimal estimation and completion of matrices with biclustering structures. The Journal of Machine Learning Research, 17(1):5602–5630, 2016.

[GMZZ17] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Achieving optimal misclassification proportion in stochastic block models. The Journal of Machine Learning Research, 18(1):1980–2024, 2017.

[GRESS16] A. Gonen, D. Rosenbaum, Y. C. Eldar, and S. Shalev-Shwartz. Subspace learning with partial information. The Journal of Machine Learning Research, 17(1):1821–1841, 2016.

[Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. IEEE Transactions on Information Theory, 57(3):1548–1566, March 2011.

[GRY11] S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. Inverse Problems, 27(2):025010, 2011.

[GSSB17] M. Ghassemi, Z. Shakeri, A. D. Sarwate, and W. U. Bajwa. STARK: Structured dictionary learning through rank-one tensor recovery. In 2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pages 1–5. IEEE, 2017.

[GV16] O. Guédon and R. Vershynin. Community detection in sparse networks via grothendieck’s inequality. Probability Theory and Related Fields, 165(3-4):1025–1049, 2016.

[HSSS16] S. B. Hopkins, T. Schramm, J. Shi, and D. Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In Proceedings of the forty-eighth annual ACM symposium on Theory of Computing, pages 178–191. ACM, 2016.

[HWX16] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. IEEE Transactions on Information Theory, 62(10):5918–5937, 2016.

[HZC18] B. Hao, A. Zhang, and G. Cheng. Sparse and low-rank tensor estimation via cubic sketchings. arXiv preprint arXiv:1801.09326, 2018.

[JCSX11] A. Jalali, Y. Chen, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. In ICML, volume 11, pages 1001–1008, 2011.

[JH12] J. Josse and F. Husson. Handling missing values in exploratory multivariate data analysis methods. Journal de la Société Française de Statistique, 153(2):79–99, 2012.

[JL09] I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. Journal of the American Statistical Association, 104(486):682–693, 2009.

[JMRT16] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi. Phase transitions in semidefinite relaxations. Proceedings of the National Academy of Sciences, 113(16):E2218–E2223, 2016.

[JNS13] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In ACM symposium on Theory of computing, pages 665–674, 2013.

[JO14] P. Jain and S. Oh. Provable tensor factorization with missing data. In Advances in Neural Information Processing Systems, pages 1431–1439, 2014.
I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, 29(2):295–327, 2001.

H. A. Kiers. Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika*, 62(2):251–266, 1997.

V. Koltchinskii and K. Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. In *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 1976–2013. Institut Henri Poincaré, 2016.

R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, June 2010.

H.-J. Kim, E. Ollila, V. Koivunen, and C. Croux. Robust and sparse estimation of tensor decompositions. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 965–968. IEEE, 2013.

V. Koltchinskii and D. Xia. Perturbation of linear forms of singular vectors under gaussian noise. In *High Dimensional Probability VII*, pages 397–423. Springer, 2016.

L. Lei, P. J. Bickel, and N. El Karoui. Asymptotics for high dimensional regression m-estimates: fixed design results. *Probability Theory and Related Fields*, 172(3-4):983–1079, 2018.

D. B. Larremore, A. Clauset, and A. Z. Jacobs. Efficiently inferring community structure in bipartite networks. *Physical Review E*, 90(1):012805, 2014.

L. Lei. Unified $\ell_{2,\rightarrow \infty}$ eigenspace perturbation theory for symmetric random matrices. *arXiv preprint arXiv:1909.04798*, 2019.

D. N. Lawley and A. E. Maxwell. Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 12(3):209–229, 1962.

J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):208–220, 2013.

M. Lelarge, L. Massoulié, and J. Xu. Reconstruction in the labelled stochastic block model. *IEEE Transactions on Network Science and Engineering*, 2(4):152–163, 2015.

K. Lounici. Sparse principal component analysis with missing observations. In *High dimensional probability VI*, pages 327–356. Springer, 2013.

K. Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664, 06 2012.

Z. Ma. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

L. Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703. ACM, 2014.

C. Mu, B. Huang, J. Wright, and D. Goldfarb. Square deal: Lower bounds and improved relaxations for tensor recovery. In *International conference on machine learning*, pages 73–81, 2014.
| Reference | Authors | Title | Journal/Conference | Year |
|----------|---------|-------|--------------------|------|
| [MNS14]  | E. Mossel, J. Neeman, and A. Sly. | Consistency thresholds for binary symmetric block models. | arXiv preprint arXiv:1407.1591, 3(5), 2014. |
| [MNS15]  | E. Mossel, J. Neeman, and A. Sly. | Reconstruction and estimation in the planted partition model. | Probability Theory and Related Fields, 162(3-4):431–461, 2015. |
| [MS18]   | A. Montanari and N. Sun. | Spectral algorithms for tensor completion. | Communications on Pure and Applied Mathematics, 71(11):2381–2425, 2018. |
| [MSC17]  | X. Mao, P. Sarkar, and D. Chakrabarti. | Estimating mixed memberships with sharp eigenvector deviations. | arXiv preprint arXiv:1709.00407, 2017. |
| [MWCC17] | 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. | arXiv preprint arXiv:1711.10467, 2017. |
| [Nad08]  | B. Nadler. | Finite sample approximation results for principal component analysis: A matrix perturbation approach. | The Annals of Statistics, 36(6):2791–2817, 2008. |
| [OVW18]  | S. O'Rourke, V. Vu, and K. Wang. | Random perturbation of low rank matrices: Improving classical bounds. | Linear Algebra and its Applications, 540:26–59, 2018. |
| [Pau07]  | D. Paul. | Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. | Statistica Sinica, pages 1617–1642, 2007. |
| [PO19]   | E. Pavez and A. Ortega. | Covariance matrix estimation with non uniform and data dependent missing observations. | arXiv preprint arXiv:1910.00667, 2019. |
| [PS17]   | A. Potechin and D. Steurer. | Exact tensor completion with sum-of-squares. | In Conference on Learning Theory, pages 1619–1673, 2017. |
| [PW19]   | A. Pananjady and M. J. Wainwright. | Value function estimation in Markov reward processes: Instance-dependent $\ell_\infty$-bounds for policy evaluation. | arXiv preprint arXiv:1909.08749, 2019. |
| [RCY11]  | K. Rohe, S. Chatterjee, and B. Yu. | Spectral clustering and the high-dimensional stochastic blockmodel. | The Annals of Statistics, 39(4):1878–1915, 2011. |
| [RM14]   | E. Richard and A. Montanari. | A statistical model for tensor PCA. | In Advances in Neural Information Processing Systems, pages 2897–2905, 2014. |
| [RPP13]  | B. Romera-Paredes and M. Pontil. | A new convex relaxation for tensor completion. | In Advances in Neural Information Processing Systems, pages 2967–2975, 2013. |
| [RV15]   | M. Rudelson and R. Vershynin. | Delocalization of eigenvectors of random matrices with independent entries. | Duke Mathematical Journal, 164(13):2507–2538, 2015. |
| [SCC17]  | P. Sur, Y. Chen, and E. J. Candes. | The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. | Probability Theory and Related Fields, pages 1–72, 2017. |
| [SHKM14] | O. Semerci, N. Hao, M. E. Kilner, and E. L. Miller. | Tensor-based formulation and nuclear norm regularization for multienergy computed tomography. | IEEE Transactions on Image Processing, 23(4):1678–1693, 2014. |
| [SHSS16] | Y. Shen, Q. Huang, N. Srebro, and S. Sanghavi. | Normalized spectral map synchronization. | In Advances in Neural Information Processing Systems, pages 4925–4933, 2016. |
| [Sin11]  | A. Singer. | Angular synchronization by eigenvectors and semidefinite programming. | Applied and computational harmonic analysis, 30(1):20–36, 2011. |
[SRK09] J. Salmi, A. Richter, and V. Koivunen. Sequential unfolding SVD for tensors with applications in array signal processing. *IEEE Transactions on Signal Processing*, 57(12):4719–4733, 2009.

[STFP12] D. L. Sussman, M. Tang, D. E. Fishkind, and C. E. Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

[TB77] J. M. Ten Berge. Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42(2):267–276, 1977.

[Ver12] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing, Theory and Applications*, pages 210–268, 2012.

[Vu11] V. Vu. Singular vectors under random perturbation. *Random Structures & Algorithms*, 39(4):526–538, 2011.

[Vu18] V. Vu. A simple SVD algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.

[Wed72] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[XY17] D. Xia and M. Yuan. On polynomial time methods for exact low-rank tensor completion. *Foundations of Computational Mathematics*, pages 1–49, 2017.

[XYZ17] D. Xia, M. Yuan, and C.-H. Zhang. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *arXiv preprint arXiv:1711.04934*, 2017.

[XZ19] D. Xia and F. Zhou. The sup-norm perturbation of HOSVD and low rank tensor denoising. *Journal of Machine Learning Research*, 20(61):1–42, 2019.

[YP14] S.-Y. Yun and A. Proutiere. Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint arXiv:1412.7335*, 2014.

[YP16] S.-Y. Yun and A. Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.

[YWS14] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2014.

[YZ16] M. Yuan and C.-H. Zhang. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics*, 16(4):1031–1068, 2016.

[YZ17] M. Yuan and C.-H. Zhang. Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Transactions on Information Theory*, 63(10):6753–6766, 2017.

[ZA18] Z. Zhou and A. A. Amini. Optimal bipartite network clustering. *arXiv preprint arXiv:1803.06031*, 2018.

[ZB18] Y. Zhong and N. Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

[ZCW18] A. Zhang, T. T. Cai, and Y. Wu. Heteroskedastic PCA: Algorithm, optimality, and applications. *arXiv preprint arXiv:1810.08316*, 2018.

[ZWS19] Z. Zhu, T. Wang, and R. J. Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.

[XZ18] A. Zhang and D. Xia. Tensor SVD: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.