ASYMPTOTICS AND STATISTICAL INFERENCE IN HIGH-DIMENSIONAL LOW-RANK MATRIX MODELS

by
Joshua Agterberg

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland
March 2023

Abstract

High-dimensional matrix and tensor data is ubiquitous in machine learning and statistics and often exhibits low-dimensional structure. With the rise of these types of data is the need to develop statistical inference procedures that adequately address the low-dimensional structure in a principled manner. In this dissertation we study asymptotic theory and statistical inference in structured low-rank matrix models in high-dimensional regimes where the column and row dimensions of the matrix are allowed to grow, and we consider a variety of settings for which structured low-rank matrix models manifest.

Chapter 1 establishes the general framework for statistical analysis in high-dimensional low-rank matrix models, including introducing entrywise perturbation bounds, asymptotic theory, distributional theory, and statistical inference, illustrated throughout via the matrix denoising model. In Chapter 2, Chapter 3, and Chapter 4 we study the entrywise estimation of singular vectors and eigenvectors in different structured settings, with Chapter 2 considering heteroskedastic and dependent noise, Chapter 3 sparsity, and Chapter 4 additional tensor structure. In Chapter 5 we apply previous asymptotic theory to study a two-sample test for equality of distribution in network analysis, and in Chapter 6 we study a model for shared community memberships across multiple networks, and we propose and analyze a joint spectral clustering algorithm that leverages newly developed asymptotic theory for this setting.

Throughout this dissertation we emphasize tools and techniques that are data-driven, nonparametric, and adaptive to signal strength, and, where applicable, noise distribution. The contents of Chapters 2-6 are based on the papers Agterberg et al. (2022b); Agterberg and Sulam (2022); Agterberg and Zhang (2022); Agterberg et al. (2020a) and Agterberg

et al. (2022a) respectively, and Chapter 1 contains several novel results.

Acknowledgements

First and foremost I would like to thank my family, in particular my parents Swati and Daniel for their steadfast support throughout my PhD, without which completing would not have been possible. I cannot overstate their help and love throughout my PhD. I would also like to thank my siblings Sita and Kieran for their support, always finding ways to see me, either in Baltimore, Milwaukee, or elsewhere. Finally, thank you to Hannah, who keeps me honest about what I know and don't know (the answer: I know a lot about this dissertation, but very little else).

My advisor Carey Priebe has been a constant source of support throughout my PhD, always saying that he didn't want to "ruin the good ones." It was through Carey that I first learned *what* types of problems to even consider, how to give a good talk or write a good paper, and how to write emails (the answer: quickly and cryptically). He always supported me teaching my own courses, pursuing my own research interests, traveling to conferences, and was always ready to write a letter of recommendation the day before it was due. Carey is also a fountain of knowledge in classical statistics, and from my first few lectures of his graduate statistical theory class I knew I wanted to work with him. Conversations with Carey in Whitehead 306, on his front porch in the later years, and at dinners with speakers were some of the best times of my PhD.

I would also like to thank all of my collaborators throughout the years, who helped me solidify research problems, fixed my typos, and put up with my work style. In loosely chronological order, they are Dan Sussman, Vince Lyzinski, Youngser Park, Minh Tang, Joshua Cape, Jeremias Sulam, Zachary Lubberts, Jesús Arroyo, and Anru Zhang. I would also like to thank my fellow PhD students, who have been a constant source of ideas. Of note,

I would like to thank Jason Miller, Zachary Pisano, Vittorio Loprinzo, Yashil Sukurdeep, and Philip Kerger for being friends throughout my time at Hopkins. Furthermore, I would like to thank all the wonderful faculty I have interacted with and learned from at Hopkins, some of whom are James Fill, Avanti Athreya, Daniel Naiman, Daniel Robinson, Rene Vidal, Amitabh Basu, Laurent Younes, John Wierman, Chikako Mese, Donniell Fishkind, Mauro Maggioni, Soledad Villar, Ben Grimmer, and Fadil Santosa. I would also like to thank the many wonderful people and researchers I have had the opportunity to interact with outside of Hopkins. They are too numerous to name here.

Finally, I would like to thank my committee Carey Priebe, Zachary Lubberts, Rene Vidal, Jeremias Sulam, and Anru Zhang.

Dedication

This dissertation is dedicated to my grandparents Aba, Ajee, Oma, and Opa.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

Consider the Gaussian sequence model, where one observes n samples of the form

$$x_i = \mu + \sigma\varepsilon_i,$$

where $\varepsilon_i \in \mathbb{R}^d$ are d-dimensional standard Gaussian random variables and $\sigma > 0$ is a standard deviation parameter. To keep the exposition straightforward, throughout this section we will keep our discussion informal, though where necessary we will provide appropriate citations.

Define the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. It is possible to show that $\bar{x}$ satisfies

$$\|\bar{x} - \mu\| \leq C\sigma\sqrt{\frac{d\log(n)}{n}}, \tag{1.1}$$

with high probability (in n), where C is a universal constant (see, e.g., Theorem 3.1.1 of Vershynin (2018)). Moreover, it is straightforward to check that

$$\mathbb{E}\|\bar{x} - \mu\|^2 = \sigma^2\frac{d}{n}.$$

When d is small or finite, these bounds are sufficient to assert that $\|\bar{x} - \mu\|$ is converging to zero as $n \to \infty$ (almost surely and in mean squared error). In addition, it holds that (e.g.,

example 15.16 of Wainwright (2019))

$$\inf_{\text{estimators } \widetilde{x} \text{ of } \mu} \sup_{\mu \in \mathbb{R}^d} \mathbb{E}\|\widetilde{x} - \mu\|^2 \geq c\sigma^2 \frac{d}{n},$$

where the infimum is taken over all estimators $\widetilde{x}$ of μ. Moreover, when d is finite and fixed, it is well known (e.g., van der Vaart (2000)) that $\bar{x}$ coincides with the maximum likelihood estimate for μ, and hence is also asymptotically efficient. In essence, these results in tandem suggest that $\bar{x}$ is optimal for estimating μ (for some appropriate notion of optimality).

However, in many modern applications d can be large relative to n (say $d \approx n$). Therefore, despite its optimalty, $\bar{x}$ may no longer be consistent for μ, as the bound $\sigma^2 \frac{d}{n}$ may not tend to zero as $n \to \infty$. In order to obtain consistent estimators for μ, a common theme in high-dimensional statistics is to impose low-dimensional structural assumptions on the data-generating mechanism. One common such mechanism is via *sparsity*, which in this setting imposes the assumption that at most s of the entries of μ are nonzero. Since μ now belongs to a restricted subset of $\mathbb{R}^d$ (i.e., the set of vectors with at most s nonzero coordinates), it may be that with this additional knowledge one can find an estimator $\bar{x}^{(\text{sparsity})}$ that is consistent for μ.

It can be shown (see Exercise 10.3.8 of Vershynin (2018)) that there is an estimator $\bar{x}^{(\text{sparsity})}$ that satisfies

$$\mathbb{E}\|\bar{x}^{(\text{sparsity})} - \mu\| \leq C\sigma\sqrt{\frac{s\log(ed/s)}{n}};$$

and, moreover, it holds that (Example 15.16, Wainwright (2019))

$$\inf_{\text{Estimators } \widetilde{x} \text{ of } \mu} \sup_{\substack{\mu : \mu \text{ has at most } s \\ \text{nonzero coordinates}}} \mathbb{E}\|\widetilde{x} - \mu\|^2 \geq c\sigma^2 \frac{s\log(ed/s)}{n}.$$

Consequently, as $s \to d$, one obtains the same rate as in the nonsparse setting, whereas the rate improves when s is very small relative to d.

Therefore, we see that a key property emerges: by imposing low-dimensional structural assumptions, one can maintain consistency (in a minimax sense) even in high dimensions.

Informally, this setting refers to the fact that the underlying parameter of interest has low *intrinsic* dimension, while living in a high *ambient* dimension. In modern data science the data need no longer be Euclidean; for example, one may observe matrix or even tensor data instead of vector valued data. Nonetheless, the story above still continues to be valid even in these non-Euclidean settings, with the proviso that the low-dimensional structural assumption now may be tailored to the data type. In many practical settings these low-dimensional structural assumptions are actually motivated by concrete problem instances; we shall see several examples in the following sections.

One common low-dimensional structural assumption to place on matrix data is the assumption of low-rankedness. Here, in lieu of the mean *vector* being sparse, one instead assumes that the mean *matrix* is *low rank*, where "low rank" means that its rank is "small" relative to the row and column dimensions. In many settings it may not be the entire underlying low-rank matrix in which we are interested, but rather its *eigenvectors* and *singular vectors*, or perhaps its individual entries. Therefore, a major focus of this dissertation will be studying the estimation of eigenvectors, singular vectors, and related quantities in asymptotic regimes where the row and column dimensions of the matrix are growing but the rank remains small (either fixed or growing slowly). The focus here will be on obtaining *fine-grained* (e.g., entrywise) bounds in the presence of noise that go beyond previous (coarse-grained) bounds.

Beyond consistency, we are often also interested in performing *hypothesis testing* or obtaining *confidence intervals* for the underlying low-rank matrix of interest (or its associated properties). Therefore, another major focus of this work will be in developing the requisite asymptotic theory and statistical inference procedures for these types of data. In what follows we will elucidate the model more thoroughly, as well as provide several examples that help to contextualize the problems that are studied in this dissertation.

Notation: Within this chapter and the corresponding appendix, for two sequences of numbers a_n and b_n, we write $a_n \lesssim b_n$ if there is a universal constant C such that $a_n \leq Cb_n$, and we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. In addition, we write $a_n = O(b_n)$ if $a_n \lesssim b_n$, and we write $a_n \ll b_n$ if $a_n/b_n \to 0$ as $n \to \infty$. We also write $a_n = \widetilde{O}(b_n)$ if there exists some

constant $c > 0$ such that $a_n = O(b_n \log^c(n))$; i.e., $\widetilde{O}(\cdot)$ hides logarithmic factors. Occasionally we will use the letter C to denote some constant that is allowed to change from line to line. We write $\|\cdot\|$ for the spectral norm or usual Euclidean norm on matrices and vectors respectively, and we let bold letters $\mathbf{U}$ and $\mathbf{S}$ denote matrices. We let e_i denote a standard basis vector, and write $\mathbf{S}_{i\cdot}$ as the i'th row of a matrix $\mathbf{S}$, viewed as a column vector. We write $\|\cdot\|_F$ as the Frobenius norm on matrices, and $\mathbf{I}_r$ denotes the $r \times r$ identity. We write $\mathcal{N}(0, \eta)$ for a Gaussian distribution with variance η. Finally, we write $\|\cdot\|_{\psi_2}$ as the Orlicz norm of a random variable (see Appendix C.3).

1.1 High-Dimensional Low-Rank Matrix Models

In this work we consider a matrix generalization of the Gaussian sequence model, the "signal plus noise" model, where we observe $\widehat{\mathbf{S}}$ of the form

$$\widehat{\mathbf{S}} = \mathbf{S} + \mathbf{N}. \tag{1.2}$$

Typically $\mathbf{S}$ is taken to be a matrix containing important population information and $\mathbf{N}$ is noise, but $\mathbf{S}$ may instead be a tensor, or perhaps a collection of matrices. For now we focus on the setting that $\mathbf{S}$ is a matrix.

By analogy to the example in the previous section, when $\widehat{\mathbf{S}}$ is high-dimensional (i.e., its row and column dimensions are growing), without additional structural assumptions on $\mathbf{S}$, it may not be possible to obtain consistent estimation for $\mathbf{S}$. However, in many practical settings $\mathbf{S}$ naturally exhibits low rank structure, and hence it is possible to retain consistency. The particular definitions of $\mathbf{S}$ and $\mathbf{N}$ change from problem to problem, but the underlying mechanism remains the same – typically $\mathbf{S}$ has underlying low-rank structure (in addition to possibly other structure) that is informed by the problem at hand, and $\mathbf{N}$ consists of noise.

In many problems of interest we are not directly interested in $\mathbf{S}$ itself but rather its eigenvectors, singular vectors, or some function thereof, as these, rather than $\mathbf{S}$ itself, are the parameter of interest or contain important population information. Moreover, as $\mathbf{S}$ is low-rank, if one has a sufficiently strong estimate of its eigenvectors, then one may be able

to translate these results to obtain a strong estimate of $\mathbf{S}$ itself.

In order to formalize and contextualize the types of statistical models we will be discussing, we consider several examples exhibiting the structure in (1.2).

- **Matrix Denoising**: Throughout this chapter we illustrate our main ideas via the canonical problem of *matrix denoising*, which, while perhaps somewhat artificial, is also arguably the simplest model that instills many of the main ideas behind the signal plus noise model in (1.2). Assume that the matrix $\mathbf{S}$ is positive semidefinite and rank r with eigendecomposition

$$\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^\top,$$

where $\mathbf{U} \in \mathbb{R}^{n\times r}$ is a matrix with r othornormal columns and $\Lambda \in \mathbb{R}^{r\times r}$ is a diagonal matrix consisting of the nonzero eigenvalues of $\mathbf{S}$. The noise matrix $\mathbf{N}$ has entries drawn according to

$$\mathbb{E}\mathbf{N}_{ij} = 0; \qquad \mathbb{E}\mathbf{N}_{ij}^2 = \sigma^2,$$

with $\mathbf{N}_{ij}$ independent for $i \leq j$, with $\|\mathbf{N}_{ij}\|_{\psi_2} \leq \sigma$ (e.g., $\mathbf{N}_{ij}$ are Gaussian with variance σ^2). We are interested in estimating the matrix $\mathbf{U}$ in a regime where $n \to \infty$ (and perhaps $r \to \infty$ slowly).

- **High-Dimensional Mixture Models**: The mixture model is a standard statistical model for high-dimensional community data (Amini and Razaee, 2021; Abbe et al., 2022; Zhang and Zhou, 2022; Abbe et al., 2022; Amini and Razaee, 2021; Löffler et al., 2021; Schiebinger et al., 2015; Vempala and Wang, 2004; von Luxburg, 2007; Ding and Sun, 2019; Li et al., 2020a; Little et al., 2020). Suppose one has n observations, where each observation i belongs to one of K different communities (that are unobserved). Denote $z : [n] \to [K]$ as the assignment vector associating each index i to its associated community; i.e. $z(i) = k$ if observation i belongs to community k. Suppose one

observes

$$X_i = \mu_{z(i)} + Y_i,$$

where $\mu_1, \ldots, \mu_K$ are K different mean vectors (associated to each community), and Y_i is a mean-zero independent random variable. To translate this model into the signal plus noise setting, let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the matrix whose rows are X_i, and similarly for $\mathbf{M}$ and $\mathbf{Y}$ respectively. Then one observes

$$\mathbf{X} = \mathbf{M} + \mathbf{Y},$$

where $\mathbf{M}$ has at most K unique rows (and hence is rank at most K).

When d is small or finite, it is feasible to estimate $\mathbf{M}$, but in high dimensions (e.g. $d \gtrsim n$), estimating $\mathbf{M}$ becomes infeasible (information-theoretically) unless the noise is prohibitively small. However, suppose one is only interested in estimating the *communities* (as opposed to the matrix of means $\mathbf{M}$ itself). It can be shown that if $\mathbf{M}$ has singular value decomposition $\mathbf{M} = \mathbf{U}\Lambda\mathbf{V}^\top$, then it holds that $\mathbf{U} = \mathbf{ZR}$, where $\mathbf{Z} \in \{0,1\}^{n \times K}$ is a matrix with exactly one nonzero entry in each row (corresponding to the community membership of each observation), and $\mathbf{R}$ is a $K \times r$ matrix with $r = \text{rank}(\mathbf{M})$ with non-repeated rows. Consequently, knowledge of the matrix $\mathbf{U}$ is sufficient for community recovery, even in high dimensions.

In Chapter 2 we consider application of our theoretical results in a general model to this problem when the rows of $\mathbf{Y}$ have heterogeneous covariances.

- **Principal Component Analysis**: Suppose one has n observations $X_i \in \mathbb{R}^d$ such that

$$\mathbb{E}X_i = 0; \qquad \mathbb{E}X_i X_i^\top = \mathbf{\Sigma}.$$

Assume that

$$\mathbf{\Sigma} = \mathbf{\Sigma}_0 + \mathbf{\Sigma}_\perp,$$

where $\mathbf{\Sigma}_0$ is a rank r matrix of the form $\mathbf{U}\Lambda\mathbf{U}^\top$, where Λ is the diagonal matrix of r leading eigenvalues of $\mathbf{\Sigma}$, and $\mathbf{\Sigma}_\perp$ consists of the bottom $d-r$ eigenvectors and eigenvalues of $\mathbf{\Sigma}$.

In *principal component analysis* one wishes to estimate the matrix $\mathbf{U}$, whose columns can be interpreted as the *directions* with the largest component of the variance. Consider estimating $\mathbf{\Sigma}$ with the sample covariance $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$. Then to translate this into the signal plus noise model, consider the decomposition

$$\widehat{\mathbf{\Sigma}} = \underbrace{\mathbf{\Sigma}_0}_{\mathbf{S}} + \underbrace{\mathbf{\Sigma}_\perp + \left(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right)}_{\mathbf{N}}.$$

Unlike the matrix denoising model, the noise matrix $\mathbf{N}$ is not exactly mean zero, but instead satisfies $\mathbb{E}\mathbf{N}\mathbf{U} = 0$; that is, $\mathbf{N}$ is only mean-zero after projection onto the subspace spanned by the leading eigenvectors of $\mathbf{\Sigma}$. When the eigenvalues of $\mathbf{\Sigma}_0$ are sufficiently separated from the eigenvalues of $\mathbf{\Sigma}_\perp$ (informally, a large portion of the variance is contained in the first r directions), it is possible to estimate $\mathbf{U}$.

In Chapter 3 we consider the estimation of $\mathbf{U}$ under the additional assumption that $\mathbf{U}$ is *sparse*, which is known as *sparse principal component analysis*. See Chapter 3 for further details.

- **Statistical Network Analysis**: In network data one only observes pairwise interactions (edges) between nodes (vertices). Define the *adjacency matrix* $\mathbf{A}$ associated to a network by setting $\mathbf{A}_{ij}$ to be one if there is an edge between vertex i and vertex j, where there are n vertices. Suppose that the graph is undirected, so that $\mathbf{A}$ is symmetric.

  In the *latent space model*, each vertex i has associated to it a low-dimensional Euclidean vector $X_i \in \mathbb{R}^d$. A natural probabilistic model is that the edge probabilities are formed by a (pseudo) inner product between the latent vectors of the given vertices.

Explicitly, one has $\mathbb{P}(\mathbf{A}_{ij} = 1) = X_i^\top X_j$. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ denote the matrices whose rows are the latent vectors X_i. Then we can write

$$\mathbf{A} = \underbrace{\mathbf{X}\mathbf{X}^\top}_{\mathbf{S}} + \underbrace{\left(\mathbf{A} - \mathbf{X}\mathbf{X}^\top\right)}_{\mathbf{N}},$$

Here the matrix $\mathbf{X}\mathbf{X}^\top$ is low rank whenever $d \ll n$, and the noise $\mathbf{A} - \mathbf{X}\mathbf{X}^\top$ is centered Bernoulli noise. One can consider estimating the matrix $\mathbf{X}$ via the leading eigenvectors of the adjacency matrix $\mathbf{A}$, scaled by the square roots of their corresponding eigenvalues.

In Chapter 5 we consider the setting where the inner product is replaced by an pseudo inner product and the latent vectors X_i are drawn from a distribution F_X. We consider observing two networks with adjacency matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ with latent position distributions F_X and F_Y respectively, and we consider testing whether $F_X = F_Y$ (up to identifiability).

In addition, in Chapter 6 we consider observing L networks on the same vertices from a model where the vectors X_i are shown to lie on *rays*. Here the unique rays can be interpreted as communities, where two vertices are in the same community in one network if their latent vectors lie on the same ray. We consider a setting wherein the directions of the rays may change from network to network but the communities remain the same, and we consider jointly estimating the communities with a spectral clustering algorithm.

- **Low-Rank Tensors** In low-rank tensor data, each *matricization* of the underlying signal tensor is low-rank, where we defer the formal definition of matricization to Chapter 4. Unlike the matrix singular value decomposition, tensor singular value decomposition is ill-defined in general. However, for certain low-rank tensor structures there are efficient algorithms, and one such algorithm to estimate the singular vectors is via the *higher-order orthogonal iteration* (HOOI) algorithm, which uses both low-rank structure and the additional tensor structure. In Chapter 4 we study a community-based model for tensor data and propose an algorithm to estimate community memberships

based on HOOI and the underlying spectral geometry.

In all of the above models it is the singular vectors or related quantities in which we are interested. Moreover, in many of the settings above we are not just interested in consistent estimation, but rather in obtaining valid statistical procedures for hypothesis testing and uncertainty quantification. To obtain such procedures, one must necessarily have a good understanding of the limiting asymptotic and distributional properties of given estimates. Therefore, a focus of this dissertation is in establishing asymptotic and distributional theory for eigenvectors, singular vectors, and related quantities in high-dimensional regimes, where "asymptotic and distributional theory" refers to studying estimation error *rates* and limit theorems in appropriate asymptotic regimes.

In order to discuss the primary technical motivation of this dissertation, in the following section we provide a review of matrix perturbation theory and its application to a statistical context. In Section 1.3 we discuss how entrywise perturbation bounds can be used to obtain a more fine-grained understanding of algorithms and techniques, which are the major focuses of Chapter 3 and Chapter 4. Next, in Section 1.4 we discuss different types of asymptotic and distributional theory for these contexts, the first of which is an important tool for proving the main results in Chapter 6 and the second of which forms the main results in Chapter 2. Finally, in Section 1.5 we discuss statistical inference, which is the primary focus of Chapter 5, and in Section 1.6 we discuss the main contributions of this thesis.

Throughout all of this chapter we use matrix denoising as a running example to be able to put these ideas in context. While matrix denoising is perhaps too simple of a model to use in practice, as a mathematical tool it is useful to help understand the underlying ideas. In some cases the results presented will be novel (having not previously appeared in the literature besides in this dissertation).

1.2 Primer on Matrix Perturbation Theory

Given a matrix-valued observation

$$\widehat{\mathbf{S}} = \mathbf{S} + \mathbf{N},$$

two canonical questions in classical matrix perturbation theory are:

1. Can we quantify how much the eigenvalues of $\widehat{\mathbf{S}}$ change as a function of $\mathbf{N}$?
2. Can we quantify how much the eigenvectors of $\widehat{\mathbf{S}}$ change as a function $\mathbf{N}$?

If $\mathbf{S}$ is rectangular or non-symmetric, we can also consider the singular values or singular vectors instead.

A partial answer to the first question is given by the famous *Weyl's inequality*, which states that the eigenvalues (resp. singular values) are bounded via

$$|\widehat{\lambda}_i - \lambda_i| \leq \|\mathbf{N}\|,$$

where $\widehat{\lambda}_i$ are the (ordered) eigenvalues of $\widehat{\mathbf{S}}$ and λ_i are the eigenvalues of $\mathbf{S}$.

To answer the second question in a quantitative manner, one must first quantify in what sense the eigenvectors are converging. One way to quantify eigenvector perturbation is via the $\sin\Theta$ distance between subspaces defined as follows. For a given matrix $\mathbf{U} \in \mathbb{R}^{n\times r}$ with orthonormal columns, let $\mathbf{U}_\perp \in \mathbb{R}^{n\times n-r}$ denote the matrix with orthonormal columns satisfying $\mathbf{U}_\perp^\top \mathbf{U} = 0$ (note that $\mathbf{U}_\perp$ is not necessarily unique). The $\sin\Theta$ distance between two matrices of eigenvectors $\mathbf{U}$ and $\widehat{\mathbf{U}}$ is given by

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\| = \|\mathbf{U}_\perp^\top \widehat{\mathbf{U}}\|;$$

this quantity is unique (despite the non-uniqueness of $\mathbf{U}_\perp$). One can also replace the operator norm $\|\cdot\|$ with any unitarily invariant norm (e.g. $\|\cdot\|_F$). It can be shown that (see Lemma 1 of Cai and Zhang (2018) or Lemma 31) that the (spectral) $\sin\Theta$ distance satisfies

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\| \leq \inf_{\mathbf{W}:\mathbf{W}\mathbf{W}^\top = \mathbf{I}_r} \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}\| \leq \sqrt{2}\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\|.$$

Similar inequalities (with slightly different constants) can be obtained when using the Frobenius norm instead of the spectral norm. We refer the reader to Bhatia (1997) for the details. In essence, the inequalities above reflect the fact that subspaces are defined up to a global rotation (hence the appearance of the additional orthogonal matrix), the multidimensional

analogue of sign-flips.

Remark 1 (The Minimizing Orthogonal Matrix)**.** *For the spectral norm, the orthogonal matrix minimizing* $\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}\|$ *cannot be computed in closed form. However, the Frobenius norm admits an analytic expression. Define the orthogonal matrix*

$$\mathbf{W}_* := \underset{\mathbf{W}:\mathbf{W}\mathbf{W}^\top = \mathbf{I}_r}{\arg\min} \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}\|_F.$$

It can be shown that $\mathbf{W}_* = \operatorname{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})$*, where* $\operatorname{sgn}(\cdot)$ *is the* matrix sign function*, defined as follows. For a given square matrix* $\mathbf{M}$*, let* $\mathbf{W}_1 \Sigma \mathbf{W}_2^\top$ *denote its singular value decomposition. Then the matrix sign function is defined via* $\operatorname{sgn}(\mathbf{M}) := \mathbf{W}_1 \mathbf{W}_2^\top$*. The matrix* $\mathbf{W}_* = \operatorname{sgn}(\mathbf{U}^\top \widehat{\mathbf{U}})$ *has many appealing properties that will be useful in the subsequent sections and chapters.*

Henceforth we consider $\widehat{\mathbf{U}}$ the leading r eigenvectors of $\widehat{\mathbf{S}}$ and $\mathbf{U}$ the leading r eigenvectors of $\mathbf{S}$.

A classic perturbation bound for eigenvectors is then given by the famous Davis-Kahan Theorem (Bhatia, 1997), which states that

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\| \leq \frac{\|\mathbf{N}\|}{\delta},$$

where $\delta = \max\{|\lambda_r - \widehat{\lambda}_{r+1}|, |\lambda_{r+1} - \widehat{\lambda}_r|$, provided $\delta \geq 2\|\mathbf{N}\|$. The result also holds for any other unitarily invariant norm. Extensions to singular vectors are possible via Wedin's Theorem or Theorem 1 of Cai and Zhang (2018).

Now consider the setting that $\mathbf{S}$ is rank r; i.e., $\lambda_i = 0$ for $i \geq r+1$. Suppose that $\lambda_r/4 \geq \|\mathbf{N}\|$. Then by Weyl's inequality,

$$|\widehat{\lambda}_r - \lambda_r| \leq \|\mathbf{N}\| \leq \lambda_r/2,$$

so that $\widehat{\lambda}_r \geq \lambda_r/2$. Similarly,

$$|\widehat{\lambda}_{r+1}| = |\widehat{\lambda}_{r+1} - \lambda_{r+1}| \leq \|\mathbf{N}\|,$$

so that $\widehat{\lambda}_{r+1} \leq \lambda_r/4$. Consequently,

$$\delta = \max\{|\lambda_r - \widehat{\lambda}_{r+1}|, |\lambda_{r+1} - \widehat{\lambda}_r|\} \geq \lambda_r/2,$$

and hence the Davis-Kahan Theorem implies that

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\| \leq 2\frac{\|\mathbf{N}\|}{\lambda_r}.$$

In a statistical context when $\mathbf{N}$ is random, such a bound is useful as it transfers a statement about the random quantity δ (which depends on $\mathbf{N}$ through the eigenvalues of $\widehat{\mathbf{S}}$) to the nonrandom quantity λ_r.

Beyond the Davis-Kahan Theorem, explicit series expansions for the projection matrices $\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top$ about $\mathbf{U}\mathbf{U}^\top$ (from which versions of the Davis-Kahan Theorem can be derived) have been studied in Kato (1995) (for the case of a single eigenvector), and, more recently, in Xia (2021) (for multiple eigenvectors). In Chapter 2 we apply this expansion to establish our main results.

1.2.1 Probabilistic Bounds and Application to Matrix Denoising

In many situations the bounds in the previous subsection can be combined with probabilistic concentration inequalities to yield high-probability upper bounds for the convergence of $\widehat{\mathbf{U}}$ to $\mathbf{U}$ in $\sin\Theta$ distance. In the matrix denoising setting, suppose that $\mathbf{N}$ consists of independent Gaussian noise with common variance σ. Then by Bandeira and Handel (2016), it holds that

$$\|\mathbf{N}\| \leq C\sigma\sqrt{n}$$

with probability at least $1 - n^{-20}$, provided n is sufficiently large. If one further assumes that $\lambda_r \geq C\sigma\sqrt{n}$, then the Davis-Kahan Theorem implies that

$$\|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U})\| \leq C\frac{\sigma\sqrt{n}}{\lambda_r}.$$

Consequently, whenever $\lambda_r/\sigma \gg \sqrt{n}$, the eigenvectors of $\widehat{\mathbf{S}}$ are consistent for the eigenvectors of $\mathbf{S}$ in $\sin\Theta$ distance as $n \to \infty$.

Remark 2 (The Signal to Noise Ratio)**.** *Observe that the term λ_r/σ is invariant to simultaneous rescaling of the magnitude of the noise and the eigenvalues of $\mathbf{S}$. In the matrix denoising setting λ_r/σ can be understood as the* signal-to-noise ratio *(SNR), where the condition $\lambda_r/\sigma \gg \sqrt{n}$ is required for consistency in $\sin\Theta$ distance. In the subsequent chapters we will encounter other notions of the SNR that depend on different problem parameters (as quantified through some notion of signal strength and some notion of noise standard deviation). While the particular definition of SNR may vary from problem to problem, the fundamental concept remains the same.*

Remark 3 (Optimality of Davis-Kahan Theorem for Matrix Denoising)**.** *A natural question is whether the condition $\lambda_r/\sigma \gg \sqrt{n}$ is not just* sufficient*, but also* necessary *for consistency. It can be shown that the minimax rate for subspace estimation in $\sin\Theta$ distance can be lower bounded by $c\frac{\sqrt{n}}{(\lambda_r/\sigma)}$ (Cai and Zhang, 2018), where c is some universal constant. Consequently, the upper bound implied by the Davis-Kahan Theorem (in the context of Matrix Denoising) is actually* optimal *up to universal constants.*

1.3 Entrywise Perturbation Bounds

Thus far we have seen that by combining classical (deterministic) eigenvector perturbation theory with concentration inequalities, one can show that the eigenvectors and singular vectors are consistent in $\sin\Theta$ distance. In some practical settings, such as in high-dimensional mixture models, these types of perturbation bounds can also yield upper bounds on, say, the misclustering error rate of K-means applied to the rows of the estimated singular vector matrices (as in, e.g., Lei and Rinaldo (2015)). However, in clustering problems, the bounds implied by the Davis-Kahan Theorem (and many similar bounds) may be suboptimal. Roughly speaking, for clustering problems the minimax lower bounds are often *exponential* in the signal to noise ratio, and the misclustering error rates implied by the Davis-Kahan Theorem are typically *polynomial* in the signal to noise ratio. In this manner the Davis-Kahan Theorem suffices to prove *consistency*, though it may yield a slower

convergence rate than is information-theoretically attainable.

Furthermore, the misclustering error bounds implied by the Davis-Kahan Theorem may be too weak to guarantee, for example, *perfect clustering*, or guaranteeing that *every* community is estimated correctly with probability tending to one (misclustering rates instead typically show *all but a vanishing fraction* of communities are estimated correctly). In addition, $\sin\Theta$ distances are not metrics on *matrices*, but rather *subspaces*, and hence do not necessarily imply that the individual entries of $\widehat{\mathbf{U}}$ are converging to $\mathbf{U}$, but rather that the angles between the subspaces corresponding to these matrices are tending to zero. Consequently, $\sin\Theta$ bounds are often too coarse to understand how precisely the estimated eigenvector matrices are converging, and therefore cannot be applied to study subsequent inference with eigenvectors, singular vectors, or related quantities. For example, in network analysis problems such as in Chapter 5, finer-grained characterizations of eigenvectors are important to establish the consistency of hypothesis testing procedures.

In entrywise perturbation theory one considers the *entrywise* perturbation of the eigenvectors (up to an orthogonal transformation). Two canonical entrywise norms to use are the entrywise max norm and the $\ell_{2,\infty}$ norm, with the latter defined via

$$\|\mathbf{M}\|_{2,\infty} := \max_i \|\mathbf{M}_{i\cdot}\|.$$

The $\ell_{2,\infty}$ norm exhibits a number of appealing geometric properties discussed in Cape et al. (2019b), and, since for any matrix $\mathbf{M}$ it holds that $\|\mathbf{M}\|_{\max} \leq \|\mathbf{M}\|_{2,\infty}$, the $\ell_{2,\infty}$ perturbation has been more often considered in the literature.

Beyond simply providing perfect clustering results, entrywise perturbation bounds can be used for the following reasons:

- **A refined understanding of the effect of noise on signal**: While the deterministic bounds discussed in the previous sections can provide consistency in $\sin\Theta$ distance, they only do so through the ratio of the overall magnitude of the noise term $\|\mathbf{N}\|$ and the smallest nonzero singular value λ_r. Entrywise perturbation bounds can be used to establish a more refined understanding of how the noise $\mathbf{N}$ affects estimation through the notion of *incoherence*, defined formally for symmetric matrices in the following

subsection.

- **Guarantees for nonconvex algorithm initializations**: A number of nonconvex algorithms begin with a so-called *spectral initialization*, which uses the eigenvectors of an appropriate matrix as an initial starting point for an iterative sequence (as in Chapter 4). Entrywise perturbation bounds can be used to study the entrywise convergence of nonconvex algorithms via an inductive argument.

- **Implicit regularization**: A number of nonconvex algorithms often require not only that the initializations lie within a region of contraction, but also that the initializations have certain desirable properties. Many algorithms explicitly regularize to induce these properties – Theorem 1 can be used to show that these properties hold *automatically*, without requiring the need for explicit regularization.

- **Precursors to asymptotic theory and distributional theory:** Entrywise perturbation bounds and their respective proofs are often required en route to developing asymptotic theory and distributional theory, both major focuses of this dissertation. These further results can be used to establish stronger upper bounds in applications, such as exponential misclustering error rates (e.g. Chapter 6).

- **Subsequent inference**: Finally, entrywise perturbation and the following asymptotic and distributional theory can be used to justify statistical inference procedures built from eigenvectors and singular vectors, as studied in Chapter 5.

Both Cape et al. (2019b) and Abbe et al. (2020) initiated the entrywise analysis of eigenvectors of symmetric matrices, with the first focusing on general deterministic tools and techniques, and the second focusing on a general "leave-one-out" analysis technique. Beyond these two, extensions and refinements have been considered for rectangular matrices (Cai et al., 2021a), network analysis problems (Jin et al., 2019), principal component analysis (Yan et al., 2021), and kernel spectral clustering (Abbe et al., 2022). A general procedure for obtaining entrywise eigenvector perturbation bounds is described in Chen et al. (2021c), and some parts of the subsequent subsection are motivated by their discussions. We refer the interested reader to the related work sections of Chapters 2, 3, and 4.

1.3.1 Entrywise Perturbation Bounds for Matrix Denoising

We now detail a general entrywise perturbation bound for matrix denoising. In order to do so we first discuss the notion of incoherence for symmetric matrices.

The *incoherence parameter* of a symmetric $n \times n$ matrix $\mathbf{S} = \mathbf{U}\Lambda\mathbf{U}^\top$ of rank r with matrix of eigenvectors $\mathbf{U} \in \mathbb{R}^{n\times r}$ is defined as the smallest number μ_0 such that

$$\max_i \|\mathbf{U}_{i\cdot}\| \le \mu_0 \sqrt{\frac{r}{n}}.$$

If $\mathbf{S}$ is the matrix with one nonzero entry λ, then it holds that $\mu_0 = \sqrt{n}$, and if $\mathbf{S}$ is the constant matrix with entries λ/n, it holds that $\mu_0 = 1$. In this manner the incoherence of a matrix is a measure of *spikiness* of the underlying matrix. Informally speaking, if $\mathbf{S}$ is "too spiky" (i.e., with large μ_0), then it may be difficult to obtain consistent *row-wise* estimation of $\mathbf{U}$. This intuition manifests in the following result, establishing the entrywise perturbation of $\widehat{\mathbf{U}}$ (the leading r eigenvectors of $\widehat{\mathbf{S}}$).

Theorem 1 (Entrywise Perturbation Bounds for Symmetric Matrix Denoising)**.** *Consider the matrix denoising problem and suppose that $\lambda_r/\sigma \ge C\sqrt{n\log(n)}$ for some sufficiently large constant C. Suppose that $\mathbf{N}_{ij}$ are subgaussian random variables with ψ_2 norm at most σ and variance σ^2 for $i \le j$. Define $\mathbf{W}_* := \operatorname{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}})$, and suppose $\mathbf{S}$ is incoherent with incoherence constant μ_0. Then there exists a universal constant C' such that with probability at least $1 - O(n^{-20})$*

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} \le C' \frac{\mu_0\sqrt{r\log(n)}}{\lambda_r/\sigma}.$$

We now discuss the consequences of this bound in the context of the previous section.

- **A refined understanding of the effect of noise on signal**: In comparison to the bound implied by the Davis-Kahan Theorem, which states that with high probability

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_F = \inf_{\mathbf{W}} \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}\|_F \le \sqrt{r}\inf_{\mathbf{W}} \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}\| \lesssim \frac{\sigma\sqrt{rn}}{\lambda_r},$$

  Theorem 1 is smaller by a factor of $\mu_0\sqrt{\log(n)}/\sqrt{n}$. When $\mu_0 = O(1)$ (i.e., $\mathbf{S}$ is

"not too spiky"), this result shows that the errors are *spread out* amongst the rows of $\widehat{\mathbf{U}}$. Furthermore, as μ_0 transitions from 1 to $\sqrt{n}$ (or $\mathbf{S}$ becomse "more spiky"), we see that we eventually match the bound implied by the Davis-Kahan Theorem (up to logarithmic terms), which demonstrates how the spectral properties of $\mathbf{S}$ interact with the entrywise noise through the incoherence parameter μ_0. In contrast, the $\sin\Theta$ perturbation bound does not depend on the incoherence parameter.

- **Implicit Regularization**: Observe that by Theorem 1 it holds that

$$\|\widehat{\mathbf{U}}\|_{2,\infty} \leq \|\mathbf{U}\mathbf{W}_*\|_{2,\infty} + \frac{\mu_0\sqrt{r}}{\sqrt{n}}\left(C'\frac{\sqrt{n\log(n)}}{\lambda_r/\sigma}\right) \leq 2\mu_0\sqrt{\frac{r}{n}},$$

which shows that $\widehat{\mathbf{S}}$ has incoherence parameter at most $2\mu_0$ (here we have used the fact that since $\mathbf{W}_*$ is orthogonal, $\|\mathbf{U}\mathbf{W}_*\|_{2,\infty} = \|\mathbf{U}\|_{2,\infty}$).

Many nonconvex algorithms proceed with an initial estimate of the eigenvectors of an appropriate matrix, but they occasionally require an additional thresholding step obtained by setting any rows of the eigenvectors that are "too large" to some tuning parameter δ (e.g., Jing et al. (2021)). Typically the additional thresholding step is performed to guarantee that the empirical eigenvectors are also incoherent, and hence Theorem 1 shows that no additional thresholding is needed to maintain incoherence. In other words, no explicit regularization (in the form of thresholding) is required to induce incoherence.

- **Precursors to limit theorems, distributional theory, and subsequent inference**: The proof and statement of Theorem 1 are important steps necessary to establish the results of Theorem 2, Theorem 3, and Theorem 4 which will be stated in the next sections. These results study limit theorems, distributional theory, and subsequent inference with eigenvectors.

Remark 4 (Comparison to Previous Bounds)**.** *A similar result in this context can be found in Theorem 4.2 of Chen et al. (2021c), who demonstrate that*

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} \lesssim \frac{\mu_0\sigma\kappa\sqrt{r}}{\lambda_r} + \frac{\sigma\sqrt{r\log(n)}}{\lambda_r},$$

where $\kappa = \lambda_1/\lambda_r$ *is the* condition number *of* $\mathbf{S}$. *Observe that this bound is slightly weaker if* $\mu_0 = O(1)$, *and, moreover, their bound holds only for almost surely bounded noise. In contrast, our result above has no dependence on the condition number* κ *when* $\mu_0 = O(1)$. *Furthermore, our proof also lends itself easily to deriving the asymptotic expansions and distributional theory that we will study in the subsequent sections.*

Remark 5 (Optimality)**.** *Recall that a minimal information-theoretical threshold for consistency in* $\sin\Theta$ *distance is that* $\lambda_r/\sigma \gg \sqrt{n}$. *Theorem 1 requires that* $\lambda_r/\sigma \geq C\sqrt{n\log(n)}$ *for some sufficiently large constant* C, *and proves consistency in* $\ell_{2,\infty}$ *distance if* $\lambda_r/\sigma \gg \sqrt{n\log(n)}$. *Consequently, the condition in Theorem 1 is optimal up to the* $\sqrt{\log(n)}$ *factor.*

1.4 Asymptotics and Distributional Theory

While Theorem 1 establishes that the entries of $\widehat{\mathbf{U}}$ are converging to the entries of $\mathbf{U}$ (up to orthogonal transformation) when μ_0 and r are sufficiently small, it falls short of providing the types of asymptotic expansions and distributional theory needed for statistical inference. For example, in Chapter 6 we study the misclustering error rate of a proposed clustering algorithm using estimated eigenvectors, and Theorem 1 is unable to provide such a rate. Consequently, in many settings, it is useful to further refine the notion in which $\widehat{\mathbf{U}}$ is converging to $\mathbf{U}$. Throughout this section $\widehat{\mathbf{U}}$ and $\mathbf{U}$ can be replaced with empirical and population estimates of eigenvectors, singular vectors, or scaled eigenvectors (see Chapter 5).

In classical univariate Gaussian mean estimation (with variance σ^2), it holds that $\bar{x} \to \mu$ as $n \to \infty$, and, furthermore, $\sqrt{n}(\bar{x} - \mu) \to \mathcal{N}(0, \sigma^2)$ in distribution as $n \to \infty$. The additional scaling of $\sqrt{n}$ demonstrates that by "exploding" the errors by $\sqrt{n}$, one obtains a Gaussian distribution. Therefore, by analogy, we are interested in studying under what scaling $\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*$ is Gaussian. Since $\widehat{\mathbf{U}}$ is a random variable whose dimensions are growing with n, it is useful to study only its individual rows. By studying distributional theory, we can determine how precisely the noise and signal interact, and interesting phenomena can manifest. See Chapter 2 for an example of this.

Beyond distributional theory, it is also useful to study *how precisely* the error rate in Theorem 1 comes about via the notion of asymptotic expansions. Roughly speaking, in

this dissertation we refer to "asymptotic expansions" as a means to express $\widehat{\mathbf{U}}$ as $\mathbf{U}$ plus a first-order correction term that is linear or nearly linear in the noise. These types of asymptotic expansions can yield an even more refined understanding of the bounds in the previous section, allowing for stronger results. For example, in Chapter 6 by deriving an asymptotic expansion for the output of our proposed estimator we are able to establish an exponential misclustering error rate (as opposed to polynomial). Furthermore, asymptotic expansions can be used to develop other subsequent inference procedures as in Chapter 5.

1.4.1 Asymptotics for Matrix Denoising

We now consider an asymptotic expansion for the eigenvectors of $\widehat{\mathbf{U}}$ in the matrix denoising context.

Theorem 2 (First-Order Asymptotic Expansion for Matrix Denoising)**.** *Consider the matrix denoising problem and suppose that* $\lambda_r/\sigma \geq C\sqrt{n\log(n)}$ *for some sufficiently large constant* C*. Suppose that* $\mathbf{N}_{ij}$ *are subgaussian random variables with* ψ_2 *norm at most* σ *and variance* σ^2 *for* $i \leq j$*. Define* $\mathbf{W}_* := \operatorname{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}})$*, and suppose* $\mathbf{S}$ *is incoherent with incoherence constant* μ_0*. There is an event* $\mathcal{E}$ *satisfying* $\mathbb{P}(\mathcal{E}) \geq 1 - n^{-10}$ *such that*

$$\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U} = \mathbf{N}\mathbf{U}\Lambda^{-1} + \Gamma,$$

where

$$\|\Gamma\|_{2,\infty} \lesssim \frac{\mu_0\big(r + \sqrt{r\log(n)}\big)}{\sqrt{n}(\lambda_r/\sigma)} + \frac{\mu_0\sqrt{rn}\log(n)}{(\lambda_r/\sigma)^2}.$$

Recall that Theorem 1 yields an entrywise bound of order $\widetilde{O}\left(\frac{1}{\lambda_r/\sigma}\right)$ when $r, \mu_0 = O(1)$. Theorem 2 demonstrates that the eigenvectors $\widehat{\mathbf{U}}$ are (up to orthogonal transformation) equal to $\mathbf{U}$ plus a term that is linear in the noise (the term $\mathbf{N}\mathbf{U}\Lambda^{-1}$) and a second-order residual term. In contrast to Theorem 1, which only reflects the order of the leading-order term $\mathbf{N}\mathbf{U}\Lambda^{-1}$, Theorem 2 explicitly characterizes the structure of the leading-order term. In addition, since $\lambda_r/\sigma \geq C\sqrt{n\log(n)}$, the higher-order term Γ is of significantly smaller order than the bound obtained in Theorem 1. The proof of Theorem 2 relies on the proof

of Theorem 1.

Remark 6 (Comparison to Previous Results)**.** *To the best of our knowledge, Theorem 2 is novel in this particular setting, though similar results have been obtained previously. In Yan et al. (2021) the authors demonstrate the the singular vectors admit a similar asymptotic expansion, but their results do not hold for eigenvectors as ours do here; consequently, their results are not immediately applicable.*

We next consider developing distributional theory for the rows of $\widehat{\mathbf{U}}$. For convenience we consider the setting that $\mu_0, r, \kappa \asymp 1$; this condition can be relaxed with more careful tabulation of error terms (here $\kappa = \lambda_1/\lambda_r$ is the *condition number* of $\mathbf{S}$).

Theorem 3 (Central Limit Theorem for Rows in Matrix Denoising)**.** *Consider the matrix denoising setting with r fixed, where each $\mathbf{N}_{ij}$ are subgaussian random variables with ψ_2 norm at most σ and variance σ^2. Suppose that the condition number κ of $\mathbf{S}$ is bounded, and that $\mu_0 = O(1)$. Suppose further that $\lambda_r/\sigma \gg \sqrt{n}\log(n)$. Then it holds that*

$$\frac{\Lambda}{\sigma}\left(\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U}\right)_{i\cdot} \to \mathcal{N}(0, \mathbf{I}_r)$$

in distribution as $n \to \infty$.

This result, which is based on the technical tools used to develop Theorem 1, further refines the notion in which the error decays at an order λ_r/σ. In classical (univariate, Gaussian) mean estimation, the error rate for estimation is roughly of order $\widetilde{O}(\frac{\sigma}{\sqrt{n}})$, and, by "exploding" the errors by $\sqrt{n}/\sigma$, one obtains a standard Gaussian distribution. Similarly, Theorem 3 demonstrates that by "exploding" the error by λ_r/σ, one obtains a standard Gaussian distribution.

Remark 7 (Asymptotic Variance)**.** *Observe that the asymptotic covariance for the i'th row of $\widehat{\mathbf{U}}$ is given by the matrix $\frac{1}{\sigma}\Lambda$. In other words, the i, l entry of $\widehat{\mathbf{U}}$ is approximately Gaussian with mean $\mathbf{U}_{il}$ and variance λ_l/σ (up to orthogonal transformation). Roughly speaking, the standard deviations of the entries of the l'th singular vector are proportional to the l'th eigenvalue. Consequently, this phenomenon showcases the fact that the variance increases*

when moving further into the spectrum. In Chapter 2 we note a similar phenomenon occurs for the estimated singular vectors.

Remark 8 (Optimality)**.** *The condition that* $\lambda_r/\sigma \gg \sqrt{n}\log(n)$ *is stronger than the condition* $\lambda_r/\sigma \gg \sqrt{n}$ *required for consistency in* $\sin\Theta$ *distance by a factor of* $\log(n)$*. In the proof of this result we did not put much effort into eliminating logarithmic terms, and it may be that further improvements are possible down to the condition* $\lambda_r/\sigma \gg \sqrt{n}$*. However, such refinements are cumbersome and not the main point of this section.*

1.5 Statistical Inference

Thus far we have considered entrywise convergence *rates* (Theorem 1), asymptotic expansions (Theorem 2), and distributional theory (Theorem 3) for eigenvectors, singular vectors, and related quantities. Many similar (albeit more complicated) results form the cornerstone of the main results in subsequent chapters. However, in many settings, while these results are elegant and mathematically interesting, they are not immediately practical, as one may need to estimate additional parameters of interest or perform some hypothesis test that depends on an asymptotic distribution that requires knowledge of some parameter of the distribution. These analyses are intimately tied to the particular inference problem at hand, and, as such, may require case-by-case analysis.

Potential inferential problems of interest may include:

- **Confidence intervals and regions for eigenvectors and singular vector estimates**: In many settings one may require a high degree of confidence in an estimate, such as in applications to the medical sciences. While results concerning asymptotic and distributional theory may reveal the structure of how the noise interacts with the signal through its low-dimensional structure, they often fail to provide a means for obtaining confidence intervals or regions for estimates.
- **Testing vertex memberships in network analysis**: Consider a network model where vertices are permitted to belong to a convex combination of communities; i.e., each vertex has a membership vector describing the intensity of membership in each

community (see, e.g., Chapter 4 for a similar model). Alternatively, consider a setting wherein each vertex has associated to it a latent low-dimensional Euclidean vector. In either setting one may be interested in testing whether two vertices have the same membership vectors or the same latent vectors. The results in the subsequent section are motivated by several works studying this hypothesis test (Fan et al., 2022; Du and Tang, 2022).

- **Global network hypothesis testing**: Given a single network or multiple networks, there are a number of reasonable tests of practical interest. Consider, for example, a goodness-of-fit test, or the two-sample test of equality of distribution for a network. The results in the previous section do not establish whether test statistics constructed from estimated quantities can yield consistent hypothesis testing in general. In Chapter 5 we demonstrate that a test statistic built from scaled eigenvectors yields consistent testing by applying previous asymptotic theory to the particular setting we consider.

- **Uncertainty quantification and hypothesis testing for algorithms and techniques**: Beyond simply running an algorithm, one may be interested obtaining confidence intervals for the final output. The tools and techniques in the previous section can be used to derive confidence intervals and testing guarantees for estimates constructed from observations, providing a data-driven manner to quantify the uncertainty of an algorithm.

In all of these settings, while additional analysis may be required to demonstrate consistency of a given inference procedure, it is through the analysis in the previous sections that we are able to arrive at the intuition behind developing these tools.

1.5.1 Hypothesis Testing in Matrix Denoising

For our final application to matrix denoising, we consider a hypothesis test statistic inspired by some of the ideas in Fan et al. (2022) and Du and Tang (2022), where we consider testing

whether two given rows of $\mathbf{S}$ are equal. Explicitly, consider the hypotheses

$$H_0 : \mathbf{S}_{i\cdot} = \mathbf{S}_{j\cdot};$$
$$H_A : \mathbf{S}_{i\cdot} \neq \mathbf{S}_{j\cdot}.$$

Note that since $\mathbf{S}$ is low rank, under the null it holds that $(\mathbf{U}\Lambda)_{i\cdot} = (\mathbf{U}\Lambda)_{j\cdot}$. Our test statistic is motivated by this idea.

Since we do not have access to $\mathbf{U}\Lambda$, we propose to use $\widehat{\mathbf{U}}\widehat{\Lambda}$, and we construct a test statistic defined via

$$T_{ij}^2 := \frac{1}{2\sigma^2}\|(\widehat{\mathbf{U}}\widehat{\Lambda})_{i\cdot} - (\widehat{\mathbf{U}}\widehat{\Lambda})_{j\cdot}\|^2,$$

where we assume that σ is known for convenience. The following result demonstrates the asymptotic distribution of T_{ij}^2 under the null and local alternatives respectively.

Theorem 4. *Instate the conditions of Theorem 3 and define*

$$T_{ij}^2 := \frac{1}{2\sigma^2}\|(\widehat{\mathbf{U}}\widehat{\Lambda})_{i\cdot} - (\widehat{\mathbf{U}}\widehat{\Lambda})_{j\cdot}\|^2.$$

Then:

- *(Consistency under the null) If $\mathbf{S}_{i\cdot} = \mathbf{S}_{j\cdot}$, it holds that $T_{ij}^2 \to \chi_r^2$, where χ_r^2 denotes a χ^2 random variable with r degrees of freedom.*
- *(Consistency under local alternatives) If it holds that $\|\mathbf{S}_{i\cdot} - \mathbf{S}_{j\cdot}\| \gg \sigma$, then it holds that $\mathbb{P}(T_{ij}^2 > C) \to 1$ for any $C > 0$. If instead it holds that $\frac{1}{2\sigma^2}\|\mathbf{S}_{i\cdot} - \mathbf{S}_{j\cdot}\|^2 \to \mu > 0$, then $T_{ij}^2 \to \chi_r^2(\mu)$, where $\chi^2(\mu)$ denotes a noncentral χ^2 distribution with noncentrality parameter μ and r degrees of freedom.*

Theorem 4 demonstrates that the test statistic T_{ij}^2 is asymptotically consistent under the null and the local alternative $\|\mathbf{S}_{i\cdot} - \mathbf{S}_{j\cdot}\| \gg \sigma$. In essence, Theorem 4 demonstrates that despite only having access to the matrix $\widehat{\mathbf{S}}$, one can still perform valid and principled statistical inference by harnessing the low-rank structure. This observation is used as a guiding principle for the rest of this dissertation. The hypothesis test considered in this

section is by no means the only test or statistical inference problem of interest; however, it is useful to establish the spirit of the types of problems that may be practically important, but still statistically principled.

Remark 9 (Relationship to Previous Results)**.** *The test statistic in this section is closely inspired by similar test statistics for related problems considered in Fan et al. (2022) and Du and Tang (2022), both of whom study a related problem in network analysis. To the best of our knowledge, such a result in a general matrix denoising setting has not been considered in the literature. Since we assume homoskedasticity and knowledge of σ^2, our test statistic requires no estimation of the covariance matrix, whereas these other two require estimation of the limiting covariance. Furthermore, our local power result is likely optimal given the local power of the Hotelling T^2 test statistic in r dimensions, something that is obfuscated in the network setting (with heteroskedastic and non (sub)Gaussian noise).*

1.6 Contributions of This Thesis

We now outline in more detail the contributions of this thesis, as well as some potential avenues for future work.

- **Chapter 2**: In this chapter we consider a generalization of the matrix denoising model, where the signal matrix is rectangular and low rank, and where the noise matrix is permitted to have dependence within rows and heteroskedasticity between them. We present an estimator for the left singular vectors of the signal matrix, and we show that our estimator is asymptotically Gaussian around the true left singular vectors (modulo orthogonal transformation), with limiting covariance dependent on how the signal and noise interact with each other. We also apply our results to high-dimensional mixture models, establishing consistency of data-driven confidence regions in the fixed-rank setting. With respect to this chapter, the main results in Chapter 2 are most closely related to Theorem 3.
- **Chapter 3**: This chapter considers the *sparse principal component analysis* (sparse PCA) model, where the leading few eigenvectors of the population covariance are

assumed to be sparse. Under an assumption of *sparsistency* (see Chapter 3), we establish $\ell_{2,\infty}$ perturbation bounds of a similar form to Theorem 1 for this model. These results reveal how the low-rank structure, sparsity structure, and noise interact.

- **Chapter 4**: In this chapter we consider low Tucker rank tensors, which are a form of low-rank matrix model with additional tensorial structures. We study a model for community membership for tensor data, and we propose an estimator for the membership matrices. To prove our main results we also establish $\ell_{2,\infty}$ perturbation bounds for the output of *higher-order orthogonal iteration*, an algorithm for computing the tensor singular value decomposition that uses the additional tensorial structure. Our main results are most similar to Theorem 1, albeit for the output of an iterative algorithm instead of simply the singular vectors. We also apply our procedure to several different datasets.
- **Chapter 5**: This chapter considers a two-sample network hypothesis test to test whether two networks have the same distribution. Leveraging the results of Rubin-Delanchy et al. (2022) (who establish asymptotic expansions for the *scaled* eigenvectors of a similar form to Theorem 2) we show that a two-sample test statistic is consistent for this hypothesis. The test statistic uses the technical machinery of optimal transport combined with a maximum mean discrepancy (Gretton et al., 2012) computed using the scaled eigenvectors.
- **Chapter 6**: In this chapter we study a statistical model for multilayer networks on the same vertex set, wherein each network has the same community structure. We propose a joint spectral clustering algorithm that leverages information from all the networks but still respects their individual heterogeneity. We establish an expected misclustering error rate that improves exponentially with mutiple networks, and we apply our algorithm to US flight data. To establish our main results we provide an asymptotic expansion for the output of our algorithm similar in spirit to Theorem 2, but now relying explicitly on the individual network-level parameters.

It is also worth noting again that several of the results in this chapter are in fact novel. In particular, Theorem 2 and Theorem 3 have not been established in this setting in the

literature previously to the best of my knowledge, and while Theorem 4 is similar to and inspired by results in Fan et al. (2022) and Du and Tang (2022), the result in this setting and for this null hypothesis have not been studied.

1.6.1 Future Work

Finally, while this dissertation makes an initial foray into establishing asymptotic and distributional theory for low-rank matrix models, there are there are still a number of potential future works, including, but not limited to:

- **Distributional Theory and Uncertainty Quantitification for Tensor Data**: While Chapter 4 considers the $\ell_{2,\infty}$ estimation of population singular vectors, it falls short of providing distributional theory and uncertainty quantification in the spirit if Theorem 3. In ongoing work we are developing this theory and applying it to a number of potential inferential problems of interest, such as a vertex testing problem.
- **Statistical Inference for Sparse Models**: The results of Chapter 3 rely on an assumption that the nonzero support of the sparse eigenvectors are found with high probability. In many settings one may wish to *test* whether an individual row is zero to obtain a higher degree of confidence for the output. In sparse models it is common to have to *debias* the output of a given algorithm, but such theory is limited in the sparse PCA model (to the best of our knowledge, only Janková and van de Geer (2021) study this problem, and only for the rank one setting). It would be useful to provide hypothesis testing guarantees in a general setting.
- **Fine-grained estimation bounds and statistical inference for nonconvex algorithms**: In Chapter 4 we study the output of HOOI, a nonconvex algorithm for estimating tensor singular vectors. However, there are a number of different nonconvex algorithms tailored to different settings; for example, one may have algorithms tailored to other types of tensor structures, algorithms allowing for outliers, or algorithms that allow for additional sparse structure (e.g., as in sparse PCA). It is of interest to provide fine-grained perturbation bounds in the spirit of Theorem 1 for these algorithms, potentially as a precursor to developing inference techniques.

- **Multilayer network analysis**: While Chapter 5 and Chapter 6 consider two problems related to multiple network analysis, there are still a number of interesting problems remaining. For example, it may be interesting to study other joint community models that generalize the problem studied in Chapter 6, and it may be interesting to determine the fundamental lower bounds for these models. Furthermore, existing techniques ignore new vertices, often causing one to throw away potentially useful information; it is of interest to come up with methodology that accounts for these new vertices.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Entrywise Estimation of Singular Vectors of Low-Rank Matrices with Heteroskedasticity and Dependence

2.1 Introduction

Consider the signal-plus-noise model

$$\widehat{M} = M + E,$$

where $M \in \mathbb{R}^{n \times d}$ is a deterministic rank r signal matrix and E is a mean-zero noise matrix. We assume M has the singular value decomposition

$$M = U \Lambda V^\top,$$

where U is an $n \times r$ matrix with orthonormal columns, Λ is an $r \times r$ diagonal matrix whose entries are the r nonzero singular values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ of M, and V is an $d \times r$ matrix with orthonormal columns. Letting $E_i^\top$ denote the i'th row of the noise matrix E,

we suppose each noise vector E_i is of the form

$$E_i = \Sigma_i^{1/2} Y_i,$$

where $\mathbb{E}Y_i = 0$, $\Sigma_i \in \mathbb{R}^{d\times d}$ is a positive semidefinite matrix allowed to depend on the row i, and the coordinates $Y_{i\alpha}$ of the vector Y_i are independent subgaussian random variables satisfying $\mathbb{E}Y_{i\alpha}^2 = 1$. Define the *signal to noise ratio*

$$\text{SNR} := \frac{\lambda_r}{\sigma\sqrt{rd}},$$

where $\sigma^2 := \max_i \|\Sigma_i\|$ is the maximum spectral norm of the covariances of each row.

In this work we study the entrywise estimation of the matrix U (or column space of M) in the "quasi-asymptotic regime" wherein n and d are assumed to be large but finite. Because we allow the rows of E to have different covariances, the "vanilla SVD" can be biased, so we propose a bias-corrected estimator for the matrix U and analyze its limiting distribution when the signal to noise ratio is large enough relative to the dimension of the problem. Furthermore, we do not assume distinct singular values for M.

Our main contributions are the following:

- We provide a nonasymptotic Berry-Esseen Theorem (Theorem 5) for the entries of our proposed estimator for U under general assumptions on the signal matrix M and the noise matrix E;
- We study the $\ell_{2,\infty}$ approximation (Theorem 6) of our proposed estimator to U which matches previous $\ell_{2,\infty}$ bounds in the special case of independent noise.
- We apply our results to the particular setting of a K-component subgaussian mixture model (Corollary 2) and show that one can accurately estimate the resulting limiting covariance of the rows of our estimator within each community (Corollary 3), allowing us to derive data-driven, asymptotically valid confidence regions.

Since we allow for dependence within each row of the noise matrix E, our asymptotic results highlight the geometric relationship between the covariance structure and the singular subspaces of M. Furthermore, our estimator is based off the HeteroPCA algorithm proposed

in Zhang et al. (2022), so as a byproduct of our results, we also provide a more refined analysis of this algorithm, leading to an upper bound on the $\ell_{2,\infty}$ error that scales with the noise.

The organization of the paper is as follows. In the next subsection we discuss related work, and we end this section with notation and terminology we will use throughout. In Section 2.2, we recall the HeteroPCA algorithm of Zhang et al. (2022), and use this to define our estimator $\widehat{U}$ for U. We also discuss alternative approaches to the problem and some of the shortcomings of those approaches which have motivated the present work. In Section 2.3, we state our main theorems, namely our $\ell_{2,\infty}$ concentration and Berry-Esseen theorems. We also discuss the various assumptions of our model, and compare these to previous work. In Section 2.3.2, we discuss the statistical implications of our results for mixture distributions. We further illustrate our results in simulations in Section 2.4. Discussion of these results and potential future work is in Section 2.5. Finally, Section 2.6 contains the proof of Theorem 6 and a proof sketch of Theorem 5. The full proof of Theorem 5 and additional supplementary lemmas are contained in the Appendices.

2.1.1 Related Work

Spectral methods, which refer to a collection of tools and techniques informed by matrix analysis and eigendecompositions, underpin a number of methods used in high-dimensional multivariate statistics and data science, including but not limited to network analysis (Abbe et al., 2022, 2020; Agterberg et al., 2020a; Athreya et al., 2018; Cai et al., 2021a; Fan et al., 2022; Jin et al., 2019; Lei, 2019; Lei and Rinaldo, 2015; Mao et al., 2020; Rubin-Delanchy et al., 2020; Zhang et al., 2020b), principal components analysis, (Cai et al., 2021a; Cai and Zhang, 2018; Cai et al., 2021b; Koltchinskii et al., 2020; Koltchinskii and Lounici, 2017; Koltchinskii and Xia, 2016; Koltchinskii and Lounici, 2016; Wang and Fan, 2017; Johnstone and Lu, 2009; Lounici, 2013, 2014; Xie et al., 2022; Zhu et al., 2019), and spectral clustering (Abbe et al., 2022, 2020; Amini and Razaee, 2021; Cai et al., 2021a; Lei, 2019; Löffler et al., 2021; Schiebinger et al., 2015; Srivastava et al., 2021). In addition, eigenvectors or related quantities can be used as a "warm start" for optimization methods (Chen et al. (2019b, 2021c); Chi et al. (2019); Lu and Li (2017); Ma et al. (2020); Xie and Xu (2020); Xie (2021)), yielding provable convergence to quantities of interest provided the initialization

is sufficiently close to the optimum. The model we consider includes as a submodel the high-dimensional K-component mixture model. High-Dimensional mixture models play an important role in the analysis of spectral clustering (Abbe et al., 2022; Amini and Razaee, 2021; Löffler et al., 2021; Schiebinger et al., 2015; Vempala and Wang, 2004; von Luxburg, 2007) and classical multidimensional scaling (Ding and Sun, 2019; Li et al., 2020a; Little et al., 2020).

To analyze these methods, researchers have used existing results on matrix perturbation theory such as the celebrated Davis-Kahan $\sin\Theta$ Theorem (Bhatia, 1997; Yu et al., 2014; Chen et al., 2021c), which provides a deterministic bound on the difference between the eigenvectors of a perturbed matrix and the eigenvectors of the unperturbed matrix, provided the perturbation is sufficiently small. Unfortunately, the Davis-Kahan Theorem and classical approaches to matrix perturbation analysis may fail to provide entrywise guarantees for estimated eigenvectors, though there has been work on studying the subspace distances in the presence of random noise (Li and Li, 2018; Xia, 2021; Bao et al., 2021; O'Rourke et al., 2018; Ding, 2020).

Entrywise eigenvector analysis plays an important role in furthering the understanding of spectral methods in a number of statistical problems (Abbe et al., 2020, 2022; Cai et al., 2021a; Cape et al., 2019a,b; Chen et al., 2021c; Damle and Sun, 2020; Eldridge et al., 2018; Fan et al., 2018; Lei, 2019; Luo et al., 2020; Mao et al., 2020; Rohe and Zeng, 2020; Xia and Yuan, 2020; Xie et al., 2022; Zhong and Boumal, 2018; Zhu et al., 2019). A number of works have studied entrywise eigenvector or singular vector analysis when the noise matrix E consists of independent entries (Abbe et al., 2020, 2022; Chen et al., 2021c; Cape et al., 2019a; Cai et al., 2021a; Lei, 2019), and some authors have also studied the estimation of linear forms of eigenvectors (Chen et al., 2021b; Cheng et al., 2021; Fan et al., 2020; Koltchinskii and Xia, 2016; Koltchinskii and Lounici, 2017; Koltchinskii and Xia, 2016; Koltchinskii, 1998; Koltchinskii et al., 2020; Li et al., 2021). In the present work, we extend existing results on entrywise analysis by allowing for dependence in the noise matrix. We address this more general setting using a combination of matrix series expansions (Cape et al., 2019a; Chen et al., 2021b; Eldridge et al., 2018; Xia, 2019, 2021; Xia and Yuan, 2020; Xie et al., 2022), leave-one-out analysis (Abbe et al., 2020, 2022; Chen et al., 2021c; Lei,

2019), and careful conditioning arguments.

While entrywise statistical guarantees refine classical deterministic perturbation techniques, they do not necessarily allow for studying the distributional properties of the eigenvectors. Several works have studied the asymptotic distributions of individual eigenvectors (Cape et al., 2019a; Fan et al., 2020) or $\sin\Theta$ distances (Bao et al., 2021; Ding, 2020; Li and Li, 2018; Xia, 2021), but there are very few finite-sample results on the distribution of the individual entries of singular vectors rather than eigenvectors, and existing results often depend on independence of the noise. We explicitly characterize the distribution of the individual entries of the estimated singular vectors, which showcases the effect that the geometric relationship of the covariance structure of the rows of E and the spectral structure of the signal matrix M has on this distribution.

Finally, the bias of the singular value decomposition in the presence of heteroskedastic noise has been addressed in a number of works (Cai et al., 2021a; Abbe et al., 2022; Florescu and Perkins, 2016; Koltchinskii and Gine, 2000; Leeb and Romanov, 2021; Lei and Lin, 2022; Lounici, 2014). A common method for addressing this is the diagonal deletion algorithm, for which an entrywise analysis is carried out in Cai et al. (2021a) for several different statistical problems. However, as identified in Zhang et al. (2022), in some situations diagonal deletion incurs unnecessary additional error, leading them to propose the HeteroPCA algorithm which we further study here. We include detailed comparisons to existing work in Section 2.3.

2.1.2 Notation

We use capital letters for both matrices and vectors, where the distinction will be clear from context, except for the letter C, which we use to denote constants. For a matrix M, we write M_{ij} as its i,j entry, $M_{\cdot j}$ for its j'th column and $M_{j\cdot}$ for its j'th row. The symbol e_i represents the standard basis vector in the appropriate dimension. We use $\|\cdot\|$ to denote the spectral norm for matrices and the Euclidean norm for vectors, and $\|\cdot\|_F$ as the Frobenius norm for matrices. We write the $\ell_{2,\infty}$ norm of a matrix as $\|U\|_{2,\infty} = \max_i \|U_{i\cdot}\|$ which is the maximum Euclidean row norm. We consider the set of orthogonal $r \times r$ matrices as $\mathbb{O}(r)$, and we exclusively use the letter $\mathcal{O}$ to mean an element of $\mathbb{O}(r)$. We write $\langle \cdot, \cdot \rangle$ as the

standard Euclidean inner product.

For a random variable X taking values in $\mathbb{R}$, we let $\|X\|_{\psi_2}$ denote its ψ_2 (subgaussian) Orlicz norm; that is,

$$\|X\|_{\psi_2} := \sup_{z \in \mathbb{R}} \{\mathbb{E} \exp(X^2/z) \leq 2\},$$

and for a random variable Y taking values in $\mathbb{R}^d$, we let $\|Y\|_{\psi_2}$ denote $\sup \|\langle Y, u \rangle\|_{\psi_2}$, where the supremum is over all unit vectors u. Similarly we denote $\|\cdot\|_{\psi_1}$ as the ψ_1 (subexponential) Orlicz norm. For more details on relationships between these norms, see Vershynin (2018).

For two equal-sized matrices U_1 and U_2 with orthonormal columns, we denote the $\sin\Theta$ distance between them as

$$\|\sin\Theta(U_1, U_2)\| := \|U_1 U_1^\top - U_2 U_2^\top\|.$$

For details on the $\sin\Theta$ distance between subspaces, see Bhatia (1997), Chen et al. (2021c), Cape et al. (2019b), or Cape (2020). For a square matrix M, we write $\Gamma(M)$ to be the hollowed version of M; that is, $\Gamma(M)_{ij} = M_{ij}$ for $i \neq j$ and $\Gamma(M)_{ii} = 0$. We write $G(M) := M - \Gamma(M)$.

Occasionally we will write $f(n,d) \lesssim g(n,d)$ if there exists a constant C sufficiently large such that $f(n,d) \leq Cg(n,d)$ for sufficiently large n and d. We also write $f(n,d) \ll g(n,d)$ if $f(n,d)/g(n,d)$ tends to zero as n and d tend to infinity. We write $f(n,d) = O(g(n,d))$ if $f(n,d) \lesssim g(n,d)$. For two numbers a and b, we write $a \vee b$ to denote the maximum of a and b. Finally we write $f(n,d) \asymp g(n,d)$ if $f(n,d) = O(g(n,d))$ and $g(n,d) = O(f(n,d))$.

2.2 Background and Methodology

We consider a rectangular matrix $M \in \mathbb{R}^{n \times d}$ in the "high-dimensional regime" wherein n and d are large and comparable, though we are interested in the setting that d is larger than n. We observe $\widehat{M} = M + E$ where the additive error matrix E has rows $E_i^\top$ which are independent with covariance matrices $\Sigma_i \in \mathbb{R}^{d \times d}$ that are allowed to vary between rows. We write the singular value decomposition of M as $U\Lambda V^\top$, where $U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{d \times r}$ are

matrices with orthonormal columns and $\Lambda \in \mathbb{R}^{r \times r}$ is diagonal with nonincreasing positive diagonal entries $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ that are not necessarily distinct. We note that this decomposition of M is not unique, so we fix some choice of U, which necessarily fixes a choice of V also. Our results account for the nonuniqueness of U by aligning the estimator $\widehat{U}$ to U through right-multiplication by an $r \times r$ orthogonal matrix. For more details on this alignment procedure, see, for example, Chen et al. (2021c) or Cape (2020).

Algorithm 1 HeteroPCA (Algorithm 1 of Zhang et al. (2022))

Require: Input matrix $A + Z$, rank r, and maximum number of iterations T
1: Let $N_0 := \Gamma(A + Z)$; $T = 0$
2: **repeat**
3: Take SVD of $N_T := \sum_i \lambda_i^{(T)} U_{\cdot i}^{(T)} (V_{\cdot i}^{(T)})^\top$
4: Set $\widetilde{N}_T := \sum_{i \leq r} \lambda_i^{(T)} U_{\cdot i}^{(T)} (V_{\cdot i}^{(T)})^\top$ the best rank r approximation of N_T
5: Set $N_{T+1} := G(\widetilde{N}_T) + \Gamma(N_T)$
6: $T = T + 1$
7: **until** convergence or maximum number of iterations reached
8: **return** $\widehat{U} := U^{(T)}$

Since U may equivalently be understood to be the matrix of eigenvectors of $A = MM^\top$ corresponding to its r nonzero eigenvalues, it is natural to consider A and its noisy counterpart $A + Z$, where $Z = ME^\top + EM^\top + EE^\top$, sometimes referred to as the "sample Gram matrix" in the literature (e.g. (Cai et al., 2021a)). The matrix $\mathbb{E}[Z]$ is diagonal with diagonal entries $\mathrm{Tr}(\Sigma_i)$. When d is large and the rows of E are heteroskedastic, this means that the eigenvectors of $A + \mathbb{E}[Z]$ may not well-approximate those of A.

Authors have suggested hollowing the matrix $A + Z$ as a method to correct this bias, which amounts to using the eigenvectors of $\Gamma(A + Z)$ as the estimator for those of A. An analysis of this approach is given in Cai et al. (2021a) and Abbe et al. (2022), though this is not the primary focus of the latter. Unfortunately, while the eigenvectors of $\Gamma(A + Z)$ may be closer to the eigenvectors of A than those of $A + Z$, they still incur a nontrivial, deterministic bias owing to the loss of information along the diagonal of A. In Zhang et al. (2022), the authors provide an example where the eigenvectors of $\Gamma(A + Z)$ do not yield a consistent estimator for U in the regime that n and d tend to infinity with $d \asymp n$. This motivates their alternative approach to correcting the bias, the HeteroPCA algorithm, which we review in Algorithm 1. In essence, the algorithm proceeds by iteratively re-scaling the

diagonals, attempting to fit the off-diagonal entries of $A + Z$ while maintaining the low rank r. We refer to the output of this algorithm after sufficiently many iterations as $\widehat{A}$ (see Theorem 5 for a quantitative condition on the number of iterations), and prove in Lemma 2 that it well-approximates the "idealized" perturbation of A given by $\widetilde{A} = A + \Gamma(Z)$ in spectral norm. This latter matrix has mean A, meaning that the bias along the diagonal has been accounted for. Removing this bias is also important for our distributional results, since otherwise one must center around the eigenvectors of $\Gamma(A)$ rather than those of A, but our interest is in the latter quantity.

The leading eigenvectors $\widehat{U}$ of this matrix $\widehat{A}$ serve as our estimator for U, and our concentration and distributional results demonstrate the quality of this estimator. This builds on the work in Zhang et al. (2022), where they prove a bound on the $\sin\Theta$ distance between $\widehat{U}$ and U, showing that $\widehat{U}$ is a consistent estimator for U. While a bound on the discrepancy between $\widehat{U}$ and U in this metric is a first step towards the entrywise concentration results we obtain here, our results require much tighter control of the error between them. In addition, measuring the error in this way does not lend itself to the distributional results we consider in Theorem 5.

In the case that M is well-conditioned and has highly incoherent singular vectors U and V, the bias associated with diagonal deletion is not too severe, and the difference in performance between the eigenvectors of $\widehat{A}$ and $\Gamma(A + Z)$ may not be significant, especially when n and d are large. On the other hand, for moderate n and d, or for moderate levels of incoherence in U and V, the bias incurred by diagonal deletion is highly significant. We consider an example of this in Figure 2.1, in the setting of estimating memberships in a Gaussian mixture model, under heteroskedasticity and dependence. The noise level in this problem is relatively small, which suggests that consistent estimation should be possible for the right estimator, however the moderate incoherence causes a severe breakdown in the performance of the estimate obtained by diagonal deletion, while the HeteroPCA algorithm continues to perform well.

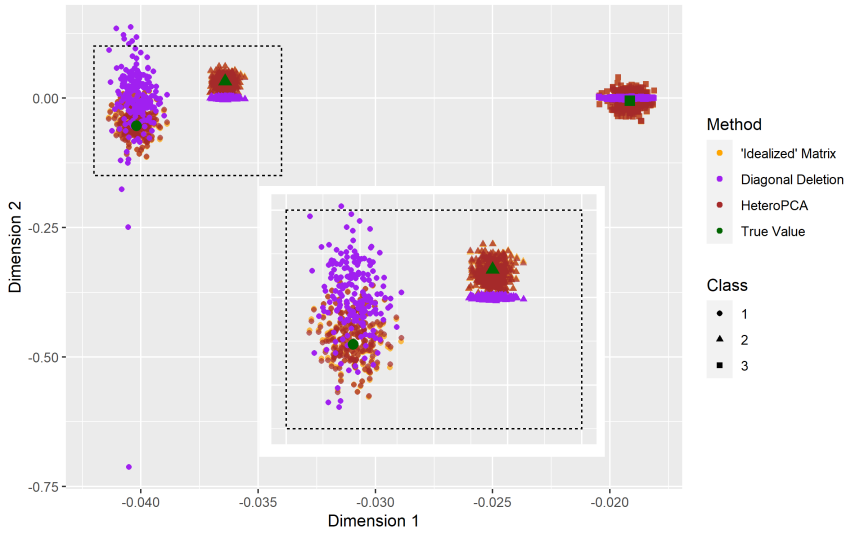


Figure 2.1: Comparison of the estimators for U given by the eigenvectors of $\Gamma(A+Z)$ (diagonal deletion) and those of $\widehat{A}$ (the output of the HeteroPCA algorithm). For convenience, we also plot the eigenvectors of the "idealized" matrix $A+\Gamma(Z)$ that the HeteroPCA algorithm is approximating as well as a zoomed-in reference for classes one and two. For details on the experimental setup, see Section 2.4.

2.3 Main Results

Before presenting our main results, we discuss the various assumptions required, and their role in the analysis. Our first assumption concerns the noise matrix E.

Assumption 2.1 (Noise). *The noise matrix E has rows $E_i^\top$ that can be written in the form $E_i = \Sigma_i^{1/2} Y_i$, where each Y_i is a vector of independent mean-zero subgaussian random variables with unit variance and ψ_2 norm uniformly bounded by 1, and Σ_i is positive semidefinite.*

The following assumption ensures that there is sufficient signal to consistently identify the eigenvectors U. We let $\kappa := \lambda_1/\lambda_r$ denote the condition number of M.

Assumption 2.2 (Enough Signal). *The signal-to-noise ratio satisfies* $\text{SNR} \geq C_{SNR}\kappa\sqrt{\log(n \vee d)}$, *for a sufficiently large constant C_{SNR}.*

We remark that in the case of independent noise, $\text{SNR} \to \infty$ is required in order for consistency (e.g. Cai and Zhang (2018); Zhang et al. (2022); Xia (2021)).

The following assumption ensures that we have sufficiently many samples for our concentration results to hold.

Assumption 2.3 (High Dimensional Regime, Low Rank)**.** *There exists a constant* c_1 *and a sufficiently small constant* c_2 *such that such that* $d \geq c_1 n$*, and* $r \leq c_2 n$*. In addition,* $\log(d) \leq n$*.*

The next assumption concerns the incoherence of the matrix M, which measures the "spikiness" of the matrix. We say M is μ_0-incoherent if

$$\max\left\{\|U\|_{2,\infty}\sqrt{\frac{n}{r}}, \|V\|_{2,\infty}\sqrt{\frac{d}{r}}\right\} \leq \mu_0.$$

When $\mu_0 = O(1)$, then the entries of U and V are spread, which corresponds to M being fully incoherent. For more details, see (Chen et al., 2021c; Chi et al., 2019).

Assumption 2.4 (Incoherence)**.** *The matrix* U *of left singular vectors of* M *satisfies* $\|U\|_{2,\infty} \leq \mu_0\sqrt{\frac{r}{n}}$*, where* μ_0 *satisfies* $\kappa^2\mu_0 \leq \sqrt{n}$*. In addition, there exists a constant* C_I *sufficiently large such that* $\|V\|_{2,\infty} \leq C_I\|U\|_{2,\infty}$*.*

We note that in much of the literature one assumes that both V and U are both μ_0-incoherent, whereas Assumption 2.4 is slightly stronger as we assume that $\|V\|_{2,\infty} \leq C_I\|U\|_{2,\infty}$. As we assume that $d \gtrsim n$, this assumption is not stringent, but rather we make this assumption for convenience as this results in a simple statement of Theorem 6 in terms of $\|U\|_{2,\infty}$. In order to apply our results to mixture distributions (see Section 2.3.2), we need that $\|\widehat{U}\mathcal{O}_* - U\|_{2,\infty} \ll \|U\|_{2,\infty}$ in order to guarantee sufficient cluster separation. Consequently, assuming both U and V are μ_0-incoherent is not quite sufficient for these purposes; instead we must additionally assume that $\|V\|_{2,\infty} \leq C_I\|U\|_{2,\infty}$.

Our final assumption concerns the relationship between the covariance of each row and the singular subspace of M. It ensures that the covariance matrices Σ_i are not too ill-conditioned on the signal subspace of M.

Assumption 2.5 (Covariance Condition Number)**.** *There exists a constant* κ_σ *such that*

$$0 < \frac{1}{\kappa_\sigma} \leq \min_{i,j} \frac{\sigma_i}{\|\Sigma_i^{1/2}V_{\cdot j}\|} \leq \max_{i,j} \frac{\sigma}{\|\Sigma_i^{1/2}V_{\cdot j}\|} \leq \kappa_\sigma < \infty.$$

Informally, Assumption 2.5 requires that Σ_i does not act "adversarially" along V in the sense that the action of Σ_i along the subspace V is well-behaved. Note that if $\Sigma_i = \sigma^2 I_d$ for

all i, then this condition is automatically satisfied with $\kappa_\sigma = 1$. Our main result (see the forthcoming Theorem 5) is stated in terms of a fixed i and j. While we make the simplifying assumption that κ_σ is uniformly bounded in both i and j, we note that if instead κ_σ is allowed to depend on i and j and satisfies

$$0 < \frac{1}{\kappa_\sigma} \leq \min_k \frac{\sigma_k}{\|\Sigma_i^{1/2} V_{\cdot j}\|} \leq \frac{\sigma}{\|\Sigma_i^{1/2} V_{\cdot j}\|} \leq \kappa_\sigma < \infty,$$

then our asymptotic normality results continue to hold with the proviso that κ_σ depends on both i and j (with appropriate modifications in the setting of Corollary 2 and 3).

We are now ready to present our main result concerning the distribution of the entrywise difference between $\widehat{U}$ and U.

Theorem 5. *Define*

$$\sigma_{ij}^2 := \|\Sigma_i^{1/2} V_{\cdot j}\|^2 \lambda_j^{-2}.$$

Suppose Assumptions 2.1, 2.2, 2.3, 2.4 and 2.5 hold. Let $\widehat{U}$ be the output of the HeteroPCA algorithm after $T = \Theta\left(\frac{\lambda_r^2}{\|U\|_{2,\infty}\|\Gamma(Z)\|}\right)$ iterations. Then there exist absolute constants C_1, C_2 and C_3 and an orthogonal matrix $\mathcal{O}_*$ such that

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{1}{\sigma_{ij}} e_i^\top \left(\widehat{U}\mathcal{O}_* - U \right) e_j \leq x \right) - \Phi(x) \right| \leq C_1 \frac{\|\Sigma_i^{1/2} V_{\cdot j}\|_3^3}{\|\Sigma_i^{1/2} V_{\cdot j}\|^3} + C_2 \kappa^3 \kappa_\sigma \mu_0 \frac{r \log(n \vee d)}{\mathrm{SNR}}$$
$$+ C_3 \kappa^2 \kappa_\sigma \mu_0 \sqrt{\frac{r}{n}} \left(\sqrt{\log(n \vee d)} + \mu_0 \kappa^2 \sqrt{r} \right).$$

One should interpret Theorem 5 as stating that the entries of $\widehat{U}$ are approximately Gaussian about their corresponding population counterparts modulo the nonidentifiability in the singular subspace stemming from the repeated singular values. One can also generalize our analysis for the *rows* of $\widehat{U}\mathcal{O}_*$ to obtain the joint distribution; see Corollary 2 for an application of this for fixed r.

Suppose $\kappa, \kappa_\sigma, \mu_0 = O(1)$. Then we see that asymptotic normality holds as long as

$$\max\left\{\frac{\|\Sigma_i^{1/2} V_{\cdot j}\|_3^3}{\|\Sigma_i^{1/2} V_{\cdot j}\|^3}, \frac{r\log(n\vee d)}{\mathrm{SNR}}, \frac{r\log(n\vee d)}{\sqrt{n}}\right\} \to 0.$$

When Σ_i is the identity and $V_{\cdot j} = \frac{1}{\sqrt{d}}\mathbf{1}$, then the first term is exactly equal to $\frac{1}{\sqrt{d}}$. Moreover, if $\mathrm{SNR} \geq \sqrt{n}$, then we obtain the classical parametric rate up to logarithmic factors. If instead $\mu_0 \log(n \vee d) \ll \min(\sqrt{n}, \mathrm{SNR})$, then asymptotic normality still holds. Therefore, in the "high-signal" regime with $\mathrm{SNR} \gg \sqrt{n}$, asymptotic normality holds as long as $\mu_0 \log(n \vee d) = o(\sqrt{n})$.

We remark that the additional logarithmic factors stem from our $\ell_{2,\infty}$ result in Theorem 6 below. It may be possible that these logarithmic factors can be eliminated with more refined analysis, but we leave this for future work, since our primary focus is on studying asymptotic normality in the presence of dependence. Furthermore, our results allow the covariances to be (strictly) positive semidefinite as long as the vector $V_{\cdot j}$ is not too close to the null space of the matrix Σ_i. A simple example is if $\Sigma_i \propto VV^\top$, then asymptotic normality still holds.

Note that Theorem 5 (and Assumption 2.5) depends on the fact that $\sigma_{ij} \neq 0$ (and, moreover, does not shrink to zero relative to the overall noise σ). If instead $\sigma_{ij} = 0$, then one must consider the higher-order asymptotics, in which case the dominant term contributing to the asymptotic normality becomes a higher order noise term, resulting in a different scaling for the asymptotic normality. While it may be possible to obtain asymptotic normality in this setting, Theorem 6 still holds regardless, thereby yielding strong concentration for $\widehat{U}$.

The following corollary specializes Theorem 5 to the case of scalar matrices Σ_i. We note that in this case, the variances of the entries of the j'th singular vector estimate are proportional to the inverse of the j'th singular value of $MM^\top$. Furthermore, the leading term in Theorem 5 simplifies to be $\|V\|_{2,\infty} \leq C_I \|U\|_{2,\infty} \lesssim \mu_0 \sqrt{\frac{r}{n}}$,which is smaller than the rightmost term by Assumption 2.4. In particular, this result shows that the variance of the i, j entry of $\widehat{U}$ is asymptotically equal to to σ_i^2/λ_j^2, which reflects the fact that the variance increases deeper into the spectrum of M.

Corollary 1. *Assume the setting of Theorem 5, with $\Sigma_i = \sigma_i^2 I_d$ for all i. Then*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\lambda_j}{\sigma_i}e_i^\top\left(\widehat{U}\mathcal{O}_* - U\right)e_j \le x\right) - \Phi(x)\right| \le C_1'\kappa^3\kappa_\sigma\mu_0\frac{r\log(n\vee d)}{\mathrm{SNR}} + C_2'\kappa^2\kappa_\sigma\mu_0\sqrt{\frac{r}{n}}\left(\sqrt{\log(n\vee d)} + \mu_0\kappa^2\sqrt{r}\right).$$

Our second theorem is an $\ell_{2,\infty}$ concentration result for the matrix $\widehat{U}$ as an estimator for the matrix U. Theorem 6 is a consequence of a deterministic bound concerning the HeteroPCA algorithm (see Theorem 8) and a bound concerning the idealized (random) perturbation $A + \Gamma(Z)$ (see Theorem 7). As part of our proof of Theorem 6, we prove additional tight concentration for several residual terms that we rely on in the proof of Theorem 5.

Theorem 6. *Suppose Assumptions 2.1, 2.2, 2.3 and 2.4 hold. Let $\widehat{U}$ be the output of the HeteroPCA algorithm after $T = \Theta\left(\frac{\lambda_r^2}{\|U\|_{2,\infty}\|\Gamma(Z)\|}\right)$ iterations. Then there exists a universal constant $C > 0$ such that with probability at least $1 - 2(n\vee d)^{-4}$*

$$\inf_{\mathcal{O}\in\mathbb{O}(r)}\|\widehat{U} - U\mathcal{O}\|_{2,\infty} \le C\left(\frac{\sqrt{rnd}\log(n\vee d)\sigma^2}{\lambda_r^2} + \frac{\sqrt{rn\log(n\vee d)}\kappa\sigma}{\lambda_r}\right)\|U\|_{2,\infty}.$$

Suppose $\kappa, \kappa_\sigma, \mu_0, r = O(1)$. The upper bound in Theorem 6 holds as long as $\mathrm{SNR} \gtrsim \sqrt{\log(n\vee d)}$, whereas the asymptotic normality in Theorem 5 holds when $\mathrm{SNR} \gg \log(n\vee d)$. It is possible that with additional work the asymptotic normality in Theorem 5 holds in the regime $\sqrt{\log(n\vee d)} \lesssim \mathrm{SNR} \lesssim \log(n\vee d)$, but this is not the focus of the present paper.

2.3.1 Comparison to Prior Work

Our Theorem 5 is the first distributional result for the entries of the singular vector estimator in the setting of dependent, heteroskedastic noise. In Xia (2021), the author derives Berry-Esseen Theorems for the $\sin\Theta$ distance between the singular vectors under the assumption that the noise consists of independent Gaussian random variables with unit variance. See also Bao et al. (2021) for a similar result in slightly different regime. In Xia and Yuan (2020), the authors use the techniques in Xia (2021) to develop Berry-Esseen Theorems for linear forms of the matrix M. They also develop $\ell_{2,\infty}$ bounds (see their Theorem 4) en route

to their main results that are similar to the bounds we obtain. While our approach and that of Xia (2021) and Xia and Yuan (2020) share a common core, our main results require additional technical considerations due to the heteroskedasticity and dependence. We also include a deterministic analysis of the HeteroPCA algorithm, which is not needed in Xia and Yuan (2020) as the entries all have the same variance.

The works Koltchinskii and Lounici (2016); Koltchinskii et al. (2020) study estimating general linear forms of the eigenvectors of a sample covariance matrix. More specifically, Theorem 7 of Koltchinskii and Lounici (2016) derive the asymptotic normality of the linear form

$$\sqrt{n}\left\langle \widehat{U}_{\cdot i} - \sqrt{1+b(n)}U_{\cdot i}, a \right\rangle,$$

where $b(n)$ is a bias term, a is a unit vector, and $\widehat{U}_{\cdot i}$ is the i'th estimated eigenvector of the sample covariance matrix. If $a = e_i$, then our results are similar in spirit to those of Koltchinskii and Lounici (2016). However, there are a few key differences:

- Koltchinskii and Lounici (2016) study covariance estimation, whereas we study the signal-plus-noise model, where the signal is assumed to be deterministic. Viewing the matrix X as the matrix whose rows are the observations, in effect Koltchinskii and Lounici (2016) study the right singular subspace of X assuming that the vectors are mean-zero. However, our analysis allows for dependence within *rows*, whereas Koltchinskii and Lounici (2016) study iid observations of sample vectors, which corresponds to dependence within *columns*.
- The results of Koltchinskii and Lounici (2016) rely heavily on the fact that the corresponding eigenvalue is simple, whereas our results allow for repeated singular values. Consequently, our results only hold up an orthogonal transformation $\mathcal{O}_*$ that accounts for the nonidentifiability of the associated singular spaces, whereas Koltchinskii and Lounici (2016) are able to directly analyze the corresponding empirical eigenvector up to a global sign flip.
- Koltchinskii and Lounici (2016) obtain asymptotic normality for a general linear form

of the eigenvectors, whereas we obtain asymptotic normality for only the individual entries. However, Koltchinskii and Lounici (2016) consider iid *Gaussian* random variables, whereas we allow general subgaussian tail conditions. Consequently, Koltchinskii and Lounici (2016) are able to make use of powerful techniques tailored specifically to Gaussian random variables, whereas our analysis uses a combination of leave-one-out arguments and conditioning.

Therefore, the results of Koltchinskii and Lounici (2016), while closely related, are not immediately comparable to our results.

In the context that M is a symmetric, square matrix and E is a matrix of independent noise along the upper triangle, Cape et al. (2019a) developed asymptotic normality results for the rows of the leading eigenvectors, which is similar in spirit to our main result in Theorem 5. Similarly, Fan et al. (2020) studied general bilinear forms of eigenvectors in this setting. Our results do not apply in these settings even if $n = d$ since we require that E_{ij} and E_{ji} are independent if $j \neq i$. On the other hand, their results cannot be applied in our setting either because of the dependence structure.

Regarding our $\ell_{2,\infty}$ concentration results, perhaps the most similar results appear in Cai et al. (2021a), in which the authors study the performance of the diagonal deletion algorithm in the presence of independent heteroskedastic noise. Our work differs from theirs in several ways:

- We allow for dependence among the rows of the noise matrix, whereas Cai et al. (2021a) requires the entries of the noise matrix to be independent random variables.
- Cai et al. (2021a) allow for missingness in the matrix M, whereas we assume that the matrix M is fully observed.
- Assumption 2.2 requires that $\sqrt{d \log(n \vee d)} \lesssim \lambda_r/\sigma$, whereas the theory in Cai et al. (2021a) covers the setting $(nd)^{1/4} \lesssim \lambda_r/\sigma$, which is a broader range than ours for the SNR regime. However, we note that our $\ell_{2,\infty}$ concentration holds under the weaker assumption $\lambda_r/\sigma \gtrsim \kappa(nd)^{1/4}\sqrt{r \log(n \vee d)}$ (which can be seen from the main proof in Section 2.6).

- Our main results include both asymptotic normality and $\ell_{2,\infty}$ concentration, whereas Cai et al. (2021a) only obtain $\ell_{2,\infty}$ concentration.
- Our $\ell_{2,\infty}$ concentration result does not incur the "diagonal deletion effect" in the upper bound of Theorem 1 in Cai et al. (2021a), since this has been accounted for using the HeteroPCA algorithm (see Theorem 8). This allows our upper bound to scale with the noise.

The most important of these differences is perhaps the last point, since eliminating the diagonal deletion effect is crucial for our asymptotic normality analysis. Besides us removing this term, our upper bound agrees with theirs up to a $\sqrt{r}$ factor. Also, since we have rid ourselves of the error coming from diagonal deletion, we are able to achieve the minimax lower bound for this problem given in Theorem 2 of Cai et al. (2021a) up to log factors when $r, \mu_0, \kappa \asymp 1$.

Shortly after posting our manuscript to ArXiv and submitting for publication, a very closely related manuscript (Yan et al., 2021) was also posted, studying a very similar setting to ours. In Yan et al. (2021), the authors study statistical inference for Heteroskedastic PCA under the spiked covariance model, where the spike component is assumed to be low rank; moreover, they also use the HeteroPCA algorithm of Zhang et al. (2022). They also obtain $\ell_{2,\infty}$ concentration and asymptotic normality, though their asymptotic normality results are not directly comparable, as they focus on statistical inference for the spike component (as opposed to the singular subspace directly). The key difference is that our asymptotic normality results allow for dependence within rows, which is a setting not covered by Yan et al. (2021). However, our $\ell_{2,\infty}$ concentration result is markedly similar to theirs (see their Theorem 10) and agrees up to factors of r and κ. In Yan et al. (2021), the authors study the regime $(nd)^{1/4} \lesssim \lambda_r/\sigma$ (ignoring logarithmic terms, factors of r, and factors of κ). While our $\ell_{2,\infty}$ concentration continues to hold in this regime, our asymptotic normality result may not hold. Since our main focus in this paper is the entrywise estimation of singular subspaces under both dependence and heteroskedasticity, we leave deriving limit theory and asymptotic normality in this regime to future work.

Finally, in Abbe et al. (2022), the authors study exact recovery in the case of the two-

component mixture model. Their results are similar to ours in that they allow for dependence and heteroskedasticity, but they do not study the limiting distribution of their diagonal deletion estimator. Moreover, their results are not directly comparable, as they use a different definition of incoherence and do not study the explicit dependence of their bound on the noise parameters and the spectral structure of the matrix M, but instead find conditions on their signal-to-noise ratio such that their upper bound tends to zero. On the other hand, their theory covers the weak-signal regime, and they extend their results to Hilbert spaces. In principle, since our results depend only on the properties of the Gram matrix, they could also be extended to a general Hilbert space, but we do not pursue such an extension here.

2.3.2 Application to Mixture Distributions

Consider the following submodel. Suppose we observe n observations of the form $X_i = M_i + E_i \in \mathbb{R}^d$, where there are K unique vectors $\mu_1, \dots, \mu_K$. Let $\widehat{M}$ be the matrix whose i'th row is $X_i^\top$. If M is rank K, by Lemma 2.1 of Lei and Rinaldo (2015), there are K unique rows of the matrix U, where each row i corresponds to the membership of the vector X_i. We then have the following Corollary to Theorem 5 in this setting.

Corollary 2. *Let $X_i = M_i + E_i$, and define the matrix*

$$S_i := \Lambda^{-1} V^\top \Sigma_i V \Lambda^{-1}. \tag{2.1}$$

Suppose $\kappa, \kappa_\sigma, \mu_0$ are bounded and r stays fixed as n and d tend to infinity, and suppose

$$\frac{\log(n \vee d)}{\text{SNR}} \to 0.$$

Then

$$(S_i)^{-1/2}\left(\widehat{U}\mathcal{O}_* - U\right)_{i\cdot} \to N(0, I_r)$$

as n and d tend to infinity with $d \geq n \geq \log(d)$.

We remark that the result above allows E_i to have an i-dependent covariance matrix. In

the setting that the covariance matrix of E_i depends only on the vector μ_k, where k is such that $M_i = \mu_k$, the following result shows that we can leverage this structure to consistently estimate the matrix S_i, which is the same within each community. We assume that one can accurately estimate the cluster memberships with probability tending to one, which holds by the signal-to-noise ratio condition, the $\ell_{2,\infty}$ bound in Theorem 6, and the setting for Corollary 2 since the eigenvector difference $\|\widehat{U} - U\mathcal{O}_*\|_{2,\infty} \ll \|U\|_{2,\infty}$, which implies that the rows of $\widehat{U}$ are asymptotically separated (see e.g. Lei and Rinaldo (2015)).

Corollary 3. *Suppose the setting for Corollary 2, and assume that there are K different communities with each community having mean μ_k and covariance matrix $\Sigma^{(k)}$ (that is, $\Sigma_i = \Sigma^{(k)}$ for all i in community k). Let n_k denote the number of observations in community k, and suppose that $n_k \asymp n$. Suppose $\bar{U}^{(k)}$ is the estimate for the centroid of the k-th mean, and let C_k denote the set of indices such that $M_i = \mu_k$. Define the estimate*

$$\widehat{S}^{(k)} := \frac{1}{n_k} \sum_{i \in C_k} \left(\widehat{U}_{i\cdot} - \bar{U}^{(k)} \right) \left(\widehat{U}_{i\cdot} - \bar{U}^{(k)} \right)^\top.$$

Then for the orthogonal matrix $\mathcal{O}_*$ appearing in Corollary 1,

$$\|(S^{(k)})^{-1} \mathcal{O}_*^\top \widehat{S}^{(k)} \mathcal{O}_* - I_r\| \to 0$$

in probability, where $S^{(k)}$ is the community-wise covariance defined in (2.1).

We note that the appearance of the orthogonal matrix $\mathcal{O}_*$ is of no inferential consequence, since Gaussianity is preserved by orthogonal transformation. This result implies that one can consistently estimate the covariance matrix for the corresponding row, which immediately implies that one can derive a pivot for the i'th row in the mixture setting described above, by setting

$$\widehat{T}_i := (\widehat{S}^{(k)})^{-1/2} (\widehat{U}_i - \bar{U}^{(k)}).$$

Corollaries 2 and 3, the Continuous Mapping Theorem, and Slutsky's Theorem imply that $\widehat{T}_i \to N(0, I_r)$ as n and d tend to infinity, which provides an asymptotically valid confidence

region. We remark that in the asymptotic regime in Corollaries 2 and 3, when r is fixed, any fixed finite collection of rows can be shown to be asymptotically independent, and hence the confidence region is simultaneously valid for any fixed set of rows; for example, for one row each of each community.

2.4 Numerical Results

We consider the following mixture model setup. We let M be the matrix whose first n_1, n_2, and n_3 rows are μ_1, μ_2, and μ_3 respectively where

$$\mu_1 := (10, 10, \dots, 10, 12, \dots, 12)^\top$$
$$\mu_2 := (10, 10, \dots, 10, 10, \dots, 10)^\top$$
$$\mu_3 := (5, 5, \dots, 5, 5.5, \dots, 5.5)^\top$$

The matrix M is readily seen to be rank 2, since $\mu_3 = .25\mu_1 + .25\mu_2$.

We compare $\widehat{U}$ to the diagonal deletion estimator in this setting in Figure 2.1 of Section 2.2, where we can clearly see the bias that comes from deleting the diagonal entries in the final approximation of the singular vectors. We consider the class-wise covariances

$$\Sigma^{(k)} := \sigma_k^2 F_k F_k^\top + I_d,$$

where $\sigma_1 = 15$, $\sigma_2 = 10$, $\sigma_3 = 7.5$, and F_k is drawn uniformly on the Stiefel manifold of dimensions $100, 50$, and 200 respectively, and $n_1 = 200$, and $n_2 = n_3 = 400$, with $d = 1000$. For the smallest class with mean μ_1 we clearly see that the reduced incoherence severely impacts the estimation with the diagonal deletion estimator, while the effect on $\widehat{U}$ is relatively small. On the latter two classes, we observe that the rows of $\widehat{U}$ are very close to those of the idealized matrix $A + \Gamma(Z)$, and the covariances are comparable between these. On the other hand, the rows of the diagonal deletion estimator do not preserve the covariance structure of this idealized matrix, since the diagonal deletion estimator is approximating the eigenvectors of $\Gamma(A+Z)$. Since $A+\Gamma(Z)$ is an unbiased perturbation of A while $\Gamma(A+Z)$ is not, we consider this to be the more natural object of comparison, and our theory supports

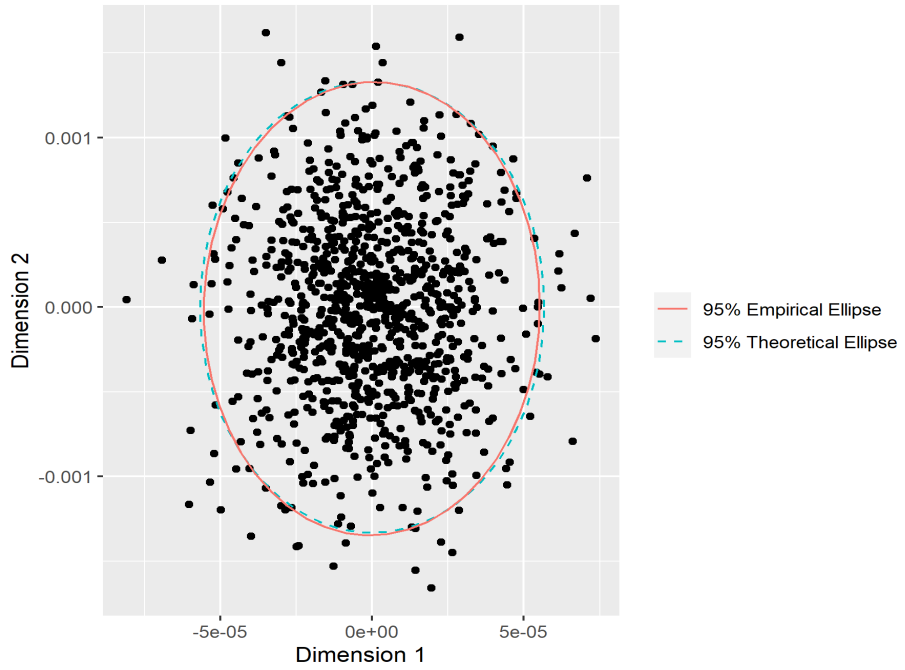


Figure 2.2: Plot of 1000 Monte Carlo iterations of the first row of $\widehat{U}\mathcal{O}_* - U$ with the same M matrix as above and $n = d = 1800$ with $n_1 = n_2 = n_3 = 600$. The covariances here are spherical within each mixture component, though they differ between components. The dotted line represents the theoretical 95 percent confidence ellipse from Theorem 5 and Corollary 2. The solid line is the estimated ellipse. The different scalings on the two axes arise because the variances in these two dimensions are proportional to the first two squared singular values of M, as seen in Corollary 1. Further details are in Section 2.4.

this view.

To study the effect of larger n and d, in Figure 2.2 we plot 1000 Monte Carlo iterations of $(\widehat{U}\mathcal{O}_* - U)_1$. where $\mathcal{O}_*$ is estimated using a Procrustes alignment between $\widehat{U}$ and U on $n = d = 1500$ points with $n_1 = n_2 = n_3 = 500$, where we use the balanced case to ensure incoherence. The solid line represents the estimated 95% confidence ellipse, and the dotted line represents the 95% confidence ellipse implied by Theorem 5. We consider the spherical noise setting, with class-wise covariances $.1I_d$, $.2I_d$, and $.3I_d$ for each component respectively. The empirical and theoretical ellipses are readily seen to be close.

2.4.1 Elliptical Versus Spherical Covariances

In Figure 2.3 we examine the effect of elliptical covariances on the limiting distribution predicted by Corollary 2. We consider the same M matrix and mixture sizes as in Figure

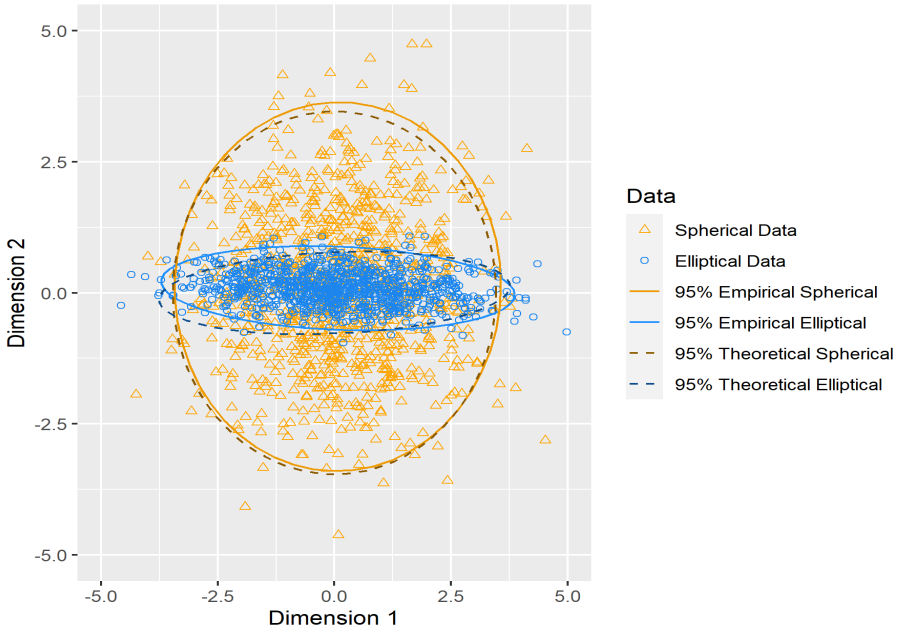


Figure 2.3: Comparison of $\Lambda(\widehat{U}\mathcal{O}_* - U)_1$. for the same mixture distribution as above, only we modify the covariance for the first mixture component between each data set. Details are in Section 2.4.1.

2.1, only we fix the covariances as

$$\Sigma_E^{(1)} := F_1 F_1^\top + .1 I_d$$
$$\Sigma^{(2)} := F_2 F_2^\top + I_d$$
$$\Sigma^{(3)} := 2 I_d,$$

where F_1 and F_2 are square matrices with entries drawn independently from uniform distributions on $[0, .003]$ and $[0, .001]$ respectively. We also consider the spherical case $\Sigma_S^{(1)} = VV^\top + I_d$. We run 1000 iterations of this simulation and examine the first row of the matrix $(\widehat{U}\mathcal{O}_* - U)\Lambda$, where again $\mathcal{O}_*$ is estimated using the procrustes difference between $\widehat{U}$ and U. The only thing we change between each dataset is the covariance; i.e., we draw $E_1 \in \mathbb{R}^{n_1 \times d}$ as a random Gaussian matrix with independent entries and multiply by $(\Sigma_E^{(1)})^{1/2}$ or $(\Sigma_S^{(1)})^{1/2}$ to obtain the first n_1 rows of the matrix E. We keep the other $n - n_1$ rows fixed within each Monte Carlo iteration, so the only randomness for each iteration is in drawing the nd Gaussian random variables. We scale $\widehat{U}\mathcal{O}_* - U$ by Λ in order to explicitly showcase the

covariance structure. The way we create $\Sigma_E^{(1)}$ yields that the limiting covariance $S_E^{(1)}$ is approximately

$$S_E^{(1)} := \Lambda^{-1} \begin{pmatrix} 2.35 & 0.083 \\ 0.083 & 0.104 \end{pmatrix} \Lambda^{-1},$$

so that $\Lambda(\widehat{U}\mathcal{O}_* - U)_{1\cdot}$ is approximately Gaussian with covariance $\begin{pmatrix} 2.35 & 0.083 \\ 0.083 & 0.104 \end{pmatrix}$. On the other hand, when we consider the spherical case, we see that $S_S^{(1)}$ is of the form

$$S_S^{(1)} = 2\Lambda^{-2}$$

so that $\Lambda(\widehat{U}\mathcal{O}_* - U)_{1\cdot}$ is approximately Gaussian with covariance $2I_r$. Figure 2.3 shows these differences, where we plot the empirical 95% confidence ellipses with respect to both the estimated covariance (solid line) and theoretical covariance (dashed line).

To study the relationship between Σ_i and V in more detail, we also consider the following setting. We consider the class-wise covariances again

$$\begin{aligned} \Sigma^{(1)} &:= 15F_1F_1^\top + .1I_d \\ \Sigma_\theta^{(2)} &:= 5V_{\cdot 1}V_{\cdot 1}^\top + 5V_2^\theta(V_2^\theta)^\top + .1I_d \\ \Sigma^{(3)} &:= 10F_3F_3^\top + .1I_d, \end{aligned}$$

where again F_1 and F_3 are drawn uniformly from the Stiefel manifold of dimension 100 and 200 respectively, with $n_1 = 200$ and $n_2 = n_3 = 400$. We also change $\mu_3 = (5, \ldots, 5, 6, \ldots 6)$ to better separate the clusters. The vector V_2^θ is orthogonal to $V_{\cdot 1}$ and satisfies $\langle V_2^\theta, V_{\cdot 2}\rangle = \theta$ for $\theta \in \{.9, .5, .1\}$. As θ decreases, the limiting covariance matrix in Corollary 2 will change along the second dimension only. Figure 2.4 reflects this theory, where we plot 1000 Monte Carlo runs of $\Lambda(\widehat{U}\mathcal{O}_* - U)_{(n_1+1)\cdot}$. The variance stays fixed along the first dimension, but it shrinks along the second dimension, showcasing the geometric relationship between V and $\Sigma^{(2)}$ as suggested by Corollary 2.

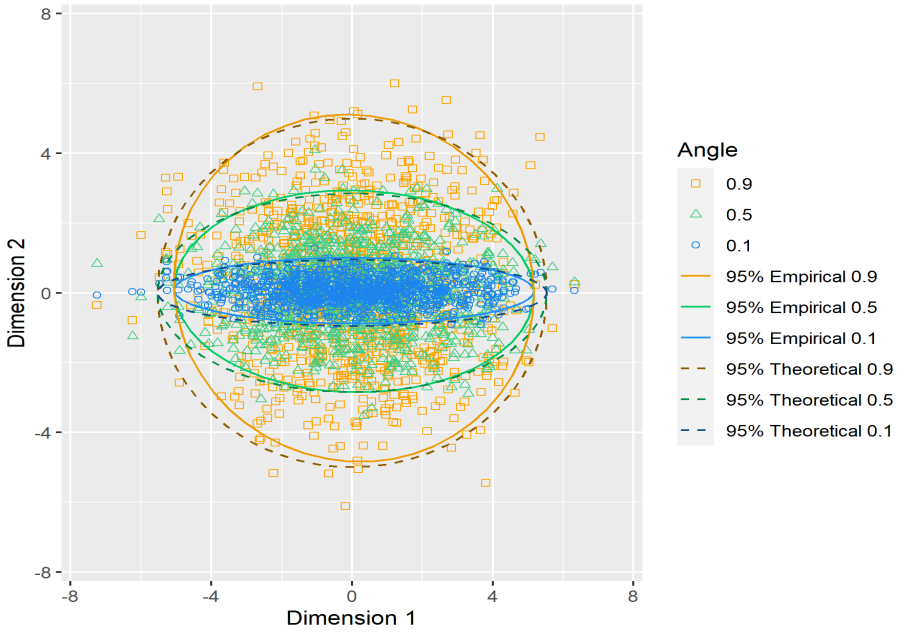


Figure 2.4: Comparison of $\Lambda(\widehat{U}\mathcal{O}_* - U)_{(n_1+1)\cdot}$ for the same mixture distribution as above, only we modify the covariance for the second mixture component between each data set by changing the angle between the leading covariance and the second mixture component. Details are in Section 2.4.1.

2.5 Discussion

We have shown that under general model assumptions for the noise matrix, allowing for heteroskedasticity between rows and dependence within them, that the entries of the left singular vectors of the output from HeteroPCA are consistent estimators for those of the original signal matrix M, and the errors are asymptotically normally-distributed in a natural high-dimensional regime. Furthermore, our Berry-Esseen theorem makes clear the rate at which this asymptotic approximation becomes valid, revealing the effect of the relationship between the noise covariances and the spectral structure of the signal matrix on the distributional convergence. In the particular case that the individual covariances are scalar multiples of the identity, our results also show that the variances of the entries of the j'th estimated singular vector are proportional to the inverse j'th singular value of the signal matrix. In particular, this means that estimating additional singular vectors in this model becomes more challenging and requires more data, since the variance in this estimation grows with j.

In this paper, we assume the rank r is known *a priori*, but in practice one may need to estimate r using, for example, the methods proposed in Zhu and Ghodsi (2006), Han et al. (2020), or Yang et al. (2020). In addition, while our results highlight the interplay of the dependence structure of the noise with the signal matrix, additional work is required to make the upper bound computable from observed data. For example, our results do not imply consistent estimation for the row-wise covariance matrices Σ_i, though we do show if one has covariances that are only distinct between clusters, then one can estimate the limiting covariance matrix S_i for each row of $\widehat{U}$. Our techniques could therefore be appropriately modified to develop two-sample asymptotically valid confidence regions or test statistics, such as in deriving a Hotelling T^2 analogue for the singular vectors as in Fan et al. (2022); Du and Tang (2022). Furthermore, one possibility for further inference would be to consider drawing several matrices $M+E$ independently from the same distribution, assuming the rows are matched together between samples, in which case one could leverage existing statistical methodology to conduct two-sample tests of hypothesis.

Another possible extension is estimating linear forms of singular vectors under the dependence structure we consider, which has been studied in other settings under independent noise. Our results naturally extend to sufficiently sparse linear forms (i.e. linear forms T such that $\|TU\| \leq b\|U\|_{2,\infty}$), but studying linear forms for which $\|TU\| \asymp 1$ would require additional methods, as the entrywise analysis methods we use would not be applicable. Finally, our results hold for a natural subgaussian mixture model, but many high-dimensional datasets contain outlier vectors or heavier tails, in which case additional techniques are required.

2.6 Proof Architecture for Theorems 5 and 6

In this section we state several intermediate lemmas, prove Theorem 6, and sketch the proof of Theorem 5. Full proofs are in the appendices. First we collect some initial spectral norm bounds that are useful in the sequel. The first is a bound on the noise error $\|\Gamma(Z)\|$, the proof of which is adapted from Theorem 2 in Amini and Razaee (2021). Throughout this section and all the proofs, we allow constants C to change from line to line.

Lemma 1 (Spectral Norm Concentration). *Under assumption 2.1, there exists a universal constant $C_{spectral}$ such that with probability at least $1 - 4(n \vee d)^{-6}$*

$$\|\Gamma(EM^\top + ME^\top + EE^\top)\| \leq C_{spectral}\bigg(\sigma^2(n + \sqrt{nd}) + \sigma\sqrt{n}\kappa\lambda_r\bigg).$$

The next bound shows that the approximation of $\widetilde{A} = A + \Gamma(Z)$ to $\widehat{A}$, the output of the HeteroPCA algorithm, is much smaller than the approximation of $\widetilde{A}$ to A. Recall that N_T denotes the approximation of A from Algorithm 1 after T iterations. We note that the existence of T_0 in the statement of this lemma follows from Zhang et al. (2022) and Assumptions 2.3 and 2.4. In particular, if we take $T_0 \geq C\log\left(\frac{\lambda_r^2}{\|\Gamma(Z)\|}\right)$, then by the proof of Theorem 7 in Zhang et al. (2022), it holds that $\|N_T - A\| \leq 3\|\Gamma(Z)\|$.

Lemma 2. *Define T_0 as the first iteration such that $\|N_T - A\| \leq 3\|\Gamma(Z)\|$. Let $\rho = 10\|\Gamma(Z)\|/\lambda_r^2$, and suppose Assumption 2.2. Define $\widetilde{K}_T := \|N_T - \widetilde{A}\|$. Then for all $T \geq T_0$ and n large enough, on the event in Lemma 1, we have that $\rho < \frac{1}{2}$, and*

$$\widetilde{K}_T \leq 4\rho^{T-T_0}\|\Gamma(Z)\| + \frac{20}{1-\rho}\|U\|_{2,\infty}\|\Gamma(Z)\|.$$

Consequently, when $T = \Theta\left(\log\left(\frac{\lambda_r^2}{\|U\|_{2,\infty}\|\Gamma(Z)\|}\right)\right)$, it holds that

$$\|\widehat{A} - \widetilde{A}\| \leq 41\|U\|_{2,\infty}\|\Gamma(Z)\|.$$

Assumption 2.2 implies that

$$\lambda_r^2 \geq C\bigg(\sigma^2 rd\log(n \vee d) + \sigma\sqrt{nr\log(n \vee d)}\kappa\lambda_r\bigg),$$

which ensures that that there is an eigengap on the event in Lemma 1. Therefore, a standard application of the Davis-Kahan Theorem (e.g. Chen et al. (2021c)) and Lemma 1 immediately implies that $\widehat{U}\widehat{U}^\top - UU^\top$ tends to zero in spectral norm.

To prove our main $\ell_{2,\infty}$ result, we analyze the statistical error and algorithmic error

separately. Define $H := U^\top \widetilde{U}$, and $\widetilde{H} := \widetilde{U}^\top \widehat{U}$. First, write $\widehat{U} - U\mathcal{O}$ as

$$\widehat{U} - U\mathcal{O} = (\widehat{U} - \widetilde{U}\widetilde{H}) + (\widetilde{U} - UH)\widetilde{H} + U(H\widetilde{H} - \mathcal{O}).$$

The first term captures the algorithmic error between the eigenvectors of the output of Algorithm 1, $\widehat{A}$, and those of the matrix it approximates, $\widetilde{A}$. The next term is the statistical error between the matrix $\widetilde{A}$ approximated by the algorithm and the true matrix of interest A. Finally, we have a correction term which accounts for the fact that $\widetilde{H}$ and H are contractions rather than orthogonal matrices.

Since the bound on the algorithmic error depends on the properties of $\widetilde{U}$, we first prove the bound on the middle term, or the statistical error between $\widetilde{U}$ and U. Then we bound the algorithmic error and finally the correction term. The following result bounds the statistical error between $\widetilde{U}$ and UH with high probability.

Theorem 7. *Under Assumptions 2.1, 2.2, 2.3 and 2.4, we have that there exists a universal constant C_R such that*

$$\|\widetilde{U} - UH\|_{2,\infty} \leq C_R\left(\frac{\sqrt{rnd}\log(\max(n,d))\sigma^2}{\lambda_r^2} + \frac{\sqrt{rn\log(n\vee d)}\kappa\sigma}{\lambda_r}\right)\|U\|_{2,\infty}$$

with probability at least $1 - (n \vee d)^{-4}$.

In order to prove this result, we use the matrix series expansion developed in Xia (2021) to write the difference of projection matrices in terms of the noise matrix, each term of which requires careful considerations due to the dependence between columns of the noise matrix E.

Consequently, by Assumption 2.2, the result in Theorem 7 implies that

$$\begin{aligned}\|\widetilde{U}\|_{2,\infty} &\leq \|\widetilde{U} - UH\|_{2,\infty} + \|U\|_{2,\infty} \\ &\lesssim \|U\|_{2,\infty},\end{aligned}$$

which shows that the matrix $\widetilde{U}$ is just as incoherent as U up to constant factors. We also have the following result for the deterministic analysis.

Theorem 8. *Define* $\widetilde{H} := \widetilde{U}^\top \widehat{U}$. *Suppose the event in Theorem 7 holds. Then, in the setting of Lemma 2, we have that there exists a universal constant* C_D *such that*

$$\|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty} \leq C_D \kappa^2 \frac{\|U\|_{2,\infty}^2 \|\Gamma(Z)\|}{\lambda_r^2}.$$

Finally, we consider the correction term.

Lemma 3. *There exists an orthogonal matrix* $\mathcal{O}_*$ *and a universal constant* C *such that under Assumptions 2.2 and 2.4, the event in Lemma 1, and* $T = \Theta\left(\frac{\lambda_r^2}{\|U\|_{2,\infty}\|\gamma(Z)\|}\right)$,

$$\|UH\widetilde{H} - U\mathcal{O}_*^\top\|_{2,\infty} \leq C\|U\|_{2,\infty}\frac{\|\Gamma(Z)\|^2}{\lambda_r^4}.$$

We now have all the pieces to prove Theorem 6.

Proof of Theorem 6. We note that all the events that determine Lemmas 2, 3, and Theorem 8 are the event in Lemma 1, and the event in Theorem 7. Taking a union bound, these events occur simultaneously with probability at least $1-(n\vee d)^{-4}-4(n\vee d)^{-6} \geq 1-2(n\vee d)^{-4}$; henceforth we operate on the intersection of these events. Since Theorem 7 gives the stated upper bound for $\|\widetilde{U} - UH\|_{2,\infty}$ on this set, by increasing the constants if necessary, we need only show that this bound holds for the algorithmic error $\|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty}$, and the correction term $\|UH\widetilde{H} - U\mathcal{O}_*\|_{2,\infty}$.

Under assumption 2.4, we have that

$$\begin{aligned}\kappa^2\|U\|_{2,\infty} &\leq \kappa^2\mu_0\sqrt{r/n} \\ &\leq \sqrt{r\log(n\vee d)}.\end{aligned}$$

Hence, the bound in Theorem 8 becomes

$$\|\widehat{U} - \widetilde{U}\widehat{H}\|_{2,\infty} \leq C_D\|U\|_{2,\infty}\sqrt{r\log(n\vee d)}\frac{\|\Gamma(Z)\|}{\lambda_r^2}. \tag{2.2}$$

In addition, on the event of Lemma 1,

$$\|\Gamma(Z)\| \leq C_{\text{spectral}}\bigg(\sigma^2(n+\sqrt{nd}) + \sigma\sqrt{n}\kappa\lambda_r\bigg).$$

This gives the desired bound for the algorithmic error. For the correction term, we note that the upper bound in Lemma 3 does not exceed that of Equation (2.2), which has already been bounded.

Combining these bounds, there is a constant $C > 0$ such that with probability at least $1 - 2(n \vee d)^{-4}$,

$$\inf_{\mathcal{O}\in\mathbb{O}(r)} \|\widehat{U} - U\mathcal{O}\|_{2,\infty} \leq C\bigg(\frac{\sqrt{rnd}\log(n\vee d)\sigma^2}{\lambda_r^2} + \frac{\sqrt{rn\log(n\vee d)}\kappa\sigma}{\lambda_r}\bigg)\|U\|_{2,\infty},$$

as advertised. □

In order to prove Theorem 5, we show that

$$e_i^\top\left(\widehat{U}\mathcal{O}_* - U\right)e_j = \langle E_i, V_{\cdot j}\rangle\lambda_j^{-1} + R, \tag{2.3}$$

where $V_{\cdot j}$ is the j'th column of V and R is a residual term that we bound using similar ideas to the proof of Theorem 6. A straightforward calculation reveals that $\mathbb{E}\bigg(\langle E_i, V_{\cdot j}\rangle\lambda_j^{-1}\bigg)^2 = V_{\cdot j}^\top\Sigma_i V_{\cdot j}\lambda_j^{-2}$, and hence if we define $\sigma_{ij} := \|\Sigma_i^{1/2}V_{\cdot j}\|\lambda_j^{-1}$, we have that

$$\begin{aligned}\frac{\langle E_i, V_{\cdot j}\rangle\lambda_j^{-1}}{\sigma_{ij}} &= \frac{\langle E_i, V_{\cdot j}\rangle}{\|\Sigma_i^{1/2}V_{\cdot j}\|} \\ &= \frac{\langle Y_i, \Sigma_i^{1/2}V_{\cdot j}\rangle}{\|\Sigma_i^{1/2}V_{\cdot j}\|},\end{aligned}$$

which by Assumption 2.1 is a sum of d independent mean-zero random variables to which the classical Berry-Esseen Theorem (Berry, 1941) can be applied. The residual term R consists of higher-order terms stemming from a matrix series expansion (see Lemma 14 in Appendix B.1). However, we have already bounded many of these residual terms as part of the proof of Theorem 7, so we need only show that dividing the residual terms by σ_{ij} yields convergence to zero. We then use the Lipschitz property of the Gaussian cumulative

distribution function to complete the proof of the theorem.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Entrywise Bounds for Sparse PCA via Sparsistent Algorithms

3.1 Introduction

Principal component analysis (PCA) is a standard statistical technique for dimensionality reduction of data in an unsupervised manner. Given i.i.d mean-zero observations $X_1, \ldots, X_n \in \mathbb{R}^p$ with covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$, the goal of PCA is to estimate the leading k-dimensional subspace of Σ, which has the interpretation of representing each observation as a linear combination of *principal components*, where each principal component is a direction of maximal variance. The classical theory of PCA (e.g. Anderson (2003)) shows that if the number of covariates p is fixed and the number of samples n tends to infinity, then the leading eigenvectors of the sample covariance approximate the leading eigenvectors of the population covariance well.

In the modern era of big data, it is often unrealistic to assume that p remains fixed in n. In the seminal work of Johnstone and Lu (2009), the authors introduced the *spiked covariance model* where the leading eigenvalue of the population covariance satisfies $\lambda_1 > 1$, while all other eigenvalues are all 1. In Johnstone and Lu (2009), the authors showed that if $\widehat{u}_1$ is the leading eigenvector of the sample covariance and u_1 is the leading eigenvector of the population covariance, then $\langle \widehat{u}_1, u_1 \rangle$ need not tend to 1 as p and n tend to infinity unless

either $p/n \to 0$ or the leading eigenvalue λ_1 tends to infinity. They then went on to show that if λ_1 remains bounded away from infinity but the leading eigenvector is *sparse* then a simple thresholding estimator could yield consistent estimation. Since then, there has been much work on generalizing the model in Johnstone and Lu (2009) to settings where either the leading eigenvalues tend to infinity (Bao et al., 2020; Cai et al., 2020, 2021a; Fan et al., 2020; Yan et al., 2021) or the leading eigenvectors are sparse (Amini and Wainwright, 2009; d'Aspremont et al., 2007; Cai et al., 2013; Gao et al., 2017; Gataric et al., 2020; Gu et al., 2014; Lei and Vu, 2015; Ma, 2013; Yang et al., 2015).

In this paper we consider the setting where the leading eigenvalues of the covariance matrix are bounded away from zero and infinity, but the leading k eigenvectors are s-sparse as n and p tend to infinity. There have been substantial theoretical (Banks et al., 2018; Cai et al., 2013; Krauthgamer et al., 2015; Vu and Lei, 2013; Wang et al., 2016) and methodological (Berthet and Rigollet, 2013; Chen and Rohe, 2020; Gataric et al., 2020; Ma, 2013; Rohe and Zeng, 2020; Xie et al., 2022) developments in sparse PCA. In Vu et al. (2013) the authors propose a semidefinite program enforcing sparsity to estimate the leading eigenvectors of the population covariance given only the sample covariance, and in Lei and Vu (2015) the authors provide general results for which the algorithm in Vu et al. (2013) selects the correct support. Similarly, Gu et al. (2014) propose a nonconvex algorithm that selects the correct support with high probability.

In many of the existing theoretical results on sparse PCA, authors are primarily concerned with subspace estimation error in spectral or Frobenius norm (e.g. Cai et al. (2013); Vu et al. (2013); Vu and Lei (2013)). However, in many situations entrywise guarantees can lead to more refined results which can be useful for downstream inference. In this paper, building upon a host of recent works on entrywise guarantees for eigenvectors (Abbe et al., 2022, 2020; Agterberg et al., 2022b; Cai et al., 2021a; Cape et al., 2019a,b; Charisopoulos et al., 2020; Chen et al., 2021c; Damle and Sun, 2020; Fan et al., 2018; Jin et al., 2019; Lei, 2019; Mao et al., 2020; Xia and Yuan, 2020; Xie et al., 2022; Xie, 2022; Yan et al., 2021), we study entrywise guarantees for sparse PCA for a very general class of models. Our main results hold for any *sparsistent* algorithm, i.e. one that selects the correct support for the eigenvectors, with high probability. Sparsistency has also been studied in other contexts in

high-dimensional statistics, such as in sparse linear models (Fan and Li, 2001; Wainwright, 2009; Zhao and Yu, 2006). See Bühlmann and van de Geer (2011) for a more comprehensive overview.

The literature on entrywise eigenvector analysis includes a suite of tools and techniques to bound the entries of eigenvectors in ways that classical matrix perturbation theory (e.g. Horn and Johnson (2012); G. W. Stewart and J.-G. Sun (1990); Bhatia (1997)) fails to address. The Davis-Kahan Theorem (Yu et al., 2014) provides a useful benchmark for eigenvector analysis, but this can lead to suboptimal entrywise bounds. The primary reason for the lack of optimality is due to the fact that the Davis-Kahan Theorem can be somewhat coarse, as it fails to take into account the probabilistic nature of empirical eigenvectors in statistical settings. Therefore, entrywise eigenvector bounds require careful probabilistic and matrix analysis techniques that go beyond what the Davis-Kahan Theorem and classical matrix perturbation theory can do. See Chen et al. (2021c) for an accessible introduction to entrywise eigenvector estimation. The only other work on entrywise eigenvector analysis in sparse PCA is in Xie et al. (2022), which is a Bayesian setting under the relatively stringent spiked model. In this paper we develop entrywise bounds for sparse PCA under a much more general model class. More specifically, our results hold for models satisfying a mild eigengap requirement (see Assumption 3.4) that includes the spiked model.

The rest of this paper is organized as follows. In Section 3.2 we provide the requisite background for sparse PCA and existing results on sparsistency. In Section 3.3 we provide our main results, and Section 3.4 includes the discussion. We include a sketch of our main proof in Section C.1, but the full proofs are relegated to the supplementary material.

3.1.1 Notation

We use capital letters to denote matrices and random vectors, which will be clear from context, and lower case letters to denote fixed vectors. We let $X_1, \dots, X_n$ denote a collection of n random variables in $\mathbb{R}^p$. For a generic real-valued random variable X, its ψ_α *Orlicz norm of order* α (or just ψ_α norm) is defined via $\|X\|_{\psi_\alpha} := \inf\{t > 0 : \mathbb{E}\exp(|X|^\alpha/t) \leq 1\}$. Random variables with finite ψ_2 norm are called *subgaussian* and random variables with finite ψ_1 norm are called *subexponential*. More discussion on Orlicz norms is included in

Appendix C.3 in the supplementary material.

For $d_1 \geq d_2$, we define the set of matrices $U \in \mathbb{R}^{d_1 \times d_2}$ with orthonormal columns as $\mathbb{O}(d_1, d_2)$ and when $d = d_1 = d_2$, we denote this set as $\mathbb{O}(d)$. We use $\|\cdot\|$ as the spectral norm on matrices and the Euclidean norm on vectors, $\|\cdot\|_F$ as the Frobenius norm, and $\|\cdot\|_{\max}$ for the maximum entry norm. Except for the spectral norm, we write $\|\cdot\|_{p\to q}$ as the operator norm from $\ell_p \to \ell_q$; that is $\|M\|_{p\to q} := \sup_{\|x\|_p=1} \|Mx\|_q$. Of particular importance is the $2 \to \infty$ norm, which is the maximum row norm of a matrix. Except for the maximum entry norm, we write $\|\cdot\|_p$ to denote the entrywise p norm of a matrix viewed as a long vector. For a matrix M, $\mathrm{diag}(M)$ extracts its diagonal, and $\mathrm{Tr}(M)$ is its trace. For two symmetric matrices A and B, we write $A \succcurlyeq B$ if $A - B$ is positive semidefinite. For a matrix M, $M_{j\cdot}$ and $M_{\cdot i}$ denote its j'th row and i'th column respectively. For a collection of indices J, M_{JJ} denotes the principal submatrix of M found by taking its columns and rows corresponding to indices in J, and for a vector x, $x[J]$ denotes the components of x corresponding to indices in J. For a matrix M, the operator $\mathrm{supp}(M)$ denotes its support, i.e. the indices corresponding to nonzero components in M. We denote the *reduced condition number* of Σ (with respect to the dimension k) as $\kappa := \frac{\lambda_1}{\lambda_k}$.

For two functions $f(n)$ and $g(n)$, we write $f(n) \lesssim g(n)$ or $f(n) = O(g(n))$ if there exists a constant C such that $f(n) \leq Cg(n)$ for all n sufficiently large, and we write $f(n) \ll g(n)$ or $f(n) = o(g(n))$ if $f(n)/g(n) \to 0$ as $n \to \infty$. In the proofs, a generic constant C may change from line to line.

3.2 Sparse PCA and Sparsistency

Suppose $\{X_i\}_{i=1}^n \in \mathbb{R}^p$ are mean-zero random variables with covariance matrix Σ and eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Define the empirical covariance $\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$, which is just the usual method of moments estimator. We assume that Σ has a *sparse* k-dimensional leading subspace, meaning that its leading k eigenvectors are s-sparse, in the sense that there is a set $J \subset \{1, \ldots, p\}$ with cardinality at most s, with each eigenvector's nonzero support restricted to indices in J. In the language of Vu and Lei (2013), this setting refers to *row*-sparsity (as opposed to *column*-sparsity). See Vu and Lei (2013) for a comparison.

We denote the $p \times k$ matrix U as the matrix of k orthonormal eigenvectors of Σ. Since U is assumed row-sparse, it has at most s nonzero *rows*. Concretely, this means that the nonzero support of each column of U is restricted to rows with indices in J. A useful interpretation of the set J is that it corresponds to the subset of covariates that contribute to the directions of maximum variance. In order for Σ to have a well-defined (sparse) leading k-dimensional subspace, it must have an eigengap, meaning that $\lambda_k - \lambda_{k+1} > 0$. In Section 3.3, Assumption 3.4 offers a slightly more quantitative condition on this eigengap.

The *sparse PCA problem* consists of estimating the matrix U from the observations $\{X_i\}_{i=1}^n$. There have been a number of approaches, including, but not limited to semidefinite programming Amini and Wainwright (2009); d'Aspremont et al. (2007), Fantope Projection and Selection algorithm (Vu et al., 2013; Lei and Vu, 2015), nonconvex approaches (Gu et al., 2014), Bayesian approaches (Xie et al., 2022), amongst others (Gataric et al., 2020; Chen and Rohe, 2020; Wang et al., 2014; Ma, 2013). In this paper we consider any algorithm that selects the correct support with high probability (see Assumption 3.2) in an asymptotic regime where $k \ll s \ll n \lesssim p$. From a practical standpoint, it is useful to consider the regime where k stays fixed but s tends to infinity as n and p at a rate $s = o(n)$. This regime is similar to that studied in the literature on high-dimensional sparse linear models, where one assumes that the coefficients are s-sparse with $s \ll n$. While it is possible to use analogous techniques to those in sparse linear models to study sparse PCA (e.g. Janková and van de Geer (2021)), the unsupervised problem of sparse PCA is markedly distinct from the *supervised* setting of sparse linear regression, and often requires additional considerations.

Note that if Π is a permutation matrix, then $\Pi\Sigma\Pi^\top(\Pi U) = \Pi\Sigma U = \Pi U\Lambda$, where Λ is the $k \times k$ diagonal matrix of leading eigenvalues of Σ. This shows that ΠU are eigenvectors of $\Pi\Sigma\Pi^\top$. Therefore, given the set of nonzero indices J, without loss of generality, we can assume $J = \{1, \dots, s\}$ by permuting Σ if necessary. We can partition Σ via

$$\Sigma := \begin{pmatrix} \Sigma_{JJ} & \Sigma_{JJ^c} \\ \Sigma_{JJ^c}^\top & \Sigma_{J^cJ^c} \end{pmatrix};$$

a similar partition holds for $\widehat{\Sigma}$ and U. Under the assumption that the leading eigenvectors

Algorithm 2 "Debiased" Sparse PCA

Require: Sparsistent sparse PCA algorithm `SparsePCA`, empirical covariance matrix $\widehat{\Sigma}$

1: Run `SparsePCA` algorithm on $\widehat{\Sigma}$, obtaining support set estimate $\widehat{J} \subset \{1, \dots, p\}$.
2: Define $\widetilde{U}_{\widehat{J}}$ as the leading k eigenvectors of $\widehat{\Sigma}_{\widehat{J}\widehat{J}}$.
3: **return** Full matrix $\widetilde{U}$, where

$$\widetilde{U}_{i\cdot} = \begin{cases} (\widetilde{U}_{\widehat{J}})_{i\cdot} & i \in \widehat{J} \\ 0 & i \notin \widehat{J} \end{cases}$$

of Σ are sparse, we have from the eigenvector equation that

$$\Sigma U = \begin{pmatrix} \Sigma_{JJ} & \Sigma_{JJ^c} \\ \Sigma_{JJ^c}^\top & \Sigma_{J^cJ^c} \end{pmatrix} \begin{pmatrix} U_J \\ 0 \end{pmatrix} = \begin{pmatrix} \Sigma_{JJ} U_J \\ \Sigma_{JJ^c}^\top U_J \end{pmatrix} = \begin{pmatrix} U_J \\ 0 \end{pmatrix} \Lambda$$

which shows also that U_J is orthogonal to the matrix $\Sigma_{JJ^c}^\top$ and that the leading k eigenvectors and eigenvalues of Σ_{JJ} are exactly the leading k eigenvectors of Σ with the zeros removed.

An important property of any sparse PCA algorithm is identifying the support J with high probability. Suppose $\widehat{U}$ is any estimator for U (or, equivalently, $\widehat{UU^\top}$ is any estimator for $UU^\top$). In this work we consider a "debiased" version of sparse PCA under the assumption that $\widehat{U}$ and U contain the same set of nonzero components, which implies that the estimator $\widehat{U}$ equivalently estimates the support J. We defer the particular details of this assumption to Assumption 3.2. Our estimator is then defined as the following modification on any sparsistent algorithm: given any set J, let $\widetilde{U}_J$ be the $s \times k$ matrix of eigenvectors of the principal submatrix $\widehat{\Sigma}_{JJ}$, and define $\widetilde{U} := \begin{pmatrix} \widetilde{U}_J \\ 0 \end{pmatrix}$. If the algorithm is sparsistent, then the correct set J will be selected with high probability. In this way, the particular choice of sparse PCA algorithm can be viewed as a variable selection procedure as opposed to an estimation procedure. The full procedure is presented in Algorithm 2.

A natural question is whether sparsistent algorithms for sparse PCA exist. The answer is positive: in Theorem 1 of Lei and Vu (2015), the authors provide deterministic conditions on Σ guaranteeing that the Fantope Projection and Selection estimator is unique and has support set J with probability at least $1 - O(p^{-2})$ when $s\sqrt{\frac{\log(p)}{n}} \to 0$. Their conditions require an error bound on $\|\widehat{\Sigma} - \Sigma\|_{\max}$ as well as conditions on the magnitudes of the eigengaps

and entries of the projection matrices. Similarly, Gu et al. (2014) provide general conditions on Σ (in terms of the magnitudes of the entries) so that their (nonconvex) algorithm obtains the support set J with probability at least $1 - O(n^{-2})$ when $\frac{sk\log(p)}{n} \to 0$. In general, sparsistency is a property of an algorithm, and the particular structure of Σ must be taken into account. Therefore, our results will hold for general matrices Σ with only mild conditions, and can be coupled with additional structural assumptions and algorithms to yield improved recovery guarantees.

3.3 Main Results

In order to state our main results, we need a few assumptions. Our main results will be stated for large n with p, s and k functions of n. We have the following assumption on the dimensions.

Assumption 3.1 (Sample Size and Dimension)**.** *The sample size n and dimension p satisfy*

$$s\log(p) \ll n; \qquad k \ll s.$$

The assumption that $s\log(p) \ll n$ is weaker than the assumption $s \lesssim \sqrt{n/\log(p)}$ as is the condition in Lei and Vu (2015) for sparsistency. However, this still allows $p/n \to \infty$; e.g. $p = n^c$ for any $c \geq 1$. The second condition $k \ll s$ is not explicitly required, but it does rule out the degenerate case $k = O(s)$, since $k \leq s$ by definition. In many works $k = 1$ (e.g. Amini and Wainwright (2009); Elsener and van de Geer (2019); Janková and van de Geer (2021)).

The next assumption imposes the condition that whatever variable selection procedure we use selects the correct support set J with high probability.

Assumption 3.2 (Sparsistency)**.** *The algorithm is* sparsistent*, meaning that with probability $1 - \delta$ the correct set J is chosen.*

Note that Theorem 1 of Lei and Vu (2015) provides sufficient conditions for Assumption 3.2 to hold, as does Theorem 1 of Gu et al. (2014). In general, this assumption is the hardest to check as it depends on the particular variable selection algorithm. In Lei and Vu (2015),

the authors show that $\delta = O(p^{-2})$ when $s\sqrt{\frac{\log(p)}{n}} \to 0$ (in addition to some other conditions omitted here). Similarly, Gu et al. (2014) show that $\delta = O(n^{-2})$ when $\frac{s\log(p)}{n} \to 0$ (in addition to other conditions omitted here). Typically the other conditions include some "signal-strength" requirements, such as the magnitudes of the entries of Σ being sufficiently large. The particular details for these requirements can be found in Lei and Vu (2015) and Gu et al. (2014) respectively.

The following assumption imposes general tail conditions on the distribution of the observations $X_1, \ldots, X_n$.

Assumption 3.3 (Randomness)**.** *The variables X_i are mean zero and satisfy $X_i = \Sigma^{1/2} Y_i$ for independent random variables Y_i with independent coordinates with unit variance. Furthermore, the ψ_2 norm of each coordinate Y_{ij} satisfies $\|Y_{ij}\|_{\psi_2} = 1$.*

This assumption says that the X_i's are linear combinations of Y_i's whose entries are independent. In general, assuming that each observation is a linear combination of independent random variables is a little stringent, but still common in the random matrix theory literature (e.g. El Karoui (2010); Knowles and Yin (2017); Bao et al. (2020); Ding and Yang (2021); Yang (2019, 2020)). While a more general result may be possible, Assumption 3.3 includes the setting that the Y_i's are i.i.d. Gaussians with identity covariance.

The following assumption imposes a quantitative condition on the eigengap (note that the existence of an eigengap is required for identifiability).

Assumption 3.4 (Eigenvalues)**.** *The top eigenvalues of Σ satisfy*

$$C\lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}}\right) + \frac{\lambda_{k+1}}{8} \leq \frac{\lambda_k}{8}$$

for some sufficiently large constant C. In addition, for all p, we have that $2\lambda_{k+1} < (1-\varepsilon)\lambda_k$ for some $\varepsilon > \frac{1}{64}$.

The requirement $\varepsilon > \frac{1}{64}$ is somewhat arbitrary and can be relaxed in general to any constant strictly greater than zero. The other part of the assumption is required to obtain enough signal on the top k eigenvalues of Σ, and hence Σ_{JJ}. Furthermore, in light of Lemma 4 (our principal submatrix concentration bound), this ensures that the top k eigenvalues of

$\widehat{\Sigma}_{JJ}$ "track" those of Σ_{JJ}. In lieu of stronger assumptions, such as in a spiked model, this is the minimum requirement to guarantee that leading eigenvectors of $\widehat{\Sigma}_{JJ}$ are well-defined.

The main results will be stated in terms of the $2 \to \infty$ norm of the difference of two matrices. Recall that for a matrix $M \in \mathbb{R}^{p \times k}$, we have that

$$\|M\|_{2\to\infty} = \max_{1\le i\le p} \|M_{i\cdot}\|_2;$$

that is, $\|M\|_{2\to\infty}$ is the maximum (Euclidean) row norm of the matrix M. Moreover, the $2 \to \infty$ norm has some attractive geometrical properties; for example, for two matrices A and B, we have that $\|AB\|_{2\to\infty} \le \|A\|_{2\to\infty}\|B\|$. More discussion on these relationships can be found in Cape et al. (2019b).

The following assumption concerns the *incoherence* of the matrix U, which is defined as $\|U\|_{2\to\infty}$. This assumption is only included to ease interpretation and is not explicitly required. A more general – albeit more complicated – result is provided in the supplementary material.

Assumption 3.5 (Incoherence and Conditioning)**.** *Suppose* $\|U\|_{2\to\infty} \lesssim \left(\frac{k}{s}\right)^{1/2}$*, that* $k \lesssim \sqrt{s}$*, and that the eigenvalues satisfy*

$$\lambda_{k+1} \le \frac{\lambda}{2} < \lambda \le \lambda_k \le \lambda_1 \le \kappa\lambda$$

for some parameters κ *and* λ*.*

The requirement $k \lesssim \sqrt{s}$ is only needed to simplify terms. The incoherence assumption states that the matrix Σ_{JJ} is incoherent in the usual sense. In this paper we do not worry about the particular incoherence constant as long as it is $O(1)$, whereas in the matrix completion literature (Candes and Plan, 2010; Candes and Tao, 2010; Chen et al., 2020, 2019b) one often studies the precise dependence on the incoherence constant. If one desires a more refined understanding of incoherence, our more general result in the supplementary material shows how our upper bound depends explicitly on the incoherence of U.

In addition, Assumption 3.5 should not be confused with Assumption 3.4 on the eigengap. The parameter κ is the *reduced condition number* of the leading k-dimensional subspace

of Σ, and can be much smaller than the usual (full) condition number of Σ, especially when the leading k eigenvalues are of comparable order (or "spiked") relative to the bottom $p-k$ eigenvalues. Assumption 3.4 in fact implies an upper bound on κ of order at most $\sqrt{n/(s\log(p))}$, but it is useful to think of the setting that $\kappa = O(1)$, which corresponds to the case where the leading k eigenvalues are of comparable order. In the setting that the eigenvalues are uniformly bounded away from zero and infinity, this assumption is not particularly strong; moreover, if the leading k eigenvalues grow sufficiently fast as a function of n and p, then the leading k eigenvectors are consistent without additional assumptions. Consequently, the primary technical condition in Assumption 3.5 is on the incoherence, i.e. $\|U\|_{2\to\infty} \lesssim \left(\frac{k}{s}\right)^{1/2}$.

Before stating the main theorem, we will require some notions from subspace perturbation theory (Bhatia, 1997; G. W. Stewart and J.-G. Sun, 1990). For $V, V' \in \mathbb{O}(p,k)$, the quantity

$$d_F(V,V') = \inf_{W\in\mathbb{O}(k)} \|V - V'W\|_F \tag{3.1}$$

defines a metric on k-dimensional subspaces invariant to choice of basis. Therefore, by analogy, one might wish to study the quantity

$$d_{2\to\infty}(V,V') := \inf_{W\in\mathbb{O}(k)} \|V - V'W\|_{2\to\infty}. \tag{3.2}$$

Unfortunately, for fixed V, V', one cannot necessarily compute the minimizer in (3.2) in closed form. However, for fixed V, V' the minimizer of (3.1) is attained using the singular value decomposition of $V^\top V'$. That is, let $W_1 D W_2^\top$ be the singular value decomposition of $V^\top V'$. Then the minimizer of (3.1), denoted W_*, satisfes $W_* := W_1 W_2^\top$. In addition,

$$d_{2\to\infty}(V,V') \le \|V - V'W_*\|_{2\to\infty}.$$

Therefore, the results will be stated in terms of the *existence* of an orthogonal matrix $W_* \in \mathbb{O}(k)$ that provides an upper bound for the $2\to\infty$ distance. In the proof, we show that W_* is actually a specific Frobenius-optimal orthogonal matrix. For convenience, we also

include more information on subspace distances in the supplementary material (Appendix C.3).

We are now prepared to state our main result.

Theorem 9. *Suppose Assumptions 3.1, 3.2, 3.3, 3.4, and 3.5 are satisfied, and let $\widetilde{U}$ be the output of Algorithm 2. Then with probability at least $1 - \delta - p^{-2}$, there exists an orthogonal matrix $W_* \in \mathbb{O}(k)$ such that*

$$\max_{1 \le i \le n} \|\widetilde{U}_{i\cdot} - (UW_*)_{i\cdot}\| \lesssim \kappa^2 \sqrt{\frac{k \log(p)}{n}} + \kappa^3 \frac{s \log(p)}{n}.$$

Consequently, if $\kappa = O(1)$, then

$$\max_{1 \le i \le n} \|\widetilde{U}_{i\cdot} - (UW_*)_{i\cdot}\| \lesssim \sqrt{\frac{k \log(p)}{n}} + \frac{s \log(p)}{n}.$$

As a brief remark, the dependence on the reduced condition number κ here may be suboptimal and could potentially be improved – we believe this is primarily an artifact of our proof technique and not a fundamental requirement. Recall that in the regime that the eigenvalues are bounded away from zero and infinity, when the leading k eigenvalues are of comparable order, it holds that $\kappa = O(1)$.

Note that by taking $\delta = O(p^{-2})$ and the conditions in Lei and Vu (2015) needed for sparsistency, the above bound holds with probability at least $1 - O(p^{-2})$; similarly, under the conditions needed for sparsistency in Gu et al. (2014), one has $\delta = O(n^{-2})$, in which case the bound holds with probability at least $1 - O(n^{-2})$.

3.4 Discussion

In the regime that the eigenvalues are uniformly bounded away from zero and infinity in n, then Theorem 9 shows that we have the error rate

$$\max_{1 \le i \le n} \|\widetilde{U}_{i\cdot} - (UW_*)_{i\cdot}\| \lesssim \max\left(\sqrt{\frac{k \log(p)}{n}}, \frac{s \log(p)}{n}\right).$$

In contrast, under the same conditions, in Frobenius norm, it has been shown in Cai et al. (2013) that the minimax rate satisfies

$$\|\widetilde{U} - UW_*\|_F \lesssim \sqrt{\frac{s\log(p)}{n}},$$

so Theorem 9 improves upon this. Moreover, our result improves greatly upon the Frobenius norm bound in Vu et al. (2013), as well as the Frobenius minimax rates studied in Cai et al. (2013) and Vu and Lei (2013). To the best of our knowledge, this is the first $2 \to \infty$ guarantee for sparse PCA under a generic sparsistency requirement. A similar result was found in Xie et al. (2022) for spiked sparse covariance matrices, but here the only assumption on the spike is Assumption 3.4, which is a much weaker assumption.

Our bounds can also be compared to the spiked covariance matrix setting $\Sigma = U\Lambda U^\top + \sigma^2 I$, where U is no longer sparse but $\lambda_k \to \infty$ in n and p. In this setting the eigenvectors $\widehat{U}$ of $\widehat{\Sigma}$ are consistent in the following sense. Define the *effective rank* $\mathfrak{r}(\Sigma) := \frac{\mathrm{Tr}(\Sigma)}{\lambda_1}$. Theorem 1 of Cape et al. (2019b) (see also Yan et al. (2021) and Cai et al. (2021a)) shows that if $\lambda_1 \gtrsim d/k$, $\mathfrak{r}(\Sigma) = o(n)$, $\kappa = O(1)$, and $\lambda_k - \sigma^2 \gtrsim \lambda_k$, then

$$\max_{1\le i\le n} \|\widehat{U}_{i\cdot} - (UW_*)_{i\cdot}\| \lesssim \sqrt{\frac{\max\{\mathfrak{r}(\Sigma), \log(d)\}}{n}}\sqrt{\frac{k^3}{p}}.$$

Here the primary error is no longer in *detecting* the leading eigenvectors (as the assumption that $\lambda_1 \gtrsim d/k$ implies large enough separation), but rather in the inherent statistical error implicit from the difference $\widehat{\Sigma} - \Sigma$. Our upper bound requires that J is either known or can be estimated consistently (Assumption 3.2), so that our error depends on the inherent statistical error from $\widehat{\Sigma}_{JJ} - \Sigma_{JJ}$. In contrast, we do not optimize for factors of λ_1 in our upper bound, as the setting for sparse PCA typically assumes that the eigenvalues remain bounded in n and p. We instead need only the (milder) eigenvalue separation in Assumption 3.4.

Suppose instead of just observing $X_1, \ldots, X_n \in \mathbb{R}^p$, one also observes response variables $Y_i \in \mathbb{R}$. Consider the linear model $Y_i = X_i^\top \beta + \varepsilon_i$, where ε_i is a mean-zero error term with variance σ^2. Suppose one first performs unsupervised dimensionality reduction on the data

matrix via sparse PCA and then computes $\widehat{\beta}$ using ordinary least squares with the reduced data matrix. The $2 \to \infty$ bound in Theorem 9 could provide a partial answer to the out-of-sample prediction performance using a variable selection procedure. To be concrete, define $\widehat{\beta}$ as the output of ordinary least squares by regressing Y_i along $X\widetilde{U}\widetilde{U}^\top$, where $\widetilde{U}$ is the output of the sparse PCA procedure in Algorithm 2 and X is the $n \times p$ matrix of predictors. Following Huang et al. (2020a), we can bound the risk of a new sample point (x_*, Y_*) via

$$\mathbb{E}\|Y_* - x_*^\top \widetilde{\beta}_{\mathrm{SPCA}}\|^2 | X \leq \beta^\top (I - \widetilde{U}\widetilde{U}^\top)\Sigma(I - \widetilde{U}\widetilde{U}^\top)\beta + \frac{\sigma^2}{n}\mathrm{Tr}\left[\left(\frac{1}{n}\widetilde{U}\widetilde{U}^\top X^\top X \widetilde{U}\widetilde{U}^\top\right)^\dagger \Sigma\right] + \sigma^2,$$

where the first term represents the bias, the second term represents the variance, and the third term (σ^2) is the noise intrinsic to the problem. The bias term can be expanded further via

$$\begin{aligned}\beta^\top (I - \widetilde{U}\widetilde{U}^\top)\Sigma(I - \widetilde{U}\widetilde{U}^\top)\beta = \beta^\top (\widetilde{U}\widetilde{U}^\top - UU^\top)\Sigma(\widetilde{U}\widetilde{U}^\top - UU^\top)\beta \\ + 2\beta^\top (\widetilde{U}\widetilde{U}^\top - UU^\top)\Sigma(I - UU^\top)\beta \\ + \lambda_{k+1}\|\beta\|_2^2.\end{aligned}$$

Consider the second term. This could be bounded by noting that

$$\begin{aligned}|\beta^\top (\widetilde{U}\widetilde{U}^\top - UU^\top)\Sigma(\widetilde{U}\widetilde{U}^\top - UU^\top)\beta| &\leq \|\beta^\top (\widetilde{U}\widetilde{U}^\top - UU^\top)\|_\infty \|\Sigma(I - UU^\top)\beta\|_1 \\ &\leq \lambda_{k+1}\|\beta\|_\infty \|\beta\|_1 \|\widetilde{U}\widetilde{U}^\top - UU^\top\|_{2\to\infty}.\end{aligned}$$

This bound has a factor of $\|\widetilde{U}\widetilde{U}^\top - UU^\top\|_{2\to\infty}$, which, while not exactly the same as what appears in Theorem 9, is closely related to it by appealing to notions in subspace perturbation theory (see, e.g. Lemma 1 of Cai and Zhang (2018)). Therefore, through similar analysis, one could obtain bounds for the other bias and variance terms with respect to the eigenvalues of Σ, the quantity $\|\widetilde{U}\widetilde{U}^\top - UU^\top\|_{2\to\infty}$ and the quantities $\|\beta\|_1$ and $\|\beta\|_\infty$. Consequently, these bounds would complement those in Theorem 1 of Huang et al. (2020a) as sparse PCA is typically needed in a regime when $\mathfrak{r}(\Sigma) \gtrsim n$, whereas Huang et al. (2020a) study the setting that $\mathfrak{r}(\Sigma) = o(n)$.

Finally, our upper bound depends on the debiased estimator $\widetilde{U}_J$, which is the matrix of eigenvectors of $\widehat{\Sigma}_{JJ}$. A key requirement is that any algorithm obtains the correct set J with probability at least $1-\delta$. In general, one must consider the output of an optimization procedure to determine whether a specific algorithm obtains the correct set J. If one additionally wanted to *test* whether a certain row of U is equal to zero (i.e., whether $i \in J$), then one would need to construct a different debiased estimator as in Janková and van de Geer (2021) that uses the first-order necessary optimality conditions. This procedure therefore relies heavily on the particular algorithm used, whereas our bounds hold for generic algorithms.

3.5 Overview of the Proof of Theorem 9

The full proof of Theorem 9 is in the supplementary material, though we include a brief overview here. First, our main upper bound holds without Assumption 3.5, and we provide this general upper bound in Theorem 22 (stated in the supplementary material C.1) and show how Theorem 9 can be deduced from Assumption 3.5. To prove Theorem 22, we first show the following *principal submatrix concentration* bound.

Lemma 4 (Principal Submatrix Concentration)**.** *Let J be an index set of $\{1, ..., p\}$ of size s. Then*

$$\|\widehat{\Sigma}_{JJ} - \Sigma_{JJ}\| \lesssim \lambda_1 \left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \right)$$

with probability at least $1 - O(p^{-4})$.

The proof is somewhat standard and primarily follows arguments detailed in Wainwright (2019) via ε-nets and concentration, though we include it in Section C.2.1 for completeness. It is also very similar to a result in Amini and Wainwright (2009) for Gaussian random variables. To the best of our knowledge, there is no general result of this form in the literature for subgaussian random vectors. The following Lemma shows that the leading k eigenvalues of $\widehat{\Sigma}_{JJ}$ are well-separated from its bottom eigenvalues.

Lemma 5 (Existence of an Eigengap). *Under the event in Lemma 4 and Assumption 3.4, the eigenvalues of $\widehat{\Sigma}_{JJ}$ and Σ_{JJ} satisfy*

$$\lambda_k - \widetilde{\lambda}_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8}; \qquad \widetilde{\lambda}_k - \lambda_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8};$$
$$\widetilde{\lambda}_k \geq \frac{\lambda_k}{4}.$$

Consequently, this bound holds with probability at least $1 - O(p^{-4})$.

We also note that $\lambda_{k+1}(\Sigma_{JJ}) \leq \lambda_{k+1}$ by the Cauchy interlacing inequalities (Horn and Johnson, 2012), and the top k eigenvalues of Σ_{JJ} are the same as those of Σ by the eigenvector equation. These lemmas set the stage for our main analysis.

As an immediate consequence of Lemmas 4 and 5, we can obtain the following proposition concerning the spectral proximity of $U_J U_J^\top$ to $\widetilde{U}_J \widetilde{U}_J^\top$, ensuring that $\widetilde{U}_J$ (and hence $\widetilde{U}$) is well-defined.

Proposition 1 (Spectral Proximity). *Under the assumptions of Theorem 22, we have that*

$$\|U_J U_J^\top - \widetilde{U}_J \widetilde{U}_J^\top\| \lesssim \frac{\lambda_1}{\lambda_k - \lambda_{k+1}} \left[\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \right]$$

with probability at least $1 - O(p^{-4})$.

We use this bound several times in our subsequent analysis. After these preliminary bounds, which are restated for convenience in the supplementary material, we develop an expansion for the difference $\widetilde{U}_J - U_J W_*$ in terms of the error matrix $(\Sigma - \widehat{\Sigma})$ and deterministic quantities depending only on Σ. Informally, we show that we have the "first-order" approximation

$$\widetilde{U}_J - U_J W_* = (\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ}) \widetilde{U}_J \widetilde{\Lambda}^{-1} + R,$$

where R is a residual term and $\widetilde{\Lambda}$ is the diagonal matrix of the k leading eigenvalues of $\widehat{\Sigma}_{JJ}$. Lemma 5 ensures that the eigenvalues of $\widetilde{\Lambda}$ can be bounded with respect to the eigenvalues of Σ. The residual term R (the terms T_1, T_2, and T_3 in the supplementary material) is bounded in Lemmas 24, 25, and 26 with tools from complex analysis (Greene and Krantz, 2006),

matrix perturbation theory (Bhatia, 1997), and high-dimensional probability (Wainwright, 2019; Vershynin, 2018).

To bound the leading term in $2 \to \infty$ norm, we show that it can be further decomposed into two terms, that we dub J_1 and J_2, by the decomposition

$$\begin{aligned}(\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\widetilde{U}_J &= (\widehat{\Sigma}_{JJ} - \Sigma_{JJ})\widetilde{U}_J + U_\perp \Lambda_\perp U_\perp^\top \widetilde{U}_J \\ &:= J_1 + J_2,\end{aligned}$$

where $U_\perp$ is the $s \times (s-k)$ matrix such that $[U_J, U_\perp]$ is an orthogonal matrix. The first term reflects the error from the randomness and the leading subspace U_J and the second term reflects the influence of $U_\perp$ on $\widetilde{U}_J$.

The term $J_2 = U_\perp \Lambda_\perp U_\perp^\top \widetilde{U}_J$ is bounded using a matrix series expansion for the matrix $\widetilde{U}_J$ (Lemma 27). More explicitly, we define the perturbation $E := \widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top$, and we show that we can write

$$\widetilde{U}_J = \sum_{m=0}^{\infty} E^m (U_J \Lambda U_J^\top)\widetilde{U}_J \Lambda^{-m+1}.$$

We then analyze each term in $2 \to \infty$ norm, take a union bound for the first $O(\log(n))$ terms and bound the remaining part of the series coarsely using the spectral norm. Similar techniques have been used in Cape et al. (2019a); Xie et al. (2022); Tang (2018) and Tang et al. (2017c), but our analysis requires additional considerations due to the fact that we do not have a mean-zero perturbation since $\mathbb{E}E = U_\perp U_\perp^\top \Sigma_{JJ} U_\perp U_\perp^\top$. However, the matrix EU_J *is* mean-zero since $U_\perp^\top U_J = 0$.

The remaining term $J_1 = (\widehat{\Sigma}_{JJ} - \Sigma_{JJ})\widetilde{U}_J$ is then analyzed directly through its block-structure (Equation (C.5)). Letting X be the $n \times p$ matrix whose rows are the observations, by Assumption 3.3, $X = Y\Sigma^{1/2}$, where Y is an $n \times p$ matrix of independent random variables

with unit variance. Then the empirical covariance $\widehat{\Sigma} = \frac{1}{n}X^\top X$ and hence

$$\widehat{\Sigma}_{JJ} = \frac{1}{n}\Big((\Sigma^{1/2})_{JJ} Y_J^\top Y_J (\Sigma^{1/2})_{JJ} + \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} + (\Sigma^{1/2})_{JJ} Y_J^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top + \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top \Big)$$

where we have abused the notation

$$\Sigma^{1/2}_{JJ^c} = (\Sigma^{1/2})_{JJ^c}.$$

Above, the $n \times p$ matrix Y is partitioned via $Y = [Y_J, Y_{J^c}]$, where Y_J and Y_{J^c} are the variables corresponding to J and its complement, J^c, respectively. This term is bounded in Lemmas 28, 29, and 30. Lemmas 28 and 29 are standard applications of matrix perturbation theory (via Proposition 1) and standard concentration inequalities such as Bernstein's inequality, but Lemma 30 requires studying the spectral properties of the matrix Σ_{JJ^c} and its relation to U_J (Proposition 10).

Our proof is then completed by combining and aggregating all of these bounds. Throughout the proof we make heavy use of several important concentration inequalities and notions from subspace perturbation theory, so Appendix C.3 in the supplementary material contains additional information on these topics.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Estimating Higher-Order Mixed Memberships via $\ell_{2,\infty}$ Tensor Perturbation Bounds

4.1 Introduction

Higher-order multiway data, i.e., tensor data, is ubiquitous in modern machine learning and statistics, and there is a need to develop new methodologies for these types of data that succinctly capture the underlying structures. In a variety of scenarios, tensor data may exhibit community-like structures, where each component (node) along each different mode is associated with a certain community/multiple communities. High-order clustering aims to partition each mode of a dataset in the form of a tensor into several discrete groups. In many settings, the assumption that groups are discrete, or that each node belongs to only one group, can be restrictive, particularly if there is a domain-specific reason that groups need not be distinct. For example, in the multilayer flight data we consider in Section 4.3.3, one observes flights between airports over time. Imposing the assumption that the underlying tensor has discrete communities assumes that each time index can be grouped into distinct "buckets" – however, time is continuous, and individual time points can belong to multiple primary communities. Similarly, airports do not need to belong to discrete communities – if

communities loosely correspond to geographical location, then airports may belong to some combination of geographical locations, as the location is a continuous parameter.

To ameliorate this assumption of distinct communities, in this paper we propose the *tensor mixed-membership blockmodel*, which relaxes the assumption that communities are discrete. Explicitly, we assume that each entry of the underlying tensor $\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ can be written via the decomposition

$$\mathcal{T}_{i_1 i_2 i_3} = \sum_{l_1=1}^{r_1} \sum_{l_2=1}^{r_2} \sum_{l_3=1}^{r_3} \mathcal{S}_{l_1 l_2 l_3} (\mathbf{\Pi}_1)_{i_1 l_1} (\mathbf{\Pi}_2)_{i_2 l_2} (\mathbf{\Pi}_3)_{i_3 l_3}, \tag{4.1}$$

where $\mathbf{\Pi}_k \in [0,1]^{p_k \times r_k}$ satisfies $\sum_{l=1}^{r_k} (\mathbf{\Pi}_k)_{il} = 1$ and $\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is a mean tensor. In words, the model (4.1) associates to each index along each mode a $[0,1]$-valued membership vector. For each index i of each mode k, the entries of its membership vector $(\mathbf{\Pi}_k)_{i\cdot}$ correspond to one of the r_k latent underlying communities, with the magnitude of the entry governing the intensity of membership within that community. The entry i_1, i_2, i_3 of the underlying tensor is then a weighted combination of the entries of the mean tensor $\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, with weights corresponding to three different membership vectors $(\mathbf{\Pi}_1)_{i_1\cdot}, (\mathbf{\Pi}_2)_{i_2\cdot}, (\mathbf{\Pi}_2)_{i_3\cdot}$.

In the previous example of airport flights, considering just the mode corresponding to time (in months), the mixed-membership tensor blockmodel posits that there are latent "pure" months and each individual time is a convex combination of these pure months. For a given index i, each entry of the i'th row of the membership matrix $(\mathbf{\Pi}_{\text{time}})_{i\cdot}$ corresponds to how much the time index i reflects each of the latent "pure communities." When the matrices $\mathbf{\Pi}_k$ are further assumed to be $\{0,1\}$-valued, every index is "pure" and this model reduces to the tensor blockmodel considered in Han et al. (2021); Chi et al. (2020); Wu et al. (2016), and Wang and Zeng (2019).

The factorization in (4.1) can be related to the so-called *Tucker decomposition* of the tensor $\mathcal{T}$. A tensor $\mathcal{T}$ is said to be of Tucker rank (r_1, r_2, r_3) if it can be written via

$$\mathcal{T} = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_3 \times_3 \mathbf{U}_3,$$

where $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is a *core tensor* and $\mathbf{U}_k \in \mathbb{R}^{p_k \times r_k}$ are orthonormal *loading matrices*

(see Section 4.1.2 for details). In this paper, we consider estimating $\mathbf{\Pi}_k$ (i.e., the community memberships) by considering the explicit relationship between the decomposition (4.1) and the loading matrices $\mathbf{U}_k$ in its Tucker decomposition. Our main contributions are as follows:

- We provide conditions for the identifiability of the model (4.1) (Proposition 2) and we relate the decomposition in (4.1) to the Tucker decomposition of the underlying tensor (Proposition 3 and Lemma 6).
- We propose an algorithm to estimate the membership matrices $\mathbf{\Pi}_k$ obtained by combining the *higher-order orthogonal iteration* (HOOI) algorithm with the corner-finding algorithm of Gillis and Vavasis (2014), and we demonstrate a high-probability *per-node* error bound for estimation of the membership matrices $\mathbf{\Pi}_k$ in the presence of heteroskedastic, subgaussian noise (Theorem 10).
- To prove our main results, we develop a new $\ell_{2,\infty}$ perturbation bound for the HOOI algorithm in the presence of heteroskedastic, subgaussian noise (Theorem 11) that may be of independent interest.
- We conduct numerical simulations and apply our algorithm to three different datasets:
  - In the first dataset consisting of global flights by airline, we find a "separation phenomenon" between American airports and airlines and Chinese airports and airlines.
  - In the next dataset consisting of flights between US airports over time, we obtain evidence of a "seasonality effect" that vanishes at the onset of the COVID-19 pandemic.
  - In the final dataset consisting of global trading for different food items, we find that global food trading can be grouped by region, with European countries grouped more closely together than other regions.

Our main technical result, Theorem 11, relies only on spectral properties of the underlying tensor and holds for nearly-optimal signal-to-noise ratio conditions such that a polynomial-time estimator exists. Our proof is based on a careful leave-one-out construction that requires additional "tensorial" considerations. See Section 4.4 for further details on our proof

techniques. For ease of presentation, this paper focuses on the order-three setting. Our results and methodology naturally extend to the higher-order setting, and we give an informal statement of the extension to higher-order in Section 5.5.

The rest of this paper is organized as follows. In Section 4.1.1 we review related works, and in Section 4.1.2 we set notation and review tensor algebra. In Section 4.2 we provide our main estimation algorithm and present our main theoretical results, including our per-node estimation errors, and in Section 4.3 we present our results on simulated and real data. In Section 4.4 we provide an overview of the proof of our main $\ell_{2,\infty}$ perturbation bound, and we finish in Section 5.5 with a discussion. We defer the full proofs of our main results to the appendices.

4.1.1 Related Work

Tensors, or multidimensional arrays, arise in problems in the sciences and engineering (Hore et al., 2016; Koniusz and Cherian, 2016; Schwab et al., 2019; Zhang et al., 2020a), and there is a need to develop principled statistical theory and methodology for these data. Tensor data analysis techniques are closely tied to spectral methods, which have myriad applications in high-dimensional statistics (Chen et al., 2021c), including in principal component analysis, spectral clustering, and as initializations for nonconvex algorithms (Chi et al., 2019). With the ubiquity of spectral methods, there has also been a development of both theory and methodology for fine-grained statistical inference with spectral methods, though the existing theory is limited to specific settings, and may not be applicable to tensors.

Algorithms for high-order clustering have relied on convex relaxations (Chi et al., 2020) or spectral relaxation (Wu et al., 2016). Perhaps the most closely related results for high-order clustering are in Han et al. (2021), who consider both statistical and computational thresholds for perfect cluster recovery. Their proposed algorithm HLloyd is a generalization of the classical Lloyd's algorithm for K-Means clustering to the tensor setting. Similarly, Luo and Zhang (2022) consider the statistical and computational limits for clustering, but they focus on expected misclustering error. Unlike these previous works, our model allows for mixed-memberships, and hence we are not estimating discrete memberships.

The tensor mixed-membership blockmodel is also closely related to and inspired by the

mixed-membership stochastic blockmodel proposed by Airoldi et al. (2008), and our estimation algorithm is closely related to the algorithm proposed in Mao et al. (2021), who propose estimating mixed memberships in networks by studying the relationship between the leading eigenvectors and the membership matrix. Similar to Mao et al. (2021) we also repurpose the algorithm proposed in Gillis and Vavasis (2014) for membership estimation, and we obtain our main results by obtaining sharp $\ell_{2,\infty}$ perturbation bounds for the estimated singular vectors. However, unlike Mao et al. (2021), our analysis requires studying the output of the higher-order orthogonal iteration (HOOI) algorithm, whereas Mao et al. (2021) need only consider the $\ell_{2,\infty}$ perturbation of the leading eigenvectors. Nearly optimal perturbation bounds for the matrix mixed-membership blockmodel have also been obtained in Xie (2022), and we provide a comparison of our results to both Mao et al. (2021) and Xie (2022), demonstrating the effect of higher-order "tensorial" structure on estimation accuracy. Our $\ell_{2,\infty}$ perturbation bounds are *not* simply extensions of previous bounds for matrices, and instead require additional novel theoretical techniques; see Section 4.4 for details.

Considering general perturbation results for tensors, Cai et al. (2022) focuses on symmetric tensors of low CP rank, and they consider the performance of their noisy tensor completion algorithm obtained via vanilla gradient descent, and they prove entrywise convergence guarantees and $\ell_{2,\infty}$ perturbation bounds. Our analysis differs in a few key ways: first, we consider tensors of low Tucker rank, which generalizes the CP rank; next, our analysis holds for *asymmetric* tensors under general subgaussian noise, and, perhaps most crucially, we analyze the HOOI algorithm, which can be understood as power iteration (as opposed to gradient descent). Therefore, while the results in Cai et al. (2022) may be qualitatively similar, the results are not directly comparable. Similarly, Wang et al. (2021) consider the entrywise convergence of their noiseless tensor completion algorithm for symmetric low Tucker rank tensors; our analysis is somewhat similar, but we explicitly characterize the effect of noise, which is a primary technical challenge in the analysis.

Besides Cai et al. (2022) and Wang et al. (2021), entrywise perturbation bounds for tensors are still lacking in general, though there are several generalizations of classical matrix perturbation bounds to the tensor setting. A sharp (deterministic) $\sin\Theta$ upper bound for tensor SVD was obtained in Luo et al. (2021), and Auddy and Yuan (2022b) consider per-

turbation bounds for orthogonally decomposable tensors. Zhang and Han (2019) considered tensor denoising when some of the modes have sparse factors. Zhang and Xia (2018) established statistical and computational limits for tensor SVD with Gaussian noise; our work builds off of their analysis by analyzing the tensor SVD algorithm initialized with diagonal deletion. Finally, Richard and Montanari (2014) and Auddy and Yuan (2022a) also consider estimating low CP-rank tensors under Gaussian and heavy-tailed noise respectively.

Our main $\ell_{2,\infty}$ bound is also closely related to a series of works developing fine-grained entrywise characterizations for eigenvectors and singular vectors, such as Abbe et al. (2020, 2022); Agterberg and Sulam (2022); Agterberg et al. (2022b); Cape et al. (2019b,a); Cai et al. (2021a); Koltchinskii and Xia (2015); Yan et al. (2021), for example. The monograph Chen et al. (2021c) gives an introduction to spectral methods from a statistical point of view, with the final chapter focusing on entrywise bounds and distributional characterizations for estimates constructed from eigenvectors. Several works on entrywise singular vector analyses have also applied their results to tensor data, such as Xia and Zhou (2019); Cai et al. (2021a), though these analyses often fail to take into account the additional structure arising in tensor data.

From a technical point of view, our work uses the "leave-one-out" analysis technique, first pioneered for entrywise eigenvector analysis in Abbe et al. (2020), though the method had been used previously to analyze nonconvex algorithms (Chi et al., 2019; Ma et al., 2020; Chen et al., 2020, 2021d,e), M-estimators (El Karoui et al., 2013; Sur et al., 2019; Sur and Candès, 2019), among others (Ding and Chen, 2020; Zhong and Boumal, 2018). The leave-one-out technique for singular vectors and eigenvectors has been further refined to analyze large rectangular matrices (Cai et al., 2021a), kernel spectral clustering (Abbe et al., 2022), to obtain distributional guarantees for spectral methods (Yan et al., 2021), and to study the performance of spectral clustering (Zhang and Zhou, 2022). A comprehensive survey on the use of this technique can be found in Chen et al. (2021c). Our work bridges the gap between analyzing nonconvex algorithms and analyzing spectral methods: HOOI performs a low-dimensional SVD at each iteration, thereby requiring both singular vector analyses and algorithmic considerations. Our proof technique also demonstrates an implicit regularization effect in tensor SVD – provided the initialization is sufficiently incoherent, tensor SVD

maintains this level of incoherence at each iteration. Finally, our proof of the spectral initialization also slightly improves upon the bound in Cai et al. (2021a) (for the singular vectors of rectangular matrices) by a factor of the condition number; see Theorem 25.

4.1.2 Notation and Preliminaries

For two functions f and g viewed as functions of some increasing index n, we say $f(n) \lesssim g(n)$ if there exists a uniform constant $C > 0$ such that $f(n) \leq Cg(n)$, and we say $f(n) \asymp g(n)$ if $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$. We write $f(n) \ll g(n)$ if $f(n)/g(n) \to 0$ as the index n increases. We also write $f(n) = O(g(n))$ if $f(n) \lesssim g(n)$, and we write $f(n) = \widetilde{O}(g(n))$ if $f(n) = O(g(n)\log^c(n))$ for some value c (not depending on n).

We use bold letters $\mathbf{M}$ to denote matrices, we let $\mathbf{M}_{i\cdot}$ and $\mathbf{M}_{\cdot j}$ denote its i'th row and j'th column, both viewed as column vectors, and we let $\mathbf{M}^\top$ denote its transpose. We denote $\|\cdot\|$ as the spectral norm for matrices and the Euclidean norm for vectors, and we let $\|\cdot\|_F$ denote the Frobenius norm. We let e_i denote the i'th standard basis vector and $\mathbf{I}_k$ denote the $k \times k$ identity. For a matrix $\mathbf{M}$ we let $\|\mathbf{M}\|_{2,\infty} = \max_i \|e_i^\top \mathbf{M}\|$. For two orthonormal matrices $\mathbf{U}$ and $\mathbf{V}$ satisfying $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$, we let $\|\sin\Theta(\mathbf{U},\mathbf{V})\|$ denote their $\sin\Theta$ (spectral) distance; i.e., $\|\sin\Theta(\mathbf{U},\mathbf{V})\| = \|(\mathbf{I}_r - \mathbf{U}\mathbf{U}^\top)\mathbf{V}\|$. For an orthonormal matrix $\mathbf{U}$ we let $\mathbf{U}_\perp$ denote its orthogonal complement; that is, $\mathbf{U}_\perp$ satisfies $\mathbf{U}_\perp^\top\mathbf{U} = 0$. We denote the $r \times r$ orthogonal matrices as $\mathbb{O}(r)$.

For multi-indices $\mathbf{r} = (r_1, r_2, r_3)$ and $\mathbf{p} = (p_1, p_2, p_3)$, we let $r_{-k} = \prod_{j\neq k} r_j$, and we define p_{-k} similarly. We also denote $p_{\min} = \min p_k$ and $p_{\max} = \max p_k$, with $r_{\min}$ and $r_{\max}$ defined similarly. A tensor $\mathcal{T} \in \mathbb{R}^{p_1\times p_2\times p_3}$ is a multidimensional array. We let calligraphic letters $\mathcal{T}$ denote tensors, except for the letter $\mathcal{M}$, for which $\mathcal{M}_k(\mathcal{T})$ denotes its *matricization* along the k'th mode; i.e., $\mathcal{M}_k(\mathcal{T})$ satisfies

$$\mathcal{M}_k(\mathcal{T}) \in \mathbb{R}^{p_k\times p_{-k}}; \qquad \big(\mathcal{M}_k(\mathcal{T})\big)_{i_k,j} = \mathcal{T}_{i_1 i_2 i_3}; \qquad j = 1 + \sum_{\substack{l=1\\ l\neq k}}^{d}\Big\{(i_l - 1)\prod_{\substack{m=1\\ m\neq k}} p_m\Big\},$$

for $1 \leq i_l \leq p_l$, $l = 1, 2, 3$. See Kolda and Bader (2009) for more details on matricizations. We also reserve the calligraphic letter $\mathcal{P}$ for either permutations or projections, as will be

clear from context. For an orthonormal matrix $\mathbf{U}$, we let $\mathcal{P}_{\mathbf{U}}$ denote its corresponding orthogonal projection $\mathcal{P}_{\mathbf{U}} = \mathbf{U}\mathbf{U}^\top$.

We denote the multilinear rank of a tensor $\mathcal{T}$ as a tuple $\mathbf{r} = (r_1, r_2, r_3)$, where r_k is the rank of the k'th matricization of $\mathcal{T}$. A tensor $\mathcal{T}$ of rank $\mathbf{r}$ has a Tucker decomposition

$$\mathcal{T} = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_3 \times_3 \mathbf{U}_3,$$

where $\mathcal{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is the core tensor and $\mathbf{U}_k$ are the $p_k \times r_k$ left singular vectors of the matrix $\mathcal{M}_k(\mathcal{T})$. Here the mode 1 product of a tensor $\mathcal{T} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ with a matrix $\mathbf{U} \in \mathbb{R}^{p_1 \times r_1}$ is denoted by $\mathcal{T} \times_k \mathbf{U}^\top \in \mathbb{R}^{r_1 \times p_2 \times p_3}$ and is given by

$$(\mathcal{T} \times_1 \mathbf{U}^\top)_{ji_2i_3} = \sum_{i_1=1}^{p_k} \mathcal{T}_{i_1 i_2 i_3} \mathbf{U}_{i_1 j}.$$

The other mode-wise multiplications are defined similarly. For two matrices $\mathbf{U}$ and $\mathbf{V}$, we denote $\mathbf{U} \otimes \mathbf{V}$ as their Kronecker product. For a tensor $\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ and matrices $\mathbf{U}_k$ of appropriate sizes, the following identity holds (see e.g., Kolda (2006)):

$$\mathcal{M}_1(\mathcal{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3) = \mathbf{U}_1 \mathcal{M}_1(\mathcal{S})\big(\mathbf{U}_2^\top \otimes \mathbf{U}_3^\top\big),$$

with similar identities holding for the other modes. For a matrix $\mathbf{M}$ we write $\mathrm{SVD}_r(\mathbf{M})$ to denote the leading r singular vectors of $\mathbf{M}$. Concretely, for a tensor of Tucker rank $\mathbf{r} = (r_1, r_2, r_3)$, it holds that $\mathbf{U}_k = \mathrm{SVD}_{r_k}(\mathcal{M}_k(\mathcal{T}))$.

For a tensor $\mathcal{T}$ with Tucker decomposition $\mathcal{T} = \mathcal{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$, we denote its incoherence parameter μ_0 as the smallest number such that

$$\max_k \sqrt{\frac{p_k}{r_k}} \|\mathbf{U}_k\|_{2,\infty} \leq \mu_0.$$

For a nonsquare matrix $\mathbf{M}$ of rank r, we let $\lambda_{\min}(\mathbf{M})$ denote its smallest nonzero singular value, and we denote its singular values as $\lambda_k(\mathbf{M})$. For a square matrix $\mathbf{M}$, we let $\lambda_{\min}(\mathbf{M})$ denote its smallest nonzero eigenvalue and $\sigma_{\min}(\mathbf{M})$ denote its smallest nonzero singular value, with other eigenvalues and singular values defined similarly. For a tensor $\mathcal{T}$ of rank

$\mathbf{r} = (r_1, r_2, r_3)$, we let $\lambda_{\min}(\mathcal{T})$ denote its smallest nonzero singular value along all of its matricizations; that is,

$$\lambda_{\min}(\mathcal{T}) = \min_k \lambda_{\min}(\mathcal{M}_k(\mathcal{T})).$$

We let the condition number of a tensor $\mathcal{T}$ be denoted as κ, defined as

$$\kappa := \max_k \frac{\|\mathcal{M}_k(\mathcal{T})\|}{\lambda_{\min}(\mathcal{M}_k(\mathcal{T}))}.$$

Finally, for a random variable X, we let $\|X\|_{\psi_2}$ denote its subgaussian Orlicz norm; that is,

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}.$$

See Chapter 2 of Vershynin (2018) for more details on Orlicz norms and subgaussian random variables.

4.2 Main Results

We now describe our model in detail. Assume that one observes

$$\widehat{\mathcal{T}} = \mathcal{T} + \mathcal{Z} \in \mathbb{R}^{p_1 \times p_2 \times p_3},$$

where $\mathcal{Z}$ consists of independent mean-zero subgaussian noise satisfying $\|\mathcal{Z}_{ijk}\|_{\psi_2} \leq \sigma$ (note that $\mathcal{Z}$ is not assumed to be homoskedastic). Assume further that the underlying tensor $\mathcal{T}$ admits the following factorization:

$$\mathcal{T} = \mathcal{S} \times_1 \mathbf{\Pi}_1 \times_2 \mathbf{\Pi}_2 \times_3 \mathbf{\Pi}_3, \tag{4.2}$$

where $\mathbf{\Pi}_k \in [0,1]^{p_k \times r_k}$ is a membership matrix with rows that sum to one, and $\mathcal{S} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ is a *mean tensor*. The matrices $\mathbf{\Pi}_k$ can be interpreted as follows: $(\mathbf{\Pi}_k)_{i_k l}$ denotes how much the i_k'th node along the k'th mode belongs to community l.

For a node i_k along mode k, we say i_k is a *pure node* if $(\mathbf{\Pi}_k)_{i_k \cdot} \in \{0,1\}^{r_k}$; that is, exactly

one entry of the i_k'th row of $\mathbf{\Pi}_k$ is nonzero (and hence equal to one). Intuitively, a pure node is a node that belongs to one and only one community. Observe that if all nodes are pure nodes, then one recovers the tensor blockmodel. As in the matrix setting (Mao et al., 2021), the existence of pure nodes is intimately related to the identifiability of the model (4.1). The following result establishes the identifiability of the tensor mixed-membership blockmodel when $\mathcal{S}$ is rank r_k along each mode and there is a pure mode for each community along each direction. We note that it is also possible to establish identifiability in the case that $\mathcal{S}$ has some mode with a rank less than r_k, but this is beyond the scope of this paper.

Proposition 2 (Identifiability)**.** *Consider the model (4.2). Assume the following two conditions hold.*

- *Each matricization of $\mathcal{S}$ is rank r_k respectively with $r_k \leq r_{-k}$;*
- *For each mode k, there is at least one pure node for each community.*

Then if there exists another set of parameters $\mathcal{S}', \mathbf{\Pi}_1', \mathbf{\Pi}_2'$, and $\mathbf{\Pi}_3'$ such that $\mathcal{T} = \mathcal{S}' \times_1 \mathbf{\Pi}_1' \times_2 \mathbf{\Pi}_2' \times_3 \mathbf{\Pi}_3'$ it must hold that $\mathbf{\Pi}_k = \mathbf{\Pi}_k' \mathcal{P}_k$, where $\mathcal{P}_k$ is an $r_k \times r_k$ permutation matrix and $\mathcal{S} = \mathcal{S}' \times_1 \mathcal{P}_1 \times_2 \mathcal{P}_2 \times_3 \mathcal{P}_3$.

Next, suppose that each matricization of $\mathcal{S}$ is rank r_k respectively with $r_k \leq r_{-k}$. Suppose that $\mathbf{\Pi}_k$ is identifiable up to permutation; i.e., any other $\mathbf{\Pi}_k'$ generating the same tensor $\mathcal{T}$ must satisfy $\mathbf{\Pi}_k' = \mathbf{\Pi}_k \mathcal{P}_k$ for some permutation $\mathcal{P}_k$ and $\mathcal{S}' = \mathcal{S} \times_k \mathcal{P}_k$. Then there must be at least one pure node for each community along mode k.

Therefore, we see that when the underlying tensor is full rank and there is at least one pure node for each community, the model will be identifiable up to permutation of the communities.

In order to describe our estimation procedure in the following subsection, we provide the following crucial observation relating the tensor mixed-membership blockmodel to its Tucker factorization.

Proposition 3. *Suppose $\mathcal{T}$ is a tensor mixed-membership blockmodel of the form in (4.2), and suppose that each matricization of $\mathcal{S}$ is rank r_k respectively with $r_k \leq r_{-k}$ for each k. Suppose further that there is a pure node for each community along each mode. Let*

$\mathcal{T} = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_3 \times_3 \mathbf{U}_3$ *denote its rank* (r_1, r_2, r_3) *Tucker factorization. Then it holds that*

$$\mathbf{U}_k = \mathbf{\Pi}_k \mathbf{U}_k^{(\text{pure})},$$

where $\mathbf{U}_k^{(\text{pure})} \in \mathbb{R}^{r_k \times r_k}$ *is rank* r_k *and contains the rows of* $\mathbf{U}_k$ *corresponding to pure nodes.*

Consequently, Proposition 3 shows that the singular vectors $\mathbf{U}_k$ of the underlying tensor $\mathcal{T}$ belong to a simplex with vertices given by $\mathbf{U}_k^{(\text{pure})}$, or the rows of $\mathbf{U}_k$ corresponding to pure nodes. The connection between the membership matrix $\mathbf{\Pi}_k$ and the singular vectors $\mathbf{U}_k$ has previously been considered in the matrix setting in Mao et al. (2021).

4.2.1 Estimation Procedure

We now detail our estimation procedure. In light of Proposition 3, the singular vectors of the tensor $\mathcal{T}$ and the matrices $\mathbf{\Pi}_k$ are intimately related via the matrix $\mathbf{U}_k^{(\text{pure})}$. Therefore, given estimated tensor singular vectors $\widehat{\mathbf{U}}_k$ obtained from the observed tensor $\widehat{\mathcal{T}}$, we propose to estimate the pure nodes by applying the corner-finding algorithm of Gillis and Vavasis (2014) to the rows of $\widehat{\mathbf{U}}_k$ to obtain estimated pure nodes. Consequently, in order to run the corner-finding algorithm, we will require the estimated tensor singular vectors $\widehat{\mathbf{U}}_k$.

However, unlike the matrix SVD, tensor SVD is not well-defined in general. For low Tucker rank tensors, a common algorithm to estimate the singular vectors of tensors is via the higher-order orthogonal iteration (HOOI) algorithm (De Lathauwer et al., 2000). Under the specific Gaussian additive model, this algorithm has been analyzed and minimax optimal error bounds in $\sin\Theta$ distances were established in Zhang and Xia (2018), which is the main impetus behind using HOOI to estimate the singular vectors. However, a major technical challenge in analyzing our estimator is in providing a fine-grained understanding of the output of HOOI for tensor SVD in order to ensure that the correct pure nodes are found. See Section 4.2.4 for further details. Algorithm 3 includes full pseudo-code for HOOI.

In order to initialize HOOI, since we do not assume homoskedastic noise we propose initializing via diagonal-deletion; namely, we define $\widehat{\mathbf{U}}_k^{(0)}$ as the leading r_k eigenvectors of

Algorithm 3 Higher-Order Orthogonal Iteration (HOOI)

Input: $\widehat{\mathcal{T}} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, Tucker rank $\mathbf{r} = (r_1, r_2, r_3)$, initialization $\widehat{\mathbf{U}}_2^{(0)}, \widehat{\mathbf{U}}_3^{(0)}$.
repeat
  Let $t = t + 1$
  For $k = 1, 2, 3$

$$\widehat{\mathbf{U}}_k^{(t)} = \mathrm{SVD}_{r_k}\left(\mathcal{M}_k\left(\widehat{\mathcal{T}} \times_{k'<k} (\widehat{\mathbf{U}}_{k'}^{(t)})^\top \times_{k'>k} (\widehat{\mathbf{U}}_{k'}^{(t-1)})^\top\right)\right).$$

until Convergence or the maximum number of iterations is reached.
return : $\widehat{\mathbf{U}}_k^{(t_{\max})}$.

the matrix

$$\Gamma\left[\mathcal{M}_k(\widehat{\mathcal{T}})\mathcal{M}_k(\widehat{\mathcal{T}})^\top\right],$$

where $\Gamma(\cdot)$ is the *hollowing operator*: for a square matrix $\mathbf{M}$, $\Gamma(\mathbf{M})$ sets its diagonal entries to zero, i.e.,

$$[\Gamma(\mathbf{M})]_{ij} = \begin{cases} [\mathbf{M}]_{ij} & i \neq j; \\ 0 & i = j. \end{cases}$$

Algorithm 4 provides the full pseudo-code for this initialization procedure.

We now have all the pieces to our estimation procedure. First, using the initializations $\widehat{\mathbf{U}}_k^{(0)}$, we plug these into Algorithm 3 to estimate the tensor singular vectors. Next, given the estimates $\widehat{\mathbf{U}}_k$ for $k = 1, 2, 3$, we obtain the index sets J_k containing the estimated pure nodes via the algorithm proposed in Gillis and Vavasis (2014), and we set $\widehat{\mathbf{U}}_k^{(\mathrm{pure})} := \big(\widehat{\mathbf{U}}_k\big)_{J_k \cdot}$. Finally, we estimate $\widehat{\mathbf{\Pi}}_k$ via $\widehat{\mathbf{\Pi}}_k = \widehat{\mathbf{U}}_k\big(\widehat{\mathbf{U}}_k^{(\mathrm{pure})}\big)^{-1}$. The full procedure is stated in Algorithm 5. In practice we have found that there are occasionally negative or very small values of $\widehat{\mathbf{\Pi}}_k$; therefore, our actual implementation thresholds small values and re-normalizes the rows of $\widehat{\mathbf{\Pi}}_k$, though the theory discussed in the following sections will be for the implementation without this additional step.

4.2.2 Technical Assumptions

To develop the theory for our estimation procedure, we will require several assumptions. In light of Proposition 2 and to induce regularity into the community memberships, we impose

Algorithm 4 Diagonal-Deletion Initialization

Input: $\widehat{\mathcal{T}} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, Tucker rank $\mathbf{r} = (r_1, r_2, r_3)$.
for $k = 2, 3$ **do**
  Set $\widehat{\mathbf{U}}_k^{(0)}$ as the leading r_k eigenvectors of the matrix $\widehat{\mathbf{G}}$, with

$$\widehat{\mathbf{G}} := \Gamma\big(\mathcal{M}_k(\widehat{\mathcal{T}})\mathcal{M}_k(\widehat{\mathcal{T}})^\top\big), \qquad \text{where } \Gamma(\cdot) \text{ is the } \textit{hollowing operator} \text{ that sets the diagonal of "}\cdot\text{" to zero;}$$

end for
return : $\widehat{\mathbf{U}}_k^{(0)}$.

Algorithm 5 Successive Projection Algorithm for Tensor Mixed-Membership Estimation

Input: tensor $\widehat{\mathcal{T}} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$, ranks r_1, r_2, r_3
Compute the estimated loading matrices $\{\widehat{\mathbf{U}}_k\}_{k=1}^3$ via Algorithm 3, initialized via Algorithm 4.
for $k = 1, 2, 3$ **do**
  $\mathbf{R} := \widehat{\mathbf{U}}_k$, $J_k = \{\}, j = 1$
  while $\mathbf{R} \neq 0_{n \times r_k}$ and $j \leq r_k$ **do**
    Set $j^* = \arg\max \|e_j^\top \mathbf{R}\|^2$. If there are ties, set j^* as the smallest index.
    Set $\mathbf{v}_j := e_{j^*}^\top \mathbf{R}$
    Set $\mathbf{R} = \mathbf{R}\big(\mathbf{I}_{r_k} - \frac{\mathbf{v}_j \mathbf{v}_j^\top}{\|\mathbf{v}_j\|^2}\big)$
    $J_k = J_k \cup \{j^*\}$
    $j = j + 1$
  end while
  Define $\widehat{\mathbf{\Pi}}_k := \widehat{\mathbf{U}}_k(\widehat{\mathbf{U}}_k[J_k, \cdot])^{-1}$
end for
return : three membership matrices $\{\widehat{\mathbf{\Pi}}_k\}_{k=1}^3$.

the following assumption.

Assumption 4.1 (Regularity and Identifiability)**.** *The community membership matrices* $\mathbf{\Pi}_k$ *satisfy*

$$\frac{p_k}{r_k} \lesssim \lambda_{\min}\left(\mathbf{\Pi}_k^\top \mathbf{\Pi}_k\right) \leq \lambda_{\max}\left(\mathbf{\Pi}_k^\top \mathbf{\Pi}_k\right) \lesssim \frac{p_k}{r_k}.$$

In addition, each matricization of $\mathcal{S}$ *is rank* r_k *respectively, and there is at least one pure node for each community for every mode.*

The condition above implies that each community is approximately the same size. When $\mathcal{T}$ is a tensor blockmodel, the matrix $\mathbf{\Pi}_k^\top \mathbf{\Pi}_k$ is a diagonal matrix with diagonal entries equal to the community sizes; Assumption 4.1 states then that the community sizes are each of

order p_k/r_k, which is a widely used condition in the literature on clustering (Löffler et al., 2021; Han et al., 2021; Wu and Yang, 2020; Hu and Wang, 2022).

Tensor SVD is feasible only with certain signal strength (Zhang and Xia, 2018). In order to quantify the magnitude of the signal strength, we introduce an assumption on the signal-to-noise ratio (SNR), as quantified in terms of singular values of $\mathcal{S}$ and maximum variance σ.

Assumption 4.2 (Signal Strength)**.** *The smallest singular value of $\mathcal{S}$, $\Delta = \lambda_{\min}(\mathcal{S})$, satisfies*

$$\frac{\Delta^2}{\sigma^2} \gtrsim \frac{\kappa^2 p_{\max}^2 \log(p_{\max}) r_1 r_2 r_3}{p_1 p_2 p_3 p_{\min}^{1/2}}.$$

Here κ denotes the condition number of $\mathcal{S}$. When $p_k \asymp p$ and $r_k = O(1)$, Assumption 4.2 is equivalent to the assumption

$$\frac{\Delta^2}{\sigma^2} \gtrsim \frac{\kappa^2 \log(p)}{p^{3/2}}.$$

Remark 10 (Relationship to Tensor Blockmodel)**.** *In the tensor blockmodel setting, Han et al. (2021) define the signal-strength parameter*

$$\widetilde{\Delta}^2 := \min_k \min_{i \neq j} \| \big(\mathcal{M}_k(\mathcal{S})\big)_{i\cdot} - \big(\mathcal{M}_k(\mathcal{S})\big)_{j\cdot} \|^2;$$

i.e., the worst case row-wise difference between any two rows of each matricization of $\mathcal{S}$. Han et al. (2021) explicitly consider settings where $\mathcal{S}$ is rank degenerate; however, if one further assumes that $\mathcal{S}$ is rank (r_1, r_2, r_3), then it is straightforward to check that both $\widetilde{\Delta}$ and Δ coincide up to a factor of the condition number, since

$$\| \big(\mathcal{M}_k(\mathcal{S})\big)_{i\cdot} - \big(\mathcal{M}_k(\mathcal{S})\big)_{j\cdot} \|^2 = (e_i - e_j)^\top \big(\mathcal{M}_k(\mathcal{S})\big)\big(\mathcal{M}_k(\mathcal{S})\big)^\top (e_i - e_j) \geq \|e_i - e_j\|^2 \Delta^2 \asymp \Delta^2,$$

and

$$\| \big(\mathcal{M}_k(\mathcal{S})\big)_{i\cdot} - \big(\mathcal{M}_k(\mathcal{S})\big)_{j\cdot} \|^2 = (e_i - e_j)^\top \big(\mathcal{M}_k(\mathcal{S})\big)\big(\mathcal{M}_k(\mathcal{S})\big)^\top (e_i - e_j) \leq \|e_i - e_j\|^2 \kappa^2 \Delta^2 \asymp \kappa^2 \Delta^2.$$

Han et al. (2021) demonstrate that the condition $\frac{\Delta^2}{\sigma^2} \gtrsim \frac{1}{p^{3/2}}$ *is required to obtain perfect cluster recovery in polynomial time if the number of cluster centroids is assumed constant and* $p_k \asymp p$. *Therefore, our assumption that* $\frac{\Delta^2}{\sigma^2} \gtrsim \frac{\kappa^2 \log(p)}{p^{3/2}}$ *is optimal up to logarithmic factors and factors of* κ *in order for a polynomial time estimator to achieve exact community detection (albeit in the simple setting that* $\mathcal{S}$ *is full rank along each mode). However, in contrast to Han et al. (2021), our model permits mixed memberships, so the two conditions are not directly comparable.*

Finally, our analysis relies heavily on the following lemma relating the signal strength parameter Δ to the smallest singular value of the core tensor $\mathcal{C}$ in the Tucker decomposition of $\mathcal{T}$.

Lemma 6. *Let* $\mathcal{T}$ *be a Tensor Mixed- Membership Blockmodel of the form (4.2), let* $\mathcal{T} = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$ *denote its Tucker decomposition, and let* $\lambda = \lambda_{\min}(\mathcal{C})$ *denote its smallest singular value. Suppose further that Assumptions 4.1 and 4.2 hold with* $r_k \le r_{-k}$ *for each* k. *Then it holds that*

$$\lambda \asymp \Delta \frac{(p_1 p_2 p_3)^{1/2}}{(r_1 r_2 r_3)^{1/2}}; \quad \mu_0 = O(1).$$

Furthermore, $\mathbf{U}_k^{(\text{pure})} (\mathbf{U}_k^{(\text{pure})})^\top = \left(\mathbf{\Pi}_k^\top \mathbf{\Pi}_k\right)^{-1}$.

4.2.3 Estimation Errors

The following theorem characterizes the errors in estimating $\mathbf{\Pi}_k$.

Theorem 10 (Uniform Estimation Error)**.** *Suppose that* $r_{\max} \lesssim p_{\min}^{1/2}$, *that* $r_{\max} \asymp r$ *with* $r \lesssim r_{\min}$, *and that* $\kappa^2 \lesssim p_{\min}^{1/4}$. *Suppose further that Assumptions 4.1 and 4.2 hold. Let* $\widehat{\mathbf{\Pi}}_k$ *be the output of Algorithm 5 with* t *iterations for* $t \asymp \log\left(\frac{\kappa p_{\max}(r_1 r_2 r_3)^{1/2}}{(\Delta/\sigma)(p_1 p_2 p_3)^{1/2}}\right) \vee 1$. *Then with probability at least* $1 - p_{\max}^{-10}$ *there exists a permutation matrix* $\mathcal{P} \in \mathbb{R}^{r_k \times r_k}$ *such that*

$$\max_{1 \le i \le p_k} \left\| \left(\mathbf{\Pi}_k - \widehat{\mathbf{\Pi}}_k \mathcal{P}\right)_{i\cdot} \right\| \lesssim \frac{\kappa \sqrt{r^3 \log(p)}}{(\Delta/\sigma)(p_{-k})^{1/2}}.$$

Consequently, when $p_k \asymp p$, it holds that

$$\max_{1\le i\le p_k} \|(\mathbf{\Pi}_k - \widehat{\mathbf{\Pi}}_k\mathcal{P})_{i\cdot}\| \lesssim \frac{\kappa\sqrt{r^3\log(p)}}{(\Delta/\sigma)p}.$$

Observe that Theorem 10 establishes a *uniform* error bound for the estimated communities; that is, the *worst case error* over all nodes i.

Remark 11 (Relationship to Matrix Mixed-Membership Blockmodels)**.** *Theorem 10 is related to similar bounds in the literature for the matrix setting. Mao et al. (2021) considers estimating the membership matrix with the leading eigenvectors of the observed matrix. Explicitly, they define the $p \times p$ matrix*

$$\mathbf{M} = \rho_n \mathbf{\Pi}\mathbf{B}\mathbf{\Pi}^\top,$$

and they assume one observes

$$\widehat{\mathbf{M}} = \mathbf{M} + \mathbf{E},$$

where $\mathbf{E}$ consists of mean-zero Bernoulli noise with $\mathbb{E}\mathbf{E}_{ij}^2 = \mathbf{M}_{ij}(1-\mathbf{M}_{ij})$. Their main result (Theorem 3.5) demonstrates an upper bound of the form

$$\|\mathbf{\Pi} - \widehat{\mathbf{\Pi}}\mathcal{P}\|_{2,\infty} = \widetilde{O}\left(\frac{r^{3/2}}{\sqrt{p\rho_n}\lambda_{\min}(\mathbf{B})}\right),$$

where the $\widetilde{O}(\cdot)$ hides logarithmic terms (we assume for simplicity that $\mathbf{B}$ is positive definite). See also Xie (2022) for a similar bound for sparse Bernoulli noise.

The term $\sqrt{\rho_n}\lambda_{\min}(B)$ can be informally understood as the signal-to-noise ratio (SNR) in the Bernoulli noise setting[1]*. Consequently, when $\kappa, r \asymp 1$, and $p_k \asymp p$, one has the bounds*

$$\text{(Matrix setting)} \qquad \|\mathbf{\Pi} - \widehat{\mathbf{\Pi}}\mathcal{P}\|_{2,\infty} = \widetilde{O}\left(\frac{1}{\mathrm{SNR}\times\sqrt{p}}\right);$$

$$\text{(Tensor setting)} \qquad \|\mathbf{\Pi} - \widehat{\mathbf{\Pi}}\mathcal{P}\|_{2,\infty} = \widetilde{O}\left(\frac{1}{\mathrm{SNR}\times p}\right).$$

[1]If the noise were truly Gaussian with mean $\mathbf{M}_{ij}$ and variance bounded by ρ_n, then our definition of Δ would be simply $\rho_n\lambda_{\min}(\mathbf{B})$, resulting in the SNR of $\sqrt{\rho_n}\lambda_{\min}(\mathbf{B})$.

Therefore, Theorem 10 can be understood as providing an estimation improvement of order $\sqrt{p}$ compared to the matrix setting – one may view this extra $\sqrt{p}$ factor as stemming from the higher-order tensor structure. However, the technical arguments required to prove Theorem 10 require analyzing the output of HOOI, and, consequently, Theorem 10 is not *simply an extension of previous results to the tensor setting.*

Since the rows of $\mathbf{\Pi}_k$ can be understood as weight vectors, a natural metric to use in this setting is the average ℓ_1 norm. Theorem 10 then implies the following corollary.

Corollary 4 (Average ℓ_1 Error)**.** *In the setting of Theorem 10, with probability at least $1 - p_{\max}^{-10}$, one has for each k,*

$$\inf_{\text{Permutations } \mathcal{P}} \frac{1}{p_k} \sum_{i=1}^{p_k} \|\left(\widehat{\mathbf{\Pi}}_k - \mathbf{\Pi}_k \mathcal{P}\right)_{i\cdot}\|_1 \lesssim \frac{r^2 \kappa \sqrt{\log(p_{\max})}}{(\Delta/\sigma)(p_{-k})^{1/2}}.$$

4.2.4 Key Tool: $\ell_{2,\infty}$ Tensor Perturbation Bound

In this section we introduce the new $\ell_{2,\infty}$ tensor perturbation bound, which serves as a key tool for developing the main results of this paper. Other $\ell_{2,\infty}$ bounds for HOOI in this setting have not appeared in the literature to the best of our knowledge. Unlike the matrix SVD, HOOI (Algorithm 3) is an iterative algorithm that proceeds by updating the estimates at each iteration. Therefore, analyzing the output of HOOI requires carefully tracking the interplay between noise and estimation error at each iteration as a function of the spectral properties of the underlying tensor.

In what follows, recall we define the *incoherence* of a tensor $\mathcal{T}$ as the smallest number μ_0 such that

$$\max_k \sqrt{\frac{p_k}{r_k}} \|\mathbf{U}_k\|_{2,\infty} \leq \mu_0.$$

By way of example, for a $p \times p \times p$ tensor $\mathcal{T}$, observe that when $\mathcal{T}$ contains only one large nonzero entry, it holds that $\mu_0 = \sqrt{p}$, whereas when $\mathcal{T}$ is the tensor with constant entries, it holds that $\mu_0 = 1$. Consequently, μ_0 can be understood as a measure of "spikiness" of the underlying tensor, with larger values of μ_0 corresponding to more "spiky" $\mathcal{T}$.

In addition, we will present bounds for the estimation of $\mathbf{U}_k$ up to right mulitplication of an orthogonal matrix $\mathbf{W}_k$. The appearance of the orthogonal matrix $\mathbf{W}_k$ occurs due to the fact that we do not assume that singular values are distinct, and hence singular vectors are only identifiable up to orthogonal transformation. Previous results of this type also typically include an additional orthogonal matrix (Agterberg and Sulam, 2022; Agterberg et al., 2022b; Cai et al., 2021a; Abbe et al., 2020; Cape et al., 2019b).

The following result establishes the $\ell_{2,\infty}$ perturbation bound for the estimated singular vectors from the HOOI algorithm (initialized via diagonal deletion, Algorithm 4) under the general tensor denoising model. Note that the setting considered in this section is more general than the setting considered in the previous sections.

Theorem 11. *Suppose that $\widehat{\mathcal{T}} = \mathcal{T} + \mathcal{Z}$, where $\mathcal{Z}_{ijk}$ are independent mean-zero subgaussian random variables satisfying $\|\mathcal{Z}_{ijk}\|_{\psi_2} \leq \sigma$. Let $\mathcal{T}$ have Tucker decomposition $\mathcal{T} = \mathcal{C} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$, and suppose that $\mathcal{T}$ is incoherent with incoherence constant μ_0. Suppose that $\lambda/\sigma \gtrsim \kappa\sqrt{\log(p_{\max})}p_{\max}/p_{\min}^{1/4}$, $\mu_0^2 r \lesssim p_{\min}^{1/2}$, that $\kappa^2 \lesssim p_{\min}^{1/4}$, and that $r \lesssim r_{\min}$, where $\lambda = \lambda_{\min}(\mathcal{C})$. Let $\widehat{\mathbf{U}}_k^{(t)}$ denote the output of HOOI (Algorithm 3) after t iterations initialized via diagonal deletion (Algorithm 4). Then there exists an orthogonal matrix $\mathbf{W}_k \in \mathbb{O}(r_k)$ such that after t iterations with $t \asymp \log(\sigma\kappa p_{\max}/\lambda) \vee 1$, with probability at least $1 - p_{\max}^{-10}$:*

$$\|\widehat{\mathbf{U}}_k^{(t)} - \mathbf{U}_k\mathbf{W}_k\|_{2,\infty} \lesssim \frac{\kappa\mu_0\sqrt{r_k \log(p_{\max})}}{\lambda/\sigma}.$$

Remark 12 (Signal Strength Condition)**.** *The condition $\lambda/\sigma \gtrsim \kappa\sqrt{\log(p_{\max})}p_{\max}/p_{\min}^{1/4}$ is only slightly stronger than the condition $\lambda/\sigma \gtrsim p_{\max}\sqrt{r/p_{\min}}$ when $r \lesssim p_{\min}^{1/2}$. It has been shown in Luo et al. (2021) that this second condition implies a bound of the form $\|\sin\Theta(\widehat{\mathbf{U}}_k, \mathbf{U}_k)\| \lesssim \frac{\sqrt{p_k}}{\lambda/\sigma}$, which matches the minimax lower bound established in Zhang and Xia (2018) when $p_k \asymp p$. Therefore, the condition $\lambda/\sigma \gtrsim \kappa\sqrt{\log(p_{\max})}p_{\max}/p_{\min}^{1/4}$ allows for different orders of p_k without being too strong. Perhaps one way to understand Theorem 11 is that after sufficiently many iterations, the $\sin\Theta$ upper bound is of order $\sqrt{p_k}/(\lambda/\sigma)$, and Theorem 11 demonstrates that these errors are approximately* spread out *amongst the entries of $\widehat{\mathbf{U}}_k$ when $\mathbf{U}_k$ is incoherent. It is perhaps of interest to study the $\ell_{2,\infty}$ errors under different signal strength conditions; however, we leave this setting to future work.*

Next for simplicity, consider the regime $p_k \asymp p$. First, we assume that $\mu_0^2 r \lesssim \sqrt{p}$ which allows r to grow. In addition, in this regime the SNR condition translates to the condition $\lambda/\sigma \gtrsim \kappa p^{3/4}\sqrt{\log(p)}$, which is optimal up to a factor of $\kappa\sqrt{\log(p)}$ for a polynomial-time estimator to exist (Zhang and Xia, 2018). The work Abbe et al. (2022) suggests that without an additional logarithmic factor it may not be possible to obtain $\ell_{2,\infty}$ bounds, so it is possible that this additional logarithmic factor is actually necessary.

Remark 13 (Optimality)**.** *It was shown in Zhang and Xia (2018) that the minimax rate for tensor SVD satisfies*

$$\inf_{\bar{\mathbf{U}}_k} \sup_{\mathcal{T}\in\mathcal{F}_{p,r}(\lambda)} \mathbb{E}\|\sin\Theta(\bar{\mathbf{U}}_k, \mathbf{U}_k)\| \gtrsim \frac{\sqrt{p_k}}{\lambda/\sigma},$$

where $\mathcal{F}_{p,r}(\lambda)$ is an appropriate class of low-rank signal tensors and the infimum is over all estimators of $\mathbf{U}_k$. By properties of the $\sin\Theta$ distance and the $\ell_{2,\infty}$ norm, it holds that

$$\begin{aligned}\inf_{\bar{\mathbf{U}}_k} \sup_{\mathcal{T}\in\mathcal{F}_{p,r}(\lambda)} \mathbb{E} \inf_{\mathbf{W}\in\mathbb{O}(r)} \|\bar{\mathbf{U}}_k - \mathbf{U}_k\mathbf{W}\|_{2,\infty} &\geq \frac{1}{\sqrt{p_k}} \inf_{\bar{\mathbf{U}}_k} \sup_{\mathcal{T}\in\mathcal{F}_{p_k,r}(\lambda)} \mathbb{E} \inf_{\mathbf{W}\in\mathbb{O}(r)} \|\bar{\mathbf{U}}_k - \mathbf{U}\mathbf{W}\|_2 \\ &\gtrsim \frac{1}{\sqrt{p_k}} \inf_{\bar{\mathbf{U}}_k} \sup_{\mathcal{T}\in\mathcal{F}_{p_k,r}(\lambda)} \mathbb{E}\|\sin\Theta(\bar{\mathbf{U}}_k, \mathbf{U}_k)\| \gtrsim \frac{1}{\lambda/\sigma}.\end{aligned}$$

Consequently, when $\kappa, \mu_0, r \asymp 1$, Theorem 11 shows that HOOI attains the minimax rate for the $\ell_{2,\infty}$ norm up to a logarithmic term. Such a result is new to the best of our knowledge.

Remark 14 (Relationship to Matrices)**.** *Considering again $p_k \asymp p$, under the conditions of Theorem 11, we prove (see Theorem 25) that the diagonal-deletion initialization satisfies the high-probability upper bound*

$$\|\widehat{\mathbf{U}}_k^S - \mathbf{U}_k\mathbf{W}_k\|_{2,\infty} \lesssim \left(\underbrace{\frac{\kappa\sqrt{p\log(p)}}{\lambda/\sigma}}_{\text{linear error}} + \underbrace{\frac{p^{3/2}\log(p)}{(\lambda/\sigma)^2}}_{\text{quadratic error}} + \underbrace{\kappa^2\mu_0\sqrt{\frac{r}{p}}}_{\text{bias term}}\right)\mu_0\sqrt{\frac{r}{p}}. \tag{4.3}$$

The full proof of this result is contained in Appendix D.1.5; it should be noted that this result slightly improves upon the bound of Cai et al. (2021a) by a factor of κ^2 (though we do not include missingness as in Cai et al. (2021a)). This quantity in (4.3) consists of three terms: the first term is the "linear error" that appears in Theorem 11, the second term is the

"quadratic error", and the third term is the error stemming from the bias induced by diagonal deletion. In the high noise regime $\lambda/\sigma \asymp p^{3/4}\text{polylog}(p)$*, the quadratic error can dominate the linear error; and, moreover, the bias term does not scale with the noise of the problem. In Zhang and Xia (2018), it was shown that tensor SVD removes the "quadratic" error for the* $\sin\Theta$ *distance for homoskedastic noise; Theorem 11 goes one step further to show that these errors are evenly spread out, particularly when* μ_0 *is sufficiently small.*

Remark 15 (Adaptivity of HOOI to Heteroskedasticity)**.** *The bias term in (4.3), which does not scale with the noise* σ*, arises naturally due to the fact that one deletes the diagonal of both the noise and the underlying low-rank matrix. In the setting that the noise is heteroskedastic, Zhang et al. (2022) showed that a form of bias-adjustment is necessary for many settings; moreover, they showed that their algorithm* `HeteroPCA` *eliminates this bias factor in* $\sin\Theta$ *distance – in essence, the* `HeteroPCA` *algorithm is a debiasing procedure for the diagonal of the Gram matrix. The follow-on works Agterberg et al. (2022b) and Yan et al. (2021) have shown that this algorithm also eliminates the bias term in* $\ell_{2,\infty}$ *distance, implying that it is possible to obtain a bound that scales with the noise, albeit at the cost of additional computation. In contrast, Theorem 11 shows that the HOOI algorithm, when initialized via diagonal deletion, does not require any additional bias-adjustment to combat heteroskedasticity in order to obtain a bound that scales with the noise. In effect, this result demonstrates that HOOI is* adaptive *to heteroskedasticity.*

Remark 16 (Implicit Regularization)**.** *Theorem 11 and its proof also reveal an implicit regularization effect in tensor SVD with subgaussian noise – when the underlying low-rank tensor is sufficiently incoherent (e.g.,* $\mu_0 = O(1)$*) and the signal-to-noise ratio is sufficiently strong, all of the iterations are also incoherent with parameter* μ_0*. Several recent works have proposed incoherence-regularized tensor SVD (Ke et al., 2020; Jing et al., 2021), and Theorem 11 suggests that this regularization may not be needed. However, these prior works have focused on the setting of Bernoulli noise, for which the subgaussian variance proxy can be of much larger order than the standard deviation, particularly for sparse networks. Moreover, these previous works have included some form of symmetry, whereas this work considers the completely asymmetric setting. Nevertheless, it may be possible to extend our*

work to the multilayer network or hypergraph setting using other forms of concentration inequalities for certain "tensorial terms" arising in the analysis (see Section 4.4 for details on the proof techniques).

4.2.5 The Cost of Ignoring Tensorial Structure

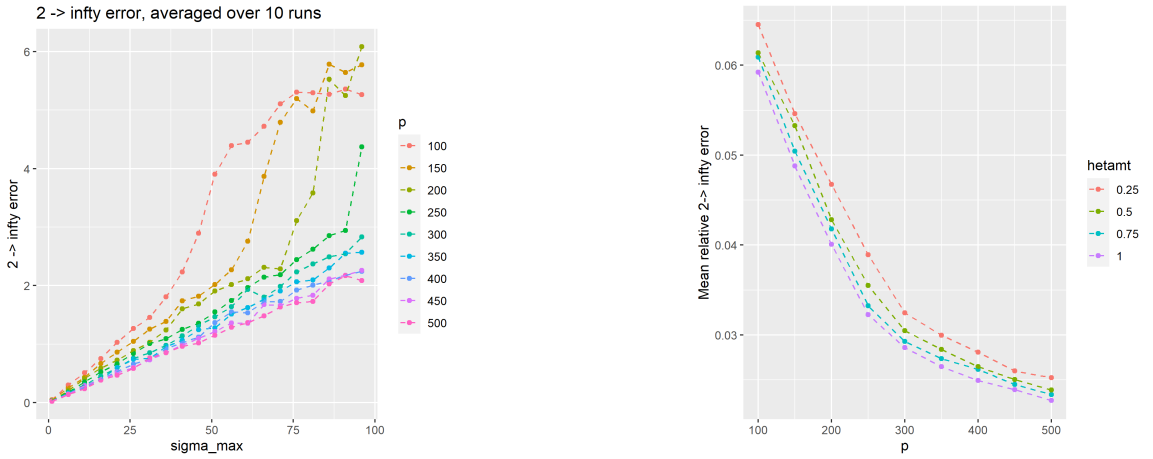
Perhaps the simplest tensor singular vector estimation procedure for Tucker low-rank tensors is the HOSVD algorithm, which simply takes the singular vectors of each matricization of $\widehat{\mathcal{T}}$ and outputs these as the estimated singular vectors. We discuss briefly why HOSVD-like procedures instead of HOOI in Algorithm 5 may not yield the same estimation error as in Theorem 10, particularly in the high-noise setting. For simplicity we focus on the regime $\kappa, r, \mu_0 \asymp 1$ and $p_k \asymp p$.

Recall that we do not assume that the noise is homoskedastic. It has previously been demonstrated that in the presence of heteroskedastic noise, HOSVD can yield biased estimates (Zhang et al., 2022). Therefore, in order to combat bias, one could modify the HOSVD procedure and instead use a procedure that eliminates the bias term stemming from either vanilla HOSVD or diagonal-deleted SVD (Algorithm 4). It was shown in Agterberg et al. (2022b) and Yan et al. (2021) that the output of the `HeteroPCA` algorithm proposed in Zhang et al. (2022) after sufficiently many iterations yields the high-probability upper bound

$$\|\widehat{\mathbf{U}}_k^{(\text{HeteroPCA})} - \mathbf{U}_k \mathbf{W}_k\|_{2,\infty} \lesssim \frac{\sqrt{\log(p)}}{\lambda/\sigma} + \frac{p \log(p)}{(\lambda/\sigma)^2},$$

where $\widehat{\mathbf{U}}_k^{(\text{HeteroPCA})}$ denotes the estimated singular vectors obtained by applying the `HeteroPCA` algorithm after sufficiently many iterations. Note that this upper bound does not suffer from any bias; in essence, this is the sharpest $\ell_{2,\infty}$ bound in the literature for any procedure that ignores tensorial structure.

Suppose one uses the estimate $\widehat{\mathbf{U}}_k^{(\text{HeteroPCA})}$ to estimate $\mathbf{\Pi}_k$ via Algorithm 5, and let $\widehat{\mathbf{\Pi}}_k^{(\text{HeteroPCA})}$ denote the output of this procedure. Arguing as in our proof of Theorem 10, by applying the results of Gillis and Vavasis (2014) and Lemma 6, using this bound we will



(a) $\ell_{2,\infty}$ estimation error defined via $\text{err} = \min_{\mathcal{P}} \|\mathbf{\Pi}_1 - \widehat{\mathbf{\Pi}}_1\mathcal{P}\|_{2,\infty}$ with varying values of $\sigma = \sigma_{\max}$ for heteroskedastic noise averaged over 10 runs.

(b) Relative $\ell_{2,\infty}$ error defined via err/σ averaged across each p for different amounts of heteroskedasticity, ranging from 1 (least heteroskedastic) to .25 (most heteroskedastic).

Figure 4.1: Simulated maximum node-wise errors, as described in Section 4.3.1.

obtain that

$$\|\widehat{\mathbf{\Pi}}_k^{(\text{HeteroPCA})} - \mathbf{\Pi}_k\mathcal{P}\|_{2,\infty} \lesssim \frac{\sqrt{\log(p)}}{(\Delta/\sigma)p} + \frac{\log(p)}{(\Delta/\sigma)^2 p^{3/2}}.$$

In the challenging regime $\Delta/\sigma \asymp \frac{\sqrt{\log(p)}}{p^{3/4}}$ (recall that by Assumption 4.2 we must have that $\frac{\Delta^2}{\sigma^2} \gtrsim \frac{\log(p)}{p^{3/2}}$), the above bound translates to

$$\|\widehat{\mathbf{\Pi}}_k^{(\text{HeteroPCA})} - \mathbf{\Pi}_k\mathcal{P}\|_{2,\infty} \lesssim \frac{1}{p^{1/4}} + 1 \asymp 1,$$

which does not tend to zero as $p \to \infty$. Therefore, in this high-noise regime, the estimates obtained via `HeteroPCA` (or any similar procedure that ignores the tensorial structure) may not even be consistent. In contrast, Theorem 10 shows that in this regime our proposed estimation procedure yields the upper bound

$$\|\widehat{\mathbf{\Pi}}_k - \mathbf{\Pi}_k\mathcal{P}\|_{2,\infty} \lesssim \frac{1}{p^{1/4}},$$

which still yields consistency, even in the high-noise regime.

4.3 Numerical Results

We now consider the numerical performance of our proposed procedure. In Section 4.3.1 we provide simulation results for several examples, including varying levels of heteroskedasticity. We then apply our algorithm to three different flight data sets – the first is the data described in Han et al. (2021), which measures global flights, the second is USA flight data, available from the Bureau of Transportation Statistics[2], and the third is a global trade dataset, available from De Domenico et al. (2015).

4.3.1 Simulations

In this section we consider the maximum row-wise estimation error for the tensor mixed membership blockmodel via Algorithm 5 for simulated data. In each data setup we generate the underlying tensor by first generating the mean tensor $\mathcal{S} \in \mathbb{R}^{3\times 3\times 3}$ with $N(0,1)$ entries and then adjusting the parameter Δ to 10. We then draw the memberships by manually setting the first three nodes along each mode to be pure nodes, and then drawing the other vectors from a random Dirichlet distribution. We generate the noise as follows. First, we generate the standard deviations $\{\sigma_{ijk}\}$ via $\sigma_{ijk} \sim \sigma_{\max} \times \beta(\alpha, \alpha)$, where β denotes a β distribution. The parameter α governs the heteroskedasticity, with $\alpha = 1$ corresponding to uniformly drawn standard deviations and $\alpha \to 0$ corresponding to "highly heteroskedastic" standard deviations. We then generate the noise via $\mathcal{Z}_{ijk} \sim N(0, \sigma^2_{ijk})$.

In Fig. 4.1(a) we examine the $\ell_{2,\infty}$ error as a function of $\sigma = \sigma_{\max}$ for Algorithm 5 applied to this noisy tensor averaged over 10 runs with $\alpha = 1$. Here we keep the mean matrix fixed but re-draw the memberships, variances, and noise each run. We vary σ from 1 to 96 by five. We see a clear linear relationship in the error for each value of p from 100 to 500 by 50, with larger values of $\sigma_{\max}$ being significantly less accurate for smaller values of p.

In Fig. 4.1(b) we consider the mean relative $\ell_{2,\infty}$ error defined as follows. First, for each value of $\sigma_{\max}$ we obtain an estimated $\ell_{2,\infty}$ error err averaged over 10 runs. We then divide this error by $\sigma_{\max}$ to put the errors on the same scale. Finally, we average this

[2] https://www.transtats.bts.gov/

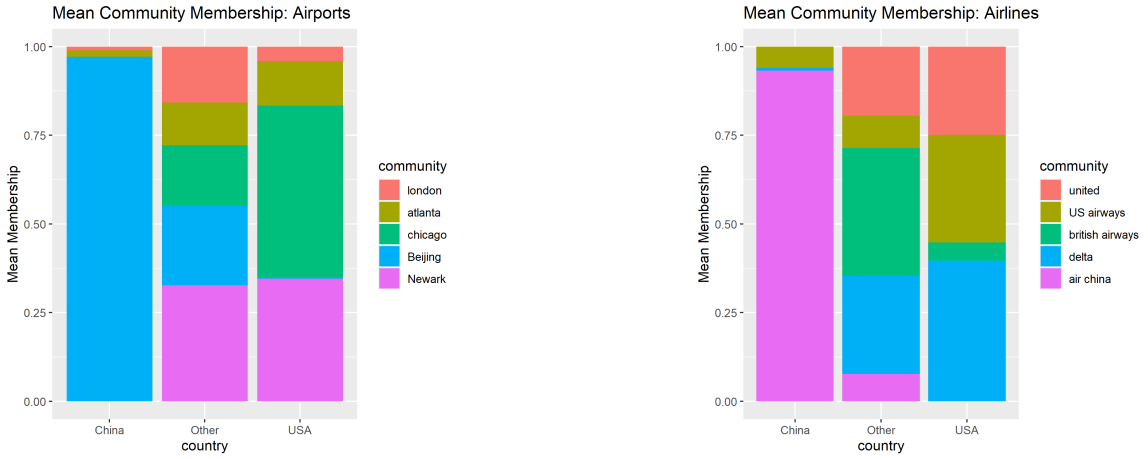


Figure 4.2: Community memberships for airports (left) and airlines (right), separated according to country to emphasize "disconnectedness" between Chinese airports and airlines with American airports and airlines.

error for all values of $\sigma_{\max}$ and plot the value as a function of p for different amounts of heteroskedasticity. We see that the error decreases in p as anticipated, and slightly more heteroskedasticity results in slightly worse performance.

4.3.2 Application to Global Flight Data

We now apply our mixed-membership estimation to the flight data described in Han et al. (2021). There are initially 66,765 global flight routes from 568 airlines and 3,309 airports[3], and we preprocess similar to Han et al. (2021) by considering only the top 50 airports with the highest numbers of flight routes. We end up with a tensor $\widehat{\mathcal{T}}$ of size $39 \times 50 \times 50$, where each entry $\widehat{\mathcal{T}}_{i_1 i_2 i_3}$ is one if there is a flight route from airport i_2 to i_3 in airline i_1 and zero otherwise. We use the same choice of $\mathbf{r} = \{5, 5, 5\}$ as in Han et al. (2021), chosen via the Bayesian information criterion for block models (Wang and Zeng, 2019) from candidate r values ranging from 3 to 6. When running our algorithm, occasionally there are negative or very small values of $\widehat{\mathbf{\Pi}}_k$; we therefore threshold and re-normalize in order to obtain our estimates.

First, our algorithm relies on identifying pure nodes along each mode. For the airports, the pure nodes are London, Atlanta, Chicago, Beijing, and Newark. For airlines, we find the pure nodes to be United, US airways, British Airways, Delta, and Air China. When ana-

[3] https://openflights.org/data.html#route

lyzing the output, we found that airlines and airports associated to the USA had extremely low membership in Chinese-associated pure nodes, and vice versa for Chinese airlines and airports. Therefore, in Fig. 4.2 we plot the average membership of each airport and airline associated to its home country, whether it is in China, the USA, or elsewhere. This figure demonstrates that the USA has less membership in the airline and airport communities based outside the USA; in particular almost no membership in Chinese communities, and China has almost entirely pure membership in Chinese airport and airline communities. The other countries have nearly equal membership in each community.

Furthermore, we observe that the USA airlines have zero membership in the "Air China" pure node, and the China airlines have primarily membership in the "Air China" pure node. We find a similar phenomenon in the airports as well. Interestingly, other airports (i.e., non-Chinese and non-American) do not exhibit this phenomenon. In Han et al. (2021) five clusters were found, including one that contains Beijing, which is a pure node here. This analysis suggests that perhaps the Beijing cluster might be much more distinct from the USA cluster than the other clusters are from each other. Airports and airlines in other countries do not exhibit such a trend – they have memberships in all other clusters equally. This observation is not identifiable in settings with discrete memberships, since either a node belongs to a community or does not, whereas in the tensor mixed membership blockmodel setting we can examine the strength of the membership.

4.3.3 Application to USA Flight Data

We also apply our methods to USA flight data publicly available from the Bureau of Transportation Statistics[4] and also analyzed in Agterberg et al. (2022a). We focused on the largest connected component, resulting in 343 airports with counts of flights between airports for each month from January 2016 to September 2021, resulting in 69 months of data and a $343 \times 343 \times 69$ dimensional tensor. To choose the embedding dimension, we apply the "elbow" method of Zhu and Ghodsi (2006). First, we apply the elbow procedure to the square roots of the nonnegative eigenvalues of the diagonal-deleted Gram matrix, which is the matrix we use for our initialization. This yields $r_1 = r_2 = 3$ for the airport mode, but

[4] https://transtats.bts.gov/

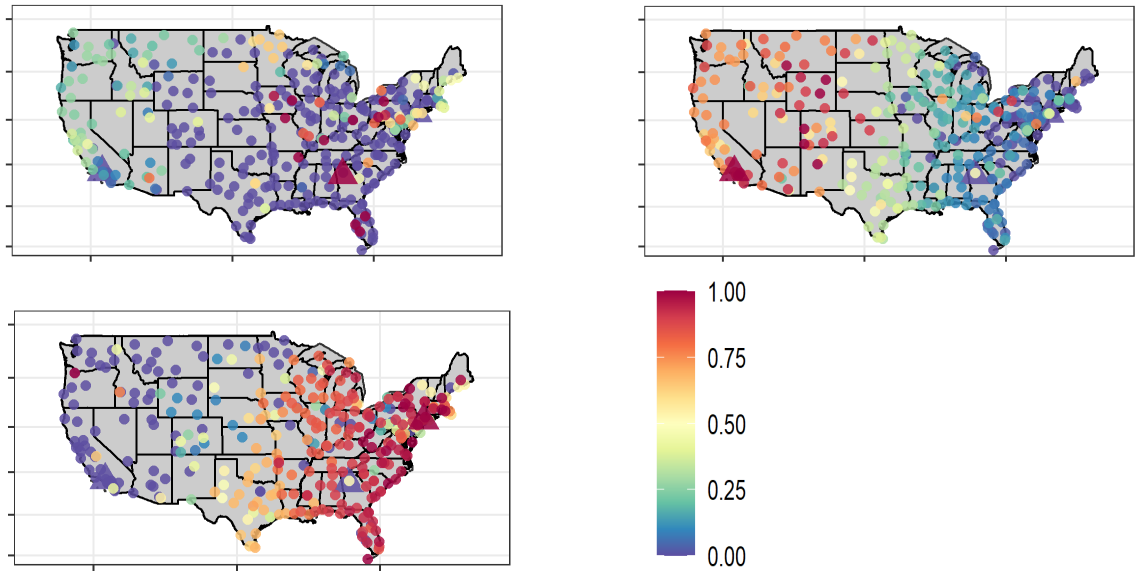


Figure 4.3: Pure node memberships for the airport mode, with pure nodes ATL (top left), LAX (top right), and LGA (bottom left). Red demonstrates high membership and purple demonstrates low membership within that particular community. The pure nodes are drawn with large triangles.

for the mode corresponding to time, this procedure resulted in only two nonnegative eigenvalues. Therefore, we ran the elbow method on the vanilla singular values instead, resulting in elbows at 1 and 4. We therefore chose $r_3 = 4$ to perform our estimation.

Plotted in Fig. 4.3 are the membership intensities in each of the communities associated to the three different pure nodes, with red corresponding to high membership and purple corresponding to low memberships. The three pure nodes were found to be ATL (Atlanta), LAX (Los Angeles), and LGA (New York). From the figure it is evident that the LGA community is associated with flights on the eastern half of the country, and LAX is associated with flights on the western half of the country. Based on the colors, the ATL community has memberships primarily from some airports on both the east and west coasts, but less directly in central USA. Therefore, it seems that the ATL community serves as a "hub" community connecting airports in the west coast to airports in the east coast – this intuition is justified by noting that ATL has the largest number of destinations out of any airport in the USA.

Plotted in Fig. 4.4 are the memberships in each of the four time communities, where the pure nodes were found to be August 2016, March 2020, January 2021, and August 2021. The blue lines correspond to the yearly smoothed values (using option `loess` in the

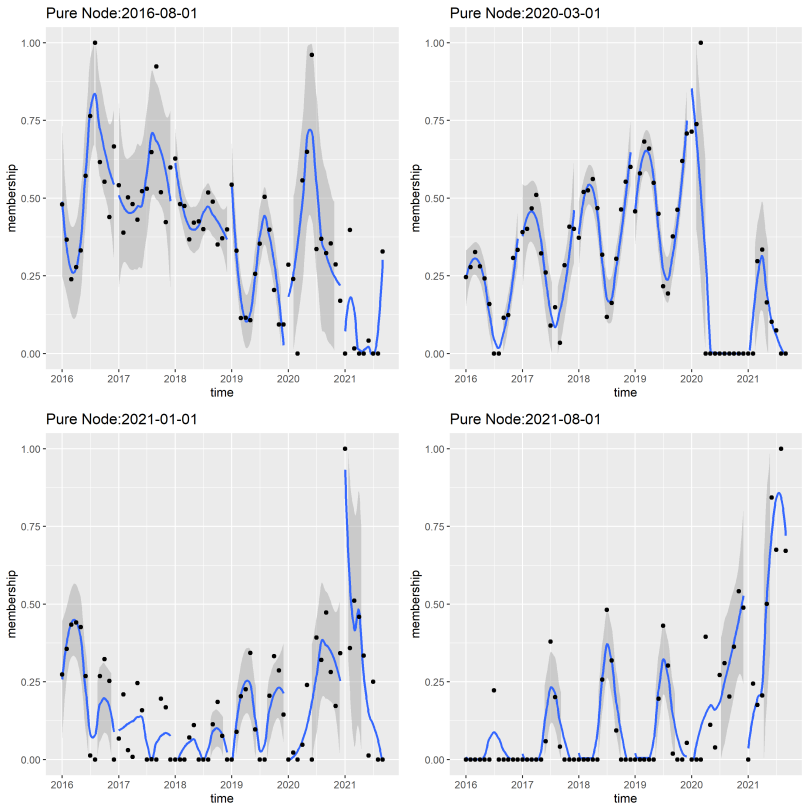


Figure 4.4: Pure node memberships for the time mode, with higher values corresponding to stronger membership intensity. Data are smoothed within each year to emphasize the effect of seasonality

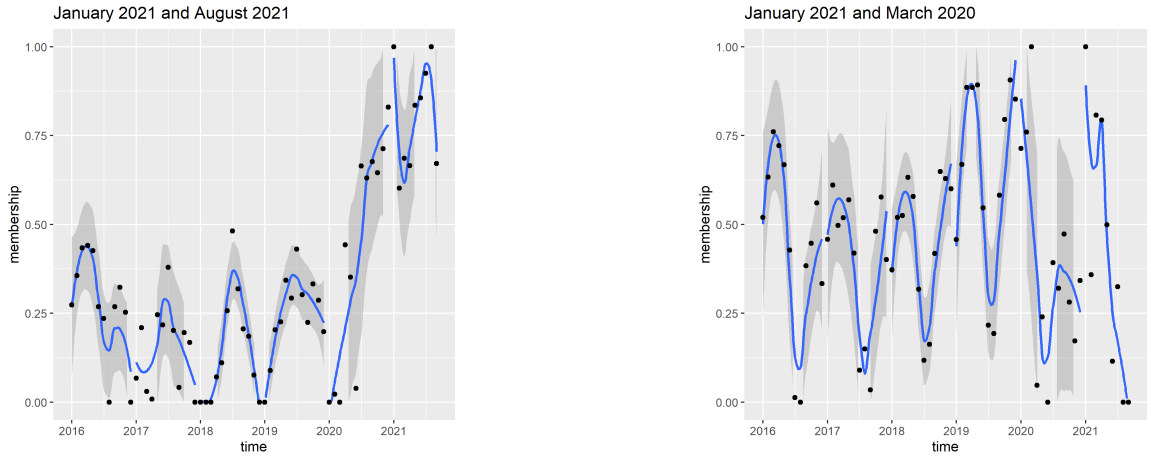


Figure 4.5: Joint plot emphasizing COVID-19 (left) and seasonality (right) effects of the January 2021 community.

`R` programming language), and the grey regions represent confidence bands. We chose to smooth within each year in order to emphasize seasonality. Immediately one notices the pure

node associated to March 2020 yields strong seasonality (demonstrating a "sinusoidal" curve within each year), only for it to vanish at the onset of the COVID-19 lockdowns in the USA, which began on March 15th, 2020. The seasonality effect seems to mildly recover in 2021, which roughly corresponds to the reopening timeline. The pure nodes associated to August seem to demonstrate a seasonality effect, with August 2021 also including a COVID-19 effect (as the membership in 2020 increases) – note that vaccines in the USA became available to the general public beginning in May 2021, so the community associated to August 2021 may include some of the "normal" seasonal effects.

Finally, the January 2021 community seems to exhibit a combination of a form of seasonality together with COVID-19, though it is perhaps not as pronounced as the March 2020 seasonality effect, nor is it as pronounced as the August 2021 COVID-19 effect. To emphasize these effects, we plot this mode by combining it (i.e., adding it together) with March 2020 (to emphasize seasonality) and August 2021 (to emphasize the COVID-19 effect) in Fig. 4.5. When combined with August 2021, the COVID-19 effect becomes more pronounced during and after 2020. When combined with March 2020, the seasonality effect becomes even more pronounced before March 2020, with larger swings within each year. Both combinations further corroborate our finding that the January 2021 community exhibits both of these effects.

4.3.4 Application to Global Trade Data

Next, we apply our algorithm to the global trade network dataset collected in De Domenico et al. (2015)[5] and further analyzed in Jing et al. (2021), which consists of trading relationships for 364 different goods between 214 countries in the year 2010, with weight corresponding to amount traded. Here each individual network corresponds to the trading relationships between countries for a single good. To preprocess, we first convert each network to undirected, and we keep the networks with largest connected component of at least size 150, which results in a final tensor of dimension $59 \times 214 \times 214$. Note that unlike Jing et al. (2021), we do not delete or binarize edges, nor do we only use the largest connected component within each network. To select the ranks we use the same method as in the

[5] http://www.fao.org

previous section, resulting in $\mathbf{r} = (5, 4, 4)$.

In Fig. 4.6 we plot each of the memberships associated to the "country" mode, where the pure nodes are found to be USA, Japan, Canada, and Germany. For the communities corresponding to Germany and Japan, we see that the weight of the corresponding countries roughly corresponds to geographical location, with closer countries corresponding to higher membership intensity. In particular, Germany's memberships are highly concentrated in Europe and Africa, with the memberships of all European countries being close to one. On the other hand, for the pure nodes associated to the USA and Canada, we see that the membership is relatively dispersed outside of Europe, which provides evidence that European trade communities are "closer-knit" than other communities. Since the USA and Canada likely have similar trading patterns, in Fig. 4.7 we combine these two values by adding them together, and we see that the memberships are fairly global besides Europe, though the intensity in any one area is not as strong as the intensities for the other pure nodes.

Next we consider the pure nodes corresponding to the different goods. The pure nodes were found to be maize (corn), crude materials, distilled alcoholic beverages, food prep nes (not elsewhere specfied), and whole cow milk cheese. It was found in Jing et al. (2021) that communities roughly correspond to either prepared or unprepared food; we also found food prep nes as one of the pure nodes, which gives further evidence to this finding. This community is also the "largest" community – the mean membership in this mode is .4147. To better understand the separation between processed and unprocessed food, we combine the "processed" communities food prep nes, distilled alcoholic beverages, and whole cow milk cheese into one community and group the other two communities together. Below is a summary of the communities with greater than .7 membership intensity in either group, as well as those with smaller than .7 intensity in both communities.

- **Processed** $> .7$: Tobacco products nes, Butter (cowmilk), Tomatoes, Milk (skimmed, dried), Tobacco (unmanufactured), Spices (nes), Fruit (prepared nes), Cigarettes, Potatoes, non alcoholic Beverages, Vegetables (frozen), Oil (essential nes), Oil (vegetable origin nes), Nuts (prepared (exc. groundnuts)), Sugar Raw Centrifugal, Vegetables (fresh nes), Waters (ice, etc.), Flour, wheat, Nuts nes, Tomato paste, Macaroni,

Sugar refined, Food prep nes, Cheese (whole cow milk), Chocolate products nes, Beer of barley, Beverages (distilled alcoholic), Bread, Cereals (breakfast), Coffee extracts, Coffee (roasted), Fruit (dried nes), Apples, Flour (maize), Pastry, Sugar confectionery, Wine, Sugar nes.

- **Unprocessed** $> .7$: Crude materials, Maize, palm oil, Sesame seed, Wheat
- **Neither**: milled Rice, dehydrated Vegetables, Pepper (piper spp.), chicken, Infant food, Fruit (fresh nes), Tea, Beans (dry), Coffee (green), dry Chillies and peppers, orange juice (single strength), soybean oil, fruit Juice nes, Milk (whole dried), Vegetables (preserved nes), Honey (natural).

By examining these "communities," it seems that the processed foods are more similar than the unprocessed foods, since many more foods have higher memberships in communities associated to processed foods. Moreover, the "neither" category also contains some "mildly processed foods" (e.g., dried milk), which shows how the mixture model here is more representative of the data. We leave further investigations to future work.

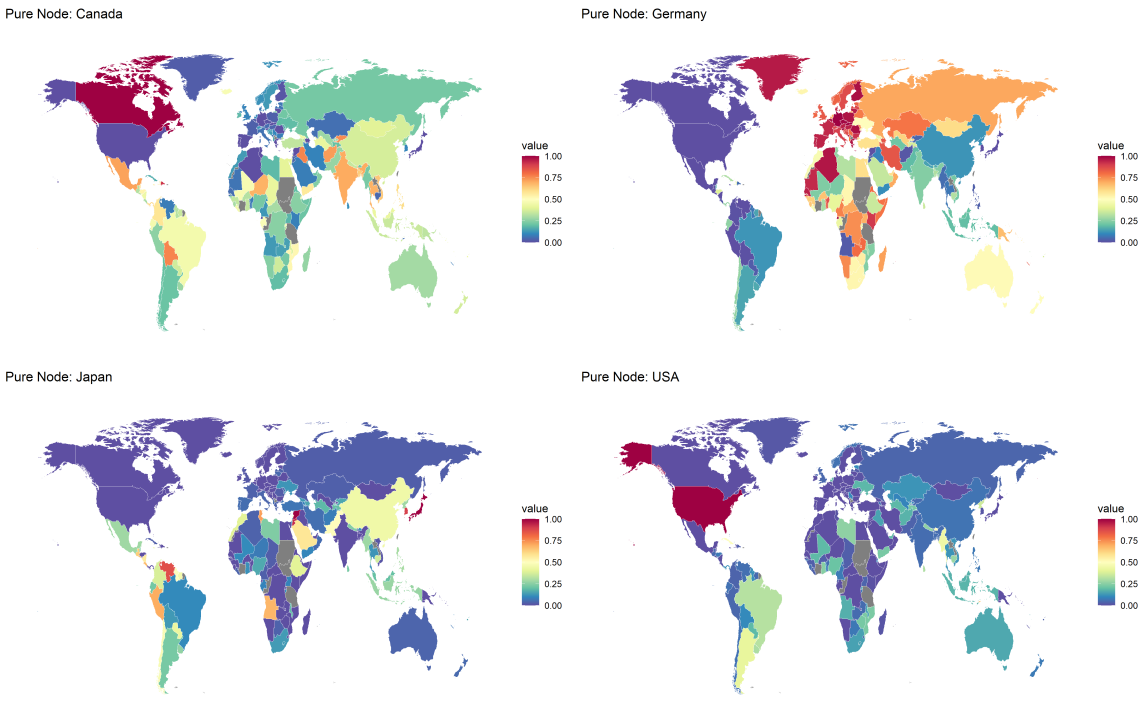


Figure 4.6: Pure node memberships for the countries, with red corresponding to higher membership intensity. Grey corresponds to countries that were not included in the analysis.

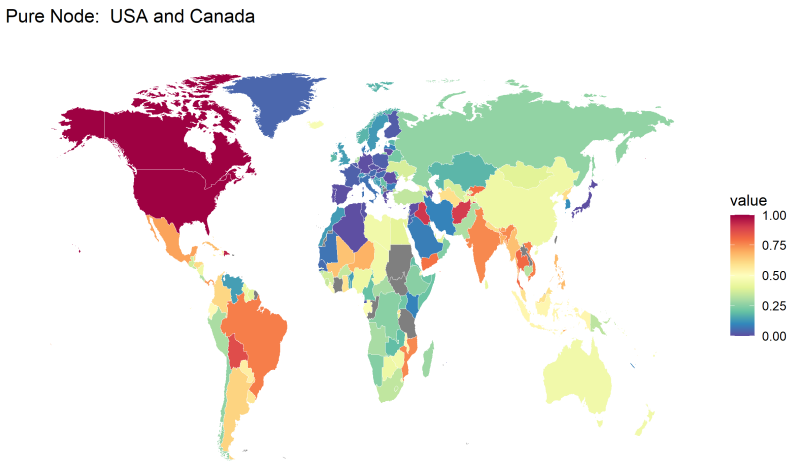


Figure 4.7: Combined memberships for the pure nodes associated to the USA and Canada.

4.4 Overview of the Proof of Theorem 11

In this section we provide a high-level overview and highlight the novelties of the proof of Theorem 11, the main theorem of this paper. As mentioned previously, the proof idea is based on a leave-one-out analysis. Different versions of leave-on-out analysis has been used in, for example, Yan et al. (2021); Abbe et al. (2022, 2020); Chen et al. (2021c). However, a key difference is that these previous works focus on the eigenvectors or singular vectors of the perturbed matrix, and subsequently do not have to analyze additional iterates. A series of related works in nonconvex optimization have studied the iterates of algorithms using a leave-one-out sequence, but these analyses typically focus on gradient descent or similar optimization techniques (Ma et al., 2020; Chen et al., 2021e,d; Zhong and Boumal, 2018; Ding and Chen, 2020; Cai et al., 2022), and hence do not need to consider additional singular vectors besides the first step. In contrast to these works, our upper bounds rely on an inductive argument in the number of iterations, where we show that a certain upper bound holds for each iteration, so our analysis uses primarily matrix perturbation tools.

To be concrete, without loss of generality, assume that $\sigma = 1$. For simplicity assume that $p_k \asymp p$ throughout this section. Our proof proceeds by showing that at each iteration

t with probability at least $1 - 3tp^{-15}$ one has the bound

$$\|\widehat{\mathbf{U}}_k^{(t)} - \mathbf{U}_k \mathbf{W}_k^{(t)}\|_{2,\infty} \leq \left(\frac{\delta_{\mathrm{L}}}{\lambda} + \frac{1}{2^t}\right)\mu_0\sqrt{\frac{r}{p}}, \tag{4.4}$$

where we define δ_{L} as the *linear error*

$$\delta_{\mathrm{L}} \coloneqq C_0\kappa\sqrt{p\log(p)},$$

with C_0 some fixed constant. Here the matrix $\mathbf{W}_k^{(t)}$ is defined via

$$\mathbf{W}_k^{(t)} := \underset{\mathbf{W}:\mathbf{W}\mathbf{W}^\top = \mathbf{I}_{r_k}}{\arg\min} \|\widehat{\mathbf{U}}_k^{(t)} - \mathbf{U}_k\mathbf{W}\|_F;$$

i.e., it is the orthogonal matrix most closely aligning $\mathbf{U}_k$ and $\widehat{\mathbf{U}}_k^{(t)}$ in Frobenius norm. A key property of the matrix $\mathbf{W}_k^{(t)}$ is that it can be computed analytically from the singular value decomposition of $\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)}$ as follows. Let $\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)}$ have singular value decomposition $\mathbf{W}_1\Sigma\mathbf{W}_2^\top$; then $\mathbf{W}_k^{(t)} = \mathbf{W}_1\mathbf{W}_2^\top$. The matrix $\mathbf{W}_k^{(t)}$ is also known as the *matrix sign function* of $\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)}$, denoted as $\mathrm{sgn}(\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)})$.

To show that the bound in (4.4) holds, we consider a fixed mth row to note that

$$\begin{aligned} e_m^\top\left(\widehat{\mathbf{U}}_k^{(t)} - \mathbf{U}_k\mathbf{W}_k^{(t)}\right) &= e_m^\top\left(\widehat{\mathbf{U}}_k^{(t)} - \mathbf{U}_k\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)} + \mathbf{U}_k(\mathbf{U}_k\widehat{\mathbf{U}}_k^{(t)} - \mathbf{W}_k^{(t)})\right) \\ &= e_m^\top\left((\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^\top)\widehat{\mathbf{U}}_k + \mathbf{U}_k(\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)} - \mathbf{W}_k^{(t)})\right). \end{aligned}$$

The second term is easily handled since $\mathbf{W}_k^{(t)}$ is close to $\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)}$ (see Lemma 33).

The first term requires additional analysis. For ease of exposition, consider the case $k = 1$. Recall that $\widehat{\mathbf{U}}_1^{(t)}$ are defined as the left singular vectors of the matrix

$$\widehat{\mathbf{T}}_1\left(\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}\right) = \mathbf{T}_1\left(\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}\right) + \mathbf{Z}_1\left(\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}\right),$$

where $\mathbf{T}_1 = \mathcal{M}_1(\mathcal{T})$ and $\mathbf{Z}_1 = \mathcal{M}_1(\mathcal{Z})$. To analyze $\widehat{\mathbf{U}}_1^{(t)}$ more directly, we consider the

corresponding Gram matrix. Then $\widehat{\mathbf{U}}_1^{(t)}$ are the *eigenvectors* of the matrix

$$\mathbf{T}_1\Big(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\Big)\mathbf{T}_1^\top+\mathbf{Z}_1\Big(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\Big)\mathbf{T}_1^\top$$
$$+\mathbf{T}_1\Big(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\Big)\mathbf{Z}_1^\top+\mathbf{Z}_1\Big(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\Big)\mathbf{Z}_1^\top.$$

Let this matrix be denoted $\widehat{\mathbf{T}}_1^{(t)}$, and note that $\widehat{\mathbf{U}}_1^{(t)}$ satisfies

$$\widehat{\mathbf{U}}_1^{(t)}=\widehat{\mathbf{T}}_1^{(t)}\widehat{\mathbf{U}}_1^{(t)}(\widehat{\mathbf{\Lambda}}_1^{(t)})^{-2},$$

where $\widehat{\mathbf{\Lambda}}_1^{(t)}$ are the singular values of the matricizations of the iterates (and hence $(\widehat{\mathbf{\Lambda}}_1^{(t)})^2$ are the eigenvalues of $\widehat{\mathbf{T}}_1^{(t)}$). With this identity together with the fact that $(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^\top)\mathbf{T}_1=0$ yields the identity

$$\begin{aligned}
e_m^\top\Big(\widehat{\mathbf{U}}_k^{(t)}-\mathbf{U}_k\mathbf{W}_k^{(t)}\Big)&=e_m^\top(\mathbf{I}-\mathbf{U}_k\mathbf{U}_k^\top)\widehat{\mathbf{U}}_k^{(t)}+e_m^\top\mathbf{U}_k(\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)}-\mathbf{W}_k^{(t)})\\
&=e_m^\top(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^\top)\mathbf{Z}_1\big(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\big)\mathbf{T}_1^\top\widehat{\mathbf{U}}_k^{(t)}(\widehat{\mathbf{\Lambda}}_k^{(t)})^{-2}\\
&\quad+e_m^\top(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^\top)\mathbf{Z}_1\big(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\big)\mathbf{Z}_1^\top\widehat{\mathbf{U}}_1^{(t)}(\widehat{\mathbf{\Lambda}}_1^{(t)})^{-2};\\
&\quad+e_m^\top\mathbf{U}_k(\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)}-\mathbf{W}_k^{(t)})\\
&=:e_m^\top\mathbf{L}_1^{(t)}+e_m^\top\mathbf{Q}_1^{(t)}+e_m^\top\mathbf{U}_k(\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)}-\mathbf{W}_k^{(t)}),
\end{aligned}$$

where the first two terms represent the *linear error* and *quadratic error* respectively (dubbed so because each term is linear and quadratic in the noise matrix $\mathbf{Z}_1$ respectively). Again for ease of exposition, consider the linear error. To analyze a row of the linear error, one needs to analyze the vector

$$e_m^\top(\mathbf{I}-\mathbf{U}_1\mathbf{U}_1^\top)\mathbf{Z}_1\big(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\big)\mathbf{T}_1^\top\widehat{\mathbf{U}}_k^{(t)}(\widehat{\mathbf{\Lambda}}_k^{(t)})^{-2}.$$

Ideally we would like to argue that this behaves like a sum of independent random variables in $\mathbf{Z}_1$, but this is not true, as there is nontrivial dependence between $\mathbf{Z}_1$ and the projection matrix $\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}$.

The primary argument behind the leave-one-out analysis technique is to define a se-

quence that is independent from the random variables in $e_m^\top \mathbf{Z}_1$. Hand waving slightly, the prototypical leave-one-out analysis for eigenvector and singular vector analyses (e.g., Cai et al. (2021a); Abbe et al. (2020)) argues that

$$\begin{aligned}\|e_m^\top \mathbf{Z}_1 \widehat{\mathbf{U}}\| &\le \|e_m^\top \mathbf{Z}_1 (\widehat{\mathbf{U}} - \mathbf{U}^{\mathrm{LOO}})\| + \|e_m^\top \mathbf{Z}_1 \mathbf{U}^{\mathrm{LOO}}\| \\ &\lesssim \|\mathbf{Z}_1\| \|\sin\Theta(\widehat{\mathbf{U}}, \mathbf{U}^{\mathrm{LOO}})\| + \|e_m^\top \mathbf{Z}_1 \mathbf{U}^{\mathrm{LOO}}\|,\end{aligned}$$

where $\mathbf{U}^{\mathrm{LOO}}$ is the matrix obtained by zeroing out the m'th row of $\mathbf{Z}_1$. By independence, the second term can be handled via standard concentration inequalities, and the first term is typically handled by standard tools from matrix perturbation theory, which is sufficient as the leave-one-out sequence is extremely close to the true sequence $\widehat{\mathbf{U}}$.

For tensors, this approach suffers from two drawbacks: one is that the spectral norm of $\mathbf{Z}_1$ may be very large in the tensor setting – of order $\sqrt{p_2 p_3} \asymp p$ when $\sigma = 1$. However, this can be managed by carefully using the Kronecker structure that arises naturally in the tensor setting; in fact, Zhang and Xia (2018) showed that

$$\sup_{\|\mathbf{U}_1\|=1, \|\mathbf{U}_2\|=1, \mathrm{rank}(\mathbf{U}_1, \mathbf{U}_2) \le 2r} \|\mathbf{Z}_1 (\mathbf{U}_1 \otimes \mathbf{U}_2)\| \lesssim \sqrt{pr}$$

when $\sigma = 1$, which eliminates the naive upper bound of order p when $r \ll p$. Note that this bound holds only for subgaussian noise, as it relies quite heavily on an ε-net argument, which is in general suboptimal for heavy-tailed or Bernoulli noise[6]. This term turns out to be a major hurdle in analyzing Bernoulli noise; see e.g., Jing et al. (2021), Ke et al. (2020), Yuan and Zhang (2017) for methods to handle similar terms for other types of noise.

The second drawback is a bit more subtle, but it has to do with the leave-one-out sequence definition. Suppose one defines the leave-one-out sequence by removing the m'th row of $\mathbf{Z}_1$ and running HOOI with this new noise matrix, as one may be most tempted to do. Let $\check{\mathbf{U}}_k^{(t)}$ denote the output of this sequence for the k'th mode (with m fixed). Then the $\sin\Theta$ distance between the true sequence and the leave-one-out sequence for mode 1 will

[6] By way of analogy, for an $n \times n$ mean-zero Bernoulli noise matrix $\mathbf{E}$ with entrywise variance at most ρ_n, an ε-net argument only yields $\|\mathbf{E}\| \lesssim \sqrt{n}$, whereas a more refined argument as in Bandeira and Handel (2016) yields $\|\mathbf{E}\| \lesssim \sqrt{n\rho_n}$.

depend on the difference matrix

$$\begin{aligned}\Big(\mathbf{T}_1\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}+\mathbf{Z}_1\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}\Big)-\Big(\mathbf{T}_1\check{\mathbf{U}}_2^{(t-1)}\otimes\check{\mathbf{U}}_3^{(t-1)}+\mathbf{Z}_1^{1-m}\check{\mathbf{U}}_2^{(t-1)}\otimes\check{\mathbf{U}}_3^{(t-1)}\Big)\\ =\mathbf{T}_1\Big((\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)})-(\check{\mathbf{U}}_2^{(t-1)}\otimes\check{\mathbf{U}}_3^{(t-1)})\Big)+\mathbf{Z}_1\Big((\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)})-(\check{\mathbf{U}}_2^{(t-1)}\otimes\check{\mathbf{U}}_3^{(t-1)})\Big)\\ +\begin{pmatrix}\cdots 0\cdots\\ e_m^\top\mathbf{Z}_1\\ \cdots 0\cdots\end{pmatrix}(\check{\mathbf{U}}_2^{(t-1)}\otimes\check{\mathbf{U}}_3^{(t-1)}),\end{aligned}\tag{4.5}$$

where we define $\mathbf{Z}_1^{1-m}$ as the matrix $\mathbf{Z}_1$ with the m'th row set to zero, and the final term is only nonzero in its m'th row. The second two terms (containing $\mathbf{Z}_1$) can be shown to be quite small by appealing to spectral norm bounds together with the Kronecker structure (e.g., Lemma 48). However, the first term depends on both the matrix $\mathbf{T}_1$ and the proximity of the leave-one-out sequence to the true sequence. If one simply bounds this term in the spectral norm, the $\sin\Theta$ distance may end up *increasing* with respect to the condition number κ, and hence may be much larger than the concentration for $\mathbf{Z}_1$ (and may not shrink to zero sufficiently quickly). Therefore, in order to eliminate this problem, we carefully construct a modified leave-one-out sequence $\widetilde{\mathbf{U}}_k^{(t)}$ that can eliminate the dependence on $\mathbf{T}_1$ as follows.

First, we set the initialization $\widetilde{\mathbf{U}}_k^{(0,1-m)}$ as one may expect: $\widetilde{\mathbf{U}}_k^{(0,1-m)}$ are the left singular vectors obtained via the diagonal deletion of $\widehat{\mathcal{T}}$, only with the m'th row of $\mathbf{Z}_1$ set to zero. Let $\mathbf{Z}_k^{1-m}$ denote the matrix $\mathbf{Z}_k$ with the entries associated to the m'th row of $\mathbf{Z}_1$ set to zero (note that in this manner $\mathbf{Z}_k-\mathbf{Z}_k^{1-m}$ will consist of sparse nonzero *columns*). We set $\widetilde{\mathbf{U}}_k^{(0,1-m)}$ as the leading r_k eigenvectors of the matrix

$$\Gamma\big(\mathbf{T}_k\mathbf{T}_k^\top+\mathbf{Z}_k^{1-m}\mathbf{T}_k^\top+\mathbf{T}_k(\mathbf{Z}_k^{1-m})^\top+\mathbf{Z}_k^{1-m}(\mathbf{Z}_k^{1-m})^\top\big),$$

so that $\widetilde{\mathbf{U}}_k^{(0,1-m)}$ is independent from the m'th row of $\mathbf{Z}_1$ (for each k). We now set $\widetilde{\mathbf{U}}_k^{(t,1-m)}$ inductively via

$$\widetilde{\mathbf{U}}_k^{(t,1-m)}:=\mathrm{SVD}_{r_k}\big(\mathbf{T}_k+\mathbf{Z}_k^{1-m}\widetilde{\mathcal{P}}_k^{t,1-m}\big),$$

which is independent from $e_m^\top \mathbf{Z}_1$. Here, we set $\widetilde{\mathcal{P}}_k^{t,1-m}$ inductively as the projection matrix

$$\widetilde{\mathcal{P}}_k^{t,1-m} := \begin{cases} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t-1,1-m)} \otimes \widetilde{\mathbf{U}}_3^{(t-1,1-m)}} & k = 1; \\ \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t-1,1-m)} \otimes \widetilde{\mathbf{U}}_1^{(t,1-m)}} & k = 2; \\ \mathcal{P}_{\widetilde{\mathbf{U}}_1^{(t,1-m)} \otimes \widetilde{\mathbf{U}}_2^{(t,1-m)}} & k = 3, \end{cases}$$

which is the projection matrix corresponding to the previous two iterates of the leave-one-out sequence. Note that this matrix is *still* independent from the m'th row of $\mathbf{Z}_1$, and hence each sequence $\widetilde{\mathbf{U}}_k^{(t,1-m)}$ is independent from the m'th row of $\mathbf{Z}_1$. Moreover, with this choice of matrix of singular vectors, the *projection* matrix $\widetilde{\mathbf{U}}_1^{(t,1-m)}(\widetilde{\mathbf{U}}_1^{(t,1-m)})^\top$ is also the projection onto the dominant left singular space of the matrix

$$\left(\mathbf{T}_1 + \mathbf{Z}_1^{1-m}\widetilde{\mathcal{P}}_1^{t,1-m}\right)\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}$$

as long as an eigengap condition is met (Lemma 36). Then the true sequence and the leave-one-out sequence depend only on the difference matrix

$$\left(\mathbf{Z}_1^{1-m}\widetilde{\mathcal{P}}_1^{t,1-m} - \mathbf{Z}_1\right)(\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}),$$

which depends only on the random matrix $\mathbf{Z}_1$, and hence is approximately mean-zero. In particular, with this careful leave-one-out construction, we can eliminate the extra terms containing $\mathbf{T}_k$ in (4.5) to obtain good bounds on the $\sin\Theta$ distance between the leave-one-out sequence and the true sequence (c.f., Lemma 37).

Finally, the exposition above has focused on the case $k = 1$. Since there are three modes and we prove the result by induction, we actually need to construct *three separate* leave-one-out sequences (each one corresponding to each mode), and we control each of these sequences simultaneously at each iteration. Therefore, at each iteration, we show that (4.4) holds as well as controlling the three separate leave-one-out sequences. Our final proof requires careful tabulation of the probabilities of the events defined by each of these separate sequences. To ease the analysis, we first bound each term deterministically under eigengap conditions, and then further obtain probabilistic bounds by induction using the

leave-one-out sequences.

4.5 Discussion

In this paper, we have considered the tensor mixed-membership blockmodel, which generalizes the tensor blockmodel to settings where communities are no longer discrete. By studying the $\ell_{2,\infty}$ perturbation of the HOOI algorithm, we can obtain a consistent estimator for the memberships provided there are pure nodes along each mode. By applying our proposed algorithm to several different datasets, we have identified phenomena that are not feasible to obtain in the discrete community setting.

It is natural to consider estimating the mixed memberships of the higher-order tensors. Suppose one observes a tensor $\widehat{\mathcal{T}} \in \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_d}$. Our algorithm and methodology naturally extend to this setting, with the only modification being the implementation of the HOOI algorithm, which is straightforward to adapt to the higher-order setting. By adapting our main arguments, we can prove the following informal result.

Theorem 12 (Estimation of mixed memberships for higher-order tensors; informal)**.** *Suppose that $r_{\max} \lesssim p_{\min}^{1/(d-1)}$, that $r_{\max} \asymp r$ with $r \lesssim r_{\min}$, and that $\kappa^2 \lesssim p_{\min}^{1/(2(d-1))}$. Suppose that the smallest singular value of $\mathcal{S}$ satisfies $\Delta^2/\sigma^2 \gtrsim \frac{\kappa^2 p_{\max}^2 r_1 \cdots r_d}{p_1 \cdots p_d \sqrt{p_{\min}^{(d-1)/(d-2)}}}$. Let $\widehat{\mathbf{\Pi}}_k$ be the output of Algorithm 5 (with HOOI adapted to order d) with t iterations for $t \asymp \log\left(\frac{\kappa p_{\max}(r_1 \cdots r_d)^{1/2}}{(\Delta/\sigma)(p_1 \cdots p_d)^{1/2}}\right) \vee 1$. Then with probability at least $1 - p_{\max}^{-10}$, there exist d permutation matrices $\mathcal{P}_k \in \mathbb{R}^{r_k \times r_k}$ such that for each k*

$$\max_{1 \le i \le p_k} \|\left(\mathbf{\Pi}_k - \widehat{\mathbf{\Pi}}_k \mathcal{P}_k\right)_{i\cdot}\| \lesssim_d \frac{\kappa \sqrt{r^d \log(p_{\max})}}{(\Delta/\sigma)(p_{-k})^{1/2}}.$$

Consequently, when $p_k \asymp p$, it holds that

$$\max_{1 \le i \le p_k} \|\left(\mathbf{\Pi}_k - \widehat{\mathbf{\Pi}}_k \mathcal{P}_k\right)_{i\cdot}\| \lesssim_d \frac{\kappa \sqrt{r^d \log(p)}}{(\Delta/\sigma)p^{(d-1)/2}}.$$

Here $a \lesssim_d b$ means that the implicit constant depends on the number of modes d.

As in the order three setting, we see an improvement in the error rate of order $\sqrt{p}$ for

each additional mode, albeit at the cost of a slightly stronger signal-strength condition and condition on r. In future work it may be interesting to determine the dependence of the implicit constants on the order d and to study the regime where $d \to \infty$.

In other future work, it may be natural to extend the mixed-membership tensor block-model to allow degree corrections as in Jin et al. (2019) or Hu and Wang (2022). Furthermore, the analysis in this paper focuses on subgaussian noise, whereas many multilayer network datasets have Bernoulli noise, as well as different types of symmetry, and a natural extension would encompass noise and structures of this form, and would also perhaps include missingness.

Beyond these natural extensions of the model it is of interest to extend the $\ell_{2,\infty}$ perturbation theory covered herein to other regimes of signal strength, as well as provide entrywise bounds for other types of low-rank structures beyond the Tucker low-rank assumption. Furthermore, it may be relevant to develop distributional theory for the outputs of tensor SVD, and to obtain principled confidence intervals for the outputs of HOOI.

Chapter 5

Nonparametric Two-Sample Hypothesis Testing for Random Graphs with Negative and Repeated Eigenvalues

5.1 Introduction

Network data arises naturally in several fields, including neuroscience (Bullmore and Sporns, 2009; Bullmore and Bassett, 2011; Vogelstein et al., 2013; Finn et al., 2015; Priebe et al., 2019; Arroyo-Relión et al., 2019) and social networks (Newman et al., 2002; Newman, 2006; Carrington et al., 2005) among others. With the introduction of network data in the various sciences, there is a need for developing corresponding statistical methodology and theory. Often one wishes to determine whether or not two graphs exhibit similar distributional properties for some notion of similarity between distributions on networks. Furthermore, as in classical statistics, one may wish to analyze graph data with only a few assumptions on the probability distributions. For example, for Euclidean data, given i.i.d. observations $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m \in \mathbb{R}^d$ with cumulative distribution functions denoted F_X and F_Y respectively, a model-agnostic way to test whether $F_X = F_Y$ is use nonparametric methods, such as Mood

(1954); Anderson et al. (1994); Gretton et al. (2012); Székely and Rizzo (2013), and Chen and Friedman (2017).

For the two-sample test we consider, we take the perspective that a single network constitutes an observation as is often the case in the statistical network analysis literature. A number of works have studied hypothesis testing when the graphs are on the same vertex set (Tang et al., 2017a; Li and Li, 2018; Levin et al., 2019; Ghoshdastidar et al., 2020; Draves and Sussman, 2020), analogous to the "matched pairs" paradigm in Euclidean data. However, in many settings, there is not necessarily an *a priori* matching between vertices; for example, one could introduce and remove vertices without fundamentally altering the network structure (e.g. Cai and Li (2015)). In this paper we study a nonparametric two-sample test without assuming that the graphs have an alignment between the vertices.

We assume only that the two networks have conditionally independent edges and that the graphs' edge-probability matrices are low rank (see Section 5.2.1 for a formal description). We consider a latent-space model (Hoff et al., 2002) introduced in Rubin-Delanchy et al. (2020), wherein each vertex has a latent vector in Euclidean space associated to it. The latent-space framework is specific enough to allow for a meaningful notion of similarity between graphs on different vertex sets, and it general enough to allow for arbitrary low-rank graphs. Low rank, conditionally edge-independent random graphs include a number of submodels including the stochastic blockmodel (Holland et al., 1983), the random dot product graph (Athreya et al., 2018), and finite-rank graphons (Lovász, 2012). In addition, other tests have been designed for fixed models and related problems, such as Lei (2016); Bickel and Sarkar (2016) and Fan et al. (2022). A more thorough discussion of related literature is in Section 5.3.3.

A major difficulty in allowing for negative eigenvalues in the graphs requires an understanding the relationship between the latent-space Euclidean geometry and the indefinite geometry induced by the negative eigenvalues. Nevertheless, we show that despite the underlying geometry, consistent testing can be performed given access only to the adjacency matrices. More specifically, we show that a test procedure based on a two-sample U-statistic with radial kernel κ applied to rotated adjacency spectral embeddings yields a consistent test. Since we only assume that the graphs have a low-rank structure, our analysis includes

random graphs whose edge-probability matrices have negative eigenvalues and a vanishing eigengap amongst the nonzero eigenvalues, and we conduct our study under different sparsity regimes. In particular, our proposed test procedure and its theoretical properties depend on a careful analysis of the interplay between indefinite orthogonal transformations, optimal transport, and convergence of degenerate U-statistics.

We further show that under the null hypothesis, for sufficiently dense graphs, the non-degenerate limiting distribution of our test statistic can be related to that of the U-statistic evaluated at suitably transformed latent vectors, and we provide additional results for sparser graphs. The convergence of our test statistic is analogous to that of Anderson et al. (1994); Gretton et al. (2012) in the Euclidean setting. Furthermore, our sparsity requirement is consistent with a number of works on network bootstraps for nondegenerate U-statistics (Levin and Levina, 2019; Lunde and Sarkar, 2019; Zhang and Xia, 2020; Lin et al., 2020a,b), which occur as subgraph frequencies. An important aspect of our results is that our test statistic is a *degenerate* (two-sample) U-statistic (e.g. Serfling (1980)).

The paper is organized as follows. In the following subsection, we motivate the problem more thoroughly, and in Section 5.2, we give the relevant definitions and describe our setting. In Section 5.3 we state our main theoretical results for sparse, indefinite random graphs with negative and repeated eigenvalues and describe a modification to handle repeated eigenvalues. In Section 5.4 we show our results on simulated data, and in Section 5.5 we discuss our results. Section E.1 contains the proofs of our main results.

Notation: We use capital letters to denote random vectors $X \in \mathbb{R}^d$, bold lowercase letters to denote fixed vectors, and bold capital letters for fixed or random matrices (which will be clear from context). The distribution of a random vector X will be denoted by F_X, and for $X_1, \ldots, X_n$ i.i.d. some distribution F_X, we use $\mathbf{X}$ to denote the $n \times d$ matrix with its rows the vectors $X_1, \ldots, X_n$. In many occasions, given $X_1, \ldots, X_n$ i.i.d. F_X, we let X denote a realization from F that is independent from $\{X_i\}_{i=1}^n$. We write $\|\cdot\|$ for the usual Euclidean norm on vectors and the spectral norm on matrices and $\|\cdot\|_F$ for the Frobenius norm. For a matrix $\mathbf{M}$ we write its ℓ_2 to ℓ_∞ operator norm via $||\mathbf{M}||_{2,\infty} \equiv \max_i ||\mathbf{M}_{i\cdot}||$, where $\mathbf{M}_{i\cdot}$ are the rows of $\mathbf{M}$. We use $\mathbf{I}_k$ to denote the $k \times k$ identity matrix. For a matrix $\mathbf{M}$, the operator

$\text{diag}(\mathbf{M})$ extracts its diagonal as a matrix, and for two matrices $\mathbf{M}$ and $\mathbf{N}$, the operator $\text{bdiag}(\mathbf{M}, \mathbf{N})$ constructs the block-diagonal matrix $\begin{pmatrix} \mathbf{M} & 0 \\ 0 & \mathbf{N} \end{pmatrix}$. We write $f(n) = O(g(n))$ if there exists a constant $C > 0$ such that $f(n) \leq Cg(n)$ for all n sufficiently large, and $f(n) = \omega(g(n))$ if there exists a constant $c > 0$ such that $cg(n) \leq f(n)$ for all n sufficiently large. We also write $f(n) \gg g(n)$ if $g(n)/f(n) \to 0$ as $n \to \infty$.

5.1.1 Motivating Example

Suppose there are n and m vertices in two different graphs respectively, and suppose the vertices can be partitioned into three disjoint sets, called communities, where each vertex belongs to community k, $k \in \{1, 2, 3\}$, with probability $1/3$. Suppose further that for vertices in the same community, the probability of connection is a and for vertices in different communities the probability of connection is b; such a model is referred to as the *three-block balanced homogenous stochastic blockmodel* in the literature. The matrix

$$\mathbf{B} := \begin{pmatrix} a & b & b \\ b & a & b \\ b & b & a \end{pmatrix}$$

has three eigenvalues; one always positive eigenvalue of $a + 2b$ and a repeated eigenvalue $a - b$, which is negative when $b > a$. Let $\mathbf{Z}^{(1)} \in \{0, 1\}^{n \times 3}$ be the matrix such that $\mathbf{Z}^{(1)}_{ik} = 1$ if vertex i belongs to community k and 0 otherwise, and similarly for $\mathbf{Z}^{(2)} \in \{0, 1\}^{m \times 3}$. Define

$$\mathbf{P}^{(1)} := \mathbf{Z}^{(1)}\mathbf{B}(\mathbf{Z}^{(1)})^\top; \qquad \mathbf{P}^{(2)} := \mathbf{Z}^{(2)}\mathbf{B}(\mathbf{Z}^{(2)})^\top.$$

Now, consider the eigendecomposition of $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$. In the setting that exactly $1/3$ of the vertices belong to each community respectively, the eigenvalues are simply scaled by $n/3$ or $m/3$ for graphs one and two respectively. Furthermore, if one scales the eigenvectors by the square roots of the absolute values of these eigenvalues, by viewing each row as a point, one obtains three distinct points on $\mathbb{R}^3$ that remain constant in n and m. For example, the

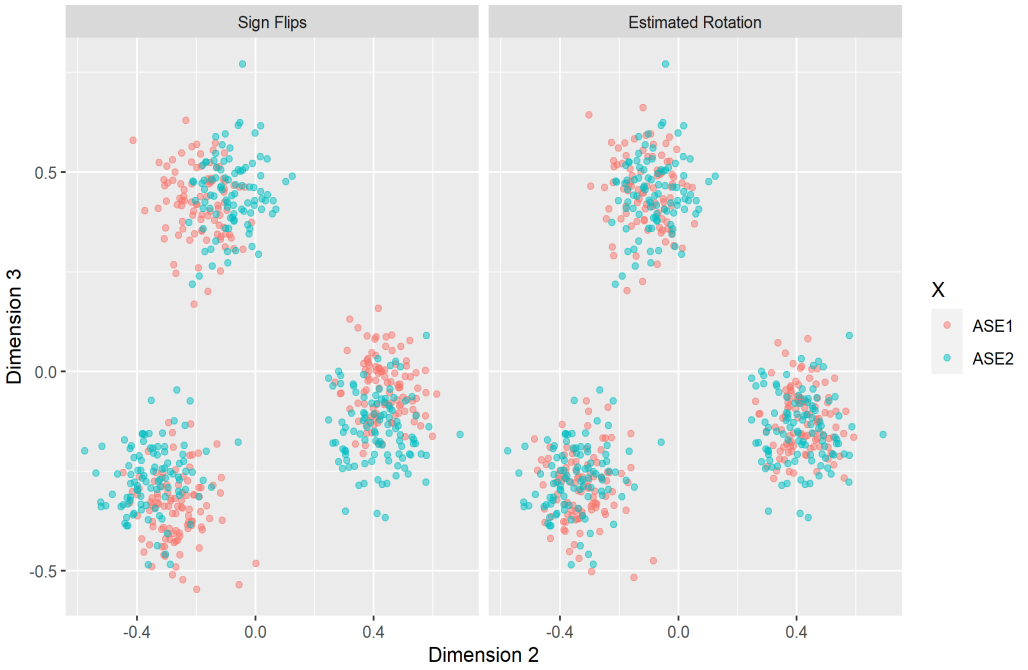


Figure 5.1: Comparisons of the naïve sign-flip alignment procedure (left) and the optimal transport alignment procedure (right) for two adjacency spectral embeddings for the stochastic blockmodel. On the left hand side, we see that that visually the clusters do not lie on top of each other, and on the right hand side, the clusters appear to lie on top of each other.

first eigenvector scaled by the $\sqrt{\frac{n(a+2b)}{3}}$ yields the vector whose entries are all $\sqrt{\frac{a+2b}{3}}$.

Elementary copmputation shows that the second and third eigenvector correspond to the same eigenvalue $\frac{n(a-b)}{3}$ and correspond to a two-dimensional subspace. Hence, even though scaling the eigenvector by $\sqrt{\frac{n|a-b|}{3}}$ (or $\sqrt{\frac{m|a-b|}{3}}$) yields a term that does not change as n and m increase, it is not defined uniquely because of the repeated eigenvalue. In fact, since the second and third eigenvectors correspond to any choice of basis for the subspace corresponding to the eigenvalue $\frac{n(a-b)}{3}$, one can arbitrarily rotate the second and third eigenvectors by any 2×2 orthogonal transformation and still obtain eigenvectors.

Suppose one observes two graphs $\mathbf{A}^{(1)} \in \{0,1\}^{n \times n}$ and $\mathbf{A}^{(2)} \in \{0,1\}^{m \times m}$ with $\mathbf{A}^{(1)}$ independent of $\mathbf{A}^{(2)}$, where each $\mathbf{A}^{(1)}_{ij} \sim \text{Bernoulli}(\mathbf{P}^{(1)}_{ij})$ for $i \leq j$, with $\mathbf{A}^{(1)}_{ij} = \mathbf{A}^{(1)}_{ji}$ for $j \leq i$ and similarly for $\mathbf{A}^{(2)}$. A common way to estimate the scaled eigenvectors of the matrices $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ given observed graphs $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ is the *adjacency spectral embedding*, which here is just the leading 3 eigenvectors scaled by the square roots of the absolute values of their eigenvalues (see Definition 4). Even if the adjacency spectral embedding is

consistent, it will only be consistent up to some transformation that takes into account the nonidentifiability of the second and third eigenvectors.

As we will make clear in Section 5.2.1, when we refer to (in)equality in distribution we are referring to the fact that two stochastic blockmodels could be different in both the block assignment probabilities and the block probabilities matrix $\mathbf{B}$. If one knew an *a priori* correspondence between the vertices, one could simply perform orthogonal Procrustes, which has a closed-form solution. Unfortunately, for two graphs of different sizes, there are many situations in which one need not have a correspondence between the graphs. Therefore, though the two graphs do not have an *a priori* alignment, under the null hypothesis that the two graphs have the same distribution there is a block-orthogonal matrix $\mathbf{W}$ that will approximately align the supports of the empirical distributions of the rows of the adjacency spectral embeddings (see Proposition 5).

Motivated by this problem and the literature on optimal transport, we show that estimating the orthogonal matrix by aligning the *support* of the empirical distributions suffices to obtain consistency; in particular, we propose minimizing the *Orthogonal Wasserstein Distance*, which outputs a block-orthogonal matrix $\widehat{\mathbf{W}}$ (see Section 5.3.2). This remedies the nonidentifiability of the second and third eigenvectors above.

For a generic eigenvector, if the corresponding eigenvalue has multiplicity one, then the only freedom in selecting the eigenvector is the choice of sign. In general, ignoring the orthogonal transformations would yield a test statistic that minimizes over all possible sign-flip combinations. In Figure 5.1, we plot the second and third dimensions of the adjacency spectral embeddings for adjacency matrices simulated from a stochastic blockmodel with connectivity matrix

$$\mathbf{B} = \begin{pmatrix} .5 & .8 & .8 \\ .8 & .5 & .8 \\ .8 & .8 & .5 \end{pmatrix},$$

and probability of community membership $\frac{1}{3}$ for each community. In the left figure, we plot the second and third dimensions of the adjacency spectral embeddings for two graphs on $n = 300$ vertices, where the first embedding is rotated using only the best sign flip. Here

"best" corresponds to the minimum value of the test statistic. In the right figure, we plot the second and third dimensions of the two embeddings after using the optimal transport-based alignment we outline in Section 5.3.2.

From a purely visual standpoint, when the distributions for each graph are the same, the empirical distributions should lie approximately on top of one another; however, we see that sign flips fail to recover this correspondence. The left hand side shows visually how the second and third dimensions are not approximately aligned, and the right hand illustrates how the supports of the distributions approximately lie on top of one another, showing that estimating the rotation approximately recovers the implicit distributional correspondence. This figure demonstrates an important point for spectral methods in statistical network analysis: simply ignoring repeated eigenvalues could yield inconsistent testing. Section 5.4 contains further simulations and quantitative comparisons under more general model settings.

5.2 Preliminaries

We will now situate the hypothesis test described in the previous section in the general setting in which we will be performing our hypothesis test.

5.2.1 Setting

We use the latent position framework of the *generalized random dot product graph* proposed in Rubin-Delanchy et al. (2020) and closely related to that in Lei (2020b). First, we discuss the notion of a (p, q) admissible distribution. In what follows, the matrix $\mathbf{I}_{p,q}$ is defined as $\mathbf{I}_{p,q} := \text{diag}(\mathbf{I}_p, -\mathbf{I}_q)$.

Definition 1. We say F_X with support $\Omega \subset \mathbb{R}^d$ is a (p, q) admissible distribution if for all $\mathbf{x}, \mathbf{y} \in \Omega$, $\mathbf{x}^\top \mathbf{I}_{p,q} \mathbf{y} \in [0, 1]$.

For a fixed (p, q) admissible distribution, we consider the generalized random dot product graph as follows.

Definition 2 (Rubin-Delanchy et al. (2020))**.** We say a graph $\mathbf{A} \in \{0,1\}^{n\times n}$ is a *generalized random dot product graph on n vertices* with (p,q)-admissible distribution F_X, sparsity factor α_n, and latent positions $\mathbf{X}$ if the matrix $\mathbf{A}$ is symmetric, and the entries $\mathbf{A}_{ij}$ are conditionally independent given $\mathbf{X}$ and Bernoulli random variables with

$$\mathbb{P}(\mathbf{A}_{ij} = 1|\mathbf{X}) = \alpha_n X_i^\top \mathbf{I}_{p,q} X_j$$

with $\mathbf{A}_{ij} = \mathbf{A}_{ji}$, and $X_1, \ldots, X_n \sim F_X$ are i.i.d. We write $(\mathbf{A}, \mathbf{X}) \sim GRDPG(F_X, n, \alpha_n)$.

The introduction of the matrix $\mathbf{I}_{p,q}$ is to model large in magnitude negative eigenvalues of the adjacency matrix. The GRDPG model allows for arbitrary low-rank models, so is sufficiently agnostic to provide a meaningful setting for nonparametric inference. In a nonparametric setting, the parameters for the GRDPG model are simply the signature (p,q), the sparsity parameter α_n, and the distribution F_X (which may or may not be parametric). For this work, we assume the signature (p,q) is known. Practically speaking, this model is equivalent to assuming only that the probability generating matrix is low-rank, and there are several works showing that low-rank models can approximate full-rank models arbitrarily well (Tang et al., 2013; Xu, 2018; Udell and Townsend, 2019; Lei, 2020b; Rubin-Delanchy, 2020). Finally, in the setting $q = 0$, one recovers the *random dot product graph* (RDPG) model (Athreya et al., 2018), which assumes a low rank, positive semidefinite probability matrix.

One potential issue with the GRDPG model is that it exhibits nonidentifiability. To be explicit, suppose $\mathbf{Q}$ is a matrix such that $\mathbf{Q}\mathbf{I}_{p,q}\mathbf{Q}^\top = \mathbf{I}_{p,q}$ (this is known as the *indefinite orthogonal group* $\mathbb{O}(p,q)$). Define the distribution $\widetilde{F}_X := F_X \circ \mathbf{Q}$, where $F_X \circ \mathbf{Q}$ means that one generates $X_i \sim F_X$ and then left multiplies the vectors X_i by $\mathbf{Q}^\top$. Then the probabilities

of each edge are fixed since

$$\begin{aligned}\mathbb{P}_{\widetilde{F}_X}(\mathbf{A}_{ij} = 1|\mathbf{X}) &= \alpha_n(\mathbf{Q}^\top X_i)^\top \mathbf{I}_{p,q}(\mathbf{Q}^\top X_j) \\ &= \alpha_n X_i^\top(\mathbf{Q}\mathbf{I}_{p,q}\mathbf{Q}^\top)X_j \\ &= \alpha_n X_i^\top \mathbf{I}_{p,q} X_j \\ &= \mathbb{P}_{F_X}(\mathbf{A}_{ij} = 1|\mathbf{X}).\end{aligned}$$

Hence, the distribution of the graph remains unchanged if one transforms the support of F_X by any indefinite orthogonal transformation $\mathbf{Q}$. Therefore, any nonparametric test of equality of distribution must allow equality up to indefinite orthogonal transformations. This motivates the following definition.

Definition 3. Let F_X and F_Y be two (p, q) admissible distributions. We say F_X and F_Y are *equal up to indefinite orthogonal transformation* if there exists a matrix $\mathbf{Q} \in \mathbb{O}(p, q)$ such that

$$F_X = F_Y \circ \mathbf{Q}.$$

In this case, we write $F_X \simeq F_Y$.

We note that in the RDPG model, $F_X \simeq F_Y$ is equivalent to saying the distributions are equivalent up to orthogonal transformation, since when $q = 0$ the nonidentifiability is of the form of orthogonal matrices.

We are now ready to formally describe our hypothesis test under the generalized random dot product graph framework. Suppose we observe two graph adjacency matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ such that $(\mathbf{A}^{(1)}, \mathbf{X}) \sim GRDPG(F_X, n, \alpha_n)$ and $(\mathbf{A}^{(2)}, \mathbf{Y}) \sim GRDPG(F_Y, m, \beta_m)$ are mutually independent and have the same signature (p, q). We consider the hypothesis test

$$\begin{aligned}H_0 &: F_Y \simeq F_X \\ H_A &: F_Y \not\simeq F_X.\end{aligned}$$

Again, we assume throughout that (p, q) is known and fixed in n and m. In general, we

do not assume (α_n, β_m) are known, but, for ease of exposition we shall first assume that they are known. They can be estimated consistently, so we will revisit these issues later (see Corollary 6).

Remark 17 (Equivalence to Section 5.1.1)**.** *Although the above hypothesis test seems to suffer from a lack of identifiability, the nonidentifiability is primarily an artifact of working in the framework of the GRDPG model. Were we to reformulate the test in Section 5.1.1 in terms of the stochastic block model with generating matrices* $\mathbf{B}^{(1)}$ *and* $\mathbf{B}^{(2)}$ *and probability vectors* $\pi^{(1)}$ *and* $\pi^{(2)}$*, this test would just be determining whether both* $\mathbf{B}^{(1)} = \mathbf{B}^{(2)}$ *and* $\pi^{(1)} = \pi^{(2)}$ *up to permutation of the communities. To see the explicit equivalence, one can transform any stochastic blockmodel with connectivity matrix* $\mathbf{B}$ *into a GRDPG model by letting* $\mathbf{B} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ *and fixing* K *vectors* $\nu_1, \ldots, \nu_K$ *as the rows* $\mathbf{V}|\mathbf{D}|^{\mathbf{1/2}}$*, where* $\mathbf{V}$ *is chosen arbitrarily if there are repeated eigenvalues. Then the GRDPG model in question is just a mixture of point masses corresponding to the entries of* π*, where the vectors are the rows of* $\mathbf{V}|\mathbf{D}|^{\mathbf{1/2}}$.

This test also allows one graph to be a submodel of the other. For example, this test includes the situation that $\mathbf{A}^{(1)}$ comes from a stochastic blockmodel and $\mathbf{A}^{(2)}$ comes from a model that may be broader, such as the mixed-membership stochastic blockmodel.

In practice, one observes only the graph adjacency matrix, and therefore must estimate the latent position matrix $\mathbf{X}$. The statistical properties of the scaled eigendecomposition, referred to as the *adjacency spectral embedding* (ASE), are investigated in Rubin-Delanchy et al. (2020). The definition is given below.

Definition 4 (Adjacency Spectral Embedding)**.** Suppose $(\mathbf{A}, \mathbf{X}) \sim GRDPG\ (F_X, n, \alpha_n)$, and write the eigendecomposition of $\mathbf{A}$ as $\sum_{i=1}^{n} \lambda_i \mathbf{u_i} \mathbf{u_i}^\top$, where the λ_i are ordered by magnitude; $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$, and the $\mathbf{u_i}$ are orthonormal. Form the $d \times d$ matrix $\mathbf{\Lambda_A}$ by taking the d largest (in magnitude) eigenvalues of $\mathbf{A}$ sorted by positive and then negative components, and the $n \times d$ matrix $\mathbf{U_A}$ with columns consisting of the eigenvectors associated to the eigenvalues in $\mathbf{\Lambda_A}$. The *adjacency spectral embedding* of $\mathbf{A}$ is the $n \times d$ matrix

$$\widehat{\mathbf{X}} := \mathbf{U_A}|\mathbf{\Lambda_A}|^{\mathbf{1/2}},$$

where the operator $|\cdot|$ takes the absolute values of the entries.

5.2.2 A Kernel Estimator

To describe our test statistic, we must first define *mean embedding* of a distribution. Consider a symmetric positive-definite kernel $\kappa(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with associated reproducing kernel Hilbert space $\mathcal{H}$. The *mean embedding* of a distribution function F with support Ω is defined via

$$\mu[F] := \int_\Omega \kappa(\cdot, x) dF(x).$$

A kernel κ is called *characteristic* (Sriperumbudur et al., 2011) if the embedding $\mu[F]$ is injective, so that $F = G$ if and only if $\mu[F] = \mu[G]$. Examples of characteristic kernels include the Gaussian kernel $\kappa(x,y) = \exp\left(-\frac{1}{\sigma^2}\|x-y\|^2\right)$ and the Laplace kernel $\kappa(x,y) = \exp\left(-\frac{1}{\sigma}\|x-y\|_1\right)$.

Since κ is a function of two variables, given independent samples $\mathbf{X} = \{X_i\}_{i=1}^n$ and $\mathbf{Y} = \{Y_j\}_{j=1}^m$, we define the (two-sample) U-statistic

$$U_{n,m}(\mathbf{X},\mathbf{Y}) := \frac{1}{n(n-1)} \sum_{j \neq i} \kappa\left(X_i, X_j\right) - \frac{2}{mn} \sum_{i=1}^n \sum_{k=1}^m \kappa\left(X_i, Y_k\right) + \frac{1}{m(m-1)} \sum_{l \neq k} \kappa\left(Y_k, Y_l\right).$$

In the asymptotic regime that n and m tend to infinity, under the assumption $\frac{m}{n+m} \to \rho \in (0,1)$, Gretton et al. (2012) showed that

$$U_{n,m}(\mathbf{X},\mathbf{Y}) \to \|\mu[F_X] - \mu[F_Y]\|_{\mathcal{H}}^2$$

almost surely, and, when κ is characteristic, then $\|\mu[F_X] - \mu[F_Y]\|_{\mathcal{H}}^2 = 0$ if and only if $F_X = F_Y$. Moreover, they showed that $(n+m)U_{n,m}(\mathbf{X},\mathbf{Y})$ has a nondegenerate limiting distribution under the null hypothesis $F_X = F_Y$. The scaling by $(n+m)$ is due to the fact that the U-statistic is degenerate, where degeneracy of a U-statistic with kernel h of two variables means that $\mathbb{E}_{F_X}(h(X,\cdot))$ is constant. See, for example, Serfling (1980) for more details on the theory of degenerate U-statistics.

5.3 Hypothesis Testing With Negative and Repeated Eigenvalues

We now present a detailed asymptotic analysis of our two-sample test statistic. Given two graphs $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ on n and m vertices respectively our test statistic is defined via

$$U_{n,m}(\widehat{\mathbf{X}}, \widehat{\mathbf{Y}}) := \frac{1}{n(n-1)} \sum_{j \neq i} \kappa\left(\widehat{X}_i, \widehat{X}_j\right) - \frac{2}{mn} \sum_{i=1}^{n} \sum_{k=1}^{m} \kappa\left(\widehat{X}_i, \widehat{Y}_k\right) + \frac{1}{m(m-1)} \sum_{l \neq k} \kappa\left(\widehat{Y}_k, \widehat{Y}_l\right),$$

where $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ are the adjacency spectral embeddings of $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ respectively. In what follows, all of our asymptotic results are stated as n and m tend to infinity. We will require some assumptions on the kernel κ.

Assumption 1. The kernel κ is characteristic, radial, and twice continuously differentiable on $\mathbb{R}^d$.

The assumption that κ is radial is so that our results can be expressed in terms of individual orthogonal matrices that may themselves be products of several orthogonal matrices. For example, we have the identity $U(\widehat{\mathbf{X}}\mathbf{W}_1, \widehat{\mathbf{Y}}\mathbf{W}_2) = U(\widehat{\mathbf{X}}\widetilde{\mathbf{W}}, \widehat{\mathbf{Y}})$, where $\widetilde{\mathbf{W}} = \mathbf{W}_1\mathbf{W}_2^\top$. Our main results can be modified slightly to hold without this assumption at the penalty of introducing more orthogonal matrices. Differentiability is a relatively mild requirement, and the assumption of κ being characteristic is satisfied by continuous kernels whose embeddings are dense in $\mathcal{H}$ equipped with the supremum norm, since the support of F_X and F_Y can be taken to be compact (see Theorem 26), and hence any universal kernel defined on $\mathbb{R}^d$ is characteristic for the problem we consider herein.

Since real-world graphs are sparse, we conduct a more thorough study of our test statistic under sparsity. First, we make assumptions on the sparsity for which our more general results hold. We implicitly assume that either $\alpha_n, \beta_n \to 0$ or that $\alpha_n \equiv \beta_m \equiv 1$, since if α_n or β_m are converging to some constant greater than zero, one can just rescale the distribution F_X or F_Y.

Assumption 2a. The sparsity parameters for the graphs satisfy

$$\min(n\alpha_n, m\beta_m) = \omega(\log^4(n)),$$

and

$$\frac{m\beta_m}{m\beta_m + n\alpha_n} \to \rho \in (0,1).$$

If instead we have a slightly denser graph, we make the following assumption.

Assumption 2b. The sparsity parameters for the graphs satisfy for some $\eta > 0$,

$$\min(n\alpha_n, m\beta_m) = \omega(n^{1/2}\log^{1+\eta}(n)),$$

and

$$\frac{m}{m+n} \to \rho \in (0,1).$$

In both asymptotic regimes, there are two competing factors: the first is in the approximation of the *unperturbed* U-statistic to the population value; that is, the U-statistic obtained given access to the latent vectors $X_1, \ldots, X_n$, and the second is in the approximation of the *estimated* U-statistic to the unperturbed U-statistic. In the first asymptotic regime, the primary difficulty stems from the approximation of $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ to $\mathbf{X}$ and $\mathbf{Y}$ (up to indefinite orthogonal transformation) because the graph sparsity makes estimation difficult. In the second asymptotic regime, the primary difficulty comes from the approximation of the U-statistic to the maximum mean discrepancy between two appropriately defined distributions. The asymptotic regime in Assumption 2b up to the logarithmic term has been assumed in the literature (Tang et al., 2017c; Jones and Rubin-Delanchy, 2021) and is a common assumption in the theory of bootstrapped U-statistics for random graphs, particularly as they pertain to subgraph counts. See Levin and Levina (2019); Lunde and Sarkar (2019); Zhang and Xia (2020); Lin et al. (2020a), and Lin et al. (2020b) for details.

When considering negative eigenvalues, if one uses the GRDPG model framework, one

necessarily has to contend with indefinite orthogonal transformations. From the equation $\mathbf{Q}\mathbf{I}_{p,q}\mathbf{Q}^\top = \mathbf{I}_{p,q}$ we have that $|\det(\mathbf{Q})| = 1$, and hence $\mathbf{Q}$ is invertible and $\mathbf{Q}^{-1} \in \mathbb{O}(p,q)$ as well. We also note that the set $\mathbb{O}(p,q)$ includes block-diagonal orthogonal matrices; i.e. if we have $\mathbf{W}_p$ and $\mathbf{W}_q$ for $p \times p$ and $q \times q$ orthogonal matrices, then

$$\begin{pmatrix} \mathbf{W}_p & 0 \\ 0 & \mathbf{W}_q \end{pmatrix} \mathbf{I}_{p,q} \begin{pmatrix} \mathbf{W}_p^\top & 0 \\ 0 & \mathbf{W}_q^\top \end{pmatrix} = \begin{pmatrix} \mathbf{W}_p\mathbf{W}_p^\top & 0 \\ 0 & -\mathbf{W}_q\mathbf{W}_q^\top \end{pmatrix} = \mathbf{I}_{p,q}.$$

We refer to the subgroup $\mathbb{O}(p,q) \cap \mathbb{O}(d)$ as the subgroup of block-orthogonal matrices. Note that $\|\mathbf{Q}\| = 1$ for any block-orthogonal $\mathbf{Q}$, whereas for any finite $M > 0$, there exists $\mathbf{Q} \in \mathbb{O}(p,q) \setminus \mathbb{O}(d)$ with $\|\mathbf{Q}\| > M$.

Therefore, allowing for negative eigenvalues involves studying matrices $\mathbf{Q} \in \mathbb{O}(p,q)$ that could be badly behaved (in a spectral norm sense). Nevertheless, using the limiting results in Agterberg et al. (2020b), by subtly passing between indefinite and Euclidean geometry, we can show that when using the adjacency spectral embeddings, one does not even have to consider indefinite orthogonal matrices. Our first proposition shows that the U-statistic applied to the rows of the adjacency spectral embedding yields a consistent test. All proofs are deferred to Section E.1.

Proposition 4. *Let* $(\mathbf{A}^{(1)}, \mathbf{X}) \sim GRDPG(F_X, n, \alpha_n)$ *and* $(\mathbf{A}^{(2)}, \mathbf{Y}) \sim GRDPG(F_Y, m, \beta_m)$ *be independent. Suppose Assumption 2a or Assumption 2b is satisfied, and suppose further that* κ *satisfies Assumption 1. Set* $\mathbf{\Delta_X} := \mathbb{E}(XX^\top)$*, and similarly for* $\mathbf{\Delta_Y}$*. Suppose that both* $\mathbf{\Delta_X}\mathbf{I}_{p,q}$ *and* $\mathbf{\Delta_Y}\mathbf{I}_{p,q}$ *have distinct eigenvalues. Then*

$$U_{n,m}(\widehat{\mathbf{X}}/\alpha_n^{1/2}, \widehat{\mathbf{Y}}/\beta_m^{1/2}) \to \begin{cases} 0 & F_X \simeq F_Y \\ c > 0 & F_X \not\simeq F_Y, \end{cases}$$

almost surely.

Our main requirements are the uniqueness of a certain indefinite orthogonal matrix from Agterberg et al. (2020b), which is given by the assumption that $\mathbf{\Delta_X}\mathbf{I}_{p,q}$ and $\mathbf{\Delta_Y}\mathbf{I}_{p,q}$ have distinct eigenvalues which corresponds to distinct eigenvalues of $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$. Without

distinct eigenvalues, one is still able to obtain consistency up to a block orthogonal transformation.

Proposition 5. *Let* $(\mathbf{A}^{(1)}, \mathbf{X}) \sim GRDPG(F_X, n, \alpha_n)$ *and* $(\mathbf{A}^{(2)}, \mathbf{Y}) \sim GRDPG(F_Y, m, \beta_m)$ *be independent. Suppose Assumption 2a or Assumption 2b is satisfied, and suppose further that* κ *satisfies Assumption 1. If* $F_X \simeq F_Y$*, there exists a sequence of block orthogonal matrices* $\widehat{\mathbf{W}}_n \in \mathbb{O}(p,q) \cap \mathbb{O}(d)$ *such that*

$$U_{n,m}(\widehat{\mathbf{X}}\widehat{\mathbf{W}}_n/\alpha_n^{1/2}, \widehat{\mathbf{Y}}/\beta_m^{1/2}) \to 0$$

almost surely. However, if $F_X \not\simeq F_Y$*, then for any sequence of orthogonal matrices* $\widehat{\mathbf{W}}_n \in \mathbb{O}(p,q) \cap \mathbb{O}(d)$*, there exists a constant* $C > 0$ *depending only on* F_X *and* F_Y *such that almost surely*

$$\liminf_{n,m} U_{n,m}(\widehat{\mathbf{X}}\widehat{\mathbf{W}}_n/\alpha_n^{1/2}, \widehat{\mathbf{Y}}/\beta_m^{1/2}) \geq C.$$

We emphasize that these results are desirable since the matrices $\widehat{\mathbf{W}}_n$ are block orthogonal matrices and not indefinite orthogonal matrices, and hence the only estimation required is the matrix $\widehat{\mathbf{W}}_n$. Proposition 5 suggests that if one can estimate the matrices $\widehat{\mathbf{W}}_n$ consistently, then we can devise a consistent test procedure through a permutation test and bootstrapping the test statistic distribution.

5.3.1 Main Results

Our main results include a more refined study of our test statistic. Define

$$\mathbf{P}^{(1)} := \alpha_n \mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top; \qquad \mathbf{P}^{(2)} := \beta_m \mathbf{Y}\mathbf{I}_{p,q}\mathbf{Y}^\top,$$

and let $\mathbf{U}_\mathbf{X}\mathbf{\Lambda}_\mathbf{X}\mathbf{U}_\mathbf{X}^\top$ and $\mathbf{U}_\mathbf{Y}\mathbf{\Lambda}_\mathbf{Y}\mathbf{U}_\mathbf{Y}^\top$ be their respective eigendecompositions, with $\mathbf{\Lambda}_\mathbf{X}$ and $\mathbf{\Lambda}_\mathbf{Y}$ arranged with the p positive eigenvalues first and q negative eigenvalues second. Define

$$\widetilde{\mathbf{X}} := \mathbf{U}_\mathbf{X}|\mathbf{\Lambda}_\mathbf{X}|^{1/2}; \qquad \widetilde{\mathbf{Y}} := \mathbf{U}_\mathbf{Y}|\mathbf{\Lambda}_\mathbf{Y}|^{1/2}.$$

The matrices $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ can be viewed as surrogates for the matrices $\alpha_n^{1/2}\mathbf{X}$ and $\beta_m^{1/2}\mathbf{Y}$ up to indefinite orthogonal transformation. In fact, the proof of Proposition 4 reveals that under the distinct eigenvalues assumption there exists a block-orthogonal matrix $\mathbf{W}_n$ such that

$$U_{n,m}(\widehat{\mathbf{X}}/\sqrt{\alpha_n}, \widehat{\mathbf{Y}}/\sqrt{\beta_m}) - U_{n,m}(\widetilde{\mathbf{X}}\mathbf{W}_n/\alpha_n^{1/2}, \widetilde{\mathbf{Y}}/\beta_m^{1/2}) \to 0$$

and, furthermore,

$$U_{n,m}(\widetilde{\mathbf{X}}\mathbf{W}_n/\alpha_n^{1/2}, \widetilde{\mathbf{Y}}/\beta_m^{1/2}) \to \|\mu[F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}] - \mu[F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}]\|_{\mathcal{H}}^2,$$

where $\widetilde{\mathbf{Q}}_{\mathbf{X}}$ and $\widetilde{\mathbf{Q}}_{\mathbf{Y}}$ are indefinite orthogonal matrices defined only in terms of the distributions F_X, F_Y and the signature (p, q). Therefore, we analyze the convergence of a scaled U-statistic.

Theorem 13. *Let $(\mathbf{A}^{(1)}, \mathbf{X}) \sim GRDPG(F_X, n, \alpha_n)$ and $(\mathbf{A}^{(2)}, \mathbf{Y}) \sim GRDPG(F_Y, m, \beta_m)$ be independent, and suppose Assumption 2a is satisfied. Suppose $\mathbf{\Delta}_{\mathbf{X}}\mathbf{I}_{p,q}$ and $\mathbf{\Delta}_{\mathbf{Y}}\mathbf{I}_{p,q}$ have distinct eigenvalues, and let κ satisfy Assumption 1. Then, under the null hypothesis $F_X \simeq F_Y$, there exists a sequence of block-orthogonal matrices $\mathbf{W}_n \in \mathbb{O}(d) \cap \mathbb{O}(p, q)$ such that*

$$(m\beta_m + n\alpha_n)\left(U_{n,m}\left(\frac{\widehat{\mathbf{X}}}{\sqrt{\alpha_n}}, \frac{\widehat{\mathbf{Y}}}{\sqrt{\beta_m}}\right) - U_{n,m}\left(\frac{\widetilde{\mathbf{X}}\mathbf{W}_n}{\sqrt{\alpha_n}}, \frac{\widetilde{\mathbf{Y}}}{\sqrt{\beta_m}}\right)\right) \to 0$$

almost surely. If instead $F_X \not\simeq F_Y$,

$$\frac{(m\beta_m + n\alpha_n)}{\log(n)}\left(U_{n,m}\left(\frac{\widehat{\mathbf{X}}}{\sqrt{\alpha_n}}, \frac{\widehat{\mathbf{Y}}}{\sqrt{\beta_m}}\right) - U_{n,m}\left(\frac{\widetilde{\mathbf{X}}\mathbf{W}_n}{\sqrt{\alpha_n}}, \frac{\widetilde{\mathbf{Y}}}{\sqrt{\beta_m}}\right)\right) \to 0$$

almost surely.

The hypotheses of the previous theorem include that $\mathbf{\Delta}\mathbf{I}_{p,q}$ has distinct eigenvalues. Similar to Proposition 5, we can actually remove this assumption if we are willing to include additional orthogonal matrices.

Theorem 14. *Consider the setting of Theorem 13, but suppose that $\mathbf{\Delta}\mathbf{I}_{p,q}$ does not necessarily have distinct eigenvalues. Then, we have that under the null $F_X \simeq F_Y$ there exist two*

sequences of block-orthogonal matrices $\widehat{\mathbf{W}}_n$ *and* $\mathbf{W}_n$ *such that*

$$(m\beta_m + n\alpha_n)\left(U_{n,m}\left(\frac{\widehat{\mathbf{X}}\widehat{\mathbf{W}}_n}{\sqrt{\alpha_n}}, \frac{\widehat{\mathbf{Y}}}{\sqrt{\beta_m}}\right) - U_{n,m}\left(\frac{\widetilde{\mathbf{X}}\mathbf{W}_n}{\sqrt{\alpha_n}}, \frac{\widetilde{\mathbf{Y}}}{\sqrt{\beta_m}}\right)\right) \to 0$$

almost surely. If instead $F_X \not\asymp F_Y$*, then*

$$\frac{(m\beta_m + n\alpha_n)}{\log(n)}\left(U_{n,m}\left(\frac{\widehat{\mathbf{X}}\widehat{\mathbf{W}}_n}{\sqrt{\alpha_n}}, \frac{\widehat{\mathbf{Y}}}{\sqrt{\beta_m}}\right) - U_{n,m}\left(\frac{\widetilde{\mathbf{X}}\mathbf{W}_n}{\sqrt{\alpha_n}}, \frac{\widetilde{\mathbf{Y}}}{\sqrt{\beta_m}}\right)\right) \to 0$$

almost surely.

In Theorem 14, the additional orthogonal matrix $\widehat{\mathbf{W}}_n$ appears because without distinct eigenvalues assumption $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ need to be simultaneously aligned to each other as well as to $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. Again, *a priori* there are indefinite orthogonal transformations to contend with, but, as we show in Section E.1, we can effectively bypass these transformations through careful analyses of their convergence. Similar to the proof of Proposition 4, the proof of Theorem 14 reveals that $U_{n,m}(\widetilde{\mathbf{X}}\mathbf{W}_n/\alpha_n^{1/2}, \widetilde{\mathbf{Y}}/\beta_m^{1/2})$ is converging to a quantity that depends only on population parameters; however, without the distinct eigenvalues assumption the matrices $\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}$ and $\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}$ are no longer unique, so the convergence analysis and its statement require careful tabulation of additional block-orthogonal matrices.

If instead Assumption 2b holds, one can obtain a similar limiting result without including the sparsity in the scaling under the null hypothesis.

Corollary 5. *Suppose the setting of Theorem 14, but suppose instead that Assumption 2b is satisfied. Under the null hypothesis, we have that*

$$(m+n)\left(U_{n,m}\left(\frac{\widehat{\mathbf{X}}\widehat{\mathbf{W}}_n}{\sqrt{\alpha_n}}, \frac{\widehat{\mathbf{Y}}}{\sqrt{\beta_m}}\right) - U_{n,m}\left(\frac{\widetilde{\mathbf{X}}\mathbf{W}_n}{\sqrt{\alpha_n}}, \frac{\widetilde{\mathbf{Y}}}{\sqrt{\beta_m}}\right)\right) \to 0$$

almost surely, for the same sequences of orthogonal matrices $\widehat{\mathbf{W}}_n$ *and* $\mathbf{W}_n$ *as in Theorem 14.*

Finally, we note that in general the sparsity factors α_n and β_m are not known. If instead we wish to use the estimated sparsity factors, we have the following result.

Corollary 6. *Assume that* $\mathbb{E}(X_1^\top \mathbf{I}_{p,q} X_2) = 1$ *and that* $\alpha_n, \beta_m \to 0$. *Define*

$$\widehat{\alpha}_n := \binom{n}{2}^{-1} \sum_{i<j} \mathbf{A}^{(1)}_{ij}; \qquad \widehat{\beta}_m := \binom{m}{2}^{-1} \sum_{i<j} \mathbf{A}^{(2)}_{ij}.$$

Then the limiting results in Theorem 13, Theorem 14, and Corollary 5 all hold under their respective conditions with $\widehat{\mathbf{X}}/\alpha_n^{1/2}$ *and* $\widehat{\mathbf{Y}}/\beta_m^{1/2}$ *replaced with* $\widehat{\mathbf{X}}/\widehat{\alpha}_n^{1/2}$ *and* $\widehat{\mathbf{Y}}/\widehat{\beta}_m^{1/2}$ *respectively and the almost sure convergence replaced with convergence in probability.*

The condition $\mathbb{E}(X_1 \mathbf{I}_{p,q} X_2) = 1$ is used only for identifiability of α_n and β_m when they need to be estimated. See e.g., Lunde and Sarkar (2019) for an identical condition in the setting of graphons.

Interpretation

There are several different alignment matrices that appear in order to show the convergence in Theorems 13 and 14. However, in our analysis we are able to show that only the indefinite orthogonal matrices that are simultaneously orthogonal have any effect on the limiting values. Given Propositions 4 and 5, the main results in Theorems 13 and 14 further detail that under the null hypothesis $F_X \simeq F_Y$ one can perform consistent testing given access to only the graphs $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$. The results of Gretton et al. (2012) imply that $(m+n)U_{n,m}(\mathbf{X}, \mathbf{Y})$ has a nondegenerate limiting distribution under the null hypothesis. For graphs with average degree growing faster than $n^{1/2}\text{polylog}(n)$, Corollary 5 says that the same scaling occurs under the null hypothesis with $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ as replacements for $\mathbf{X}$ and $\mathbf{Y}$. For almost surely dense graphs; i.e. graphs with $\alpha_n = \beta_m = 1$, Theorems 13 and 14 also provide a result under the alternative.

As mentioned in Section 5.3.1, under the distinct eigenvalues assumption, the proof of Proposition 4 reveals that

$$U_{n,m}(\widetilde{\mathbf{X}}\mathbf{W}_n/\sqrt{\alpha_n}, \widetilde{\mathbf{Y}}/\sqrt{\beta_m}) \to \|\mu[F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}] - \mu[F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}]\|_{\mathcal{H}}^2,$$

where $\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}$ and $\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}$ are deterministic quantities depending only on F_X and F_Y. A similar convergence happens without the distinct eigenvalues assumption but with additional block-

orthogonal matrices. Lemma 50 shows that the rate of the approximation of $\widetilde{\mathbf{X}}\mathbf{W}_n/\sqrt{\alpha_n}$ and $\widetilde{\mathbf{Y}}/\sqrt{\beta_m}$ to $\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}$ and $\mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}$ is of order $\sqrt{\log(n)/n}$, which, in general, is not fast enough to guarantee the convergence of

$$(n+m)\left(U_{n,m}(\widetilde{\mathbf{X}}\mathbf{W}_n/\alpha_n^{1/2},\widetilde{\mathbf{Y}}/\beta_m^{1/2})-U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}},\mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}})\right)$$

to zero. However, Propositions 4 and 5 show that the U-statistic evaluated at $\widehat{\mathbf{X}}\widehat{\mathbf{W}}_n$ and $\widehat{\mathbf{Y}}$ still tends to zero under the null hypothesis and to a constant under the alternative, which, together with Theorems 13 and 14, suggests that $(n+m)U_{n,m}(\widetilde{\mathbf{X}}\mathbf{W}_n/\alpha_n^{1/2},\widetilde{\mathbf{Y}}/\beta_m^{1/2})$ has a nondegenerate limiting distribution.

For testing purposes, the lack of distributional results is of no consequence, since if one can reliably estimate the orthogonal transformation $\widehat{\mathbf{W}}_n$ appearing in Theorem 14 (and Proposition 5) then one can perform consistent testing through a bootstrapped permutation test; see the following section. Furthermore, the limiting distribution for the maximum mean discrepancy between two distributions F_X and F_Y will not be independent of F_X and F_Y in general, so one may have to use a permutation test to approximate the null distribution anyways.

For sparser graphs, the almost sure convergence in Theorems 13 and 14 under the null hypothesis requires the scaling $m\beta_m+n\alpha_n$, which, if $n\asymp m$ and $\alpha_n\asymp\beta_m$ is slower than the convergence in Gretton et al. (2012) by a factor of α_n. The reason for this stems primarily from the fact that for sparse graphs it is much more difficult to estimate $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. The sparsity factor here brings down the effective sample size; one observes only $O(n\alpha_n)$ edges on average for sparse graphs instead of $O(n)$ edges for dense graphs. Therefore, though the scaling may not be optimal, Theorems 13 and 14 provide a more refined study of the test statistic, since Propositions 4 and 5 already imply consistent testing.

We also note that our results can be adapted to the conditional mixed-membership and degree-corrected stochastic blockmodel. Consider a deterministic sequence of matrices $\mathbf{P}=\mathbf{P}_n$, where $\mathbf{P}$ has the structure

$$\mathbf{P}=\alpha_n\mathbf{\Theta}\mathbf{Z}\mathbf{B}\mathbf{Z}^\top\mathbf{\Theta},$$

where $\mathbf{Z} \in [0,1]^{n \times K}$ is a membership matrix whose rows sum to 1, and $\mathbf{\Theta} \in (0,1)^{n \times n}$ is a diagonal matrix of degree-correction parameters. For identifiability, assume $\max_i \mathbf{\Theta}_{ii} = 1$. Let $\mathbf{B} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ be the eigendecomposition of $\mathbf{B}$, and let p and q denote the number of positive and negative entries of $\mathbf{D}$ respectively. Define $\mathbf{X} = \mathbf{\Theta}\mathbf{Z}\mathbf{V}|\mathbf{D}|^{1/2}$; then $\mathbf{P} = \alpha_n \mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top$. Though Theorem 14 is not immediately applicable as the rows of $\mathbf{X}$ are no longer drawn i.i.d., the proof can be modified as long as $\mathbf{\Theta}$ and $\mathbf{Z}$ both converge in the sense that the limit $\frac{1}{n}\mathbf{Z}^\top\mathbf{\Theta}^\top\mathbf{\Theta}\mathbf{Z}\mathbf{B}$ exists. If $\mathbf{B}$ is full rank and does not change in n, the notion of repeated eigenvalues then refers to the eigenvalues of the limit of $\frac{1}{n}\mathbf{Z}^\top\mathbf{\Theta}^\top\mathbf{\Theta}\mathbf{Z}\mathbf{B}$, which depends on the particular sequence of $\mathbf{Z}$ and $\mathbf{\Theta}$ matrices.

5.3.2 Optimal Transport for Repeated Eigenvalues

We note that thus far, we have demonstrated that negative eigenvalues do not affect limiting results despite *a priori* having to consider indefinite orthogonal transformations. Such a result is desirable, as one does not have to resort to numerical algorithms optimizing over $\mathbb{O}(p,q)$, which could be unstable due to the ill-conditioning inherent in indefinite orthogonal transformations. Furthermore, we have shown that any modification to our test need estimate only the matrix $\widehat{\mathbf{W}}_n$ from Theorem 14. We now draw our attention to estimating $\widehat{\mathbf{W}}_n$.

Let $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ be the adjacency spectral embeddings of $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$. One can view a collection of points as a distribution by assigning equal point mass to each point. Define $\widehat{F}_{\widehat{X}}$ as the empirical distribution for $\widehat{\mathbf{X}}$ and define $\widehat{F}_{\widehat{Y}}$ as the empirical distribution for $\widehat{\mathbf{Y}}$; that is $\widehat{F}_{\widehat{X}}$ places point mass of $\frac{1}{n}$ at each $\widehat{X}_i$, and $\widehat{F}_{\widehat{Y}}$ places point mass of $\frac{1}{m}$ at each $\widehat{Y}_j$. Let $d_2(\cdot,\cdot)$ denote the Wasserstein ℓ^2 distance between two distributions; that is, given two distributions F and G, we define

$$d_2(F,G) := \inf_{\Gamma_{F,G}} (\mathbb{E}_{(X,Y)\sim\Gamma_{F,G}} \|X - Y\|_2^2)^{1/2},$$

where $\Gamma_{F,G}$ is the set of distributions whose marginals are F and G. The set $\Gamma_{F,G}$ is called the set of *couplings* of F and G. If F and G are empirical distributions on n and m points respectively, the couplings can be represented by matrices whose rows and columns sum to

$\frac{1}{m}$ and $\frac{1}{n}$; these are the *assignment matrices*.

In light of Theorem 14, we propose finding the orthogonal matrix $\widehat{\mathbf{W}}_n$ that solves the problem

$$\inf_{\mathbf{W}\in\mathbb{O}(d)\cap\mathbb{O}(p,q)} d_2(\widehat{F}_{\widehat{X}/\widehat{\alpha}_n^{1/2}}, \widehat{F}_{\widehat{Y}/\widehat{\beta}_m^{1/2}}\circ\mathbf{W}), \tag{5.1}$$

where $\widehat{F}_{\widehat{Y}/\widehat{\beta}_m^{1/2}}\circ\mathbf{W}$ is the empirical distribution $\widehat{F}_{\widehat{Y}/\widehat{\beta}_m^{1/2}}$ transformed by an orthogonal matrix $\mathbf{W}$. The above distance is considered in both Lei (2020b) and Levin and Levina (2019) as the *orthogonal Wasserstein distance*.

The problem in expression 5.1 is simultaneously an optimal transport problem in finding the minimum over couplings and a Procrustes problem in finding the minimum over orthogonal matrices. Define the matrix $\mathbf{C}_{\mathbf{W}}\in\mathbb{R}^{n\times m}$ as the *cost matrix with respect to* $\mathbf{W}$ by setting

$$(\mathbf{C}_{\mathbf{W}})_{ij} := ||\widehat{X}_i - \mathbf{W}\widehat{Y}_j||^2.$$

Then expression 5.1 can be written as

$$\min_{\mathbf{W},\mathbf{\Pi}} \langle \mathbf{\Pi}, \mathbf{C}_{\mathbf{W}}\rangle \tag{5.2}$$

where the inner product is the Frobenius (matrix) inner product, $\mathbf{W}$ is a block-orthogonal matrix, and $\mathbf{\Pi}$ satisfies $\mathbf{\Pi}\mathbf{1} = \frac{1}{m}\mathbf{1}$ and $\mathbf{\Pi}^\top\mathbf{1} = \frac{1}{n}\mathbf{1}$; that is, $\mathbf{\Pi}$ is an assignment matrix. We have the following proposition.

Proposition 6. *Assume that* $\mathbb{E}(X_1^\top\mathbf{I}_{p,q}X_2) = 1$ *and that* $\alpha_n, \beta_m \to 0$. *Suppose* $\widehat{\mathbf{W}}_n$ *minimizes* $\langle\mathbf{\Pi}, \mathbf{C}_{\mathbf{W}}\rangle$ *over the block-orthogonal matrices. Suppose further that* $F_X \simeq F_Y$*; that is, the null hypothesis holds. Then there exists constants* $c > 0$ *and* $C > 0$ *possibly depending on* d *such that with probability at least* $1 - c(n^{-2} + m^{-2})$,

$$d_2(\widehat{F}_{\widehat{X}/\widehat{\alpha}_n^{1/2}}, \widehat{F}_{\widehat{Y}/\widehat{\beta}_m^{1/2}}\circ\widehat{\mathbf{W}}_n) \le C\left(\frac{\log^{1/d}(n)}{n^{1/d}} + \frac{\log^{1/d}(m)}{m^{1/d}} + \frac{\log(n)}{(n\alpha_n)^{1/2}} + \frac{\log(m)}{(m\beta_m)^{1/2}}\right).$$

We also show that the orthogonal Wasserstein distance does not tend to zero under the alternative.

Proposition 7. *Assume that* $\mathbb{E}(X_1^\top \mathbf{I}_{p,q} X_2) = 1$ *and that* $\alpha_n, \beta_m \to 0$. *Suppose* $\widehat{\mathbf{W}}_n$ *minimizes* $\langle \mathbf{\Pi}, \mathbf{C}_{\mathbf{W}} \rangle$ *over the block-orthogonal matrices. Suppose that* $F_X \not\approx F_Y$. *Then there exists a constant* $C > 0$ *depending on* F_X *and* F_Y *such that*

$$\liminf_{n,m} d_2(\widehat{F}_{\widehat{X}/\sqrt{\widehat{\alpha}_n}}, \widehat{F}_{\widehat{Y}/\sqrt{\widehat{\beta}_m}} \circ \widehat{\mathbf{W}}_n) \geq C$$

almost surely.

Again, the assumption $\mathbb{E}(X_1^\top \mathbf{I}_{p,q} X_2) = 1$ is for identifiability of α_n and β_m. If instead one assumes that $\alpha_n = \beta_m = 1$, the result still holds without the sparsity factors.

The above theory shows that given two adjacency matrices $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$, calculating the adjacency spectral embeddings $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$, aligning them with an orthogonal matrix by solving Equation 5.2, and calculating the corresponding U-statistic, $U_{n,m}(\widehat{\mathbf{X}}, \widehat{\mathbf{Y}}\widehat{\mathbf{W}}_n)$ yields a consistent test statistic. From there, one can bootstrap the null distribution of $U_{n,m}$ to get an approximate p-value. The procedure is summarized in Algorithm 6.

Algorithm 6 Nonparametric Two-Graph Hypothesis Testing

Require: $\mathbf{A}^{(1)} \in \mathbb{R}^{n \times n}, \mathbf{A}^{(2)} \in \mathbb{R}^{m \times m}$

1: Embed $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ into $\mathbb{R}^d$ using the adjacency spectral embeddings, obtaining $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ and sparsity estimates $\widehat{\alpha}_n^{1/2}$ and $\widehat{\beta}_m^{1/2}$.
2: Find $\widehat{\mathbf{W}}_n$ minimizing Equation 5.2 above using Algorithm 7;
3: Calculate the value of the U-statistic $U_{n,m}(\widehat{\mathbf{X}}/\widehat{\alpha}_n^{1/2}, \widehat{\mathbf{Y}}\widehat{\mathbf{W}}_n/\widehat{\beta}_m^{1/2})$;
4: Bootstrap the U-statistic distribution assuming the null hypothesis;
5: Calculate the empirical probability of observing $U_{n,m}(\widehat{\mathbf{X}}/\widehat{\alpha}_n^{1/2}, \widehat{\mathbf{Y}}\widehat{\mathbf{W}}_n/\widehat{\beta}_m^{1/2})$ under the bootstrapped null distribution.
6: **return** Estimated p-value.

To solve for the matrix $\widehat{\mathbf{W}}_n$ in practice, we use the method proposed in Alvarez-Melis et al. (2019) tailored to our specific problem, in which the authors propose solving an entropy-regularized version of the problem which can be done efficiently. Define the auxiliary expression

$$\inf_{\mathbf{\Pi}, \mathbf{W}} \langle \mathbf{\Pi}, C_{\mathbf{W}} \rangle + \varepsilon H(\mathbf{\Pi}) \tag{5.3}$$

where $H(\mathbf{\Pi})$ is the entropy of the distribution given by $\mathbf{\Pi}$. For a fixed ε, Equation 5.3

can be computed efficiently via the Sinkhorn algorithm (Cuturi, 2013). We then alternately minimize over $\mathbf{W}$ and $\mathbf{\Pi}$ to find the solution. Finally, given a fixed orthogonal matrix $\mathbf{W}$, we project $\mathbf{W}$ onto the block-orthogonal matrices. The (Frobenius) projection is given by the following proposition. We summarize the entire procedure in Algorithm 7.

Proposition 8. *Let* $\mathbf{W} \in \mathbb{O}(d)$. *Then*

$$\inf_{\mathbf{R} \in \mathbb{O}(d) \cap \mathbb{O}(p,q)} \|\mathbf{R} - \mathbf{W}\|_F$$

is attained by taking the orthogonal components of the singular value decomposition of the top $p \times p$ *block of* $\mathbf{W}$ *and the bottom* $q \times q$ *block of* $\mathbf{W}$.

Algorithm 7 Optimal Transport-Procrustes

Require: $\widehat{\mathbf{X}}, \widehat{\mathbf{Y}}$ initial guesses $\mathbf{W} \in \mathbb{O}(d) \cap \mathbb{O}(p,q)$, $\mathbf{\Pi} = \frac{1}{mn}\mathbf{1}\mathbf{1}^\top$, dimension p and $q = d - p$.

1: **repeat**
2: Set $\mathbf{M} := \mathbf{\Pi}\widehat{\mathbf{X}}\mathbf{W}(\widehat{\mathbf{Y}})^\top$ with singular value decomposition $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, set $\mathbf{W}' := \mathbf{U}\mathbf{V}^\top$
3: Calculate $C_{\mathbf{W}'}$ via $(C_{\mathbf{W}'})_{ij} := ||\widehat{X}_i - \mathbf{W}'\widehat{Y}_j||_2^2$
4: Set $\varepsilon > 0$ as some positive number and solve for $\mathbf{\Pi}$ via the Sinkhorn algorithm
5: Set $\mathbf{W} := \mathbf{W}'$
6: **until** max number of iterations
7: Define $\mathbf{W_p} := \mathbf{W}[1 : p, 1 : p]$; $\mathbf{W_q} := \mathbf{W}[p+1 : d, p+1 : d]$ with singular value decompositions $\mathbf{U_p}\mathbf{\Sigma_p}\mathbf{V_p}^\top$ and $\mathbf{U_q}\mathbf{\Sigma_q}\mathbf{V_q}^\top$
8: Set $\widehat{\mathbf{W}} = \text{bdiag}(\mathbf{U_p}\mathbf{V_p}^\top, \mathbf{U_q}\mathbf{V_q}^\top)$
9: **return** $\widehat{\mathbf{W}}$

In essence, the algorithm alternates between solving for an orthogonal transformation given a fixed assignment matrix and solving for the assignment matrix given the orthogonal transformation.

Close Eigenvalues

Before moving on, we provide some intuition as to why estimating a rotation can be beneficial even when one does not have exactly repeated eigenvalues. We focus on the positive semidefinite case for convenience, though the analysis for the indefinite case is similar.

Suppose $\mathbb{E}(XX^\top)$ has d distinct eigenvalues, and let $\mathbf{U_A}$ and $\mathbf{U_P}$ be the leading d

eigenvectors of $\mathbf{A}$ and $\mathbf{P} = \alpha_n \mathbf{X}\mathbf{X}^\top$. Let $\mathbf{W}_*$ be defined via

$$\mathbf{W}_* = \inf_{\mathbf{W} \in \mathbb{O}(d)} \|\mathbf{U}_\mathbf{A} - \mathbf{U}_\mathbf{P}\mathbf{W}\|_F,$$

which has a closed-form solution in terms of the left and right singular vectors of the matrix $\mathbf{U}_\mathbf{A}^\top \mathbf{U}_\mathbf{P}$. Since $\mathbb{E}(XX^\top)$ has distinct eigenvalues, without loss of generality assume that the columns of $\mathbf{U}_\mathbf{A}$ are chosen so that the inner product between the columns of $\mathbf{U}_\mathbf{A}$ and $\mathbf{U}_\mathbf{P}$ are positive. Then the sequence of matrices $\mathbf{W}_*$ is converging to the identity, which also provides the uniqueness (up to sign) of the matrices $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$.

Define

$$\delta := \min_{1 \le i \le d} \Big(\lambda_i(\mathbb{E}(XX^\top)) - \lambda_{i+1}(\mathbb{E}(XX^\top)) \Big),$$

where $\lambda_{d+1} := -\infty$ by convention. It can be shown (see Appendix E.2) that

$$\|\mathbf{W}_* - \mathbf{I}\|_F = O\left(\frac{\log(n)}{n\alpha_n \delta} \right),$$

where the big $O(\cdot)$ notation hides dependence on the dimension d. Hence, even though the right hand side tends to zero as $n\alpha_n \to \infty$, so the eigenvectors of $\mathbf{A}$ and $\mathbf{P}$ are well-aligned (up to sign), the rate of convergence of the orthogonal matrix depends on n, α_n, and the corresponding eigengap.

In practice, one observes only the two graphs, and the eigenvalues must be estimated from the eigenvalues of $\mathbf{A}$. Therefore, even though the orthogonal matrix is converging to the identity, for any finite n, it may not be close if the eigengap is small relative to n. So if one observes two graphs from the same model, but both n and m are small relative to δ, then one may still need to estimate a rotation to align $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$, despite asymptotically having distinct eigenvalues.

5.3.3 Relation to Previous Results

There have been several tests proposed assuming that the graphs have the same set of vertices, such as Tang et al. (2017a); Ghoshdastidar et al. (2017); Li and Li (2018); Levin et al. (2019) and Draves and Sussman (2020). In Tang et al. (2017a); Levin and Levina (2019)

and Draves and Sussman (2020), the authors work under the random dot product graph model, though they require that the expected degree grows as $\omega(n)$ (that is, the sparsity parameter is constant). In Li and Li (2018), working under the stochastic blockmodel, the authors are able to derive more explicit limiting results for their test statistic. In Ghoshdastidar et al. (2020), the authors allow for arbitrary distributions on two graphs, but again require that the graphs be on the same set of nodes. In contrast to all of these works, we do not assume that the two graphs are on the same set of vertices.

Our test statistic is based on a two-sample U-statistic using the rows of $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$. In Levin and Levina (2019), the authors consider bootstrapping nondegenerate U-statistics for random dot product graphs by estimating the latent positions. In addition, there have been a number of works on U-statistics for graphs in the more general graphon model (Lunde and Sarkar, 2019; Zhang and Xia, 2020; Lin et al., 2020a,b), but these involve bootstrapping moments of the underlying graphon, which can be computationally infeasible in practice. In this paper, we study a *degenerate* two-sample test statistic, which is not considered in any of these works.

Both Ghoshdastidar et al. (2017) and Tang et al. (2017b) consider a similar test as in this paper. In Ghoshdastidar et al. (2017), the authors introduce a formalism for two-sample testing under the assumption one observes only the adjacency matrices. Although our broad setting is similar to theirs, the model we study has more structural assumptions, allowing us to construct a test statistic using estimated latent positions. In addition, since our population test statistic is injective, we obtain a universally consistent and computationally tractable test statistic for GRDPGs, whereas the test statistic in Ghoshdastidar et al. (2017) will not necessarily be universally consistent or computationally tractable in general. Furthermore, the structure in the setting we consider allows for a much simpler bootstrapping procedure.

In Tang et al. (2017b), the authors consider a similar test statistic as in our setting under the assumption that $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$ are both from the random dot product graph model, the sparsity parameters are constant, and the matrices $\mathbb{E}(XX^\top)$ and $\mathbb{E}(YY^\top)$ have distinct eigenvalues. Leveraging previous results for random dot product graphs, the authors show that if κ is a radial kernel, and $\frac{m}{n+m} \to \rho \in (0,1)$, then there exists a sequence of orthogonal

matrices $\mathbf{W}_n$ such that

$$(n+m)\left(U_{n,m}(\widehat{\mathbf{X}},\widehat{\mathbf{Y}}) - U_{n,m}(\mathbf{X},\mathbf{Y}\mathbf{W}_n)\right) \to 0$$

almost surely under the null hypothesis, where $U_{n,m}(\widehat{\mathbf{X}},\widehat{\mathbf{Y}})$ is the U-statistic defined using the estimates $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ from the adjacency spectral embeddings of $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$. Moreover, for the sequence of orthogonal matrices $\mathbf{W}_n$ we have

$$U_{n,m}(\mathbf{X},\mathbf{Y}\mathbf{W}_n) \to \begin{cases} 0 & F_X \simeq F_Y \\ c > 0 & F_X \not\simeq F_Y, \end{cases}$$

where recall for RDPGs $F_X \simeq F_Y$ means that F_X and F_Y are the equivalent up to orthogonal transformation.

While at first glance the test statistic proposed in Tang et al. (2017b) is similar to our test statistic, analyzing the test statistic in a general low-rank setting involves substantial theoretical and methodological considerations with respect to indefinite orthogonal transformations, optimal transport, and sparsity. In addition, the assumption that $\mathbb{E}(XX^\top)$ and $\mathbb{E}(YY^\top)$ have repeated eigenvalues precludes testing in the case of the K-block balanced homogeneous stochastic blockmodel from Section 5.1.1 even if the $\mathbf{B}$ matrix is positive definite. Our results include the random dot product graph as a special case, though the analysis and our proposed methodology are not simply trivial extensions of the results in Tang et al. (2017b).

We remark that Propositions 6 and 7 provide similar bounds to Theorem 5 of Levin and Levina (2019) and Theorem 4.4 of Lei (2020b), both of which consider convergence of empirical distributions to the corresponding latent position distribution under the single-graph setting, though in both of these works they require that the orthogonal matrix in the eigendecomposition of $\mathbb{E}(XX^\top)$ is block diagonal. As a counterexample, consider the following GRDPG model. Let $\mathbf{B} \in [0,1]^{K\times K}$ be a symmetric connectivity matrix of rank K, and let $\mathbf{V}\mathbf{D}\mathbf{V}^\top$ be its eigendecomposition. Let $Z_i \sim$Dirichlet(α) for some $\alpha \in \mathbb{R}^K$, and define $X_i = \mathbf{V}|\mathbf{D}|^{1/2}Z_i$, which is a valid GRDPG distribution. Then the matrix $\mathbb{E}(XX^\top)$

has a block-orthogonal eigendecomposition if and only if $\mathbf{V}|\mathbf{D}|^{1/2}\mathbb{E}(ZZ^\top)|\mathbf{D}|^{1/2}\mathbf{V}^\top$ does, where $\mathbb{E}(ZZ^\top)$ is the second moment matrix for a Dirichlet random variable. If α is the all ones vector, then

$$\mathbf{V}|\mathbf{D}|^{1/2}\mathbb{E}(ZZ^\top)|\mathbf{D}|^{1/2}\mathbf{V}^\top = \frac{K}{K+1}\mathbf{V}|\mathbf{D}|\mathbf{V}^\top + \frac{1}{K(K+1)}\mathbf{V}|\mathbf{D}|^{1/2}\mathbf{1}\mathbf{1}^\top|\mathbf{D}|^{1/2}\mathbf{V}^\top.$$

The example $\mathbf{B} = -.1\mathbf{I} + .2\mathbf{1}\mathbf{1}^\top$ yields an orthogonal matrix that is not block-diagonal. Hence, even though assuming the eigendecomposition of $\mathbb{E}(XX^\top)$ has a block diagonal structure is an attractive assumption amenable to theoretical analysis, this assumption can be violated by many different models.

Because of the prevalence of spectral methods in the literature, estimation of $\widehat{\mathbf{W}}_n$ arises often in related inference tasks. For example, Zhang (2018) proposes solving a smooth function of the Laplace distance between distributions to estimate $\widehat{\mathbf{W}}_n$, and Li and Li (2018), operating under the stochastic blockmodel, consider estimating $\widehat{\mathbf{W}}_n$ by minimizing over the community memberships. Indeed, both methods are practically similar to ours, and may provide comparable results in practice, though we believe we are the first to apply it to nonparametric hypothesis testing and to provide asymptotic statistical guarantees under both the null and alternative hypotheses. Furthermore, Optimal Transport-Procrustes has been used to some success in the literature on natural language processing. Though our methodology is similar to Alvarez-Melis et al. (2019), other methods have been proposed for numerically solving the problem (e.g. Grave et al. (2019)).

Finally, we mention that though our algorithm is based on entropy-regularized Wasserstein distance, our results are stated in terms of the unregularized Wasserstein distance. While it may be possible to extend results on regularized optimal transport (e.g. Gangrade et al. (2019); Bigot et al. (2019)) to the mixed continuous and discrete setting implicitly required for our purposes, such an extension would require nontrivial analysis of the regularized Sinkhorn distance between $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$.

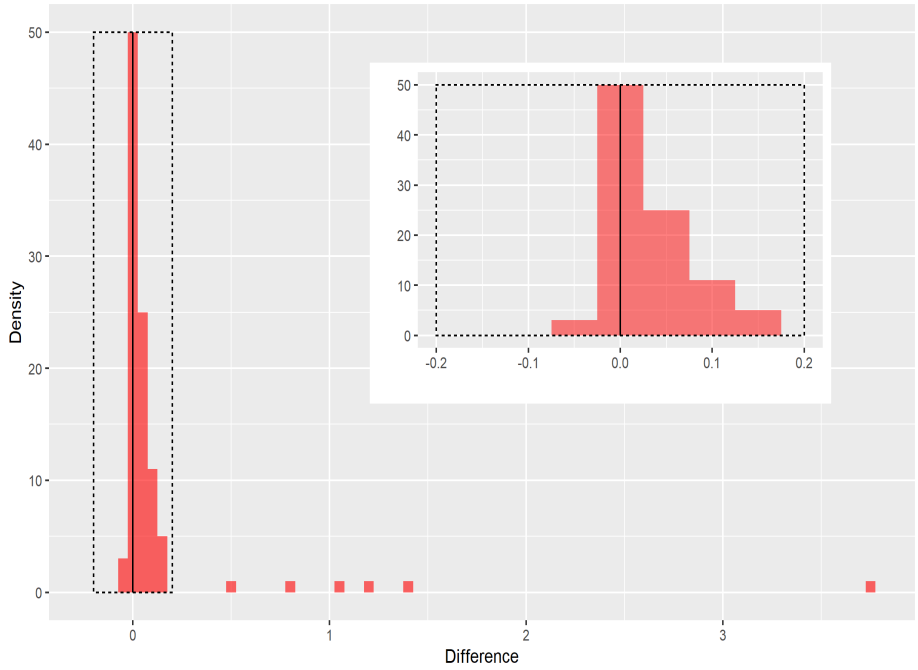


Figure 5.2: Density plot of the difference $\widehat{U}_{\text{sign flips}} - \widehat{U}_{\text{rotation}}$ 100 Monte Carlo iterations of a stochastic blockmodel. A Wilcoxon test gives a p-value of $< .0001$ for testing whether the estimated rotation is better than sign flips.

5.4 Simulations

Recall the motivating example in Section 5.1.1, given by the balanced homogeneous stochastic blockmodel, in which for vertices in the same community, the probability of an edge is a and between communities the probability of an edge is b. When $b > a$ and the probability of belonging to a community is $\frac{1}{K}$ if there are K communities, then this model has both repeated and negative eigenvalues. With the same model as in Section 5.1.1, where

$$\mathbf{B} = \begin{pmatrix} .5 & .8 & .8 \\ .8 & .5 & .8 \\ .8 & .8 & .5 \end{pmatrix}$$

and the probability of community membership is $1/3$, we simulate 100 Monte Carlo iterations on $n = 300$ vertices from this stochastic blockmodel, and we calculate the value of our test statistic with both the naïvely rotated versions and the output of Algorithm 7.

Under the distinct eigenvalues assumption, choosing the sign of $\widehat{\mathbf{X}}$ to match those of $\widehat{\mathbf{Y}}$

suffices to give convergence as in Theorem 13; we dub this naïve alignment procedure the *sign flips procedure*. In Figure 5.2, we plot the density of the difference of our estimated test statistic using the Gaussian kernel with both the sign flips procedure and the estimated rotation from Algorithm 7. Since this is a finite sample simulation, we find the alignment $\widehat{\mathbf{W}}_n$ by running Algorithm 7 from multiple different initializations, and we take the value $\widehat{\mathbf{W}}$ that minimizes the test statistic, where $\widehat{\mathbf{W}}$ are the local minimums from Algorithm 7 for the estimated rotation and $\widehat{\mathbf{W}}$ are the sign matrices for $\widehat{U}_{\text{sign flip}}$. We see that the density lies almost completely to the right of zero which suggests that the naïve estimate is nearly always larger than the estimated rotation. Moreover, there are some situations in which the difference is quite large, which demonstrates that the test statistic estimated using only sign flips need not necessarily converge to zero under the null hypothesis.

Under the null hypothesis, the test statistic should be tending to zero almost surely, and we see that the value of the test statistic evaluated using the estimated rotation is much more concentrated about zero. Moreover, a Wilcoxon test gives a p-value of less than 0.0001 for testing whether the estimated rotation test statistic is smaller than the sign-flips.

In Figure 5.3, using the same $\mathbf{B}$ matrix as in the previous example, we also allow for independent degree correction parameters $\theta_i \sim .5 \times U(0,1) + .5$, where $U(0,1)$ denotes the Uniform distribution on $(0,1)$. For two vertices i and j with communities k and l respectively, the probability of an edge is defined as $\theta_i\theta_j\mathbf{B}_{kl}$. We plot the second and third dimensions of $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ with both sign flips and the estimated orthogonal matrix output from Algorithm 7. By the independence of the degree-correction parameters, the matrix $\mathbb{E}(XX^\top)\mathbf{I}_{p,q}$ still has repeated eigenvalues, since it will be a scalar times the corresponding second moment matrix for a stochastic blockmodel. Here we use $n = m = 500$ to encourage the convergence of the second moment matrix. We see that visually the corresponding clusters lie on top of each other despite the added noise from the degree correction. Note that the clusters are "elongated" relative to the stochastic blockmodel in Figure 5.1; this is due to the fact that the latent position distribution for a degree-corrected stochastic blockmodel is supported on a ray, since the degree correction parameters change the magnitude of the latent positions but not the direction.

In Figure 5.4 we plot the density of the test statistic with and without the rotation, and

Figure 5.3: Comparisons of the naïve sign-flip alignment procedure and the optimal transport alignment procedure for two adjacency spectral embeddings for the degree-corrected stochastic blockmodel. The left hand side shows the naïve alignment, and visually the clusters are not on top of each other, and the right hand side shows that using Algorithm 7 places the clusters approximately on top of each other

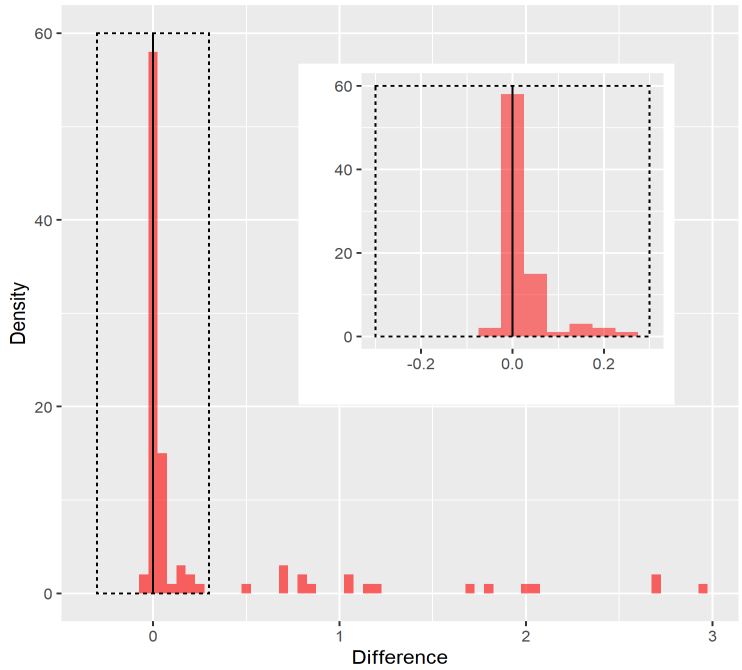


Figure 5.4: Density plot of the difference $\widehat{U}_{\text{sign flips}} - \widehat{U}_{\text{rotation}}$ for 100 Monte Carlo iterations of a degree-corrected stochastic blockmodel. A Wilcoxon test gives a p-value of $< .0001$ for testing whether the estimated rotation is better than the Sign Flips.

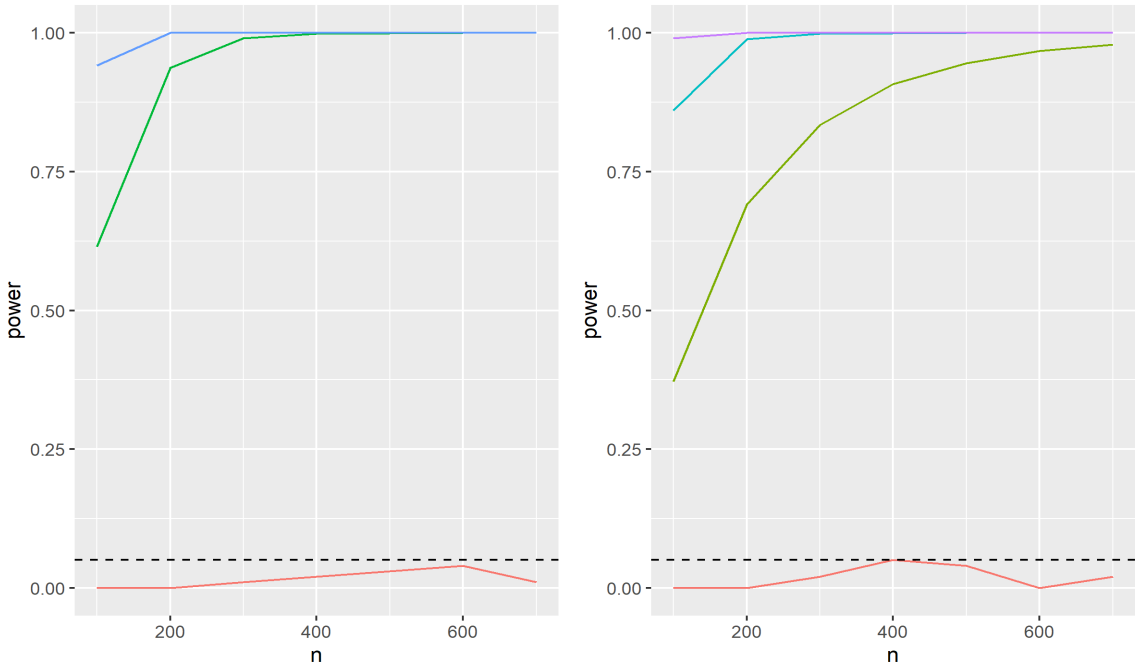


Figure 5.5: Estimated power curves for the stochastic blockmodel (left) and degree-corrected stochastic blockmodel (right) alternatives. In both graphs, red corresponds to the null hypothesis and the other colors correspond to various alternatives as discussed in Section 5.4.1.

a Wilcoxon test gives a p-value of less than .0001 for whether the rotated lies to the left of the naïve alignment. We see that the distribution of the estimated rotation lies to the left of the distribution of sign flips.

5.4.1 Simulated Power Analysis

For the stochastic blockmodel in the previous section, the left hand side of Figure 5.5 gives estimated power curves for the following setup. First, we generate $\mathbf{A}^{(1)}$ as a stochastic blockmodel with $\mathbf{B}^{(1)}$ as before. We then generate $\mathbf{A}^{(2)}$ independently from $\mathbf{A}^{(1)}$ with probability matrix $\mathbf{B}^{(1)} + \varepsilon \mathbf{I}$ for $\varepsilon \in \{0, .1, .2\}$ in red, green, and blue respectively. Note that all these choices of ε still yield a probability generating matrix with negative eigenvalues. We run 100 simulated permutation tests with 500 permutations and choose the critical value for $\alpha = .05$.

Similarly, the right hand side of Figure 5.5 shows estimated power curves for the following setup. First, we generate $\mathbf{A}^{(1)}$ from a stochastic blockmodel with $\mathbf{B}$ as before. Then

we generate $\mathbf{A}^{(2)}$ as the degree-corrected stochastic blockmodel with degree-correction parameters chosen independently as $\beta \times U(0,1) + (1-\beta)$ for various choices of β. When $\beta = 0$ there are no degree-correction parameters, and the null hypothesis holds. As in the previous example, we run this simulation 100 times with 500 permutations per run. We consider $\beta \in \{0, .1, .2, .3\}$ in red, green, blue, and purple respectively. Larger values of β can be understood qualitatively as moving further away from the null hypothesis. We see that under the alternative for larger values of n the estimated power tends to 1, and the Type I error remains below .05 under the null hypothesis.

5.5 Discussion

We have shown that a test statistic defined by using the maximum mean discrepancy applied to the rows of the adjacency spectral embedding yields a consistent test in a natural asymptotic regime as the number of vertices tends to infinity. The methodology we propose shows that solving the optimal transport problem estimates the orthogonal matrix stemming from the eigenvalue multiplicity of the matrices $\mathbb{E}(XX^\top)\mathbf{I}_{p,q}$ and $\mathbb{E}(YY^\top)\mathbf{I}_{p,q}$. While our optimization scheme alternates between points, we note that we have not proven that it yields a globally optimum solution in general, and many different initializations may be required to find the global minimizer. In addition, we note that using the resulting orthogonal transformation, if globally minimized, does not asymptotically affect power in the case of distinct eigenvalues, as the orthogonal transformation it is approximating is a sign matrix.

Our results show that the U-statistic associated to the reproducing kernel yields consistent testing under appropriate edge density; determining the exact nondegenerate limiting distribution is yet an open problem. The proof reveals that it will depend on the asymptotic distribution of the difference of indefinite orthogonal matrices $\sqrt{n}(\mathbf{Q}_\mathbf{X} - \widetilde{\mathbf{Q}}_\mathbf{X})$ in Section E.1, but for practical purposes, this is irrelevant, as the resulting limit will not be independent of F_X and F_Y in general. While exact derivation of the limiting distribution is complicated, we note that our procedure yields a consistent test through a simple bootstrapping procedure. Our main results have demonstrated that only repeated eigenvalues (and not negative eigenvalues) require any modification to obtain consistency for two graph hypothesis testing.

As in Tang et al. (2017b), one can also extend our methodology to determine whether $F_X \simeq F_Y \circ c$ for some constant $c > 0$, or for the setting $F_X \circ \pi \simeq F_Y \circ \pi$, where π is the projection onto the sphere. Here $F_X \simeq F_Y \circ c$ means that $F_X \simeq F_{cY}$, and $F_X \circ \pi \simeq F_Y \circ \pi$ means that $F_{\pi(X)} \simeq F_{\pi(Y)}$. For c appropriately defined so that $F_Y \circ c$ is a valid (p, q)-admissible distribution, one can use the estimates

$$\widehat{s}_X = n^{-1/2}\|\widehat{\mathbf{X}}\|_F; \widehat{s}_Y = m^{-1/2}\|\widehat{\mathbf{Y}}\|_F,$$

and hence, by Lemma 51, these are consistent estimates of the parameters

$$s_X = n^{-1/2}\|\widetilde{\mathbf{X}}\|_F; s_Y = m^{-1/2}\|\widetilde{\mathbf{Y}}\|_F.$$

Similarly, one can project the estimates $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ to the unit sphere to test whether $F_Y \circ \pi \simeq F_Y \circ \pi$.

It remains an open question as to whether our results can be extended to graphons or other random graph models. In addition, while we have demonstrated that estimation of sparsity is sufficient to obtain consistency, graphs below the $\sqrt{n}$ threshold may require additional analysis.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Joint Spectral Clustering for Multilayer Degree-Corrected Stochastic Blockmodels

6.1 Introduction

Community detection, or the problem of clustering the vertices of a network into distinct groups (communities) in a coherent manner that somehow reflects the structure of the network, has become a fundamental tool for the analysis of network data, with many applications in fields such as neuroscience (Sporns and Betzel, 2016), biology (Luo et al., 2007), social sciences (Conover et al., 2011), among others.

In order to understand community detection in networks from a statistical perspective, a number of models have been proposed that characterize edge connectivity probabilities according to some notion of ground-truth communities.

A workhorse community-based statistical model for networks is the *stochastic blockmodel*, which posits that vertices belong to latent communities and that edges are drawn independently, with edge probability determined by the community memberships of each vertex (Holland et al., 1983). A number of works have studied community detection from the lens of the stochastic blockmodel, including deriving information-theoretical limits (Zhang and

Zhou, 2016) and phase transition phenomena (Abbe, 2017). Of the various algorithms proposed for community detection in stochastic blockmodels, *spectral clustering procedures* (von Luxburg, 2007; Rohe et al., 2011; Lei and Rinaldo, 2015), which are collections of clustering techniques that use matrix factorizations such as eigendecompositions and singular value decompositions, have been shown to exhibit good performance both in practice and theoretically, including achieving perfect clustering down to the information-theoretical threshold (Lyzinski et al., 2014; Lei, 2019; Abbe et al., 2020; Su et al., 2020).

One potential drawback of the stochastic blockmodel is that vertices are assumed to be "equivalent" within communities; i.e., edge probabilities are determined *solely* by community memberships. To relax this assumption, in the *degree-corrected stochastic blockmodel* (Karrer and Newman, 2011) each vertex has associated to it a *degree correction parameter* intended to shrink edge probabilities according to its magnitude. On the one hand, the degree-corrected stochastic blockmodel allows for vertex heterogeneity within communities, but on the other hand the model is more general than the stochastic blockmodel, often requiring more sophisticated procedures to recover communities. A number of variants of spectral clustering algorithms for community detection in this model have been considered (Lyzinski et al., 2014; Lei and Rinaldo, 2015; Jin, 2015; Gao et al., 2018), intended to ameliorate the "nuisance" degree correction parameters.

Many modern datasets deal with observations that consists of multiple networks on the same vertex set (Kivelä et al., 2014; Bazzi et al., 2020), denoted as *layers*, such as multiedges or multiview data, networks with time-varying structure, or multiple network observations. Community detection in these data presents additional challenges, as it is important to take advantage of a shared structure in the collection of graphs while respecting individual levels of idiosyncrasy. For these types of network data, which we refer to as *multilayer networks*, perhaps the simplest community-based statistical model is the *multilayer stochastic blockmodel* (Holland et al., 1983). This model posits that communities are shared across networks but that edge probabilities change between networks. Spectral algorithms for multilayer stochastic blockmodels and generalizations have been considered in Lei and Lin (2022); Jing et al. (2021); Pensky and Wang (2021); Pensky and Zhang (2019); Arroyo et al. (2021); Han et al. (2015), though we defer a more thorough discussion of closely related

work to Section 6.1.1.

A key aspect of the multilayer stochastic blockmodel is that it allows for *network heterogeneity* via the possibly changing edge probabilities. However, as in the single network setting, vertices in the multilayer stochastic blockmodel are essentially equivalent; i.e., given their community memberships and the block probability matrices, their edge probabilities are entirely determined. In the *multilayer degree-corrected stochastic blockmodel* that we consider in this work, individual vertices have network-specific degree correction parameters, so that there is *global* network heterogeneity (via the connection probabilities), and *local* vertex heterogeneity (via the degree correction parameters).

Our main contributions are as follows:

- We establish necessary and sufficient conditions for community identifiability of the multilayer degree-corrected stochastic blockmodel and propose a spectral clustering algorithm to estimate community memberships under this model. Our necessary and sufficient conditions for identifiability also hold for the single network setting.
- We obtain an expected misclustering error that improves exponentially with the number of networks, and we demonstrate perfect clustering under sufficient signal strength. Our technical results rely only on signal strength conditions of each network and hold under severe degree heterogeneity within and between networks.
- In simulated data, we demonstrate that our method is competitive in multiple scenarios. Meanwhile, when there is severe heterogeneity across the network layers, state-of-the-art community detection methods can fail in recovering the correct community structure of the model.
- We illustrate the flexibility of the model and methodology in a time series of United States flight network data from January 2016 to September 2021, identifying trends in airport popularity and the influence of COVID-19 on travel both at the local (vertex) and global (community) level.

Our proposed algorithm consists of two stages: first, we compute individual (network-level) spectral embeddings, and then we compute a joint embedding by aggregating the output of the first stage. To prove our main technical results, we develop two separate first-order

entrywise expansions for each stage of our algorithm that explicitly depend on all of the parameters of the model, including degree-corrections. The proof is based on combining the "leave-one-out" analysis technique (Chen et al., 2021c) together with matrix-analytic concentration arguments that carefully track the dependence on the degree-corrections. Furthermore, as a byproduct of our main results we establish an exponential error rate for single network spectral clustering on the scaled eigenvectors with spherical normalization (Theorem 18) that matches the error rate of state-of-the-art spectral methods for single-network degree-corrected blockmodels (Jin, 2015; Jin et al., 2021). This result holds under slightly weaker conditions on the degree-correction parameters, and slightly stronger conditions on the community separation.

The rest of this paper is structured as follows. In Section 6.1.1 we consider closely related work, and in Section 6.1.2 we set notation. We present our model and identifiability in Section 6.2, and we present our algorithm in Section 6.2.1. The main results are presented in Section 6.3, and our simulations and real data analysis are presented in Section 6.4 and Section 6.5 respectively. We finish in Section 6.6 with a discussion, and we prove our main results in Section 6.7. The full proofs of all of our results are in the appendices.

6.1.1 Related Work

Community detection in the single network setting has received widespread attention in recent years (Abbe, 2017; Fortunato and Newman, 2022). A number of works have studied community detection in the stochastic blockmodel, including consistency (Rohe et al., 2011; Zhao et al., 2012; Lei and Rinaldo, 2015), phase transition phenomena (Abbe et al., 2020) and minimax rates (Gao et al., 2018). Beyond the stochastic blockmodel, a number of inference techniques have been considered for generalizations, such as the mixed-membership blockmodel (Airoldi et al., 2008; Mao et al., 2021), the random dot product graph (Athreya et al., 2018) and generalised random dot product graph (Rubin-Delanchy et al., 2022). This work is closely related to the literature on degree-corrected stochastic blockmodels (Karrer and Newman, 2011). The work Jin (2015) considered community detection in degree-corrected stochastic blockmodels using SCORE, or spectral clustering on ratios of eigenvectors, and several refinements, generalizations, and applications of this procedure have been consid-

ered, including Jin et al. (2019, 2021); Ke and Wang (2022) and Fan et al. (2022). Our main results are perhaps most similar to Jin et al. (2021), who obtain an exponential error rate for spectral clustering with the SCORE procedure for a single network.

Turning to community detection in multilayer networks, several procedures have been considered for the multilayer stochastic blockmodel, including spectral methods Han et al. (2015); Bhattacharyya and Chatterjee (2018, 2020); Huang et al. (2020b); Lei and Lin (2022), matrix factorization approaches (Paul and Chen, 2020; Lei et al., 2020), the expectation-maximization algorithm (De Bacco et al., 2017), and efficient MCMC approaches Peixoto (2015); Bazzi et al. (2020). Extensions have also been considered, such as Chen et al. (2021a), which allows some members of each community to switch between networks. Furthermore, Jing et al. (2021); Pensky and Wang (2021), and Noroozi and Pensky (2022) all consider generalizations of the multilayer stochastic blockmodel where there are a few different possible community configurations. Other spectral methods for multilayer data have considered different low-rank models Levin et al. (2019); Draves and Sussman (2021); Jones and Rubin-Delanchy (2021); Pantazis et al. (2022); Arroyo et al. (2021); Zheng and Tang (2022); MacDonald et al. (2022). Although spectral methods are competitive in terms of computation and accuracy, existing methods are limited in handling heterogeneous degree correction parameters. Both Bhattacharyya and Chatterjee (2020) and Bhattacharyya and Chatterjee (2018) consider degree-corrections for each network, but they require that the degree-corrections remain the same across networks, making the analysis feasible. Our work is perhaps most closely connected to the works Arroyo et al. (2021) and Zheng and Tang (2022), which consider the estimation of a common invariant subspace, but the model we consider in this paper is substantially different, and we provide finer theoretical results to analyze misclustering rates.

From a technical point of of view, our analysis is also closely related to the literature on entrywise eigenvector analysis of random matrices (Abbe et al., 2020; Chen et al., 2021c), of which there has been applications to networks (Cape et al., 2019a; Abbe et al., 2020; Su et al., 2020; Mao et al., 2021; Jin et al., 2019, 2021; Ke and Wang, 2022; Abbe et al., 2022), principal component analysis (Cape et al., 2019b; Cai et al., 2021a; Yan et al., 2021; Agterberg and Sulam, 2022), high-dimensional mixture models (Abbe et al., 2022; Agterberg

et al., 2022b; Zhang and Zhou, 2022), randomized algorithms (Zhang and Tang, 2022), ranking from paired data (Chen et al., 2019a, 2022), tensor data analysis (Cai et al., 2021a; Xia and Zhou, 2019), among others. In the single-network community detection setting, several authors have previously considered the entrywise analysis of the eigenvectors of a single degree-corrected stochastic blockmodel, such as Lyzinski et al. (2014); Jin et al. (2019, 2021); Su et al. (2020) and Ke and Wang (2022). Here we provide an entrywise analysis of the *scaled* eigenvectors of degree-corrected stochastic blockmodels with explicit dependence on the degree-correction parameters that is needed for the analysis of the multilayer embedding.

6.1.2 Notation

We use bold or greek capital letters $\mathbf{M}$ or Λ for matrices, and we let $\mathbf{M}_{i\cdot}$ and $\mathbf{M}_{\cdot j}$ denote the i'th row and j'th column respectively, where we view both as column vectors. We let $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_{2,\infty}$ denote its spectral and $\ell_{2,\infty}$ norm, where the latter is defined as $\max_i \|\mathbf{M}_{i\cdot}\|$, where $\|\mathbf{M}_{i\cdot}\|$ is the usual (vector) Euclidean norm. For a vector x we let $\|x\|_1, \|x\|_\infty$ denote its vector ℓ_1 and ℓ_∞ norms respectively. We let $\mathbf{I}_r$ denote the $r \times r$ identity. For two orthonormal matrices $\mathbf{U}$ and $\mathbf{V}$ satisfying $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$, we let $\|\sin\boldsymbol{\Theta}(\mathbf{U},\mathbf{V})\|$ denote their (spectral) $\sin\boldsymbol{\Theta}$ distance, defined as $\|\sin\boldsymbol{\Theta}(\mathbf{U},\mathbf{V})\| = \|(\mathbf{I}-\mathbf{U}\mathbf{U}^\top)\mathbf{V}\|$. We write $\mathbb{O}(r)$ to denote the set of $r\times r$ orthogonal matrices; i.e. $\mathbf{W} \in \mathbb{O}(r)$ if $\mathbf{W}\mathbf{W}^\top = \mathbf{I}_r$. We also denote e_i as the standard basis vector, and, where appropriate, we view $e_i^\top\mathbf{M}$ as a column vector. We let $\mathbb{I}\{\cdot\}$ denote the indicator function, and $\mathbb{R}_+$ denote the strictly positive real numbers.

We will also use asymptotic notation throughout this paper. For two functions $f(n)$ and $g(n)$, we write $f(n) \lesssim g(n)$ if there exists some constant $C > 0$ such that $f(n) \le Cg(n)$, and we write $f(n) \ll g(n)$ if $f(n)/g(n) \to 0$ as $n \to \infty$. We denote by $f(n) \asymp g(n)$ the case where both $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$. We also write $f(n) = O(g(n))$ if $f(n) \lesssim g(n)$.

We denote $[n] = \{1, 2, \ldots, n\}$, and we use L and l to refer to individual networks, i and j to refer to nodes, and K, r, and s to refer to communities.

6.2 The Multilayer Degree-Corrected Stochastic Blockmodel

Suppose one observes a collection of L adjacency matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(L)}$ of size $n \times n$, with the vertices of the corresponding graphs aligned across the collection. For simplicity of the presentation and the theory, we assume that the adjacency matrices represent simple undirected graphs, hence these matrices are symmetric with binary entries, and we allow the networks to have self-edges (loops), but the main results are not materially different if loops are not permitted. Much of the theory and methodology we consider here is also applicable in the settings of weighted or directed networks, but we focus on the binary and undirected setting since our primary concern in the present work is to quantify the misclustering error rate as a function of the degree parameters.

The model considered in this paper assumes a shared community structure across all the graphs, but allows for idiosyncrasy in the edge probabilities across the collection of graphs by letting the global and local individual parameters of each graph to be different. In particular, we consider a multilayer version of the degree-corrected stochastic blockmodel (Karrer and Newman, 2011), in which both the block connectivity matrices and the vertex degree parameters can be different for each network. Some versions of this model have appeared in Peixoto (2015); Bazzi et al. (2020); Bhattacharyya and Chatterjee (2020), but to be precise, we will use the following definition.

Definition 5 (Multilayer Degree-Corrected Stochastic Blockmodel)**.** A collection of L graphs $\{\mathbf{A}^{(l)}\}_{l=1}^{L}$ on n vertices are drawn from the *multilayer degree-corrected stochastic blockmodel* (multilayer DCSBM) if:

- each vertex i belongs to one of K communities. Let $z : [n] \to [K]$ be the community membership function satisfying $z(i) = r$ if vertex i belongs to community r;
- $\theta_1^{(l)}, \ldots, \theta_n^{(l)} \in \mathbb{R}_+$ are the *degree correction parameters* associated to each vertex i in network l;
- $\mathbf{B}^{(1)}, \ldots, \mathbf{B}^{(L)} \in \mathbb{R}_+^{K \times K}$ are symmetric *block connectivity matrices*;
- the edges of the networks are mutually independent, and their expected values (prob-

abilities) are described by

$$\mathbb{E}[\mathbf{A}_{ij}^{(l)}] = \theta_i^{(l)}\theta_j^{(l)}\mathbf{B}_{z(i),z(j)}^{(l)}, \qquad l \in [L],\ i,j \in [n], i \geq j.$$

The degree correction parameters denote a local connectivity component and the block connection probability matrices characterize a global connectivity component, both of which can vary from graph to graph, while the community memberships remain constant. Since the edges are binary, the expected value also denotes the probability of the corresponding edge, but this definition can be used in other distributions (e.g. Poisson (Karrer and Newman, 2011)).

It is convenient to represent the multilayer DCSBM using matrix notation. Denote the collection of matrices that encode the edge expectations by $\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(L)} \in [0,1]^{n\times n}$, such that $\mathbb{E}[\mathbf{A}_{ij}^{(l)}] = \mathbf{P}_{ij}^{(l)}$ for each $l \in [L]$ and $i,j \in [n], i \geq j$. Then we can write

$$\mathbf{P}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}\mathbf{Z}^\top\mathbf{\Theta}^{(l)}, \qquad l \in [L], \tag{6.1}$$

where $\mathbf{\Theta}^{(l)} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $\mathbf{\Theta}_{ii}^{(l)} = \theta_i^{(l)} > 0$, $\mathbf{Z} \in \{0,1\}^{n\times K}$ is a binary matrix indicating community memberships ($\mathbf{Z}_{ir} = 1$ if $z(i) = r$, and $\mathbf{Z}_{ir} = 0$ otherwise), and $\mathbf{B}^{(l)} \in \mathbb{R}_+^{K\times K}$ is a symmetric matrix proportional to the connectivity between and within communities in the graph l. We assume that $\operatorname{rank}(\mathbf{B}^{(l)}) = K_l$, and we allow K_l to be less than K.

The multilayer DCSBM model is flexible enough to represent heterogeneous structures both at the vertex and the community levels, while retaining a joint community structure across the graphs. Due to these local and global idiosyncrasies, distinguishing between local and global graph structure at the single and multilayer level becomes important, as it is possible to formulate parameterizations of the model that give equivalent characterizations. For instance, one may group high degree vertices in their own community according to degree correction parameters alone. To ensure identifiability and maintain a parsimonious model, we assume that the number of communities K is the smallest possible that can represent the communities uniquely (up to label permutations). Our first result establishes

the identifiability of the communities in the model.

Theorem 15 (Community membership identifiability)**.** *Suppose that* $\{\mathbf{P}^{(l)}\}_{l=1}^{L} \in \mathbb{R}^{n\times n}$ *are matrices such that*

$$\mathbf{P}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}\mathbf{Z}^{\top}\mathbf{\Theta}^{(l)}, \qquad l = 1, \dots, L,$$

where $\mathbf{Z} \in \{0,1\}^{n\times K}$ *is a binary block membership matrix with at least one vertex in each community (* $\sum_{r=1}^{K}\mathbf{Z}_{ir} = 1, i \in [n]$*, and* $\sum_{i=1}^{n}\mathbf{Z}_{ir} \geq 1$*,* $r \in [K]$*),* $\{\mathbf{B}^{(l)}\}_{l=1}^{L}$ *are symmetric matrices with entries in* $\mathbb{R}_{+}$*, and* $\{\mathbf{\Theta}^{(l)}\}_{l=1}^{L}$ *are diagonal matrices with positive entries on the diagonal. Let* $\mathbf{B}^{(l)} = \mathbf{V}^{(l)}\mathbf{D}^{(l)}(\mathbf{V}^{(l)})^{\top}$ *be the eigendecomposition of* $\mathbf{B}^{(l)}$*, with* $\mathbf{V}^{(l)} \in \mathbb{R}^{K\times K_l}$ *a matrix with orthonormal columns and* $\mathbf{D}^{(l)} \in \mathbb{R}^{K_l}$ *a diagonal matrix, and rank*$(\mathbf{B}^{(l)}) = K_l$*. Write* $\mathbf{Q}^{(l)}$ *as the matrix with normalized rows of* $\mathbf{V}^{(l)}$*, i.e.,* $\mathbf{Q}^{(l)}_{r\cdot} = \frac{1}{\|\mathbf{V}^{(l)}_{r\cdot}\|}\mathbf{V}^{(l)}_{r\cdot}$*, and let* $\mathbf{Q} = [\mathbf{Q}^{(1)}, \cdots, \mathbf{Q}^{(L)}]$*. The membership matrix* $\mathbf{Z}$ *is identifiable (up to label permutations) if and only if* $\mathbf{Q}$ *has no repeated rows.*

In essence, the identifiability condition requires that the matrices $\{\mathbf{B}^{(l)}\}$ have exactly K jointly distinguishable rows, which determine the community memberships. The condition $\mathbf{Q}$ having no repeated rows implies that there are precisely K unique *directions* associated to the rows of $[\mathbf{V}^{(1)}, \cdots, \mathbf{V}^{(L)}]$. Therefore, if $\widetilde{\mathbf{Q}}^{(l)}$ is defined by normalizing the rows of $\mathbf{V}^{(l)}$ in any other way, then as long as $\widetilde{\mathbf{Q}}$ has K distinct rows the communities will be identifiable. A key feature of Theorem 15 is that it also holds for $L = 1$, thereby establishing both necessary and sufficient conditions for identifiability in the single network model.

The matrix $\mathbf{B}^{(l)}$ is often assumed to be full rank (Qin and Rohe, 2013; Jin et al., 2022), in which case there are exactly K identifiable communities. Theorem 15 requires a milder condition to allow flexibility in modeling multiple networks, as the number of identifiable communities in each layer may be smaller than K. The identifiable communities in the joint model are given by the different directions taken by the combined rows of $\mathbf{B}^{(l)}$ across all the layers. For instance, two communities may have the same row directions in a particular network layer, but when considered as an ensemble, all of the communities have distinct directions. Since this condition is also necessary for identifiability, this value of K gives the most parsimonious representation in terms of the number of communities.

Identifiability of the degree correction and block connectivity parameters requires additional constraints, as it is otherwise possible to change their values up to a multiplicative constant. Multiple characterizations have been used previously for the single-network setting, and these immediately extend to the multilayer setting. For instance, if for all i, we have $\mathbf{B}^{(l)}_{ii} = 1$, (e.g. Jin et al. (2022)) then the other model parameters are identifiable as well. We adopt this identifiability constraint to facilitate the presentation of the theoretical results in Section 6.3, as it allows us to isolate the effect of the degree correction parameters. Nevertheless, to ease interpretation, in Section 6.5 we adopt a different constraint, namely, that the sum of degree corrections within each community is equal to 1. Both parameterizations are equivalent, and the corresponding normalizing constants can either be absorbed into the degree corrections or the block connectivity matrices.

Remark 18 (Relationship to (Generalized) Random Dot Product Graphs)**.** *We note that, in the single-network setting, Theorem 15 can also be interpreted through the lens of the (generalized) random dot product graph (GRDPG) (Rubin-Delanchy et al., 2022; Athreya et al., 2018). Under the GRDPG model, each vertex is associated with a latent position in low-dimensional Euclidean space. Under this framework, Theorem 15 shows that the communities in the DCSBM are identifiable if and only if the latent positions associated to each vertex lie on exactly K unique rays emanating from the origin.*

6.2.1 DC-MASE: Degree-Corrected Multiple Adjacency Spectral Embedding

In order to obtain a statistically principled, computationally efficient, and practical algorithm for community detection, we will consider a spectral clustering procedure. General spectral clustering approaches for one network typically proceed in a standard manner: first, using a few leading eigenvectors of the adjacency matrix (or related quantities, such as the graph Laplacian), obtain individual vertex representations by considering the rows of the matrices; we will refer to this first step as obtaining an *embedding*. Then, the communities are estimated by clustering the rows of this matrix using a clustering algorithm such as K-means or K-medians.

For multilayer networks with shared community structure, the general procedure is similar, only now the requirement is to use all of the networks to obtain individual vertex representations in a low-dimensional space. For the multilayer stochastic blockmodel, a typical approach is to simply consider a few leading eigenvectors of the average adjacency matrix $\bar{\mathbf{A}} = \frac{1}{L}\sum_l \mathbf{A}^{(l)}$ (Tang et al., 2009; Han et al., 2015). However, as discussed in e.g. Paul and Chen (2020); Lei and Lin (2022), this procedure is only guaranteed to work when there is certain level of homogeneity in the block connectivity matrices, and it can fail if the $\mathbf{B}^{(l)}$ matrices are different. Lei and Lin (2022) proposed to rectify this by considering a bias-corrected version of the sum of the squared adjacency matrices. Alternatively, one can look at an embedding obtained by aggregating the projections onto the principal subspaces of each graph (Paul and Chen, 2020; Arroyo et al., 2021). In these situations, the population probability matrices $\{\mathbf{P}^{(l)}\}$ share a common singular subspace, and running the relevant algorithm on those reveals the community memberships. Unfortunately, this is not the case in the setting considered herein, but with some modification, a certain matrix can be shown to have a left singular subspace that reveals the community memberships.

Our proposal to find an embedding is based on several observations concerning the joint spectral geometry of the matrices $\{\mathbf{P}^{(l)}\}$, some of which have been considered before in the single-network literature (Qin and Rohe, 2013; Lyzinski et al., 2014; Lei and Rinaldo, 2015; Jin, 2015; Su et al., 2020).

- **Observation 1**: *The rows of the K_l scaled eigenvectors of $\mathbf{P}^{(l)}$ are supported on at most K different rays in $\mathbb{R}^{K_l}$, with each ray corresponding to a distinct community, and magnitude of each row determined by the magnitude of its corresponding degree-correction parameter.*

Suppose that each $\mathbf{P}^{(l)}$ has eigendecomposition $\mathbf{U}^{(l)}\Lambda^{(l)}(\mathbf{U}^{(l)})^\top$, where $\mathbf{U}^{(l)}$ is an $n \times K_l$ orthonormal matrix and $\Lambda^{(l)}$ is the matrix of eigenvalues of $\mathbf{P}^{(l)}$. Define

$$\widetilde{\mathbf{X}}^{(l)} := \mathbf{U}^{(l)}|\Lambda^{(l)}|^{1/2}, \tag{6.2}$$

where $|\cdot|$ is the entrywise absolute value. It can be shown (see the proof of Proposition 9 below) that $\widetilde{\mathbf{X}}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{M}^{(l)}$, where $\mathbf{M}^{(l)} \in \mathbb{R}^{K \times K_l}$ has K_l' unique rows, with $K_l \leq K_l' \leq K$.

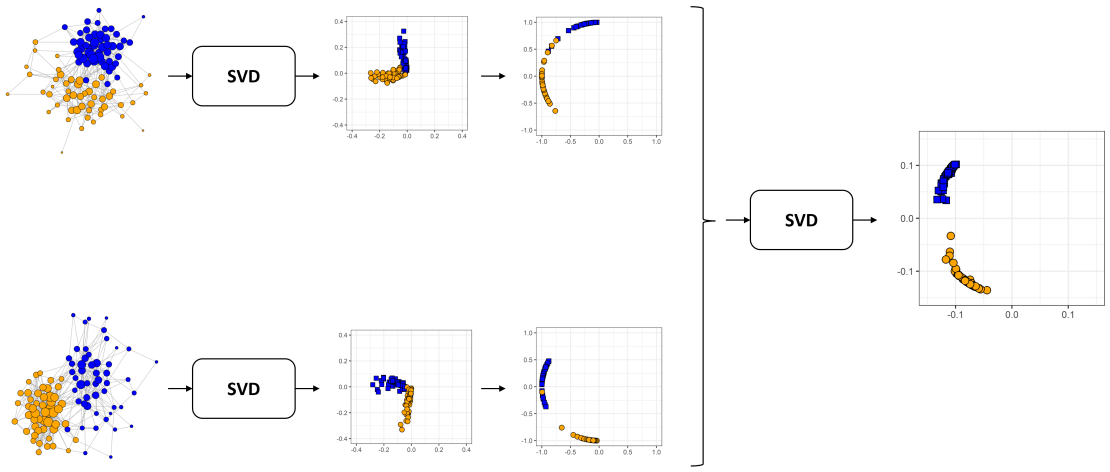


Figure 6.1: Pictorial representation of Algorithm 8.

Explicitly, this observation implies that that each row i of $\widetilde{\mathbf{X}}^{(l)}$ satisfies

$$\widetilde{\mathbf{X}}_{i\cdot}^{(l)} = \theta_i^{(l)} \mathbf{M}_{z(i)\cdot}^{(l)}. \tag{6.3}$$

- **Observation 2**: *Projecting each row of $\widetilde{\mathbf{X}}^{(l)}$ to the sphere results in a matrix of at most K unique rows, with each row corresponding to community membership.*

Define $\mathbf{Y}^{(l)}$ via

$$\mathbf{Y}_{i\cdot}^{(l)} = \frac{\widetilde{\mathbf{X}}_{i\cdot}^{(l)}}{\|\widetilde{\mathbf{X}}_{i\cdot}^{(l)}\|}.$$

By (6.3), it holds that $\mathbf{Y}_{i\cdot}^{(l)} = \frac{\mathbf{M}_{z(i)\cdot}^{(l)}}{\|\mathbf{M}_{z(i)\cdot}^{(l)}\|}$. In particular, there are only $K_l' \leq K$ unique rows of $\mathbf{Y}^{(l)}$, with each row corresponding to community membership.

- **Observation 3**: *The left singular subspace of $\mathcal{Y} = [\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \cdots, \mathbf{Y}^{(L)}] \in \mathbb{R}^{n \times \sum_l K_l}$ reveals the community memberships.*

Suppose that $\mathcal{Y}$ has singular value decomposition given by $\mathcal{Y} = \mathbf{U}\Sigma\mathbf{V}^\top$. It can be shown (see Proposition 9) that under the condition of Theorem 15, $\text{rank}(\mathcal{Y}) = \widetilde{K} \leq K$ and $\mathbf{U} \in \mathbb{R}^{n \times \widetilde{K}}$

satisfies

$$\mathbf{U} = \mathbf{ZM},$$

where $\mathbf{M} \in \mathbb{R}^{K \times \widetilde{K}}$ is some matrix without repeated rows. Explicitly, this says that there are only K unique rows of $\mathbf{U}$, with each row i of $\mathbf{U}$ corresponding to community membership of vertex i. Moreover, since $\mathbf{U}$ is obtained via the singular value decomposition of $\mathcal{Y}$, it contains information from all the networks. Clustering the rows of the matrix $\mathbf{U}$ therefore reveals the community memberships.

The observations presented above lead to a joint spectral clustering algorithm applied to the sample adjacency matrices, which is summarized in Algorithm 8 and a pictorial representation is shown in Fig. 6.1. Without the row-normalization step, one obtains the *scaled* multiple adjacency spectral embedding (MASE) algorithm of Arroyo et al. (2021), who consider the COSIE (COmmon Subspace Independent Edge) model where each "population" network shares a common invariant subspace (which includes the multilayer stochastic blockmodel as a special case). Due to the different degree correction parameters, the multilayer DCSBM model is not a particular instance of the COSIE model, but our algorithm can be viewed as a normalized version of the MASE algorithm, so we dub it DC-MASE, or degree-corrected multiple adjacency spectral embedding. Introducing this normalization step is crucial in the presence of heterogeneous degree correction parameters and makes this methodology applicable to a much more flexible model. The following proposition formalizes the three arguments to construct the algorithm. The proof of this result can be found in Appendix F.1.

Proposition 9. *Under the conditions of Theorem 15, Algorithm 8 applied to the collection of matrices* $\mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(L)}$ *recovers the community memberships exactly.*

Variations of Algorithm 8 can be obtained by changing the initial embedding, row-normalization, or clustering procedures, for which we conjecture that similar results to Proposition 9 may hold, but we do not undertake a complete analysis of these different choices in the present work. In particular, here we consider the adjacency spectral embedding that uses the scaled eigenvectors in Eq. (6.2) due to their interpretation as the

Algorithm 8 Degree-corrected multiple adjacency spectral embedding (DC-MASE)

Require: Collection of adjacency matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(L)}$; individual ranks $K_1, \ldots, K_L$, joint rank $\widetilde{K}$, number of communities K.

1. For each graph $l \in [L]$,
   (a) Let $\widehat{\mathbf{X}}^{(l)} \in \mathbb{R}^{n \times K_l}$ be defined $\widehat{\mathbf{X}}^{(l)} := \widehat{\mathbf{U}}^{(l)} |\widehat{\Lambda}^{(l)}|^{1/2}$, where $\widehat{\mathbf{U}}^{(l)}$ is the matrix containing the K_l eigenvectors associated to the K_l largest eigenvalues (in magnitude) of $\mathbf{A}^{(l)}$ and $\widehat{\Lambda}^{(l)}$ are the corresponding eigenvalues;
   (b) let $\widehat{\mathbf{Y}}^{(l)} \in \mathbb{R}^{n \times K_l}$ be the matrix containing the rows of $\widehat{\mathbf{X}}^{(l)}$ projected to the sphere, defined as
$$\widehat{\mathbf{Y}}^{(l)}_{i\cdot} = \frac{\widehat{\mathbf{X}}^{(l)}_{i\cdot}}{\|\widehat{\mathbf{X}}^{(l)}_{i\cdot}\|}$$
2. Form the matrix $\widehat{\mathcal{Y}} = [\widehat{\mathbf{Y}}^{(1)}, \cdots \widehat{\mathbf{Y}}^{(L)}]$ by concatenating the row-scaled eigenvector matrices.
3. Let $\widehat{\mathbf{U}} \in \mathbf{R}^{n \times \widetilde{K}}$ be the matrix containing the $\widetilde{K}$ leading left singular values of $\widehat{\mathcal{Y}}$.
4. Assign community memberships as the clusters of the rows of $\widehat{\mathbf{U}}$ into K groups via K-means.

return Community memberships.

estimated latent positions of a (generalized) random dot product graph (RDPG) (Sussman et al., 2012; Rubin-Delanchy et al., 2022). Other variations can be obtained by changing the embedding, for example, to unscaled eigenvectors or using the Laplacian matrix; the normalization procedure, for example, by using SCORE (Jin, 2015); or by changing the clustering procedure to K-medians (Lei and Rinaldo, 2015) or Gaussian mixture modeling (Athreya et al., 2016).

6.2.2 Estimating the Number of Communities

Choosing the number of communities in the multilayer DCSBM via DC-MASE is an important yet challenging problem, as one is required to estimate the individual and joint embedding dimensions for each adjacency matrix, as well as the total number of communities in the joint model. Throughout this paper, we assume that these numbers are known or can be estimated appropriately, but we discuss here some approaches for choosing these parameters in practice.

The first step of Algorithm 8 requires the selection of K_l, which corresponds to the rank

of the matrix $\mathbf{P}^{(l)} = \mathbb{E}[\mathbf{A}^{(l)}]$, and hence this corresponds to a rank estimation problem. A common practical approach is to look for an elbow in the scree plot of the eigenvalues of the adjacency matrix (Zhu and Ghodsi, 2006). Similarly, to estimate $\widetilde{K}$, one can look for elbows in the scree plot of the singular values obtained from the concatenated matrix $\widehat{\mathcal{Y}}$, as this matrix concentrates around a population matrix that has rank exactly equal to $\widetilde{K}$. In simulations, we have observed that overestimating these parameters typically does not have a significant effect on the performance of the clustering method.

The choice of K is more important, as it controls the number of communities in the joint model. Several existing methods assume that the matrix $\mathbf{B}^{(l)}$ has full rank, in which case the value of K_l corresponds to the number of communities in the degree-corrected SBM for each network $l \in [L]$. A number of methods exist for estimating the communities in a single-layer DCSBM (Wang and Bickel, 2017; Ma et al., 2021; Le and Levina, 2022; Li et al., 2020b; Han et al., 2020), including recent work by Jin et al. (2022), who achieves the optimal phase transition under this assumption. Alternatively, one can use an appropriate criterion for choosing the number of clusters via K-means.

6.3 Main Results

Having described our algorithm in detail, we are now prepared to discuss the associated community recovery guarantees. In order to do so, we first must state some assumptions on the regularity of each network. For simplicity of analysis and to facilitate interpretation, we assume that $\mathbf{B}^{(l)}_{rr} = 1$ for all $r \in [K], l \in [L]$, and that each $\mathbf{B}^{(l)}$ is rank K, but our main results continue to hold as long as the K-th smallest singular value of $\mathcal{Y}$ grows sufficiently quickly (see the supplementary materials for the details).

Assumption 6.1 (Regularity Conditions)**.** *Let $\mathcal{C}(r)$ denote the indices associated to community r; i.e., the set of i such that $z(i) = r$. It holds that $|\mathcal{C}(r)| \asymp |\mathcal{C}(s)|$ for $r \neq s$ and $K\|\theta^{(l)}_{\mathcal{C}(r)}\|^2 \asymp \|\theta^{(l)}\|^2$ for all $r \in [K]$. In addition, each matrix $\mathbf{B}^{(l)}$ is rank K with unit diagonals; let $\lambda^{(l)}_t$ denote its ordered eigenvalues. Then $|\lambda^{(l)}_K| \geq \lambda^{(l)}_{\min}$ for some $\lambda^{(l)}_{\min} \in (0, 1)$ and $\|\mathbf{B}^{(l)}\| = \lambda^{(l)}_1 \asymp 1$.*

The first part of Assumption 6.1 essentially requires that the communities and degree

corrections within each community are balanced, and it is commonly imposed in the analysis of the DCSBM Jin et al. (2022); Su et al. (2020), but it can be relaxed by keeping track of these constants. We also assume for simplicity that $\|\mathbf{B}^{(l)}\| \leq C$, which is not strictly required but facilitates analysis.

We have introduced the parameter $\lambda_{\min}^{(l)}$, which can be understood as a proxy for the community separation. For example, consider the matrix $\mathbf{B}^{(l)} = \begin{pmatrix} 1 & 1-\eta \\ 1-\eta & 1 \end{pmatrix}$. Then it holds that $\lambda_{\min}^{(l)} = \eta$. As $\eta \to 0$, the communities become more similar. We also assume for simplicity that $\lambda_{\min}^{(l)} \in (0,1)$; when this is not the case, the communities are well-separated, so the problem is qualitatively easier. Therefore, the assumption that $\lambda_{\min}^{(l)} \in (0,1)$ restricts our analysis to the models for which community detection is more difficult. Also, as seen in the example just above, when $\mathbf{B}^{(l)}$ is positive semidefinite, we necessarily have $\lambda_{\min}^{(l)} \in (0,1)$.

Next we introduce some assumptions on the individual network-level signal strengths and degree homogeneity. Let $\theta_{\min}^{(l)} := \min_i \mathbf{\Theta}_{ii}^{(l)}$, and let $\theta_{\max}^{(l)}$ be defined similarly. Define also the following average minimum eigenvalue parameter:

$$\bar{\lambda} := \frac{1}{L}\sum_{l=1}^{L} \lambda_{\min}^{(l)} \in (0,1).$$

The following is our main technical assumption on the individual network signal strengths.

Assumption 6.2 (Network-Level Signal Strengths)**.** *There exist constants C and c (with C depending on the community sizes) such that each network l satisfies*

$$C\left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\right)\frac{K^8\theta_{\max}^{(l)}\|\theta^{(l)}\|_1\log(n)}{(\lambda_{\min}^{(l)})^2\|\theta^{(l)}\|^4} \leq \bar{\lambda}; \qquad \textit{(Signal Strength)}$$

$$\frac{\theta_{\min}^{(l)}}{\theta_{\max}^{(l)}} \geq \sqrt{\frac{\log(n)}{n}} \qquad \textit{(Degree Heterogeneity)}$$

$$\theta_{\min}^{(l)}\|\theta^{(l)}\|_1 \geq c\log(n). \qquad \textit{(Logarithmic Degree Growth).}$$

To build intuition we consider several examples.

Example 6.1 (Degree-Correction Heterogeneity)**.** We consider a setting with $\lambda_{\min} \asymp 1$, $K \asymp 1$ and we suppose that $\theta_i^{(l)} = a$, for $1 \leq i \leq \gamma n$ and $\theta_i^{(l)} = b > a$ for $\gamma n + 1 \leq i \leq n$. It

is easy to show that Assumption 6.2 holds if $\frac{b^2(\gamma a+(1-\gamma)b)}{a(\gamma a^2+(1-\gamma)b^2)^2} \lesssim \frac{n}{\log(n)}$ and $a/b \gtrsim \sqrt{\frac{\log(n)}{n}}$. For example, if $\gamma n = 1$ (an outlier model) and $b \gg a$, the first condition reduces to $ab \gtrsim \log(n)/n$. If $b = 1$, $a = \sqrt{\log(n)/n}$ satisfies the degree heterogeneity assumption.

Example 6.2 (Close Communities with Homogeneous Degree Corrections)**.** We consider a setting with all $\theta_i^{(l)} \asymp \sqrt{\rho_n}$, $K \asymp 1$, and $\lambda_{\min}^{(l)} \asymp \lambda_{\min}$ for all l. Then we require $\lambda_{\min}^3 \gtrsim \frac{\log(n)}{n\rho_n}$. If only $o(L)$ networks have $\lambda_{\min}^{(l)} \asymp \lambda_{\min}$, and all others have $\lambda_{\min}^{(l)} \asymp 1$, then we have the weaker condition $\lambda_{\min}^2 \gtrsim \frac{\log(n)}{n\rho_n}$. Then so long as the majority of networks have strong signal, we can tolerate even weaker signal in the worst-behaved layers.

Assumption 6.2 is markedly similar to the main technical assumptions in Jin et al. (2021) for a single network. When $L = 1$, our condition in Assumption 6.2 is only slightly stronger than that of Jin et al. (2021) in terms of $\bar{\lambda}$ and slightly weaker in terms of $\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}$, though we include a more detailed comparison in Section 6.3.2. To understand the intuition behind the signal-strength condition in terms of $\bar{\lambda}$ in Assumption 6.2, recall that the second step of DC-MASE requires taking the left singular vectors of the matrix $\mathcal{Y}$. When $\bar{\lambda}$ is small, the average community separation is small, and hence the rays associated to each (unscaled) embedding $\widetilde{\mathbf{X}}^{(l)}$ (see (6.3)) will be nearly colinear. Consequently, in this regime, the K unique rows of $\mathcal{Y}$ can be quite close, so $\mathcal{Y}$ is nearly rank degenerate, and hence the second SVD step will not be as stable. Therefore, in order for the SVD step to succeed, we will require sufficient separation of the communities, which is why Assumption 6.2 concerns $\bar{\lambda}$. This phenomenon will also manifest in our main results in the following subsection.

6.3.1 Misclustering Error Rate and Perfect Clustering for Multilayer Networks

With these assumptions in hand, we are now prepared to state our main results. For technical reasons we use $(1+\varepsilon)$ K-means to obtain our estimate. Let $\widehat{z}$ denote the estimated clustering by applying $(1+\varepsilon)$ K-means to DC-MASE; i.e. $\widehat{z}(i) = r$ if node i is estimated to belong to community r. Let z denote the true clustering. We define

$$\ell(\widehat{z}, z) := \inf_{\text{Permutations } \mathcal{P}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\widehat{z}(i) \neq \mathcal{P}(z(i))\}.$$

In other words $\ell(\widehat{z}, z)$ is the misclustering error up to label permutations. The following theorem is our main technical result, demonstrating an upper bound on the misclustering error.

Theorem 16. *Suppose that Assumption 6.1 and Assumption 6.2 are satisfied, and suppose that $L \lesssim n^5$. Define*

$$\mathrm{err}_{\mathrm{ave}}^{(i)} := \frac{1}{L} \sum_{l} \frac{\|\theta^{(l)}\|_3^3}{\theta_i^{(l)} \|\theta^{(l)}\|^4 \lambda_{\min}^{(l)}}; \tag{6.4}$$

$$\mathrm{err}_{\max}^{(i)} := \max_{l} \frac{\theta_{\max}^{(l)}}{\theta_i^{(l)} \|\theta^{(l)}\|^2 (\lambda_{\min}^{(l)})^{1/2}}. \tag{6.5}$$

Then there exists a sufficiently small constant c depending on the implicit constants in the assumptions such that the expected misclustering error is

$$\mathbb{E}\ell(\widehat{z}, z) \leq \frac{2K}{n} \sum_{i=1}^{n} \exp\left(- cL \min\left\{ \frac{\bar{\lambda}^2}{K^4 \mathrm{err}_{\mathrm{ave}}^{(i)}}, \frac{\bar{\lambda}}{K^2 \mathrm{err}_{\max}^{(i)}} \right\} \right) + O(n^{-10}).$$

The assumption that $L \lesssim n^5$ is primarily for technical convenience; this is made so that we can take a union bound over all L networks. If L is larger but still polynomial in n, the result can still hold at the cost of increasing all of the implicit constants in the assumptions. However, once L is sufficiently large relative to n, the exponent can be made to be smaller than e^{-cn} for some constant c, and hence by Markov's inequality one obtains that $\mathbb{P}(\ell(\widehat{z}, z) \geq \frac{1}{n}) \leq ne^{-cn} + O(n^{-9}) = O(n^{-9})$, which shows that all vertices are recovered correctly with high probability. Therefore, while our theory only covers L growing polynomially with n, for all practical purposes this assumption is irrelevant, as perfect clustering will be guaranteed once L is larger than some polynomial of n.

Theorem 16 makes precise the sense in which DC-MASE aggregates information across all of the networks. In the bound there are two factors: one is the worst-case error for each network $\mathrm{err}_{\max}^{(i)}$, and one is the average-case error $\mathrm{err}_{\mathrm{ave}}^{(i)}$. In order to further consider the rate of improvement relative to L, we also consider the following application in the regime that the signal strengths are comparable.

Corollary 7 (Network Homogeneity)**.** *Instate the conditions of Theorem 16, and suppose*

that $\lambda_{\min}^{(l)} = \lambda_{\min}$ *and* $\theta_i^{(l)} = \theta_i$ *for all* l. *Then there exists a sufficiently small constant* c *depending on the implicit constants in the assumptions such that*

$$\mathbb{E}\ell(\widehat{z}, z) \leq \frac{2K}{n} \sum_{i=1}^{n} \exp\left(- cL\theta_i \min\left\{ \frac{\|\theta\|^4 \lambda_{\min}^3}{K^4 \|\theta\|_3^3}, \frac{\|\theta\|^2 \lambda_{\min}^{3/2}}{K^2 \theta_{\max}} \right\} \right) + O(n^{-10}).$$

When $L = 1$, Corollary 7 nearly matches the rate obtained in Theorem 1 of Jin et al. (2021) up to factors of K and $\lambda_{\min}$. However, Corollary 7 further elucidates the sense in which DC-MASE aggregates information from multiple networks: the error rate includes a gain of L but penalties of $\lambda_{\min}$ (relative to which term is the minimizer in the rate). In particular, if networks have extreme degree heterogeneity but well-separated communities, then the error rate for DC-MASE highly improves upon the corresponding rate for single networks.

To further ease interpretation, we consider the setting that $\theta_{\max}^{(l)} \leq C\theta_{\min}^{(l)}$ with with $\theta_{\max}^{(l)} \asymp \sqrt{\rho_n}$ for each l as in Example 6.2. In this setting the term $n\rho_n$ can be interpreted as the order of the average expected degree of each vertex, and hence larger ρ_n corresponds to denser networks. We then have the following corollary.

Corollary 8 (Homogeneous Degrees)**.** *Suppose that the conditions of Theorem 16 hold and that* $\theta_{\max}^{(l)} \leq C\theta_{\min}^{(l)}$ *for all* l, *and suppose that* $\theta_{\max}^{(l)} \asymp \sqrt{\rho_n}$ *for some* ρ_n. *Suppose also that* $\lambda_{\min}^{(l)} \asymp \lambda_{\min}$ *for all* l. *Then*

$$\mathbb{E}\ell(\widehat{z}, z) \leq 2K \exp\left(- cL \frac{n\rho_n \lambda_{\min}^3}{K^4} \right) + O(n^{-10}).$$

In the regime considered in Corollary 7, Jin et al. (2021) demonstrated that the SCORE clustering procedure with $L = 1$ yields the error rate of order $\exp(-c\lambda_{\min}^2 n\rho_n) + o(n^{-3})$, where we have ignored factors of K. In contrast, when L is large, we see that Corollary 8 demonstrates an error rate of order $\exp(-cL\lambda_{\min}^3 n\rho_n) + O(n^{-10})$. Therefore, we see that in this regime DC-MASE benefits whenever $\lambda_{\min} \gg \frac{1}{L}$, *even if each network is very sparse.*

Our final technical result shows that under sufficient global signal strength DC-MASE yields perfect clustering with high probability. Define the following signal-to-noise ratio

parameter vector

$$\mathrm{SNR}_l := \left(\frac{\theta^{(l)}_{\min}}{\theta^{(l)}_{\max}}\right)^{1/2} (\lambda^{(l)}_{\min})^{1/2} \|\theta^{(l)}\|. \tag{6.6}$$

When $\theta_{\max} \asymp \theta_{\min} \asymp \sqrt{\rho_n}$, it holds that $\mathrm{SNR}_l \asymp \sqrt{\lambda^{(l)}_{\min} n \rho_n}$. Our result will be stated in terms of a condition on SNR_l.

Theorem 17 (Perfect Clustering)**.** *Suppose that the conditions of Theorem 16 hold, and that* $\min_l \mathrm{SNR}_l^2 \geq C \frac{K^8 \log(n)}{L\bar{\lambda}^2}$*,where* C *is some sufficiently large constant. Then running* K*-means on the output of DC-MASE yields perfect recovery with probability at least* $1 - O(n^{-9})$*.*

Theorem 17 demonstrates that if the layer-wise SNR is sufficiently strong relative to $\bar{\lambda}$, we achieve perfect clustering. We note that Assumption 6.2 already implies that $\mathrm{SNR}_l^2 \gtrsim \left(\frac{\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|^2}\right) \frac{K^8 \log(n)}{\bar{\lambda}\lambda^{(l)}_{\min}}$as well as imposing a lower bound on $\bar{\lambda}$. If $(L\bar{\lambda})^{-1} \lesssim \frac{\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|^2\lambda^{(l)}_{\min}}$ for all l, then this condition is already met. Therefore, since the term $\frac{\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|^2\lambda^{(l)}_{\min}}$ is always larger than one (by assumption), the condition in Theorem 17 is only more stringent whenever $\bar{\lambda} \ll \frac{1}{L}$, which can only happen in the moderate L regime, since Assumption 6.2 already imposes a lower bound on $\bar{\lambda}$. At an intuitive level, this condition further reflects the idea that the second SVD step may not perform as well when $\bar{\lambda}$ is small.

6.3.2 Spherical Clustering for Single Networks

In the previous subsection we have compared our results to the best-known expected misclustering error for spectral clustering without refinement for degree-corrected stochastic blockmodels; i.e., the result in Jin et al. (2021). However, DC-MASE uses the spherical normalization, and the result in Jin et al. (2021) uses the SCORE normalization. While Jin et al. (2021) demonstrate that the SCORE procedure exhibits an exponential misclustering rate, to the best of our knowledge there is no similarly strong error rate for vanilla spectral clustering with the spherical normalization, though there are polynomial upper bounds (Lei and Rinaldo, 2015; Qin and Rohe, 2013), as well as some perfect clustering results (Lyzinski et al., 2014; Su et al., 2020). Conveniently, as a byproduct of our analysis we characterize the rows of $\widehat{\mathbf{Y}}^{(l)}$, and we are able to apply the same proof strategy for Theorem 16 to analyze

the result of running K-means on these rows.

The following theorem demonstrates an exponential error rate for single network clustering. For simplicity, we suppress the dependence of the parameters on the index l.

Theorem 18 (Single Network Misclustering Rate: Spherical Normalization)**.** *Assume that Assumption 6.1 and Assumption 6.2 hold (with $\bar{\lambda} = \lambda_{\min}$). Then the output of $(1+\varepsilon)$ K-means on the rows of $\widehat{\mathbf{Y}}$ satisfies*

$$\mathbb{E}\ell(\widehat{z}, z) \leq \frac{2K}{n}\sum_{i=1}^{n}\exp\left(-c\theta_i \min\left\{\frac{\|\theta\|^4\lambda_{\min}^2}{K^3\|\theta\|_3^3}, \frac{\|\theta\|^2\lambda_{\min}}{K^{3/2}\theta_{\max}}\right\}\right) + O(n^{-10}).$$

This rate exactly matches the rate obtained in Jin et al. (2021). However, the technical condition requires a stronger condition on $\lambda_{\min}$, but a weaker condition on the degree heterogeneity (relative to $\theta_{\max}/\theta_{\min}$). The main assumption behind Theorem 2.1 of Jin et al. (2021) requires that

$$\frac{K^8\theta_{\max}\|\theta\|_1\log(n)}{\|\theta\|^4\lambda_{\min}^2}\left(\frac{\theta_{\max}}{\theta_{\min}}\right)^2 \lesssim 1.$$

In contrast, we require that

$$\frac{K^8\theta_{\max}\|\theta\|_1\log(n)}{\|\theta\|^4\lambda_{\min}^3}\frac{\theta_{\max}}{\theta_{\min}} \lesssim 1.$$

Therefore, our condition is weaker whenever $\frac{1}{\lambda_{\min}} \lesssim \frac{\theta_{\max}}{\theta_{\min}}$. This regime corresponds to high degree heterogeneity relative to the community separation. For example, if the network is sparse (e.g. $\|\theta\| \asymp \sqrt{\log(n)}$), then it must be that $\lambda_{\min} \asymp 1$ (or else the assumption fails). It is not clear if this assumption (either ours or that of Jin et al. (2021)) is necessary or an artifact of our proof technique, and it would be interesting to try to weaken them.

6.4 Simulation Results

We evaluate the performance of different methods for community detection in networks generated from the multilayer DCSBM. The experiments focus on the effect of the number of graphs L for recovering the communities under different parameter setups. An im-

plementation of the code is available at https://github.com/jesusdaniel/dcmase. The performance measure reported in the experiments is the *adjusted rand index* (ARI) (Hubert and Arabie, 1985), which is a number that indicates the similarity between the estimated community labels and the true ones, and it is equal to 1 if the two partitions are the same (up to label permutations).

The benchmarks considered include spectral-based, optimization-based and likelihood-based clustering algorithms for multilayer networks. For spectral methods, the list comprises clustering on the embeddings defined as (i) the leading eigenvectors of the aggregated sum of the adjacency matrices $\sum_l \mathbf{A}^{(l)}$ (Han et al., 2015; Bhattacharyya and Chatterjee, 2020), (ii) the leading eigenvectors of the bias-adjusted sum-of-squared (SoS) adjacency matrices of Lei and Lin (2022), and (iii) an estimate of the common invariant subspace of the adjacency matrices obtained via multiple adjacency spectral embedding (MASE) from Arroyo et al. (2021). Existing methods and theoretical results for multilayer community detection with the aforementioned embedding procedures typically consider K-means clustering on the rows of these embeddings to obtain communities, but this clustering scheme is not expected to work well under high degree heterogeneity even for a single network. Thus, to isolate the performance of the embedding from the clustering method adopted, we employed spherical spectral clustering by normalizing the rows of the embeddings before performing K-means clustering (Lei and Rinaldo, 2015; Bhattacharyya and Chatterjee, 2020), as we observed better empirical performance compared to the unnormalized version. We also consider the orthogonal linked matrix factorization (OLMF) of Paul and Chen (2020), and an optimized Monte Carlo Markov Chain approach (Peixoto, 2014a, 2015) implemented via the graph-tool package (Peixoto, 2014b).

All the simulated graphs are generated using the multilayer DCSBM with $n = 150$ vertices and $K = 3$ equal sized communities, for which we assume that the membership matrix $\mathbf{Z}$ is such that vertices in the same community have adjacent rows. We focus on studying the effect of number of graphs L in the presence of different types of parameter heterogeneity. For that goal, we consider scenarios in which the block connectivity matrices $\{\mathbf{B}^{(l)}\}$ or the degree correction parameters $\{\mathbf{\Theta}^{(l)}\}$ are the same or different across the collection of graphs. For the block connectivity matrices, we generate these parameters as follows:

- *Same connectivity matrices:* the matrices $\mathbf{B}^{(l)}, l \in [L]$ are all set to be equal and defined as $\mathbf{B}^{(l)}_{rr} = 1, r \in [K]$, and $\mathbf{B}^{(l)}_{rs} = 0.4$, $r \neq s$.
- *Different connectivity matrices:* each matrix $\mathbf{B}^{(l)}, l \in [L]$ is generated independently with its entries equal to $\mathbf{B}^{(l)}_{rr} = p^{(l)} \sim \text{Unif}(0,1)$, for $r \in [K]$, and $\mathbf{B}^{(l)}_{rs} = q^{(l)} \sim \text{Unif}(0,1)$ for $r \neq s$.

In terms of the degree correction parameters, we consider scenarios as follows:

- *Same degree corrections:* the diagonal entries of the matrices satisfy $\mathbf{\Theta}^{(l)}_{ii} = \theta_i$ and are generated from a shifted exponential distribution such that $\theta_1, \dots, \theta_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)+0.2$.
- *Different degree corrections:* the parameters are generated in a similar fashion to the previous scenario, but now each matrix has its own parameters $\theta^{(l)}_1, \dots, \theta^{(l)}_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(1) + 0.2.$.
- *Alternating degrees:* the vertices within each community are split into two equal sized groups, and each group alternates between having low and high degrees on each network, that is, $\theta^{(l)}_i = 0.8$ if either l and i are odd or l and i are even numbers, and $\theta^{(l)}_i = 0.15$ otherwise.

The expected adjacency matrices are then defined as $\mathbf{P}^{(l)} = \alpha^{(l)}\mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}\mathbf{Z}\mathbf{\Theta}^{(l)}$ similar to Eq. (6.1), and the constant $\alpha^{(l)}$ is introduced to keep the average expected degree equal to 10. For each parameter setup, the experiments are repeated 100 times, and the average results are reported.

The results are shown in Figure 6.2. As expected, the accuracy of the methods generally improves with more graphs, and although there is no specific method that dominates in all the scenarios considered, we observe that DC-MASE is the only one that consistently improves its performance with L until perfect clustering is achieved. When the degree correction parameters are the same (left column), most of the methods perform accurately, especially in the setting with the same connectivity matrices. In particular, spectral methods perform well due to the fact that the singular subspace is shared in the expected adjacency matrices, and the population version of the matrix in which the embedding is performed captures the community structure after further correcting for degree heterogeneity via spherical

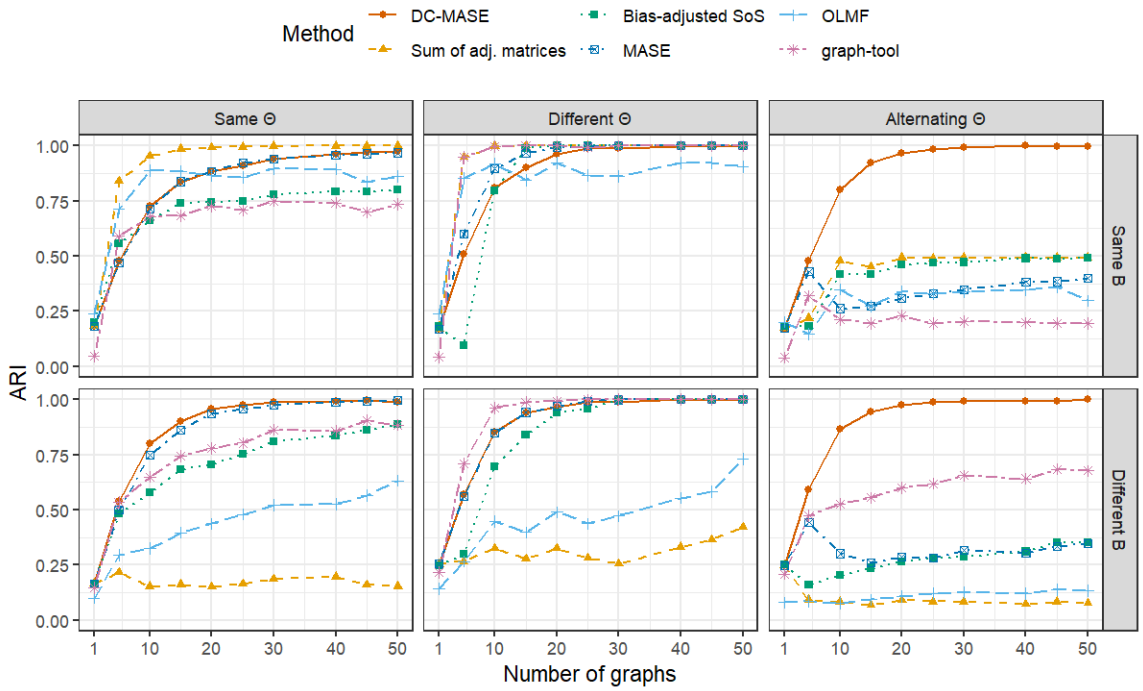


Figure 6.2: Community detection accuracy of different methods (measured via adjusted Rand index (ARI), averaged over 100 replications) as a function of the number of graphs. See Section 6.4 for a discussion of the setups.

normalization. In the scenario with different but random degree corrections (middle column) several methods are still able to perform accurately even when the population matrix does not have the correct clustering structure, possibly due to an averaging effect of the degree-correction parameters generated independently at random for each graph[1]. Aggregation methods, such as the sum of the adjacency matrices, perform very well when the global structure of the graphs is the same, but are not able to identify the correct structure in the presence of severe parameter heterogeneity. Notably, in the alternating degrees scenario (right column), DC-MASE is the only method that performs accurately, whereas other methods struggle to identify the model communities and instead cluster the vertices based on degree similarity.

[1]Since *all* of the degree-corrections are i.i.d., there is still a notion of an "average matrix" with the correct clustering structure. E.g., denoting $\mathbb{E}_\theta \mathbf{P}^{(l)}$ as the expectation of the matrix $\mathbf{P}^{(l)}$ by integrating with respect to the independence of the θ_i's, we see that $\mathbb{E}_\theta(\mathbf{P}^{(l)})_{ij} = c(\mathbf{Z}\mathbf{B}^{(l)}\mathbf{Z}^\top)_{ij}$ for $i \neq j$.

6.5 Analysis of US Airport Network

We evaluate the performance of the method in a time series of networks encoding the number of flights between airports in the United States within a given month for the period of January 2016 to September 2021. A multilayer degree-corrected SBM allows us to track the flight dynamics both at the airport and community levels to characterize the effect of the Covid 19 pandemic in flight connectivity. The data are publicly available and were downloaded from the US Bureau of Transportation Statistics (Bureau of Transportation Statistics, 2022).

The vertices of the networks correspond to some of the airports located within the 48 contiguous states in the US. For each network, the weighted edges contain the total number of flights of class F (scheduled passenger/cargo service) between each pair of airports within a given month. We restricted the analysis to the vertices in the intersection of the largest connected components of all the networks, resulting in a total of $n = 343$ airports. The period of the study contains 69 months (number of graphs).

To identify communities of airports with similar connectivity patterns in the data, we apply DC-MASE to the collection of adjacency matrices. The number of communities was selected to be $K = 4$ to facilitate interpretation and based on the scree plots of the individual network embeddings and the concatenated matrix, as described in Section 6.2.2. Figure 6.3(a) shows the estimated community memberships of the airports. Three of the communities identified (communities 2, 3 and 4) appear to be related to the geographical area, (west, east and southwest, respectively), whereas community 1 contains most of the hub airports in the east side of the country, as well as other smaller airports that are mostly connected to these hubs.

To characterize the dynamics in community and airport connectivity, we estimate the block connectivity matrices and degree correction parameters of the multilayer DCSBM. As the edges count the total number of flights between pairs of locations, the adjacency matrices are weighted, and thus, the parameters of the model describe the expected adjacency matrix $\mathbb{E}[\mathbf{A}^{(l)}] = \boldsymbol{\Theta}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}\mathbf{Z}^{\top}\boldsymbol{\Theta}^{(l)}$. For ease of interpretation, we adopt a similar identifiability condition as in Karrer and Newman (2011) by constraining the sum of the degree correction

parameters within each community to be equal to the size of the community, that is, if vertex i is in community r then

$$\sum_{i \in \mathcal{C}(r)} \theta_i^{(l)} = |\mathcal{C}(r)|, \qquad \text{for } r \in [K], l \in [L].$$

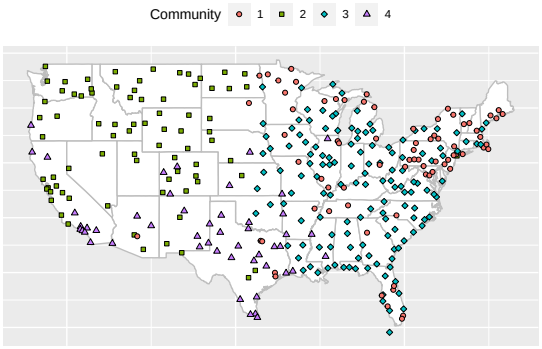
With this parameterization, we have the following relations. Let $d_i^{(l)} = \sum_{j=1}^{n} \mathbb{E}[\mathbf{A}_{ij}^{(l)}]$ be the expected degree of node i in network l. Then, for every $i \in [n]$, $r, s \in [K]$ and $l \in [L]$ we have

$$\theta_i^{(l)} = \frac{d_i^{(l)}}{\frac{1}{|\mathcal{C}(r)|} \sum_{j \in \mathcal{C}(r)} d_j^{(l)}}, \qquad \mathbf{B}_{rs}^{(l)} = \frac{1}{|\mathcal{C}(r)|\ |\mathcal{C}(s)|} \sum_{i \in \mathcal{C}(r), j \in \mathcal{C}(s)} \mathbb{E}[\mathbf{A}_{ij}^{(l)}]. \tag{6.7}$$

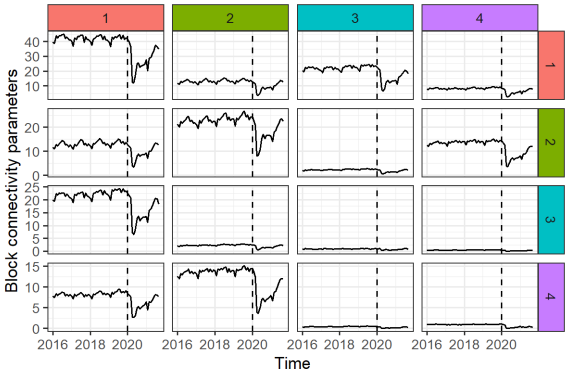
Under this parameterization, the degree correction parameters are on average equal to 1, and large values can be interpreted as higher individual connectivity of the corresponding vertex relative to other vertices in the community. Meanwhile, the block connectivity simply calculates the average number of edges within and between each pair of communities. When comparing the values of these parameters across time, this parameterization allows us to split global and local dynamics into the block connectivity matrices and degree corrections, respectively. We obtain plug-in estimates of the model parameters by using $\mathbf{A}^{(l)}$ rather than $\mathbb{E}[\mathbf{A}^{(l)}]$, and by using the estimated community memberships, which under certain edge distributions (e.g. Poisson) coincides with the maximum profile likelihood estimates given the fitted community memberships.

The multilayer DCSBM estimated parameters shown in Figure 6.3(b) track the changes in airport connectivity at the community level, which are mostly related to regional dynamics. By contrast, Figure 6.4(a) shows the individual airport popularity relative to airports within its community over time. While the overall number of flights within and between communities decreased after the pandemic started, the impact on the airport traffic was not homogeneous, and this is captured by the changes in degree correction parameters. Figure 6.4(b) explores these changes in more detail for community 1, which includes some of the largest hubs, such as ATL, DFW and ORD. These became relatively more prominent with respect to other airports in their community at the start of the pandemic in the US. Meanwhile, the airports in the New York City area (EWR and LGA) were relatively

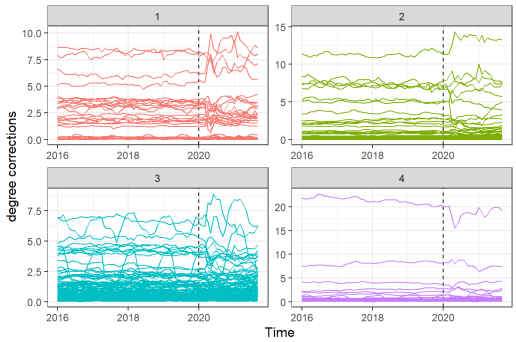
more negatively affected, possibly due to the pandemic dynamics and related closures. This analysis illustrates the flexibility of the multilayer DCSBM model for tracking local and community-level dynamics with changes over time.



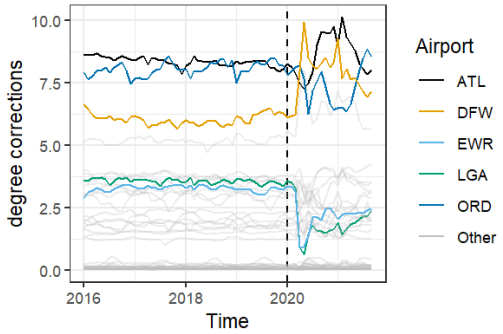
(a) Map of US airports colored according to the communities discovered by DC-MASE.



(b) Time series plot of the estimated block connectivity matrices in the multilayer DCSBM for the US airports data, with the communities discovered by DC-MASE. Each cell represents an entry of these matrices over time. The vertical line indicates January 1st, 2020.



(a) Degree correction parameter estimates in the US airport data. Each line corresponds to the parameter for some specific airport over time; the collection is divided according to the communities discovered by the algorithm. The variability in the parameter estimates suggests the need for a DCSBM, allowing different degree correction parameters at each time point.



(b) Degree correction parameter estimates for community 1, with some major airports highlighted. While some hub airports became relatively more prominent within the community after the pandemic started, the NYC airports (EWR and LGA) were relatively more negatively impacted.

We compared the performance of DC-MASE with the other spectral clustering algorithms considered in Section 6.4. In the absence of ground truth communities, we measure the performance in terms of out-of-sample mean squared error (MSE) for a given graph l and

some number of communities K, defined as

$$\text{MSE}(K, l) = \frac{1}{n^2} \|\mathbf{A}^{(l)} - \widehat{\mathbf{P}}^{(l)}_{\widehat{\mathbf{Z}}^{(-l,K)}}\|_F^2.$$

Here, $\widehat{\mathbf{Z}}^{(-l,K)}$ indicates the estimated community memberships obtained from a particular method fitted on the set of of graphs indexed by $[L] \setminus \{l\}$ with K communities. Given $\widehat{\mathbf{Z}}$, the value of the expected adjacency matrix is estimated as $\widehat{\mathbf{P}}^{(l)}_{\widehat{\mathbf{Z}}} = \widehat{\mathbf{\Theta}}^{(l)}_{\widehat{\mathbf{Z}}} \widehat{\mathbf{Z}} \widehat{\mathbf{B}}^{(l)}_{\widehat{\mathbf{Z}}} \widehat{\mathbf{Z}}^\top \widehat{\mathbf{\Theta}}^{(l)}_{\widehat{\mathbf{Z}}}$, where $\widehat{\mathbf{\Theta}}^{(l)}_{\widehat{\mathbf{Z}}}$ and $\widehat{\mathbf{B}}^{(l)}_{\widehat{\mathbf{Z}}}$ are the plug-in estimates defined via Eq. (6.7) using the communities defined by $\widehat{\mathbf{Z}}$. As the expected value of the average MSE is minimized by the expected adjacency matrices calculated with the correct communities, small values of this quantity are a proxy for the quality of the community estimates.

After calculating the MSE for all the graphs in the data and for different values of K, we performed a paired comparison via the MSE difference between the results for a given method and DC-MASE for each value of K and l. Figure 6.5 shows boxplots of these differences across all values of $l \in [L]$ and as a function of the number of communities. Notably, the MSE differences are positive for almost all graphs in the data and all values of K, indicating that the communities obtained by DC-MASE generally have smaller generalization error than the ones obtained by the other spectral methods considered.

6.6 Discussion

In this work we have considered the multilayer degree-corrected stochastic blockmodel, established its identifiability, and proposed a joint spectral clustering algorithm based on clustering the rows of a matrix that appropriately aggregates information about the communities in the model. The proposed method is simple and efficient, while the most expensive computations (required to estimate the leading eigenvalues and eigenvectors of each network) are able to be performed in parallel. This allows the methodology to scale to large datasets, both in terms of network size and in the number of graphs or layers. Our main results demonstrate that the method can effectively leverage the information across the graphs to obtain an improvement in community estimation, particularly when the number of networks

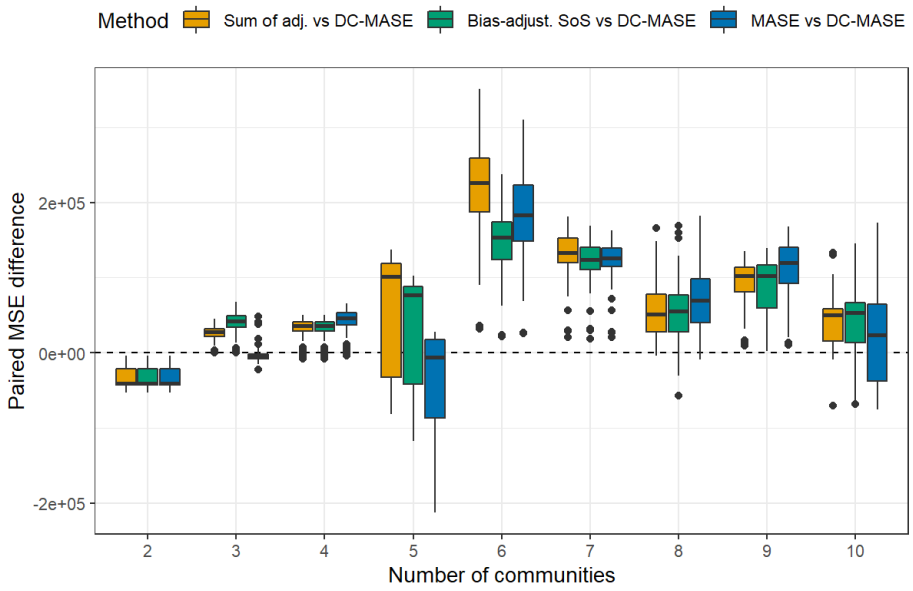


Figure 6.5: Paired out-of-sample mean squared error (MSE) difference for the Frobenius error of the estimated expected adjacency matrices obtained by each method and DC-MASE. Positive values indicate that the MSE of the respective method is larger than the MSE of DC-MASE.

L is large, even in the presence of significant vertex and layer heterogeneity. In our simulations, we observe that clustering with DC-MASE performs consistently well in various scenarios, and it is competitive with other state-of-the-art methods for multilayer community detection, particularly in situations with extreme degree heterogeneity. In our flight data studies, we see that the multilayer DCSBM is a flexible but succinct model, allowing us to identify clusters, track degree corrections, and observe block connectivity over time.

While the multilayer DCSBM is a flexible model, our main results require sufficient community separation for the proposed spectral method to succeed. This is partly due to the nature of the method, which aggregates the data after performing an eigendecomposition of each graph individually, and this requires enough signal on each graph to succeed. It has been argued that methods that perform aggregation in earlier stages can outperform late fusion techniques (Paul and Chen, 2020; Jing et al., 2021; Lei et al., 2020), and often perform better even under very sparse regimes (Lei and Lin, 2022). However, the model considered here requires one to appropriately handle the parameter heterogeneity before aggregation, and we observed in simulations that other spectral methods may fail to achieve

this goal. With respect to the theory, it may be possible to relax the community separation assumption using *singular gap-free perturbation bounds* as in, e.g., Löffler et al. (2021); Han et al. (2021); Zhang and Zhou (2022), but this is beyond the scope of this work as these other works rely on the assumption that the noise is Gaussian. Finally, the recent work Ke and Wang (2022) demonstrates that the eigenvectors of the *regularized Laplacian* can yield optimal mixed-membership estimation under extreme degree heterogeneity; it would be interesting to study the multilayer DCSBM in this regime as well as extend the methodology to mixed-membership models.

6.7 Proof Ingredients and Proof of Theorem 16

This section details the main ingredients required for the proof of Theorem 16. Our proof requires three key steps, each proved sequentially.

In what follows we let $\widehat{\mathbf{Y}}^{(l)}$ be defined in Algorithm 8, and we let $\mathbf{Y}^{(l)}$ denote the corresponding matrix associated to the population matrix $\mathbf{P}^{(l)}$. We also recall $\widehat{\mathcal{Y}} = [\widehat{\mathbf{Y}}^{(1)}, \ldots, \widehat{\mathbf{Y}}^{(L)}]$, and we let $\mathcal{Y} = [\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(L)}]$. Finally, we let $\mathbf{U}$ and Σ denote the leading K left singular vectors and singular values of $\mathcal{Y}$, and we let $\widehat{\mathbf{U}}$ and $\widehat{\Sigma}$ be defined similarly. For simplicity of notation, we assume that $\widehat{\mathbf{Z}}$ and $\widehat{z}$ satisfy

$$\|\widehat{\mathbf{Z}} - \mathbf{Z}\|_F = \min_{\mathbf{P}} \|\widehat{\mathbf{Z}} - \mathbf{Z}\mathbf{P}\|_F,$$

$$\sum_{i=1}^{n} \mathbb{I}\{\widehat{z}(i) \neq z(i)\} = \min_{\mathcal{P}} \sum_{i=1}^{n} \mathbb{I}\{\widehat{z}(i) \neq \mathcal{P}(z(i))\},$$

where the minimum is taking among all permutations $\mathcal{P}$ and permutation matrices $\mathbf{P}$.

Step 1: First Stage Asymptotic Expansion

In Theorem 19 we show that the initial estimates $\widehat{\mathbf{Y}}^{(l)}$ satisfy

$$\widehat{\mathbf{Y}}^{(l)}\mathbf{W}_*^{(l)} - \mathbf{Y}^{(l)} = \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)}) + \mathcal{R}^{(l)}_{\text{Stage I}},$$

where $\mathcal{L}(\cdot)$ is a linear function and $\mathcal{R}^{(l)}_{\text{Stage I}}$ is a residual with small $\ell_{2,\infty}$ error. Here $\mathbf{W}_*^{(l)}$ is

the $K \times K$ orthogonal matrix aligning the leading K eigenvectors of $\mathbf{A}^{(l)}$ and $\mathbf{P}^{(l)}$. Unlike previous results of this type (Du and Tang, 2022; Fan et al., 2022), our residual bounds depend explicitly on the degree corrections, which is what enables us to analyze the second stage of Algorithm 8 and obtain the exponential clustering rate in Theorem 16.

Step 2: Second Stage $\sin\Theta$ Perturbation Bounds

In the second step, we prove Theorem 20, applying Theorem 19 to obtain concentration in $\sin\boldsymbol{\Theta}$ distance for the empirical singular vectors $\widehat{\mathbf{U}}$ to the true singular vectors $\mathbf{U}$ that reveal the community memberships. In particular, by virtue of our first-order expansion, since $\mathcal{L}(\cdot)$ is linear in the noise, we are able to obtain stronger concentration for $\sin\boldsymbol{\Theta}$ distance than if one were to simply apply the naïve concentration using the triangle inequality, which would not yield improvement with L.

Step 3: Second Stage Asymptotic Expansion

In our final step, we prove Theorem 21, which combines these previous results as well as several additional concentration bounds to establish an asymptotic expansion for $\widehat{\mathbf{U}}\mathbf{W}_* - \mathbf{U}$ of the form

$$\widehat{\mathbf{U}}\mathbf{W}_* - \mathbf{U} = \sum_l \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2} + \mathcal{R}_{\text{Stage II}},$$

where $\mathcal{R}_{\text{Stage II}}$ can be understood as an "overall residual term" containing all of the residuals from all the networks, as well as the second stage of the algorithm, and $\mathbf{W}_*$ is the orthogonal matrix most closely aligning $\widehat{\mathbf{U}}$ and $\mathbf{U}$. Here $\mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})$ is the same linear operator as in the first step (Theorem 19), and $\mathbf{W}_*$ is an orthogonal matrix aligning $\widehat{\mathbf{U}}$ and $\mathbf{U}$. Using this final result, we then have all the ingredients to prove Theorem 16.

In the following subsections we formally state these results and then prove Theorem 16. After characterizing each result we discuss how it relates to previous literature.

6.7.1 First Stage Characterization

In the first step of the proof, we derive the following asymptotic expansion result for the individual networks. Recall that $\widehat{\mathbf{X}}^{(l)}$ and $\widetilde{\mathbf{X}}^{(l)}$ denote the scaled eigenvectors of $\mathbf{A}^{(l)}$ and $\mathbb{E}\mathbf{A}^{(l)} = \mathbf{P}^{(l)}$, respectively, and we let $\widehat{\mathbf{U}}^{(l)}$ and $\mathbf{U}^{(l)}$ be the leading K eigenvectors of $\mathbf{A}^{(l)}$ and $\mathbf{P}^{(l)}$ respectively. We let $\mathbf{I}_{p,q}^{(l)}$ denote the diagonal matrix with elements ± 1, where 1 appears p times and -1 appears q times, with p corresponding to the number of positive eigenvalues of $\mathbf{P}^{(l)}$ and q corresponding to the number of negative eigenvalues of $\mathbf{P}^{(l)}$. Equivalently, p and q count the number of positive and negative eigenvalues of $\mathbf{B}^{(l)}$. We let $\Lambda^{(l)}$ denote the nonzero eigenvalues of $\mathbf{P}^{(l)}$, and $\widehat{\Lambda}^{(l)}$ denote the leading p positive and q negative eigenvalues of $\mathbf{A}^{(l)}$, arranged in decreasing order by magnitude after splitting according to positive and negative.

The following result characterizes the rows of $\widehat{\mathbf{Y}}^{(l)}$.

Theorem 19 (Asymptotic Expansion: Stage I)**.** *Suppose that Assumption 6.1 and Assumption 6.2 hold. Fix a given $l \in [L]$. Let $\mathbf{W}_*^{(l)}$ denote the orthogonal matrix satisfing*

$$\mathbf{W}_*^{(l)} := \underset{\mathbf{W} \in \mathbb{O}(K)}{\arg\min} \|\widehat{\mathbf{U}}^{(l)} - \mathbf{U}^{(l)}\mathbf{W}_*^{(l)}\|_F.$$

Then there is an event $\mathcal{E}^{(l)}_{\text{Stage I}}$ with $\mathbb{P}(\mathcal{E}^{(l)}_{\text{Stage I}}) \geq 1 - O(n^{-15})$ such that the following expansion holds:

$$\widehat{\mathbf{Y}}^{(l)}(\mathbf{W}_*^{(l)})^\top - \mathbf{Y}^{(l)} = \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)}) + \mathcal{R}^{(l)}_{\text{Stage I}},$$

where the matrix $\mathcal{R}^{(l)}_{\text{Stage I}}$ satisfies

$$\|\mathcal{R}^{(l)}_{\text{Stage I}}\|_{2,\infty} \lesssim \frac{K^2\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\lambda^{(l)}_{\min}\|\theta^{(l)}\|^4}\left(\log(n)\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}} + \frac{1}{\lambda^{(l)}_{\min}} + \left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\frac{K^{5/2}\log(n)}{(\lambda^{(l)}_{\min})^{1/2}}\right),$$

and the matrix $\mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})$ has rows given by

$$\mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})_{i\cdot} = \frac{1}{\|\widetilde{\mathbf{X}}^{(l)}_{i\cdot}\|}\left(\mathbf{I} - \frac{\widetilde{\mathbf{X}}^{(l)}_{i\cdot}(\widetilde{\mathbf{X}}^{(l)}_{i\cdot})^\top}{\|\widetilde{\mathbf{X}}^{(l)}_{i\cdot}\|^2}\right)\left((\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}^{(l)}_{p,q}\right)_{i\cdot}.$$

Explicitly, Theorem 19 provides an entrywise expansion for the rows of $\widehat{\mathbf{Y}}^{(l)}$ about their corresponding population counterparts, up to the orthogonal transformation most closely aligning $\widehat{\mathbf{U}}^{(l)}$ and $\mathbf{U}^{(l)}$.

We remark briefly how Theorem 19 is related to and generalizes several previous results for single network analysis. In Du and Tang (2022), the authors consider the rows of $\widehat{\mathbf{Y}}$ to test if $\mathbf{Z}_{i\cdot} = \mathbf{Z}_{j\cdot}$ (under a mixed-membership model). To prove their main result, they establish a similar asymptotic expansion to Theorem 19. Our asymptotic linear term is the same as theirs, but our residual term exhibits a much finer characterization of the dependence on degree correction parameters, as they implicitly assume that $\theta_{\max} \asymp \theta_{\min}$, whereas we allow significant degree heterogeneity and extremely weak signals (Du and Tang (2022) also implicitly assume that $\lambda_{\min}^{(l)} \asymp 1$). Similarly, Fan et al. (2022) consider the asymptotic normality of rows of the SCORE-normalized eigenvectors for testing equality of membership in degree-corrected stochastic blockmodels. However, they also require that $\theta_{\max} \asymp \theta_{\min}$, which again eliminates the possibility of severe degree correction. Moreover, our results also allow K to grow and $\lambda_{\min}$ to shrink to zero sufficiently slowly, provided this is compensated for elsewhere in the signal strength, and previous results require much stronger conditions on these parameters. Finally, a similar asymptotic expansion (with explicit degree corrections and dependencies) was used implicitly to prove the main result in Jin et al. (2021), albeit for the SCORE normalization (as opposed to spherical normalization). Therefore, our results complement theirs by providing an analysis of the spherical normalization often used in practice. We will also apply Theorem 19 in the proof of Theorem 18.

The following result will be used as an intermediate bound in the proof of Theorem 18, demonstrating a concentration inequality for $\|\widehat{\mathbf{Y}}^{(l)} - \mathbf{Y}^{(l)}\mathbf{W}_*^{(l)}\|_{2,\infty}$.

Corollary 9. *With probability at least* $1 - O(n^{-15})$*, it holds that*

$$\|\widehat{\mathbf{Y}}^{(l)} - \mathbf{Y}^{(l)}\mathbf{W}_*^{(l)}\|_{2,\infty} \lesssim \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\right)^{1/2} \frac{K\sqrt{\log(n)}}{\|\theta^{(l)}\|(\lambda_{\min}^{(l)})^{1/2}} + \frac{K^2\theta_{\max}^{(l)}\|\theta^{(l)}\|_1}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^4}\left(\log(n)\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} + \frac{1}{\lambda_{\min}^{(l)}} + \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\right)^{1/2}\frac{K^{5/2}\log(n)}{(\lambda_{\min}^{(l)})^{1/2}}\right).$$

The proof follows from Lemma 58 (see Appendix F.2) and Theorem 19.

6.7.2 Second Stage Characterization I: $\sin\boldsymbol{\Theta}$ Bound

With the strong upper bounds for the first stage in Theorem 19, we can apply this result to establish $\sin\boldsymbol{\Theta}$ perturbation for the output of DC-MASE. In what follows we denote SNR^{-1} as the entrywise inverse of the vector SNR defined in Equation 6.6.

Theorem 20 ($\sin\boldsymbol{\Theta}$ Perturbation Bound)**.** *Suppose the conditions in Theorem 16 hold. Define*

$$\alpha_{\max} = \frac{K^2\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\lambda^{(l)}_{\min}\|\theta^{(l)}\|^4}\left(\log(n)\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}} + \frac{1}{\lambda^{(l)}_{\min}} + \left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\frac{K^{5/2}\log(n)}{(\lambda^{(l)}_{\min})^{1/2}}\right);$$

i.e., $\alpha_{\max}$ is the residual upper bound from Theorem 19. Then with probability at least $1 - O(n^{-10})$, it holds that

$$\begin{aligned}\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\| \lesssim K^2\sqrt{\log(n)}\frac{\left(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\right)^{1/2}}{\sqrt{L\bar{\lambda}}} + K^3\log(n)\frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} \\ + K^2\sqrt{\log(n)}\frac{\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K\alpha_{\max}}{\bar{\lambda}}.\end{aligned}$$

In particular, under the conditions of Theorem 16 it holds that

$$\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\| \lesssim \frac{1}{K}.$$

We note that the first bound provided in Theorem 20 may actually be much stronger than the upper bound of $\frac{1}{K}$, which is all that is needed for the proof of Theorem 16. First, by combining Assumption 6.2 and the definition of SNR_l in Equation 6.6 it is straightforward to check that each term is smaller than one, since we require that

$$C\frac{K^8\theta^{(l)}_{\max}\|\theta^{(l)}\|_1\log(n)}{\|\theta^{(l)}\|^2\mathrm{SNR}_l^2} \leq \bar{\lambda},$$

for some large constant C. Since $\frac{\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|^2}$ is always larger than one, we see that Assumption 6.2 is a stronger assumption than each term in Theorem 20 being smaller than one.

For ease of interpretation, when all l have $\lambda^{(l)}_{\min} \asymp 1$, and $\theta^{(l)}_{\max} \asymp \theta^{(l)}_{\min} \asymp \sqrt{\rho_n}$ and $K \asymp 1$,

we have that

$$\begin{aligned}\|\mathrm{SNR}^{-1}\|_\infty &\lesssim \frac{1}{\sqrt{n\rho_n}};\\ \alpha_{\max} &\lesssim \frac{\log(n)}{n\rho_n};\\ \frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2 &\lesssim \frac{1}{n\rho_n}.\end{aligned}$$

Therefore, the $\sin\boldsymbol{\Theta}$ bound simplifies to

$$\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\| \lesssim \frac{\sqrt{\log(n)}}{\sqrt{Ln\rho_n}} + \frac{\log(n)}{n\rho_n}.$$

This final bound shows that $\widehat{\mathbf{U}}$ concentrates in $\sin\boldsymbol{\Theta}$ distance about $\mathbf{U}$ as n increases by a factor that improves with $\sqrt{L}$ when $L \lesssim n\rho_n/\log(n)$. For a single stochastic blockmodel without degree corrections, the $\sin\boldsymbol{\Theta}$ distance between $\widehat{\mathbf{U}}$ and $\mathbf{U}$ can be upper bounded as $\sqrt{\frac{\log(n)}{n\rho_n}}$ (Lei and Rinaldo, 2015). Therefore, Theorem 20, which utilizes the information from all the networks and allows degree heterogeneity, already demonstrates improvement from multiple networks by a factor of $\max\{\frac{1}{\sqrt{L}}, \frac{\sqrt{\log(n)}}{\sqrt{n\rho_n}}\}$ relative to the single-network setting.

6.7.3 Second Stage Characterization II: Asymptotic Expansion

In essence, we require Theorem 20 to demonstrate that the clusters are correctly identified (see the proof of Theorem 16 in Section 6.7.4), but it falls short of providing a fine-grained characterization for the rows of $\widehat{\mathbf{U}}$, which is what is needed for the exponential error rate.

The following result demonstrates a first-order asymptotic expansion for the singular vectors in the second stage of our algorithm. The proof is given in the Appendix.

Theorem 21 (Asymptotic Expansion: Stage II)**.** *Suppose the conditions of Theorem 16 hold. Define*

$$\mathbf{W}_* := \operatorname*{arg\,min}_{\mathbf{W}\in\mathbb{O}(K)} \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}\|_F.$$

There is an event $\mathcal{E}_{\text{Stage II}}$ *satisfying* $\mathbb{P}(\mathcal{E}_{\text{Stage II}}) \geq 1 - O(n^{-10})$ *such that on this event, we*

have the asymptotic expansion

$$\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U} = \sum_l \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2} + \mathcal{R}_{\text{Stage II}},$$

where $\mathcal{L}(\cdot)$ *is the operator from Theorem 19 and the residual satisfies*

$$\|\mathcal{R}_{\text{Stage II}}\|_{2,\infty} \lesssim \frac{K^3\sqrt{\log(n)}}{nL\bar{\lambda}}\|\text{SNR}^{-1}\|_2 + \frac{K^4\log(n)}{L^2\sqrt{n}\bar{\lambda}^2}\|\text{SNR}^{-1}\|_2^2 + \frac{K^{7/2}\log(n)}{\sqrt{n}\bar{\lambda}}\|\text{SNR}^{-1}\|_\infty^2 + \frac{\alpha_{\max}}{\sqrt{n}\bar{\lambda}}.$$

Here $\alpha_{\max}$ *is as Theorem 20. In particular, under the assumptions of Theorem 16, it holds that*

$$\|\mathcal{R}_{\text{Stage II}}\|_{2,\infty} \leq \frac{1}{16\sqrt{n_{\max}}}.$$

Theorem 21 establishes a first-order expansion for the rows of the difference matrix $\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U}$, which is the main technical tool required to establish Theorem 16. The proof of Theorem 21 relies on both Theorem 19 and Theorem 20, but requires a number of additional considerations to bound the residual term $\mathcal{R}_{\text{Stage II}}$ in $\ell_{2,\infty}$ norm.

6.7.4 Proof of Theorem 16 and Theorem 17

With all of these ingredients in place, we are nearly prepared to prove Theorem 16. In the proof we will also require several results concerning the population parameters, which we state in the following two lemmas. The proofs can be found in Appendix F.1.

Lemma 7 (Population Properties: Stage I)**.** *Suppose Assumption 6.1 holds, and let* $\lambda_r^{(l)}$ *denote the eigenvalues of* $\mathbf{P}^{(l)}$ *and let* $\lambda_r(\mathbf{B}^{(l)})$ *denote the eigenvalues of* $\mathbf{B}^{(l)}$*. Then for all*

$1 \leq r \leq K$,

$$\theta_i^{(l)} \lesssim \|\widetilde{\mathbf{X}}_{i\cdot}^{(l)}\| \lesssim \theta_i^{(l)};$$
$$\|\mathbf{U}_{i\cdot}^{(l)}\| \lesssim \sqrt{K}\frac{\theta_i^{(l)}}{\|\theta^{(l)}\|};$$
$$\lambda_r^{(l)} \asymp \frac{\|\theta^{(l)}\|^2}{K}\lambda_r(\mathbf{B}^{(l)}).$$

Next, the following result establishes the population properties of the the second stage; in particular demonstrating a lower bound on the smallest eigenvalue of the population matrix $\mathcal{Y}\mathcal{Y}^\top$ in terms of $\bar{\lambda}$.

Lemma 8 (Population Properties: Stage II)**.** *Suppose that $\mathcal{Y}$ is rank K, and let $\mathcal{Y} = \mathbf{U}\Sigma\mathbf{V}^\top$ be its (rank K) singular value decomposition. Then it holds that*

$$\mathbf{U} = \mathbf{Z}\mathbf{M},$$

where $\mathbf{M} \in \mathbb{R}^{K\times K}$ is some invertible matrix satisfying

$$\|\mathbf{M}_{r\cdot} - \mathbf{M}_{s\cdot}\| = \sqrt{n_r^{-1} + n_s^{-1}}.$$

In addition, when $n_{\min} \asymp n_{\max}$, it holds that

$$\lambda_Y^2 := \lambda_{\min}\left(\sum_l \mathbf{Y}^{(l)}(\mathbf{Y}^{(l)})^\top\right) \gtrsim \frac{n}{K}L\bar{\lambda}.$$

Armed with these lemmas as well as Theorems 19, 20, and 21, we are prepared to prove Theorem 16.

Proof of Theorem 16. We follow the analysis technique developed in Jin et al. (2021) to derive an exponential rate for the output of $(1+\varepsilon)$ K-means. First will use the the $\sin\boldsymbol{\Theta}$ bound (Theorem 20) together with Lemma 5.3 of Lei and Rinaldo (2015) to demonstrate a Hamming error of order strictly less than $\frac{n_{\min}}{4}$, so that each cluster has at a majority of its true members. This allows us to associate each empirical cluster centroid to a true cluster centroid. Next, we will study the empirical centroids of these clusters to show that they

are strictly closer to their corresponding true cluster centroid than they are to each other. Finally, we decompose the expected error into individual node-wise errors, where we apply the asymptotic expansion in Theorem 21 to obtain the exponential error rate.

In what follows, let $\mathcal{E}_{\sin\boldsymbol{\Theta}}$ denote the event

$$\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\| \leq \frac{\beta}{8K\sqrt{C_\varepsilon}},$$

where $\beta \in (0,1)$ is such that $n_{\min} \geq \beta n_{\max}$ and C_ε is a constant to be defined in the subsequent analysis. We note that by Theorem 20 the event $\mathcal{E}_{\sin\boldsymbol{\Theta}}$ holds with probability at least $1 - O(n^{-10})$. We also let $(\widehat{\mathbf{Z}}, \widehat{\mathbf{M}})$ denote the output of $(1+\varepsilon)$ K-means on the rows of $\widehat{\mathbf{U}}$, where $\widehat{\mathbf{Z}} \in \{0,1\}^{n\times K}$ and $\widehat{\mathbf{M}} \in \mathbb{R}^{K\times K}$.

Step 1: Initial Hamming Error

First by Lemma 8 it holds that $\mathbf{U} = \mathbf{Z}\mathbf{M}$ where $\mathbf{M}$ has K unique rows satisfying

$$\frac{1}{\sqrt{n_{\max}}} \leq \|\mathbf{M}_{r\cdot} - \mathbf{M}_{s\cdot}\| \leq \frac{\sqrt{2}}{\sqrt{n_{\min}}}.$$

Define the matrix $\widehat{\mathbf{V}} := \widehat{\mathbf{Z}}\widehat{\mathbf{M}}$. Define $S_r := \{i \in \mathcal{C}(r) : \|\mathbf{W}_*\widehat{\mathbf{V}}_{i\cdot} - \mathbf{U}_{i\cdot}\| \geq \delta_r/2\}$, where $\delta_r = \frac{1}{\sqrt{n_r}}$. By Lemma 5.3 of Lei and Rinaldo (2015), it holds that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{\widehat{z}(i) \neq z(i)\} \leq \frac{1}{n}\sum_{r=1}^{K}|S_r| \leq \sum_{r=1}^{K}\frac{|S_r|}{n_r} = \sum_{r=1}^{K}|S_r|\delta_r^2 \\ \leq C_\varepsilon\|\widehat{\mathbf{U}}\mathbf{W}_* - \mathbf{U}\|_F^2 \leq C_\varepsilon\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\|_F^2 \leq C_\varepsilon K\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\|^2.\end{aligned}$$

Therefore, on the event $\mathcal{E}_{\sin\boldsymbol{\Theta}}$, it holds that

$$\sum_{i=1}^{n}\mathbb{I}\{\widehat{z}(i) \neq z(i)\} \leq C_\varepsilon K n\frac{\beta^2}{64K^2C_\varepsilon} \leq \frac{\beta}{64}n_{\min},$$

since $n \leq Kn_{\max} \leq \frac{K}{\beta}n_{\min}$. Therefore, since this error is strictly less than $\beta n_{\min}/64 \leq n_{\min}n_r/(64n_{\max})$, each cluster r has at least $n_r - \beta n_{\min}/64 \geq (1 - n_{\min}/(64n_{\max}))n_r \geq (63/64)n_r$ of its true members. This implies that we can associate each empirical cluster to

a true cluster – let these empirical clusters be denoted $\widehat{\mathcal{C}}(r)$. Observe that we must have that $|\widehat{\mathcal{C}}(r)| \geq (1-\beta/64)n_{\min}$ and that $|\widehat{\mathcal{C}}(r) \setminus \mathcal{C}(r)| \leq \beta n_{\min}/64$.

Step 2: Properties of Empirical Centroids

Recall that the cluster centroid associated to $\widehat{\mathcal{C}}(r)$ is equal to $\widehat{\mathbf{M}}_{r\cdot}$. Then by definition,

$$\widehat{\mathbf{M}}_{r\cdot} = \frac{1}{|\widehat{\mathcal{C}}(r)|}\sum_{i\in\widehat{\mathcal{C}}(r)} \widehat{\mathbf{U}}_{i\cdot}.$$

Recall that $\mathbf{U}$ consists of K unique rows of $\mathbf{M}$. Without loss of generality assume that $\mathbf{M}_{r\cdot}$ is associated to $\mathcal{C}(r)$. Then

$$\begin{aligned}
\|\mathbf{W}_*\widehat{\mathbf{M}}_{r\cdot} - \mathbf{M}_{r\cdot}\| &= \frac{1}{|\widehat{\mathcal{C}}(r)|}\Big\|\sum_{i\in\widehat{\mathcal{C}}(r)}(\mathbf{W}_*\widehat{\mathbf{U}}_{i\cdot} - \mathbf{M}_{r\cdot})\Big\| \\
&\leq \frac{1}{|\widehat{\mathcal{C}}(r)|}\Big\|\sum_{i\in\widehat{\mathcal{C}}(r)}(\mathbf{W}_*\widehat{\mathbf{U}}_{i\cdot} - \mathbf{U}_{i\cdot})\Big\| + \frac{1}{|\widehat{\mathcal{C}}(r)|}\Big\|\sum_{i\in\widehat{\mathcal{C}}(r)}(\mathbf{U}_{i\cdot} - \mathbf{M}_{r\cdot})\Big\| \\
&\leq \frac{1}{|\widehat{\mathcal{C}}(r)|}\left\|\sum_{i\in\widehat{\mathcal{C}}(r)}(\mathbf{W}_*\widehat{\mathbf{U}}_{i\cdot} - \mathbf{U}_{i\cdot})\right\| + \frac{1}{|\widehat{\mathcal{C}}(r)|}\left\|\sum_{i\in\widehat{\mathcal{C}}(r)\setminus\mathcal{C}(r)}(\mathbf{U}_{i\cdot} - \mathbf{M}_{r\cdot})\right\|.
\end{aligned}$$

We observe that for $i \notin \mathcal{C}(r)$, it holds that

$$\frac{1}{\sqrt{n_{\max}}} \leq \|\mathbf{U}_{i\cdot} - \mathbf{M}_{r\cdot}\| \leq \frac{\sqrt{2}}{\sqrt{n_{\min}}}$$

by Lemma 8. Therefore,

$$\begin{aligned}
\|\mathbf{W}_*\widehat{\mathbf{M}}_{r\cdot} - \mathbf{M}_{r\cdot}\| &\leq \frac{1}{|\widehat{\mathcal{C}}(r)|}\Big\|\sum_{i\in\widehat{\mathcal{C}}(r)}(\mathbf{W}_*\widehat{\mathbf{U}}_{i\cdot} - \mathbf{U}_{i\cdot})\Big\| + \frac{1}{|\widehat{\mathcal{C}}(r)|}\Big\|\sum_{i\in\widehat{\mathcal{C}}(r)\setminus\mathcal{C}(r)}(\mathbf{U}_{i\cdot} - \mathbf{M}_{r\cdot})\Big\| \\
&\leq \frac{1}{|\widehat{\mathcal{C}}(r)|^{1/2}}\|\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U}\|_F + \frac{|\widehat{\mathcal{C}}(r)\setminus\mathcal{C}(r)|}{|\widehat{\mathcal{C}}(r)|}\frac{\sqrt{2}}{\sqrt{n_{\min}}} \\
&\leq \frac{\sqrt{2K}}{\sqrt{n_{\min}(1-\beta/64)}}\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\| + \frac{\beta n_{\min}}{64(1-\beta/64)n_{\min}}\frac{\sqrt{2}}{\sqrt{n_{\min}}} \\
&\leq \frac{\sqrt{2K}}{\sqrt{\beta n_{\max}(1-\beta/64)}}\frac{\beta}{8KC_\varepsilon^{1/2}} + \frac{\beta}{64(1-\beta/64)}\frac{\sqrt{2}}{\sqrt{\beta n_{\max}}} \\
&\leq \frac{1}{\sqrt{n_{\max}}}\left(\frac{\sqrt{2\beta}}{8K^{1/2}C_\varepsilon^{1/2}\sqrt{1-\beta/64}} + \frac{\beta^{1/2}\sqrt{2}}{64(1-\beta/64)}\right) \\
&\leq \frac{1}{8\sqrt{n_{\max}}},
\end{aligned}$$

since $n_{\min} \geq \beta n_{\max}$, $K \geq 1$ and $\beta < 1$, as well as the assumption $C_\varepsilon \geq 4$. Therefore, on the event $\mathcal{E}_{\sin\boldsymbol{\Theta}}$ it holds that

$$\max_{1\leq r\leq K}\|\mathbf{W}_*\widehat{\mathbf{M}}_{r\cdot} - \mathbf{M}_{r\cdot}\| \leq \frac{1}{8\sqrt{n_{\max}}}.$$

Step 3: Applying The Asymptotic Expansion

In this section we will use the previous bound on the cluster centroids and Theorem 21 to obtain the desired bound. Recall that by Theorem 20, $\mathbb{P}(\mathcal{E}^c_{\sin\boldsymbol{\Theta}}) = O(n^{-10})$. It then holds that

$$\begin{aligned}
\mathbb{E}\ell(\widehat{z}, z) &= \frac{1}{n}\sum_{i=1}^n \mathbb{P}(\mathbf{Z}_{i\cdot} \neq \widehat{\mathbf{Z}}_{i\cdot}) \\
&\leq \frac{1}{n}\sum_{i=1}^n \mathbb{P}(\mathbf{Z}_{i\cdot} \neq \widehat{\mathbf{Z}}_{i\cdot}, \mathcal{E}_{\sin\boldsymbol{\Theta}}) + O(n^{-10}).
\end{aligned}$$

Suppose that $\|(\widehat{\mathbf{U}}\mathbf{W}_*^\top)_{i\cdot} - \mathbf{U}_{i\cdot}\| \leq \frac{1}{4\sqrt{n_{\max}}}$, and suppose the i'th node is in community r.

Then on the event $\mathcal{E}_{\sin\boldsymbol{\Theta}}$

$$\begin{aligned}\|\mathbf{W}_*\widehat{\mathbf{U}}_{i\cdot} - \mathbf{W}_*\widehat{\mathbf{M}}_{r\cdot}\| &\leq \|(\widehat{\mathbf{U}}\mathbf{W}_*^\top)_{i\cdot} - \mathbf{U}_{i\cdot}\| + \|\mathbf{U}_{i\cdot} - \widehat{\mathbf{M}}_{r\cdot}\| \\ &\leq \frac{1}{4\sqrt{n_{\max}}} + \max_s \|\mathbf{W}_*\widehat{\mathbf{M}}_{s\cdot} - \mathbf{M}_{s\cdot}\| \\ &\leq \frac{3}{8\sqrt{n_{\max}}}.\end{aligned}$$

In addition, for any $s \neq r$, we have that

$$\begin{aligned}\|\mathbf{W}_*\widehat{\mathbf{U}}_{i\cdot} - \mathbf{W}_*\widehat{\mathbf{M}}_{s\cdot}\| &\geq \|\mathbf{M}_{r\cdot} - \mathbf{M}_{s\cdot}\| - \|\mathbf{W}_*\widehat{\mathbf{U}}_{i\cdot} - \mathbf{U}_{i\cdot}\| - \|\mathbf{W}_*\widehat{\mathbf{M}}_{s\cdot} - \mathbf{M}_{s\cdot}\| \\ &\geq \frac{1}{\sqrt{n_{\max}}} - \frac{1}{4\sqrt{n_{\max}}} - \frac{1}{8\sqrt{n_{\max}}} \\ &\geq \frac{5}{8\sqrt{n_{\max}}}.\end{aligned}$$

Therefore, node i must belong to cluster $\widehat{\mathcal{C}}(r)$, so that there is no error on node i. Therefore,

$$\begin{aligned}\mathbb{P}(\mathbf{Z}_{i\cdot} \neq \widehat{\mathbf{Z}}_{i\cdot}, \mathcal{E}_{\sin\boldsymbol{\Theta}}) &\leq \mathbb{P}(\|(\widehat{\mathbf{U}}\mathbf{W}_*^\top)_{i\cdot} - \mathbf{U}_{i\cdot}\| \geq \frac{1}{4\sqrt{n_{\max}}}) \\ &\leq \mathbb{P}(\|(\widehat{\mathbf{U}}\mathbf{W}_*^\top)_{i\cdot} - \mathbf{U}_{i\cdot}\| \geq \frac{1}{4\sqrt{n_{\max}}}, \mathcal{E}_{\text{Stage II}}) + O(n^{-10}),\end{aligned}$$

where $\mathcal{E}_{\text{Stage II}}$ is the event in Theorem 21. On the event $\mathcal{E}_{\text{Stage II}}$ it holds that

$$\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U} = \sum_l \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2} + \mathcal{R}_{\text{Stage II}},$$

with

$$\|\mathcal{R}_{\text{Stage II}}\|_{2,\infty} \leq \frac{1}{16\sqrt{n_{\max}}}.$$

Therefore,

$$\begin{aligned}\mathbb{P}(\|(\widehat{\mathbf{U}}\mathbf{W}_*^\top)_{i\cdot} - \mathbf{U}_{i\cdot}\| \geq \frac{1}{4\sqrt{n_{\max}}}, \mathcal{E}_{\text{Stage II}}) &\leq \mathbb{P}\left(\left\|e_i^\top \sum_l \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}\right\| \geq \frac{1}{8\sqrt{n_{\max}}}\right) \\ &\leq \mathbb{P}\left(\left\|e_i^\top \sum_l \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}\right\| \geq C\frac{\sqrt{K}}{\sqrt{n}}\right) \\ &\leq K \max_k \mathbb{P}\left(\left|\sum_l e_i^\top \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\big(\mathbf{Y}^{(l)}\big)^\top \mathbf{U}\Sigma^{-2} e_k\right| \geq C\frac{1}{\sqrt{n}}\right).\end{aligned}$$

We will apply the Bernstein inequality now. We have that

$$\sum_l e_i^\top \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2} e_k = \sum_l \sum_j (\mathbf{A}^{(l)} - \mathbf{P}^{(l)})_{ij} \left(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}\mathbf{J}(\widetilde{\mathbf{X}}_{i\cdot})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}\right)_{jk}.$$

Using Lemma 7 and Lemma 8, the variance v of this quantity is upper bounded by

$$\begin{aligned}v &\leq \sum_l \sum_j \theta_i^{(l)}\theta_j^{(l)} \|e_j^\top \mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}\mathbf{J}(\widetilde{\mathbf{X}}_{i\cdot})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}\|^2 \\ &\leq \sum_l \sum_j \theta_i^{(l)}\theta_j^{(l)} \|e_j^\top \mathbf{U}^{(l)}\|^2 \||\Lambda^{(l)}|^{-1/2}\|^2 \|\mathbf{J}(\widetilde{\mathbf{X}}_{i\cdot})\|^2 \|(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}\|^2 \\ &\leq C\sum_l \sum_j \theta_i^{(l)}\theta_j^{(l)} \frac{K(\theta_j^{(l)})^2}{\|\theta^{(l)}\|^2} \frac{K}{\|\theta^{(l)}\|^2\lambda_{\min}^{(l)}} \frac{1}{\|\widetilde{\mathbf{X}}_{i\cdot}\|^2} \frac{nK^2}{n^2L^2\bar{\lambda}^2} \\ &\leq C\frac{K^4}{nL^2\bar{\lambda}^2} \sum_l \sum_j \theta_i^{(l)}\theta_j^{(l)} \frac{(\theta_j^{(l)})^2}{\|\theta^{(l)}\|^4\lambda_{\min}^{(l)}(\theta_i^{(l)})^2} \\ &\leq C\frac{K^4}{nL^2\bar{\lambda}^2} \sum_l \frac{\|\theta^{(l)}\|_3^3}{\theta_i^{(l)}\|\theta^{(l)}\|^4\lambda_{\min}^{(l)}}.\end{aligned}$$

In addition, each term satisfies

$$\begin{aligned}\max_{l,j} \|e_j^\top \mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}\mathbf{J}(\widetilde{\mathbf{X}}_{i\cdot})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}\| &\leq \max_{l,j} C\frac{\theta_j^{(l)}K}{\theta_i^{(l)}\|\theta^{(l)}\|^2(\lambda_{\min}^{(l)})^{1/2}} \frac{K}{\sqrt{n}L\bar{\lambda}} \\ &\leq C\frac{K^2}{\sqrt{n}L\bar{\lambda}} \max_l \frac{\theta_{\max}^{(l)}}{\theta_i^{(l)}\|\theta^{(l)}\|^2(\lambda_{\min}^{(l)})^{1/2}}.\end{aligned}$$

By Bernstein's inequality,

$$\begin{aligned}
&\mathbb{P}\bigg(\bigg|\sum_l e_i^\top \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}e_k\bigg| \geq C\frac{1}{\sqrt{n}}\bigg) \\
&\quad \leq 2\exp\bigg(-\frac{C^2\frac{1}{128n}}{C_1\frac{K^4}{nL^2\bar{\lambda}^2}\sum_l \frac{\|\theta^{(l)}\|_3^3}{\theta_i^{(l)}\|\theta^{(l)}\|^4\lambda_{\min}^{(l)}} + C_1\frac{1}{\sqrt{n}}\frac{K^2}{\sqrt{n}L\bar{\lambda}}\max_l \frac{\theta_{\max}^{(l)}}{\theta_i^{(l)}\|\theta^{(l)}\|^2(\lambda_{\min}^{(l)})^{1/2}}}\bigg) \\
&\quad \leq 2\exp\bigg(-\frac{C_1}{C_2\frac{K^4}{L^2\bar{\lambda}^2}\sum_l \frac{\|\theta^{(l)}\|_3^3}{\theta_i^{(l)}\|\theta^{(l)}\|^4\lambda_{\min}^{(l)}} + \frac{1}{24}C_2\frac{K^2}{L\bar{\lambda}}\max_l \frac{\theta_{\max}^{(l)}}{\theta_i^{(l)}\|\theta^{(l)}\|^2(\lambda_{\min}^{(l)})^{1/2}}}\bigg) \\
&\quad \leq 2\exp\bigg(-C_3\min\bigg\{\frac{\bar{\lambda}^2 L}{K^4}\bigg(\frac{1}{L}\sum_l \frac{\|\theta^{(l)}\|_3^3}{\theta_i^{(l)}\|\theta^{(l)}\|^4\lambda_{\min}^{(l)}}\bigg)^{-1}, \frac{L\bar{\lambda}}{K^2}\min_m \frac{\theta_i^{(l)}\|\theta^{(l)}\|^2(\lambda_{\min}^{(l)})^{1/2}}{\theta_{\max}^{(l)}}\bigg\}\bigg) \\
&\quad \leq 2\exp\bigg(-cL\min\bigg\{\frac{\bar{\lambda}^2}{K^4\mathrm{err}_{\mathrm{ave}}^{(i)}}, \frac{\bar{\lambda}}{K^2\mathrm{err}_{\max}^{(i)}}\bigg\}\bigg),
\end{aligned}$$

where $\mathrm{err}_{\mathrm{ave}}^{(i)}$ and $\mathrm{err}_{\max}^{(i)}$ are as defined in Eqs. (6.4) and (6.5). This completes the proof. □

Proof of Theorem 17

Proof of Theorem 17. The proof proceeds from partway through the proof of Theorem 16. We have already shown that on the event $\mathcal{E}_{\sin\boldsymbol{\Theta}}$ if $\|(\widehat{\mathbf{U}}\mathbf{W}_*^\top)_{i\cdot} - \mathbf{U}_{i\cdot}\| \leq \frac{1}{4\sqrt{n_{\max}}}$ then node i must be classified correctly. By repeating the argument in step 3 of the proof of Theorem 16, it holds that

$$\mathbb{P}(\mathbf{Z}_{i\cdot} \neq \widehat{\mathbf{Z}}_{i\cdot}) \leq 2K\exp\bigg(-cL\min\bigg\{\frac{\bar{\lambda}^2}{K^4\mathrm{err}_{\mathrm{ave}}^{(i)}}, \frac{\bar{\lambda}}{K^2\mathrm{err}_{\max}^{(i)}}\bigg\}\bigg) + O(n^{-10}).$$

In order for the exponential to be strictly less than $O(n^{-10})$, we require that

$$\min\bigg\{\frac{\bar{\lambda}^2}{K^4\mathrm{err}_{\mathrm{ave}}^{(i)}}, \frac{\bar{\lambda}}{K^2\mathrm{err}_{\max}^{(i)}}\bigg\} \geq \frac{C\log(n)}{L},$$

where C is a sufficiently large constant. Recalling the definitions of $\text{err}_{\text{ave}}^{(i)}$ and $\text{err}_{\text{max}}^{(i)}$, we see that we must have

$$\frac{\bar{\lambda}}{K^2} \geq \frac{C\log(n)}{L} \max_l \frac{\theta_{\text{max}}^{(l)}}{\theta_{\text{min}}^{(l)}} \frac{1}{\|\theta^{(l)}\|^2 (\lambda_{\text{min}}^{(l)})^{1/2}};$$
$$\frac{\bar{\lambda}^2}{K^4} \geq \frac{C\log(n)}{L} \left(\frac{1}{L} \sum_l \frac{\|\theta^{(l)}\|_3^3}{\theta_{\text{min}}^{(l)} \|\theta^{(l)}\|^4 \lambda_{\text{min}}^{(l)}} \right).$$

Considering the first term and rearranging, we see that we require that

$$\frac{\bar{\lambda}}{K^2} \min_l \left(\frac{\theta_{\text{min}}^{(l)}}{\theta_{\text{max}}^{(l)}} \right) \|\theta^{(l)}\|^2 (\lambda_{\text{min}}^{(l)})^{1/2} \geq \frac{C\log(n)}{L}.$$

A sufficient condition is that

$$\min_l \text{SNR}_l^2 \geq \frac{CK^8 \log(n)}{L\bar{\lambda}}$$

As for the second term, by upper bounding $\|\theta^{(l)}\|_3^3 \leq \theta_{\text{max}}^{(l)} \|\theta^{(l)}\|^2$, we see that it sufficient to have that

$$\frac{\bar{\lambda}^2}{K^4} \geq \frac{C\log(n)}{L} \left(\frac{1}{L} \sum_l \left(\frac{\theta_{\text{max}}^{(l)}}{\theta_{\text{min}}^{(l)}} \right) \frac{1}{\|\theta^{(l)}\|^2 \lambda_{\text{min}}^{(l)}} \right). \tag{6.8}$$

Therefore, rearranging (6.8) yields the sufficient condition

$$\left(\frac{1}{L} \sum_l \frac{1}{\text{SNR}_l^2} \right)^{-1} \geq C \frac{K^8 \log(n)}{L\bar{\lambda}^2}.$$

It is straightforward to check that the condition in Theorem 17 is sufficient for the result to hold. □

Appendix A

Proofs from Chapter 1

A.1 Proofs of Matrix Denoising Results (Theorems 1, 2, and 3)

In this section we provide proofs of the main results in this chapter. Our proofs rely on a number of auxiliary lemmas; these are proven in the following subsection. We prove each theorem sequentially, as each theorem relies on the previous results and their proofs as well..

Proof of Theorem 1. At the outset, we note that by Theorem 4.4.5 of Vershynin (2018), $\|\mathbf{N}\| \lesssim \sigma\sqrt{n}$ with probability at least $1 - e^{-cn}$. Let $\widehat{\lambda}_r$ denotes the r'th largest eigenvalue of $\widehat{\mathbf{S}}$. Then by Weyl's inequality it holds that

$$|\widehat{\lambda}_r - \lambda_r| \lesssim \sigma\sqrt{n},$$

which implies that $\widehat{\lambda}_r \geq \lambda_r/2$ for n sufficiently large, since $\lambda_r \geq C\sigma\sqrt{n\log(n)}$ by assumption. We will use this result without additional reference in the subsequent analysis.

We now expand $\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*$ via

$$\begin{aligned}\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_* &= \mathbf{N}\widehat{\mathbf{U}}\widehat{\Lambda}^{-1} + \mathbf{U}\mathbf{U}^\top\mathbf{N}\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1} + \mathbf{U}\mathbf{U}^\top\mathbf{N}\mathbf{U}_\perp\mathbf{U}_\perp^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1} + \mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*) \\ &:= \mathbf{N}\widehat{\mathbf{U}}\widehat{\Lambda}^{-1} + (I) + (II) + (III),\end{aligned}$$

with

$$(I) := \mathbf{U}\mathbf{U}^\top \mathbf{N}\mathbf{U}\mathbf{U}^\top \widehat{\mathbf{U}}\widehat{\Lambda}^{-1}$$
$$(II) := \mathbf{U}\mathbf{U}^\top \mathbf{N}\mathbf{U}_\perp \mathbf{U}_\perp^\top \widehat{\mathbf{U}}\widehat{\Lambda}^{-1}$$
$$(III) := \mathbf{U}(\mathbf{U}^\top \widehat{\mathbf{U}} - \mathbf{W}_*).$$

The following lemma bounds the terms $(I), (II)$, and (III) respectively.

Lemma 9. *The following bounds hold with probability at least* $1 - O(n^{-30})$*:*

$$\|\mathbf{U}\mathbf{U}^\top \mathbf{N}\mathbf{U}\mathbf{U}^\top \widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\|_{2,\infty} \lesssim \frac{\sigma}{\lambda_r}\mu_0\sqrt{\frac{r}{n}}\Big(\sqrt{r} + \sqrt{\log(n)}\Big);$$
$$\|\mathbf{U}\mathbf{U}^\top \mathbf{N}\mathbf{U}_\perp \mathbf{U}_\perp^\top \widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r}{n}}\frac{\sigma^2 n}{\lambda_r^2};$$
$$\|\mathbf{U}(\mathbf{U}^\top \widehat{\mathbf{U}} - \mathbf{W}_*)\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r}{n}}\frac{\sigma^2 n}{\lambda_r^2}.$$

Therefore, by Lemma 9, we obtain that with probability at least $1 - O(n^{-30})$

$$\big(\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\big)_{i\cdot} = \big(\mathbf{N}\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\big)_{i\cdot} + O\bigg(\frac{\mu_0\big(r + \sqrt{r\log(n)}\big)}{\sqrt{n}(\lambda_r/\sigma)} + \frac{\mu_0\sqrt{rn}}{(\lambda_r/\sigma)^2}\bigg). \tag{A.1}$$

It therefore suffices to focus on the first term. Observe that

$$\|\big(\mathbf{N}\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\big)_{i\cdot}\| \lesssim \frac{1}{\lambda_r}\|\big(\mathbf{N}\widehat{\mathbf{U}}\big)_{i\cdot}\|.$$

Consequently, it suffices to analyze the term $\|\big(\mathbf{N}\widehat{\mathbf{U}}\big)_{i\cdot}\|$. This requires the use of the leave-one-out strategy from Abbe et al. (2020) (see also the proofs in Chapter 6 and Chapter 4), though other methods are available (e.g., see the proofs from Chapter 2 or Chapter 3). For now, we simply state the following lemma.

Lemma 10. *Let* i *be fixed and instate the conditions of Theorem 1. Then, with probability at least* $1 - O(n^{-30})$ *it holds that*

$$\|\big(\mathbf{N}\widehat{\mathbf{U}}\big)_{i\cdot}\| \lesssim \sigma\sqrt{n\log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty}.$$

Therefore, by Lemma 10 and (A.1), we obtain that with probability at least $1-O(n^{-30})$,

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} \lesssim \frac{\sigma\sqrt{n\log(n)}}{\lambda_r}\|\widehat{\mathbf{U}}\|_{2,\infty} + \mu_0\sqrt{\frac{r}{n}}\left(\frac{\sqrt{r}+\sqrt{\log(n)}}{\lambda_r/\sigma} + \frac{\sigma^2 n}{(\lambda_r/\sigma)^2}\right).$$

As a result,

$$\begin{aligned}\|\widehat{\mathbf{U}}\|_{2,\infty} &\leq \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} + \|\mathbf{U}\|_{2,\infty} \\ &\leq \frac{\sigma\sqrt{n\log(n)}}{\lambda_r}\|\widehat{\mathbf{U}}\|_{2,\infty} + \mu_0\sqrt{\frac{r}{n}}.\end{aligned}$$

Since $\lambda_r/\sigma \geq C\sqrt{n\log(n)}$, it holds that $\|\widehat{\mathbf{U}}\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r}{n}}$. Putting it all together yields that with probability at least $1 - O(n^{-30}) \geq 1 - n^{-20}$,

$$\begin{aligned}\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} &\leq \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r} + \mu_0\sqrt{\frac{r}{n}}\left(\frac{r+\sqrt{r\log(n)}}{\lambda_r/\sigma} + \frac{\sigma^2 n}{(\lambda_r/\sigma)^2}\right) \\ &\asymp \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r} + \frac{r+\sqrt{r\log(n)}}{\sqrt{n}(\lambda_r/\sigma)} + \frac{\mu_0\sigma^2\sqrt{rn}}{(\lambda_r/\sigma)^2} \\ &\asymp \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r},\end{aligned}$$

which is the desired bound in Theorem 1, which completes the proof. □

Proof of 2. We start midway through the proof of Theorem 1 to demonstrate that by (A.1) we have the expansion

$$\left(\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U}\right)_{i\cdot} = e_i^\top \mathbf{N}\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\mathbf{W}_*^\top + O\left(\frac{\mu_0\left(r+\sqrt{r\log(n)}\right)}{\sqrt{n}(\lambda_r/\sigma)} + \frac{\mu_0\sqrt{rn}}{(\lambda_r/\sigma)^2}\right) \tag{A.2}$$

In the previous part of the proof we bounded the first term directly; we now expand it out further. It holds that

$$\begin{aligned}e_i^\top \mathbf{N}\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\mathbf{W}_*^\top &= e_i^\top \mathbf{N}\mathbf{U}\Lambda^{-1} + e_i^\top \mathbf{N}\left[\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\right]\widehat{\Lambda}^{-1}\mathbf{W}_*^\top \\ &\quad + e_i^\top \mathbf{N}\mathbf{U}\left[\Lambda^{-1}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\right]\mathbf{W}_*^\top + e_i^\top \mathbf{N}\mathbf{U}\Lambda^{-1}(\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}).\end{aligned}$$

The following lemmas bound these three terms. The first term requires the use of leave-one-out sequences, so we state it as its own independent lemma.

Lemma 11. *In the context of Theorem 2, the following bound holds with probability at least* $1 - O(n^{-20})$*:*

$$\|\mathbf{N}\big[\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\widehat{\Lambda}^{-1}\mathbf{W}_*^\top\|_{2,\infty} \lesssim \frac{\sigma^2\mu_0\sqrt{rn}\log(n)}{\lambda_r^2}$$

Finally, the next lemma bounds the remaining two terms.

Lemma 12. *In the context of Theorem 2, the following bounds hold with probability at least* $1 - O(n^{-20})$ *:*

$$\|\mathbf{N}\mathbf{U}\big[\Lambda^{-1}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\big]\mathbf{W}_*^\top\|_{2,\infty} \lesssim \frac{\sigma^2\mu_0\sqrt{rn\log(n)}}{\lambda_r^2};$$
$$\|\mathbf{N}\mathbf{U}\Lambda^{-1}(\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}})\|_{2,\infty} \lesssim \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r}\frac{\sigma^2 n}{\lambda_r}.$$

Therefore, by (A.2), Lemma 11,Lemma 12, and the fact that $\lambda_r/\sigma \geq C\sqrt{n\log(n)}$, it holds that

$$\begin{aligned}\big(\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U}\big)_{i\cdot} &= e_i^\top\mathbf{N}\mathbf{U}\Lambda^{-1} + O\bigg(\frac{\mu_0\big(r+\sqrt{r\log(n)}\big)}{\sqrt{n}(\lambda_r/\sigma)} + \frac{\mu_0\sqrt{rn}}{(\lambda_r/\sigma)^2} + \frac{\mu_0\sqrt{rn}\log(n)}{(\lambda_r/\sigma)^2} + \frac{\mu_0\sqrt{rn\log(n)}}{(\lambda_r/\sigma)^2}\bigg)\\ &= e_i^\top\mathbf{N}\mathbf{U}\Lambda^{-1} + O\bigg(\frac{\mu_0\big(r+\sqrt{r\log(n)}\big)}{\sqrt{n}(\lambda_r/\sigma)} + \frac{\mu_0\sqrt{rn}\log(n)}{(\lambda_r/\sigma)^2}\bigg).\end{aligned}$$

Therefore, we have shown that there is an event $\mathcal{E}$ satisfying $\mathbb{P}\big(\mathcal{E}\big) \geq 1 - n^{-10}$ such that

$$\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U} = \mathbf{N}\mathbf{U}\Lambda^{-1} + \Gamma,$$

where

$$\|\Gamma\|_{2,\infty} \lesssim \frac{\mu_0\big(r+\sqrt{r\log(n)}\big)}{\sqrt{n}(\lambda_r/\sigma)} + \frac{\mu_0\sqrt{rn}\log(n)}{(\lambda_r/\sigma)^2}.$$

This completes the proof. □

Proof of Theorem 3. When r is fixed, we note that

$$e_i^\top \mathbf{N}\mathbf{U}\Lambda^{-1} = \sum_j \mathbf{N}_{ij}\mathbf{U}_{j\cdot}\Lambda^{-1},$$

and it is not hard to see that when $\mathbb{E}\mathbf{N}_{ij}^2 = \sigma^2$ this term has covariance $\sigma^2\Lambda^{-2}$. Consequently, by the Lindeberg-Feller Central Limit Theorem it holds that

$$\frac{1}{\sigma}\Lambda\Big(\mathbf{N}\mathbf{U}\Lambda^{-1}\Big)_{i\cdot} \to \mathcal{N}(0, \mathbf{I}_r),$$

as long as $\mu_0 = O(1)$. The result is then completed by Slutsky's Theorem and the fact that

$$\frac{1}{\sigma}\|\Lambda\|\Gamma \le \frac{\kappa\mu_0\big(r + \sqrt{r\log(n)}\big)}{\sqrt{n}} + \frac{\kappa\mu_0\sqrt{rn}\log(n)}{\lambda_r/\sigma} \to 0,$$

since $\lambda_r/\sigma \gg \kappa\mu_0\sqrt{rn}\log(n)$ when $\kappa, \mu_0, r = O(1)$. □

Proof of Theorem 4. First we note that $\|\widehat{\Lambda}\| \lesssim \kappa\lambda_r$ on the high probability event $\|\mathbf{N}\| \lesssim \sigma\sqrt{n}$. Next, observe that by Theorem 2

$$\begin{aligned}
&\Big\|\widehat{\mathbf{U}}\widehat{\Lambda}\mathbf{W}_*^\top - \mathbf{U}\Lambda - \mathbf{N}\mathbf{U}\Big\|_{2,\infty} \\
&\le \Big\|\widehat{\mathbf{U}}\Big[\widehat{\Lambda}(\mathbf{W}_*^\top - \widehat{\mathbf{U}}^\top\mathbf{U}) \\
&\qquad + \widehat{\Lambda}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}^\top\mathbf{U}\Lambda + (\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{W}_*^\top)\Lambda\Big]\Big\|_{2,\infty} + \Big\|\big(\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U} - \mathbf{N}\mathbf{U}\Lambda^{-1}\big)\Lambda\Big\|_{2,\infty} \\
&\le \|\widehat{\mathbf{U}}\|_{2,\infty}\Big(\kappa\lambda_r\|\mathbf{W}_*^\top - \widehat{\mathbf{U}}^\top\mathbf{U}\| + \|\widehat{\Lambda}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}^\top\mathbf{U}\Lambda\|\Big) + \|\Gamma\Lambda\|_{2,\infty}. \qquad \text{(A.3)}
\end{aligned}$$

The proof of Lemma 9 reveals that

$$\|\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*\| \lesssim \frac{\sigma^2 n}{\lambda_r^2}.$$

Similarly, the proof of Lemma 12 reveals that

$$\|\widehat{\Lambda}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}^\top\mathbf{U}\Lambda\| \lesssim \sigma\sqrt{r} + \sigma\sqrt{\log(n)} + \sigma\sqrt{n}\frac{\sigma\sqrt{n}}{\lambda_r}.$$

By Theorem 1 it holds that $\|\widehat{\mathbf{U}}\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r}{n}}$. Plugging in these bounds to (A.3) reveals that

$$\begin{aligned}
\left\|\widehat{\mathbf{U}}\widehat{\Lambda}\mathbf{W}_*^\top - \mathbf{U}\Lambda - \mathbf{N}\mathbf{U}\right\|_{2,\infty} &\lesssim \mu_0\sqrt{\frac{r}{n}}\left(\kappa\lambda_r\frac{\sigma^2 n}{\lambda_r^2} + \sigma\sqrt{r} + \sigma\sqrt{\log(n)} + \sigma\sqrt{n}\frac{\sigma\sqrt{n}}{\lambda_r}\right) + \|\Gamma\Lambda\|_{2,\infty} \\
&\lesssim \frac{\mu_0\kappa\sigma^2\sqrt{nr}}{\lambda_r} + \frac{\mu_0\sigma r}{\sqrt{n}} + \frac{\mu_0\sigma\sqrt{r\log(n)}}{\sqrt{n}} + \frac{\mu_0\sigma^2\sqrt{rn}}{\lambda_r} + \kappa\lambda_r\|\Gamma\|_{2,\infty} \\
&\lesssim \frac{\mu_0\kappa\sigma^2\sqrt{nr}}{\lambda_r} + \frac{\mu_0\sigma r}{\sqrt{n}} + \frac{\mu_0\sigma\sqrt{r\log(n)}}{\sqrt{n}} + \frac{\mu_0\sigma^2\sqrt{rn}}{\lambda_r} \\
&\quad + \kappa\lambda_r\left(\frac{\mu_0\sigma(r+\sqrt{r\log(n)})}{\lambda_r\sqrt{n}} + \frac{\mu_0\sigma^2\sqrt{rn}\log(n)}{\lambda_r^2}\right) \\
&\asymp \sigma\left(\frac{\kappa\big(\mu_0 r + \mu_0\sqrt{r\log(n)}\big)}{\sqrt{n}} + \frac{\mu_0\kappa\sqrt{rn}\log(n)}{\lambda_r/\sigma}\right) \\
&= o(\sigma)
\end{aligned}$$

which holds since $\kappa, \mu_0, r = O(1)$ and $\lambda_r/\sigma \gg \sqrt{n}\log(n)$ by assumption. Therefore, by Slutsky's Theorem, these results demonstrate that

$$\frac{1}{\sigma}e_i^\top\left(\widehat{\mathbf{U}}\widehat{\Lambda}\mathbf{W}_*^\top - \mathbf{U}\Lambda\right) \to \mathcal{N}(0, \mathbf{I}_r)$$

in distribution. Furthermore, $e_i^\top\left(\widehat{\mathbf{U}}\widehat{\Lambda}\mathbf{W}_*^\top\right)$ is asymptotically independent from $e_j^\top\left(\widehat{\mathbf{U}}\widehat{\Lambda}\mathbf{W}_*^\top\right)$ (since they only depend on the shared diagonal element, which is negligible). Together this implies that

$$\frac{1}{2\sigma^2}(e_i - e_j)^\top\left(\widehat{\mathbf{U}}\widehat{\Lambda}\mathbf{W}_*^\top - \mathbf{U}\Lambda\right) \to \mathcal{N}\big(0, \mathbf{I}_r\big)$$

in distribution.

We now analyze the test statistic under the null and alternative respectively. When $\mathbf{S}_{i\cdot} = \mathbf{S}_{j\cdot}$, it holds that $(e_i - e_j)^\top\mathbf{U}\Lambda = 0$, since $\|\mathbf{S}_{i\cdot} - \mathbf{S}_{j\cdot}\| = \|(e_i - e_j)^\top\mathbf{U}\Lambda\mathbf{U}^\top\| = \|(e_i - e_j)^\top\mathbf{U}\Lambda\|$. Therefore, under the null hypothesis it holds that

$$\frac{1}{2\sigma^2}(e_i - e_j)^\top\left(\widehat{\mathbf{U}}\widehat{\Lambda}\mathbf{W}_*^\top\right) \to \mathcal{N}\big(0, \mathbf{I}_r\big),$$

and hence the continuous mapping theorem implies that

$$T_{ij}^2 = \frac{1}{2\sigma^2}\|(e_i - e_j^\top)\widehat{\mathbf{U}}\widehat{\Lambda}\|^2 = \frac{1}{2\sigma^2}\|(e_i - e_j)^\top \widehat{\mathbf{U}}\widehat{\Lambda}\mathbf{W}_*^\top\|^2 \to \chi_r^2$$

in distribution.

Under the local alternative

$$\|\mathbf{S}_{i\cdot} - \mathbf{S}_{j\cdot}\| \gg \sigma,$$

it holds that

$$\|(e_i - e_j)^\top \mathbf{U}\Lambda\|^2 \gg \sigma^2,$$

and hence it holds that

$$T_{ij}^2 = \frac{1}{2\sigma^2}\|(e_i - e_j^\top)\widehat{\mathbf{U}}\widehat{\Lambda}\|^2 = \frac{1}{2\sigma^2}\|(e_i - e_j)^\top \widehat{\mathbf{U}}\widehat{\Lambda}\mathbf{W}_*^\top\|^2 \to \infty$$

in probability. Under the weaker condition

$$\frac{1}{2\sigma^2}\|\mathbf{S}_{i\cdot} - \mathbf{S}_{j\cdot}\|^2 \to \mu < \infty,$$

the Continuous Mapping Theorem implies that

$$\frac{1}{2\sigma^2}T_{ij}^2 \to \chi_r^2(\mu),$$

as desired. This completes the proof. □

A.2 Proofs of Auxiliary Lemmas

A.2.1 Proof of Lemma 9

Proof of Lemma 9. Throughout we use the fact that $\widehat{\lambda}_r \gtrsim \lambda_r$ and $\|\mathbf{N}\| \lesssim \sigma\sqrt{n}$ with overwhelming probability.

For the first term, observe that

$$\begin{aligned}\|\mathbf{U}\mathbf{U}^\top\mathbf{N}\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\|_{2,\infty} &\lesssim \|\mathbf{U}\|_{2,\infty}\|\mathbf{U}^\top\mathbf{N}\mathbf{U}\|\|\widehat{\Lambda}^{-1}\| \\ &\lesssim \mu_0\sqrt{\frac{r}{n}}\frac{\|\mathbf{U}^\top\mathbf{N}\mathbf{U}\|}{\lambda_r}.\end{aligned}$$

It now follows from a straightforward ε-net argument(e.g., the proof of Theorem 4.4.5 of Vershynin (2018)) that $\|\mathbf{U}^\top\mathbf{N}\mathbf{U}\| \lesssim \sigma\big(\sqrt{r}+\sqrt{\log(n)}\big)$ with probability at least $1-O(n^{-30})$. Therefore, with this same probability,

$$\|\mathbf{U}\mathbf{U}^\top\mathbf{N}\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\|_{2,\infty} \lesssim \frac{\sigma}{\lambda_r}\mu_0\sqrt{\frac{r}{n}}\bigg(\sqrt{r}+\sqrt{\log(n)}\bigg).$$

For the next term, it holds that

$$\begin{aligned}\|\mathbf{U}\mathbf{U}^\top\mathbf{N}\mathbf{U}_\perp\mathbf{U}_\perp^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\|_{2,\infty} &\lesssim \mu_0\sqrt{\frac{r}{n}}\frac{\|\mathbf{N}\|\|\mathbf{U}_\perp^\top\widehat{\mathbf{U}}\|}{\lambda_r} \\ &\lesssim \mu_0\sqrt{\frac{r}{n}}\frac{\sigma\sqrt{n}}{\lambda_r}\|\mathbf{U}_\perp^\top\widehat{\mathbf{U}}\|.\end{aligned}$$

Observe that $\|\mathbf{U}_\perp^\top\widehat{\mathbf{U}}\| = \|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\|$. Consequently, by the Davis-Kahan Theorem,

$$\|\mathbf{U}_\perp^\top\widehat{\mathbf{U}}\| \lesssim \frac{\sigma\sqrt{n}}{\lambda_r}.$$

Putting it together yields that

$$\|\mathbf{U}\mathbf{U}^\top\mathbf{N}\mathbf{U}_\perp\mathbf{U}_\perp^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\|_{2,\infty} \lesssim \frac{\sigma^2 n}{\lambda_r^2}.$$

For the final term, we note that

$$\|\mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}}-\mathbf{W}_*)\|_{2,\infty} \le \mu_0\sqrt{\frac{r}{n}}\|\mathbf{U}^\top\widehat{\mathbf{U}}-\mathbf{W}_*\|.$$

It is straightforward to check that since $\mathbf{W}_* = \mathrm{sgn}(\mathbf{U}^\top\widehat{\mathbf{U}})$, one has the bound

$$\|\mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}}-\mathbf{W}_*)\|_{2,\infty} \le \mu_0\sqrt{\frac{r}{n}}\|\mathbf{U}^\top\widehat{\mathbf{U}}-\mathbf{W}_*\| \le \|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\|^2 \lesssim \frac{\sigma^2 n}{\lambda_r^2}.$$

For details see, for example, Lemma 4.15 of Chen et al. (2021c). Combining these bounds completes the proof. □

A.2.2 Proof of Lemma 10

Proof of Lemma 10. Let $\widetilde{\mathbf{U}}^{(i)}$ be the estimate obtained by setting the i'th row and column of $\mathbf{N}$ to zero. Then

$$\begin{aligned}\|\big(\mathbf{N}\widehat{\mathbf{U}}\big)_{i\cdot}\| &\leq \|\Big(\mathbf{N}(\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top\widehat{\mathbf{U}})\Big)_{i\cdot}\| + \|\big(\mathbf{N}\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top\widehat{\mathbf{U}}\big)_{i\cdot}\| \\ &\leq \|\mathbf{N}\|\|\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top\widehat{\mathbf{U}}\| + \|\big(\mathbf{N}\widetilde{\mathbf{U}}^{(i)}\big)_{i\cdot}\|. \end{aligned} \tag{A.4}$$

Let $\widetilde{\lambda}_{r+1}^{(i)}$ denote the $r+1$'st eigenvalue of $\mathbf{S} + \mathbf{N}^{(i)}$, where $\mathbf{N}^{(i)}$ has its i'th row and column set to zero. Then it holds that

$$\|\widehat{\mathbf{S}} - \mathbf{S} - \mathbf{N}^{(i)}\| \leq \|\mathbf{N} - \mathbf{N}^{(i)}\| \leq 3\|(\mathbf{N})_{i\cdot}\| \leq \|\mathbf{N}\| \lesssim \sigma\sqrt{n},$$

and, hence, since $\widehat{\lambda}_r \geq \lambda_r/2$, by Weyl's inequality it holds that

$$\|\widehat{\lambda}_r - \widetilde{\lambda}_{r+1}| \geq \lambda_r/2 - C\sigma\sqrt{n} \geq \lambda_r/4 \gtrsim \lambda.$$

Therefore, we can apply the Davis-Kahan Theorem (Theorem 2.7 of Chen et al. (2021c)) to yield that

$$\begin{aligned}
\|\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top \widehat{\mathbf{U}}\| &\leq \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\| \\
&\lesssim \|\sin\Theta(\widehat{\mathbf{U}}, \widetilde{\mathbf{U}}^{(i)})\| \\
&\lesssim \frac{\|(\mathbf{N} - \mathbf{N}^{(i)})\widetilde{\mathbf{U}}^{(i)}\|}{\lambda_r} \\
&\lesssim \frac{\|e_i^\top \mathbf{N}\widetilde{\mathbf{U}}\| + \|\mathbf{N}e_i e_i^\top \widetilde{\mathbf{U}}\| + \|e_i^\top \mathbf{N} e_i e_i^\top \widetilde{\mathbf{U}}\|}{\lambda_r} \\
&\lesssim \frac{\|e_i^\top \mathbf{N}\widetilde{\mathbf{U}}^{(i)}\|}{\lambda_r} + \frac{\|\mathbf{N}\|}{\lambda_r}\|e_i^\top \widetilde{\mathbf{U}}^{(i)}\| \\
&\lesssim \frac{\|e_i^\top \mathbf{N}\widetilde{\mathbf{U}}^{(i)}\|}{\lambda_r} + \frac{\|\mathbf{N}\|}{\lambda_r}\|\widetilde{\mathbf{U}}^{(i)}\|_{2,\infty} \\
&\lesssim \frac{\|e_i^\top \mathbf{N}\widetilde{\mathbf{U}}^{(i)}\|}{\lambda_r} + \frac{\sigma\sqrt{n}}{\lambda_r}\|\widetilde{\mathbf{U}}^{(i)}\|_{2,\infty} \qquad \text{(A.5)}
\end{aligned}$$

Observe that

$$e_i^\top \mathbf{N}\widetilde{\mathbf{U}}^{(i)} = \sum_{j=1}^{n} \mathbf{N}_{ij}\big(\widetilde{\mathbf{U}}^{(i)}\big)_{j\cdot},$$

which is sum of independent mean-zero random matrices, conditional on $\mathbf{N}^{(i)}$. By the Matrix Hoeffing inequality (Tropp, 2015), it holds that

$$\|\big(\mathbf{N}\widetilde{\mathbf{U}}^{(i)}\big)_{i\cdot}\| \lesssim \sigma\sqrt{n\log(n)}\|\widetilde{\mathbf{U}}\|_{2,\infty} \qquad \text{(A.6)}$$

with probability at least $1 - O(n^{-30})$. Plugging this bound into Eq. (A.5) and further simplifying, we obtain that

$$\begin{aligned}
\|\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top \widehat{\mathbf{U}}\| &\lesssim \frac{\sigma\sqrt{n\log(n)}}{\lambda_r}\|\widetilde{\mathbf{U}}\|_{2,\infty} + \frac{\sigma\sqrt{n}}{\lambda_r}\|\widetilde{\mathbf{U}}^{(i)}\|_{2,\infty} \\
&\asymp \frac{\sigma\sqrt{n\log(n)}}{\lambda_r}\|\widetilde{\mathbf{U}}^{(i)}\|_{2,\infty}.
\end{aligned}$$

As a byproduct of this bound, we also obtain

$$\begin{aligned}\|\widetilde{\mathbf{U}}^{(i)}\|_{2,\infty} &\leq \|\widetilde{\mathbf{U}}^{(i)} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\widetilde{\mathbf{U}}^{(i)}\|_{2,\infty} + \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\widetilde{\mathbf{U}}^{(i)}\|_{2,\infty} \\ &\leq \|\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\| + \|\widehat{\mathbf{U}}\|_{2,\infty} \\ &\lesssim \frac{\sigma\sqrt{n\log(n)}}{\lambda_r}\|\widetilde{\mathbf{U}}^{(i)}\|_{2,\infty} + \|\widehat{\mathbf{U}}\|_{2,\infty},\end{aligned}$$

which shows that by rearranging $\|\widetilde{\mathbf{U}}^{(i)}\|_{2,\infty} \leq 2\|\widehat{\mathbf{U}}\|_{2,\infty}$ since $\lambda_r/\sigma \geq C\sqrt{n\log(n)}$ for some sufficiently large constant C. Therefore, with probability at least $1 - O(n^{-30})$,

$$\|\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top\widehat{\mathbf{U}}\| \lesssim \frac{\sigma\sqrt{n\log(n)}}{\lambda_r}\|\widehat{\mathbf{U}}\|_{2,\infty}.$$

By Eq. (A.6), we have also therefore shown that

$$\|\big(\mathbf{N}\widetilde{\mathbf{U}}^{(i)}\big)_{i\cdot}\| \lesssim \sigma\sqrt{n\log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty}.$$

Plugging in these bounds to (A.4) we obtain that

$$\begin{aligned}\|\big(\mathbf{N}\widehat{\mathbf{U}}\big)_{i\cdot}\| &\leq \|\mathbf{N}\|\|\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top\widehat{\mathbf{U}}\| + \|\big(\mathbf{N}\widetilde{\mathbf{U}}^{(i)}\big)_{i\cdot}\| \\ &\lesssim \sigma\sqrt{n}\frac{\sigma\sqrt{n\log(n)}}{\lambda_r}\|\widehat{\mathbf{U}}\|_{2,\infty} + \sigma\sqrt{n\log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty} \\ &\asymp \sigma\sqrt{n\log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty},\end{aligned}$$

which all hold cumulatively with probability at least $1 - O(n^{-30})$, where the final bound follows from the fact that $\lambda_r/\sigma \geq C\sqrt{n\log(n)}$ for some sufficiently large constant C. This completes the proof. □

A.2.3 Proof of Lemma 11

Proof of Lemma 11. We will first fix the i'th row, and we will use the leave-one-out sequences from the proof of Lemma 9. Let $\widetilde{\mathbf{U}}^{(i)}$ be the leading eigenvectors of $\mathbf{S} + \mathbf{N}^{(i)}$, where $\mathbf{N}^{(i)}$ is

obtained by setting the i'th row and column of $\mathbf{N}$ to zero. For the first term, we have that

$$\begin{aligned}\|e_i^\top \mathbf{N}\big[\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\widehat{\Lambda}^{-1}\mathbf{W}_*^\top\| &\le \|e_i^\top \mathbf{N}\big[\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\widehat{\mathbf{U}}\big]\widehat{\Lambda}^{-1}\mathbf{W}_*^\top\| + \|e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\widehat{\Lambda}^{-1}\mathbf{W}_*^\top\| \\ &\lesssim \frac{1}{\lambda_r}\Big(\|e_i^\top \mathbf{N}\big[\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\widehat{\mathbf{U}}\big]\| + \|\|e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\|\Big) \\ &\lesssim \frac{1}{\lambda_r}\Big(\|\mathbf{N}\|\|\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\| + \|\|e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\|\Big) \\ &\lesssim \frac{1}{\lambda_r}\Big(\sigma\sqrt{n}\|\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\widehat{\mathbf{U}}\| + \|\|e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\|\Big).\end{aligned}$$

By repeating the argument in the proof of Lemma 10, with probability at least $1 - O(n^{-30})$ it holds that

$$\begin{aligned}\|\widehat{\mathbf{U}} - \widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top\widehat{\mathbf{U}}\| &\lesssim \frac{\sigma\sqrt{n\log(n)}}{\lambda_r}\|\widehat{\mathbf{U}}\|_{2,\infty} \\ &\lesssim \frac{\sigma\sqrt{n\log(n)}}{\lambda_r}\mu_0\sqrt{\frac{r}{n}} \\ &\asymp \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r},\end{aligned} \tag{A.7}$$

where we have used the fact that on the event in Theorem 1 it holds that $\|\widehat{\mathbf{U}}\|_{2,\infty} \le 2\mu_0\sqrt{\frac{r}{n}}$. Therefore, we have that

$$\begin{aligned}\|e_i^\top \mathbf{N}\big[\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\widehat{\Lambda}^{-1}\mathbf{W}_*^\top\| &\lesssim \frac{1}{\lambda_r}\Big(\sigma\sqrt{n}\frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r} + \|\|e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\|\Big) \\ &\asymp \frac{\sigma^2\mu_0\sqrt{rn\log(n)}}{\lambda_r^2} + \frac{\|e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\|}{\lambda_r}.\end{aligned} \tag{A.8}$$

We now observe that

$$\|e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\| = \|e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top - \mathbf{U}\mathbf{U}^\top\big]\|.$$

Observe that by construction $\widetilde{\mathbf{U}}^{(i)}$ is independent of the i'th row of $\mathbf{N}$, and hence $\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top - \mathbf{U}\mathbf{U}^\top$ is independent of the i'th row of $\mathbf{N}$. Therefore, the term $e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top - \mathbf{U}\mathbf{U}^\top\big]$ is a sum of independent mean-zero random matrices (conditional on $\mathbf{N}^{(i)}$), and by the Matrix

Hoeffding inequality (Tropp, 2015), we obtain that

$$\|e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top - \mathbf{U}\mathbf{U}^\top\big]\| \lesssim \sigma\sqrt{n\log(n)}\|\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top - \mathbf{U}\mathbf{U}^\top\|_{2,\infty}.$$

Next, we note that by (A.7),

$$\begin{aligned}\|\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top - \mathbf{U}\mathbf{U}^\top\|_{2,\infty} &\leq \|\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\| + \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_{2,\infty} \\ &\lesssim \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r} + \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_{2,\infty}.\end{aligned}$$

Next, we note that with probability at least $1 - O(n^{-30})$,

$$\begin{aligned}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\|_{2,\infty} &\leq \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{U}^\top - \mathbf{U}\mathbf{U}^\top\|_{2,\infty} + \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\|_{2,\infty} \\ &\leq \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{U}\|_{2,\infty} + \|\widehat{\mathbf{U}}\|_{2,\infty}\|\widehat{\mathbf{U}}^\top(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\| \\ &\leq \|\widehat{\mathbf{U}}(\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{W}_*^\top)\|_{2,\infty} + \|\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U}\|_{2,\infty} + \|\widehat{\mathbf{U}}\|_{2,\infty}\|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\| \\ &\leq \|\widehat{\mathbf{U}}\|_{2,\infty}\|\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*\| + \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r} + \|\widehat{\mathbf{U}}\|_{2,\infty}\frac{\sigma\sqrt{n}}{\lambda_r} \\ &\lesssim \mu_0\sqrt{\frac{r}{n}}\|\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*\| + \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r} + \mu_0\sqrt{\frac{r}{n}}\frac{\sigma\sqrt{n}}{\lambda_r}\end{aligned}$$

where in the penultimate line we have applied Theorem 1 as well as the fact that $\|\sin\Theta(\widehat{\mathbf{U}},\mathbf{U})\| \lesssim \frac{\sigma\sqrt{n}}{\lambda_r}$ with this same probability, and the final line we have used the fact that by Theorem 1 $\|\widehat{\mathbf{U}}\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r}{n}}$. By arguing as in the proof of Lemma 9, it holds that

$$\|\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}\| \lesssim \frac{\sigma^2 n}{\lambda_r}.$$

Putting it all together, we obtain

$$\begin{aligned}\|e_i^\top \mathbf{N}\big[\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top - \mathbf{U}\mathbf{U}^\top\big]\| &\lesssim \sigma\sqrt{n\log(n)}\|\widetilde{\mathbf{U}}^{(i)}(\widetilde{\mathbf{U}}^{(i)})^\top - \mathbf{U}\mathbf{U}^\top\|_{2,\infty} \\ &\lesssim \sigma\sqrt{n\log(n)}\frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r}.\end{aligned}$$

Plugging this into Eq. (A.8), we obtain the final bound

$$\begin{aligned}\|e_i^\top \mathbf{N}\big[\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big]\widehat{\Lambda}^{-1}\mathbf{W}_*^\top\| &\lesssim \frac{\sigma^2\mu_0\sqrt{rn\log(n)}}{\lambda_r^2} + \frac{1}{\lambda_r}\left(\sigma\sqrt{n\log(n)}\frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r}\right)\\ &\asymp \frac{\sigma^2\mu_0\sqrt{rn}\log(n)}{\lambda_r^2},\end{aligned}$$

which holds with probability at least $1 - O(n^{-20})$. This completes the proof. □

A.2.4 Proof of Lemma 12

Proof of Lemma 12. For both terms, we first note that the Matrix Hoeffing inequality and a union bound implies that with probability at least $1 - O(n^{-20})$,

$$\begin{aligned}\|\mathbf{N}\mathbf{U}\|_{2,\infty} &\lesssim \sigma\sqrt{n\log(n)}\|\mathbf{U}\|_{2,\infty}\\ &\lesssim \sigma\mu_0\sqrt{r\log(n)}.\end{aligned}$$

Therefore, we observe that

$$\begin{aligned}\|\mathbf{N}\mathbf{U}\big[\Lambda^{-1}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\big]\mathbf{W}_*^\top\|_{2,\infty} &\lesssim \|\mathbf{N}\mathbf{U}\|_{2,\infty}\|\Lambda^{-1}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\|\\ &\lesssim \sigma\mu_0\sqrt{r\log(n)}\|\Lambda^{-1}\big(\mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda} - \Lambda\mathbf{U}^\top\widehat{\mathbf{U}}\big)\widehat{\Lambda}^{-1}\|\\ &\lesssim \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r^2}\|\mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda} - \Lambda\mathbf{U}^\top\widehat{\mathbf{U}}\|. \qquad \text{(A.9)}\end{aligned}$$

By the eigenvector-eigenvalue equation, it holds that

$$\begin{aligned}\|\mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda} - \Lambda\mathbf{U}^\top\widehat{\mathbf{U}}\| &= \|\mathbf{U}^\top\widehat{\mathbf{S}}\widehat{\mathbf{U}} - \mathbf{U}^\top\mathbf{S}\widehat{\mathbf{U}}\|\\ &= \|\mathbf{U}^\top\mathbf{N}\widehat{\mathbf{U}}\|\\ &\le \|\mathbf{U}^\top\mathbf{N}\mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\| + \|\mathbf{U}^\top\mathbf{N}(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\widehat{\mathbf{U}}\|\\ &\le \|\mathbf{U}^\top\mathbf{N}\mathbf{U}\| + \|\mathbf{N}\|\|\sin\Theta(\mathbf{U},\widehat{\mathbf{U}})\|\\ &\lesssim \sigma\left(\sqrt{r} + \sqrt{\log(n)}\right) + \sigma\sqrt{n}\|\sin\Theta(\mathbf{U},\widehat{\mathbf{U}})\|\\ &\lesssim \sigma\sqrt{r} + \sigma\sqrt{\log(n)} + \sigma\sqrt{n}\frac{\sigma\sqrt{n}}{\lambda_r},\end{aligned}$$

which holds with probability at least $1 - O(n^{-20})$, where we have used the fact that by a simple ε-net argument with this same probability $\|\mathbf{U}^\top \mathbf{N}\mathbf{U}\| \lesssim \sigma\big(\sqrt{r} + \sqrt{\log(n)}\big)$, as well as the bounds on $\|\mathbf{N}\|$ and $\|\sin\Theta\|$ that we have used repeatedly. Plugging this bound into (A.9), we obtain

$$\begin{aligned}\|\mathbf{N}\mathbf{U}\big[\Lambda^{-1}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}\big]\mathbf{W}_*^\top\|_{2,\infty} &\lesssim \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r^2}\left(\sigma\big(\sqrt{r} + \sqrt{\log(n)}\big) + \sigma\sqrt{n}\frac{\sigma\sqrt{n}}{\lambda_r}\right) \\ &\asymp \frac{\sigma^2\mu_0\sqrt{r\log(n)}}{\lambda_r^2}\left(\sqrt{r} + \sqrt{\log(n)} + \sqrt{n}\frac{\sigma\sqrt{n}}{\lambda_r}\right) \\ &\lesssim \frac{\sigma^2\mu_0\sqrt{rn\log(n)}}{\lambda_r^2},\end{aligned}$$

since $r \le n$ and $\lambda_r \gg \sigma\sqrt{n}$.

As for the other term, arguing similarly as in the proof of Lemma 17, we have that

$$\begin{aligned}\|\mathbf{N}\mathbf{U}\Lambda^{-1}\big(\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}\big)\|_{2,\infty} &\le \frac{\|\mathbf{N}\mathbf{U}\|_{2,\infty}}{\lambda_r}\|\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}\| \\ &\lesssim \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r}\|\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}\| \\ &\lesssim \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r}\|\sin\Theta(\mathbf{U},\widehat{\mathbf{U}})\|^2 \\ &\lesssim \frac{\sigma\mu_0\sqrt{r\log(n)}}{\lambda_r}\frac{\sigma^2 n}{\lambda_r^2}.\end{aligned}$$

which holds with probability at least $1 - O(n^{-20})$. This completes the proof.

□

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B

Proofs from Chapter 2

B.1 Proof of Theorem 7

First, note that it holds that

$$\begin{aligned}\|\widetilde{U}\widehat{H} - U\widetilde{H}\widehat{H}\|_{2,\infty} &\leq \|\widetilde{U} - U\widetilde{H}\|_{2,\infty} \\ &\leq \|\widetilde{U}\widetilde{U}^\top - UU^\top\widetilde{U}\widetilde{U}^\top\|_{2,\infty} \\ &\leq \|\widetilde{U}\widetilde{U}^\top - UU^\top\|_{2.\infty}\|\widetilde{U}\widetilde{U}^\top\| \\ &\leq \|\widetilde{U}\widetilde{U}^\top - UU^\top\|_{2,\infty},\end{aligned}$$

which is a difference of projection matrices. We now wish to expand $\widetilde{U}\widetilde{U}^\top$ in terms of the noise matrix $\Gamma(EM^\top + ME^\top + EE^\top)$.

In what follows, define, for some sufficiently large constant C_0

$$\begin{aligned}\delta_1 &:= \sqrt{rnd}\log(\max(n,d))\sigma^2; \\ \delta_2 &:= \sqrt{rn\log(n \vee d)}\lambda_1\sigma; \\ \delta &:= C_0(\delta_1 + \delta_2).\end{aligned}$$

The two terms δ_1 and δ_2 appear frequently in our bounds, so this notation simplifies the statements of several of our results. With this new notation, to prove Theorem 7 it is

sufficient to show that

$$\|\widetilde{U}\widetilde{U}^\top - UU^\top\|_{2,\infty} \leq C_R \frac{\delta}{\lambda_r^2}\|U\|_{2,\infty}.$$

It is slightly more mathematically convenient to study the perturbation $EM^\top + ME^\top + \Gamma(EE^\top)$, so we introduce the matrix $\widetilde{U}_D\widetilde{U}_D^\top$ which is the projection onto the leading eigenspace of the matrix $MM^\top + ME^\top + EM^\top + \Gamma(EE^\top)$. We have the following spectral norm guarantee that shows that $\widetilde{U}_D\widetilde{U}_D^\top$ and $\widetilde{U}\widetilde{U}^\top$ are exceedingly close.

Lemma 13. *With probability at least* $1 - 2(n \vee d)^{-5}$,

$$\|\widetilde{U}_D\widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\| \leq \frac{C\delta_1}{\lambda_r^2}\|U\|_{2,\infty}$$

By Lemma 1 and Lemma 13, we have that with probability at least $1 - c(n \vee d)^{-5}$ that

$$\|EM^\top + ME^\top + \Gamma(EE^\top)\| \leq \frac{\delta}{\sqrt{r\log(n \vee d)}}$$

provided C_0 is sufficiently large.

In order to analyze the approximation of $\widetilde{U}_D\widetilde{U}_D^\top$ to $UU^\top$, we will use the projection matrix expansion in Xia (2019), restated slightly for our purposes here.

Lemma 14 (Theorem 1 from Xia (2021))**.** *Let U be the eigenvectors of A corresponding to its nonzero eigenvalues and let $\widetilde{U}_D$ be the eigenvectors of $A + W$, where $\|W\| \leq \frac{\lambda_r^2}{2}$. Then $\widetilde{U}_D$ admits the series expansion*

$$\widetilde{U}_D\widetilde{U}_D^\top = UU^\top + \sum_{p\geq 1} S_{A,k}(W),$$

where $S_{A,p}$ is defined according to Xia (2021) via

$$S_{A,p}(W) = \sum_{\mathbf{s}:s_1+\cdots+s_{p+1}=p} (-1)^{1+\tau(\mathbf{s})}\mathcal{P}^{-s_1}W\mathcal{P}^{-s_2}W\cdots W\mathcal{P}^{-s_{p+1}},$$

where $\mathcal{P}^{-p} = U\Lambda^{-2p}U^\top$, and $\mathcal{P}^0 = U_\perp U_\perp^\top$.

By Assumption 2.2, we have that $\lambda_r^2 \geq 2\|EM^\top + ME^\top + \Gamma(EE^\top)\|$, and hence the

assumptions to apply the expansion in Lemma 14 hold. Let $W := ME^\top + EM^\top + \Gamma(EE^\top)$. We note that we therefore have

$$\widetilde{U}_D\widetilde{U}_D^\top - UU^\top = \sum_{p\geq 1} S_{MM^\top,p}(W),$$

Note that if $s_1 \geq 1$, $\|U\Lambda^{-2s_1}U^\top W \cdots U\Lambda^{-2s_{p+1}}U^\top\|_{2,\infty} \leq \|U\|_{2,\infty}\lambda_r^{-2p}\|W\|^p$ which is bounded above by $\delta^p\lambda_r^{-2p}\|U\|_{2,\infty}$ by Lemma 1. Hence, it suffices to bound terms of the form

$$\|(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^p WU\|_{2,\infty}.$$

We have the following lemma characterizing terms of this form, whose proof is in Appendix B.5. This proof requires additional considerations about dependence which requires specially crafted "leave-one-out" terms that have hitherto not been considered in the literature on entrywise eigenvector analysis.

Lemma 15. *Let $W = EM^\top + ME^\top + \Gamma(EE^\top)$. There exists universal constants C_1 and C_2 such that for any $p \geq 1$, we have that with probability at least $1-(p+1)(n\vee d)^{-5}$ for all $1 \leq p_0 \leq p$ that*

$$\|(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p_0-1} WU\|_{2,\infty} \leq C_1(C_2\delta)^{p_0}\|U\|_{2,\infty},$$

Let c be some number to be chosen later. There are at most 4^p terms such that $s_1 + \cdots + s_{p+1} = p$, and hence

$$\begin{aligned}
\|\sum_{p\geq 1} S_{A,k}(W)\|_{2,\infty} &\leq \sum_{p=1}^{c\log(n\vee d)} \|S_{A,p}(W)\|_{2,\infty} + \sum_{p=c\log(n\vee d)}^{\infty} \|S_{A,p}(W)\|_{2,\infty} \\
&\leq \|U\|_{2,\infty} \sum_{p=1}^{c\log(n\vee d)} C_1\left(\frac{4C_2\delta}{\lambda_r^2}\right)^p + \sum_{p=c\log(n\vee d)}^{\infty}\left(\frac{4\|W\|}{\lambda_r^2}\right)^p \\
&\leq C\|U\|_{2,\infty}\frac{\delta}{\lambda_r^2} + \frac{1}{2^{c\log(n\vee d)}} \\
&\leq C_R\|U\|_{2,\infty}\frac{\delta}{\lambda_r^2},
\end{aligned}$$

where in the final line we have used the fact that the second term can be bounded by $2^{-c\log(n\vee d)} \leq \|U\|_{2,\infty}$ for c taken to be sufficiently large. Finally, this bound holds with probability at least $1 - c\log(n\vee d)(n\vee d)^{-5} - 4(n\vee d)^{-6} \geq 1 - (n\vee d)^{-4}$, which completes the proof of Theorem 7.

B.2 Proof of Theorem 8

First, we have the following result on the eigengap.

Lemma 16. *Suppose Assumption 2.2 holds, and suppose $T \geq T_0$, where T_0 is the first iterate such that $\|N_T - A\| \leq 3\|\Gamma(Z)\|$. Then on the event in Lemma 1, we have*

$$(\widehat{\lambda}_r^{(T)})^2 - \widetilde{\lambda}_{r+1}^2 \geq \lambda_r^2/2.$$

In particular,

$$\widehat{\lambda}_r^2 - \widetilde{\lambda}_{r+1}^2 \geq \lambda_r^2/2.$$

Proof. $\widehat{\lambda}_r^2 \geq \widetilde{\lambda}_r^2 - \|N_T - \widetilde{A}\| \geq \widetilde{\lambda}_r^2 - 4\|\Gamma(Z)\|$ once $T \geq T_0$. Then $\widehat{\lambda}_r^2 - \widetilde{\lambda}_{r+1}^2 \geq \widetilde{\lambda}_r^2 - \widetilde{\lambda}_{r+1}^2 - 4\|\Gamma(Z)\| \geq \lambda_r^2 - \lambda_{r+1}^2 - 6\|\Gamma(Z)\| = \lambda_r^2 - 6\|\Gamma(Z)\|$. On the event in Lemma 1, $\|\Gamma(Z)\| \leq C_{\text{spectral}}(\sigma^2(n+\sqrt{nd}) + \sigma\sqrt{n}\kappa\lambda_r)$, and under the signal-to-noise ratio condition of Assumption 2.2,

$$\lambda_r^2 \geq 12C_{\text{spectral}}\left(\sigma^2(n+\sqrt{nd}) + \sigma\sqrt{n}\kappa\lambda_r\right),$$

this gives

$$\|\Gamma(Z)\| \leq \frac{1}{12}\lambda_r^2,$$

meaning that

$$\widehat{\lambda}_r^2 - \widetilde{\lambda}_{r+1}^2 \geq \lambda_r^2/2.$$

□

We can now prove Theorem 8.

Proof of Theorem 8. We write:

$$\begin{aligned}\widehat{U} - \widetilde{U}\widetilde{U}^\top\widehat{U} &= P_{\widetilde{U}}\widetilde{A}[\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1} + [N_T - P_{\widetilde{U}}\widetilde{A}][\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1} \\ &\quad + [N_T - P_{\widetilde{U}}\widetilde{A}]\widetilde{U}\widetilde{H}\widehat{\Lambda}^{-1} - \widetilde{U}[\widetilde{H}\widehat{\Lambda} - \widetilde{\Lambda}\widetilde{H}]\widehat{\Lambda}^{-1} \\ &:= J_1 + J_2 + J_3 + J_4.\end{aligned}$$

We bound each term successively.

The term J_1: Note that $\|\widehat{U} - \widetilde{U}\widetilde{H}\| \leq \sqrt{2}\|\sin\Theta(\widehat{U}, \widetilde{U})\|$. Therefore, we have that

$$\begin{aligned}\|P_{\widetilde{U}}\widetilde{A}[\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1}\|_{2,\infty} &\leq \|P_{\widetilde{U}}\widetilde{A}\|_{2,\infty}\|\widehat{U} - \widetilde{U}\widetilde{H}\|\|\widehat{\Lambda}^{-1}\| \\ &\leq \|\widetilde{U}\widetilde{\Lambda}\widetilde{U}\|_{2,\infty}\|\widehat{U} - \widetilde{U}\widetilde{H}\|\widehat{\lambda}_r^{-2} \\ &\leq \frac{\widetilde{\lambda}_1^2}{\widehat{\lambda}_r^2}\|\widetilde{U}\|_{2,\infty}\|\widehat{U} - \widetilde{U}\widetilde{H}\| \\ &\leq \sqrt{2}\frac{\widetilde{\lambda}_1^2}{\widehat{\lambda}_r^2}\|\widetilde{U}\|_{2,\infty}\|\sin\Theta(\widehat{U}, \widetilde{U})\|. \qquad \text{(B.1)}\end{aligned}$$

The term J_2: We decompose via

$$\begin{aligned}[N_T - P_{\widetilde{U}}\widetilde{A}][\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1} &= \widehat{U}[\widehat{\Lambda}\widehat{U}^\top - \widetilde{H}^\top\widetilde{\Lambda}\widetilde{U}^\top][\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1} + [\widehat{U} - \widetilde{U}\widetilde{H}]\widetilde{H}^\top\widetilde{\Lambda}\widetilde{U}^\top[\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1} \\ &\quad - \widetilde{U}[I - \widetilde{H}\widetilde{H}^\top]\widetilde{\Lambda}\widetilde{U}^\top[\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1}.\end{aligned}$$

Note that $\widetilde{U}^\top[\widehat{U} - \widetilde{U}\widetilde{H}] = 0$, by the definition of $\widetilde{H}$, whence we have

$$\begin{aligned}[N_T - P_{\widetilde{U}}\widetilde{A}][\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1} &= \widetilde{U}\widetilde{H}[\widehat{\Lambda}\widehat{U}^\top - \widetilde{H}^\top\widetilde{\Lambda}\widetilde{U}^\top][\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1} + [\widehat{U} - \widetilde{U}\widetilde{H}][\widehat{\Lambda}\widehat{U}^\top - \widetilde{H}^\top\widetilde{\Lambda}\widetilde{U}^\top][\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1} \\ &= \widetilde{U}\widetilde{H}\widehat{\Lambda}[\widehat{U} - \widetilde{U}\widetilde{H}]^\top[\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1} + [\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}[I - \widetilde{H}^\top\widetilde{H}]\widehat{\Lambda}^{-1}.\end{aligned}$$

Taking norms, we have

$$\|[N_T - P_{\widetilde{U}}\widetilde{A}][\widehat{U} - \widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1}\|_{2,\infty} \leq \|\widetilde{U}\|_{2,\infty}\widehat{\lambda}_1^2\|\widehat{U} - \widetilde{U}\widetilde{H}\|^2\widehat{\lambda}_r^{-2} + \|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty}\widehat{\lambda}_1^2\|I - \widetilde{H}^\top\widetilde{H}\|\widehat{\lambda}_r^{-2}$$

Recall that $\|I-\widetilde{H}^\top\widehat{H}\| \le \|\widehat{U}\widehat{U}^\top-\widetilde{U}\widetilde{U}^\top\| \le 2\|\sin\Theta(\widehat{U},\widetilde{U})\|$, and $\|\widehat{U}-\widetilde{U}\widehat{H}\| \le \sqrt{2}\|\sin\Theta(\widehat{U},\widetilde{U})\|$. Consequently,

$$\|J_2\|_{2,\infty} \le 2\|\widetilde{U}\|_{2,\infty}\frac{\widehat{\lambda}_1^2}{\widehat{\lambda}_r^2}\|\sin\Theta(\widehat{U},\widetilde{U})\|^2 + 2\|\widehat{U}-\widetilde{U}\widetilde{H}\|_{2,\infty}\frac{\widehat{\lambda}_1^2}{\widehat{\lambda}_r^2}\|\sin\Theta(\widehat{U},\widetilde{U}\|. \tag{B.2}$$

The term J_3: Again using the fact that $\widetilde{U}^\top[\widehat{U}-\widetilde{U}\widetilde{H}]=0$,

$$\begin{aligned}[N_T - P_{\widetilde{U}}\widetilde{A}][\widehat{U}-\widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1} &= \widehat{U}\widehat{\Lambda}[I-\widetilde{H}^\top\widetilde{H}]\widehat{\Lambda}^{-1} \\ &= \widetilde{U}\widetilde{H}\widehat{\Lambda}[I-\widetilde{H}^\top\widetilde{H}]\widehat{\Lambda}^{-1} + [\widehat{U}-\widetilde{U}\widetilde{H}]\widehat{\Lambda}[I-\widetilde{H}^\top\widetilde{H}]\widehat{\Lambda}^{-1}.\end{aligned}$$

This gives

$$\begin{aligned}\|[N_T - P_{\widetilde{U}}\widetilde{A}][\widehat{U}-\widetilde{U}\widetilde{H}]\widehat{\Lambda}^{-1}\|_{2,\infty} &\le \|\widetilde{U}\|_{2,\infty}\widehat{\lambda}_1^2\|I-\widetilde{H}^\top\widetilde{H}\|\widehat{\lambda}_r^{-2} + \|\widehat{U}-\widetilde{U}\widetilde{H}\|_{2,\infty}\widehat{\lambda}_1^2\|I-\widetilde{H}^\top\widetilde{H}\|\widehat{\lambda}_r^{-2} \\ &\le 2(\|\widetilde{U}\|_{2,\infty} + \|\widehat{U}-\widetilde{U}\widetilde{H}\|_{2,\infty})\frac{\widehat{\lambda}_1^2}{\widehat{\lambda}_r^2}\|\sin\Theta(\widehat{U},\widetilde{U})\|. \end{aligned}\tag{B.3}$$

The term J_4: In $2,\infty$ norm, we note that

$$\|\widetilde{U}[\widetilde{H}\widehat{\Lambda}-\widetilde{\Lambda}\widetilde{H}]\widehat{\Lambda}^{-1}\|_{2,\infty} \le \frac{\|\widetilde{U}\|_{2,\infty}}{\widehat{\lambda}_r^2}\|\widetilde{H}\widehat{\Lambda}-\widetilde{\Lambda}\widetilde{H}\|.$$

Furthermore,

$$\begin{aligned}\|\widetilde{H}\widehat{\Lambda}-\widetilde{\Lambda}\widetilde{H}\| &= \|\widetilde{U}^\top[N_T-\widetilde{A}]\widehat{U}\| \\ &\le \|\widetilde{U}^\top[N_T-\widetilde{A}]\|.\end{aligned}$$

Consequently,

$$\|J_4\| \le \|\widetilde{U}\|_{2,\infty}\frac{1}{\widehat{\lambda}_r^2}\|\widetilde{U}^\top[N_T-\widetilde{A}]\|. \tag{B.4}$$

Putting it together: Collecting the bounds in (B.1),(B.2),(B.3), and (B.4), we see that

$$\|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty} \leq \frac{\widetilde{\lambda}_1^2}{\widehat{\lambda}_r^2}\|\widetilde{U}\|_{2,\infty}\left[\sqrt{2}\|\sin\Theta(\widehat{U},\widetilde{U})\| + \frac{\widehat{\lambda}_1^2}{\widetilde{\lambda}_1^2}(2\|\sin\Theta(\widehat{U},\widetilde{U})\|^2 + 2\|\sin(\widehat{U},\widetilde{U})\|) + \frac{\widetilde{K}_T}{\widetilde{\lambda}_1^2}\right] + \|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty}\frac{4\widehat{\lambda}_1^2}{\widehat{\lambda}_r^2}\|\sin\Theta(\widehat{U},\widetilde{U})\|,$$

where $\widetilde{K}_T := \|N_T - \widetilde{A}\|$. Applying Davis-Kahan, Lemma 2, and Lemma 16, we see that

$$\begin{aligned}\|\sin\Theta(\widehat{U},\widetilde{U})\| &\leq \frac{\|\widehat{A} - \widetilde{A}\|}{\widehat{\lambda}_r^2 - \widetilde{\lambda}_{r+1}^2} \\ &\leq \frac{41\|U\|_{2,\infty}\|\Gamma(Z)\|}{\lambda_r^2/2} =: \tau.\end{aligned}$$

By Theorem 7, for large enough n, $\|\widetilde{U}\|_{2,\infty} \leq \|U\|_{2,\infty}$ since the bound in Theorem 7 is of the form $c\|U\|_{2,\infty}$, where $c < 1$ for n sufficiently large. This gives

$$\|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty} \leq 2\frac{\widetilde{\lambda}_1^2}{\widehat{\lambda}_r^2}\|U\|_{2,\infty}\left[\sqrt{2}\tau + \frac{\widehat{\lambda}_1^2}{\widetilde{\lambda}_1^2}\left(2\tau^2 + 2\tau\right) + \frac{\tau\lambda_r^2/2}{\widetilde{\lambda}_1^2}\right] + \|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty}\frac{4\widehat{\lambda}_1^2}{\widehat{\lambda}_r^2}\tau.$$

The proof of Lemma 16 reveals that $(11/12)\lambda_1^2 \leq \widetilde{\lambda}_1^2 \leq (13/12)\lambda_1^2$, $(3/4)\lambda_r^2 \leq \widehat{\lambda}_r^2 \leq (5/4)\lambda_r^2$, with the same bounds holding for $\widehat{\lambda}_1$, also. Thus $\widetilde{\lambda}_1^2/\widehat{\lambda}_r^2 \leq (13/9)\kappa^2$, $\widehat{\lambda}_1^2/\widetilde{\lambda}_1^2 \leq 15/11$, and $\widehat{\lambda}_1^2/\widehat{\lambda}_r^2 \leq (5/3)\kappa^2$. This gives

$$\|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty} \leq 2(13/9)\kappa^2\|U\|_{2,\infty}\left[\sqrt{2}\tau + (15/11)\left(2\tau^2 + 2\tau\right) + \frac{\tau\lambda_r^2/2}{(11/12)\lambda_1^2}\right] + \|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty}4(5/3)\kappa^2\tau.$$

When $(20/3)\kappa^2\tau \leq 1/2$, which occurs for n sufficiently large under the event in Theorem 7, this gives

$$\begin{aligned}\|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty} &\leq \kappa^2\frac{500\|U\|_{2,\infty}^2\|\Gamma(Z)\|}{\lambda_r^2}\left[\sqrt{2} + (15/11)\left(2 + \frac{3}{20\kappa^2}\right) + \frac{6}{11\kappa^2}\right] \\ &= C_D\kappa^2\frac{\|U\|_{2,\infty}^2\|\Gamma(Z)\|}{\lambda_r^2}\end{aligned}$$

as required. □

B.3 Proof of Theorem 5

First, we justify Equation (2.3). Note that by Theorem 6, on that event $\|\widehat{U}\|_{2,\infty} \leq C\|U\|_{2,\infty}$, which implies that $\widehat{U}$ is as incoherent as U up to constant factors. Now following similarly in the first part of the proof Theorem 6, we have that for the same orthogonal matrix $\mathcal{O}_*$ as in Lemma 3 that

$$e_i^\top \left(\widehat{U}\mathcal{O}_* - U\right) e_j = e_i^\top \widehat{U}(\mathcal{O}_* - \widehat{U}^\top \widetilde{U}\widetilde{U}^\top U)e_j + e_i^\top \left(\widehat{U}\widehat{U}^\top \widetilde{U}\widetilde{U}^\top U - \widetilde{U}\widetilde{U}^\top U\right) e_j + e_i^\top \left(\widetilde{U}\widetilde{U}^\top U - U\right) e_j. \tag{B.5}$$

As in the proof of Theorem 7 (see Appendix B.5), let $\widetilde{U}_D$ be the matrix of eigenvectors of $MM^\top + EM^\top + ME^\top + \Gamma(EE^\top)$, and let $W := EM^\top + ME^\top + \Gamma(EE^\top)$.

Now we again apply Lemma 14 to $\widetilde{U}_D\widetilde{U}_D^\top$. First, recall the definition of $S_{MM^\top,1}(W) = U_\perp U_\perp^\top WU\Lambda^{-2}U^\top + U^\top\Lambda^{-2}U^\top WU_\perp U_\perp^\top$, and note that $S_{MM^\top,1}(W)U = U_\perp U_\perp^\top WU\Lambda^{-2}$. Now, just as in the proof of Theorem 7 we expand $\widetilde{U}_D\widetilde{U}_D^\top$ as an infinite series in W via

$$\begin{aligned}
e_i^\top \left(\widetilde{U}\widetilde{U}^\top U - U\right) e_j &= e_i^\top \left(\widetilde{U}_D\widetilde{U}_D^\top U - U\right) e_j - e_i^\top \left(\widetilde{U}_D\widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\right) Ue_j \\
&= e_i^\top S_{MM^\top,1}(W)Ue_j + e_i^\top \sum_{k\geq 2} S_{MM^\top,k}(W)Ue_j - e_i^\top \left(\widetilde{U}_D\widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\right) Ue_j \\
&= e_i^\top U_\perp U_\perp^\top \left(EM^\top + ME^\top + \Gamma(EE^\top)\right)\Lambda^{-2}e_j + e_i^\top \sum_{k\geq 2} S_{MM^\top,k}(W)Ue_j \\
&\quad - e_i^\top \left(\widetilde{U}_D\widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\right) Ue_j \\
&= e_i^\top \left(EM^\top + \Gamma(EE^\top)\right)U\Lambda^{-2}e_j - e_i^\top UU^\top \left(EM^\top + \Gamma(EE^\top)\right)U\Lambda^{-2}e_j \\
&\quad + e_i^\top \sum_{k\geq 2} S_{MM^\top,k}(W)Ue_j - e_i^\top \left(\widetilde{U}_D\widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\right) Ue_j
\end{aligned} \tag{B.6}$$

where in the penultimate line we used the fact that $U_\perp U_\perp^\top M = 0$. Hence, plugging the

expansion in (B.6) into (B.5), we see that

$$\begin{aligned} e_i^\top \left(\widehat{U}\mathcal{O}_* - U\right) e_j &= e_i^\top EM^\top U\Lambda^{-2}e_j + e_i^\top \Gamma(EE^\top)U\Lambda^{-2}e_j - e_i^\top UU^\top \left(EM^\top + \Gamma(EE^\top)\right)U\Lambda^{-2}e_j \\ &\quad + e_i^\top \sum_{k\geq 2} S_{MM^\top,k}(W)Ue_j - e_i^\top \left(\widetilde{U}_D\widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\right)Ue_j \\ &\quad + e_i^\top \widehat{U}(\mathcal{O}_* - \widehat{U}^\top\widetilde{U}\widetilde{U}^\top U)e_j + e_i^\top \left(\widehat{U}\widehat{U}^\top\widetilde{U}\widetilde{U}^\top U - \widetilde{U}\widetilde{U}^\top U\right)e_j \\ &:= e_i^\top EM^\top U\Lambda^{-2}e_j + R_0 + R_1 + R_2 + R_3 + R_4 + R_5, \end{aligned}$$

where

$$\begin{aligned} R_0 &:= e_i^\top \Gamma(EE^\top)U\Lambda^{-2}e_j; \\ R_1 &:= e_i^\top U\left(EM^\top + \Gamma(EE^\top)\right)U\Lambda^{-2}e_j; \\ R_2 &:= e_i^\top \left(\widetilde{U}_D\widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\right)Ue_j; \\ R_3 &:= e_i^\top \sum_{k\geq 2} S_{MM^\top,k}(W)Ue_j; \\ R_4 &:= e_i^\top \widehat{U}(\mathcal{O}_* - \widehat{U}^\top\widetilde{U}\widetilde{U}^\top U)e_j; \\ R_5 &:= e_i^\top \left(\widehat{U}\widehat{U}^\top\widetilde{U}\widetilde{U}^\top U - \widetilde{U}\widetilde{U}^\top U\right)e_j. \end{aligned}$$

We now characterize the residual terms.

Lemma 17. *There exist universal constants C_6 and C_7 such that the residual terms R_1,R_2, and R_3 satisfy, uniformly over i and j,*

$$\frac{1}{\sigma_{ij}}\left(|R_1| + |R_2| + |R_3|\right) \leq C_6\kappa_\sigma\kappa^2\mu_0\sqrt{\frac{r\log(n\vee d)}{n}} + C_7\kappa^3\kappa_\sigma\mu_0\frac{r\log(n\vee d)}{\mathrm{SNR}}$$

with probability at least $1 - 5(n\vee d)^{-4}$.

Lemma 18. *On the intersection of the events in Theorem 6 and Lemma 1 the residual terms R_4 and R_5 satisfy for all i and j,*

$$\frac{1}{\sigma_{ij}}\left(|R_4| + |R_5|\right) \leq C_8\kappa^3\kappa_\sigma\mu_0\frac{1}{\mathrm{SNR}} + C_9\kappa^4\kappa_\sigma\mu_0^2\frac{r}{\sqrt{n}}.$$

for some universal constants C_8 and C_9.

To bound R_0, we note that we can equivalently write

$$R_0 := \sum_{k \neq i} \langle E_i, E_k \rangle (U\Lambda^{-2})_{kj} = \sum_{k \neq i} \langle E_i, E_k \rangle U_{kj} \lambda_j^{-2}.$$

We have the following bound for R_0.

Lemma 19. *There exists a universal constant C_{10} such that with probability at least $1 - 4(n \vee d)^{-4}$*

$$\frac{1}{\sigma_{ij}} \left| \sum_{k \neq i} \langle E_i, E_k \rangle U_{kj} \lambda_j^{-2} \right| \leq C_{10} \mu_0 \kappa_\sigma \frac{\log(n \vee d)}{\mathrm{SNR}},$$

where the probability is uniform over i and j.

Let $R := R_0 + R_1 + R_2 + R_3 + R_4 + R_5$. This argument leaves us with

$$e_i^\top (\widehat{U}\mathcal{O}_* - U) e_j = e_i^\top E M^\top U \Lambda^{-2} e_j + R.$$

Note in addition that $M^\top U \Lambda^{-2} = V \Lambda U^\top U \Lambda^{-2} = V \Lambda^{-1}$ by definition. Hence, the term $e_i^\top E M^\top U \Lambda^{-2} e_j$ can be equivalently written as $\langle E_i, V_{\cdot j} \rangle \lambda_j^{-1}$ where $V_{\cdot j}$ is the j'th column of the matrix V, which justifies equation (2.3).

Now, combining the bounds for the residuals in Lemmas 17, 18, and 19, we see that with probability at least $1 - 10(n \vee d)^{-4} - 4(n \vee d)^{-6} \geq 1 - (n \vee d)^{-3}$,

$$\begin{aligned}
\frac{|R|}{\sigma_{ij}} &\leq C_6 \kappa_\sigma \kappa^2 \mu_0 \sqrt{\frac{r \log(n \vee d)}{n}} + C_7 \kappa^3 \kappa_\sigma \mu_0 \frac{r \log(n \vee d)}{\mathrm{SNR}} + C_8 \kappa^3 \kappa_\sigma \mu_0 \frac{1}{\mathrm{SNR}} \\
&\quad + C_9 \kappa^4 \kappa_\sigma \mu_0^2 \frac{r}{\sqrt{n}} + C_{10} \mu_0 \kappa_\sigma \frac{\log(n \vee d)}{\mathrm{SNR}} \\
&\leq \widetilde{C}_1 \kappa^3 \kappa_\sigma \mu_0 \frac{r \log(n \vee d)}{\mathrm{SNR}} + \widetilde{C}_2 \kappa^2 \kappa_\sigma \mu_0 \sqrt{\frac{r}{n}} \left(\sqrt{\log(n \vee d)} + \mu_0 \kappa^2 \sqrt{r} \right) \\
&=: B.
\end{aligned}$$

We can now complete the proof of Theorem 5. By the classical Berry-Esseen Theorem

(Berry, 1941), for any $x \in \mathbb{R}$, denoting $Y_{i\alpha}$ as the α entry of Y_i,

$$\left|\mathbb{P}\left(\frac{\langle E_i, V_{\cdot j}\rangle}{\|\Sigma_i^{1/2}V_{\cdot j}\|} > x\right) - \Phi(x)\right| \leq C\frac{\sum_\alpha |(\Sigma_i^{1/2}V_{\cdot j})_\alpha|^3 \mathbb{E}|Y_{i\alpha}|^3}{\|\Sigma_i^{1/2}V_{\cdot j}\|^3} \leq C\frac{\|\Sigma_i^{1/2}V_{\cdot j}\|_3^3}{\|\Sigma_i^{1/2}V_{\cdot j}\|^3}.$$

Hence, by the Lipchitz property of Φ, (e.g. Xia (2021))

$$\begin{aligned}\mathbb{P}\left(\frac{1}{\sigma_{ij}}e_i^\top\left(\widehat{U}\mathcal{O}_* - U\right)e_j \leq x\right) &\leq \mathbb{P}\left\{\frac{\langle E_i, V_{\cdot j}\rangle}{\|\Sigma_i^{1/2}V_{\cdot j}\|} \leq x + B\right\} + (n\vee d)^{-3} \\ &\leq \Phi(x+B) + C\frac{\|\Sigma_i^{1/2}V_{\cdot j}\|_3^3}{\|\Sigma_i^{1/2}V_{\cdot j}\|^3} + (n\vee d)^{-3} \\ &\leq \Phi(x) + C\frac{\|\Sigma_i^{1/2}V_{\cdot j}\|_3^3}{\|\Sigma_i^{1/2}V_{\cdot j}\|^3} + B + (n\vee d)^{-3}.\end{aligned}$$

A similar bound for the left tail also holds. Therefore, after relabeling constants and noting that $\kappa^2\kappa_\sigma\sqrt{r/n} \geq (n\vee d)^{-3}$, we conclude that

$$\begin{aligned}\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{1}{\sigma_{ij}}e_i^\top\left(\widehat{U}\mathcal{O}_* - U\right)e_j \leq x\right) - \Phi(x)\right| &\leq C\frac{\|\Sigma_i^{1/2}V_{\cdot j}\|_3^3}{\|\Sigma_i^{1/2}V_{\cdot j}\|^3} + B + (n\vee d)^{-3} \\ &\leq C_1\frac{\|\Sigma_i^{1/2}V_{\cdot j}\|_3^3}{\|\Sigma_i^{1/2}V_{\cdot j}\|^3} + C_2\kappa^3\kappa_\sigma\mu_0\frac{r\log(n\vee d)}{\mathrm{SNR}} \\ &\quad + C_3\kappa^2\kappa_\sigma\mu_0\sqrt{\frac{r}{n}}\left(\sqrt{\log(n\vee d)} + \mu_0\kappa^2\sqrt{r}\right).\end{aligned}$$

B.4 Proof of Corollaries in Section 2.3.2

Proof of Corollary 2. By the proof of Theorem 5, we have that

$$e_i^\top\left(\widehat{U}\mathcal{O}_* - U\right)e_j = \langle E_i, V_{\cdot j}\rangle\lambda_j^{-1} + R_{ij},$$

Hence. the i'th row of $\widehat{U}$ satisfies

$$e_i^\top\left(\widehat{U}\mathcal{O}_* - U\right) = E_i^\top V\Lambda^{-1} + R_i.$$

We now analyze $(S_i)^{-1/2}R_i$. However, by Lemmas 17, 18, and 19, we see that R_{ij} satisfies with probability at least $1-(n\vee d)^{-3}$

$$|R|_{ij} \leq C_2\sigma_{ij}\kappa^3\kappa_\sigma\mu_0\frac{r\log(n\vee d)}{\text{SNR}} + C_3\sigma_{ij}\kappa^2\kappa_\sigma\mu_0\sqrt{\frac{r}{n}}\left(\sqrt{\log(n\vee d)}+\mu_0\kappa^2\sqrt{r}\right).$$

In addition, note that

$$\|S_i^{-1/2}\|\sigma_{ij} \leq \kappa\kappa_\sigma.$$

Therefore,

$$S_i^{-1/2}R_i \to 0$$

in probability (and almost surely) as n and d tend to infinity, since $\|S_i^{-1/2}\|\sigma_{ij} = O(1)$ when κ and κ_σ are bounded. Furthermore, we note that

$$\mathbb{E}\left(\langle E_i, V_{\cdot j}\rangle\langle E_i, V_{\cdot k}\rangle\lambda_j^{-1}\lambda_k^{-1}\right) = (S_i)_{jk}.$$

Hence, the result holds by the Cramer-Wold device and Slutsky's Theorem. □

Proof of Corollary 3. Without loss of generality assume that $C_k = \{1,\cdots,n_k\}$, or else re-order the matrix. Furthermore, we assume that the set of indices for community k is known; under the assumptions for Theorem 6 this will be true for sufficiently large n, d since $\|\widehat{U}-U\mathcal{O}_*\| \ll \|U\|_{2,\infty}$ and each row of U reveals the community memberships by Lemma 2.1 of Lei and Rinaldo (2015).

In what follows, recall that $\Lambda^{-1}V^\top E_i$ is an r-dimensional column vector and $E_i^\top V\Lambda^{-1}$ is a row vector. For convenience, we let $\bar{U}^{(k)}$ denote the rank one matrix whose rows are all

just $\bar{U}^{(k)}$. By the expansion in the proof Theorem 5, we have that

$$\begin{aligned}\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\widehat{U}-\bar{U}^{(k)}\right)_{i\cdot}\left(\widehat{U}-\bar{U}^{(k)}\right)_{i\cdot}^{\top} &= \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\widehat{U}-U\mathcal{O}_*^{\top}\right)_{i\cdot}\left(\widehat{U}-U\mathcal{O}_*^{\top}\right)_{i\cdot}^{\top} \\ &+ \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\widehat{U}-U\mathcal{O}_*^{\top}\right)_{i\cdot}\left(U\mathcal{O}_*^{\top}-\bar{U}^{(k)}\right)_{i\cdot}^{\top} \\ &+ \frac{1}{n_k}\sum_{i=1}^{n_k}\left(U\mathcal{O}_*^{\top}-\bar{U}^{(k)}\right)_{i\cdot}\left(\widehat{U}-U\mathcal{O}_*^{\top}\right)_{i\cdot}^{\top} \\ &+ \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\bar{U}^{(k)}-U\mathcal{O}_*^{\top}\right)_{i\cdot}\left(\bar{U}^{(k)}-U\mathcal{O}_*^{\top}\right)_{i\cdot}^{\top} \\ &:= J_1+J_2+J_3\end{aligned}$$

where

$$\begin{aligned}J_1 &:= \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\widehat{U}-U\mathcal{O}_*^{\top}\right)_{i\cdot}\left(\widehat{U}-U\mathcal{O}_*^{\top}\right)_{i\cdot}^{\top}; \\ J_2 &:= \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\bar{U}^{(k)}-U\mathcal{O}_*^{\top}\right)_{i\cdot}\left(\bar{U}^{(k)}-U\mathcal{O}_*^{\top}\right)_{i\cdot}^{\top}; \\ J_3 &:= \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\widehat{U}-U\mathcal{O}_*^{\top}\right)_{i\cdot}\left(U\mathcal{O}_*^{\top}-\bar{U}^{(k)}\right)_{i\cdot}^{\top}+\frac{1}{n_k}\sum_{i=1}^{n_k}\left(U\mathcal{O}_*^{\top}-\bar{U}^{(k)}\right)_{i\cdot}\left(\widehat{U}-U\mathcal{O}_*^{\top}\right)_{i\cdot}^{\top}.\end{aligned}$$

We will show that $(S^{(k)})^{-1}J_1$ converges to I_r in probability and that the other terms tend to zero in probability.

The term J_1: Using the same expansion as in the proof for Corollary 2 we expand out

$\widehat{U} - U\mathcal{O}_*^\top$ via

$$\begin{aligned}
J_1 :&= \frac{1}{n_k}\sum_{i=1}^{n_k}\left(\widehat{U} - U\mathcal{O}_*^\top\right)_{i\cdot}\left(\widehat{U} - U\mathcal{O}_*^\top\right)_{i\cdot}^\top \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\Big(\mathcal{O}_*(\Lambda^{-1}V^\top E_i) + \mathcal{O}_* R_i\Big)\Big(\mathcal{O}_*(\Lambda^{-1}V^\top E_i) + \mathcal{O}_* R_i\Big)^\top \\
&= \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_*(\Lambda^{-1}V^\top E_i E_i^\top V\Lambda^{-1})\mathcal{O}_*^\top + \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_* R_i(E_i^\top V\Lambda^{-1})\mathcal{O}_*^\top + \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_*(\Lambda^{-1}V^\top E_i)R_i\mathcal{O}_*^\top \\
&\quad + \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_* R_i R_i^\top \mathcal{O}_*^\top \\
&:= J_{11} + J_{12} + J_{13},
\end{aligned}$$

where

$$\begin{aligned}
J_{11} &:= \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_*(\Lambda^{-1}V^\top E_i E_i^\top V\Lambda^{-1})\mathcal{O}_*^\top; \\
J_{12} &:= \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_* R_i(E_i^\top V\Lambda^{-1})\mathcal{O}_*^\top + \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_*(\Lambda^{-1}V^\top E_i)R_i\mathcal{O}_*^\top; \\
J_{13} &:= \frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_* R_i R_i^\top \mathcal{O}_*^\top.
\end{aligned}$$

Since $E_i = (\Sigma^{(k)})^{1/2}Y_i$, the term J_{11} satisfies

$$\mathcal{O}_*(S^{(k)})^{-1}\mathcal{O}_*^\top J_{11} = \mathcal{O}_*(S^{(k)})^{-1}\frac{1}{n_k}\sum_{i=1}^{n_k}\Lambda^{-1}V^\top\Sigma_i^{1/2}Y_iY_i^\top\Sigma_i^{1/2}V\Lambda^{-1}\mathcal{O}_*^\top.$$

The random variable $\Lambda^{-1}V^\top\Sigma_i^{1/2}Y_iY_i^\top\Sigma_i^{1/2}V\Lambda^{-1}$ is an $r\times r$ matrix with expectation $S^{(k)}$, and hence $\|\mathcal{O}_*(S^{(k)})^{-1}\mathcal{O}_*^\top J_1 - I_r\| \to 0$ in probability by the strong law of large numbers, the rotation invariance of the spectral norm, and the fact that each Y_i has independent subgaussian components. We now show the other terms all tend to zero in probability.

As for J_{12}, we analyze the term

$$\mathcal{O}_*(S^{(k)})^{-1}\mathcal{O}_*^\top\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_* R_i(E_i^\top V\Lambda^{-1})\mathcal{O}_*^\top = \mathcal{O}_*(S^{(k)})^{-1}\frac{1}{n_k}\sum_{i=1}^{n_k}R_i(E_i^\top V\Lambda^{-1})\mathcal{O}_*^\top.$$

The other term is similar. By the rotational invariance of the spectral norm, we may ignore

the orthogonal matrices henceforth. By the residual bounds in Lemmas 17, 18, and 19, we have that with probability at least $1 - (n \vee d)^{-3}$,

$$\max_i \|(S^{(k)})^{-1/2} R_i\| \lesssim \frac{\log(n \vee d)}{\text{SNR}} + \sqrt{\frac{\log(n \vee d)}{n}},$$

where we let the implicit constants depend on $\kappa, \mu_0, \kappa_\sigma$, and r, since they are assumed bounded in n and d. Let this event be denoted $\mathcal{E}$. Then

$$\mathbb{P}\left(\|(S^{(k)})^{-1} \frac{1}{n_k} \sum_{i=1}^{n_k} R_i (E_i^\top V \Lambda^{-1})\| > t \right) \leq \mathbb{P}\left(\|(S^{(k)})^{-1} \frac{1}{n_k} \sum_{i=1}^{n_k} R_i (E_i^\top V \Lambda^{-1})\| > t \cap \mathcal{E} \right) + (n \vee d)^{-3},$$

so it suffices to analyze this term on the event $\mathcal{E}$. Note that the vector $E_i^\top V \Lambda^{-1}$ is an $r-$dimensional random variable with covariance matrix $S^{(k)}$. Note that the condition number of $S^{(k)}$ is at most $\kappa^2 \kappa_\sigma^2$, and hence $(S^{(k)})^{1/2}$ has condition number at most $\kappa \kappa_\sigma$. Therefore,

$$\begin{aligned}
\|(S^{(k)})^{-1} \frac{1}{n_k} \sum_{i=1}^{n_k} R_i (E_i^\top V \Lambda^{-1})\| &\leq \kappa \kappa_\sigma \|(S^{(k)})^{-1/2} \frac{1}{n_k} \sum_{i=1}^{n_k} R_i (E_i^\top V \Lambda^{-1}) (S^{(k)})^{-1/2}\| \\
&\leq r \kappa \kappa_\sigma \|\frac{1}{n_k} \sum_{i=1}^{n_k} (S^{(k)})^{-1/2} R_i (E_i^\top V \Lambda^{-1}) (S^{(k)})^{-1/2}\|_{\max}.
\end{aligned}$$

Now consider the j, l entry of the above matrix, which can be written as

$$\frac{1}{n_k} \sum_{i=1}^{n_k} \left((S^{(k)})^{-1/2} R_i \right)_j \left[(E_i^\top V \Lambda^{-1}) (S^{(k)})^{-1/2} \right]_l.$$

By (the restricted) Markov's inequality,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n_k}\sum_{i=1}^{n_k}\left((S^{(k)})^{-1/2}R_i\right)_j\left[(E_i^\top V\Lambda^{-1})(S^{(k)})^{-1/2}\right]_l\right| > t \cap \mathcal{E}\right) \\
&\qquad\qquad \leq \frac{1}{t}\mathbb{E}\mathbb{I}_{\mathcal{E}}\left|\frac{1}{n_k}\sum_{i=1}^{n_k}\left((S^{(k)})^{-1/2}R_i\right)_j\left[(E_i^\top V\Lambda^{-1})(S^{(k)})^{-1/2}\right]_l\right| \\
&\qquad\qquad \leq \frac{1}{t}\max_i \mathbb{E}\mathbb{I}_{\mathcal{E}}\left|\left((S^{(k)})^{-1/2}R_i\right)_j\left[(E_i^\top V\Lambda^{-1})(S^{(k)})^{-1/2}\right]_l\right| \\
&\qquad\qquad \lesssim \frac{1}{t}\left(\frac{\log(n\vee d)}{\mathrm{SNR}} + \sqrt{\frac{\log(n\vee d)}{n}}\right)\mathbb{E}\left|\left[(E_i^\top V\Lambda^{-1})(S^{(k)})^{-1/2}\right]_l\right| \\
&\qquad\qquad \lesssim \frac{1}{t}\left(\frac{\log(n\vee d)}{\mathrm{SNR}} + \sqrt{\frac{\log(n\vee d)}{n}}\right),
\end{aligned}$$

where the final inequality is due to the fact that $E_i^\top V\Lambda^{-1}(S^{(k)})^{-1/2}$ is an isotropic r-dimensional subgaussian random variable, and hence has moments are bounded by $O(1)$. Therefore, we conclude that since the j,l entry converges to zero in probability, since r is fixed, we conclude that $(S^{(k)})^{-1}J_{12}$ converges to zero in probability.

For J_{13}, after accounting for orthogonal matrices, we note that

$$\frac{1}{n_k}\left\|(S^{(k)})^{-1}\sum_{i=1}^{n_k}R_iR_i^\top\right\| \leq \max_i \|(S^{(k)})^{-1}R_iR_i^\top\|.$$

The matrix $R_iR_i^\top$ is rank one and $S^{(k)}$ is assumed to be positive definite by Assumption 2.5. Therefore this term satisfies

$$\begin{aligned}
\max_i \|(S^{(k)})^{-1}R_iR_i^\top\| &= \max_i R_i^\top(S^{(k)})^{-1}R_i^\top \\
&= \max_i \|(S^{(k)})^{-1/2}R_i\|^2.
\end{aligned}$$

By the argument for the same term in the proof of Corollary 1, we conclude that this term tends to zero in probability, where we have implicitly used the fact that the bounds in Lemmas 17, 18 and 19 are uniform over i.

The term J_2: We note that $\bar{U}^{(k)}$ is the same for all i and by Lemma 2.1 of Lei and Rinaldo (2015), the term $U_{i\cdot}$ is the same for all i belonging to community k. Hence, again

using the asymptotic expansion as in the proof of Corollary 1, we have that

$$\begin{aligned}
\left(\bar{U}^{(k)} - U^{(k)}\mathcal{O}_*^\top\right)\left(\bar{U}^{(k)} - U^{(k)}\mathcal{O}_*^\top\right)^\top &= \left(\frac{1}{n_k}\sum_{i=1}^{n_k}(\widehat{U}_i - U_i\mathcal{O}_*^\top)\right)\left(\frac{1}{n_k}\sum_{i=1}^{n_k}(\widehat{U}_i - U_i\mathcal{O}_*^\top)\right)^\top \\
&= \left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_*(\Lambda^{-1}V^\top E_i) + \mathcal{O}_*R_i\right)\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\mathcal{O}_*(\Lambda^{-1}V^\top E_i) + \mathcal{O}_*R_i\right) \\
&= \mathcal{O}_*\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\Lambda^{-1}V^\top E_i\right)\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\Lambda^{-1}V^\top E_i\right)^\top\mathcal{O}_*^\top \\
&\quad + \mathcal{O}_*\left(\frac{1}{n_k}\sum_{i=1}^{n_k}R_i\right)\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\Lambda^{-1}V^\top E_i\right)^\top\mathcal{O}_*^\top \\
&\quad + \mathcal{O}_*\left(\frac{1}{n_k}\sum_{i=1}^{n_k}\Lambda^{-1}V^\top E_i\right)\left(\frac{1}{n_k}\sum_{i=1}^{n_k}R_i\right)^\top\mathcal{O}_*^\top \\
&\quad + \mathcal{O}_*\left(\frac{1}{n_k}\sum_{i=1}^{n_k}R_i\right)\left(\frac{1}{n_k}\sum_{i=1}^{n_k}R_i\right)^\top\mathcal{O}_*^\top \\
&:= J_{21} + J_{22} + J_{23} + J_{24}.
\end{aligned}$$

The term J_{21} satisfies

$$\|\mathcal{O}_*(S^{(k)})^{-1}\mathcal{O}_*^\top J_{21}\| \leq \kappa\kappa_\sigma\frac{1}{n_k^2}\|(S^{(k)})^{-1/2}\sum_{i=1}^{n_k}\Lambda^{-1}V^\top E_i\|^2.$$

Therefore, by Markov's inequality, since κ and κ_σ are assumed bounded, by including them

in the implicit constants, we have that

$$\begin{aligned}
\mathbb{P}\left(\|\mathcal{O}_*(S^{(k)})^{-1}\mathcal{O}_*^\top J_{21}\| > t\right) &\lesssim \frac{\mathbb{E}\left(\|(S^{(k)})^{-1/2}\sum_{i=1}^{n_k}\Lambda^{-1}V^\top E_i\|^2\right)}{n_k^2 t} \\
&= \frac{\mathbb{E}\left(\sum_{i=1}^{n_k}\Lambda^{-1}V^\top E_i\right)^\top (S^{(k)})^{-1}\left(\sum_{j=1}^{n_k}\Lambda^{-1}V^\top E_i\right)\right)}{n_k^2 t} \\
&= \sum_{i=1}^{n_k}\frac{\mathbb{E}\left[E_i^\top V\Lambda^{-1}(S^{(k)})^{-1}\left(\sum_{j=1}^{n_k}\Lambda^{-1}V^\top E_j\right)\right]}{n_k^2 t} \\
&= \sum_{i=1}^{n_k}\frac{\mathbb{E}\mathrm{Tr}\left(\sum_{j=1}^{n_k}\Lambda^{-1}V^\top E_j\right)\left[E_i^\top V\Lambda^{-1}(S^{(k)})^{-1}\right]}{n_k^2 t} \\
&= \sum_{i=1}^{n_k}\frac{\mathrm{Tr}\Lambda^{-1}V^\top\Sigma^{(k)}V\Lambda^{-1}(S^{(k)})^{-1}}{n_k^2 t} \\
&= \sum_{i=1}^{n_k}\frac{r}{n_k^2 t} \\
&= \frac{r}{n_k t},
\end{aligned}$$

which implies that $\|(S^{(k)})^{-1}J_{21}\|$ converges to zero in probability.

Note that J_{22} is a rank one matrix. Therefore

$$\begin{aligned}
\|\mathcal{O}_*(S^{(k)})^{-1}\mathcal{O}_*^\top J_{22}\| &\leq \frac{\kappa\kappa_\sigma}{n_k^2}\left|\left\langle (S^{(k)})^{-1/2}\sum_{i=1}^{n_k}R_i, (S^{(k)})^{-1/2}\sum_{i=1}^{n_k}E_i^\top V\Lambda^{-1}\right\rangle\right| \\
&\lesssim \frac{1}{n_k^2}\left\|(S^{(k)})^{-1/2}\sum_{i=1}^{n_k}R_i\right\|\left\|(S^{(k)})^{-1/2}\sum_{i=1}^{n_k}E_i^\top V\Lambda^{-1}\right\| \\
&\lesssim \left\|(S^{(k)})^{-1/2}\sum_{i=1}^{n_k}E_i^\top V\Lambda^{-1}\right\|\frac{\max_i\|(S^{(k)})^{-1/2}R_i\|}{n_k}. \qquad \text{(B.7)}
\end{aligned}$$

By the same calculation as in J_{21}, for any $t > 0$,

$$\mathbb{P}\left(\left\|(S^{(k)})^{-1/2}\sum_{i=1}^{n_k}E_i^\top V\Lambda^{-1}\right\| > t\right) \leq \frac{rn_k}{t^2}.$$

Setting $t = n^{1/2}\sqrt{\log(n)}$ shows that

$$\left\|(S^{(k)})^{-1/2}\sum_{i=1}^{n_k} E_i^\top V\Lambda^{-1}\right\| \leq \sqrt{n\log(n)} \tag{B.8}$$

with probability at least $1 - o(1)$. Furthermore, by the same argument as in the proof of Corollary 1, (i.e. the residual bounds on R_i (as in Lemmas 17, 18, and 19), we have that with probability at least $1 - (n \vee d)^{-3}$

$$\max_i \|(S^{(k)})^{-1/2}R_i\| \lesssim \frac{\log(n\vee d)}{\text{SNR}} + \sqrt{\frac{\log(n\vee d)}{n}}, \tag{B.9}$$

where we again let the implicit constants depend on $\kappa, \mu_0, \kappa_\sigma$, and r, since they are assumed bounded in n and d. Hence, combining (B.7), (B.8), and (B.9), we have that with probability at least $1 - o(1)$,

$$\|\mathcal{O}_*(S^{(k)})^{-1}\mathcal{O}_*^\top J_{22}\| \leq \frac{\sqrt{n\log(n)}}{n_k}\left(\frac{\log(n\vee d)}{\text{SNR}} + \sqrt{\frac{\log(n\vee d)}{n}}\right)$$

which converges to zero. The same exact proof works for J_{23}. For J_{24}, we note that by Cauchy-Schwarz, the term tends to zero by a similar method as in the bound for J_1.

The term J_3: Recall that J_3 consists of two terms, one of which is the transpose of the other. We analyze only the first as the other is similar, again using the same expansion

as in J_1 and J_2. We have that

$$
\begin{aligned}
&\left\|\mathcal{O}_*[S^{(k)}]^{-1}\mathcal{O}_*^\top\left[\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\widehat{U}-U\mathcal{O}_*^\top\right)_{i\cdot}\left(U\mathcal{O}_*^\top-\bar{U}^{(k)}\right)_{i\cdot}^\top\right]\right\| \\
&\qquad=\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1}\mathcal{O}_*^\top\left(\mathcal{O}_*\Lambda^{-1}V^\top E_i+\mathcal{O}_*R_i\right)\left(U\mathcal{O}_*^\top-\bar{U}^{(k)}\right)_{i\cdot}^\top\right]\mathcal{O}_*\Bigg\| \\
&\qquad\le\kappa\kappa_\sigma\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\left(\Lambda^{-1}V^\top E_i+R_i\right)\left(U\mathcal{O}_*^\top-\bar{U}^{(k)}\right)_{i\cdot}^\top\mathcal{O}_*[S^{(k)}]^{-1/2}\right]\right\| \\
&\qquad\le\kappa\kappa_\sigma\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\left(\Lambda^{-1}V^\top E_i+R_i\right)\left(\frac{1}{n_k}\sum_{m=1}^{n_k}U\mathcal{O}_*^\top-\widehat{U}\right)_{m\cdot}^\top\mathcal{O}_*[S^{(k)}]^{-1/2}\right]\right\| \\
&\qquad\le\kappa\kappa_\sigma\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\left(\Lambda^{-1}V^\top E_i+R_i\right)\left(\frac{1}{n_k}\sum_{m=1}^{n_k}\mathcal{O}_*\Lambda^{-1}V^\top E_m+\mathcal{O}_*R_m\right)^\top\mathcal{O}_*[S^{(k)}]^{-1/2}\right]\right\| \\
&\qquad\le\kappa\kappa_\sigma\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\left(\Lambda^{-1}V^\top E_i+R_i\right)\left(\frac{1}{n_k}\sum_{m=1}^{n_k}E_m^\top V\Lambda^{-1}+R_m^\top\right)[S^{(k)}]^{-1/2}\right]\right\| \\
&\qquad\lesssim\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\left(\Lambda^{-1}V^\top E_i\right)\left(\frac{1}{n_k}\sum_{m=1}^{n_k}E_m^\top V\Lambda^{-1}\right)[S^{(k)}]^{-1/2}\right]\right\| \\
&\qquad\quad+\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\left(\Lambda^{-1}V^\top E_i\right)\left(\frac{1}{n_k}\sum_{m=1}^{n_k}R_m^\top\right)[S^{(k)}]^{-1/2}\right]\right\| \\
&\qquad\quad+\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}R_i\left(\frac{1}{n_k}\sum_{m=1}^{n_k}E_m^\top V\Lambda^{-1}\right)[S^{(k)}]^{-1/2}\right]\right\| \\
&\qquad\quad+\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}R_i\left(\frac{1}{n_k}\sum_{m=1}^{n_k}R_m^\top\right)[S^{(k)}]^{-1/2}\right]\right\| \\
&\qquad:=J_{31}+J_{32}+J_{33}+J_{34}.
\end{aligned}
$$

Since the term inside of J_{31} is a product of two rank-one matrices, its spectral norm is equal to

$$
\left\|\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\Lambda^{-1}V^\top E_i\right\|^2=\frac{1}{n_k^2}\left\|\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\Lambda^{-1}V^\top E_i\right\|^2.
$$

Note that

$$
\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\Lambda^{-1}V^\top E_i\right)\left(\sum_{m=1}^{n_k}E_m^\top V\Lambda^{-1}[S^{(k)}]^{-1/2}\right)&=\sum_{i=1}^{n_k}I_r \\
&=n_kI_r.
\end{aligned}
$$

Therefore, by Markov's inequality,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n_k^2}\left\|\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\Lambda^{-1}V^\top E_i\right\|^2 > t\right) &\leq \frac{\mathbb{E}\left\|\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\Lambda^{-1}V^\top E_i\right\|^2}{n_k^2 t} \\ &= \frac{n_k \mathrm{Tr} I_r}{n_k^2 t} \\ &\leq \frac{r}{n_k t},\end{aligned}$$

so that J_{31} tends to zero in probability.

For J_{32}, using the inequality $|ab| \leq |a||b|$, we have that on the event $\mathcal{E}$,

$$\begin{aligned}&\left|\left(\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\left(\Lambda^{-1}V^\top E_i\right)\right)\left(\frac{1}{n_k}\sum_{i=1}^{n_k}R_i^\top\right)[S^{(k)}]^{-1/2}\right)_{jl}\Bigg| \\ &\qquad \leq \max_m \left|\left([S^{(k)}]^{-1/2}R_m\right)_l\right|\left|\left(\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\left(\Lambda^{-1}V^\top E_i\right)\right)_j\right| \\ &\qquad \lesssim \left(\frac{\log(n\vee d)}{\mathrm{SNR}} + \sqrt{\frac{\log(n\vee d)}{n}}\right)\left|\left(\frac{1}{n_k}\sum_{i=1}^{n_k}[S^{(k)}]^{-1/2}\left(\Lambda^{-1}V^\top E_i\right)\right)_j\right|.\end{aligned}$$

We have already shown that the outermost term tends to zero in probability, implying that J_{32} tends to zero in probability since r is fixed. The same argument also works for J_{33}. For J_{34}, the same argument as J_{13} suffices, and hence J_3 tends to zero in probability, which completes the proof. □

B.5 Proofs of Lemmas in Section B.1

In this section we prove the additional Lemmas required for the proof of Theorem 7.

B.5.1 Proof of Lemmas 1 and 13

We first prove the spectral norm concentration bound in Lemma 1. We restate it here.

Lemma 1 (Spectral Norm Concentration). *Under assumption 2.1, there exists a universal*

constant $C_{spectral}$ such that with probability at least $1 - 4(n \vee d)^{-6}$

$$\|\Gamma(EM^\top + ME^\top + EE^\top)\| \le C_{spectral}\bigg(\sigma^2(n + \sqrt{nd}) + \sigma\sqrt{n}\kappa\lambda_r\bigg).$$

Proof of Lemma 1. We will follow the proof in Theorem 2 and Lemma 3 in Amini and Razaee (2021). More specifically, defining $\nu := d + \|M\|^2/\sigma^2$, we will show that there exists a universal constant c such that for any $u \ge 0$, with probability at least $1 - 4(n \vee d)^{-6}\exp(-u^2)$,

$$\|\Gamma(EM^\top + ME^\top + EE^\top)\| \le 2\sigma^2\nu\max(\delta^2, \delta), \tag{B.10}$$

for $\delta = \sqrt{\frac{6\log(n\vee d)/c + n\log(9)/c + u^2/c}{\nu}}$. To obtain the final result, note that by choosing C sufficiently large, we have that when $u = 0$,

$$\delta \le C\sqrt{n/\nu}.$$

Furthermore, $\max(\delta^2, \delta) \le \delta^2 + \delta$ for all δ. Then the result reads that with probability at least $1 - 4(n \vee d)^{-6}$,

$$\begin{aligned}
\|\Gamma(EM^\top + ME^\top + EE^\top)\| &\le 2\sigma^2\nu(\delta^2 + \delta) \\
&\le 2\sigma^2\nu(C^2 n/\nu + C\sqrt{n/\nu}) \\
&\le 2\sigma^2 C(Cn + \sqrt{n\nu}) \\
&\le 2\sigma^2 C(Cn + \sqrt{n(d + \|M\|^2/\sigma^2)}) \\
&\le 2\sigma^2 C(Cn + \sqrt{nd}) + 2C\sqrt{n}\sigma\lambda_1 \\
&\le C_{\text{spectral}}\bigg(\sigma^2(n + \sqrt{nd}) + \sqrt{n}\sigma\kappa\lambda_r\bigg)
\end{aligned}$$

by taking C_{spectral} sufficiently large and recalling that $\kappa = \lambda_1/\lambda_r$.

We now prove the claim (B.10). Let $z \in \mathbb{R}^n$ be a unit vector, and define $Y_z := z^\top\Gamma(Z)z$. Recall $Z = EM^\top + ME^\top + EE^\top$, where the rows of E are of the form $\Sigma_i^{1/2}W_i$ for vectors W_i with independent mean-zero coordinates with unit ψ_2 norm. Define $\vec{X} \in \mathbb{R}^{nd}$ as the

vector obtained by stacking the vectors X_i. Then

$$z^\top(A+\Gamma(Z))z = \|Xz\|^2 - \sum_i z_i^2\|X_i\|^2 + z^\top G(A)z.$$

Define the matrix $\Xi_z := z^\top \otimes I_d \in \mathbb{R}^{d\times nd}$. Then $\Xi_z\vec{X} = Xz$, where X is the matrix whose columns are X_i. Let Σ be the block-diagonal matrix whose i'th block is $\Sigma_i^{1/2}$. Then

$$Xz = \Xi_z\vec{\mu} + \Xi_z\Sigma^{1/2}\vec{W} = \Xi_z\Sigma^{1/2}\vec{\xi},$$

for $\xi = \Sigma^{-1/2}\vec{\mu} + \vec{W}$. Similarly, note that

$$\sum_{i=1}^n z_i^2\|X_i\|^2 = \|\mathrm{diag}(z_iI_d)\vec{X}\|^2 = \|\mathrm{diag}(z_iI_d)\Sigma^{1/2}\vec{\xi}\|^2.$$

Therefore, we have that $z^\top Az - z^\top G(A)z = z^\top\Gamma(A)z$ and

$$\begin{aligned}
z^\top(\Gamma(Z))z &= z^\top(A+\Gamma(Z))z - z^\top Az \\
&= \|\vec{X}z\|^2 - \sum_{i=1}^n z_i^2\|X_i\|^2 + z^\top G(A)z - z^\top Az \\
&= \|\Xi_z\Sigma^{1/2}\vec{\xi}\|^2 - \|\mathrm{diag}(z)\Sigma^{1/2}\vec{\xi}\|^2 - z^\top\Gamma(A)z \\
&= \vec{\xi}^\top B_z\vec{\xi} - \vec{\xi}^\top C_z\vec{\xi} - z^\top\Gamma(A)z \\
&= \vec{\xi}^\top(B_z - C_z)\vec{\xi} - \mathbb{E}\Big(\vec{\xi}^\top(B_z-C_z)\vec{\xi}\Big),
\end{aligned}$$

where $B_z := \Sigma^{1/2}\Xi_z^\top\Xi_z\Sigma^{1/2}$; $C_z := \Sigma^{1/2}\mathrm{diag}(z_iI_d)\mathrm{diag}(z_iI_d)\Sigma^{1/2}$. We now apply the extension of the Hanson-Wright inequality (Theorem 6 in Amini and Razaee (2021)) to this quadratic form to determine that

$$\mathbb{P}\Big(|Y_z| \geq t\Big) \leq 4\exp\Big(-c\min\Big(\frac{t^2}{\|B_z-C_z\|_F^2 + \|\widetilde{M}(B_z-C_z)\|_F^2}, \frac{t}{\|B_z-C_z\|}\Big)\Big), \quad \text{(B.11)}$$

where $\widetilde{M} := (\Sigma^{-1/2}\vec{\mu})^\top$. We now bound the denominators.

We have that $B_z - C_z = \Sigma^{1/2}(\Xi_z^\top\Xi_z - \mathrm{diag}(z_i^2I_d))\Sigma^{1/2}$. Hence $\|B_z - C_z\|_F = \|\Sigma^{1/2}(\Xi_z^\top\Xi_z - \mathrm{diag}(z_i^2I_d))\Sigma^{1/2}\|_F \leq \|\Sigma^{1/2}\|^2\|(\Xi_z^\top\Xi_z - \mathrm{diag}(z_i^2I_d))\|_F$. Furthermore, by the parallelogram

law and the fact that z is unit norm,

$$\begin{aligned}\|\Xi_z^\top \Xi_z - \mathrm{diag}(z_i^2 I_d)\|_F^2 &= 2\|\Xi_z^\top \Xi_z\|_F^2 + 2\|\mathrm{diag}(z_i^2 I_d)\|_F^2 - \|\Xi_z^\top \Xi_z + \mathrm{diag}(z_i^2 I_d)\|_F^2 \\ &\leq 2d + 2d \\ &\leq 4d,\end{aligned}$$

so that $\|B_z - C_z\|_F^2 \leq 4d\sigma^4$, where $\sigma^2 = \|\Sigma\| = \max_i \|\Sigma_i\|$. Additionally,

$$\begin{aligned}\|\widetilde{M}(B_z - C_z)\|_F^2 &= \|(\Sigma^{-1/2}\vec{\mu})^\top \Sigma^{1/2}(\Xi_z^\top \Xi_z - \mathrm{diag}(z_i^2 I_d))\Sigma^{1/2}\|_F^2 \\ &= \|\vec{\mu}(\Xi_z^\top \Xi_z - \mathrm{diag}(z_i^2 I_d))\Sigma^{1/2}\|_F^2 \\ &= \|\Sigma^{1/2}(\Xi_z^\top \Xi_z - \mathrm{diag}(z_i^2 I_d))\vec{\mu}\|_2^2 \\ &\leq \sigma^2 \|(\Xi_z^\top \Xi_z - \mathrm{diag}(z_i^2 I_d))\vec{\mu}\|_2^2.\end{aligned}$$

Note that

$$\begin{aligned}\|(\Xi_z^\top \Xi_z - \mathrm{diag}(z_i^2 I_d))\vec{\mu}\|_2 &\leq \|\Xi_z^\top \Xi_z \vec{\mu}\| + \|\mathrm{diag}(z_i^2 I_d)\vec{\mu}\| \\ &\leq 2\|M\|\end{aligned}$$

using the definition of $\vec{\mu}$. Finally, for the operator norm, we have that

$$\begin{aligned}\|B_z - C_z\| &\leq \|B_z\| + \|C_z\| \\ &\leq \sigma^2\|\Xi_z\|^2 + \sigma^2\|\mathrm{diag}(z_i^2 I_d)\|^2 \\ &\leq 2\sigma^2.\end{aligned}$$

In summary,

$$\|B_z - C_z\|_F^2 \leq 4\sigma^4 d; \tag{B.12}$$

$$\|\widetilde{M}(B_z - C_Z)\|_F^2 \leq 4\sigma^2\|M\|^2; \tag{B.13}$$

$$\|B_z - C_z\| \leq 4\sigma^2. \tag{B.14}$$

Plugging (B.12), (B.13), and (B.14) into Equation B.11 and absorbing $1/4$ into the constant c yields

$$\mathbb{P}\bigg(|Y_z| \geq t\bigg) \leq 4\exp\bigg(-c\min\bigg(\frac{t^2}{\sigma^4 d + \sigma^2\|M\|^2}, \frac{t}{\sigma^2}\bigg)\bigg).$$

Define $\widetilde{t} := \sigma^2 t$. Then the inequality above becomes

$$\mathbb{P}\left(|Y_z| \geq t\sigma^2\right) \leq 4\exp\bigg(-c\min\bigg(\frac{t^2}{d + \sigma^{-2}\|M\|^2}, t\bigg)\bigg).$$

By changing t to νt, the above concentration can be written via

$$\mathbb{P}\bigg(|Y_z| \geq t\sigma^2\nu\bigg) \leq 4\exp(-c\nu\min(t^2, t)).$$

Define $\delta := \sqrt{\frac{6\log(n\vee d)+n\log(9)+u^2}{\nu c}}$. Note that regardless of the value of δ,

$$\min(\max(\delta^2, \delta)^2, \max(\delta^2, \delta)) \leq \delta^2.$$

Taking $t = \max(\delta^2, \delta)$, we arrive at

$$\begin{aligned}
\mathbb{P}(|Y_z| \geq \sigma^2\nu\max(\delta^2, \delta)) &\leq 4\exp(-c\nu\min(\max(\delta^2, \delta)^2, \max(\delta^2, \delta))) \\
&\leq 4\exp(-c\nu\delta^2) \\
&\leq 4\exp(-c\frac{6\log(n\vee d) + n\log(9) + u^2}{c}) \\
&\leq 4\exp(-6\log(n\vee d) - n\log(9) - u^2).
\end{aligned}$$

Now we follow the proof of Theorem 2 in Amini and Razaee (2021) via an ε-net. Let $\mathcal{N}$ be a $1/4$ net of the n-sphere, and hence that $|\mathcal{N}| \leq 9^n$. Then $\|\Gamma(Z)\| \leq 2\max_{z\in\mathcal{N}}|Y_z|$, so that

by a union bound,

$$\begin{aligned}\mathbb{P}\bigg(\|\Gamma(Z)\| \geq 2\sigma^2\nu\max(\delta^2,\delta)\bigg) &\leq 4\cdot 9^n\exp(-6\log(n\vee d)-n\log(9)-u^2)\\ &\leq 4\exp(n\log(9)-6\log(n\vee d)-n\log(9)-u^2)\\ &\leq 4(n\vee d)^{-6}\exp(-u^2).\end{aligned}$$

This proves the result.

□

Lemma 13. *With probability at least* $1-2(n\vee d)^{-5}$,

$$\|\widetilde{U}_D\widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\| \leq \frac{C\delta_1}{\lambda_r^2}\|U\|_{2,\infty}$$

Proof of Lemma 13. This follows because

$$\begin{aligned}\|\mathrm{diag}(EM^\top + ME^\top)\|_2 &= \max_i |(EM^\top + ME^\top)_{ii}|\\ &= 2\max_i |\langle E_i.M_i\rangle|\\ &= 2\max_i |\sum_\alpha Y_{i\alpha}(\Sigma_i^{1/2}M_i)_\alpha|.\end{aligned}$$

This is a sum of d independent random variables with ψ_2 norm bounded by $2\max_i\|\Sigma_i^{1/2}M_i\| \leq 2\sigma\max_i\|M_i\| \leq 2\sigma\lambda_1\|U\|_{2,\infty}$. Consequently, Hoeffding's inequality and a union bound that $\|\mathrm{diag}(EM^\top+ME^\top)\| \leq 2\lambda_1\sigma\|U\|_{2,\infty}\sqrt{6\log(n\vee d)}$ with probability at least $1-2(n\vee d)^{-5}$. By the Davis-Kahan theorem,

$$\begin{aligned}\|\widetilde{U}_D\widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\| &\leq C\frac{2\lambda_1}{\lambda_r^2}\sigma\|U\|_{2,\infty}\sqrt{\log(n\vee d)}\\ &\leq \frac{C\delta_1}{\lambda_r^2}\|U\|_{2,\infty}\end{aligned}$$

where we have used Weyl's inequality and Assumption 2.2. □

B.5.2 Proof of Lemma 15

In order to prove Lemma 15, we will also require the following additional lemmas.

Lemma 20. *There exists an absolute constant C_0 such that with probability at least $1-(n\vee d)^{-5}$ it holds that*

$$\|U_\perp U_\perp^\top WU\|_{2,\infty} \le C_0\bigg(\sqrt{rnd}\log(n\vee d)\sigma^2\|U\|_{2,\infty} + \sqrt{rn\log(n\vee d)}\lambda_1\sigma\bigg)\|U\|_{2,\infty},$$

Proof of Lemma 20. Note that

$$\begin{aligned}\|U_\perp U_\perp^\top WU\|_{2,\infty} &\le \|UU^\top(ME^\top + EM^\top + \Gamma(EE^\top))U\|_{2,\infty} + \|(EM^\top + ME^\top + \Gamma(EE^\top))U\|_{2,\infty} \\ &\le \|U\|_{2,\infty}\|W\| + \|(EM^\top + ME^\top + \Gamma(EE^\top))U\|_{2,\infty} \qquad \text{(B.15)}\end{aligned}$$

We will analyze the second term. Note that the i,j entry of W can be written as

$$E_i^\top M_j + M_i^\top E_j + E_i^\top E_j(1-\delta_{ij}).$$

Furthermore, note that

$$\|WU\|_{2,\infty} \le \sqrt{r}\|WU\|_{\max}.$$

Fix an index i,j. then this shows that

$$(WU)_{ij} = \sum_{k=1}^n \bigg(\langle Y_i, \Sigma_i^{1/2}M_k\rangle + \langle Y_k, \Sigma_k^{1/2}M_i\rangle + (1-\delta_{ik})\langle Y_i, \Sigma_i^{1/2}\Sigma_k^{1/2}Y_k\rangle\bigg)U_{kj}.$$

Define

$$\xi_{ik} := \begin{cases}\langle Y_i, \Sigma_i^{1/2}M_k\rangle + \langle Y_k, \Sigma_k^{1/2}M_i\rangle + \langle Y_i, \Sigma_i^{1/2}\Sigma_k^{1/2}Y_k\rangle & i\neq k \\ 2\langle Y_i, \Sigma_i^{1/2}M_i\rangle & i=k.\end{cases}$$

Then

$$(WU)_{ij} = \sum_{k=1}^{n} \xi_{ik} U_{kj}$$

Expanding out ξ_{ik}, we have that

$$\xi_{ik} = \begin{cases} \sum_{\alpha=1}^{d} (Y_{i\alpha}[\Sigma_i^{1/2} M_k]_\alpha + Y_{k\alpha}(\Sigma_k^{1/2} M_i)_\alpha) + \sum_{\alpha=1}^{d} \sum_{\beta=1}^{d} (\Sigma_k^{1/2} \Sigma_i^{1/2})_{\alpha\beta} Y_{i\alpha} Y_{k\beta} & i \neq k \\ 2 \sum_{\alpha=1}^{d} (Y_{i\alpha}[\Sigma_i^{1/2} M_i]_\alpha & i = k, \end{cases}$$

where $Y_{i\alpha}$ denotes the α coordinate of the i'th random vector Y_i. We want to write this in terms of the independent collection of random variables Y. We have that

$$\begin{aligned} \left| \sum_{k=1}^{n} \xi_{ik} U_{kj} \right| &\leq 2 \left| \sum_{k=1}^{n} U_{kj} \sum_{\alpha=1}^{d} (Y_{i\alpha}[\Sigma_i^{1/2} M_k]_\alpha] \right| + \left| \sum_{k=1, k\neq i}^{n} U_{kj} \sum_{\alpha=1}^{d} Y_{k\alpha}(\Sigma_k^{1/2} M_i)_\alpha) \right| \\ &\quad + \left| \sum_{k=1, k\neq i}^{n} U_{kj} \sum_{\alpha=1}^{d} \sum_{\beta=1}^{d} (\Sigma_k^{1/2} \Sigma_i^{1/2})_{\alpha\beta} Y_{i\alpha} Y_{k\beta} \right| \\ &= (I) + (II) + (III). \end{aligned}$$

In what follows, each term is bounded separately.

The term (I): The first term can be written via

$$\sum_{\alpha=1}^{d} Y_{i\alpha} \left[\sum_{k=1}^{n} U_{kj} [\Sigma_i^{1/2} M_k]_\alpha \right],$$

which is a sum of d mean-zero random variables. So it suffices to bound the coefficients in order to apply the Hoeffding concentration inequality for subgaussians. The coefficients can be found via

$$a_{(I)}(\alpha) := \sum_{k=1}^{n} U_{kj} [\Sigma_i^{1/2} M_k]_\alpha$$

for α ranging from 1 to d. Furthermore,

$$\begin{aligned}\sum_{\alpha=1}^{d}\Big(\sum_{k=1}^{n}U_{kj}\widetilde{M}_{ik\alpha}\Big)^2 &\le \sum_{\alpha=1}^{d}\Big(\sum_{k=1}^{n}U_{kj}^2\Big)\Big(\sum_{k=1,k\neq i}^{n}\widetilde{M}_{ik\alpha}^2\Big)\\ &\le \sum_{\alpha=1}^{d}\Big(\sum_{k=1}^{n}(\Sigma_i^{1/2}M_k)_\alpha^2\Big)\\ &\le \|\Sigma_i^{1/2}M^\top\|_F^2\\ &= \|M\Sigma_i^{1/2}\|_F^2\\ &\le n\|M\Sigma_i^{1/2}\|_{2,\infty}^2\\ &\le n\|U\|_{2,\infty}^2\lambda_1^2\sigma^2.\end{aligned}$$

By the generalized Hoeffding inequality (Theorem 2.6.3 in Vershynin (2018)) we obtain

$$\mathbb{P}\Big(|(I)|>t\Big)\le 2\exp\Big[-c\Big(\frac{t^2}{n\|U\|_{2,\infty}^2\lambda_1^2\sigma^2}\Big)\Big]$$

Taking $t = C\|U\|_{2,\infty}\lambda_1\sigma\sqrt{n(2\gamma\log(n\vee d)+2\log(r))}$ shows that this holds with probability at least $1-2r^{-1}n^{-\gamma}$. Therefore, we derive the first bound,

$$\begin{aligned}(I) &\le C\|U\|_{2,\infty}\lambda_1\sigma\sqrt{n(2\gamma\log(n\vee d)+2\log(r))}\\ &\le C\|U\|_{2,\infty}\lambda_1\sigma\sqrt{\gamma n\log(n\vee d)},\end{aligned}$$

since $\log(r)\le\log(n\vee d)$.

The term (II): By a similar argument, it suffices to bound the norm of the coefficient vector, where the coefficient vector ranges over α and $k\neq i$ with

$$(a_2)_{\alpha,k} := U_{kj}(\Sigma_k^{1/2}M_i)_\alpha).$$

Therefore, we see that

$$\begin{aligned}
\|a_2\|_2^2 :&= \sum_{\alpha=1}^{d} \sum_{k=1,k\neq i}^{n} U_{kj}^2 (\Sigma_k^{1/2} M_i)_\alpha)^2 \\
&\leq \sum_{\alpha=1}^{d} \sum_{k=1,k\neq i}^{n} (\Sigma_k^{1/2} M_i)_\alpha)^2 \\
&\leq \sum_{k=1,k\neq i}^{n} \|\Sigma_k^{1/2} M_i\|^2 \\
&\leq n \max_k \|M_i^\top \Sigma_k^{1/2}\|^2 \\
&\leq n \|M_i\|^2 \max_k \|\Sigma_k^{1/2}\|^2 \\
&\leq n \|M\|_{2,\infty} \max_k \|\Sigma_k^{1/2}\|^2 \\
&\leq n\sigma^2 \|M\|_{2,\infty}^2 \\
&\leq \|U\|_{2,\infty}^2 n\sigma^2 \lambda_1^2.
\end{aligned}$$

Therefore, we see that with probability at least $1 - 2r^{-1}n^{-\gamma}$,

$$(II) \leq \sqrt{2}\|U\|_{2,\infty}\sigma\lambda_1\sqrt{\gamma \log(n \vee d) n},$$

which matches the previous bound.

The term (III): The final quantity is of the form

$$(III) := \sum_{k=1,k\neq i}^{n} U_{kj} \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} (\Sigma_k^{1/2}\Sigma_i^{1/2})_{\alpha\beta} Y_{i\alpha} Y_{k\beta}.$$

We use a conditioning argument. First, consider the event

$$\mathcal{A} := \{\sum_{\alpha=1}^{d} Y_{i\alpha} U_{kj} (\Sigma_k^{1/2}\Sigma_i^{1/2})_{\alpha\beta} > s \text{ for any } k \text{ and } \beta\}$$

for some $s > 0$ to be determined later. Note that this event depends on the collection $\{Y_{i\alpha}\}_\alpha$ only. Conditional on this event, we see that the sum is a sum of independent mean-zero subguassian random variables with ψ_2 norm bounded by s. Since there are $(n-1)d$ such

random variables, the generalized Hoeffding inequality shows that

$$\mathbb{P}(|(III)| > t|\mathcal{A}) \leq 2\exp\bigg(-\frac{1}{2}\frac{t^2}{(n-1)ds^2}\bigg)$$

so taking $t = \sqrt{2}\sqrt{(n-1)ds^2(\gamma\log(n\vee d)+\log(r))}$ shows that this holds with probability at least $1-2r^{-1}n^{-\gamma}$. Now I will find a probabilistic bound on s.

Note that the sum in the event $\mathcal{A}$ is a sum over d independent random variables, so it suffices to estimate the term

$$\sum_{\alpha=1}^{d}(U_{kj}(\Sigma_k^{1/2}\Sigma_i^{1/2})_{\alpha\beta})^2$$

uniformly over k,β. We see that

$$\begin{aligned}\max_{k,\beta}\sum_{\alpha=1}^{d}(U_{kj}(\Sigma_k^{1/2}\Sigma_i^{1/2})_{\alpha\beta})^2 &\leq \max_{k,\beta}U_{kj}^2\sum_{\alpha=1}^{d}(\Sigma_k^{1/2}\Sigma_i^{1/2})_{\alpha\beta}^2\\ &\leq \max_k\|\Sigma_i^{1/2}\Sigma_k^{1/2}\|_{2,\infty}^2U_{kj}^2\\ &\leq \|\Sigma_i^{1/2}\|_{2,\infty}^2\|U\|_{2,\infty}^2\max_k\|\Sigma_k^{1/2}\|^2\\ &\leq \|\Sigma_i^{1/2}\|_{2,\infty}^2\|U\|_{2,\infty}^2\sigma^2.\end{aligned}$$

Therefore, we have that since there are at most nd terms with k and β,

$$\begin{aligned}\mathbb{P}(\mathcal{A}) &\leq nd\max_{k,\beta}\mathbb{P}\bigg(\sum_{\alpha=1}^{d}Y_{i\alpha}U_{kj}(\Sigma_k^{1/2}\Sigma_i^{1/2})_{\alpha\beta} > s \text{ for fixed } k \text{ and } \beta\bigg)\\ &\leq 2nd\exp\bigg(-\frac{1}{2}\frac{s^2}{\|\Sigma_i^{1/2}\|_{2,\infty}^2\|U\|_{2,\infty}^2\sigma^2}\bigg),\end{aligned}$$

so taking $s = \sqrt{2}\sqrt{(\gamma+1)\log(n\vee d)+\log(d)+\log(r)}\|U\|_{2,\infty}\|\Sigma_i^{1/2}\|_{2,\infty}\sigma$ shows that this holds with probability at least $1-2r^{-1}n^{-\gamma}$. Therefore, with this fixed choice of s, we see that

$$\begin{aligned}\mathbb{P}((III)\leq t) &\leq \mathbb{P}((III)\leq t|\mathcal{A})+\mathbb{P}(\mathcal{A}^c)\\ &\leq 2r^{-1}n^{-\gamma}+2r^{-1}n^{-\gamma}\end{aligned}$$

Doing the algebra, we see that

$$
\begin{aligned}
t &= \sqrt{2}\sqrt{(n-1)d}s\sqrt{\gamma\log(n\vee d)+\log(r)} \\
&= \sqrt{2}\sqrt{(n-1)d}\sqrt{\gamma\log(n\vee d)+\log(r)}\left(\sqrt{2}\sqrt{(\gamma+1)\log(n\vee d)+\log(d)+\log(r)}\|U\|_{2,\infty}\|\Sigma_i^{1/2}\|_{2,\infty}\sigma\right) \\
&\leq 4\sqrt{nd}\sqrt{(\gamma+1)\log(n\vee d)}\sqrt{(\gamma+1)\log(n\vee d)}\|U\|_{2,\infty}\|\Sigma_i^{1/2}\|_{2,\infty}\sigma \\
&\leq 8\sqrt{nd}\|U\|_{2,\infty}\sigma^2\gamma\log(n\vee d)
\end{aligned}
$$

Letting $\gamma \geq 20$ and taking a union over all nr entries shows that with probability at least $1-(n\vee d)^{-10}$.,

$$\|WU\|_{2,\infty} \leq C_0\left(\sqrt{rnd}\log(n\vee d)\sigma^2+\sqrt{rn\log(n\vee d)}\lambda_1\sigma\right)\|U\|_{2,\infty}.$$

Hence, the result holds by also applying Lemma 1 on the term $\|W\|$ in Equation (B.15). Since the spectral norm bound is smaller than the bound above and holds with probability at least $1-4(n\vee d)^{-6}$, the result holds by increasing the constant C_0 by a factor of 2. □

Lemma 15. *Let $W = EM^\top + ME^\top + \Gamma(EE^\top)$. There exists universal constants C_1 and C_2 such that for any $p \geq 1$, we have that with probability at least $1-(p+1)(n\vee d)^{-5}$ for all $1 \leq p_0 \leq p$ that*

$$\|(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p_0-1}WU\|_{2,\infty} \leq C_1(C_2\delta)^{p_0}\|U\|_{2,\infty},$$

Proof of Lemma 15. We will prove the result by induction. When $p=1$, the result holds by Lemma 20, where we take C_1 and C_2 in the statement of the lemma to be large, fixed constants. We now fix these constants C_1 and C_2.

Let $p>1$. Assume that with probability at least $1-p(n\vee d)^{-5}$ that for all $1\leq p_0\leq p-1$ that

$$\|U_\perp U_\perp^\top(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p_0-1}WU\|_{2,\infty} \leq C_1(C_2\delta)^{p_0}\|U\|_{2,\infty}.$$

Note that by the definition of $W := EM^\top + ME^\top + \Gamma(EE^\top)$, we have the identity

$U_\perp U_\perp^\top W U_\perp U_\perp^\top = U_\perp U_\perp^\top \Gamma(EE^\top) U_\perp U_\perp^\top$. Let $\mathcal{B}$ be the event that

$$\mathcal{B} := \left\{ \|U_\perp U_\perp^\top (U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p_0 - 1} W U\|_{2,\infty} \le C_1 (C_2 \delta)^{p_0} \|U\|_{2,\infty} \text{ for all } 1 \le p_0 \le p-1 \right\} \bigcap \left\{ \|W\| \le \frac{\delta}{\sqrt{r \log(n \vee d)}} \right\} \bigcap \left\{ \|\Gamma(EE^\top)\| \le \frac{\delta}{\sqrt{r \log(n \vee d)}} \right\}.$$

Note that $\mathbb{P}(\mathcal{B}) \ge 1 - p(n \vee d)^{-5} - 12(n \vee d)^{-6}$ by the induction hypothesis and Lemmas 1 and 13 (to get the bound on $\Gamma(EE^\top)$, we apply Lemma 1 with $M = 0$).

We note that

$$\begin{aligned} \|(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-1} W U\|_{2,\infty} &\le \|UU^\top \Gamma(EE^\top)(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} W U\|_{2,\infty} \\ &\quad + \|\Gamma(EE^\top)(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} W U\|_{2,\infty} \\ &\le \|U\|_{2,\infty} \|\Gamma(EE^\top)\| \|W\|^{p-1} + \|\Gamma(EE^\top)(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} W U\|_{2,\infty} \\ &\le \|U\|_{2,\infty} \|\Gamma(EE^\top)\| \|W\|^{p-1} + \sqrt{r} \|\Gamma(EE^\top)(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} W U\|_{\max}. \end{aligned} \tag{B.16}$$

We first will analyze the (i,j) entry of the matrix $\Gamma(EE^\top)(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} W U$. Fix an index i and consider the auxiliary matrix X^{-i} defined via

$$X^{-i} := (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-2} (I - e_i e_i^\top) W (I - e_i e_i^\top) U.$$

Note that X^{-i} is independent of the random variable E_i. Define also the matrix X via

$$X := (U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} W U.$$

We note that the i,j entry of the matrix $\Gamma(EE^\top) X$ can be written as

$$\sum_{k \neq i} \langle E_i, E_k \rangle X_{kj}^{-i} + \sum_{k \neq i} \langle E_i, E_k \rangle (X_{kj} - X_{kj}^{-i}) := I_1^{ij} + I_2^{ij}$$

where

$$I_1^{ij} := \sum_{k \neq i} \langle E_i, E_k \rangle X_{kj}^{-i};$$
$$I_2^{ij} := \sum_{k \neq i} \langle E_i, E_k \rangle (X_{kj} - X_{kj}^{-i}).$$

Let $\mathcal{A}_{ij}$ be the event that for all k, $|X_{kj}^{-i}| \leq 2C_1(C_2\delta)^{p-1}\|U\|_{2,\infty} =: A$. We first study I_1^{ij} on the event $\mathcal{A}_{ij}$.

Note that

$$\begin{aligned} I_1^{ij} &= (E_i)^\top \left(\sum_{k \neq i} E_k X_{kj}^{-i} \right) \\ &= Y_i^\top \Sigma_i^{1/2} \Big((E^{-i})^\top X^{-i} e_j \Big), \end{aligned}$$

where E^{-i} is the matrix E with the i'th row set to zero. Recall that by construction $\mathcal{A}_{ij}$ does not depend on E_i. Hence, conditional on E_k for $k \neq i$, this is a sum of d independent mean-zero random variables $Y_{i\alpha}$ with coefficients indexed by α. Define the vector $v := \Sigma_i^{1/2}(E^{-i})^\top X^{-i} e_j \in \mathbb{R}^d$. Note that

$$\begin{aligned} \|v\| &= \left\| \Sigma_i^{1/2} (E^{-i})^\top X^{-i} e_j \right\| \\ &\leq \|\Sigma_i^{1/2}\| \|E^{-i}\| \|X^{-i} e_j\|_2 \\ &\leq \sigma_i \|E^{-i}\| \|X^{-i} e_j\|_2, \end{aligned}$$

which always holds. Moreover, on the event $\mathcal{A}_{ij}$ it holds that

$$\begin{aligned} \|X^{-i} e_j\|_2 &\leq \sqrt{n} \max_k |X_{kj}^{-i}| \\ &\leq \sqrt{n} A. \end{aligned}$$

Suppose for the moment that $\mathcal{E}$ is the event that $\|E^{-i}\| \leq B$, for some bound B. Then by

independence of this event with E_i, it holds that

$$\|Y_i^\top v\|_{\psi_2}^2 = \|v\|_2^2.$$

Therefore, by Hoeffding's inequality, for some universal constant C_3,

$$\mathbb{P}\bigg(|Y_i^\top \Sigma_i^{1/2}(E^{-i})^\top X^{-i} e_j| > C_3\sqrt{n}\sigma_i AB\sqrt{20\log(n\vee d)} \cap \mathcal{A}_{ij} \cap \mathcal{E}\bigg) \leq 2(n\vee d)^{-20}. \quad \text{(B.17)}$$

We now deduce a bound B for the spectral norm of E^{-i}. Note that for any deterministic unit vectors $a \in \mathbb{R}^n$ and $b \in \mathbb{R}^d$, by independence of the $Y_{k\alpha}$'s it holds that

$$\begin{aligned}
\|a^\top E^{-i} b\|_{\psi_2}^2 &= \|\sum_{k\neq i}\sum_j a_k E_{kj}^{-i} b_j\|_{\psi_2}^2 \\
&= \bigg\|\sum_{k\neq i}\sum_j a_k (\Sigma_k^{1/2} Y_k)_j b_j\bigg\|_{\psi_2}^2 \\
&= \sum_{k\neq i} a_k^2 \bigg\|\sum_j (\Sigma_k^{1/2} Y_k)_j b_j\bigg\|_{\psi_2}^2 \\
&= \sum_{k\neq i} a_k^2 \bigg\|\sum_j\sum_\alpha (\Sigma_k^{1/2})_{j\alpha} Y_{k\alpha} b_j\bigg\|_{\psi_2}^2 \\
&= \sum_{k\neq i} a_k^2 \sum_\alpha \bigg\|Y_{k\alpha}\sum_j (\Sigma_k^{1/2})_{j\alpha} b_j\bigg\|_{\psi_2}^2 \\
&= \sum_{k\neq i} a_k^2 \sum_\alpha \|\sum_j (\Sigma_k^{1/2})_{j\alpha} b_j\|_2^2 \\
&= \sum_{k\neq i} a_k^2 \sum_\alpha [\Sigma_k^{1/2} b]_\alpha^2 \\
&= \sum_{k\neq i} a_k^2 \|\Sigma_k^{1/2} b\|_2^2 \\
&\leq \sigma^2 \sum_{k\neq i} a_k^2 \|b\|^2 \\
&\leq \sigma^2.
\end{aligned}$$

Therefore, by a standard ε-net argument (e.g. the argument in the proof of Theorem 4.4.5

in Vershynin (2018)), it holds that there exists a universal constant C_4 such that

$$\mathbb{P}\bigg(\|E^{-i}\| > C_4\sigma(\sqrt{d}+u)\bigg) \le 2\exp(-u^2).$$

Here we implicitly used Assumption 2.3, or that $d \ge cn$. Define the event

$$\mathcal{E} := \bigg\{\|E^{-i}\| \le C_4\sigma(\sqrt{d}+\sqrt{20\log(n\vee d)})\bigg\},$$

so that $\mathbb{P}(\mathcal{E}^c) \le 2(n\vee d)^{-20}$. On the event $\mathcal{E}$ it holds that $\|E^{-i}\| \le C_4\sigma\sqrt{20d\log(n\vee d)}$. By Equation (B.17),

$$\begin{aligned}\mathbb{P}\bigg(|Y_i^\top\Sigma_i^{1/2}(E^{-i})^\top X^{-i}e_j| >& 20C_3C_4\sigma^2\sqrt{nd}A\log(n\vee d)\cap\mathcal{A}_{ij}\bigg)\\ &\le \mathbb{P}\bigg(|Y_i^\top\Sigma_i^{1/2}(E^{-i})^\top X^{-i}e_j| > 20C_3C_4\sigma^2\sqrt{nd}A\log(n\vee d)\cap\mathcal{A}_{ij}\cap\mathcal{E}\bigg)+\mathbb{P}(\mathcal{E}^c)\\ &\le 4(n\vee d)^{-20}.\end{aligned}$$

Furthermore, $A := 2C_1(C_2\delta)^{p-1}\|U\|_{2,\infty}$. Recall that δ satisfies, for some sufficiently large absolute constant C_0,

$$\delta = C_0\bigg(\sqrt{rnd}\log(n\vee d)\sigma^2+\sqrt{rn\log(n\vee d)}\sigma\lambda_1\bigg),$$

so that $\frac{\delta}{\sqrt{r}} \ge \log(n\vee d)\sqrt{nd}\sigma^2+\lambda_1\sqrt{n\log(n\vee d)}\sigma$. Therefore,

$$\mathbb{P}\bigg(|I_1^{ij}| > \frac{C_1(C_2\delta)^p}{4\sqrt{r}}\|U\|_{2,\infty}\cap\mathcal{A}_{ij}\bigg) \le 4(n\vee d)^{-20}$$

as long as $C_2 \ge 20C_3C_4$, which is true by taking C_2 large since C_3 and C_4 are universal

constants. Therefore, from equation (B.16),

$$\begin{aligned}
&\mathbb{P}\bigg(\|U\|_{2,\infty}\|\Gamma(EE^\top)\|\|W\|^{p-1} + \sqrt{r}\|\Gamma(EE^\top)(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2}WU\|_{\max} > C_1(C_2\delta)^p\|U\|_{2,\infty}\bigg) \\
&\quad \le \mathbb{P}\bigg(\|U\|_{2,\infty}\|\Gamma(EE^\top)\|\|W\|^{p-1} + \sqrt{r}\|\Gamma(EE^\top)(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2}WU\|_{\max} > C_1(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B}\bigg) \\
&\quad \le \mathbb{P}\bigg(\|\Gamma(EE^\top)(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2}WU\|_{\max} > \frac{C_1}{2\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B}\bigg) + \mathbb{P}(\mathcal{B}^c) \\
&\quad \le \mathbb{P}\bigg(\bigcup_{i,j} \Big| e_i^\top \Gamma(EE^\top)(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2}WUe_j \Big| > \frac{C_1}{2\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B}\bigg) + \mathbb{P}(\mathcal{B}^c) \\
&\quad \le nr \max_{i,j} \mathbb{P}\bigg(|\Gamma(EE^\top)(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2}WU|_{ij} > \frac{C_1}{2\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B}\bigg) + \mathbb{P}(\mathcal{B}^c) \\
&\quad \le nr \max_{i,j} \mathbb{P}\bigg(|I_1^{ij} + I_2^{ij}| > \frac{C_1}{2\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B}\bigg) + \mathbb{P}(\mathcal{B}^c) \\
&\quad \le nr \max_{i,j} \mathbb{P}\bigg(|I_1^{ij}| > \frac{C_1}{4\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B}\bigg) + nr \max_{i,j} \mathbb{P}\bigg(|I_2^{ij}| > \frac{C_1}{4\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B}\bigg) + \mathbb{P}(\mathcal{B}^c) \\
&\quad \le nr \max_{i,j} \mathbb{P}\bigg(|I_1^{ij}| > \frac{C_1}{4\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B} \cap \mathcal{A}_{ij}\bigg) + nr \max_{i,j} \mathbb{P}(\mathcal{B} \cap \mathcal{A}_{ij}^c) \\
&\qquad + nr \max_{i,j} \mathbb{P}\bigg(|I_2^{ij}| > \frac{C_1}{4\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B}\bigg) + \mathbb{P}(\mathcal{B}^c). \qquad \text{(B.18)}
\end{aligned}$$

In Lemma 21 we show for all i and j that $\mathbb{P}(\mathcal{B} \cap \mathcal{A}_{ij}^c) = 0$ and

$$\mathbb{P}\bigg(|I_2^{ij}| > \frac{C_1}{4\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B}\bigg) = 0.$$

In addition, from our previous analysis,

$$\mathbb{P}\bigg(|I_1^{ij}| > \frac{C_1(C_2\delta)^p}{4\sqrt{r}}\|U\|_{2,\infty} \cap \mathcal{A}_{ij}\bigg) \le 4(n \vee d)^{-20}.$$

Finally, $\mathbb{P}(\mathcal{B}^c) \le p(n\vee d)^{-5} + 12(n\vee d)^{-6}$ by the induction hypothesis. Plugging these results into Expression (B.18), we obtain

$$\begin{aligned}
nr\mathbb{P}\bigg(|I_1^{ij}| > \frac{C_1}{4\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{A}_{ij}\bigg) + nr\mathbb{P}(\mathcal{B} \cap \mathcal{A}_{ij}^c) + \mathbb{P}\bigg(|I_2^{ij}| > \frac{C_1}{4\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B}\bigg) + \mathbb{P}(\mathcal{B}^c) \\
\le 4(nr)(n \vee d)^{-20} + p(n \vee d)^{-5} + 12(n \vee d)^{-6} \\
\le (p+1)(n \vee d)^{-5}
\end{aligned}$$

as desired. □

Lemma 21. *Define X^{-i} and X as in the proof of Lemma 15. Let $\mathcal{A}_{ij}$ and $\mathcal{B}$ be the events defined via*

$$\mathcal{A}_{ij} := \left\{ \max_k |X^{-i}_{kj}| \leq 2C_1(C_2\delta)^{p-1}\|U\|_{2,\infty} \right\}$$
$$\mathcal{B} := \left\{ \|U_\perp U_\perp^\top (U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p_0-1} W U\|_{2,\infty} \leq C_1(C_2\delta)^{p_0}\|U\|_{2,\infty} \text{ for all } 1 \leq p_0 \leq p-1 \right\}$$
$$\bigcap \left\{ \|W\| \leq \frac{\delta}{\sqrt{r\log(n\vee d)}} \right\}$$
$$\bigcap \left\{ \|\Gamma(EE^\top)\| \leq \frac{\delta}{\sqrt{r\log(n\vee d)}} \right\}.$$

Then $\mathcal{B}\cap\mathcal{A}_{ij}^c$ is empty for all i and j. Define also $I_2^{ij} := e_i^\top \Gamma(EE^\top)(X - X^{-i})e_j$. as in the proof of Lemma 15. Then again for all i and j,

$$\mathbb{P}\left(|I_2^{ij}| > \frac{C_1}{4\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty} \cap \mathcal{B} \right) = 0.$$

Proof of Lemma 21. The proof will be split up into steps, the first of which will be expanding out the difference $X - X^{-i}$ in terms of a matrix telescoping series, the second of which will be bounding individual terms, and the final step will prove the two results.

Step 1: A useful expansion

We have that

$$X = (U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} W U$$
$$X^{-i} = (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-2} (I - e_i e_i^\top) W (I - e_i e_i^\top) U.$$

Define the matrix $\xi := W - (I - e_i e_i^\top)W(I - e_i e_i^\top)$. We note that

$$\begin{aligned}
X - X^{-i} &= (U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} WU - (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-2} (I - e_i e_i^\top) W (I - e_i e_i^\top) U \\
&= (U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} WU - (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-2} (W - \xi) U \\
&= \Big[(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} - (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-2} \Big] WU \\
&\quad + (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-2} \xi U \\
&= \Big[(U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-2} - (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-3} (U_\perp U_\perp^\top (W - \xi) U_\perp U_\perp^\top) \Big] WU \\
&\quad + (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-2} \xi U \\
&= \Big[\Big((U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-3} - (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-3} \Big) (U_\perp U_\perp^\top W U_\perp U_\perp^\top) \Big] WU \\
&\quad + (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-3} (U_\perp U_\perp^\top \xi U_\perp U_\perp^\top) WU \\
&\quad + (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-2} \xi U.
\end{aligned}$$

Note that if $p = 2$ then we simply have ξU. Define the matrices

$$\begin{aligned}
S_\xi &:= U_\perp U_\perp^\top \xi U_\perp U_\perp^\top \\
S_{-i} &:= (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top) \\
S &:= U_\perp U_\perp^\top W U_\perp U_\perp^\top.
\end{aligned}$$

Then iterating the process above, it holds that

$$\begin{aligned}
X - X^{-i} &= \sum_{m=1}^{p-2} (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{m-1} (U_\perp U_\perp^\top \xi U_\perp U_\perp^\top) (U_\perp U_\perp^\top W U_\perp U_\perp^\top)^{p-m-2} WU \\
&\quad + (U_\perp U_\perp^\top (I - e_i e_i^\top) W (I - e_i e_i^\top) U_\perp U_\perp^\top)^{p-2} \xi U \\
&= \sum_{m=1}^{p-2} S_{-i}^{m-1} S_\xi S^{p-m-2} WU + S_{-i}^{p-2} \xi U,
\end{aligned}$$

where the sum is understood to be the empty sum if $p = 2$.

Step 2: Analyzing each term in the sum

We now analyze each individual term in the sum on the event $\mathcal{B}$, where we also analyze each row of $W(X - X^{-i})$. Let the matrix $B \in \{\Gamma(EE^\top), I\}$ be fixed. We ignore the boundary term $BS_{-i}^{p-2}\xi U$ for now. Note that the k'th row of such term can be written as

$$e_k^\top B S_{-i}^{m-1} S_\xi S^{p-m-2} W U,$$

Recall that

$$\begin{aligned}
S_\xi &= U_\perp U_\perp^\top \xi U_\perp U_\perp^\top \\
&= U_\perp U_\perp^\top \Big(e_i e_i^\top W + W e_i e_i^\top - e_i e_i^\top W e_i e_i^\top \Big) U_\perp U_\perp^\top \\
&= U_\perp U_\perp^\top e_i e_i^\top W U_\perp U_\perp^\top + U_\perp U_\perp^\top W e_i e_i^\top U_\perp U_\perp^\top - U_\perp U_\perp^\top e_i e_i^\top W e_i e_i^\top U_\perp U_\perp^\top.
\end{aligned}$$

Therefore, by homogeneity of the vector norm, we have that

$$
\begin{aligned}
&\|e_k^\top B S_{-i}^{m-1} S_\xi S^{p-m-2} WU\| \\
&\quad \le \|e_k^\top B S_{-i}^{m-1} U_\perp U_\perp^\top e_i e_i^\top W U_\perp U_\perp^\top S^{p-m-2} WU\| + \|e_k^\top B S_{-i}^{m-1} U_\perp U_\perp^\top W e_i e_i^\top U_\perp U_\perp^\top S^{p-m-2} WU\| \\
&\qquad + \|e_k^\top B S_{-i}^{m-1} U_\perp U_\perp^\top e_i e_i^\top W e_i e_i^\top U_\perp U_\perp^\top S^{p-m-2} WU\| \\
&\quad \le \|e_k^\top B S_{-i}^{m-1} U_\perp U_\perp^\top e_i\| \|e_i^\top W U_\perp U_\perp^\top S^{p-m-2} WU\| + \|e_k^\top B S_{-i}^{m-1} U_\perp U_\perp^\top W e_i\| \|e_i^\top U_\perp U_\perp^\top S^{p-m-2} WU\| \\
&\qquad + \|e_k^\top B S_{-i}^{m-1} U_\perp U_\perp^\top e_i\| \|e_i^\top W e_i\| \|e_i^\top U_\perp U_\perp^\top S^{p-m-2} WU\| \\
&\quad \le \|e_k^\top B S_{-i}^{m-1} U_\perp U_\perp^\top e_i\| \Big(\|e_i^\top U_\perp U_\perp^\top W U_\perp U_\perp^\top S^{p-m-2} WU\| + \|e_i^\top U U^\top W U_\perp U_\perp^\top S^{p-m-2} WU\| \Big) \\
&\qquad + \|e_k^\top B S_{-i}^{m-1} U_\perp U_\perp^\top W e_i\| \|e_i^\top U_\perp U_\perp^\top S^{p-m-2} WU\| \\
&\qquad + \|e_k^\top B S_{-i}^{m-1} U_\perp U_\perp^\top e_i\| \|e_i^\top W e_i\| \|e_i^\top U_\perp U_\perp^\top S^{p-m-2} WU\| \\
&\quad \le \|B S_{-i}^{m-1}\| \|U_\perp U_\perp^\top W U_\perp U_\perp^\top S^{p-m-2} WU\|_{2,\infty} + \|B S_{-i}^{m-1}\| \|U\|_{2,\infty} \|W U_\perp U_\perp^\top S^{p-m-2} WU\| \\
&\qquad + \|B S_{-i}^{m-1} U_\perp U_\perp^\top W\| \|U_\perp U_\perp^\top S^{p-m-2} WU\|_{2,\infty} + \|B S_{-i}^{m-1}\| \|W\| \|U_\perp U_\perp^\top S^{p-m-2} WU\|_{2,\infty} \\
&\quad \le \|B S_{-i}^{m-1}\| \|S^{p-m-1} WU\|_{2,\infty} + \|U\|_{2,\infty} \|B S_{-i}^{m-1}\| \|W\|^{p-m} + \|B S_{-i}^{m-1}\| \|W\| \|U_\perp U_\perp^\top S^{p-m-2} WU\|_{2,\infty} \\
&\qquad + \|B S_{-i}^{m-1}\| \|W\| \|U_\perp U_\perp^\top S^{p-m-2} WU\|_{2,\infty} \\
&\quad \le \|B S_{-i}^{m-1}\| \Big(C_1 (C_2 \delta)^{p-m} \|U\|_{2,\infty} + \|U\|_{2,\infty} \|W\|^{p-m} + 2\|W\| C_1 (C_2\delta)^{p-m-1} \|U\|_{2,\infty} \Big) \\
&\quad \le \|B S_{-i}^{m-1}\| \|U\|_{2,\infty} \Big(C_1 (C_2\delta)^{p-m} + \Big[\frac{\delta}{\sqrt{r \log(n \vee d)}} \Big]^{p-m} + 2 \frac{\delta}{\sqrt{r \log(n \vee d)}} C_1 (C_2 \delta)^{p-m-1} \Big).
\end{aligned}
$$

But for $\|BS_{-i}^{m-1}\|$, this bound is uniform in i and j. Finally, for the boundary term, we note that the same strategy can be applied in precisely the same manner, yielding the same bound.

Step 3: Putting it together

First we show the upper bound on I_2^{ij} on $\mathcal{B}$. Recall that

$$
I_2^{ij} = e_i^\top \Gamma(EE^\top)(X - X^{-i}) e_j.
$$

By the bounds on each term, we have that

$$
\begin{aligned}
&|e_i^\top \Gamma(EE^\top)(X - X^{-i})e_j| \\
&\quad \le \|e_i^\top \Gamma(EE^\top)(X - X^{-i})\| \\
&\quad \le \sum_{m=1}^{p-2} \|\Gamma(EE^\top)\|\|W\|^{m-1}\|U\|_{2,\infty}\left(C_1(C_2\delta)^{p-m} + \left[\frac{\delta}{\sqrt{r\log(n\vee d)}}\right]^{p-m} + 2\frac{\delta}{\sqrt{r\log(n\vee d)}}C_1(C_2\delta)^{p-m-1}\right) \\
&\quad \le 4C_1\|U\|_{2,\infty}\sum_{m=1}^{p-2}\|\Gamma(EE^\top)\|\|W\|^{m-1}(C_2\delta)^{p-m} \\
&\quad \le 4C_1\|U\|_{2,\infty}\sum_{m=1}^{p-2}\left(\frac{\delta}{\sqrt{r\log(n\vee d)}}\right)^m(C_2\delta)^{p-m} \\
&\quad \le \frac{4C_1\|U\|_{2,\infty}}{\sqrt{r\log(n\vee d)}}(C_2\delta)^p\sum_{m=1}^{p-2}C_2^{-m} \\
&\quad \le \frac{C_1}{4\sqrt{r}}(C_2\delta)^p\|U\|_{2,\infty}
\end{aligned}
$$

as long as $C_2 \ge 48$.

Next we show that $\mathcal{B}\cap\mathcal{A}_{ij}^c$ is empty for all i and j. More specifically, we show that

$$|X_{kj}^{-i} - X_{kj}| \le C_1(C_2\delta)^{p-1}\|U\|_{2,\infty},$$

Again upper bounding the k,j entry by the k'th row norm, we note that the k'th row of $X - X^{-i}$ can be written as $e_k^\top(X - X^{-i})$. Using the expansion and the bounds on each term in the summation, we have that

$$
\begin{aligned}
\|e_k^\top(X - X^{-i})\| &\le \sum_{m=1}^{p-1}\|W\|^{m-1}\|U\|_{2,\infty}\left(C_1(C_2\delta)^{p-m} + \left[\frac{\delta}{\sqrt{r\log(n\vee d)}}\right]^{p-m} + \frac{\delta}{\sqrt{r\log(n\vee d)}}C_1(C_2\delta)^{p-m-1}\right) \\
&\le 4C_1\|U\|_{2,\infty}(C_2\delta)^{p-1}\sum_{m=1}C_2^{-m} \\
&\le C_1\|U\|_{2,\infty}(C_2\delta)^{p-1},
\end{aligned}
$$

and hence on $\mathcal{B}$ we have that

$$\begin{aligned}\|e_k^\top X^{-i} e_j\| &\leq \|X\|_{2,\infty} + \|e_k^\top (X - X^{-i})\| \\ &\leq 2C_1(C_2\delta)^{p-1}\|U\|_{2,\infty}\end{aligned}$$

as desired. Therefore $\mathcal{B} \cap \mathcal{A}_{ij}^c$ is empty.

□

B.6 Proof of Lemmas in Section B.2

B.6.1 Proof of Lemma 2

We will need the following lemma, adapted from Zhang et al. (2022).

Lemma 22 (Lemma 1 in Zhang et al. (2022)). *Let $U, V \in \mathbb{O}(n, r)$ and let A be a fixed $n \times n$ matrix. Then*

$$\begin{aligned}\|G(UU^\top A)\| &\leq \|U\|_{2,\infty}\|A\| \\ \|G(AVV^\top)\| &\leq \|V\|_{2,\infty}\|A\|.\end{aligned}$$

Proof of Lemma 2. First, if $T_0 \geq C\log\left(\frac{\lambda_r^2}{\|\Gamma(Z)\|}\right)$, then by Zhang et al. (2021) it holds that $\|N_T - A\| \leq 3\|\Gamma(Z)\|$. In addition, supposing the result holds and that $\rho \leq \frac{1}{2}$, then when $T - T_0 \geq C\log\left(\frac{1}{\|U\|_{2,\infty}}\right)$ it holds that

$$\|\widehat{A} - \widetilde{A}\| \leq 41\|U\|_{2,\infty}\|\Gamma(Z)\|.$$

Hence, we must have that

$$\begin{aligned}T &\geq C\log\left(\frac{1}{\|U\|_{2,\infty}}\right) + T_0 \\ &\geq C\left(\log\left(\frac{1}{\|U\|_{2,\infty}}\right) + \log\left(\frac{\lambda_r^2}{\|\Gamma(Z)\|}\right)\right) \\ &\geq C\left(\log\left(\frac{\lambda_r^2}{\|U\|_{2,\infty}\|\Gamma(Z)\|}\right)\right).\end{aligned}$$

This proves the "consequently" part. We now show that

$$\widetilde{K}_T \leq \frac{4}{\rho^{T-T_0}}\|\Gamma(Z)\| + \frac{20}{1-\rho}\|U\|_{2,\infty}\|\Gamma(Z)\|.$$

We have that

$$\begin{aligned}
\|N_T - \widetilde{A}\| &= \|G(N_T - \widetilde{A})\| \\
&\leq \left\|G\left(\left(P_{U^{T-1}} - P_{\widetilde{U}}\right)\left(N_{T-1} - \widetilde{A}\right)\right)\right\| + \left\|G\left(P_{\widetilde{U}}\left(N_{T-1} - \widetilde{A}\right)\right)\right\| + \left\|G\left(P_{U_\perp^{T-1}}\widetilde{A}\right)\right\| \\
&\leq \left\|G\left(\left(P_{U^{T-1}} - P_{\widetilde{U}}\right)\left(N_{T-1} - \widetilde{A}\right)\right)\right\| + \left\|G\left(P_{\widetilde{U}}\left(N_{T-1} - \widetilde{A}\right)\right)\right\| + \left\|G\left(P_{U_\perp^{T-1}}A\right)\right\| + \|G(P_{U_\perp^{T-1}}\Gamma(Z))\| \\
&\leq \left\|G\left(\left(P_{U^{T-1}} - P_{\widetilde{U}}\right)\left(N_{T-1} - \widetilde{A}\right)\right)\right\| + \left\|G\left(P_{\widetilde{U}}\left(N_{T-1} - \widetilde{A}\right)\right)\right\| + \left\|G\left(P_{U_\perp^{T-1}}A\right)\right\| \\
&\quad + \|G(P_{U_\perp^{T-1}} - P_{\widetilde{U}_\perp})\Gamma(Z))\| + \|G(P_{\widetilde{U}_\perp}\Gamma(Z))\| \\
&= J_1 + J_2 + J_3 + J_4 + J_5.
\end{aligned}$$

we now bound each term.

The term J_1: We use the restricted-rank operator norm of G to bound this term, since $\operatorname{rank}(P_{U^{T-1}} - P_{\widetilde{U}}) \leq 2r$:

$$\|G((P_{U^{T-1}} - P_{\widetilde{U}})(N_{T-1} - \widetilde{A}))\| \leq \|P_{U^{T-1}} - P_{\widetilde{U}}\|\|N_{T-1} - \widetilde{A}\|.$$

The term J_2: By Lemma 22,

$$\|G(P_{\widetilde{U}}(N_{T-1} - \widetilde{A}))\| \leq \|\widetilde{U}\|_{2,\infty}\|N_{T-1} - \widetilde{A}\|.$$

The term J_3: By Lemmas 22 and 23,

$$\|G(P_{U_\perp^{T-1}}A)\| = \|G(P_{U_\perp^{T-1}}AP_U)\| \leq \|U\|_{2,\infty}\|P_{U_\perp^{T-1}}A\| \leq 2\|U\|_{2,\infty}\|N_{T-1} - A\|.$$

The term J_4: Since $P_{U_\perp^{T-1}} - P_{\widetilde{U}_\perp} = (I - P_{\widetilde{U}_\perp}) - (I - P_{U_\perp^{T-1}}) = P_{\widetilde{U}} - P_{U^{T-1}}$, we proceed as we did for J_1, obtaining

$$\|G((P_{U_\perp^{T-1}} - P_{\widetilde{U}_\perp})\Gamma(Z)\| \leq \|P_{U^{T-1}} - P_{\widetilde{U}}\|\|\Gamma(Z)\|.$$

The term J_5: We have $G(P_{\widetilde{U}_\perp}\Gamma(Z)) = G(\Gamma(Z)) - G(P_{\widetilde{U}}\Gamma(Z)) = -G(P_{\widetilde{U}}\Gamma(Z))$. By Lemma 22,

$$\|G(P_{\widetilde{U}_\perp}\Gamma(Z))\| \leq \|\widetilde{U}\|_{2,\infty}\|\Gamma(Z)\|.$$

Putting it together: Let $K_{T-1} := \|N_{T-1} - A\|$ and let $\widetilde{K}_{T-1} := \|N_{T-1} - \widetilde{A}\|$. Compiling these bounds, we have that

$$\begin{aligned}
J_1 &\leq \|P_{U^{T-1}} - P_{\widetilde{U}}\|\widetilde{K}_{T-1} \\
J_2 &\leq \|\widetilde{U}\|_{2.\infty}\widetilde{K}_{T-1} \\
J_3 &\leq 2\|U\|_{2,\infty}K_{T-1} \\
J_4 &\leq \|P_{U^{T-1}} - P_{\widetilde{U}}\|\|\Gamma(Z)\| \\
J_5 &\leq \|\widetilde{U}\|_{2,\infty}\|\Gamma(Z)\|.
\end{aligned}$$

These bounds hold regardless of T. Hence, we may take T such that $\|N_{T-1} - A\| \leq 3\|\Gamma(Z)\|$ (by Zhang et al. (2021), we may take $T \geq C\log\left(\frac{\lambda_r^2}{\|\Gamma(Z)\|}\right)$.The proof of Lemma 16 shows that $\|\Gamma(Z)\| \leq \lambda_r^2/12$, so $\widetilde{K}_{T-1} \leq 4(\lambda_r^2/12) \leq \lambda_r^2/2$, and by the Davis-Kahan theorem, we have that

$$\|P_{U^{T-1}} - P_{\widetilde{U}}\| \leq 2\frac{\widetilde{K}_{T-1}}{\lambda_r^2}.$$

Applying this to the above bounds, we see that

$$J_1 \leq 2\frac{\widetilde{K}_{T-1}^2}{\lambda_r^2}$$
$$J_4 \leq 2\frac{\widetilde{K}_{T-1}}{\lambda_r^2}\|\Gamma(Z)\|.$$

Moreover, we have the trivial bound

$$K_{T-1} \leq \widetilde{K}_{T-1} + \|\Gamma(Z)\|.$$

Hence, we see that

$$\begin{aligned}
J_1 &\leq 2\frac{\widetilde{K}_{T-1}^2}{\lambda_r^2} \\
J_2 &\leq \|\widetilde{U}\|_{2.\infty}\widetilde{K}_{T-1} \\
J_3 &\leq 2\|U\|_{2,\infty}\left(\widetilde{K}_{T-1} + \|\Gamma(Z)\|\right) \\
J_4 &\leq 2\frac{\widetilde{K}_{T-1}}{\lambda_r^2}\|\Gamma(Z)\| \\
J_5 &\leq \|\widetilde{U}\|_{2,\infty}\|\Gamma(Z)\|.
\end{aligned}$$

Now, for T_0 such that $K_{T_0} \leq 3\|\Gamma(Z)\|$, $\widetilde{K}_{T_0} \leq 4\|\Gamma(Z)\|$, and we see that we have the initial bound

$$\widetilde{K}_{T_0+1} \leq 40\frac{\|\Gamma(Z)\|^2}{\lambda_r^2} + 5\|\widetilde{U}\|_{2,\infty}\|\Gamma(Z)\| + 10\|U\|_{2,\infty}\|\Gamma(Z)\|.$$

On the event in Theorem 7 and Assumption 2.2, once n and d are large enough, $\|\widetilde{U} - UU^\top\widetilde{U}\|_{2,\infty} \leq \|U\|_{2,\infty}$, so

$$\|\widetilde{U}\|_{2,\infty} \leq \|\widetilde{U} - UU^\top\widetilde{U}\|_{2,\infty} + \|UU^\top\widetilde{U}\|_{2,\infty} \leq 2\|U\|_{2,\infty},$$

since $\|U^\top\widetilde{U}\| \leq 1$. This gives

$$\widetilde{K}_{T_0+1} \leq 40\frac{\|\Gamma(Z)\|^2}{\lambda_r^2} + 20\|U\|_{2,\infty}\|\Gamma(Z)\|.$$

Let $\rho = 10\|\Gamma(Z)\|/\lambda_r^2$, and suppose that for $T-1 \geq T_0$ we have the bound

$$\widetilde{K}_{T-1} \leq 4\rho^{T-1-T_0}\|\Gamma(Z)\| + \frac{20}{1-\rho}\|U\|_{2,\infty}\|\Gamma(Z)\|.$$

Clearly for T_0 we have this bound. With the recursion, and using $\widetilde{K}_{T-1} \leq 4\|\Gamma(Z)\|$ and $\|\widetilde{U}\|_{2,\infty} \leq 2\|U\|_{2,\infty}$, we get

$$\begin{aligned}
\widetilde{K}_T &\leq 10\frac{\widetilde{K}_{T-1}}{\lambda_r^2}\|\Gamma(Z)\| + 20\|U\|_{2,\infty}\|\Gamma(Z)\| \\
&\leq \frac{10}{\lambda_r^2}\left[4\rho^{T-1-T_0}\|\Gamma(Z)\| + \frac{20}{1-\rho}\|U\|_{2,\infty}\|\Gamma(Z)\|\right]\|\Gamma(Z)\| + 20\|U\|_{2,\infty}\|\Gamma(Z)\| \\
&\leq 4\rho^{T-T_0}\|\Gamma(Z)\| + \left(1 + \frac{\rho}{1-\rho}\right)20\|U\|_{2,\infty}\|\Gamma(Z)\| \\
&= 4\rho^{T-T_0}\|\Gamma(Z)\| + \frac{20}{1-\rho}\|U\|_{2,\infty}\|\Gamma(Z)\|
\end{aligned}$$

as required. □

B.7 Proof of Lemmas in Section B.3

Recall Lemma 17.

Lemma 17. *There exist universal constants C_6 and C_7 such that the residual terms R_1,R_2, and R_3 satisfy, uniformly over i and j,*

$$\frac{1}{\sigma_{ij}}\Big(|R_1| + |R_2| + |R_3|\Big) \leq C_6\kappa_\sigma\kappa^2\mu_0\sqrt{\frac{r\log(n\vee d)}{n}} + C_7\kappa^3\kappa_\sigma\mu_0\frac{r\log(n\vee d)}{\mathrm{SNR}}$$

with probability at least $1 - 5(n\vee d)^{-4}$.

Proof of Lemma 17. Recall the definitions of R_1, R_2 and R_3 via

$$\begin{aligned}
R_1 &:= e_i^\top UU^\top\Big(EM^\top + \Gamma(EE^\top)\Big)U\Lambda^{-2}e_j; \\
R_2 &:= e_i^\top\Big(\widetilde{U}_D\widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\Big)Ue_j; \\
R_3 &:= e_i^\top\sum_{k\geq 2}S_{MM^\top,k}(W)Ue_j.
\end{aligned}$$

We analyze each in turn.

The term R_1: We will split this into two terms, R_{11} and R_{12}. From the identity $M^\top U = V\Lambda$, the i, j entry of R_1 can be written as

$$\frac{1}{\lambda_j} e_i^\top U U^\top E V_{\cdot j} + \frac{1}{\lambda_j^2} \sum_{k \neq l} (UU^\top)_{ik} U_{lj} \langle E_k, E_l \rangle := R_{11} + R_{12},$$

where

$$R_{11} := \frac{1}{\lambda_j} \sum_{k=1}^n (UU^\top)_{ik} \langle E_k, V_{\cdot j} \rangle;$$
$$R_{12} := \frac{1}{\lambda_j^2} \sum_{k \neq l} (UU^\top)_{ik} U_{lj} \langle E_k, E_l \rangle.$$

Dividing R_{11} by σ_{ij} reveals it is of the form

$$\frac{1}{\|\Sigma_i^{1/2} V_{\cdot j}\|} \sum_{k=1}^n (UU^\top)_{ik} \langle E_k, V_{\cdot j} \rangle.$$

To calculate an upper bound, we need to calculate the ψ_2 norm squared:

$$\begin{aligned} \frac{1}{\|\Sigma_i^{1/2} V_{\cdot j}\|^2} \sum_{k=1}^n (UU^\top)_{ik}^2 V_{\cdot j}^\top \Sigma_k V_{\cdot j} &\leq \frac{\sigma^2}{\|\Sigma_i^{1/2} V_{\cdot j}\|^2} \sum_{k=1}^n (UU^\top)_{ik}^2 \\ &\leq \kappa_\sigma^2 \|U\|_{2,\infty}^2 \end{aligned}$$

which by Hoeffding's inequality shows that this is less than $\widetilde{C}_{R_1} \kappa_\sigma \|U\|_{2,\infty} t$ with probability at least $1 - 2\exp(-ct^2)$. Hence, we obtain the bound $C_{R_1} \kappa_\sigma \|U\|_{2,\infty} \sqrt{\log(n \vee d)}$ with probability at least $1 - 2(n \vee d)^{-4}$.

We now analyze R_{12}. Note that

$$R_{12} := \frac{1}{\lambda_j^2} \sum_{k \neq l} (UU^\top)_{ik} U_{lj} \langle E_k, E_l \rangle$$

resembles the random variable in the Hanson-Wright inequality (e.g. Vershynin (2020); Chen and Yang (2020)). By the generalized Hanson-Wright inequality (e.g. Exercise 6.2.7

in Vershynin (2018)), we have that

$$\mathbb{P}\Big\{\Big|\sum_{k\neq l} B_{kl}\langle E_k, E_l\rangle\Big| \geq t\Big\} \leq 2\exp\Big[-c\min\Big(\frac{t^2}{\sigma^4 d\|B\|_F^2}, \frac{t}{\sigma^2\|B\|}\Big)\Big]$$

where $B_{kl} := (UU^\top)_{il}U_{kj}$. Note that its Frobenius norm can be evaluated via

$$\begin{aligned}\|B\|_F^2 &= \sum_{k\neq l}(UU^\top)_{il}^2 U_{kj}^2 \\ &\leq \sum_{l=1}^n (UU^\top)_{il}^2 \sum_{k=1}^n U_{kj}^2 \\ &\leq \sum_{l=1}^n (UU^\top)_{il}^2 \\ &\leq \|U\|_{2,\infty}^2.\end{aligned}$$

Similarly,

$$\begin{aligned}\|B\| :&= \sup_{\|x\|=1,\|y\|=1}\sum_{k=1}^n x_k(UU^\top)_{ik}\sum_{l=1}^n U_{lj}y_l \\ &\leq \sup_{\|x\|=1}\sum_{k=1}^n x_k(UU^\top)_{ik} \\ &= \|(UU^\top)_i\|_2 \\ &\leq \|U\|_{2,\infty}.\end{aligned}$$

Therefore,

$$\mathbb{P}\Big\{\Big|\sum_{k\neq l} B_{kl}\langle E_k E_l\rangle\Big| \geq t\Big\} \leq 2\exp\Big[-c\min\Big(\frac{t^2}{\sigma^4 d\|U\|_{2,\infty}^2}, \frac{t}{\sigma^2\|U\|_{2,\infty}}\Big)\Big].$$

Taking $t = s\sqrt{d}\|U\|_{2,\infty}\sigma^2$ shows that

$$\mathbb{P}\Big\{\Big|\sum_{k\neq l} B_{kl}\langle E_k E_l\rangle\Big| \geq s\sqrt{d}\|U\|_{2,\infty}\sigma^2\Big\} \leq 2\exp\Big(-c\min\Big(s^2, s\sqrt{d}\Big)\Big),$$

and hence taking $s = \frac{1}{\sqrt{c}}\sqrt{4\log(n \vee d)}$ we see that with probability at least $1 - 2(n \vee d)^{-4}$

$$\left| \frac{1}{\lambda_j^2} \sum_{k \neq l} B_{kl} \langle E_k E_l \rangle \right| \leq C_{R_1} \frac{\sigma^2}{\lambda_j^2} \|U\|_{2,\infty} \sqrt{d \log(n \vee d)}.$$

Dividing by σ_{ij} reveals that with this same probability,

$$\begin{aligned} |R_{22}/\sigma_{ij}| &\leq C_{R_1} \kappa_\sigma \kappa \frac{\sigma\sqrt{d}}{\lambda_j} \|U\|_{2,\infty} \sqrt{\log(n \vee d)} \\ &\leq C_{R_1} \kappa_\sigma \kappa \mu_0 \frac{\sigma\sqrt{rd}}{\lambda_r} \sqrt{\frac{\log(n \vee d)}{n}} \\ &\leq C_{R_1} \kappa_\sigma \kappa \mu_0 \frac{1}{\mathrm{SNR}} \sqrt{\frac{\log(n \vee d)}{n}}. \end{aligned}$$

The term R_2: Note that $\widetilde{U}_D$ and $\widetilde{U}$ were already analyzed in Lemma 13, which shows that

$$\begin{aligned} \|\widetilde{U}_D \widetilde{U}_D^\top - \widetilde{U}\widetilde{U}^\top\| &\leq \frac{C_{R_2} \lambda_1 \sigma \sqrt{\log(n \vee d)}}{\lambda_r^2} \|U\|_{2,\infty} \\ &\leq C_{R_2} \|U\|_{2,\infty} \kappa \frac{\sigma}{\lambda_r} \sqrt{\log(n \vee d)} \end{aligned}$$

with probability at least $1 - 2(n \vee d)^{-4}$. Multiplying by $\frac{1}{\sigma_{ij}}$ yields the upper bound

$$C_{R_2} \kappa_\sigma \|U\|_{2,\infty} \kappa^2 \sqrt{\log(n \vee d)} \leq C_{R_2} \kappa_\sigma \mu_0 \kappa^2 \sqrt{\frac{r \log(n \vee d)}{n}}.$$

with probability at least $1 - 2(n \vee d)^{-4}$.

The term R_3: First, note that

$$\begin{aligned} \left| e_i^\top \sum_{k \geq 2} S_{MM^\top, k}(W) U e_j \right| &\leq \| \sum_{k \geq 2} S_{MM^\top, k}(W) U \|_{2,\infty} \\ &\leq \sum_{k \geq 2} \| S_{MM^\top, k}(W) U \|_{2,\infty}. \end{aligned}$$

Examining the proof of Theorem 7 shows that the exact same result holds, only now we start the summation at $k = 2$. Consequently, using the definition of δ as in the proof of

Theorem 7, by Lemma 15 we have that with probability $1 - c\log(n \vee d)(n \vee d)^{-5}$ that

$$\sum_{k=2}^{\infty} \|S_{MM^\top,k}(W)U\|_{2,\infty} \leq \|U\|_{2,\infty} \sum_{k=2}^{c\log(n\vee d)} C_1 \left(\frac{4C_2\delta}{\lambda_r^2}\right)^k + \sum_{k=c\log(n\vee d)}^{\infty} \left(\frac{4\delta}{\lambda_r^2}\right)^k$$
$$\leq C_{R_3}\|U\|_{2,\infty}\frac{\delta^2}{\lambda_r^4}.$$

Hence, with probability at least $1 - (n \vee d)^{-4}$,

$$\begin{aligned}
\frac{1}{\sigma_{ij}}|R_3| &\leq C_{R_3}\|U\|_{2,\infty}\frac{\delta^2}{\lambda_r^4}\frac{1}{\sigma_{ij}} \\
&\leq C_{R_3}\|U\|_{2,\infty}\frac{1}{\sigma_{ij}\lambda_r^4}\left(\sqrt{rnd}\log(\max(n,d))\sigma^2 + \sqrt{rn\log(n)}\lambda_1\sigma\right)^2 \\
&\leq C_{R_3}\|U\|_{2,\infty}\frac{rnd\log^2(d)\sigma^4 + rn\log(n\vee d)\lambda_1^2\sigma^2 + rn\sqrt{d}\log^{3/2}(n)\sigma^3}{\sigma_{ij}\lambda_r^4} \\
&\leq C_{R_3}\kappa\kappa_\sigma\|U\|_{2,\infty}\frac{rnd\log^2(d)\sigma^3 + rn\log(n\vee d)\lambda_1^2\sigma + rn\sqrt{d}\log^{3/2}(n)\sigma^2}{\lambda_r^3} \\
&\leq C_{R_3}\kappa\kappa_\sigma\|U\|_{2,\infty}\left(\frac{rnd\log^2(n)\sigma^3 + rn\sqrt{d}\log^{3/2}(n)\sigma^2}{\lambda_r^3} + \frac{rn\log(n\vee d)\kappa^2\sigma}{\lambda_r}\right) \\
&\leq C_{R_3}\kappa\kappa_\sigma\mu_0\left(\frac{r^{3/2}\sqrt{n}d\sigma^3}{\lambda_r^3}\log^2(n\vee d) + \frac{r^{3/2}\sqrt{nd}\sigma^2}{\lambda_r^3}\log^{3/2}(n\vee d) + \frac{r^{3/2}\sqrt{n}\kappa^2\sigma}{\lambda_r}\log(n\vee d)\right) \\
&\leq C_{R_3}\kappa\kappa_\sigma\mu_0\left(\frac{\log^2(n)}{\mathrm{SNR}^3} + \frac{\sqrt{r}}{\lambda_r}\frac{\log^{3/2}(n\vee d)}{\mathrm{SNR}^2} + \frac{\kappa^2 r\log(n\vee d)}{\mathrm{SNR}}\right) \\
&\leq C_{R_3}\kappa^3\kappa_\sigma\mu_0\frac{r\log(n\vee d)}{\mathrm{SNR}},
\end{aligned}$$

where we have absorbed extra constants into C_{R_3} since $\mathrm{SNR} \geq \kappa\sqrt{\log(n\vee d)}$ by Assumption 2.2. Therefore, summing up the probabilities and absorbing the constants, we see that with probability at least $1 - 5(n\vee d)^{-4}$ that

$$\begin{aligned}
\frac{1}{\sigma_{ij}}\left(|R_1| + |R_2| + |R_3|\right) &\leq C_{R_1}\kappa_\sigma\kappa\mu_0\frac{1}{\mathrm{SNR}}\sqrt{\frac{\log(n\vee d)}{n}} + C_{R_2}\kappa_\sigma\mu_0\kappa^2\sqrt{\frac{r\log(n\vee d)}{n}} \\
&\quad + C_{R_3}\kappa^3\kappa_\sigma\mu_0\frac{r\log(n\vee d)}{\mathrm{SNR}} \\
&\leq C_6\kappa_\sigma\kappa^2\mu_0\sqrt{\frac{r\log(n\vee d)}{n}} + C_7\kappa^3\kappa_\sigma\mu_0\frac{r\log(n\vee d)}{\mathrm{SNR}}
\end{aligned}$$

as required. □

Lemma 18. *On the intersection of the events in Theorem 6 and Lemma 1 the residual terms*

R_4 and R_5 satisfy for all i and j,

$$\frac{1}{\sigma_{ij}}\Big(|R_4| + |R_5|\Big) \le C_8\kappa^3\kappa_\sigma\mu_0\frac{1}{\mathrm{SNR}} + C_9\kappa^4\kappa_\sigma\mu_0^2\frac{r}{\sqrt{n}}.$$

for some universal constants C_8 and C_9.

Proof of Lemma 18. Recall the definitions of R_4 and R_5 via

$$\begin{aligned} R_4 &:= e_i^\top \widehat{U}(\mathcal{O}_* - \widehat{U}^\top\widetilde{U}\widetilde{U}^\top U)e_j; \\ R_5 &:= e_i^\top\Big(\widehat{U}\widehat{U}^\top\widetilde{U}\widetilde{U}^\top U - \widetilde{U}\widetilde{U}^\top U\Big)e_j. \end{aligned}$$

On the event in Theorem 6, $\|\widehat{U}\|_{2,\infty} \le \|U\|_{2,\infty} + \inf_{\mathcal{O}_*}\|\widehat{U} - U\mathcal{O}_*\|_{2,\infty} \le C\|U\|_{2,\infty}$ by Assumption 2.2. Therefore, the term R_4 can be bounded in a similar manner to the proof of Lemma 3 (see Appendix B.8) via

$$\begin{aligned} \Big|e_i^\top \widehat{U}(\mathcal{O}_* - \widehat{U}^\top\widetilde{U}\widetilde{U}^\top U)e_j\Big| &\le \|\widehat{U}\|_{2,\infty}\|\mathcal{O}_* - \widehat{U}^\top\widetilde{U}\widetilde{U}^\top U\| \\ &\le C\|U\|_{2,\infty}\Big(\|\sin(\widehat{U},\widetilde{U})\|^2 + \|\sin(\widetilde{U},U)\|^2\Big) \\ &\le C\|U\|_{2,\infty}\frac{\|\Gamma(Z)\|^2}{\lambda_r^4}, \end{aligned}$$

where the final inequality is by the Davis-Kahan Theorem and Lemmas 1 and 2. On the event in Lemma 1, the numerator can be bounded by

$$C_{\mathrm{spectral}}^2\Big(\sigma^2(n+\sqrt{nd}) + \sigma\lambda_1\sqrt{n}\Big)^2.$$

Consequently, there exists a universal constant C_{R_4} such that

$$\begin{aligned}
\frac{1}{\sigma_{ij}}|R_4| &\leq \frac{1}{\sigma_{ij}} C_{R_4} \|U\|_{2,\infty} \frac{\sigma^3 n\sqrt{d} + \sigma^2\lambda_1^2 n + \sigma^4 nd}{\lambda_r^4} \\
&= C_{R_4} \frac{\lambda_j}{\|\Sigma_i^{1/2} V_{\cdot j}\|} \|U\|_{2,\infty} \frac{\sigma^3 n\sqrt{d} + \sigma^2\lambda_1^2 n + \sigma^4 nd}{\lambda_r^4} \\
&\leq C_{R_4}\kappa\kappa_\sigma \|U\|_{2,\infty} \frac{\sigma^2 n\sqrt{d} + \sigma\lambda_1^2 n + \sigma^3 nd}{\lambda_r^3} \\
&\leq C_{R_4}\kappa\kappa_\sigma\mu_0 \frac{\sigma^2\sqrt{rnd} + \sigma\lambda_1^2\sqrt{rn} + \sigma^3\sqrt{rn}d}{\lambda_r^3} \\
&\leq C_{R_4}\kappa\kappa_\sigma\mu_0 \left(\frac{1}{\lambda_r \mathrm{SNR}^2} + \frac{\kappa^2}{\mathrm{SNR}} + \frac{1}{\mathrm{SNR}^3}\right) \\
&\leq C_8 \kappa^3 \kappa_\sigma \mu_0 \frac{1}{\mathrm{SNR}}.
\end{aligned}$$

The term R_5 satisfies

$$\begin{aligned}
\frac{1}{\sigma_{ij}} \left| e_i^\top \left(\widehat{U}\widehat{U}^\top \widetilde{U}\widetilde{U}^\top U - \widetilde{U}\widetilde{U}^\top U \right) e_j \right| &\leq \frac{1}{\sigma_{ij}} \|\widehat{U}\widehat{U}^\top \widetilde{U}\widetilde{U}^\top U - \widetilde{U}\widetilde{U}^\top U\|_{2,\infty} \\
&\leq \frac{1}{\sigma_{ij}} \|\widehat{U}\widehat{U}^\top \widetilde{U} - \widetilde{U}\|_{2,\infty}.
\end{aligned}$$

The definition of $\widetilde{H}$ in the proof of Theorem 8 shows that $\widetilde{H}^\top = \widehat{U}^\top \widetilde{U}$. Define

$$\mathcal{O}_1 := \operatorname*{arg\,min}_{\mathcal{O} \in \mathbb{O}(r)} \|\widehat{U} - \widetilde{U}\mathcal{O}\|_F.$$

Then

$$\begin{aligned}
\frac{1}{\sigma_{ij}}\|\widehat{U}\widehat{U}^\top\widetilde{U} - \widetilde{U}\|_{2,\infty} &= \frac{1}{\sigma_{ij}}\|\widehat{U}\widetilde{H}^\top - \widetilde{U}\|_{2,\infty} \\
&\leq \frac{1}{\sigma_{ij}}\Big(\|\widehat{U}\widetilde{H}^\top - \widehat{U}\mathcal{O}_1^\top\|_{2,\infty} + \|\widehat{U}\mathcal{O}_1^\top - \widetilde{U}\|_{2,\infty}\Big) \\
&= \frac{1}{\sigma_{ij}}\Big(\|\widehat{U}\widetilde{H}^\top - \widehat{U}\mathcal{O}_1^\top\|_{2,\infty} + \|\widehat{U} - \widetilde{U}\mathcal{O}_1\|_{2,\infty}\Big) \\
&\leq \frac{1}{\sigma_{ij}}\Big(\|\widehat{U}(\widetilde{H}^\top - \mathcal{O}_1^\top)\|_{2,\infty} + \|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty} + \|\widetilde{U}(\mathcal{O}_1 - \widetilde{H})\|_{2,\infty}\Big) \\
&\leq \frac{1}{\sigma_{ij}}\Big(\|\widehat{U}\|_{2,\infty}\|\widetilde{H} - \mathcal{O}_1\| + \|\widetilde{U}\|_{2,\infty}\|\mathcal{O}_1 - \widetilde{H}\| + \|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty}\Big) \\
&\leq \frac{1}{\sigma_{ij}}\Big(2C\|U\|_{2,\infty}\|\sin(\widehat{U},\widetilde{U})\|_2^2 + \|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty}\Big) \\
&\leq \frac{1}{\sigma_{ij}}\Big(2C\|U\|_{2,\infty}\frac{\|\Gamma(Z)\|^2}{\lambda_r^4} + \|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty}\Big),
\end{aligned}$$

where the final line follows from the Davis-Kahan Theorem. By Theorem 8 we have that

$$\|\widehat{U} - \widetilde{U}\widetilde{H}\|_{2,\infty} \leq C_D\kappa^2\|U\|_{2,\infty}^2\frac{\|\Gamma(Z)\|}{\lambda_r^2}. \tag{B.19}$$

We have already shown that

$$\frac{1}{\sigma_{ij}}\|U\|_{2,\infty}\frac{\|\Gamma(Z)\|^2}{\lambda_r^4} \leq C_8\frac{\kappa^3\kappa_\sigma\mu_0}{\mathrm{SNR}},$$

which matches the bound for R_4, so by increasing the constant C_8, we need only bound the term in Equation (B.19). We have that

$$\begin{aligned}
\frac{1}{\sigma_{ij}}C_D\kappa^2\|U\|_{2,\infty}^2\frac{\|\Gamma(Z)\|}{\lambda_r^2} &\leq \frac{C_9}{\sigma_{ij}}\kappa^2\|U\|_{2,\infty}^2\frac{\sigma^2(n+\sqrt{nd}) + \sigma\lambda_1\sqrt{n}}{\lambda_r^2} \\
&\leq C_9\kappa_\sigma\kappa^3\|U\|_{2,\infty}^2\frac{\sigma(n+\sqrt{nd}) + \lambda_1\sqrt{n}}{\lambda_r} \\
&\leq C_9\kappa_\sigma\kappa^3\mu_0^2\frac{r\sigma(n+\sqrt{nd}) + \lambda_1 r\sqrt{n}}{n\lambda_r} \\
&\leq C_9\kappa_\sigma\kappa^3\mu_0^2\Big(\frac{r\sigma}{\lambda_r} + \frac{r\sigma\sqrt{d}}{\sqrt{n}\lambda_r} + \kappa\frac{r}{\sqrt{n}}\Big) \\
&\leq C_9\kappa^4\kappa_\sigma\mu_0^2\frac{r}{\sqrt{n}}
\end{aligned}$$

which is the desired upper bound. □

Lemma 19. *There exists a universal constant C_{10} such that with probability at least $1 - 4(n \vee d)^{-4}$*

$$\frac{1}{\sigma_{ij}}\left|\sum_{k\neq i}\langle E_i, E_k\rangle U_{kj}\lambda_j^{-2}\right| \leq C_{10}\mu_0\kappa_\sigma\frac{\log(n\vee d)}{\mathrm{SNR}},$$

where the probability is uniform over i and j.

Proof of Lemma 19. First, conditionally on E_i the sum is a sum of independent random variables each with ψ_2 norm bounded by

$$\begin{aligned}\left\|\langle E_i, E_k\rangle U_{kj}\lambda_j^{-2}\right\|_{\psi_2} &\leq \max_k \|\langle E_i, E_k\rangle\|_{\psi_2}|U_{kj}|\lambda_j^{-2}\\ &\leq C\|E_i\|\sigma\|U\|_{2,\infty}\lambda_j^{-2},\end{aligned}$$

where C is a universal constant. Hence, for any $t \geq 0$, we have that

$$\left|\sum_{k\neq i}\langle E_i, E_k\rangle U_{kj}\lambda_j^{-2}\right| \leq C\sigma\sqrt{n}\|E_i\|\|U\|_{2,\infty}\lambda_j^{-2}t$$

with probability at least $1 - 2\exp(-ct^2)$. Furthermore, for some other universal constant C, $\|E_i\| \leq C\sigma_i\sqrt{d}s$ with probability at least $1 - 2\exp(-cs^2)$ (uniformly over i). Hence,

$$\left|\sum_{k\neq i}\langle E_i, E_k\rangle U_{kj}\lambda_j^{-2}\right| \leq C_{10}\sigma\sqrt{nd}\lambda_j^{-2}\|U\|_{2,\infty}\sigma_i t$$

with probability at least $1 - 4\exp(-ct)$. Recall $\sigma_{ij}^2 := \|\Sigma_i^{1/2}V_{\cdot j}\|^2\lambda_j^{-2}$. Then

$$\begin{aligned}\frac{1}{\sigma_{ij}}\left|\sum_{k\neq i}\langle E_i, E_k\rangle U_{kj}\lambda_j^{-2}\right| &\leq \frac{C_{10}}{\sigma_{ij}}\sigma\sqrt{nd}\lambda_j^{-2}\|U\|_{2,\infty}\sigma_i t\\ &\leq C_{10}\frac{\sigma\sqrt{d}}{\lambda_j}\frac{\sigma_i}{\|\Sigma_i^{1/2}V_{\cdot j}\|}\sqrt{n}\|U\|_{2,\infty}t\\ &\leq C_{10}\kappa_\sigma\frac{\sigma\sqrt{d}}{\lambda_j}\sqrt{n}\|U\|_{2,\infty}t\end{aligned}$$

with probability at least $1 - 4e^{-ct}$, since $\sigma_i / \|\Sigma_i^{1/2} V_{\cdot j}\| \leq \kappa_\sigma$ remains bounded away from zero and infinity. Taking $t = C(4/c)\log(n \vee d)$ and absorbing the constants shows that with probability at least $1 - 4(n \vee d)^{-4}$, uniformly over i and j that

$$\begin{aligned}\frac{1}{\sigma_{ij}}\left|\sum_{k \neq i} \langle E_i, E_k \rangle U_{kj} \lambda_j^{-2}\right| &\leq C_{10}\kappa_\sigma \frac{\sigma\sqrt{d}}{\lambda_j}\sqrt{n}\|U\|_{2,\infty}\log(n \vee d) \\ &\leq C_{10}\mu_0\kappa_\sigma \frac{\log(n \vee d)}{\mathrm{SNR}}\end{aligned}$$

as required. □

B.8 Proof of Auxiliary Lemmas

First, recall Lemma 3.

Lemma 3. *There exists an orthogonal matrix $\mathcal{O}_*$ and a universal constant C such that under Assumptions 2.2 and 2.4, the event in Lemma 1, and $T = \Theta\left(\frac{\lambda_r^2}{\|U\|_{2,\infty}\|\gamma(Z)\|}\right)$,*

$$\|UH\widetilde{H} - U\mathcal{O}_*^\top\|_{2,\infty} \leq C\|U\|_{2,\infty}\frac{\|\Gamma(Z)\|^2}{\lambda_r^4}.$$

Proof of Lemma 3. We have that for any orthogonal matrices $\mathcal{O}$ and $\widetilde{\mathcal{O}}$ that

$$\begin{aligned}\|UU^\top\widetilde{U}\widetilde{U}^\top\widehat{U} - U\mathcal{O}\|_{2,\infty} &\leq \|U\|_{2,\infty}\|U^\top\widetilde{U}\widetilde{U}^\top\widehat{U} - \mathcal{O}\| \\ &\leq \|U\|_{2,\infty}\Big(\|U^\top\widetilde{U}\widetilde{U}^\top\widehat{U} - \widetilde{\mathcal{O}}\widetilde{U}^\top\widehat{U}\| + \|\mathcal{O} - \widetilde{\mathcal{O}}\widetilde{U}^\top\widehat{U}\|\Big) \\ &\leq \|U\|_{2,\infty}\Big(\|U^\top\widetilde{U} - \widetilde{\mathcal{O}}\| + \|\mathcal{O} - \widetilde{\mathcal{O}}\widetilde{U}^\top\widehat{U}\|\Big).\end{aligned}$$

Let $\mathcal{O}^{(1)}$ and $\mathcal{O}^{(2)}$ be the orthogonal matrices satisfying

$$\begin{aligned}\mathcal{O}^{(1)} &:= \arg\min \|U - \widetilde{U}\mathcal{O}\|_F; \\ \mathcal{O}^{(2)} &:= \arg\min \|\widetilde{U} - \widehat{U}\mathcal{O}\|_F.\end{aligned}$$

Define $\mathcal{O}_* := (\mathcal{O}^{(1)}\mathcal{O}^{(2)})^\top$. Letting $\widetilde{\mathcal{O}} = \mathcal{O}^{(1)}$ and $\mathcal{O} = \mathcal{O}_*^\top$ implies that this is less than or

equal to

$$\|U\|_{2,\infty}\bigg(\|\sin\Theta(U,\widetilde{U})\|^2+\|\sin\Theta(\widetilde{U},\widehat{U})\|^2\bigg).$$

By the Davis-Kahan Theorem, under Assumptions 2.2 and 2.4 and under the event in Lemma 1 by Lemmas 2 and 1 we have that for some constant C,

$$\begin{aligned}\|U\|_{2,\infty}\bigg(\|\sin\Theta(U,\widetilde{U})\|^2+\|\sin\Theta(\widetilde{U},\widehat{U})\|^2\bigg) &\leq C\|U\|_{2,\infty}\bigg(\frac{\|\Gamma(Z)\|^2}{\lambda_r^4}+\|U\|_{2,\infty}^4\frac{\|\Gamma(Z)\|^2}{\lambda_r^4}\bigg)\\ &\leq C\|U\|_{2,\infty}\frac{\|\Gamma(Z)\|^2}{\lambda_r^4}\end{aligned}$$

since $\|\Gamma(Z)\|/\lambda_r^2\leq C$ under these assumptions and the event in Lemma 1. □

Lemma 23. *If M has rank r, and $\widehat{U}$ is the projector onto the top r left singular vectors of $M+E$, and if $\lambda_r\geq 2\|E\|$, then*

$$\|(I-P_{\widehat{U}})M\|\leq 2\|E\|.$$

Proof. We have that

$$\begin{aligned}\|(I-P_{\widetilde{U}})M\| &\leq \|(I-P_{\widetilde{U}})(M+E)\|+\|(I-P_{\widetilde{U}})E\|\\ &\leq \lambda_{r+1}(M+E)+\|E\|\\ &\leq 2\|E\|\end{aligned}$$

by Weyl's inequality. □

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix C

Proofs from Chapter 3

C.1 Proof of Theorem 9

In this section we prove Theorem 9. First, Theorem 9 is actually a consequence of the following more general theorem that does not require Assumption 3.5. Section C.1.1 develops the preliminary bounds in terms of principal submatrix and eigenvalue concentration (Lemmas 4 and 5), and in Section C.1.2 we prove Theorem 22. In Section C.1.3 we show how Theorem 9 can be deduced by combining Theorem 22 with Assumption 3.5. En route to the proof of Theorem 22 we introduce several technical lemmas; we prove these in Section C.2. Recall that we denote $\kappa := \frac{\lambda_1}{\lambda_k}$ as the (reduced) condition number of Σ.

Theorem 22. *Suppose Assumptions 3.1, 3.2, 3.3, and 3.4 are satisfied. Then with probability at least* $1 - \delta - p^{-2}$*, there exists an orthogonal matrix* $W_* \in \mathbb{O}(k)$ *such that*

$$\max_{1 \le i \le p} \|\widetilde{U}_{i\cdot} - (UW_*)_{i\cdot}\| \lesssim \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4 + \mathcal{E}_5$$

where

$$\mathcal{E}_1 := \frac{\kappa\lambda_1}{\lambda_k - \lambda_{k+1}} \frac{s\log(p)}{n} \|U\|_{2\to\infty} + \kappa k \sqrt{\frac{\log(p)}{n}} \|U\|_{2\to\infty}$$
$$\mathcal{E}_2 := \frac{\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2} \frac{s\log(p)}{n} \|U\|_{2\to\infty}$$
$$\mathcal{E}_3 := \sqrt{\frac{s\log(p)}{n}} \frac{\kappa\lambda_1^{1/2}}{\lambda_k - \lambda_{k+1}} \min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1} \|U\|_{2\to\infty} \right)$$
$$\mathcal{E}_4 := \frac{\lambda_{k+1}}{\lambda_k} \kappa^2 \sqrt{\frac{k\log(p)}{n}} + \frac{\lambda_{k+1}}{\lambda_k} \kappa^3 \frac{s\log(p)}{n};$$
$$\mathcal{E}_5 := \frac{\kappa\lambda_1}{\lambda_k - \lambda_{k+1}} \frac{s\log(p)}{n} + \kappa \sqrt{\frac{k\log(p)}{n}}.$$

C.1.1 Preliminary Bounds

Note that by Assumption 3.2, we need only examine the $s \times k$ matrix of eigenvectors of $\widehat{\Sigma}_{JJ}$ and Σ_{JJ} respectively. We will develop an expansion for the difference $\widetilde{U}_J - U_J W_*$ by viewing $\widehat{\Sigma}_{JJ}$ as a perturbation of Σ_{JJ}. For convenience we restate the initial preliminary bounds in the main paper. Except for Proposition 1, the proofs are in Section C.2.1. The first is the following principal submatrix concentration bound.

Lemma 4 (Principal Submatrix Concentration)**.** *Let J be an index set of $\{1, ..., p\}$ of size s. Then*

$$\|\widehat{\Sigma}_{JJ} - \Sigma_{JJ}\| \lesssim \lambda_1 \left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}} \right)$$

with probability at least $1 - O(p^{-4})$.

Henceforth, we assume the correct support set J is known; it is the correct set J with probability at least $1 - \delta$ by Assumption 3.2. As discussed in the main paper, using Lemma 4, we can derive the following eigenvalue bound, which we present as a lemma below.

Lemma 5 (Existence of an Eigengap)**.** *Under the event in Lemma 4 and Assumption 3.4,*

the eigenvalues of $\widehat{\Sigma}_{JJ}$ *and* Σ_{JJ} *satisfy*

$$\lambda_k - \widetilde{\lambda}_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8}; \qquad \widetilde{\lambda}_k - \lambda_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8};$$
$$\widetilde{\lambda}_k \geq \frac{\lambda_k}{4}.$$

Consequently, this bound holds with probability at least $1 - O(p^{-4})$.

Finally, we have the following $\sin\Theta$ distance error between U_J and $\widetilde{U}_J$.

Proposition 1 (Spectral Proximity)**.** *Under the assumptions of Theorem 22, we have that*

$$\|U_J U_J^\top - \widetilde{U}_J \widetilde{U}_J^\top\| \lesssim \frac{\lambda_1}{\lambda_k - \lambda_{k+1}}\left[\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}}\right]$$

with probability at least $1 - O(p^{-4})$.

Proof of Proposition 1. By the Davis-Kahan Theorem (Bhatia, 1997; Yu et al., 2014) and Lemma 5,

$$\begin{aligned}\|U_J U_J^\top - \widetilde{U}_J \widetilde{U}_J^\top\| &\leq \frac{\|\widehat{\Sigma}_{JJ} - \Sigma_{JJ}\|}{\lambda_k - \widetilde{\lambda}_{k+1}}. \\ &\lesssim \frac{\|\widehat{\Sigma}_{JJ} - \Sigma_{JJ}\|}{\lambda_k - \lambda_{k+1}} \end{aligned} \tag{C.1}$$

By Lemma 4, with probability at least $1 - O(p^{-4})$, the numerator can be bounded by

$$\|\widehat{\Sigma}_{JJ} - \Sigma_{JJ}\| \leq \lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}}\right)$$

Combining this and Equation (C.1) gives the result. □

In the proofs that follow, we will use the fact that by Proposition 1, we have that

$$\|U_J U_J^\top - \widetilde{U}_J \widetilde{U}_J^\top\| \lesssim \frac{\lambda_1}{\lambda_k - \lambda_{k+1}}\sqrt{\frac{s\log(p)}{n}},$$

which is a little more amenable to downstream analysis. In addition, we use several equivalent expressions for the spectral norm of the difference of projections; see Lemma 31 in Appendix C.3 for a discussion of how to equate these.

C.1.2 Proof of Theorem 22

We now proceed with the proof. At a high level, the argument consists of a deterministic matrix decomposition, each term of which we bound in probability. Define $\widetilde{\Lambda}$ as the diagonal $k \times k$ matrix of eigenvalues of $\widehat{\Sigma}_{JJ}$. Define W_* to be the matrix

$$W_* := \operatorname*{arg\,min}_{W \in \mathbb{O}(k)} \|\widetilde{U}_J - U_J W\|_F.$$

Is is well-known that W_* can be computed from the singular value decomposition of $U_J^\top \widetilde{U}_J$ (e.g. Abbe et al. (2020); Cape et al. (2019b); Chen et al. (2021c)).

We now expand the difference via

$$\begin{aligned}
\widetilde{U}_J - U_J W_* &= \widetilde{U}_J - U_J U_J^\top \widetilde{U}_J - U_J(W_* - U_J^\top \widetilde{U}_J) \\
&= \widetilde{U}_J - U_J \Lambda U_J^\top \widetilde{U}_J \widetilde{\Lambda}^{-1} + U_J \Lambda U_J^\top \widetilde{U}_J \widetilde{\Lambda}^{-1} - U_J U_J^\top \widetilde{U}_J - U_J(W_* - U_J^\top \widetilde{U}_J) \\
&= \widetilde{U}_J - U_J \Lambda U_J^\top \widetilde{U}_J \widetilde{\Lambda}^{-1} + U_J(\Lambda U_J^\top \widetilde{U}_J - U_J^\top \widetilde{U}_J \widetilde{\Lambda})\widetilde{\Lambda}^{-1} - U_J(W_* - U_J^\top \widetilde{U}_J) \\
&= \widetilde{U}_J \widetilde{\Lambda} \widetilde{\Lambda}^{-1} - U_J \Lambda U_J^\top \widetilde{U}_J \widetilde{\Lambda}^{-1} + U_J(\Lambda U_J^\top \widetilde{U}_J - U_J^\top \widetilde{U}_J \widetilde{\Lambda})\widetilde{\Lambda}^{-1} - U_J(W_* - U_J^\top \widetilde{U}_J) \\
&= (\widetilde{U}_J \widetilde{\Lambda} - U_J \Lambda U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1} + U_J(\Lambda U_J^\top \widetilde{U}_J - U_J^\top \widetilde{U}_J \widetilde{\Lambda})\widetilde{\Lambda}^{-1} - U_J(W_* - U_J^\top \widetilde{U}_J) \\
&= A + T_1 - T_2, \qquad \text{(C.2)}
\end{aligned}$$

where

$$\begin{aligned}
A &:= (\widetilde{U}_J \widetilde{\Lambda} - U_J \Lambda U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1}; \\
T_1 &:= U_J(\Lambda U_J^\top \widetilde{U}_J - U_J^\top \widetilde{U}_J \widetilde{\Lambda})\widetilde{\Lambda}^{-1} \\
T_2 &:= U_J(W_* - U_J^\top \widetilde{U}_J).
\end{aligned}$$

Both T_1 and T_2 are analyzed concisely in Lemmas 24 and 25 as follows. Their proofs are in Section C.2.2. The proof of Lemmas 24 and 25 are both rather straightforward and based on previous results in entrywise eigenvector analysis (Abbe et al., 2022, 2020; Agterberg and Sulam, 2022; Cai et al., 2021a; Cape et al., 2019a,b; Chen et al., 2021c; Tang et al., 2017c; Xia and Yuan, 2020; Xie et al., 2022; Xie, 2022; Yan et al., 2021).

Lemma 24 (Bound on T_1)**.** *We have that*

$$\begin{aligned}\|U_J(\Lambda U_J^\top \widetilde{U}_J - U_J^\top \widetilde{U}_J \widetilde{\Lambda})\widetilde{\Lambda}^{-1}\|_{2\to\infty} &\lesssim \frac{\|U\|_{2\to\infty}\lambda_1^2}{\lambda_k(\lambda_k - \lambda_{k+1})}\frac{s\log(p)}{n} \\ &\quad + \frac{k\lambda_1\|U\|_{2\to\infty}}{\lambda_k}\sqrt{\frac{\log(p)}{n}} \\ &\equiv \mathcal{E}_1\end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Lemma 25 (Bound on T_2)**.** *We have that*

$$\begin{aligned}\|U_J(W_* - U_J^\top \widetilde{U}_J)\|_{2\to\infty} &\lesssim \frac{\|U\|_{2\to\infty}\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2}\frac{s\log(p)}{n} \\ &\equiv \mathcal{E}_2\end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Expanding Equation (C.2) into T_3 and T_4:

We further expand the remaining term in (C.2) by viewing $\widehat{\Sigma}_{JJ}$ as a perturbation of $U_J U_J^\top \Sigma_{JJ}$ and using the eigenvector-eigenvalue equation via

$$\begin{aligned}A &= (\widetilde{U}_J\widetilde{\Lambda} - U_J\Lambda U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1} \\ &= (\widehat{\Sigma}_{JJ}\widetilde{U}_J - \Sigma_{JJ}U_J U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1} \\ &= (U_J U_J^\top \Sigma_{JJ}\widetilde{U}_J + (\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\widetilde{U}_J - \Sigma_{JJ}U_J U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1} \\ &= U_J U_J^\top \Sigma_{JJ}(\widetilde{U}_J - U_J U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1} + (\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\widetilde{U}_J\widetilde{\Lambda}^{-1} \\ &= T_3 + T_4,\end{aligned}$$

where

$$T_3 := U_J U_J^\top \Sigma_{JJ}(\widetilde{U}_J - U_J U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1}$$
$$T_4 := (\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\widetilde{U}_J \widetilde{\Lambda}^{-1}.$$

The term T_3 can be analyzed via techniques from complex analysis. We present this bound as a lemma, but defer the proof to Section C.2.3.

Lemma 26 (Bound on T_3)**.** *We have that*

$$\|U_J U_J^\top \Sigma_{JJ}(\widetilde{U}_J - U_J U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1}\|_{2\to\infty} \lesssim \sqrt{\frac{s\log(p)}{n}} \frac{\lambda_1^{3/2}}{\lambda_k(\lambda_k - \lambda_{k+1})} \min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty}\right)$$
$$\equiv \mathcal{E}_3$$

with probability at least $1 - O(p^{-3})$.

Expanding T_4 in terms of J_1 and J_2:

We now proceed to bound T_4. We have by Lemma 5 and properties of the $2 \to \infty$ norm that

$$\|(\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\widetilde{U}_J \widetilde{\Lambda}^{-1}\|_{2\to\infty} \le \frac{1}{\widetilde{\lambda}_k}\|(\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\widetilde{U}_J\|_{2\to\infty}$$
$$\lesssim \frac{1}{\lambda_k}\|(\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\widetilde{U}_J\|_{2\to\infty}. \tag{C.3}$$

Note that $\widetilde{U}_J$ is the matrix of eigenvectors of $\widehat{\Sigma}_{JJ}$ and hence is not independent of the error matrix $\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ}$, so one cannot bound the maximum row norm of the matrix above with standard concentration techniques. Let $U_\perp$ be the matrix such that $[U_J, U_\perp]$ is an $s \times s$ orthogonal matrix, and let $\widetilde{U}_\perp$ be defined similarly. Define also $\Lambda_\perp$ and $\widetilde{\Lambda}_\perp$ as the matrix of

bottom $s-k$ eigenvalues of Σ_{JJ} and $\widehat{\Sigma}_{JJ}$ respectively. Since $\widetilde{U}_\perp^\top \widetilde{U}_J = 0$, we have that

$$\begin{aligned}
&\frac{1}{\lambda_k}\|(\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ})\widetilde{U}_J\|_{2\to\infty} \\
&\qquad = \frac{1}{\lambda_k}\left\|\left(\widetilde{U}_J\widetilde{\Lambda}\widetilde{U}_J^\top + \widetilde{U}_\perp\widetilde{\Lambda}_\perp\widetilde{U}_\perp^\top - U_J\Lambda_J U_J^\top\right)\widetilde{U}_J\right\|_{2\to\infty} \\
&\qquad \le \frac{1}{\lambda_k}\left\|\left(\widetilde{U}_J\widetilde{\Lambda}\widetilde{U}_J^\top + \widetilde{U}_\perp\widetilde{\Lambda}_\perp\widetilde{U}_\perp^\top - U_J\Lambda_J U_J^\top - U_\perp\Lambda_\perp U_\perp^\top\right)\widetilde{U}_J\right\|_{2\to\infty} + \frac{1}{\lambda_k}\|U_\perp\Lambda_\perp U_\perp^\top\widetilde{U}_J\|_{2\to\infty} \\
&\qquad \le \frac{1}{\lambda_k}\|(\widehat{\Sigma}_{JJ} - \Sigma_{JJ})\widetilde{U}_J\|_{2\to\infty} + \frac{1}{\lambda_k}\|U_\perp\Lambda_\perp U_\perp^\top\widetilde{U}_J\|_{2\to\infty} \\
&\qquad := \frac{1}{\lambda_k}\|J_1\|_{2\to\infty} + \frac{1}{\lambda_k}\|J_2\|_{2\to\infty} \qquad \text{(C.4)}
\end{aligned}$$

where

$$\begin{aligned}
J_1 &:= (\widehat{\Sigma}_{JJ} - \Sigma_{JJ})\widetilde{U}_J; \\
J_2 &:= U_\perp\Lambda_\perp U_\perp^\top\widetilde{U}_J.
\end{aligned}$$

The term J_2 can be bounded in the following lemma, but it is rather technical; moreover, it requires some analysis that is relatively novel in the subspace estimation literature; in particular, we combine some ideas from Xia and Yuan (2020) as well as Cape et al. (2019a); Xie et al. (2022); Tang (2018); Tang et al. (2017c). The proof is in Section C.2.4.

Lemma 27 (Bound on J_2)**.** *The term J_2 satisfies*

$$\begin{aligned}
\|U_\perp\Lambda_\perp U_\perp^\top\widetilde{U}_J\|_{2\to\infty} &\lesssim \kappa^2\lambda_{k+1}\sqrt{\frac{k\log(p)}{n}} + \lambda_{k+1}\kappa^3\frac{s\log(p)}{n} \\
&\lesssim \mathcal{E}_4\lambda_k
\end{aligned}$$

with probability at least $1 - O(p^{-3})$.

Further expanding the term J_1:

What remains is to bound the first term of (C.4); i.e. the term J_1. First, note that by Assumption 3.3, each vector $X_i \in \mathbb{R}^p$ is of the form $X_i = \Sigma^{1/2}Y_i$, where $\mathbb{E}Y_iY_i^\top = I$. Let X be the $n \times p$ matrix whose rows are the X_i's; it follows that $X = Y\Sigma^{1/2}$. Let Y be

partitioned via $Y = [Y_J, Y_{J^c}]$, where Y_J is the $n \times s$ matrix of variables corresponding to those in J, and Y_{J^c} is the $n \times p - s$ matrix of the other variables. Define through slight abuse of notation the matrix $\Sigma^{1/2}_{JJ^c} := (\Sigma^{1/2})_{JJ^c}$. With these notations in place, we observe that since $\widehat{\Sigma} = \frac{1}{n}(X^\top X)$ we have the block structure

$$\widehat{\Sigma}_{JJ} = \frac{1}{n}\Big((\Sigma^{1/2})_{JJ} Y_J^\top Y_J (\Sigma^{1/2})_{JJ} + \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} + (\Sigma^{1/2})_{JJ} Y_J^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top + \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top\Big).$$

Therefore, we observe that

$$\begin{aligned}(\widehat{\Sigma}_{JJ} - \Sigma_{JJ})\widetilde{U}_J = \frac{1}{n}\Big((\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ} + \Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} \\ + (\Sigma^{1/2})_{JJ} Y_J^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top + \Sigma^{1/2}_{JJ^c}(Y_{J^c}^\top Y_{J^c} - nI)(\Sigma^{1/2}_{JJ^c})^\top\Big)\widetilde{U}_J.\end{aligned} \tag{C.5}$$

Here the identity matrices are of size s and $p-s$ respectively in order of appearance. In light of the structure in (C.5), define the matrices

$$\begin{aligned}K_1 &:= \frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\widetilde{U}_J; \\ K_2 &:= \frac{1}{n}\Sigma^{1/2}_{JJ^c} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ}\widetilde{U}_J; \\ K_3 &:= \frac{1}{n}(\Sigma^{1/2})_{JJ} Y_J^\top Y_{J^c} (\Sigma^{1/2}_{JJ^c})^\top \widetilde{U}_J; \\ K_4 &:= \frac{1}{n}\Sigma^{1/2}_{JJ^c}(Y_{J^c}^\top Y_{J^c} - nI)(\Sigma^{1/2}_{JJ^c})^\top \widetilde{U}_J.\end{aligned}$$

Then

$$J_1 = (\widehat{\Sigma}_{JJ} - \Sigma_{JJ})\widetilde{U}_J = K_1 + K_2 + K_3 + K_4.$$

We have lemmas that bound the $2 \to \infty$ norms of each of these matrices. Each of these bounds follows essentially the same set of steps:

1. Bound the $2 \to \infty$ norm using properties of the $2 \to \infty$ norm in terms of the maximum entry.
2. Write each entry as a sum of mean-zero subexponential random variables and use

either Bernstein's inequality or the Hanson-Wright inequality (see Appendix C.3) to bound the result.

The proofs for these lemmas are in Sections C.2.5 and C.2.6. Recall that we define $\mathcal{E}_5$ via

$$\begin{aligned}\mathcal{E}_5 :&= \frac{\kappa\lambda_1}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} + \kappa\sqrt{\frac{k\log(p)}{n}} \\ &\equiv \frac{1}{\lambda_k}\left(\frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} + \lambda_1\sqrt{\frac{k\log(p)}{n}}\right).\end{aligned}$$

Lemma 28 (The matrix K_1)**.** *The matrix K_1 satisfies*

$$\begin{aligned}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\widetilde{U}\|_{2\to\infty} &\lesssim \frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} + \lambda_1\sqrt{\frac{k\log(p)}{n}} \\ &\lesssim \mathcal{E}_5\lambda_k\end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Lemma 29 (The matrix K_2)**.** *The matrix K_2 satisfies*

$$\begin{aligned}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}\widetilde{U}\|_{2\to\infty} &\lesssim \lambda_1\sqrt{\frac{k\log(p)}{n}} + \frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} \\ &\lesssim \mathcal{E}_5\lambda_k\end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Lemma 30 (The matrices K_3 and K_4)**.** *The matrices K_3 and K_4 satisfy*

$$\begin{aligned}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}Y_J^\top Y_{J^c}(\Sigma^{1/2}_{JJ^c})^\top\widetilde{U}_J\|_{2\to\infty} &\lesssim \frac{s\log(p)}{n}\frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}} \\ &\lesssim \mathcal{E}_5\lambda_k; \\ \|\frac{1}{n}\Sigma^{1/2}_{JJ^c}(Y_{J^c}^\top Y_{J^c} - nI)(\Sigma^{1/2}_{JJ^c})^\top\widetilde{U}\|_{2\to\infty} &\lesssim \frac{s\log(p)}{n}\frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}} \\ &\lesssim \mathcal{E}_5\lambda_k\end{aligned}$$

with probability at least $1 - O(p^{-3})$.

Putting it together:

We are now ready to complete the proof. We have that

$$\begin{aligned}
\|\widetilde{U}_J - U_J W_*\|_{2\to\infty} &\le \left\| \left(\widetilde{U}_J\widetilde{\Lambda} - U_J\Lambda U^\top \widehat{U}\right)\widetilde{\Lambda}^{-1}\right\|_{2\to\infty} + \|T_1\|_{2\to\infty} + \|T_2\|_{2\to\infty} \\
&\le \left\| \left(\widetilde{U}_J\widetilde{\Lambda} - U_J\Lambda U^\top \widehat{U}\right)\widetilde{\Lambda}^{-1}\right\|_{2\to\infty} + \mathcal{E}_1 + \mathcal{E}_2 \\
&\le \|T_3\|_{2\to\infty} + \|T_4\|_{2\to\infty} + \mathcal{E}_1 + \mathcal{E}_2 \\
&\le \|T_4\|_{2\to\infty} + \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 \\
&\lesssim \frac{\|J_1\|_{2\to\infty} + \|J_2\|_{2\to\infty}}{\lambda_k} + \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 \\
&\lesssim \frac{\|J_1\|_{2\to\infty}}{\lambda_k} + \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4 \\
&\lesssim \frac{1}{\lambda_k}\Big(\|K_1\|_{2\to\infty} + \|K_2\|_{2\to\infty} + \|K_3\|_{2\to\infty} + \|K_4\|_{2\to\infty}\Big) + \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4 \\
&\le \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 + \mathcal{E}_4 + \mathcal{E}_5
\end{aligned}$$

with probability at least $1 - O(p^{-3})$. Consequently, by the union bound and Assumption 3.2, this bound holds with probability at least $1 - \delta - p^{-2}$ as desired.

C.1.3 Proof of Theorem 9

In this section we show how Theorem 9 can be deduced from Theorem 22. We simply bound $\mathcal{E}_1$ through $\mathcal{E}_5$ using the additional assumptions introduced in Assumption 3.5.

Note that under Assumption 3.5, we have that $\lambda_{k+1} \le \frac{\lambda}{2}$ and $\lambda_k \ge \lambda$, implying that $\lambda_k - \lambda_{k+1} \ge \frac{\lambda}{2}$. In addition $\lambda_1 \le \kappa\lambda$. Therefore,

$$\begin{aligned}
\frac{\lambda_1}{\lambda_k} &\le \frac{\kappa\lambda}{\lambda} \le \kappa; \\
\frac{\lambda_1}{\lambda_k - \lambda_{k+1}} &\lesssim \frac{\kappa\lambda}{\lambda} \le \kappa.
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathcal{E}_1 &= \frac{\kappa\lambda_1}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n}\|U\|_{2\to\infty} + \kappa k\sqrt{\frac{\log(p)}{n}}\|U\|_{2\to\infty} \\ &\lesssim \kappa^2\frac{s\log(p)}{n}\|U\|_{2\to\infty} + \kappa k\sqrt{\frac{\log(p)}{n}}\|U\|_{2\to\infty} \\ &\lesssim \kappa^2\frac{\sqrt{sk}\log(p)}{n} + \kappa\frac{k^{3/2}}{\sqrt{s}}\sqrt{\frac{\log(p)}{n}} \\ &\lesssim \kappa^2\frac{s\log(p)}{n} + \kappa\sqrt{\frac{k\log(p)}{n}}, \end{aligned} \tag{C.6}$$

where the penultimate inequality comes from the fact that $\|U\|_{2\to\infty} \lesssim (k/s)^{1/2}$ and that $k \lesssim \sqrt{s}$. Similarly,

$$\begin{aligned}\mathcal{E}_2 :&= \frac{\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2}\frac{s\log(p)}{n}\|U\|_{2\to\infty} \\ &\lesssim \kappa^2\frac{s\log(p)}{n}\|U\|_{2\to\infty} \\ &\lesssim \kappa^2\frac{\sqrt{sk}\log(p)}{n} \\ &\lesssim \kappa^2\frac{s\log(p)}{n}. \end{aligned} \tag{C.7}$$

For $\mathcal{E}_3$,

$$\begin{aligned}\mathcal{E}_3 &= \sqrt{\frac{s\log(p)}{n}}\frac{\kappa\lambda_1^{1/2}}{\lambda_k - \lambda_{k+1}}\min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty}\right) \\ &\lesssim \sqrt{\frac{s\log(p)}{n}}\frac{\kappa\lambda_1^{1/2}}{\lambda_k - \lambda_{k+1}}\sqrt{\lambda_1}\|U\|_{2\to\infty} \\ &\lesssim \kappa^2\sqrt{\frac{s\log(p)}{n}}\|U\|_{2\to\infty} \\ &\lesssim \kappa^2\sqrt{\frac{k\log(p)}{n}} \end{aligned} \tag{C.8}$$

since $\|U\|_{2\to\infty} \lesssim (k/s)^{1/2}$. For $\mathcal{E}_4$, we have that since $\lambda_{k+1} < \lambda_k$, then

$$\begin{aligned}\mathcal{E}_4 &= \kappa^2\frac{\lambda_{k+1}}{\lambda_k}\sqrt{\frac{k\log(p)}{n}} + \frac{\lambda_{k+1}}{\lambda_k}\kappa^3\frac{s\log(p)}{n} \\ &\lesssim \kappa^2\sqrt{\frac{k\log(p)}{n}} + \kappa^3\frac{s\log(p)}{n}. \end{aligned} \tag{C.9}$$

Finally, for $\mathcal{E}_5$, we see that

$$\begin{aligned}\mathcal{E}_5 :&= \frac{\kappa\lambda_1}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} + \kappa\sqrt{\frac{k\log(p)}{n}}. \\ &\lesssim \kappa^2\frac{s\log(p)}{n} + \kappa\sqrt{\frac{k\log(p)}{n}}. \end{aligned} \tag{C.10}$$

The condition number is always larger than 1. Hence, combining (C.6),(C.7),(C.8),(C.9) and (C.10) completes the proof.

C.2 Proofs of Intermediate Lemmas

In this section we collect the proofs of the Lemmas needed en route to the proof of Theorem 22. All the lemmas are self-contained and repeated for convenience.

C.2.1 Proofs of Lemmas 4 and 5

First, we recall the statement of Lemma 4.

Lemma 4 (Principal Submatrix Concentration)**.** *Let J be an index set of $\{1, ..., p\}$ of size s. Then*

$$\|\widehat{\Sigma}_{JJ} - \Sigma_{JJ}\| \lesssim \lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}}\right)$$

with probability at least $1 - O(p^{-4})$.

Proof of Lemma 4. The result is similar to Amini and Wainwright (2009), but for general subgaussian ensembles as opposed to Gaussian ensembles. The proof is standard via an ε-net; we follow similarly to the proof of Theorem 6.5 in Wainwright (2019).

Let $\Delta = \widehat{\Sigma}_{JJ} - \Sigma_{JJ}$. First take an 1/8-net of the S^{s-1} sphere; denote these vectors $v_1, ..., v_N$, where $N \leq 17^s$ (see Example 5.8 in Wainwright (2019)). Then for any s-unit vector v, there exists some vector v_j of distance at most $\varepsilon = \frac{1}{8}$ to v. Therefore

$$\langle v, \Delta v\rangle = \langle v_j, \Delta v_j\rangle + 2\langle (v - v_j), \Delta v_j\rangle + \langle v - v_j, \Delta(v - v_j)\rangle.$$

Hence, we see that by the triangle inequality and Cauchy-Schwarz,

$$\begin{aligned}|\langle v, \Delta v\rangle| &\leq |\langle v_j, \Delta v_j\rangle| + 2\|\Delta\|\|v - v_j\|\|v_j\| + \|\Delta\|\|v - v_j\|^2 \\ &\leq |\langle v_j, \Delta v_j\rangle| + \frac{1}{2}\|\Delta\|,\end{aligned}$$

where the final inequality comes from the fact that v_j is at most distance $\frac{1}{8}$ to v. Letting v denote the unit vector achieving $\sup\langle v, Qv\rangle$ and rearranging we have that

$$\|\Delta\| \leq 2|\langle v_j, \Delta v_j\rangle| \leq 2\max_{1\leq i\leq n}|\langle v_i, \Delta v_i\rangle|.$$

So we therefore have that

$$\mathbb{E}(\exp(\lambda\|\Delta\|)) \leq \mathbb{E}\bigg(\exp(2\lambda\max_{1\leq i\leq N}|\langle v_i, \Delta v_i\rangle|)\bigg) \leq \sum_{i=1}^{N}\bigg(\mathbb{E}(\exp(2\lambda\langle v_i, \Delta v_i\rangle)) + \mathbb{E}(\exp(-2\lambda\langle v_i, \Delta v_i\rangle))\bigg).$$

We now bound the mgf above, which is the primary technical difference between this and Theorem 6.5 in Wainwright (2019). Denote $X_i[J]$ as the vector X_i with only the components in J, and let u be an arbitrary unit vector. From the assumption the X_i's are iid we have that

$$\begin{aligned}\mathbb{E}\exp(tu^\top\Delta u) &= \prod_{i=1}^{n}\mathbb{E}_{X_i}\bigg[\exp\bigg(\frac{t}{n}[(X_i[J]^\top u)^2 - u^\top\Sigma_{JJ}u]\bigg)\bigg] \\ &= \bigg(\mathbb{E}_{X_1}\bigg[\exp\bigg(\frac{t}{n}[(X_1[J]^\top u)^2 - u^\top\Sigma_{JJ}u]\bigg)\bigg]\bigg)^n.\end{aligned}$$

Let ε be a Rademacher random variable independent of X_1. Then by the contraction property of Rademacher random variables,

$$\begin{aligned}\mathbb{E}_{X_1}\bigg[\exp\bigg(\frac{t}{n}[(X_1[J]^\top u)^2 - u^\top\Sigma_{JJ}u]\bigg)\bigg] &\leq \mathbb{E}_{X_1,\varepsilon}\bigg[\exp\bigg(\frac{2t}{n}\varepsilon((X_1[J])^\top u)^2\bigg)\bigg] \\ &= \sum_{k=0}^{\infty}\frac{1}{k!}\bigg(\frac{2t}{n}\bigg)^k\mathbb{E}(\varepsilon^k(X_1[J]^\top u)^{2k}) \\ &= 1 + \sum_{k=1}^{\infty}\frac{1}{(2k)!}\bigg(\frac{2t}{n}\bigg)^{2k}\mathbb{E}((X_1[J]^\top u)^{4k})\end{aligned}$$

where the first is by the series expansion for the exponential, and the second is by noting

that the ε are rademacher and hence have vanishing odd moments.

Note that by assumption the X_i's can be written as $X_i = \Sigma^{1/2} Y_i$ for some independent Y_i's satisfying $\|Y_{ij}\|_{\psi_2} \leq 1$. Then $\|\Sigma^{1/2} Y_i\|_{\psi_2} \leq \sqrt{\lambda_1}$. Hence, by equivalence of the subgaussian norm, the moments satisfy

$$\mathbb{E}((X_1[J]^\top u)^{4k}) \leq \frac{(4k)!}{2^{2k}(2k)!}(\sqrt{8}e\lambda_1^{1/2})^{4k}.$$

From this, we deduce

$$\begin{aligned} 1 + \sum_{k=1}^{\infty} \frac{1}{(2k)!}\big(\frac{2t}{n}\big)^{2k}\mathbb{E}((X_1[J]^\top u)^{4k}) &\leq 1 + \sum_{k=1}^{\infty} \frac{1}{(2k)!}\left(\frac{2t}{n}\right)^{2k} \frac{(4k)!}{2^{2k}(2k)!}(\sqrt{8}e\lambda_1^{1/2})^{4k} \\ &\leq 1 + \sum_{k=1}^{\infty}\left(\frac{16t}{n}e^2\lambda_1\right)^{2k} \end{aligned}$$

which is a geometric series. Hence, since $\frac{1}{1-a} \leq e^{2a}$ for all $a \in [0, 1/2]$, we have that

$$1 + \sum_{k=1}^{\infty}\left(\frac{16t}{n}e^2\lambda_1\right)^{2l} \leq \exp\left(2\big[\frac{16t}{n}e^2\lambda_1\big]^2\right)$$

for all $|t| < \frac{n}{32e^2\lambda_1}$. Therefore, we have shown

$$\mathbb{E}\exp(tu^\top \Delta u) \leq \exp\left(512\frac{t^2}{n}e^4\lambda_1^2\right).$$

From here, using the sum, we have that for all $|t| < \frac{n}{64e^2\lambda_1}$ that

$$\begin{aligned} \mathbb{E}(\exp(t\|\Delta\|)) &\leq \sum_{i=1}^{N}\Big(\mathbb{E}(\exp(2t\langle v_i, \Delta v_i\rangle)) + \mathbb{E}(\exp(-2t\langle v_i, \Delta v_i\rangle))\Big) \\ &\leq 2Ne^{2048\frac{t^2}{n}e^4\lambda_1^2} \\ &\leq \exp(C\frac{t^2\lambda_1^2}{n} + 4s), \end{aligned}$$

since $2(17^s) \leq e^{4s}$. Therefore, by the Chernoff bound,

$$\mathbb{P}\Big(\|\Delta\| > \eta\Big) \leq \exp\left(C\frac{t^2\lambda_1^2}{n} + 4s - \eta t\right).$$

Minimizing over t shows that

$$t = \frac{n\eta}{2C\lambda_1^2}$$

is the minimizer provided that $\eta < \frac{C\lambda_1}{32e^2}$. Plugging this value of t in yields

$$\begin{aligned}\mathbb{P}\bigg(\|\Delta\| > \eta\bigg) &\leq \exp\left(4s - \frac{\eta^2 n}{4C\lambda_1^2}\right) \\ &= \exp\left[n\left(\frac{4s}{n} - \frac{\eta^2}{4C\lambda_1^2}\right)\right].\end{aligned}$$

Suppose $\eta = C_2\lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{4\log(p)}{n}}\right)$ for some sufficiently large constant C_2. Note that Assumption 3.1 ensures that this choice of η satisfies $\eta < \frac{C\lambda_1}{32e^2}$ since $s/n = o(1)$ and $\log(p)/n = o(1)$. Therefore, with this choice of η, we have that

$$\begin{aligned}\exp\left[n\left(\frac{4s}{n} - \frac{\eta^2}{4C\lambda_1^2}\right)\right] &\leq \exp(-4\log(p)) \\ &\leq p^{-4}.\end{aligned}$$

Consequently, recalling that $\Delta = \widehat{\Sigma}_{JJ} - \Sigma_{JJ}$ we have that

$$\mathbb{P}\left[\|\widehat{\Sigma}_{JJ} - \Sigma_{JJ}\| > C_2\lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{4\log(p)}{n}}\right)\right] \leq p^{-4}$$

as desired. □

Again, we recall the statement of Lemma 5.

Lemma 5 (Existence of an Eigengap)**.** *Under the event in Lemma 4 and Assumption 3.4, the eigenvalues of $\widehat{\Sigma}_{JJ}$ and Σ_{JJ} satisfy*

$$\lambda_k - \widetilde{\lambda}_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8}; \qquad \widetilde{\lambda}_k - \lambda_{k+1} \geq \frac{\lambda_k - \lambda_{k+1}}{8};$$
$$\widetilde{\lambda}_k \geq \frac{\lambda_k}{4}.$$

Consequently, this bound holds with probability at least $1 - O(p^{-4})$.

Proof of Lemma 5. By Weyl's inequality, the event in Lemma 4 implies that for all $1 \leq i \leq s$ that

$$|\lambda_i - \widetilde{\lambda}_i| \leq C\lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}}\right).$$

Note that Σ_{JJ} is a principal submatrix of Σ; hence its eigenvalues satisfy $\lambda_i(\Sigma_{JJ}) \leq \lambda_i$ for all $i \geq k+1$ (when $1 \leq i \leq k$ we have equality). Therefore, By Assumption 3.4, we have that

$$\begin{aligned}
\lambda_k - \widetilde{\lambda}_{k+1} &\geq \lambda_k - \lambda_{k+1}(\Sigma_{JJ}) - C\lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}}\right) \\
&\geq \lambda_k - \lambda_{k+1} - C\lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}}\right) \\
&\geq \frac{7}{8}(\lambda_k - \lambda_{k+1}) \\
&\geq \frac{\lambda_k - \lambda_{k+1}}{8},
\end{aligned}$$

and similarly for $\widetilde{\lambda}_k - \lambda_{k+1}$. For the final bound,

$$\begin{aligned}
\widetilde{\lambda}_k &\geq \lambda_k - C\lambda_1\left(\sqrt{\frac{s}{n}} + \sqrt{\frac{\log(p)}{n}}\right) \\
&\geq \lambda_k - \lambda_k/8 \\
&\geq \frac{\lambda_k}{4},
\end{aligned}$$

which completes the proof. □

C.2.2 Proof of Lemmas 24 and 25

First we will recall the statement of Lemma 24.

Lemma 24 (Bound on T_1)**.** *We have that*

$$\begin{aligned}\|U_J(\Lambda U_J^\top \widetilde{U}_J - U_J^\top \widetilde{U}_J \widetilde{\Lambda})\widetilde{\Lambda}^{-1}\|_{2\to\infty} &\lesssim \frac{\|U\|_{2\to\infty}\lambda_1^2}{\lambda_k(\lambda_k - \lambda_{k+1})}\frac{s\log(p)}{n} \\ &\quad + \frac{k\lambda_1\|U\|_{2\to\infty}}{\lambda_k}\sqrt{\frac{\log(p)}{n}} \\ &\equiv \mathcal{E}_1\end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Proof of Lemma 24. Note that by properties of the $2 \to \infty$ norm, we have

$$\|U_J(\Lambda U_J^\top \widetilde{U} - U_J^\top \widetilde{U}_J \widetilde{\Lambda})\widetilde{\Lambda}^{-1}\|_{2\to\infty} \le \|U_J\|_{2\to\infty}\|\Lambda U_J^\top \widetilde{U} - U_J^\top \widetilde{U}_J\widetilde{\Lambda}\|\widehat{\lambda}_k^{-1}. \tag{C.11}$$

We note that $\lambda_k \lesssim \widetilde{\lambda}_k$ with probability $1 - O(p^{-4})$ by Lemma 5. Furthermore, by the eigenvector equation,

$$\begin{aligned}\Lambda U_J^\top \widetilde{U}_J - U_J^\top \widetilde{U}_J\widetilde{\Lambda} &= (U_J\Lambda)^\top \widetilde{U}_J - U_J^\top \widetilde{U}_J\widetilde{\Lambda} \\ &= (\Sigma_{JJ}U_J)^\top \widetilde{U} - U_J^\top \widehat{\Sigma}_{JJ}\widetilde{U}_J \\ &= U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})\widetilde{U}_J.\end{aligned}$$

In addition,

$$U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})\widetilde{U}_J = U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})U_JU_J^\top\widetilde{U}_J + U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})(I - U_JU_J^\top)\widetilde{U}_J.$$

The second term satisfies

$$\|U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})(I - U_JU_J^\top)\widetilde{U}_J\| \le \|U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})\|\|(I - U_JU_J^\top)\widetilde{U}_J\|.$$

Note that

$$\begin{aligned}\|(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})U_J\| &\le \|\Sigma_{JJ} - \widehat{\Sigma}_{JJ}\|\|U_J\| \\ &\le \|\Sigma_{JJ} - \widehat{\Sigma}_{JJ}\|\end{aligned}$$

since U_J has orthonormal columns. Therefore, by Lemma 4,

$$\|U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})\| \lesssim \lambda_1\left(\sqrt{\frac{s\log(p)}{n}}\right). \tag{C.12}$$

Note that $\|(I - U_J U_J^\top)\widetilde{U}_J\| \lesssim \|\sin\Theta(U_J, \widetilde{U}_J)\| \lesssim \|U_J U_J^\top - \widetilde{U}_J\widetilde{U}_J^\top\|$ (see Lemma 31 in Appendix C.3). Therefore, by Proposition 1, we have that

$$\|(I - U_J U_J^\top)\widetilde{U}_J\| \lesssim \frac{\lambda_1}{\lambda_k - \lambda_{k+1}}\left(\sqrt{\frac{s\log(p)}{n}}\right). \tag{C.13}$$

In summary, we have shown so far that by (C.11), (C.12), and (C.13),

$$\begin{aligned}\|U_J(\Lambda U_J^\top\widetilde{U}_J - U_J^\top\widetilde{U}_J\widetilde{\Lambda})\widetilde{\Lambda}^{-1}\|_{2\to\infty} \lesssim{}& \frac{\|U\|_{2\to\infty}\lambda_1^2}{\lambda_k(\lambda_k - \lambda_{k+1})}\frac{s\log(p)}{n} \\ &+ \frac{\|U\|_{2\to\infty}}{\lambda_k}\|U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})U_J U_J^\top\widetilde{U}_J\|.\end{aligned}$$

Therefore, we focus on bounding the term

$$\|U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})U_J U_J^\top\widetilde{U}_J\|.$$

Naively, $\|U_J^\top\widetilde{U}_J\| \le 1$ so by submultiplicativity, we have that

$$\|U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})U_J U_J^\top\widetilde{U}_J\| \le \|U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})U_J\|.$$

For any indices i and k, the entry of the above matrix can be written as

$$\frac{1}{n}\sum_{l=1}^n\langle(U_J)_{\cdot i}, (X_l X_l^\top - \mathbb{E}(X_l X_l^\top))(U_J)_{\cdot k}\rangle = \frac{1}{n}\sum_{l=1}^n\left(((U_J)_{\cdot i}^\top X_l)(X_l^\top(U_J)_{\cdot k}) - (U_J)_{\cdot i}^\top\Sigma(U_J)_{\cdot k}\right).$$

This is a sum of independent, mean-zero subexponential random variables. Therefore, to apply the generalized Bernstein inequality (see Theorem 23 in Appendix C.3), we need to

find the ψ_1 norm of the above random variable. By properties of the ψ_1 norm, we have that

$$\begin{aligned}\|((U_J)_{\cdot i}^\top X_l)(X_l^\top (U_J)_{\cdot k}) - (U_J)_{\cdot j}^\top \Sigma (U_J)_{\cdot i}\|_{\psi_1} &\leq C\|((U_J)_{\cdot i}^\top X_l)(X_l^\top (U_J)_{\cdot k})\|_{\psi_1} \\ &\leq C\|(U_J)_{\cdot i}^\top X_l\|_{\psi_2}\|X_l^\top (U_J)_{\cdot k}\|_{\psi_2} \\ &= C\|(U_J)_{\cdot i}^\top \Sigma^{1/2} Y_l\|_{\psi_2}\|Y_l^\top \Sigma^{1/2}(U_J)_{\cdot k}\|_{\psi_2} \\ &= C\sqrt{\lambda_i\lambda_k}\|(U_J)_{\cdot i}^\top Y_l\|_{\psi_2}\|(U_J)_{\cdot k}^\top Y_l\|_{\psi_2} \\ &\leq C\sqrt{\lambda_j\lambda_k} \\ &\leq C\lambda_1\end{aligned}$$

since $X_l = \Sigma^{1/2}Y_l$, the $(U_J)_{\cdot i}$ are the eigenvectors of Σ and the vectors Y are assumed to be subgaussian with unit ψ_2 norm. Therefore, by the generalized Bernstein inequality (Theorem 23), we have that for fixed i, k, that

$$\mathbb{P}\bigg(|\frac{1}{n}\sum_{l=1}^n \langle (U_J)_{\cdot i}, (X_l X_l^\top - \mathbb{E}(X_l X_l^\top))(U_J)_{\cdot k}|\rangle \geq t\bigg) \leq 2\exp\bigg[-c_0 n \min\bigg(\frac{t^2}{(\lambda_1)^2}, \frac{t}{\lambda_1}\bigg)\bigg].$$

Since $\log(k) \ll \log(p)$, taking $t = C\lambda_1\sqrt{\frac{2\log(k)+4\log(p)}{n}}$ for some constant C yields that

$$\begin{aligned}|(U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})U_J)_{ik}| &\leq C\lambda_1\sqrt{\frac{2\log(k)+4\log(p)}{n}} \\ &\lesssim \lambda_1\sqrt{\frac{\log(p)}{n}}\end{aligned}$$

with probability at least $1 - O(p^{-4}k^{-2})$. Therefore,

$$\begin{aligned}\|U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})U_J\| &\leq \|U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})U_J\|_F \\ &\leq k\|U_J^\top(\Sigma_{JJ} - \widehat{\Sigma}_{JJ})U_J\|_{\max} \\ &\leq Ck\lambda_1\sqrt{\frac{2\log(k)+4\log(p)}{n}} \\ &\lesssim k\lambda_1\sqrt{\frac{\log(p)}{n}}\end{aligned}$$

with probability at least $1 - O(p^{-4})$ by taking a union bound over all k^2 entries. Therefore,

putting it all together, we see that

$$\|U_J(\Lambda U_J^\top \widetilde{U} - U_J^\top \widetilde{U}_J \widetilde{\Lambda})\widetilde{\Lambda}^{-1}\|_{2\to\infty} \lesssim \frac{\|U\|_{2\to\infty}\lambda_1^2}{\lambda_k(\lambda_k - \lambda_{k+1})}\frac{s\log(p)}{n} + \frac{k\lambda_1\|U\|_{2\to\infty}}{\lambda_k}\sqrt{\frac{\log(p)}{n}}$$

with probability at least $1 - O(p^{-4})$ as desired. □

Now we prove Lemma 25.

Lemma 25 (Bound on T_2)**.** *We have that*

$$\|U_J(W_* - U_J^\top \widetilde{U}_J)\|_{2\to\infty} \lesssim \frac{\|U\|_{2\to\infty}\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2}\frac{s\log(p)}{n} \equiv \mathcal{E}_2$$

with probability at least $1 - O(p^{-4})$.

Proof of Lemma 25. This proof follows similarly to ideas in Cape et al. (2019a); Abbe et al. (2020); Lei (2019).

By properties of the $2 \to \infty$ norm, we have

$$\|U_J(W_* - U_J^\top \widetilde{U}_J)\|_{2\to\infty} \le \|U_J\|_{2\to\infty}\|W_* - U_J^\top \widetilde{U}_J\|.$$

We will now analyze the term inside the spectral norm. Note that W_* is the Frobenius-optimal Procrustes transformation. Let $V_1\Sigma V_2^\top$ be the SVD of $U_J^\top \widetilde{U}_J$. Then Σ contains the sines of the canonical angles between U_J and $\widetilde{U}_J$ (see Bhatia (1997) or G. W. Stewart and J.-G. Sun (1990) for details; Lemma 31 in Appendix C.3 also contains equivalent expressions

for the $\sin\Theta$ distances). Then, letting θ_j be the canonical angles and $\sigma_j = \cos(\theta_j)$,

$$\begin{aligned}
\|W_* - U_J^\top \widetilde{U}_J\| &= \|V_1 V_2^\top - V_1 \Sigma V_2^\top\| \\
&= \|I - \Sigma\| \\
&= \max_{1\le j\le k} |(1-\sigma_j)| \\
&\le \max_{1\le j\le k} (1-\sigma_j^2) \\
&= \max_j \sin^2(\theta_j) \\
&= \|U_J U_J^\top - \widetilde{U}_J \widetilde{U}_J^\top\|^2 \\
&\lesssim \frac{\lambda_1^2}{(\lambda_k - \lambda_{k+1})^2} \frac{s\log(p)}{n}.
\end{aligned}$$

with probability at least $1 - O(p^{-4})$ by Proposition 1. □

C.2.3 Proof of Lemma 26

Recall the statement of Lemma 26.

Lemma 26 (Bound on T_3)**.** *We have that*

$$\begin{aligned}
\|U_J U_J^\top \Sigma_{JJ}(\widetilde{U}_J - U_J U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1}\|_{2\to\infty} &\lesssim \sqrt{\frac{s\log(p)}{n}} \frac{\lambda_1^{3/2}}{\lambda_k(\lambda_k - \lambda_{k+1})} \min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty}\right) \\
&\equiv \mathcal{E}_3
\end{aligned}$$

with probability at least $1 - O(p^{-3})$.

Proof of Lemma 26. Note that since $U_J^\top \Sigma_{JJ} = \Lambda U_J^\top$, we have that

$$\begin{aligned}
\|U_J U_J^\top \Sigma_{JJ}(\widetilde{U}_J - U_J U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1}\|_{2\to\infty} &\le \frac{\|U\|_{2\to\infty}}{\widetilde{\lambda}_k} \|U_J^\top \Sigma_{JJ}(\widetilde{U}_J - U_J U_J^\top \widetilde{U}_J)\| \\
&\le \frac{\|U\|_{2\to\infty}}{\widetilde{\lambda}_k} \|\Lambda U_J^\top (\widetilde{U}_J \widetilde{U}_J^\top - U_J U_J^\top)\|.
\end{aligned}$$

On the other hand,

$$\begin{aligned}\|U_J U_J^\top \Sigma_{JJ}(\widetilde{U}_J - U_J U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1}\|_{2\to\infty} &\leq \frac{\|U\Lambda^{1/2}\|_{2\to\infty}}{\widetilde{\lambda}_k}\|\Lambda^{1/2}U_J^\top(\widetilde{U}_J - U_J U_J^\top \widetilde{U}_J)\| \\ &\leq \frac{\sqrt{\|\Sigma\|_{\max}}}{\widetilde{\lambda}_k}\|\Lambda^{1/2}U_J^\top(\widetilde{U}_J\widetilde{U}_J^\top - U_J U_J^\top)\|,\end{aligned}$$

where the term $\|\Sigma\|_{\max}$ comes from the fact that

$$|U_J U_J^\top \Sigma_{i,j}| = |\langle (U\Lambda^{1/2})_i, (U\Lambda^{1/2})_j\rangle|,$$

and hence that

$$\begin{aligned}\|U|\Lambda|^{1/2}\|_{2\to\infty} &= \max_i \sqrt{\langle (U\Lambda^{1/2})_i, (U\Lambda^{1/2})_i\rangle} \\ &\leq \max_i \sqrt{|(U_J U_J^\top \Sigma)_{ii}|} \\ &\leq \max_{i,j} \sqrt{|\Sigma_{ij}|},\end{aligned}$$

since the eigenvalues of Σ are all positive. Therefore,

$$\begin{aligned}\|U_J U_J^\top \Sigma_{JJ}(\widetilde{U}_J - U_J U_J^\top \widetilde{U}_J)\widetilde{\Lambda}^{-1}\|_{2\to\infty} \leq \frac{1}{\widetilde{\lambda}_k}\min\Big(&\sqrt{\lambda_1}\|U\|_{2\to\infty}\|\Lambda^{1/2}U_J^\top(\widetilde{U}_J\widetilde{U}_J^\top - U_J U_J^\top)\|, \\ &\|\Sigma\|_{\max}^{1/2}\|\Lambda^{1/2}U_J^\top(\widetilde{U}_J\widetilde{U}_J^\top - U_J U_J^\top)\|\Big)\end{aligned} \tag{C.14}$$

Therefore, what remains is to analyze

$$\|\Lambda^{1/2}U_J^\top(\widetilde{U}_J\widetilde{U}_J^\top - U_J U_J^\top)\|.$$

To find this bound, we will represent the difference $\widetilde{U}_J\widetilde{U}_J^\top - U_J U_J^\top$ using the holomorphic functional calculus as done in Lei (2019) for the spiked Wigner matrix setting. This technique has been used extensively in studying eigenvector perturbation; e.g. Mao et al. (2020); Lei (2019); Koltchinskii and Xia (2016); Xia (2021); Wahl (2019a,b). More specifically, let $\mathcal{C}$ denote a contour on the complex plane with real part ranging from $\lambda_k - \eta$ to $\lambda_1 + \eta$, and with

imaginary part ranging from $-\gamma$ to γ. Then, for a proper choice of η, the top k eigenvalues of both Σ_{JJ} and $\widehat{\Sigma}_{JJ}$ lie in $\mathcal{C}$, and one can write the difference of the spectral projections via a complex integral

$$\widetilde{U}_J\widetilde{U}_J^\top - U_JU_J^\top = -\left[\frac{1}{2\pi i}\oint_{\mathcal{C}}(\widehat{\Sigma}_{JJ}-zI)^{-1}dz - \frac{1}{2\pi i}\oint_{\mathcal{C}}(\Sigma_{JJ}-zI)^{-1}dz\right]$$

by the residue theorem (e.g. (Greene and Krantz, 2006)). Using the identity $A^{-1}-B^{-1} = B^{-1}(A-B)A^{-1}$, and assuming the real number η is chosen appropriately so that the contours are the same, the integrals can be combined to arrive at the expression

$$\widetilde{U}_J\widetilde{U}_J^\top - U_JU_J^\top = -\frac{1}{2\pi i}\oint_{\mathcal{C}}(\Sigma_{JJ}-zI)^{-1}(\widehat{\Sigma}_{JJ}-\Sigma_{JJ})(\widehat{\Sigma}_{JJ}-zI)^{-1}dz.$$

Premultiplying by $\Lambda^{1/2}U_J^\top$ yields (formally) that

$$\|\Lambda^{1/2}U_J^\top(\widetilde{U}_J\widetilde{U}_J^\top - U_JU_J^\top)\| = \frac{1}{2\pi}\left\|\oint_{\mathcal{C}}\Lambda^{1/2}U_J^\top(\Sigma_{JJ}-zI)^{-1}(\widehat{\Sigma}_{JJ}-\Sigma_{JJ})(\widehat{\Sigma}_{JJ}-zI)^{-1}dz\right\|.$$

Note that the matrix is diagonalizable by the same eigenvectors as Σ_{JJ}, so that

$$\begin{aligned}U_J^\top(\Sigma_{JJ}-zI)^{-1} &= U_J^\top(U_J(\Lambda-zI)^{-1}U_J^\top) + U_J^\top(U_\perp(\Lambda_\perp - zI)^{-1}U_\perp^\top \\ &= (\Lambda-zI)^{-1}U_J^\top\end{aligned}$$

by orthonormality, where $U_\perp$ are defined as the $s\times s$ completion of U_J such that $[U_J, U_\perp]$ is an $s\times s$ orthogonal matrix. Therefore, we have

$$\|\Lambda^{1/2}U_J^\top(\widetilde{U}_J\widetilde{U}_J^\top - U_JU_J^\top)\| = \frac{1}{2\pi}\left\|\oint_{\mathcal{C}}\Lambda^{1/2}(\Lambda-zI)^{-1}U_J^\top(\widehat{\Sigma}_{JJ}-\Sigma_{JJ})[\widetilde{U},\widetilde{U}_\perp](\widehat{\Lambda}_{all}-zI)^{-1}dz\right\|,$$

where $\widehat{\Lambda}_{all}$ is the diagonal matrix of all the eigenvalues of $\widehat{\Sigma}_{JJ}$.

The rest of the proof mirrors closely that of Lemma A.8 in Lei (2019). Recall that in order to do all these manipulations, we required that the parameter η was chosen such that the contour $\mathcal{C}$ contains the top k eigenvalues of $\widehat{\Sigma}_{JJ}$ and Σ_{JJ}. In fact, Lemma 5 shows that

the choice

$$\eta := \frac{\lambda_k - \lambda_{k+1}}{4}$$

suffices. To see this, note that by Lemmas 4 and 5,

$$|\widetilde{\lambda}_k - \lambda_k| \leq \frac{\lambda_k - \lambda_{k+1}}{8};$$
$$|\widetilde{\lambda}_{k+1} - \lambda_{k+1}| \leq \frac{\lambda_k - \lambda_{k+1}}{8},$$

so that the interval $\lambda_k \pm \eta$ contains $\widetilde{\lambda}_k$, the interval $\lambda_k \pm \eta$ does not contain $\widetilde{\lambda}_{k+1}$, and both $\widetilde{\lambda}_k$ and $\widetilde{\lambda}_{k+1}$ satisfy

$$|\lambda_k - \widetilde{\lambda}_k - \eta| \geq \eta/2$$
$$|\lambda_k - \widetilde{\lambda}_{k+1} - \eta| \geq \eta/2.$$

Therefore, the top k eigenvalues of $\widehat{\Sigma}_{JJ}$ lie within $\mathcal{C}$ with high probability and the bottom eigenvalues lie outside of it. With this particular choice of η, we can proceed to bound the integrand along the contour $\mathcal{C}$. We will conduct the rest of the analysis assuming that this event holds; it does with probability at least $1 - O(p^{-4})$.

Define $a := \lambda_k - \eta$ and $b := \lambda_1 + \eta$. We decompose the contour $\mathcal{C}$ into the following sets

$$\mathcal{C}_1 := \{z = a + xi, x \in (-\gamma, \gamma)\} \qquad \mathcal{C}_2 := \{z = x + \gamma i : x \in [a, b]\}$$
$$\mathcal{C}_3 := \{z = b + xi, x \in (-\gamma, \gamma)\} \qquad \mathcal{C}_4 := \{z = x - \gamma i : x \in [a, b]\}.$$

Let $\mathcal{I}(z)$ be the integrand. Observe that

$$\left\| \oint_{\mathcal{C}} \mathcal{I}(z)dz \right\| \leq \oint_{\mathcal{C}_1} \left\| \mathcal{I}(z)dz \right\| + \oint_{\mathcal{C}_2} \left\| \mathcal{I}(z)dz \right\| + \oint_{\mathcal{C}_4} \left\| \mathcal{I}(z)dz \right\| + \oint_{\mathcal{C}_4} \left\| \mathcal{I}(z)dz \right\|.$$

Therefore, we bound the above integrals directly. The tricky analysis will be along $\mathcal{C}_1$ and $\mathcal{C}_3$; we will show that the integral along $\mathcal{C}_2$ and $\mathcal{C}_4$ tend to zero for large γ. To this end, we

will focus on $\mathcal{C}_1$ first. Note that

$$\begin{aligned}\oint_{\mathcal{C}_1} \left\|\Lambda^{1/2}(\Lambda - zI)^{-1}U_J^\top(\widehat{\Sigma}_{JJ} - \Sigma_{JJ})[\widetilde{U}, \widetilde{U}_\perp](\widehat{\Lambda}_{all} - zI)^{-1}\right\| dz \qquad \text{(C.15)}\\ \leq \oint_{\mathcal{C}_1} \left\|\Lambda^{1/2}(\Lambda - zI)^{-1}\right\|\left\|U_J^\top(\widehat{\Sigma}_{JJ} - \Sigma_{JJ})[\widetilde{U}, \widetilde{U}_\perp]\right\|\left\|(\widehat{\Lambda}_{all} - zI)^{-1}\right\| dz \\ \leq \left\|U_J^\top(\widehat{\Sigma}_{JJ} - \Sigma_{JJ})[\widetilde{U}, \widetilde{U}_\perp]\right\| \int_{-\gamma}^{\gamma} \left\|\Lambda^{1/2}(\Lambda - (a+xi)I)^{-1}\right\|\left\|(\widehat{\Lambda}_{all} - (a+xi)I)^{-1}\right\| dx.\end{aligned}$$

First, recall the definition of $a := \lambda_k - \eta$. The term on the right-most side satisfies

$$\left\|(\widehat{\Lambda}_{all} - (a+xi)I)^{-1}\right\| \leq \frac{1}{\sqrt{(\eta)^2/4 + x^2}}$$

for all x since $(\widehat{\lambda}_i - a) \geq \eta/2$. Therefore, we are left to bound the middle term, for which we must bound

$$\max_{1\leq i\leq k} \frac{\lambda_i^{1/2}}{\sqrt{(\lambda_i - a)^2 + x^2}}.$$

Define the function

$$g(u; x, a) := \frac{u}{\sqrt{(u-a)^2 + x^2}}.$$

Then

$$\max_{1\leq i\leq k} \frac{\lambda_i^{1/2}}{\sqrt{(\lambda_i - a)^2 + x^2}} \leq \sup_{u\geq a+\eta} \Big(g(u; x, a)\Big)^{1/2} \frac{1}{(\eta^2 + x^2)^{1/4}}.$$

The details of the function g are carried out in Lei (2019); the analysis therein implies

$$\sup_{u\geq a+\eta} g(u; x, a) \leq \frac{a+\eta}{\sqrt{\eta^2 + x^2}} \mathbb{I}_{|x|\leq\sqrt{a\eta}} + \sqrt{\frac{a+\eta}{\eta}} \mathbb{I}_{|x|>\sqrt{a\eta}}.$$

Therefore the integral from (C.15) satisfies

$$\begin{aligned}
&\int_{-\gamma}^{\gamma} \|\Lambda^{1/2}(\Lambda-(a+xi)I)^{-1}\|\|(\widehat{\Lambda}_{all}-(a+xi)I)^{-1}\|dx \\
&\quad\le \int_{-\gamma}^{\gamma} \frac{1}{\sqrt{\eta^2/4+x^2}}\frac{1}{(\eta^2+x^2)^{1/4}}\left(\frac{a+\eta}{\sqrt{\eta^2+x^2}}\mathbb{I}_{|x|\le\sqrt{a\eta}}+\sqrt{\frac{a+\eta}{\eta}}\mathbb{I}_{|x|>\sqrt{a\eta}}\right)^{1/2}dx \\
&\quad\le \int_{|x|\le\sqrt{a\eta}} \frac{4}{(\eta^2+x^2)^{3/4}}\left(\frac{a+\eta}{\sqrt{\eta^2+x^2}}\right)^{1/2}dx+\int_{|x|>\sqrt{a\eta}}\frac{4}{(\eta^2+x^2)^{3/4}}\left(\sqrt{\frac{a+\eta}{\eta}}\right)^{1/2}dx \\
&\quad\le 4\sqrt{a+\eta}\int_{|x|\le\sqrt{a\eta}}\frac{1}{\eta^2+x^2}dx+4\left(\frac{a+\eta}{\eta}\right)^{1/4}\int_{|x|>\sqrt{a\eta}}\frac{1}{(\eta^2+x^2)^{3/4}}dx \\
&\quad\le 8\sqrt{a+\eta}\int_0^{\sqrt{a\eta}}\frac{1}{\eta^2+x^2}dx+8\left(\frac{a+\eta}{\eta}\right)^{1/4}\int_{\sqrt{a\eta}}^{\infty}\frac{1}{(\eta^2+x^2)^{3/4}}dx \\
&\quad\le 8\frac{\sqrt{a+\eta}}{\eta}\int_0^{\sqrt{a/\eta}}\frac{1}{1+u^2}du+8\left(\frac{a+\eta}{\eta}\right)^{1/4}\frac{1}{\eta^{1/2}}\int_{\sqrt{a/\eta}}^{\infty}\frac{1}{(1+u^2)^{3/4}}du \\
&\quad\le 8\frac{\sqrt{a+\eta}}{\eta}2\pi+8\left(\frac{a+\eta}{\eta}\right)^{1/4}\frac{1}{\eta^{1/2}}\int_{\sqrt{a/\eta}}^{\infty}\frac{1}{u^{3/2}}du \\
&\quad\le 16\pi\frac{\sqrt{a+\eta}}{\eta}+8\left(\frac{a+\eta}{\eta}\right)^{1/4}\frac{1}{\eta^{1/2}}\frac{2}{(a/\eta)^{1/2}} \\
&\quad\lesssim \frac{\sqrt{a+\eta}}{\eta}+\left(\frac{a+\eta}{a}\right)^{1/4}\frac{1}{a^{1/2}}.
\end{aligned}$$

Recall that $a+\eta=\lambda_k$; $\eta=(\lambda_k-\lambda_{k+1})/4$. With these, the bound becomes (up to constants)

$$\begin{aligned}
&\oint_{\mathcal{C}_1}\left\|\Lambda^{1/2}(\Lambda-zI)^{-1}U_J^\top(\widehat{\Sigma}_{JJ}-\Sigma_{JJ}(\widetilde{U},\widetilde{U}_\perp)(\widehat{\Lambda}_{all}-zI)^{-1}\right\|dz \\
&\qquad\lesssim \frac{\sqrt{\lambda_1}}{\lambda_k-\lambda_{k+1}}\|(\widehat{\Sigma}_{JJ}-\Sigma_{JJ})U_J\|+\kappa^{1/4}\frac{1}{\lambda_k^{1/2}}\|(\widehat{\Sigma}_{JJ}-\Sigma_{JJ})U_J\| \\
&\qquad\lesssim \frac{\sqrt{\lambda_1}}{\lambda_k-\lambda_{k+1}}\|(\widehat{\Sigma}_{JJ}-\Sigma_{JJ})U_J\|.
\end{aligned}$$

The exact same argument goes through for contour $\mathcal{C}_3$ as well. We will see that the contours along the imaginary axis tend to zero as $\gamma\to\infty$. Assuming this for the moment, by Equation (C.14), we see that the final bound is of the form

$$\begin{aligned}
&\frac{1}{\lambda_k}\|\Lambda^{1/2}U_J^\top(\widetilde{U}_J\widetilde{U}_J^\top-U_JU_J^\top)\|\min\left(\|\Sigma\|_{\max}^{1/2},\sqrt{\lambda_1}\|U\|_{2\to\infty}\right) \\
&\qquad\lesssim \frac{\|(\widehat{\Sigma}_{JJ}-\Sigma_{JJ})U_J\|}{\lambda_k}\left(\frac{\sqrt{\lambda_1}}{\lambda_k-\lambda_{k+1}}\right)\min\left(\|\Sigma\|_{\max}^{1/2},\sqrt{\lambda_1}\|U\|_{2\to\infty}\right)
\end{aligned}$$

By Lemma 4, we have that the term $\|(\widehat{\Sigma}_{JJ} - \Sigma_{JJ})U_J\|$ can be bounded via

$$\lambda_1 \sqrt{\frac{s\log(p)}{n}}$$

with probability at least $1 - O(p^{-4})$. Therefore, the bound becomes

$$\sqrt{\frac{s\log(p)}{n}} \frac{\lambda_1^{3/2}}{\lambda_k(\lambda_k - \lambda_{k+1})} \min\left(\|\Sigma\|_{\max}^{1/2}, \sqrt{\lambda_1}\|U\|_{2\to\infty} \right),$$

which is the desired bound.

It remains to show that the integrals tend to zero along the curves $\mathcal{C}_2$ and $\mathcal{C}_4$. Let $\mathcal{I}(z)$ denote the integrand. Then for sufficiently large γ,

$$\begin{aligned}
\oint_{\mathcal{C}_2} \|\mathcal{I}(z)\| dz &= \int_a^b \left\| \Lambda^{1/2}(\Lambda - (x+\gamma i)I)^{-1} U_J^\top (\widehat{\Sigma}_{JJ} - \Sigma_{JJ})[\widetilde{U}_J \widetilde{U}_\perp](\widehat{\Lambda}_{all} - (x+\gamma i)I)^{-1} \right\| dx \\
&\le (b-a) \sup_{x\in[a,b]} \left\| \Lambda^{1/2}(\Lambda - (x+\gamma i)I)^{-1} U_J^\top (\widehat{\Sigma}_{JJ} - \Sigma_{JJ})[\widetilde{U}_J \widetilde{U}_\perp](\widehat{\Lambda}_{all} - (x+\gamma i)I)^{-1} \right\| \\
&= O(\gamma^{-2}),
\end{aligned}$$

which tends to zero as $\gamma \to \infty$. □

C.2.4 Proof of Lemma 27

First, recall the statement of Lemma 27.

Lemma 27 (Bound on J_2)**.** *The term J_2 satisfies*

$$\begin{aligned}
\|U_\perp \Lambda_\perp U_\perp^\top \widetilde{U}_J\|_{2\to\infty} &\lesssim \kappa^2 \lambda_{k+1} \sqrt{\frac{k\log(p)}{n}} + \lambda_{k+1}\kappa^3 \frac{s\log(p)}{n} \\
&\lesssim \mathcal{E}_4 \lambda_k
\end{aligned}$$

with probability at least $1 - O(p^{-3})$.

Recall the definition of J_2 via

$$J_2 := U_\perp \Lambda_\perp U_\perp^\top \widetilde{U}_J.$$

Again $U_\perp$ is the matrix such that the $s \times s$ matrix $[U_J, U_\perp]$ is orthogonal.

Proof of Lemma 27. Define the matrix $E := \widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top$. Note that

$$\widetilde{U}_J \Lambda - E\widetilde{U}_J = U_J U_J^\top \Sigma_{JJ} U_J U_J^\top \widetilde{U}_J.$$

Following Cape et al. (2019a) (see also Xie et al. (2022); Tang et al. (2017c); Tang (2018)), by Assumption 3.4, the spectra of E and Λ are disjoint almost surely, so the matrix $\widetilde{U}$ can be expanded as a matrix series (Theorem VII.2.2 in Bhatia (1997)) via

$$\widetilde{U}_J = \sum_{m=0}^{\infty} E^m (U_J \Lambda U_J^\top) \widetilde{U}_J \Lambda^{-(m+1)}.$$

Therefore,

$$J_2 = U_\perp \Lambda_\perp U_\perp^\top \widetilde{U}_J = U_\perp \Lambda_\perp U_\perp^\top E U \Lambda U^\top \widetilde{U} \Lambda^{-2} + \sum_{m=2}^{\infty} U_\perp \Lambda_\perp U_\perp^\top E^m U \Lambda U_J^\top \widetilde{U}_J \Lambda^{-(m+1)}$$

since the 0-th term cancels off because $U_\perp^\top U_J = 0$. Taking the first term and setting R to be the rest of the series, we have that,

$$\|U_\perp \Lambda_\perp U_\perp^\top \widetilde{U}_J \widetilde{U}_J^\top\|_{2\to\infty} = \|U_\perp \Lambda_\perp U_\perp^\top E U_J \Lambda U_J^\top \widetilde{U} \Lambda^{-2}\|_{2\to\infty} + \|R\|_{2\to\infty}, \tag{C.16}$$

where R is the residual to be bounded. We first bound the leading term. We have that

$$\|U_\perp \Lambda_\perp U_\perp^\top E U_J \Lambda U_J^\top \widetilde{U}_J \Lambda^{-2}\|_{2\to\infty} \le \|U_\perp \Lambda_\perp U_\perp^\top E U_J\|_{2\to\infty} \lambda_k^{-1} \kappa. \tag{C.17}$$

We note that since $U_\perp^\top U_J = 0$, then

$$E U_J = (\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top) U_J = (\widehat{\Sigma}_{JJ} - \Sigma_{JJ}) U_J.$$

Define $\Sigma_{JJ}^\perp := U_\perp \Lambda_\perp U_\perp^\top$. In light of the block structure in (C.5), we see that we can write

$\Sigma_{JJ}^{\perp} E U_J$ via the sum of the terms

$$\frac{1}{n}(\Sigma_{JJ}^{\perp})\Big((\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ} + \Sigma_{JJ^c}^{1/2} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} \\ + (\Sigma^{1/2})_{JJ} Y_J^\top Y_{J^c} (\Sigma_{JJ^c}^{1/2})^\top + \Sigma_{JJ^c}^{1/2}(Y_{J^c}^\top Y_{J^c} - nI)(\Sigma_{JJ^c}^{1/2})^\top \Big) U_J.$$

Recalling that $(\Sigma_{JJ^c}^{1/2})^\top U_J = 0$, this yields the only the terms

$$\frac{1}{n}(\Sigma_{JJ}^{\perp})\Big((\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ} + \Sigma_{JJ^c}^{1/2} Y_{J^c}^\top Y_J (\Sigma^{1/2})_{JJ} \Big) U_J = \Sigma_{JJ}^{\perp}(\Sigma^{1/2})_{JJ}\Big(\frac{Y_J^\top Y_J}{n} - I \Big) U_J \Lambda^{1/2} \\ + \Sigma_{JJ}^{\perp} \Sigma_{JJ^c}^{1/2} \frac{Y_{J^c}^\top Y_J}{n} U_J \Lambda^{1/2}.$$

Define $A_{3/2} := \Sigma_{JJ}^{\perp}(\Sigma^{1/2})_{JJ}$, which satisfies $\|A_{3/2}\| \leq \sqrt{\lambda_1}\lambda_{k+1}$. In $2 \to \infty$ norm, we have that

$$\|A_{3/2}\Big(\frac{Y_J^\top Y_J}{n} - I \Big) U_J \Lambda^{1/2}\|_{2\to\infty} \leq \sqrt{k\lambda_1} \max_{i,j} \Big| \Big(A_{3/2}\Big(\frac{Y_J^\top Y_J}{n} - I \Big) U_J \Big)_{ij} \Big|.$$

Define the matrix M via $M_{kl} := (A_{3/2})_{ik} U_{lj}$. Fixing i and j, note that we can write the i,j entry above as

$$\begin{aligned} \Big| \sum_{k,l} M_{kl} \Big(\frac{1}{n} (\sum_{q=1}^n (Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk}) \Big) \Big| &= \frac{1}{n} \Big| \sum_q \sum_{k,l} M_{kl} \Big(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk} \Big) \Big| \\ &\leq \frac{1}{n} \sum_q \Big| \sum_{k,l} M_{kl} \Big(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk} \Big) \Big| \\ &\leq \max_q \Big| \sum_{k,l} M_{kl} \Big(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk} \Big) \Big|, \end{aligned}$$

which is a quadratic form in the random variables Y_{ql} (for fixed q). To bound this, we will apply the Hanson-Wright inequality (Theorem 24 in Appendix C.3), which requires

bounding the Frobenius norm of M. Note that we can bound the Frobenius norm of M via

$$\begin{aligned}\|M\|_F^2 &= \sum_{k,l} M_{kl}^2 \\ &= \sum_{k,l} (A_{3/2})_{ik}^2 U_{lj}^2 \\ &= \|A_{3/2}\|_{2\to\infty}^2 \\ &\le \left(\sqrt{\lambda_1}\lambda_{k+1}\right)^2.\end{aligned}$$

Therefore, applying the Hanson-Wright inequality shows that

$$\mathbb{P}\left(\left|\sum_{k,l} M_{kl}\left(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk}\right)\right| > t\right) \le 2\exp\left(-c\min\left\{\frac{t^2}{\|M\|_F^2}, \frac{t}{\|M\|}\right\}\right).$$

Set $t := C\sqrt{\frac{\log(s)+\log(k)+5\log(p)}{n}}\sqrt{\lambda_1}\lambda_{k+1}$. Then since $\frac{\log(p)}{n} = o(1)$, we see that with probability at least $1 - O(s^{-1}k^{-1}p^{-5})$ that

$$\left|\sum_{k,l} M_{kl}\left(Y_{ql}Y_{qk} - \mathbb{E}Y_{ql}Y_{qk}\right)\right| \lesssim \sqrt{\lambda_1}\lambda_{k+1}\sqrt{\frac{\log(p)}{n}}.$$

Taking a union bound over all n random variables shows that with probability at least $1 - O(s^{-1}k^{-1}p^{-4})$,

$$\sqrt{k\lambda_1}\left|\left(A_{3/2}\left(\frac{Y_J^\top Y_J}{n} - I\right)U_J\right)_{ij}\right| \lesssim \lambda_1\lambda_{k+1}\sqrt{\frac{k\log(p)}{n}}.$$

Taking a union bound over all s rows and k columns yields that with probability at least $1 - O(p^{-4})$,

$$\|A_{3/2}\left(\frac{Y_J^\top Y_J}{n} - I\right)U_J\Lambda^{1/2}\|_{2\to\infty} \lesssim \lambda_{k+1}\lambda_1\sqrt{\frac{k\log(p)}{n}}. \tag{C.18}$$

For the other term, proceeding similarly,

$$\|\Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2}\frac{Y_{J^c}^\top Y_J}{n}U_J\Lambda^{1/2}\|_{2\to\infty} \le \sqrt{\lambda_1 k}\max_{i,j}\left|\left((\Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2})\frac{Y_{J^c}^\top Y_J}{n}U_J\right)_{ij}\right| \\ \le \sqrt{\lambda_1 k}\max_{i,j}\max_{q}\left|\sum_{k=s+1}^{p}\sum_{l=1}^{s}(\Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2})_{ik}Y_{qk}Y_{ql}(U_J)_{lj}\right|.$$

For fixed q, i, and j, note that k ranges from $s+1$ to p and l ranges from 1 to s, so this is a sum of independent exponential random variables. We will bound these using Bernstein's inequality (Theorem 23 in Appendix C.3). Note that the ℓ_2 norm of the coefficients is bounded by

$$\sum_{k=s+1}^{p}\sum_{l=1}^{s}(\Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2})_{ik}^2(U_J)_{lj}^2 \le \|\Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2}\|_{2\to\infty}^2 \le (2\sqrt{\lambda_1}\lambda_{k+1})^2.$$

Similarly,

$$\max_{k,l}|(\Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2})_{ik}(U_J)_{lj}| \le \|U_J\|_{2\to\infty}\max_{i,k}|e_i^\top(\Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2})e_k| \\ \le 2\|U_J\|_{2\to\infty}\sqrt{\lambda_1}\lambda_{k+1}.$$

By the generalized Bernstein Inequality (Theorem 23), we have for any fixed i,j, and q that

$$\mathbb{P}\left(\left|\sum_{k=s+1}^{p}\sum_{l=1}^{s}(\Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2})_{ik}Y_{qk}Y_{ql}(U_J)_{lj}\right| > t\right) \le 2\exp\left[-c\min\left(\frac{t^2}{(\sqrt{\lambda_1}\lambda_{k+1})^2}, \frac{t}{\|U\|_{2\to\infty}\sqrt{\lambda_1}\lambda_{k+1}}\right)\right].$$

Taking $t = \sqrt{\lambda_1}\lambda_{k+1}\sqrt{\frac{\log(s)+\log(k)5\log(p)}{n}}$ shows that this holds with probability at least $1 - O(s^{-1}k^{-1}p^{-5})$. Taking a union bound over s rows, k columns, and n different random variables shows that with probability at least $1 - O(p^{-4})$ that

$$\|\Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2}\frac{Y_{J^c}^\top Y_J}{n}U_J\Lambda^{1/2}\|_{2\to\infty} \le \sqrt{\lambda_1 k}\max_{i,j}\left|\left((\Sigma_{JJ}^{\perp}\Sigma_{JJ^c}^{1/2})\frac{Y_{J^c}^\top Y_J}{n}U_J\right)_{ij}\right| \\ \lesssim \lambda_{k+1}\lambda_1\sqrt{\frac{k\log(p)}{n}} \tag{C.19}$$

Combining (C.19) and (C.18) with (C.17) yields that

$$\begin{aligned}\|U_\perp \Lambda_\perp U_\perp^\top E U_J \Lambda U_J^\top \widetilde{U}_J \Lambda^{-2} U_J^\top\|_{2\to\infty} &\lesssim \frac{\kappa}{\lambda_k}\|U_\perp \Lambda_\perp U_\perp^\top E U_J\|_{2\to\infty} \\ &\lesssim \frac{\kappa}{\lambda_k}\left(\lambda_1\lambda_{k+1}\sqrt{\frac{k\log(p)}{n}}\right) \\ &\lesssim \kappa^2\lambda_{k+1}\sqrt{\frac{k\log(p)}{n}}. \end{aligned} \tag{C.20}$$

So what remains is to bound the residual term R in (C.16). Recall the definition of R via

$$R := \sum_{m=2}^{\infty} U_\perp \Lambda_\perp U_\perp^\top E^m U_J \Lambda U_J^\top \widetilde{U}_J \Lambda^{-(m+1)}.$$

We will bound for each m, but for clarity, we will first study the case $m = 2$. We have that

$$\begin{aligned} U_\perp \Lambda_\perp U_\perp^\top E^2 U_J &= U_\perp \Lambda_\perp U_\perp^\top (\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top)(\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top) U_J \\ &= U_\perp \Lambda_\perp U_\perp^\top (\widehat{\Sigma}_{JJ} - U_J U_J^\top \Sigma_{JJ} U_J U_J^\top)(\widehat{\Sigma}_{JJ} - \Sigma_{JJ}) U_J \\ &= U_\perp \Lambda_\perp U_\perp (\widehat{\Sigma}_{JJ} - \Sigma_{JJ})^2 U_J + (U_\perp \Lambda_\perp U_\perp)^2 (\widehat{\Sigma}_{JJ} - \Sigma_{JJ}) U_J. \end{aligned}$$

The first term is readily bounded by Lemma 4, and the second term can be bounded using the techniques in the previous part of the proof of this Lemma.

We now generalize this strategy for each m, by first providing a similar identity to the one above. Define $\Delta := \widehat{\Sigma}_{JJ} - \Sigma_{JJ}$. Note that by definition $E = \Delta + U_\perp \Lambda_\perp U_\perp^\top$ and that $EU_J = \Delta U_J$. Then we have that

$$\begin{aligned} U_\perp \Lambda_\perp U_\perp^\top E^m U_J &= U_\perp \Lambda_\perp^\top U_\perp E^{m-1} \Delta U_J \\ &= U_\perp \Lambda_\perp U_\perp^\top E^{m-2}(\Delta + U_\perp \Lambda_\perp U_\perp)\Delta U_J \\ &= U_\perp \Lambda_\perp U_\perp^\top E^{m-2} \Delta U_J + U_\perp \Lambda_\perp U_\perp^\top E^{m-2} U_\perp \Lambda_\perp U_\perp^\top \Delta U_J. \end{aligned} \tag{C.21}$$

Let $\mathfrak{s}(m)$ be the set of indices such that $s_1 + ... + s_m = m$. Then for all m we have that

$$U_\perp \Lambda_\perp U_\perp^\top E^m U_J = U_\perp \Lambda_\perp U_\perp^\top \left[\sum_{\mathfrak{s}(m)} \Delta^{s_1}(U_\perp \Lambda_\perp U_\perp^\top)^{s_2} \Delta^{s_3}(U_\perp \Lambda_\perp U_\perp^\top)^{s_4} \cdots (U_\perp \Lambda_\perp U_\perp^\top)^{s_{m-1}} \Delta^{s_m}\right] U_J,$$

which is essentially a noncommutative Binomial Theorem.

First, consider the case that $s_1, ..., s_m$ has only single powers of Δ appearing. If Δ appears all the way on the right hand side; that is, $s_m = 1$, then for any integer m_0, we have that

$$\|U_\perp \Lambda_\perp^{m_0} U_\perp^\top \Delta U_J\|_{2\to\infty} \leq C\lambda_{k+1}^{m_0}\left(\lambda_1\sqrt{\frac{k\log(p))}{n}}\right),$$

with probability at least $1 - O(p^{-4})$ using analogous techniques to the steps leading up to Equation (C.20) (i.e. the case $m_0 = 1$). If Δ is not on the right hand side, suppose that its index is $s_g = 1$. Then this term is of the form

$$(U_\perp \Lambda_\perp U_\perp^\top)^{1+s_1+s_2+...+s_{g-1}} \Delta (U_\perp \Lambda_\perp U_\perp^\top)^{s_{g+1}+...+s_{m_0}} U_J \equiv 0$$

since $U_\perp^\top U_J = 0$. So the only terms that have at most one factor of Δ appearing are those that show up as ΔU_J.

Next, if $s_1, ..., s_m$ is a set of integers and at least two of the terms s_i that appear on the Δ factor are greater than 1, then

$$\|U_\perp \Lambda_\perp U_\perp^\top \Delta^{s_1} (U_\perp \Lambda_\perp U_\perp^\top)^{s_2} \Delta^{s_3} (U_\perp \Lambda_\perp U_\perp^\top)^{s_4} \cdots (U_\perp \Lambda_\perp U_\perp^\top)^{s_{m-1}} \Delta^{s_m} U_J\|_{2\to\infty} \leq \|\Delta\|^2 \lambda_{k+1}^{m-1},$$

provided that $\|\Delta\| < \lambda_{k+1}$, which happens by Assumption 3.1 and the spectral norm concentration in Lemma 4 with probability at least $1 - O(p^{-4})$. Fix this event. Then for any m, there are at most 2^m ways to select exponents with a power of at least two on the term $\|\Delta\|$. Therefore, this implies that for fixed m

$$\begin{aligned}\|U_\perp \Lambda_\perp U_\perp^\top E^m U_J\|_{2\to\infty} &\leq \|U_\perp \Lambda_\perp^m U_\perp^\top \Delta U_J\| \\ &\quad + \sum_{\{m\text{:exponent on } \|\Delta\| \text{ is at least 2}\}} \|U_\perp \Lambda_\perp U_\perp^\top \Delta^{s_1} \cdots (U_\perp \Lambda_\perp U_\perp^\top)^{s_{m-1}} \Delta^{s_m} U_J\|_{2\to\infty} \\ &\leq C\lambda_{k+1}^m \left(\lambda_1\sqrt{\frac{k\log(p))}{n}}\right) + 2^m \lambda_{k+1}^{m-1}\|\Delta\|^2.\end{aligned}$$

This bound corresponds to one such m, and hence is its own event. In order to bound for

all m, we follow a strategy in Xia and Yuan (2020). Let $\widetilde{m} := \lceil \log(p) \rceil$. Then

$$\begin{aligned}
&\| \sum_{m=2}^{\infty} U_\perp \Lambda_\perp U_\perp^\top E^m U_J \Lambda U_J^\top \widetilde{U}_J \Lambda^{-(m+1)} \|_{2\to\infty} \\
&\quad \le \sum_{m=2}^{\infty} \|U_\perp \Lambda_\perp U_\perp^\top E^m U_J\|_{2\to\infty} \frac{\lambda_1}{\lambda_k^{m+1}} \\
&\quad \le \sum_{m=2}^{\widetilde{m}} \|U_\perp \Lambda_\perp U_\perp^\top E^m U_J\|_{2\to\infty} \frac{\lambda_1}{\lambda_k^{m+1}} + \sum_{m=\widetilde{m}}^{\infty} \|U_\perp \Lambda_\perp U_\perp^\top E^m U_J\|_{2\to\infty} \frac{\lambda_1}{\lambda_k^{m+1}} \\
&\quad \le \sum_{m=2}^{\widetilde{m}} \left(C\lambda_{k+1}^m \left(\lambda_1 \sqrt{\frac{k\log(p))}{n}} \right) \frac{\lambda_1}{\lambda_k^{m+1}} \right. \\
&\qquad + \sum_{m=2}^{\widetilde{m}} \left(2^m \lambda_{k+1}^{m-1} \|\Delta\|^2 \right) \frac{\lambda_1}{\lambda_k^{m+1}} \\
&\qquad + \sum_{m=\widetilde{m}}^{\infty} \frac{\lambda_1}{\lambda_k^{m+1}} \|\Delta\| \lambda_{k+1}^{m+1}.
\end{aligned}$$

Define

$$\begin{aligned}
\delta_1 &:= C\kappa \left(\lambda_1 \sqrt{\frac{k\log(p))}{n}} \right) \\
\delta_2 &:= \kappa \lambda_k^{-1} \|\Delta\|^2
\end{aligned}$$

Then the three sums above can be written as

$$\begin{aligned}
\delta_1 \sum_{m=2}^{\widetilde{m}} \frac{\lambda_{k+1}^m}{\lambda_k^m} + \delta_2 \sum_{m=2}^{\widetilde{m}} \frac{2^m \lambda_{k+1}^{m-1}}{\lambda_k^{m-1}} + \lambda_1 \|\Delta\| \sum_{m=\widetilde{m}}^{\infty} \frac{\lambda_{k+1}^{m+1}}{\lambda_k^{m+1}} &\lesssim \delta_1 \frac{\lambda_{k+1}^2}{\lambda_k^2} + \delta_2 (1+\varepsilon) \frac{\lambda_{k+1}}{\lambda_k} + \lambda_1 \|\Delta\| \left(\frac{\lambda_{k+1}}{\lambda_k} \right)^{\log(p)} \\
&\lesssim \delta_1 \frac{\lambda_{k+1}^2}{\lambda_k^2} + \delta_2 \frac{\lambda_{k+1}}{\lambda_k} + \lambda_1^2 \sqrt{\frac{s\log(p)}{n}} (1-\varepsilon)^{\log(p)}.
\end{aligned}$$

Here, the penultimate inequality follows from the fact that by Assumption 3.4, we have that for some $\varepsilon > 1/64$, $2\lambda_{k+1}/\lambda_k < 1-\varepsilon$, and hence the second term's geometric series converges. The final inequality follows from the assumption $\lambda_{k+1}/\lambda_k < (1-\varepsilon)$. Note that this event holds with probability at least $1 - O(\log(p)p^{-4}) \ge 1 - O(p^{-3})$. Noting that

$$\|\Delta\| \lesssim \lambda_1 \sqrt{\frac{s\log(p)}{n}}$$

by Lemma 4, we see that the resulting bound for the residual satisfies

$$\begin{aligned}\|R\|_{2\to\infty} &\lesssim \delta_1(\frac{\lambda_{k+1}}{\lambda_k})^2 + \delta_2\frac{\lambda_{k+1}}{\lambda_k} + \lambda_1^2\sqrt{\frac{s\log(p)}{n}}(1-\varepsilon)^{\log(p)} \\ &\lesssim (\frac{\lambda_{k+1}}{\lambda_k})^2\kappa\left(\lambda_1\sqrt{\frac{k\log(p))}{n}}\right) + \frac{\lambda_{k+1}}{\lambda_k}\kappa\lambda_k^{-1}\|\Delta\|^2 + \lambda_1^2\sqrt{\frac{s\log(p)}{n}}(1-\varepsilon)^{\log(p)} \\ &\lesssim (\frac{\lambda_{k+1}}{\lambda_k})^2\kappa\left(\lambda_1\sqrt{\frac{k\log(p))}{n}}\right) + \frac{\lambda_{k+1}}{\lambda_k}\kappa\lambda_k^{-1}\|\Delta\|^2 \\ &\lesssim (\frac{\lambda_{k+1}}{\lambda_k})^2\kappa\left(\lambda_1\sqrt{\frac{k\log(p))}{n}}\right) + \frac{\lambda_{k+1}}{\lambda_k}\kappa\lambda_k^{-1}\lambda_1^2\frac{s\log(p)}{n} \\ &\lesssim \kappa^2\lambda_{k+1}\sqrt{\frac{k\log(p)}{n}} + \lambda_{k+1}\kappa^3\frac{s\log(p)}{n}\end{aligned}$$

with probability at least $1 - O(p^{-3})$ by the assumption $\varepsilon > \frac{1}{64}$. Combining with our initial bound in (C.20), we see that

$$\|J_2\|_{2\to\infty} \lesssim \kappa^2\lambda_{k+1}\sqrt{\frac{k\log(p)}{n}} + \lambda_{k+1}\kappa^3\frac{s\log(p)}{n}$$

with probability at least $1 - O(p^{-3})$ as desired. □

C.2.5 Proof of Lemmas 28 and 29

Recall the statement of Lemma 28.

Lemma 28 (The matrix K_1)**.** *The matrix K_1 satisfies*

$$\begin{aligned}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\widetilde{U}\|_{2\to\infty} &\lesssim \frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} + \lambda_1\sqrt{\frac{k\log(p)}{n}} \\ &\lesssim \mathcal{E}_5\lambda_k\end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Recall K_1 is given by

$$K_1 := \frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\widetilde{U}.$$

Proof of Lemma 28. Note that since $U_J U_J^\top + U_\perp U_\perp^\top = I$, we have that

$$\begin{aligned}
\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\widetilde{U}_J\|_{2\to\infty} &\leq \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_J U_J^\top \widetilde{U}_J\|_{2\to\infty} \\
&\quad + \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_\perp U_\perp^\top \widetilde{U}_J\|_{2\to\infty} \\
&\leq \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_J\|_{2\to\infty}\|U_J^\top \widetilde{U}_J\| \\
&\quad + \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_\perp\|_{2\to\infty}\|U_\perp^\top \widetilde{U}_J\| \\
&\leq \sqrt{k}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_J\|_{\max} \\
&\quad + \sqrt{s}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}U_\perp\|_{\max}\|U_\perp^\top \widetilde{U}_J\|,
\end{aligned} \tag{C.22}$$

We bound each term inside the max norm, using a strategy similar to the beginning of the proof of Lemma 27. For the first term, note that we can write the absolute value of its i, j entry via

$$\begin{aligned}
&\left|\frac{1}{n}\sum_q \sum_{k,l}\left((\Sigma^{1/2})_{JJ}\right)_{ik}(Y_{qk}Y_{ql} - \mathbb{E}Y_{qk}Y_{ql}))\left((\Sigma^{1/2})_{JJ}U_J\right)_{kj}\right| \\
&\qquad \leq \max_q \left|\sum_{k,l}\left((\Sigma^{1/2})_{JJ}\right)_{ik}(Y_{qk}Y_{ql} - \mathbb{E}Y_{qk}Y_{ql}))\left((\Sigma^{1/2})_{JJ}U_J\right)_{lj}\right|.
\end{aligned}$$

We focus on bounding for fixed q. This is a quadratic form in the random variable $\{Y_{qk}\}_{k=1}^s$. Define the matrix M via

$$M_{kl} := \left((\Sigma^{1/2})_{JJ}\right)_{ik}\left((\Sigma^{1/2})_{JJ}U_J\right)_{lj}.$$

Note that

$$\begin{aligned}
\|M\|_F^2 &= \sum_{k,l}\left((\Sigma^{1/2})_{JJ}\right)_{ik}^2\left((\Sigma^{1/2})_{JJ}U_J\right)_{lj}^2 \\
&\leq \lambda_1\|(\Sigma^{1/2})_{JJ}\|_{2\to\infty}^2 \\
&\leq \lambda_1^2.
\end{aligned}$$

Therefore, for any fixed q, i, and j, applying the Hanson-Wright inequality (Theorem 24 in

Appendix C.3),

$$\mathbb{P}\Big(\Big|\sum_{k,l}\Big((\Sigma^{1/2})_{JJ}\Big)_{ik}(Y_{qk}Y_{ql}-\mathbb{E}Y_{qk}Y_{ql}))\Big((\Sigma^{1/2})_{JJ}U_J\Big)_{lj}\Big|>t\Big)\leq 2\exp\Big(-c\min\Big\{\frac{t^2}{\lambda_1^2},\frac{t}{\|M\|}\Big\}\Big).$$

Setting $t=C\lambda_1\sqrt{\frac{\log(s)+\log(k)+5\log(p)}{n}}$ and taking a union bound for all n random variables shows that with probability at least $1-O(s^{-1}k^{-1}p^{-4})$ that

$$\max_q\Big|\sum_{k,l}\Big((\Sigma^{1/2})_{JJ}\Big)_{ik}(Y_{qk}Y_{ql}-\mathbb{E}Y_{qk}Y_{ql}))\Big((\Sigma^{1/2})_{JJ}U_J\Big)_{lj}\Big|\lesssim\lambda_1\sqrt{\frac{\log(p)}{n}}.$$

Therefore, taking a union bound over all s rows and k columns shows that with probability at least $1-O(p^{-4})$ that

$$\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}U_J\|_{\max}\lesssim\lambda_1\sqrt{\frac{\log(p)}{n}}. \tag{C.23}$$

The exact same argument yields with the same probability that

$$\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}U_\perp\|_{\max}\lesssim\lambda_1\sqrt{\frac{\log(p)}{n}}. \tag{C.24}$$

Combining (C.22) with (C.23) and (C.24) yields

$$\begin{aligned}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}\widetilde{U}_J\|_{2\to\infty}&\leq\sqrt{k}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}U_J\|_{\max}\\&\quad+\sqrt{s}\|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J-nI)(\Sigma^{1/2})_{JJ}U_\perp\|_{\max}\|U_\perp^\top\widetilde{U}_J\|\\&\lesssim\lambda_1\sqrt{\frac{k\log(p)}{n}}+\lambda_1\sqrt{\frac{s\log(p)}{n}}\|U_\perp^\top\widetilde{U}_J\|.\end{aligned}$$

So what remains is to bound the term $\|U_\perp^\top\widetilde{U}_J\|$, However, we note that this is simply (by a factor of $\sqrt{2}$) the $\sin\Theta$ distance between the subspace $U_JU_J^\top$ and $\widetilde{U}_J\widetilde{U}_J^\top$ (see Lemma 31 in Appendix C.3). Therefore, by Proposition 1, we have that this can be bounded by

$$\|U_\perp^\top\widetilde{U}_J\|\lesssim\frac{\lambda_1}{\lambda_k-\lambda_{k+1}}\sqrt{\frac{s\log(p)}{n}}.$$

Putting it all together, this yields that with probability at least $1 - O(p^{-4})$ that

$$\begin{aligned}\|K_1\|_{2\to\infty} &= \|\frac{1}{n}(\Sigma^{1/2})_{JJ}(Y_J^\top Y_J - nI)(\Sigma^{1/2})_{JJ}\widetilde{U}_J\|_{2\to\infty} \\ &\lesssim \frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} + \lambda_1\sqrt{\frac{k\log(p)}{n}},\end{aligned}$$

which is the desired bound. □

Again, we repeat the statement of Lemma 29.

Lemma 29 (The matrix K_2)**.** *The matrix K_2 satisfies*

$$\begin{aligned}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}\widetilde{U}\|_{2\to\infty} &\lesssim \lambda_1\sqrt{\frac{k\log(p)}{n}} + \frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n} \\ &\lesssim \mathcal{E}_5\lambda_k\end{aligned}$$

with probability at least $1 - O(p^{-4})$.

Recall that

$$K_2 := \Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}\widetilde{U}_J$$

Proof of Lemma 29. We have that

$$\begin{aligned}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}\widetilde{U}_J\|_{2\to\infty} \le \|&\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}U_J\|_{2\to\infty} \\ &+ \|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}U_\perp\|_{2\to\infty}\|U_\perp^\top\widetilde{U}_J\| \\ \le \sqrt{k}\|&\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}U_J\|_{\max} \\ &+ \sqrt{s}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}U_\perp\|_{\max}\|U_\perp^\top\widetilde{U}_J\|. \quad \text{(C.25)}\end{aligned}$$

We bound each norm inside the max separately. Define the random variable η_{ij} as the i, j entry of the matrix $\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}U_J$. Then

$$\eta_{ij} = \frac{1}{n}\sum_{q=1}^{n}\sum_{k=1}^{s}\sum_{l=1}^{p-s}[\Sigma^{1/2}_{JJ^c}]_{il}\xi^{(q)}_{s+l,k}\Big((\Sigma^{1/2})_{JJ}U_J\Big)_{kj},$$

where $\xi^{(q)}_{s+l,k} := Y_{q,s+l}Y_{qk}$. Following a strategy similar to the proof of Lemma 27, we have to bound both the maximum and sum of squared ψ_1 norms of the random variable

$$\alpha_{qlj} := \frac{1}{n}[\Sigma^{1/2}_{JJ^c}]_{il}\xi^{(q)}_{s+l,k}\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)_{kj}.$$

The squared entries satisfy

$$\|\frac{1}{n}[\Sigma^{1/2}_{JJ^c}]_{il}\xi^{(q)}_{s+l,k}\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)_{kj}\|^2_{\psi_1} \leq \frac{1}{n^2}([\Sigma^{1/2}_{JJ^c}]_{il})^2\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)^2_{kj}.$$

Summing up over q, l, j,

$$\begin{aligned}\sum_{q=1}^n\sum_{k=1}^s\sum_{l=1}^{p-s}\|\alpha_{qlj}\|^2_{\psi_1} &\leq \frac{1}{n}\sum_{k=1}^s\sum_{l=1}^{p-s}([\Sigma^{1/2}_{JJ^c}]_{il})^2\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)^2_{kj}\\ &\leq \frac{1}{n}\sum_{k=1}^s\bigg((\Sigma^{1/2})_{JJ}U_J\bigg)^2_{kj}\|\Sigma^{1/2}_{JJ^c}\|^2_{2\to\infty}\\ &\leq \frac{\lambda_1\|\Sigma^{1/2}_{JJ^c}\|^2_{2\to\infty}}{n}.\end{aligned}$$

Also,

$$\max_{q,l,j}\|\alpha_{qlj}\|_{\psi_1} \leq \frac{1}{n}\sqrt{\lambda_1}\|\Sigma^{1/2}_{JJ^c}\|_{2\to\infty}.$$

By the the Generalized Bernstein inequality (Theorem 23 in Appendix C.3),

$$\mathbb{P}\bigg(|\eta_{ij}| > t\bigg) \leq 2\exp\bigg(-cn\min\bigg[\frac{t^2}{\lambda_1\|\Sigma^{1/2}_{JJ^c}\|^2_{2\to\infty}}, \frac{t}{\sqrt{\lambda_1}\|\Sigma^{1/2}_{JJ^c}\|_{2\to\infty}}\bigg]\bigg).$$

Again taking $t = C\|\Sigma^{1/2}_{JJ^c}\|_{2\to\infty}\sqrt{\lambda_1}\sqrt{\frac{\log(s)+\log(k)+4\log(p)}{n}}$ shows that this holds with probability $1 - O(s^{-1}k^{-1}p^{-4})$. Taking a union over all s rows and k columns of the matrix yields that

$$\begin{aligned}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y^\top_{J^c}Y_J(\Sigma^{1/2})_{JJ}U_J\|_{\max} &\lesssim \|\Sigma^{1/2}_{JJ^c}\|_{2\to\infty}\sqrt{\lambda_1}\sqrt{\frac{\log(p)}{n}}\\ &\lesssim \lambda_1\sqrt{\frac{\log(p)}{n}}.\end{aligned}$$

Applying precisely the same argument to the other term yields with probability $1 - O(p^{-4})$ that

$$\|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}U_\perp\|_{\max} \lesssim \lambda_1\sqrt{\frac{\log(p)}{n}}.$$

Therefore, combining these bounds with the initial bound in (C.25) and Proposition 1 and the equivalent expressions for the $\sin\Theta$ distances (Lemma 31 in Appendix C.3), we have that with probability at least $1 - O(p^{-4})$,

$$\begin{aligned}\|K_2\|_{2\to\infty} = \|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}\widetilde{U}_J\|_{2\to\infty} &\le \sqrt{k}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}U_J\|_{\max} \\ &\quad + \sqrt{s}\|\frac{1}{n}\Sigma^{1/2}_{JJ^c}Y_{J^c}^\top Y_J(\Sigma^{1/2})_{JJ}U_\perp\|_{\max}\|U_\perp^\top\widetilde{U}_J\| \\ &\lesssim \lambda_1\sqrt{\frac{k\log(p)}{n}} + \frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}}\frac{s\log(p)}{n}\end{aligned}$$

as desired. □

C.2.6 Proof of Lemma 30

It will be useful to collect some properties of the matrix $\Sigma^{1/2}_{JJ^c}$, which we state as a proposition.

Proposition 10 (Properties of the Matrix $\Sigma^{1/2}_{JJ^c}$)**.** *The matrix $\Sigma^{1/2}_{JJ^c}$ satisfies*

$$\|\Sigma^{1/2}_{JJ^c}\| \le 2\sqrt{\lambda_1}$$

Furthermore, the left singular subspace of $\Sigma^{1/2}_{JJ^c}$ must contain columns of $U_\perp$.

Proof of Proposition 10. First, we note that

$$\begin{aligned}\|\Sigma^{1/2}_{JJ^c}\| &= \left\|\begin{pmatrix} 0 & (\Sigma^{1/2})_{JJ^c} \\ ((\Sigma^{1/2})_{JJ^c})^\top & 0\end{pmatrix}\right\| \\ &\le \|\Sigma\|^{1/2} + \left\|\begin{pmatrix} (\Sigma^{1/2})_{JJ} & 0 \\ 0 & \Sigma^{1/2}_{J^cJ^c}\end{pmatrix}\right\| \\ &\le 2\sqrt{\lambda_1},\end{aligned}$$

since eigenvalues bound eigenvalues of any principal submatrix. For the second claim, note that

$$\Sigma^{1/2}\begin{pmatrix} U_J \\ 0 \end{pmatrix} = \begin{pmatrix} (\Sigma^{1/2})_{JJ} & (\Sigma^{1/2})_{JJ^c} \\ ((\Sigma^{1/2})_{JJ^c})^\top & \Sigma^{1/2}_{J^cJ^c} \end{pmatrix}\begin{pmatrix} U_J \\ 0 \end{pmatrix} = \begin{pmatrix} U_J \\ 0 \end{pmatrix}\Lambda^{1/2}.$$

This shows that the matrix $(\Sigma^{1/2}_{JJ^c})^\top$ satisfies $(\Sigma^{1/2}_{JJ^c})^\top U_J = 0$, so that its null space must contain the space spanned by U_J. However, this also shows that since $(\Sigma^{1/2}_{JJ^c})^\top \in \mathbb{R}^{(p-s)\times s}$, then its rank is at most $s-k$. Hence, define $(\Sigma^{1/2}_{JJ^c})^\top = V_1 D V_2^\top$ as the reduced singular value decomposition of $(\Sigma^{1/2}_{JJ^c})^\top$. Since its rank is at most $s-k$, we have that $V_1 \in \mathbb{O}(p-s, s-k)$, $V_2 \in \mathbb{O}(s, s-k)$, and D is an $s-k \times s-k$ diagonal matrix of singular values.

Since $(\Sigma^{1/2}_{JJ^c})^\top U_J = V_1 D V_2^\top U_J = 0$, the term $V_2 \in \mathbb{O}(s, s-k)$ must span a space perpendicular to U_J. The only matrix up to choice of basis in $\mathbb{O}(s, s-k)$ satisfying $V_2^\top U_J = 0$ is the matrix $U_\perp$, which establishes the second claim. □

Therefore, all this shows that

- The left singular subspace of $\Sigma^{1/2}_{JJ^c}$ contains columns of $U_\perp$;
- Its singular values are all uniformly bounded by $2\sqrt{\lambda_1}$.

We are now prepared to prove Lemma 30.

Lemma 30 (The matrices K_3 and K_4)**.** *The matrices K_3 and K_4 satisfy*

$$\begin{aligned} \|\frac{1}{n}(\Sigma^{1/2})_{JJ} Y_J^\top Y_{J^c}(\Sigma^{1/2}_{JJ^c})^\top \widetilde{U}_J\|_{2\to\infty} &\lesssim \frac{s\log(p)}{n}\frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}} \\ &\lesssim \mathcal{E}_5\lambda_k; \\ \|\frac{1}{n}\Sigma^{1/2}_{JJ^c}(Y_{J^c}^\top Y_{J^c} - nI)(\Sigma^{1/2}_{JJ^c})^\top \widetilde{U}\|_{2\to\infty} &\lesssim \frac{s\log(p)}{n}\frac{\lambda_1^2}{\lambda_k - \lambda_{k+1}} \\ &\lesssim \mathcal{E}_5\lambda_k \end{aligned}$$

with probability at least $1 - O(p^{-3})$.

Proof of Lemma 30. Let $\Sigma^{1/2}_{JJ^c}$ have singular value decomposition $U_\perp D V^\top$, where $U_\perp \in \mathbb{O}(s, s-k)$, $D_{ii} \geq 0$, $1 \leq i \leq s-k$, $V \in \mathbb{O}(p-s, s-k)$. We will show the result for $D_{ii} > 0$, though the same proof goes through if $D_{ii} = 0$ for some i.

Then the term K_3 satisfies

$$\begin{aligned}
\|K_3\|_{2\to\infty} &= \|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n}(\Sigma^{1/2}_{JJ^c})^\top \widetilde{U}_J\|_{2\to\infty} \\
&\leq \|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n} V D U_\perp^\top \widetilde{U}_J\|_{2\to\infty} \\
&\leq \|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n} V D U_\perp^\top U_\perp\|_{2\to\infty}\|U_\perp^\top \widetilde{U}_J\| \\
&\leq \|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n} V\|_{2\to\infty}\sqrt{\lambda_1}\|U_\perp^\top \widetilde{U}_J\|. \qquad \text{(C.26)}
\end{aligned}$$

The term $\|U_\perp^\top \widetilde{U}_J\|$ can be bounded via Proposition 1 and Lemma 31 in Appendix C.3. So what remains is to bound the $2 \to \infty$ norm in (C.26). Note that the matrix V is of column dimension at most $(s-k)$. Hence, each of the s rows of the matrix $(\Sigma^{1/2})_{JJ} Y_J Y_{J^c} V$ is of dimension at most $s-k$.

Following a strategy similar to that in Lemmas 28 and 29, we have that

$$\begin{aligned}
\|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n} V\|_{2\to\infty} &\leq \sqrt{s-k}\max_{i,j}\left|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n} V\right|_{i,j} \\
&\leq \sqrt{s}\max_{i,j}\left|(\Sigma^{1/2})_{JJ}\frac{Y_J^\top Y_{J^c}}{n} V\right|_{i,j}.
\end{aligned}$$

By analogous arguments as in Lemma 29, the i,j entry is a sum of independent mean-zero subexponential random variables, each with ψ_1 norm bounded $\frac{1}{n}\sqrt{\lambda_1}$. Therefore, by Bernstein's inequality, any i,j entry is bounded by

$$C\sqrt{\lambda_1}\sqrt{\frac{\log(p)}{n}}$$

with probability at most $1 - O(p^{-3})$. Combining with Proposition 1, we have the bound

$$\begin{aligned}\|K_3\|_{2\to\infty} &\lesssim \lambda_1\sqrt{\frac{s\log(p)}{n}}\|U_\perp^\top \widetilde{U}_J\| \\ &\lesssim \lambda_1\sqrt{\frac{s\log(p)}{n}}\left(\frac{\lambda_1}{\lambda_k-\lambda_{k+1}}\sqrt{\frac{s\log(p)}{n}}\right) \\ &\lesssim \frac{\lambda_1^2}{\lambda_k-\lambda_{k+1}}\frac{s\log(p)}{n}\end{aligned}$$

as desired.

For the term K_4, we see that

$$\begin{aligned}\|K_4\|_{2\to\infty} &= \|\Sigma_{JJ^c}^{1/2}\left(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\right)VDU_\perp^\top \widetilde{U}_J\|_{2\to\infty} \\ &\leq \|\Sigma_{JJ^c}^{1/2}\left(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\right)V\|_{2\to\infty}\sqrt{\lambda_1}\|U_\perp^\top \widetilde{U}_J\| \\ &\leq \sqrt{s\lambda_1}\|\Sigma_{JJ^c}^{1/2}\left(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\right)V\|_{\max}\|U_\perp^\top \widetilde{U}_J\|. \qquad \text{(C.27)}\end{aligned}$$

We will bound the term inside the max norm for fixed i and j. Observe that

$$\begin{aligned}\left|\left(\Sigma_{JJ^c}^{1/2}\left(\frac{Y_{J^c}^\top Y_{J^c}}{n} - I\right)V\right)_{ij}\right| &= \max_{i,j}\left|\frac{1}{n}\sum_q\sum_{k,l}\left(\Sigma_{JJ^c}^{1/2}\right)_{ik}(Y_{qk}Y_{ql} - \mathbb{E}Y_{qk}Y_{ql})V_{lj}\right| \\ &\leq \max_q\left|\sum_{k,l}\left(\Sigma_{JJ^c}^{1/2}\right)_{ik}(Y_{qk}Y_{ql} - \mathbb{E}Y_{qk}Y_{ql})V_{lj}\right|.\end{aligned}$$

We will first bound the term inside the absolute value for fixed q by Hanson-Wright (Theorem 24 in Appendix C.3). Let M be the matrix defined via

$$M_{kl} := \left(\Sigma_{JJ^c}^{1/2}\right)_{ik}V_{lj}.$$

Then

$$\|M\|_F^2 = \sum_{k,l}\left(\Sigma_{JJ^c}^{1/2}\right)_{ik}^2 V_{lj}^2 = \sum_k\left(\Sigma_{JJ^c}^{1/2}\right)_{ik}^2 \leq \|\Sigma_{JJ^c}^{1/2}\|_{2\to\infty}^2 \leq 4\lambda_1.$$

Therefore, by applying the Hanson-Wright inequality, for any fixed q it holds that

$$\mathbb{P}\bigg(\bigg|\sum_{k,l}\Big(\Sigma^{1/2}_{JJ^c}\Big)_{ik}(Y_{qk}Y_{ql}-\mathbb{E}Y_{qk}Y_{ql})V_{lj}\bigg|\geq t\bigg)\leq 2\exp\bigg(-c\min\bigg\{\frac{t^2}{4\lambda_1},\frac{t}{\|M\|}\bigg\}\bigg).$$

Setting $t=C\sqrt{\lambda_1}\sqrt{\frac{\log(s)+\log(k)+5\log(p)}{n}}$ and taking a union bound over all q random variables shows that for fixed i and j, with probability at least $1-O(s^{-1}k^{-1}p^{-4})$,

$$\bigg|\bigg(\Sigma^{1/2}_{JJ^c}\bigg(\frac{Y^\top_{J^c}Y_{J^c}}{n}-I\bigg)V\bigg)_{ij}\bigg|\lesssim\sqrt{\lambda_1}\sqrt{\frac{\log(p)}{n}}.$$

Taking a union bound over s rows and k columns shows that with probability at least $1-O(p^{-4})$,

$$\|\Sigma^{1/2}_{JJ^c}\bigg(\frac{Y^\top_{J^c}Y_{J^c}}{n}-I\bigg)V\|_{\max}\lesssim\sqrt{\lambda_1}\sqrt{\frac{\log(p)}{n}}.$$

Therefore, from the initial bound in (C.27) and Proposition 1,

$$\begin{aligned}\|K_4\|_{2\to\infty}&=\|\Sigma^{1/2}_{JJ^c}\bigg(\frac{Y^\top_{J^c}Y_{J^c}}{n}-I\bigg)VDU_\perp^\top\widetilde{U}_J\|_{2\to\infty}\\&\leq\sqrt{s\lambda_1}\|\Sigma^{1/2}_{JJ^c}\bigg(\frac{Y^\top_{J^c}Y_{J^c}}{n}-I\bigg)V\|_{\max}\|U_\perp^\top\widetilde{U}_J\|\\&\lesssim\lambda_1\sqrt{\frac{s\log(p)}{n}}\|U_\perp^\top\widetilde{U}_J\|\\&\lesssim\frac{s\log(p)}{n}\frac{\lambda_1^2}{\lambda_k-\lambda_{k+1}}\end{aligned}$$

as desired. □

C.3 Background Material on Orlicz Norms, Concentration, and Subspace Perturbation

Here we briefly discuss Orlicz ψ_α Norms and Bernstein's inequality for subexponential random variables.

The *Orlicz Norm* of order α for a real-valued random variable X is defined via

$$\|X\|_{\psi_\alpha} := \inf\{t > 0 : \mathbb{E}\exp(|X|^\alpha/t) \leq 1\}.$$

Random variables with finite ψ_2 norm are called *subgaussian* and those with a finite ψ_1 norm are called *subexponential*. Generally speaking, if X is subgaussian, then X^2 is subexponential and $\|X^2\|_{\psi_1} \lesssim \|X\|_{\psi_2}^2$. One also has the "Cauchy-Schwarz" bound $\|XY\|_{\psi_1} \lesssim \|X\|_{\psi_2}\|Y\|_{\psi_2}$ (Vershynin, 2018).

For subexponential random variables, one has the following generalized Bernstein's inequality. See Theorem 2.8.2 in Vershynin (2018) for the proof.

Theorem 23 (Theorem 2.8.2 in Vershynin (2018)). *Let $X_1, ..., X_N$ be independent, mean zero subexponential random variables and let $a = (a_i)_{i=1}^N$. Then there exists a universal constant $c > 0$ such that for all $t \geq 0$, we have that*

$$\mathbb{P}\left\{\left|\sum_{i=1}^N a_i X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right]$$

where $K = \max_i \|X_i\|_{\psi_1}$.

We also make use of the Hanson-Wright Inequality. See Theorem 6.2.1 in Vershynin (2018) for the proof.

Theorem 24 (Hanson-Wright Inequality –Theorem 6.2.1 in Vershynin (2018)). *Let $X_1, \ldots, X_N$ be independent, mean-zero subgaussian random variables. Let M be some fixed $N \times N$ matrix. Then there exists a universal constant $c > 0$ such that for all $t \geq 0$, we have that*

$$\mathbb{P}\left\{\left|\sum_{k,l} M_{kl}X_kX_l - \mathbb{E}M_{kl}X_kX_l\right| \geq t\right\} \leq 2\exp\left(-c\min\left\{\frac{t^2}{K^4\|M\|_F^2}, \frac{t}{K^2\|M\|}\right\}\right),$$

where $K = \max_i \|X_i\|_{\psi_2}$.

We also use several notions from subspace perturbation theory. Suppose U and $\widehat{U}$ are two $d_1 \times d_2$ matrices with orthonormal columns with $d_2 \leq d_1$. The $\sin\Theta$ distance between the subspaces spanned by U and $\widehat{U}$ is defined as follows. Let $I - UU^\top = U_\perp U_\perp^\top$. Then the

(spectral) $\sin\Theta$ distance is defined as

$$\|\sin\Theta(U_1, U_2)\| := \|\widehat{U}^\top U_\perp\|.$$

Throughout the supplementary material, we use several equivalent terms for the $\sin\Theta$ distance. We present this here as a lemma, the statement of which is slightly modified from Lemma 1 of Cai and Zhang (2018).

Lemma 31 (Modified from Lemma 1 of Cai and Zhang (2018))**.** *The* $\sin\Theta$ *distance between two matrices satisfies*

$$\|\sin\Theta(\widehat{U}, U)\| \leq \inf_{W: WW^\top = I_{d_2}} \|\widehat{U} - UW\| \leq \sqrt{2}\|\sin\Theta(\widehat{U}, U)\|;$$
$$\|\sin\Theta(\widehat{U}, U)\| \leq \|\widehat{U}\widehat{U}^\top - UU^\top\| \leq 2\|\sin\Theta(\widehat{U}, U)\|.$$

Appendix D

Proofs from Chapter 4

D.1 Proof of Theorem 11

This section contains the full proof of Theorem 11. Without loss of generality, throughout this section we assume that $\sigma = 1$. Throughout we denote $\mathbf{T}_k = \mathcal{M}_k(\mathcal{T})$ and $\mathbf{Z}_k$ similarly. We also let $p = p_{\max}$ for convenience throughout the proofs.

Before proving our main results, we state the following results for the initialization. The proof is contained in Appendix D.1.5. It is worth noting that our $\ell_{2,\infty}$ slightly sharpens the results of Cai et al. (2021a) by a factor of κ^2 for the diagonal-deleted estimator; however, we do not consider missingness as they do. In what follows, we define the leave-one-out initialization $\widetilde{\mathbf{U}}_k^{(S,k-m)}$ as the eigenvectors of the matrix

$$\Gamma\Big(\mathbf{T}_k\mathbf{T}_k^\top + \mathbf{Z}_k^{k-m}\mathbf{T}_k^\top + \mathbf{T}_k^\top\mathbf{Z}_k^{k-m} + \mathbf{Z}_k^{k-m}(\mathbf{Z}_k^{k-m})^\top\Big),$$

where $\mathbf{Z}_k^{k-m}$ denotes the matrix $\mathbf{Z}_k$ with its m'th row set to zero (the double appearance of the index k will be useful for defining the other two leave-one-out sequences in the following subsection).

Theorem 25 (Initialization $\ell_{2,\infty}$ error)**.** *Instate the conditions of Theorem 11. Then with*

probability at least $1 - O(p^{-20})$*, it holds for each* k *that*

$$\|\widehat{\mathbf{U}}_k^S - \mathbf{U}_k \mathbf{W}_k^S\|_{2,\infty} \lesssim \frac{\kappa\mu_0\sqrt{r_1 \log(p)}}{\lambda} + \frac{\mu_0\sqrt{r_k p_{-k}}\log(p)}{\lambda^2} + \kappa^2\mu_0^2\frac{r_k}{p_k};$$
$$\max_m \|\widehat{\mathbf{U}}_k^S(\widehat{\mathbf{U}}_k^S)^\top - \widetilde{\mathbf{U}}_k^{(S,k-m)}(\widetilde{\mathbf{U}}_k^{(S,k-m)})^\top\| \lesssim \frac{\kappa\mu_0\sqrt{r_k \log(p)}}{\lambda} + \frac{\mu_0\sqrt{r_k p_{-k}}\log(p)}{\lambda^2}.$$

In Appendix D.1.1 we describe in detail the leave-one-out sequences for the iterates of tensor SVD. In Appendix D.1.2 we obtain the deterministic bounds needed en route to Theorem 11, and in Appendix D.1.3 we use these bounds to obtain high-probability guarantees on good events. Appendix D.1.4 contains the final proof of Theorem 25. Throughout we rely on several self-contained probabilistic lemmas, whose statements and proofs can be found in Appendix D.3.

D.1.1 The Leave-One-Out Sequence

In this section we formally define the leave-one-out sequence. First, we already have defined $\widehat{\mathbf{U}}_k^S$ and $\widetilde{\mathbf{U}}_k^{(S,k-m)}$ in the previous section, but we will need a few additional pieces of notation. We define $\widehat{\mathbf{U}}_k^{(t)}$ as the output of tensor power iteration after t iterations, with $\widehat{\mathbf{U}}_k^{(0)} = \widehat{\mathbf{U}}_k^S$. It will also be useful to define

$$\widehat{\mathcal{P}}_k^{(t)} := \begin{cases} \mathcal{P}_{\widehat{\mathbf{U}}_{k+1}^{(t-1)} \otimes \widehat{\mathbf{U}}_{k+2}^{(t-1)}} & k = 1; \\ \mathcal{P}_{\widehat{\mathbf{U}}_{k+1}^{(t-1)} \otimes \widehat{\mathbf{U}}_{k+2}^{(t)}} & k = 2; \\ \mathcal{P}_{\widehat{\mathbf{U}}_{k+1}^{(t)} \otimes \widehat{\mathbf{U}}_{k+2}^{(t)}} & k = 3. \end{cases}$$

The matrix $\widehat{\mathcal{P}}_k^{(t)}$ is simply the projection matrix corresponding to the previous two iterates.

We have already defined the matrix $\mathbf{Z}_j^{j-m}$ as the j'th matricization of $\mathcal{Z}$ with its m'th row set to zero. We now define $\mathcal{Z}^{j-m}$ as the corresponding tensor $\mathcal{Z}$, where the entries corresponding to the m'th row of $\mathbf{Z}_j$ are set to zero. Finally, define $\mathbf{Z}_k^{j-m} := \mathcal{M}_k(\mathcal{Z}^{j-m})$. In other words $\mathbf{Z}_k^{j-m}$ is the k'th matricization of the tensor $\mathcal{Z}$ with the entries corresponding to the m'th row of $\mathbf{Z}_j$ set to zero.

We now define $\widetilde{\mathbf{U}}_k^{(S,j-m)}$ as the leading r_k eigenvectors of the matrix

$$\Gamma\big(\mathbf{T}_k\mathbf{T}_k^\top + \mathbf{Z}_k^{j-m}\mathbf{T}_k^\top + \mathbf{T}_k(\mathbf{Z}_k^{j-m})^\top + \mathbf{Z}_k^{j-m}(\mathbf{Z}_k^{j-m})^\top\big).$$

We now show that the other leave-one-out sequence initializations are sufficiently close to the true initialization.

Lemma 32 (Proximity of the initialization leave-one-out sequences)**.** *Instate the conditions of Theorem 11. Then the initializations of the leave-one-out sequences satisfy for each k the bound*

$$\max_{1\le j\le 3}\max_{1\le m\le p_j}\|\widetilde{\mathbf{U}}_k^{(S,j-m)}(\widetilde{\mathbf{U}}_k^{(S,j-m)})^\top - \widehat{\mathbf{U}}_k^S(\widehat{\mathbf{U}}_k^S)^\top\| \lesssim \frac{\kappa\sqrt{p_k\log(p)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_j}} + \frac{(p_1p_2p_3)^{1/2}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r_1}{p_j}}$$

with probability at least $1 - O(p^{-19})$.

Lemma 32 is proven in Appendix D.1.5 after the proof of Theorem 25. To define subsequent iterates, we set $\widetilde{\mathbf{U}}_k^{(t,j-m)}$ as the outputs of tensor power iteration using these initializations, though with one modification. We now define $\widetilde{\mathbf{U}}_k^{(t,j-m)}$ as the left singular vectors of the matrix

$$\mathbf{T}_k + \mathbf{Z}_k^{j-m}\widetilde{\mathcal{P}}_k^{t,j-m},$$

which is still independent from $e_m^\top\mathbf{Z}_j$. Here, we set $\widetilde{\mathcal{P}}_k^{t,j-m}$ inductively as the projection matrix

$$\widetilde{\mathcal{P}}_k^{t,j-m} := \begin{cases} \mathcal{P}_{\widetilde{\mathbf{U}}_{k+1}^{(t-1,j-m)}\otimes\widetilde{\mathbf{U}}_{k+2}^{(t-1,j-m)}} & k=1; \\ \mathcal{P}_{\widetilde{\mathbf{U}}_{k+1}^{(t-1,j-m)}\otimes\widetilde{\mathbf{U}}_{k+2}^{(t,j-m)}} & k=2; \\ \mathcal{P}_{\widetilde{\mathbf{U}}_{k+1}^{(t,j-m)}\otimes\widetilde{\mathbf{U}}_{k+2}^{(t,j-m)}} & k=3. \end{cases}$$

Note that for each k there are 3 different leave-one-out sequences, one corresponding to each mode, by leaving out the m'th row of that mode (note that for convenience we use the index m for each leave-one-out sequence, but we slightly abuse notation as m as defined above must satisfy $1 \le m \le p_j$).

We now introduce some notation used for the remainder of our proofs. Define

$$\begin{aligned}
\mathbf{L}_k^{(t)} &\coloneqq \mathbf{U}_{k\perp}\mathbf{U}_{k\perp}^\top \mathbf{Z}_k \widehat{\mathcal{P}}_k^{(t)} \mathbf{T}_k^\top \widehat{\mathbf{U}}_k^{(t)} (\widehat{\mathbf{\Lambda}}_k^{(t)})^{-2};\\
\mathbf{Q}_k^{(t)} &\coloneqq \mathbf{U}_{k\perp}\mathbf{U}_{k\perp}^\top \mathbf{Z}_k \widehat{\mathcal{P}}_k^{(t)} \mathbf{Z}_k^\top \widehat{\mathbf{U}}_k^{(t)} (\widehat{\mathbf{\Lambda}}_k^{(t)})^{-2};\\
\tau_k &\coloneqq \sup_{\substack{\|\mathbf{U}_1\|=1, \mathrm{rank}(\mathbf{U}_1)\le 2r_{k+1}\\ \|\mathbf{U}_2\|=1, \mathrm{rank}(\mathbf{U}_2)\le 2r_{k+2}}} \|\mathbf{Z}_k\Big(\mathcal{P}_{\mathbf{U}_1}\otimes \mathcal{P}_{\mathbf{U}_2}\Big)\|;\\
\xi_k^{(t,j-m)} &\coloneqq \left\|\left(\mathbf{Z}_k^{j-m} - \mathbf{Z}_k\right)\widetilde{\mathcal{P}}_k^{t,j-m}\right\|\\
\widetilde{\xi}_k^{(t,j-m)} &\coloneqq \left\|\left(\mathbf{Z}_k^{j-m} - \mathbf{Z}_k\right)\widetilde{\mathcal{P}}_k^{t,j-m}\mathbf{V}_k\right\|\\
\eta_k^{(t,j-m)} &\coloneqq \begin{cases} \|\sin\Theta(\widetilde{\mathbf{U}}_{k+1}^{(t-1,j-m)}, \widehat{\mathbf{U}}_{k+1}^{(t-1)})\| + \|\sin\Theta(\widetilde{\mathbf{U}}_{k+2}^{(t-1,j-m)}, \widehat{\mathbf{U}}_{k+2}^{(t-1)})\| & k=1\\ \|\sin\Theta(\widetilde{\mathbf{U}}_{k+1}^{(t-1,j-m)}, \widehat{\mathbf{U}}_{k+1}^{(t-1)})\| + \|\sin\Theta(\widetilde{\mathbf{U}}_{k+2}^{(t,j-m)}, \widehat{\mathbf{U}}_{k+2}^{(t)})\| & k=2\\ \|\sin\Theta(\widetilde{\mathbf{U}}_{k+1}^{(t,j-m)}, \widehat{\mathbf{U}}_{k+1}^{(t)})\| + \|\sin\Theta(\widetilde{\mathbf{U}}_{k+2}^{(t,j-m)}, \widehat{\mathbf{U}}_{k+2}^{(t)})\| & k=3 \end{cases}\\
\eta_k^{(t)} &\coloneqq \begin{cases} \|\sin\Theta(\mathbf{U}_{k+1}, \widehat{\mathbf{U}}_{k+1}^{(t-1)})\| + \|\sin\Theta(\mathbf{U}_{k+2}, \widehat{\mathbf{U}}_{k+2}^{(t-1)})\| & k=1\\ \|\sin\Theta(\mathbf{U}_{k+1}, \widehat{\mathbf{U}}_{k+1}^{(t-1)})\| + \|\sin\Theta(\mathbf{U}_{k+2}, \widehat{\mathbf{U}}_{k+2}^{(t)})\| & k=2\\ \|\sin\Theta(\mathbf{U}_{k+1}, \widehat{\mathbf{U}}_{k+1}^{(t)})\| + \|\sin\Theta(\mathbf{U}_{k+2}, \widehat{\mathbf{U}}_{k+2}^{(t)})\| & k=3 \end{cases}.
\end{aligned}$$

First we will state results deterministically with dependence on τ_k, $\xi_k^{(t,j-m)}$ and $\eta_k^{(t,j-m)}$. Note that we already have the bound $\widetilde{\xi}_k^{(t,j-m)} \le \xi_k^{(t,j-m)}$ since $\|\mathbf{V}_k\| = 1$, but it will turn out to be slightly more useful to have the dependence on $\mathbf{V}_k$.

D.1.2 Deterministic Bounds

In this section we collect and prove deterministic bounds that we will then combine with probabilistic induction in Appendix D.1.3.

Lemma 33 (Closeness of the orthogonal matrix)**.** *Let* $\mathbf{W}_k^{(t)} = \mathrm{sgn}(\widehat{\mathbf{U}}_k^{(t)}, \mathbf{U}_k)$ *be the matrix sign of* $\widehat{\mathbf{U}}_k^{(t)}$ *and* $\mathbf{U}_k$*. Then*

$$\|\mathbf{U}_k\mathbf{W}_k^{(t)} - \mathbf{U}_k\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)}\|_{2,\infty} \le \mu_0\sqrt{\frac{r_k}{p_k}}\|\sin\Theta(\widehat{\mathbf{U}}_k^{(t)}, \mathbf{U}_k)\|^2.$$

Proof of Lemma 33. Observe that

$$\begin{aligned}\|\mathbf{U}_k\mathbf{W}_k^{(t)} - \mathbf{U}_k\mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)}\|_{2,\infty} &\leq \|\mathbf{U}_k\|_{2,\infty}\|\mathbf{W}_k^{(t)} - \mathbf{U}_k^\top\widehat{\mathbf{U}}_k^{(t)}\| \\ &\leq \mu_0\sqrt{\frac{r_k}{p_k}}\|\sin\Theta(\widehat{\mathbf{U}}_k^{(t)}, \mathbf{U}_k)\|^2.\end{aligned}$$

For details on the final inequality, see Lemma 4.6.3 of Chen et al. (2021c). □

Lemma 34 (Deterministic Bound for the Linear Term)**.** *Suppose* $\mathbf{T}_k = \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k^\top$*, and suppose that* $\lambda/2 \leq \lambda_{r_k}(\widehat{\mathbf{\Lambda}}_k^{(t)})$ *Then the linear term* $\mathbf{L}_k^{(t)}$ *satisfies*

$$\|e_m^\top\mathbf{L}_k^{(t)}\| \leq \frac{8\kappa}{\lambda}\|\mathbf{U}_k\|_{2,\infty}\left(\tau_k\eta_k^{(t)} + \|\mathbf{U}_k^\top\mathbf{Z}_k\mathbf{V}_k\|\right) + \frac{8\kappa}{\lambda}\left(\tau_k\eta_k^{(t,k-m)}\right) + \frac{4\kappa}{\lambda}\widetilde{\xi}_k^{t,k-m},$$

Proof of Lemma 34. Without loss of generality we prove the result for $k = 1$; the cases for $k = 2$ and $k = 3$ are similar by changing the index for t using the definition of $\widehat{\mathcal{P}}_k^{(t)}$.

Recall we let $\mathbf{T}_k = \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{V}_k^\top$. Then the m'th row of the linear term $\mathbf{L}_1^{(t)}$ can be written as

$$\begin{aligned}&e_m^\top\mathbf{U}_{1\perp}\mathbf{U}_{1\perp}^\top\mathbf{Z}_1\left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\mathbf{T}_1^\top\widehat{\mathbf{U}}_1^{(t)}(\widehat{\mathbf{\Lambda}}_1^{(t)})^{-2} \\ &\qquad = e_m^\top\mathbf{U}_{1\perp}\mathbf{U}_{1\perp}^\top\mathbf{Z}_1\left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\mathbf{V}_1\mathbf{\Lambda}_1\mathbf{U}_1^\top\widehat{\mathbf{U}}_1^{(t)}(\widehat{\mathbf{\Lambda}}_1^{(t)})^{-2}.\end{aligned}$$

Taking norms, we see that as long as $2\lambda^{-1} \geq (\widehat{\lambda}_{r_1}^{(t)})^{-1}$ as in the assumptions of this lemma, we have

$$\begin{aligned}&\left\|e_m^\top\mathbf{U}_{1\perp}\mathbf{U}_{1\perp}^\top\mathbf{Z}_1\left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\mathbf{V}_1\mathbf{\Lambda}_1\mathbf{U}_1^\top\widehat{\mathbf{U}}_1^{(t)}(\widehat{\mathbf{\Lambda}}_1^{(t)})^{-2}\right\| \\ &\quad\leq \frac{4\kappa}{\lambda}\left\|e_m^\top\mathbf{U}_{1\perp}\mathbf{U}_{1\perp}^\top\mathbf{Z}_1\left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\mathbf{V}_1\right\| \\ &\quad\leq \frac{4\kappa}{\lambda}\left\|e_m^\top\mathbf{Z}_1\left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\mathbf{V}_1\right\| + \frac{4\kappa}{\lambda}\left\|e_m^\top\mathbf{U}_1\mathbf{U}_1^\top\mathbf{Z}_1\left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\mathbf{V}_1\right\|.\end{aligned} \tag{D.1}$$

Thus, it suffices to analyze the two terms

$$T_1 \coloneqq \left\| e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right] \mathbf{V}_1 \right\|;$$
$$T_2 \coloneqq \left\| e_m^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right] \mathbf{V}_k \right\|;$$

for fixed m. For the term T_1, we introduce the leave-one-out sequence to observe that

$$\begin{aligned}
\left\| e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right] \mathbf{V}_1 \right\| &\leq \left\| e_m^\top \mathbf{Z}_1 \left[\left(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t-1,1-m)}} \right) \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right] \mathbf{V}_1 \right\| \\
&\quad + \left\| e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t-1,1-m)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right] \mathbf{V}_1 \right\| \\
&\leq \left\| e_m^\top \mathbf{Z}_1 \left[\left(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t-1,1-m)}} \right) \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right] \mathbf{V}_1 \right\| \\
&\quad + \left\| e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t-1,1-m)}} \otimes \left(\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t-1,1-m)}} \right) \right] \mathbf{V}_1 \right\| \\
&\quad + \left\| e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t-1,1-m)}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t-1,1-m)}} \right] \mathbf{V}_1 \right\| \\
&\leq 2\tau_1 \| \sin\Theta(\widehat{\mathbf{U}}_2^{(t-1)}, \widetilde{\mathbf{U}}_2^{(t-1,1-m)}) \| + 2\tau_1 \| \sin\Theta(\widehat{\mathbf{U}}_3^{(t-1)}, \widetilde{\mathbf{U}}_3^{(t-1,1-m)}) \| \\
&\quad + \left\| e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t-1,1-m)}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t-1,1-m)}} \right] \mathbf{V}_1 \right\|. \qquad \text{(D.2)}
\end{aligned}$$

As for T_2, we note that

$$\begin{aligned}
&\left\| e_m^\top \mathbf{U}_1 \mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right] \mathbf{V}_1 \right\| \qquad \text{(D.3)} \\
&\leq \| \mathbf{U}_1 \|_{2,\infty} \| \mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right] \mathbf{V}_1 \| \\
&\leq \| \mathbf{U}_1 \|_{2,\infty} \| \mathbf{U}_1^\top \mathbf{Z}_1 \left[\left(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} - \mathcal{P}_{\mathbf{U}_2} \right) \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right] \mathbf{V}_1 \| \\
&\quad + \| \mathbf{U}_1 \|_{2,\infty} \| \mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\mathbf{U}_2} \otimes \left(\mathcal{P}_{\mathbf{U}_3} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right) \right] \mathbf{V}_1 \| \\
&\quad + \| \mathbf{U}_1 \|_{2,\infty} \| \mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\mathbf{U}_2} \otimes \mathcal{P}_{\mathbf{U}_3} \right] \mathbf{V}_1 \| \\
&\leq 2 \| \mathbf{U}_1 \|_{2,\infty} \tau_1 \left(\| \sin\Theta(\mathbf{U}_2, \widehat{\mathbf{U}}_2^{(t-1)}) \| + \| \sin\Theta(\mathbf{U}_3, \widehat{\mathbf{U}}_3^{(t-1)}) \| \right) \\
&\quad + \| \mathbf{U}_1 \|_{2,\infty} \| \mathbf{U}_1^\top \mathbf{Z}_1 \mathbf{V}_1 \|, \qquad \text{(D.4)}
\end{aligned}$$

where the final line used the fact that $\mathcal{P}_{\mathbf{U}_2} \otimes \mathcal{P}_{\mathbf{U}_3} \mathbf{V}_1 = \mathbf{V}_1$ by definition.

We now plug in the bound for T_1 in (D.2) and T_2 in (D.4) to the initial bound in (D.1) to obtain that

$$\begin{aligned}
&\left\|e_m^\top \mathbf{U}_{1\perp}\mathbf{U}_{1\perp}^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{V}_1 \mathbf{\Lambda}_1 \mathbf{U}_1^\top \widehat{\mathbf{U}}_1^{(t)} (\widehat{\mathbf{\Lambda}}_1^{(t)})^{-2}\right\| \\
&\quad \le \frac{8\kappa}{\lambda}\|\mathbf{U}_1\|_{2,\infty}\tau_1\left(\|\sin\Theta(\mathbf{U}_2, \widehat{\mathbf{U}}_2^{(t-1)})\| + \|\sin\Theta(\mathbf{U}_3, \widehat{\mathbf{U}}_3^{(t-1)})\|\right) \\
&\qquad + \frac{4\kappa}{\lambda}\|\mathbf{U}_1\|_{2,\infty}\|\mathbf{U}_1^\top \mathbf{Z}_1 \mathbf{V}_1\| \\
&\qquad + \frac{8\kappa}{\lambda}\tau_1\|\sin\Theta(\widehat{\mathbf{U}}_2^{(t-1)}, \widetilde{\mathbf{U}}_2^{(t-1,1-m)})\| + \frac{8\kappa}{\lambda}\tau_1\|\sin\Theta(\widehat{\mathbf{U}}_3^{(t-1)}, \widetilde{\mathbf{U}}_3^{(t-1,1-m)})\| \\
&\qquad + \frac{4\kappa}{\lambda}\left\|e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t-1,1-m)}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t-1,1-m)}}\right] \mathbf{V}_1\right\| \\
&\quad \le \frac{8\kappa}{\lambda}\|\mathbf{U}_1\|_{2,\infty}\left(\tau_1\eta_1^{(t)} + \|\mathbf{U}_1^\top \mathbf{Z}_1 \mathbf{V}_1\|\right) \\
&\qquad + \frac{8\kappa}{\lambda}\left(\tau_1\eta_1^{(t,1-m)}\right) + \frac{4\kappa}{\lambda}\widetilde{\xi}_1^{t,1-m}
\end{aligned}$$

as desired. □

Lemma 35 (Deterministic Bound for the Quadratic Term)**.** *Suppose* $\lambda/2 \le \lambda_{r_k}(\widehat{\mathbf{\Lambda}}_k^{(t)})$*. Then the quadratic term* $\mathbf{Q}_k^{(t)}$ *satisfies*

$$\begin{aligned}
\|e_m^\top \mathbf{Q}_k^{(t)}\| \le \frac{4}{\lambda^2}\|\mathbf{U}_k\|_{2,\infty}\left(\tau_k\eta_k^{(t)} + \left\|\mathbf{U}_k^\top \mathbf{Z}_k\left[\mathcal{P}_{\mathbf{U}_{k+1}} \otimes \mathcal{P}_{\mathbf{U}_{k+1}}\right]\right\|\right) + \frac{16}{\lambda^2}\tau_k^2\left(\eta_k^{(t,k-m)}\right) \\
+ \frac{4}{\lambda^2}\xi_k^{t,k-m}\left(\tau_k\|\sin\Theta(\widehat{\mathbf{U}}_k^{(t)}, \mathbf{U}_k)\| + \tau_k\eta_k^{(t-1)} + \left\|\mathbf{U}_k\mathbf{U}_k^\top \mathbf{Z}_k \mathcal{P}_{\mathbf{U}_{k+1}} \otimes \mathcal{P}_{\mathbf{U}_{k+2}}\right\|\right).
\end{aligned}$$

Proof of Lemma 35. Similar to Lemma 34 we prove for $k = 1$; the case for $k = 2$ or $k = 3$ follows by modifying the index of t according to the definition of $\widehat{\mathcal{P}}_k^{(t)}$.

Recall that

$$\mathbf{Q}_1^{(t)} = \mathbf{U}_{1\perp}\mathbf{U}_{1\perp}^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)} (\widehat{\mathbf{\Lambda}}_1^{(t)})^{-2}.$$

Observe that $\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}$ is a projection matrix and hence equals its square. Therefore,

we simply decompose by noting that under the condition that $\widehat{\lambda}_{r_k}(\widehat{\mathbf{\Lambda}}_k^{(t)}) \geq \lambda/2$

$$
\begin{aligned}
\left\|e_m^\top \mathbf{Q}_1^{(t)}\right\| &= \left\|\mathbf{U}_{1\perp}\mathbf{U}_{1\perp}^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)} (\widehat{\mathbf{\Lambda}}_1^{(t)})^{-2}\right\| \\
&\leq \frac{4}{\lambda^2}\left\|e_m^\top \mathbf{U}_1\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)}\right\| + \frac{4}{\lambda^2}\left\|e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)}\right\| \\
&\leq \frac{4}{\lambda^2}\|\mathbf{U}_1\|_{2,\infty}\left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)}\right\| \\
&\quad + \frac{4}{\lambda^2}\left\|e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)}\right\| \\
&\leq \frac{4}{\lambda^2}\|\mathbf{U}_1\|_{2,\infty}\tau_1\left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\right\| \\
&\quad + \frac{4}{\lambda^2}\left\|e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)}\right\| \\
&\leq \frac{4}{\lambda^2}\|\mathbf{U}_1\|_{2,\infty}\tau_1\left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\right\| \\
&\quad + \frac{4}{\lambda^2}\left\|e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{1-m,t-1}\otimes\widehat{\mathbf{U}}_3^{(t-1)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}}\right] \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)}\right\| \\
&\quad + \frac{4}{\lambda^2}\left\|e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{1-m,t-1}\otimes\widehat{\mathbf{U}}_3^{(t-1)}}\right] \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)}\right\| \\
&\leq \frac{4}{\lambda^2}\|\mathbf{U}_1\|_{2,\infty}\tau_1\left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\right\| + \frac{8}{\lambda^2}\tau_1^2\|\sin\Theta(\widetilde{\mathbf{U}}_2^{1-m,t-1}, \widehat{\mathbf{U}}_2^{(t-1)})\| \\
&\quad + \frac{4}{\lambda^2}\left\|e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{1-m,t-1}} \otimes \left(\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_3^{1-m,t-1}}\right)\right] \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)}\right\| \\
&\quad + \frac{4}{\lambda^2}\left\|e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{1-m,t-1}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{1-m,t-1}}\right] \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)}\right\| \\
&\leq \frac{4}{\lambda^2}\|\mathbf{U}_1\|_{2,\infty}\tau_1\left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\right\| \\
&\quad + \frac{8}{\lambda^2}\tau_1^2\left(\|\sin\Theta(\widetilde{\mathbf{U}}_2^{1-m,t-1}, \widehat{\mathbf{U}}_2^{(t-1)})\| + \|\sin\Theta(\widetilde{\mathbf{U}}_3^{1-m,t-1}, \widehat{\mathbf{U}}_3^{(t-1)})\|\right) \\
&\quad + \frac{4}{\lambda^2}\left\|e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{1-m,t-1}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{1-m,t-1}}\right] \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right] \mathbf{Z}_1^\top \widehat{\mathbf{U}}_1^{(t)}\right\| \\
&\leq \frac{4}{\lambda^2}\|\mathbf{U}_1\|_{2,\infty}\tau_1\left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\right\| \\
&\quad + \frac{8}{\lambda^2}\tau_1^2\left(\|\sin\Theta(\widetilde{\mathbf{U}}_2^{1-m,t-1}, \widehat{\mathbf{U}}_2^{(t-1)})\| + \|\sin\Theta(\widetilde{\mathbf{U}}_3^{1-m,t-1}, \widehat{\mathbf{U}}_3^{(t-1)})\|\right) \\
&\quad + \frac{4}{\lambda^2}\left\|e_m^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{1-m,t-1}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{1-m,t-1}}\right]\right\|\left\|(\widehat{\mathbf{U}}_1^{(t)})^\top \mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{1-m,t-1}}\right\| \\
&\leq \frac{4}{\lambda^2}\|\mathbf{U}_1\|_{2,\infty}\tau_1\left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\right\| \\
&\quad + \frac{8}{\lambda^2}\tau_1^2\left(\eta_1^{(t,1-m)}\right) + \frac{4}{\lambda^2}\xi_1^{t,1-m}\left\|(\widehat{\mathbf{U}}_1^{(t)})^\top \mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right\|.
\end{aligned}
$$

Finally, we note that

$$\begin{aligned}
\left\|(\widehat{\mathbf{U}}_1^{(t)})^\top \mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right\| &= \left\|\widehat{\mathbf{U}}_1^{(t)}(\widehat{\mathbf{U}}_1^{(t)})^\top \mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right\| \\
&\leq \left\|\left(\widehat{\mathbf{U}}_1^{(t)}(\widehat{\mathbf{U}}_1^{(t)})^\top - \mathbf{U}_1\mathbf{U}_1^\top\right)\mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right\| \\
&\quad + \left\|\mathbf{U}_1\mathbf{U}_1^\top \mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right\| \\
&\leq \|\sin\Theta(\widehat{\mathbf{U}}_1^{(t)}, \mathbf{U}_1)\|\tau_1 + \left\|\mathbf{U}_1\mathbf{U}_1^\top \mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right\| \\
&\leq \|\sin\Theta(\widehat{\mathbf{U}}_1^{(t)}, \mathbf{U}_1)\|\tau_1 + \left\|\mathbf{U}_1\mathbf{U}_1^\top \mathbf{Z}_1 \left(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} - \mathcal{P}_{\mathbf{U}_2}\right) \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right\| \\
&\quad + \left\|\mathbf{U}_1\mathbf{U}_1^\top \mathbf{Z}_1 \mathcal{P}_{\mathbf{U}_2} \otimes \left(\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} - \mathcal{P}_{\mathbf{U}_3}\right)\right\| \\
&\quad + \left\|\mathbf{U}_1\mathbf{U}_1^\top \mathbf{Z}_1 \mathcal{P}_{\mathbf{U}_2} \otimes \mathcal{P}_{\mathbf{U}_3})\right\| \\
&\leq \tau_1\|\sin\Theta(\widehat{\mathbf{U}}_1^{(t)}, \mathbf{U}_1)\| + \tau_1\|\sin\Theta(\widehat{\mathbf{U}}_2^{(t-1)}, \mathbf{U}_2)\| + \tau_1\|\sin\Theta(\widehat{\mathbf{U}}_3^{(t-1)}, \mathbf{U}_2)\| \\
&\quad + \left\|\mathbf{U}_1\mathbf{U}_1^\top \mathbf{Z}_1 \mathcal{P}_{\mathbf{U}_2} \otimes \mathcal{P}_{\mathbf{U}_3}\right\| \\
&\leq \tau_1\|\sin\Theta(\widehat{\mathbf{U}}_1^{(t)}, \mathbf{U}_1)\| + \tau_1\eta_1^{(t)} + \left\|\mathbf{U}_1\mathbf{U}_1^\top \mathbf{Z}_1 \mathcal{P}_{\mathbf{U}_2} \otimes \mathcal{P}_{\mathbf{U}_3}\right\|,
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\right\| &\leq \left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\left(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} - \mathcal{P}_{\mathbf{U}_2}\right) \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\right\| \\
&\quad + \left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\mathbf{U}_2} \otimes \left(\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} - \mathcal{P}_{\mathbf{U}_3}\right)\right]\right\| \\
&\quad + \left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\mathbf{U}_2} \otimes \mathcal{P}_{\mathbf{U}_3}\right]\right\| \\
&\leq \tau_1\eta_1^{(t)} + \left\|\mathbf{U}_1^\top \mathbf{Z}_1 \left[\mathcal{P}_{\mathbf{U}_2} \otimes \mathcal{P}_{\mathbf{U}_3}\right]\right\|.
\end{aligned}$$

Plugging in these bounds to our initial bound completes the proof. □

Lemma 36 (Eigengaps)**.** *Suppose that* $\tau_k \leq \lambda/4$ *and that*

$$\eta_k^{(t)} \leq \frac{1}{4}.$$

Then the following bounds hold:

$$\lambda_{r_k}\left(\mathbf{T}_k\widehat{\mathcal{P}}_k^{(t)} + \mathbf{Z}_k^{j-m}\widetilde{\mathcal{P}}_k^{t,j-m}\widehat{\mathcal{P}}_k^{(t)}\right) \geq \frac{3\lambda}{4};$$
$$\lambda_{r_k+1}\left(\mathbf{T}_k\widehat{\mathcal{P}}_k^{(t)} + \mathbf{Z}_k\widehat{\mathcal{P}}_k^{(t)}\right) \leq \frac{\lambda}{4}.$$

Proof. Note that since $\mathbf{Z}_k^{j-m}$ is $\mathbf{Z}_k$ with columns (or rows if $k = j$) removed, it holds that

$$\mathbf{Z}_k^\top\mathbf{Z}_k \succcurlyeq (\mathbf{Z}_k^{j-m})^\top\mathbf{Z}_k^{j-m}$$

and hence that

$$\widetilde{\mathcal{P}}_k^{t,j-m}\mathbf{Z}_k^\top\mathbf{Z}_k\widetilde{\mathcal{P}}_k^{t,j-m} \succcurlyeq \widetilde{\mathcal{P}}_k^{t,j-m}(\mathbf{Z}_k^{j-m})^\top\mathbf{Z}_k^{j-m}\widetilde{\mathcal{P}}_k^{t,j-m}.$$

Taking norms, it holds that

$$\begin{aligned}\left\|\mathbf{Z}_k^{j-m}\widetilde{\mathcal{P}}_k^{t,j-m}\right\|^2 &= \left\|\widetilde{\mathcal{P}}_k^{t,j-m}(\mathbf{Z}_k^{j-m})^\top\mathbf{Z}_k^{j-m}\widetilde{\mathcal{P}}_k^{t,j-m}\right\| \\ &\leq \left\|\widetilde{\mathcal{P}}_k^{t,j-m}\mathbf{Z}_k^\top\mathbf{Z}_k\widetilde{\mathcal{P}}_k^{t,j-m}\right\|^2 \\ &= \left\|\mathbf{Z}_k\widetilde{\mathcal{P}}_k^{t,j-m}\right\| \\ &\leq \tau_k^2,\end{aligned}$$

where we took the supremum in the final inequality. Therefore, $\|\mathbf{Z}_k^{j-m}\widetilde{\mathcal{P}}_k^{t,j-m}\| \leq \tau_k$. Therefore, by Weyl's inequality, it holds that

$$\begin{aligned}\left|\lambda_{r_k}\left(\mathbf{T}_k\widehat{\mathcal{P}}_k^{(t)} + \mathbf{Z}_k^{j-m}\widetilde{\mathcal{P}}_k^{t,j-m}\widehat{\mathcal{P}}_k^{(t)}\right) - \lambda_{r_k}(\mathbf{T}_k\widehat{\mathcal{P}}_k^{(t)})\right| &\leq \|\mathbf{Z}_k^{j-m}\widetilde{\mathcal{P}}_k^{t,j-m}\widehat{\mathcal{P}}_k^{(t)}\| \\ &\leq \|\mathbf{Z}_k^{j-m}\widetilde{\mathcal{P}}_k^{t,j-m}\| \\ &\leq \tau_k \\ &\leq \frac{\lambda}{4}. \qquad \text{(D.5)}\end{aligned}$$

Next, when $\eta_k^{(t)} \leq \frac{1}{4}$, this implies that

$$\max\left(\|\sin\Theta(\widehat{\mathbf{U}}_{k+1}^{(t-1)}, \mathbf{U}_{k+1})\|, \|\sin\Theta(\widehat{\mathbf{U}}_{k+2}^{(t-1)}, \mathbf{U}_{k+2})\|\right) \leq \frac{1}{4},$$

and hence that

$$\begin{aligned}
\lambda_{\min}(\mathbf{T}_k\widehat{\mathcal{P}}_k^{(t)}) &= \lambda_{\min}\left(\mathbf{T}_k\mathbf{U}_{k+1}\otimes\mathbf{U}_{k+2}(\mathbf{U}_{k+1}\otimes\mathbf{U}_{k+2})^\top\widehat{\mathcal{P}}_k^{(t)}\right) \\
&\geq \lambda\lambda_{\min}\left((\mathbf{U}_{k+1}\otimes\mathbf{U}_{k+2})^\top\widehat{\mathcal{P}}_k^{(t)}\right) \\
&\geq \lambda\lambda_{\min}(\mathbf{U}_{k+1}^\top\widehat{\mathbf{U}}_{k+1}^{(t-1)})\lambda_{\min}(\mathbf{U}_{k+2}^\top\widehat{\mathbf{U}}_{k+2}^{(t-1)}) \\
&\geq \lambda(1-\frac{1}{16}) \\
&\geq \frac{15}{16}\lambda. \qquad \text{(D.6)}
\end{aligned}$$

Combining (D.6) and (D.5) gives the first claim.

For the second claim, we simply note that by Weyl's inequality,

$$\begin{aligned}
\left|\lambda_{r_k+1}\left(\mathbf{T}_k\widehat{\mathcal{P}}_k^{(t)}+\mathbf{Z}_k\widehat{\mathcal{P}}_k^{(t)}\right)-\lambda_{r_k+1}\left(\mathbf{T}_k\widehat{\mathcal{P}}_k^{(t)}\right)\right| &\leq \|\mathbf{Z}_k\widehat{\mathcal{P}}_k^{(t)}\| \\
&\leq \tau_k \leq \frac{\lambda}{4}.
\end{aligned}$$

Since $\mathbf{T}_k$ is rank r_k, it holds that

$$\lambda_{r_k+1}\left(\mathbf{T}_k\widehat{\mathcal{P}}_k^{(t)}\right) = 0,$$

which proves the second assertion. This completes the proof. □

Lemma 37 (Deterministic Bound for Leave-One-Out Sequence)**.** *Suppose that* $\tau_k \leq \frac{\lambda}{4}$ *and that* $\eta_k^{(t)} \leq \frac{1}{4}$*. Then it holds that*

$$\begin{aligned}
\|\sin\Theta(\widehat{\mathbf{U}}_k^{(t)}, \widetilde{\mathbf{U}}_k^{t,j-m})\| \leq{}& \frac{16\kappa}{\lambda}\tau_k\left(\eta_k^{(t,j-m)}\right)+\frac{16\kappa}{\lambda}\xi_k^{t,j-m}\left(\eta_k^{(t,j-m)}\right) \\
&+\frac{8\kappa}{\lambda}\widetilde{\xi}_k^{t,j-m}+\frac{16}{\lambda^2}\tau_k^2\left(\eta_k^{(t,j-m)}\right)+\frac{8}{\lambda^2}\tau_k\xi_k^{t,j-m}+\frac{4}{\lambda^2}(\xi_k^{t,j-m})^2,
\end{aligned}$$

Proof. We prove the result for $k = 1$; the result for $k = 2$ and $k = 3$ are similar by modifying the index on t.

Recall that $\widehat{\mathbf{U}}_1^{(t)}$ are the singular vectors of the matrix

$$\mathbf{T}_1 \widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)} + \mathbf{Z}_1 \widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}$$

and $\widetilde{\mathbf{U}}_1^{t,j-m}$ are the singular vectors of the matrix

$$\mathbf{T}_1 + \mathbf{Z}_1^{j-m} \widetilde{\mathcal{P}}_1^{t,j-m}$$

Consequently, the projection $\widetilde{\mathbf{U}}_1^{t,j-m} (\widetilde{\mathbf{U}}_1^{t,j-m})^\top$ is also the projection onto the dominant left singular space of the matrix

$$\left(\mathbf{T}_1 + \mathbf{Z}_1^{j-m} \widetilde{\mathcal{P}}_1^{t,j-m} \right) \widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)},$$

Therefore, both projections are projections onto the dominant eigenspaces of the matrices defined via

$$\begin{aligned}
\widehat{\mathbf{A}} &:= \mathbf{T}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top + \left[\mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top + \mathbf{T}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{Z}_1^\top + \mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{Z}_1^\top \right]; \\
\widetilde{\mathbf{A}} &:= \mathbf{T}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top + \Big[\mathbf{Z}_1^{j-m} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top \\
&\qquad + \mathbf{T}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} (\mathbf{Z}_1^{j-m})^\top \\
&\qquad + \mathbf{Z}_1^{j-m} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \mathbf{Z}_1^{j-m})^\top \Big].
\end{aligned}$$

Therefore, the perturbation $\widehat{\mathbf{A}} - \widetilde{\mathbf{A}}$ is equal to the sum of three terms, defined via

$$\begin{aligned}
\mathbf{P}_1 &\coloneqq \mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top - \mathbf{Z}_1^{j-m} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top \\
&= \left[\mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} - \mathbf{Z}_1^{j-m} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \right] \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top \\
\mathbf{P}_2 &\coloneqq \mathbf{T}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{Z}_1^\top - \mathbf{T}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} (\mathbf{Z}_1^{j-m})^\top \\
&= \mathbf{T}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \left(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{Z}_1^\top - \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} (\mathbf{Z}_1^{j-m})^\top \right) \\
\mathbf{P}_3 &\coloneqq \mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{Z}_1^\top - \mathbf{Z}_1^{j-m} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \mathbf{Z}_1^{j-m})^\top
\end{aligned}$$

where we have used the fact that $\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}}$ is a projection matrix and hence equal to its square. We now bound each term successively.

The term $\|\mathbf{P}_1\|$: Observe that

$$\begin{aligned}
\mathbf{P}_1 &= \left[\mathbf{Z}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} - \mathbf{Z}_1^{j-m} \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \right] \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top \\
&= \left[\mathbf{Z}_1 \left(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \right) \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right] \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top \\
&\quad + \left[\mathbf{Z}_1 \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \left(\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \right) \right] \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top \\
&\quad + \left[\mathbf{Z}_1 - \mathbf{Z}_1^{j-m} \right] \left(\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \right) \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \right] \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{T}_1^\top \\
&\quad + \left[\mathbf{Z}_1 - \mathbf{Z}_1^{j-m} \right] \left(\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \right) \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \left[\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \right] \mathbf{T}_1^\top \\
&\quad + \left[\mathbf{Z}_1 - \mathbf{Z}_1^{j-m} \right] \left(\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \right) \mathbf{T}_1^\top .
\end{aligned}$$

Taking norms yields

$$\begin{aligned}
\|\mathbf{P}_1\| &\leq 2\lambda_1 \tau_1 \left(\|\sin\Theta(\widehat{\mathbf{U}}_2^{(t-1)}, \widetilde{\mathbf{U}}_2^{t-1,j-m})\| + \|\sin\Theta(\widehat{\mathbf{U}}_3^{(t-1)}, \widetilde{\mathbf{U}}_3^{t-1,j-m})\| \right) \\
&\quad + 2\lambda_1 \| \left[\mathbf{Z}_1 - \mathbf{Z}_1^{j-m} \right] \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \| \|\sin\Theta(\widehat{\mathbf{U}}_2^{(t-1)}, \widetilde{\mathbf{U}}_2^{t-1,j-m})\| \\
&\quad + 2\lambda_1 \| \left[\mathbf{Z}_1 - \mathbf{Z}_1^{j-m} \right] \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \| \|\sin\Theta(\widehat{\mathbf{U}}_3^{(t-1)}, \widetilde{\mathbf{U}}_3^{t-1,j-m})\| \\
&\quad + \lambda_1 \| \left[\mathbf{Z}_1 - \mathbf{Z}_1^{j-m} \right] \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \mathbf{V}_1 \| \\
&\leq 2\lambda_1 \tau_1 \left(\eta_1^{(t,j-m)} \right) + 2\lambda_1 \xi_1^{t,j-m} \left(\eta_1^{(t,j-m)} \right) + \lambda_1 \widetilde{\xi}_1^{t,j-m} .
\end{aligned}$$

For $\mathbf{P}_2$ we proceed similarly. It holds that

$$
\begin{aligned}
\mathbf{P}_2 &= \mathbf{T}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \left(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{Z}_1^\top - \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} (\mathbf{Z}_1^{j-m})^\top \right) \\
&= \mathbf{T}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \left(\left[\mathcal{P}_{\widehat{\mathbf{U}}_{k+1}^{(t)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \right] \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \mathbf{Z}_1^\top \right) \\
&\quad + \mathbf{T}_1 \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \left(\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \left[\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} \right] \mathbf{Z}_1^\top \right) \\
&\quad + \mathbf{T}_1 \left(\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \right) \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \left(\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} (\mathbf{Z}_1 - \mathbf{Z}_1^{j-m})^\top \right) \\
&\quad + \mathbf{T}_1 \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \left(\mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}} \right) \left(\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} (\mathbf{Z}_1 - \mathbf{Z}_1^{j-m})^\top \right) \\
&\quad + \mathbf{T}_1 \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} (\mathbf{Z}_1 - \mathbf{Z}_1^{j-m})^\top.
\end{aligned}
$$

Taking norms yields the same upper bound as for $\|\mathbf{P}_1\|$.

For the the term $\mathbf{P}_3$, we note that since $\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)} \otimes \widehat{\mathbf{U}}_3^{(t-1)}}$ is a projection matrix and hence

equal to its cube, it holds that

$$
\begin{aligned}
\|\mathbf{P}_3\| \leq & \left\|\mathbf{Z}_1\left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}}\right] \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}}\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}}\mathbf{Z}_1\right\| \\
& + \left\|\mathbf{Z}_1\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \left[\mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}}\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}}\mathbf{Z}_1\right\| \\
& + \left\|\mathbf{Z}_1\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}}\left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}} - \mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}}\right] \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\mathbf{Z}_1\right\| \\
& + \left\|\mathbf{Z}_1\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}}\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \left[\mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t-1)}}\right]\mathbf{Z}_1\right\| \\
& + \left\|\mathbf{Z}_1\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}}\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\left[\mathbf{Z}_1 - \mathbf{Z}_1^{j-m}\right]^\top\right\| \\
& + \left\|\left[\mathbf{Z}_1 - \mathbf{Z}_1^{j-m}\right]\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t-1)}\otimes\widehat{\mathbf{U}}_3^{(t-1)}}\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\mathbf{Z}_1^{j-m}\right\| \\
\leq & \, 2\tau_1^2\left(\|\sin\Theta(\widehat{\mathbf{U}}_2^{(t-1)}, \widetilde{\mathbf{U}}_2^{t-1,j-m})\| + \|\sin\Theta(\widehat{\mathbf{U}}_3^{(t-1)}, \widetilde{\mathbf{U}}_3^{t-1,j-m})\|\right) \\
& + \tau_1\left\|\left[\mathbf{Z}_1 - \mathbf{Z}_1^{j-m}\right]\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\right\| \\
& + \left\|\mathbf{Z}_1^{j-m}\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\right\|\left\|\left[\mathbf{Z}_1 - \mathbf{Z}_1^{j-m}\right]\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\right\| \\
\leq & \, 4\tau_1^2\left(\eta_1^{(t,j-m)}\right) + \tau_1\xi_1^{t,j-m} + \left\|\mathbf{Z}_1^{j-m}\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\right\|\xi_1^{t,j-m} \\
\leq & \, 4\tau_1^2\left(\eta_1^{(t,j-m)}\right) + 2\tau_1\xi_1^{t,j-m} + \left\|\left(\mathbf{Z}_1 - \mathbf{Z}_1^{j-m}\right)\mathcal{P}_{\widetilde{\mathbf{U}}_2^{t-1,j-m}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{t-1,j-m}}\right\|\xi_1^{t,j-m} \\
\leq & \, 4\tau_1^2\left(\eta_1^{(t,j-m)}\right) + 2\tau_1\xi_1^{t,j-m} + \left(\xi_1^{t,j-m}\right)^2.
\end{aligned}
$$

We note that by Lemma 36, it holds that

$$
\begin{aligned}
\lambda_{r_1}(\widetilde{\mathbf{A}}) - \lambda_{r_1+1}(\widehat{\mathbf{A}}) &= \lambda_{r_1}^2\left(\mathbf{T}_1\widehat{\mathcal{P}}_1^{(t)} + \mathbf{Z}_1^{j-m}\widetilde{\mathcal{P}}_1^{t,j-m}\right) - \lambda_{r_1+1}^2\left(\mathbf{T}_1\widehat{\mathcal{P}}_1^{(t)} + \mathbf{Z}_k\widehat{\mathcal{P}}_1^{(t)}\right) \\
&\geq \left(\frac{3}{4}\lambda\right)^2 - \left(\frac{\lambda}{4}\right)^2 \\
&\geq \frac{\lambda^2}{4}.
\end{aligned}
$$

Consequently, by the Davis-Kahan Theorem, it holds that

$$
\|\sin\Theta(\widehat{\mathbf{U}}_1^{(t)}, \widetilde{\mathbf{U}}_1^{t,j-m})\| \leq \frac{4}{\lambda^2}\left(\|\mathbf{P}_1\| + \|\mathbf{P}_2\| + \|\mathbf{P}_3\|\right),
$$

which holds under the eigengap condition by Lemma 36 and the assumption $\tau_k \leq \frac{\lambda}{4}$. There-

fore,

$$\begin{aligned}
\|\sin\Theta(\widehat{\mathbf{U}}_1^{(t)}, \widetilde{\mathbf{U}}_1^{t,j-m})\| &\leq \frac{4}{\lambda^2}\Big(\|\mathbf{P}_1\| + \|\mathbf{P}_2\| + \|\mathbf{P}_3\|\Big) \\
&\leq \frac{8}{\lambda^2}\left(2\lambda_1\tau_1\left(\eta_1^{(t,j-m)}\right) + 2\lambda_1\xi_1^{t,j-m}\left(\eta_1^{(t,j-m)}\right) + \lambda_1\widetilde{\xi}_1^{t,j-m}\right) \\
&\quad + \frac{4}{\lambda^2}\left(4\tau_1^2\left(\eta_1^{(t,j-m)}\right) + 2\tau_1\xi_1^{t,j-m} + \left(\xi_1^{t,j-m}\right)^2\right) \\
&\leq \frac{16\kappa}{\lambda}\tau_1\left(\eta_1^{(t,j-m)}\right) + \frac{16\kappa}{\lambda}\xi_1^{t,j-m}\left(\eta_1^{(t,j-m)}\right) \\
&\quad + \frac{8\kappa}{\lambda}\widetilde{\xi}_1^{t,j-m} + \frac{16}{\lambda^2}\tau_1^2\left(\eta_1^{(t,j-m)}\right) + \frac{8}{\lambda^2}\tau_1\xi_1^{t,j-m} + \frac{4}{\lambda^2}(\xi_1^{t,j-m})^2
\end{aligned}$$

as desired. □

D.1.3 Probabilistic Bounds on Good Events

This section contains high-probability bounds for the terms considered in the previous subsection. Let $r = \max r_k$, $p = \max p_k$. In what follows, we denote

$$\delta_{\mathrm{L}}^{(k)} := C_0\kappa\sqrt{p_k\log(p)},$$

where C_0 is taken to be some fixed constant.

We will also recall the notation from the previous section:

$$\widehat{\mathcal{P}}_k^{(t)} := \begin{cases} \mathcal{P}_{\widehat{\mathbf{U}}_{k+1}^{(t-1)} \otimes \widehat{\mathbf{U}}_{k+2}^{(t-1)}} & k = 1; \\ \mathcal{P}_{\widehat{\mathbf{U}}_{k+1}^{(t-1)} \otimes \widehat{\mathbf{U}}_{k+2}^{(t)}} & k = 2; \\ \mathcal{P}_{\widehat{\mathbf{U}}_{k+1}^{(t)} \otimes \widehat{\mathbf{U}}_{k+2}^{(t)}} & k = 3. \end{cases}$$

$$\widetilde{\mathcal{P}}_k^{t,j-m} := \begin{cases} \mathcal{P}_{\widetilde{\mathbf{U}}_{k+1}^{(t-1,j-m)} \otimes \widetilde{\mathbf{U}}_{k+2}^{(t-1,j-m)}} & k = 1; \\ \mathcal{P}_{\widetilde{\mathbf{U}}_{k+1}^{(t-1,j-m)} \otimes \widetilde{\mathbf{U}}_{k+2}^{(t,j-m)}} & k = 2; \\ \mathcal{P}_{\widetilde{\mathbf{U}}_{k+1}^{(t,j-m)} \otimes \widetilde{\mathbf{U}}_{k+2}^{(t,j-m)}} & k = 3. \end{cases}$$

$$\mathbf{L}_k^{(t)} := \mathbf{U}_{k\perp}\mathbf{U}_{k\perp}^\top \mathbf{Z}_k \widehat{\mathcal{P}}_k^{(t)} \mathbf{T}_k^\top \widehat{\mathbf{U}}_k^{(t-1)} (\widehat{\mathbf{\Lambda}}_k^{(t-1)})^{-2};$$

$$\mathbf{Q}_k^{(t)} := \mathbf{U}_{k\perp}\mathbf{U}_{k\perp}^\top \mathbf{Z}_k \widehat{\mathcal{P}}_k^{(t)} \mathbf{Z}_k^\top \widehat{\mathbf{U}}_k^{(t-1)} (\widehat{\mathbf{\Lambda}}_k^{(t-1)})^{-2}$$

$$\tau_k := \sup_{\substack{\|\mathbf{U}_1\|=1, \mathrm{rank}(\mathbf{U}_1) \le 2r_{k+1} \\ \|\mathbf{U}_2\|=1, \mathrm{rank}(\mathbf{U}_2) \le 2r_{k+2}}} \|\mathbf{Z}_k \Big(\mathcal{P}_{\mathbf{U}_1} \otimes \mathcal{P}_{\mathbf{U}_2} \Big)\|;$$

$$\xi_k^{(t,j-m)} := \left\| \left(\mathbf{Z}_k^{j-m} - \mathbf{Z}_k \right) \widetilde{\mathcal{P}}_k^{t,j-m} \right\|$$

$$\widetilde{\xi}_k^{(t,j-m)} := \left\| \left(\mathbf{Z}_k^{j-m} - \mathbf{Z}_k \right) \widetilde{\mathcal{P}}_k^{t,j-m} \mathbf{V}_k \right\|$$

$$\eta_k^{(t,j-m)} := \begin{cases} \|\sin\Theta(\widetilde{\mathbf{U}}_{k+1}^{(t-1,j-m)}, \widehat{\mathbf{U}}_{k+1}^{(t-1)})\| + \|\sin\Theta(\widetilde{\mathbf{U}}_{k+2}^{(t-1,j-m)}, \widehat{\mathbf{U}}_{k+2}^{(t-1)})\| & k = 1 \\ \|\sin\Theta(\widetilde{\mathbf{U}}_{k+1}^{(t-1,j-m)}, \widehat{\mathbf{U}}_{k+1}^{(t-1)})\| + \|\sin\Theta(\widetilde{\mathbf{U}}_{k+2}^{(t,j-m)}, \widehat{\mathbf{U}}_{k+2}^{(t)})\| & k = 2 \\ \|\sin\Theta(\widetilde{\mathbf{U}}_{k+1}^{(t,j-m)}, \widehat{\mathbf{U}}_{k+1}^{(t)})\| + \|\sin\Theta(\widetilde{\mathbf{U}}_{k+2}^{(t,j-m)}, \widehat{\mathbf{U}}_{k+2}^{(t)})\| & k = 3 \end{cases}$$

$$\eta_k^{(t)} := \begin{cases} \|\sin\Theta(\mathbf{U}_{k+1}, \widehat{\mathbf{U}}_{k+1}^{(t-1)})\| + \|\sin\Theta(\mathbf{U}_{k+2}, \widehat{\mathbf{U}}_{k+2}^{(t-1)})\| & k = 1 \\ \|\sin\Theta(\mathbf{U}_{k+1}, \widehat{\mathbf{U}}_{k+1}^{(t-1)})\| + \|\sin\Theta(\mathbf{U}_{k+2}, \widehat{\mathbf{U}}_{k+2}^{(t)})\| & k = 2 \cdot \\ \|\sin\Theta(\mathbf{U}_{k+1}, \widehat{\mathbf{U}}_{k+1}^{(t)})\| + \|\sin\Theta(\mathbf{U}_{k+2}, \widehat{\mathbf{U}}_{k+2}^{(t)})\| & k = 3 \end{cases}$$

We will also need to define several probabilistic events. The first event $\mathcal{E}_{\mathrm{Good}}$ collects several probabilistic bounds that hold independently of t, provided $t_{\max} \le C \log(p)$ for some

constant C:

$$\begin{aligned}\mathcal{E}_{\text{Good}} \coloneqq &\left\{ \max_k \tau_k \le C\sqrt{pr} \right\} \bigcap \left\{ \|\sin\Theta(\widehat{\mathbf{U}}_k^{(t)}, \mathbf{U}_k)\| \le \frac{\delta_{\text{L}}^{(k)}}{\lambda} + \frac{1}{2^t} \text{ for all } t \le t_{\max} \text{ and } 1 \le k \le 3 \right\} \\ &\bigcap \left\{ \max_k \left\| \mathbf{U}_k^\top \mathbf{Z}_k \mathbf{V}_k \right\| \le C\left(\sqrt{r} + \sqrt{\log(p)}\right) \right\}; \\ &\bigcap \left\{ \max_k \left\| \mathbf{U}_k^\top \mathbf{Z}_k \mathcal{P}_{\mathbf{U}_{k+1}} \otimes \mathcal{P}_{\mathbf{U}_{k+2}} \right\| \le C\left(r + \sqrt{\log(p)}\right) \right\}; \\ &\bigcap \left\{ \max_k \left\| \mathbf{Z}_k \mathbf{V}_k \right\| \le C\sqrt{p_k} \right\}.\end{aligned}$$

By Lemma 48, the proof in Zhang and Xia (2018), and a standard ε-net argument, it is straightforward to show that

$$\mathbb{P}\left(\mathcal{E}_{\text{Good}}\right) \ge 1 - O(p^{-30}).$$

Note that Zhang and Xia (2018) assumes homoskedastic entries, but their proof of Theorem 1 goes through in precisely the same way under the conditions of Theorem 11, since their induction argument only requires a suitably warm initialization. Alternatively, one can apply Theorem 1 of Luo et al. (2021), where the $\sin\Theta$ bounds hold by the assumption $\lambda \gtrsim \kappa\sqrt{\log(p)}p/p_{\min}^{1/4}$.

We now define several events we use in our induction argument. Set

$$\begin{aligned}\mathcal{E}_{2,\infty}^{t,k} &\coloneqq \left\{ \|\widehat{\mathbf{U}}_k^{(t)} - \mathbf{U}_k \mathbf{W}_k^{(t)}\|_{2,\infty} \le \left(\frac{\delta_{\text{L}}^{(k)}}{\lambda} + \frac{1}{2^t}\right)\mu_0\sqrt{\frac{r_k}{p_k}} \right\}; \\ \mathcal{E}_{j-m}^{t,k} &\coloneqq \left\{ \|\sin\Theta(\widetilde{\mathbf{U}}_k^{t,j-m}, \widehat{\mathbf{U}}_k^{(t)})\| \le \left(\frac{\delta_{\text{L}}^{(k)}}{\lambda} + \frac{1}{2^t}\right)\mu_0\sqrt{\frac{r_k}{p_j}} \right\}; \\ \mathcal{E}_{\text{main}}^{t_0-1,1} &\coloneqq \bigcap_{t=1}^{t_0-1} \left\{ \bigcap_{k=1}^{3} \mathcal{E}_{2,\infty}^{t,k} \cap \bigcap_{j=1}^{3} \bigcap_{m=1}^{p_j} \mathcal{E}_{k-m}^{t,j} \right\}; \\ \mathcal{E}_{\text{main}}^{t_0-1,2} &\coloneqq \mathcal{E}_{\text{main}}^{t_0-1,1} \cap \left\{ \bigcap_{k=1}^{3} \bigcap_{m=1}^{p_k} \mathcal{E}_{k-m}^{t_0,1} \right\} \cap \mathcal{E}_{2,\infty}^{t_0,1} \\ \mathcal{E}_{\text{main}}^{t_0-1,3} &\coloneqq \mathcal{E}_{\text{main}}^{t_0-1,2} \cap \left\{ \bigcap_{k=1}^{3} \bigcap_{m=1}^{p_k} \mathcal{E}_{k-m}^{t_0,2} \right\} \cap \mathcal{E}_{2,\infty}^{t_0,2}.\end{aligned}$$

The event $\mathcal{E}_{2,\infty}^{t,k}$ concerns the desired bound, the event $\mathcal{E}_{j-m}^{t,k}$ controls the leave one out sequences, and the other events $\mathcal{E}_{\text{main}}^{t_0-1,k}$ are simply the intersection of these events, mainly

introduced for convenience.

Finally, the following event concerns the incoherence of our leave-one-out sequences:

$$\begin{aligned}\widetilde{\mathcal{E}}_{j-m}^{t,k} := \Bigg\{ &\|\widetilde{\mathcal{P}}_k^{t_0,j-m}\mathbf{V}_k\|_{2,\infty} \le \mu_0^2\frac{\sqrt{r_{-k}}}{p_j}\left(\frac{\delta_{\mathrm{L}}^{(k+1)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\left(\frac{\delta_{\mathrm{L}}^{(k+2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right); \\ &+\mu_0^2\frac{\sqrt{r_{-k}}}{\sqrt{p_jp_{k+2}}}\left(\frac{\delta_{\mathrm{L}}^{(k+1)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0^2\frac{\sqrt{r_{-k}}}{\sqrt{p_jp_{k+1}}}\left(\frac{\delta_{\mathrm{L}}^{(k+2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right) \\ &+6\mu_0^2\frac{\sqrt{r_{-k}}}{\sqrt{p_{-k}}}\left(\frac{\delta_{\mathrm{L}}^{(k+1)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+3\mu_0^2\frac{\sqrt{r_{-k}}}{\sqrt{p_{-k}}}\left(\frac{\delta_{\mathrm{L}}^{(k+2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0\sqrt{\frac{r_k}{p_{-k}}}.\Bigg\} \\ \bigcap\Bigg\{ &\|\widetilde{\mathcal{P}}_k^{t_0,j-m}\|_{2,\infty} \le \mu_0^2\frac{\sqrt{r_{-k}}}{p_j}\left(\frac{\delta_{\mathrm{L}}^{(k+1)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\left(\frac{\delta_{\mathrm{L}}^{(k+2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right) \\ &+2\mu_0^2\frac{\sqrt{r_{-k}}}{\sqrt{p_jp_{k+2}}}\left(\frac{\delta_{\mathrm{L}}^{(k+1)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+2\mu_0^2\frac{\sqrt{r_{-k}}}{\sqrt{p_jp_{k+1}}}\left(\frac{\delta_{\mathrm{L}}^{(k+2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+2\mu_0^2\frac{\sqrt{r_{-k}}}{\sqrt{p_{-k}}}.\Bigg\}.\end{aligned}$$

While the precise definition of the event $\widetilde{\mathcal{E}}_{j-m}^{t,k}$ is complicated, it is useful to keep in mind that the event will be used as an event independent of the nonzero elements in the matrices $\mathbf{Z}_k - \mathbf{Z}_k^{j-m}$, and it simply controls the incoherence of the leave-one-out sequences.

The following lemma shows that the leave-one-out sequences are incoherent whenever there are bounds on the previous iterates in $\ell_{2,\infty}$ norm.

Lemma 38. *For any fixed t_0, j, k, and m with $1 \le t_0 \le t_{\max}$, $1 \le j \le 3$, $1 \le k \le 3$, and $1 \le m \le p_j$, it holds that the set*

$$\mathcal{E}_{\mathrm{Good}} \cap \mathcal{E}_{\mathrm{main}}^{t_0-1,k} \cap \left(\widetilde{\mathcal{E}}_{j-m}^{t_0,k}\right)^c$$

is empty.

Proof. Without loss of generality, we prove the result for $k = 1$; the cases $k = 2$ and $k = 3$ are similar (in fact, the result can be made slightly sharper, but this is not needed for our purposes).

Note that when $t_0 \ge 1$, it holds that on the event $\mathcal{E}_{\mathrm{main}}^{t_0-1,1}$.

$$\|\widehat{\mathbf{U}}_1^{(t_0-1)}\|_{2,\infty} \le \|\widehat{\mathbf{U}}_1^{(t_0-1)} - \mathbf{U}_1\mathbf{W}_1^{(t_0-1)}\|_{2,\infty} + \|\mathbf{U}_1\|_{2,\infty} \le 2\mu_0\sqrt{\frac{r_1}{p_1}}.$$

Similarly,

$$\|\widehat{\mathbf{U}}_2^{(t_0-1)}\|_{2,\infty} \leq 2\mu_0\sqrt{\frac{r_2}{p_2}};$$
$$\|\widehat{\mathbf{U}}_3^{(t_0-1)}\|_{2,\infty} \leq 2\mu_0\sqrt{\frac{r_3}{p_3}}.$$

In addition, on this event it holds that

$$\|\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}}\| \leq \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\mu_0\sqrt{\frac{r_2}{p_j}};$$
$$\|\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}}\| \leq \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\mu_0\sqrt{\frac{r_3}{p_j}}. \tag{D.7}$$

Next, observe that on the events listed,

$$\begin{aligned}
\|\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} \mathbf{V}_1\|_{2,\infty} &\leq \left\| \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right] \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} \mathbf{V}_1 \right\|_{2,\infty} \\
&+ \left\| \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} \mathbf{V}_1 \right\|_{2,\infty} \\
&\leq \left\| \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right] \otimes \left[\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right] \mathbf{V}_1 \right\|_{2,\infty} \\
&+ \left\| \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right] \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \mathbf{V}_1 \right\|_{2,\infty} \\
&+ \left\| \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \otimes \left[\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right] \mathbf{V}_1 \right\|_{2,\infty} \\
&+ \left\| \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \mathbf{V}_1 \right\|_{2,\infty} \\
&\leq \left\| \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right] \otimes \left[\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right] \mathbf{V}_1 \right\|_{2,\infty} \\
&+ \left\| \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right] \otimes \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \mathbf{V}_1 \right\|_{2,\infty} \\
&+ \left\| \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \otimes \left[\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right] \mathbf{V}_1 \right\|_{2,\infty} \\
&+ \left\| \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} - \mathcal{P}_{\mathbf{U}_2} \right] \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \mathbf{V}_1 \right\|_{2,\infty} \\
&+ \left\| \mathcal{P}_{\mathbf{U}_2} \otimes \left[\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} - \mathcal{P}_{\mathbf{U}_3} \right] \mathbf{V}_1 \right\|_{2,\infty} \\
&+ \left\| \mathcal{P}_{\mathbf{U}_2} \otimes \mathcal{P}_{\mathbf{U}_3} \mathbf{V}_1 \right\|_{2,\infty} \\
&=: (I) + (II) + (III) + (IV) + (V) + (VI),
\end{aligned}$$

where

$$
\begin{aligned}
(I) &:= \left\| \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right] \otimes \left[\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right] \mathbf{V}_1 \right\|_{2,\infty}; \\
(II) &:= \left\| \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right] \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \mathbf{V}_1 \right\|_{2,\infty}; \\
(III) &:= \left\| \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \otimes \left[\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right] \mathbf{V}_1 \right\|_{2,\infty}; \\
(IV) &:= \left\| \left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} - \mathcal{P}_{\mathbf{U}_2} \right] \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \mathbf{V}_1 \right\|_{2,\infty}; \\
(V) &:= \left\| \mathcal{P}_{\mathbf{U}_2} \otimes \left[\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} - \mathcal{P}_{\mathbf{U}_3} \right] \mathbf{V}_1 \right\|_{2,\infty}; \\
(VI) &:= \left\| \mathcal{P}_{\mathbf{U}_2} \otimes \mathcal{P}_{\mathbf{U}_3} \mathbf{V}_1 \right\|_{2,\infty}.
\end{aligned}
$$

We now bound each term in turn, where we will use (D.7) repeatedly. We have that

$$
\begin{aligned}
(I) &\leq \left\| \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right] \otimes \left[\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right] \right\|_{2,\infty} \\
&\leq \left\| \mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right\| \left\| \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right\| \\
&\leq \mu_0^2 \frac{\sqrt{r_2 r_3}}{p_j} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right).
\end{aligned} \tag{D.8}
$$

Similarly,

$$
\begin{aligned}
(II) &\leq \left\| \left[\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right] \otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right\|_{2,\infty} \\
&\leq \left\| \mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \right\| \left\| \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right\|_{2,\infty} \\
&\leq \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \mu_0^2 \sqrt{\frac{r_2}{p_j}} \sqrt{\frac{r_3}{p_3}}.
\end{aligned}
$$

Next,

$$
\begin{aligned}
(III) &\leq \left\| \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} \otimes \left[\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} \right] \right\|_{2,\infty} \\
&\leq \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \mu_0^2 \sqrt{\frac{r_3}{p_j}} \sqrt{\frac{r_2}{p_2}}.
\end{aligned}
$$

For the next two terms, we note that for any orthogonal matrix $\mathbf{W} \in \mathbb{O}(r_k)$,

$$\begin{aligned}\|\widehat{\mathbf{U}}_2^{(t_0-1)}\widehat{\mathbf{U}}_2^{t_0-1\top} - \mathbf{U}_2\mathbf{U}_2^\top\|_{2,\infty} &= \|\widehat{\mathbf{U}}_2^{(t_0-1)}\mathbf{W}\mathbf{W}^\top\widehat{\mathbf{U}}_2^{t_0-1\top} - \mathbf{U}_2\mathbf{U}_2^\top\|_{2,\infty} \\ &\leq \|(\widehat{\mathbf{U}}_2^{(t_0-1)}\mathbf{W} - \mathbf{U}_2)(\widehat{\mathbf{U}}_2^{(t_0-1)}\mathbf{W})^\top\|_{2,\infty} \\ &\quad + \|\mathbf{U}_2(\widehat{\mathbf{U}}_2^{(t_0-1)}\mathbf{W} - \mathbf{U}_2)\|_{2,\infty} \\ &\leq \|\widehat{\mathbf{U}}_2^{(t_0-1)}\mathbf{W} - \mathbf{U}_2\|_{2,\infty} \\ &\quad + \|\mathbf{U}_2\|_{2,\infty}\|\widehat{\mathbf{U}}_2^{(t_0-1)}\mathbf{W} - \mathbf{U}_2\|.\end{aligned}$$

By taking the infimum over $\mathbb{O}(r_2)$, we note that by Proposition 1 of Cai and Zhang (2018)

$$\begin{aligned}\inf_{\mathbf{W}\in\mathbb{O}(r_2)}\|\widehat{\mathbf{U}}_2^{(t_0-1)}\mathbf{W} - \mathbf{U}_2\| &\leq \sqrt{2}\|\sin\Theta(\widehat{\mathbf{U}}_2^{(t_0-1)}, \mathbf{U}_2)\| \\ &\leq \sqrt{2}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right),\end{aligned}$$

where the final inequality is on the event $\mathcal{E}_{\mathrm{Good}}$ since $t_0 \leq t_{\max}$. We also note that on the event $\mathcal{E}_{\mathrm{main}}^{t_0-1,1}$ and the right-invariance of $\|\cdot\|_{2,\infty}$ to orthogonal matrices,

$$\begin{aligned}\inf_{\mathbf{W}\in\mathbb{O}(r_2)}\|\widehat{\mathbf{U}}_2^{(t_0-1)}\mathbf{W} - \mathbf{U}_2\|_{2,\infty} &\leq \inf_{\mathbf{W}\in\mathbb{O}(r_2)}\|\widehat{\mathbf{U}}_2^{(t_0-1)} - \mathbf{U}_2\mathbf{W}\|_{2,\infty} \\ &\leq \|\widehat{\mathbf{U}}_2^{(t_0-1)} - \mathbf{U}_2\mathbf{W}_2^{(t_0-1)}\|_{2,\infty} \\ &\leq \mu_0\sqrt{\frac{r_2}{p_2}}\left(\frac{\delta_{\mathrm{L}}^{((2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right).\end{aligned}$$

Therefore,

$$\|\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} - \mathcal{P}_{\mathbf{U}_2}\|_{2,\infty} \leq 3\mu_0\sqrt{\frac{r_2}{p_2}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right).$$

Similarly,

$$\|\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} - \mathcal{P}_{\mathbf{U}_3}\|_{2,\infty} \leq 3\mu_0\sqrt{\frac{r_3}{p_3}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right).$$

Therefore,

$$\begin{aligned}
(IV) &\leq \left\| \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}} - \mathcal{P}_{\mathbf{U}_2} \right\|_{2,\infty} \|\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}}\|_{2,\infty}; \\
&\leq 6\mu_0^2 \sqrt{\frac{r_3}{p_3}} \sqrt{\frac{r_2}{p_2}} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right); \\
(V) &\leq \left\| \mathcal{P}_{\mathbf{U}_2} \otimes \left[\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}} - \mathcal{P}_{\mathbf{U}_3} \right] \right\|_{2,\infty} \\
&\leq 3\mu_0^2 \sqrt{\frac{r_2}{p_2}} \sqrt{\frac{r_3}{p_3}} \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right).
\end{aligned}$$

Finally,

$$\begin{aligned}
(VI) &:= \left\| \mathcal{P}_{\mathbf{U}_2} \otimes \mathcal{P}_{\mathbf{U}_3} \mathbf{V}_1 \right\|_{2,\infty} \\
&= \|\mathbf{V}_1\|_{2,\infty} \\
&\leq \mu_0 \sqrt{\frac{r_1}{p_{-1}}}.
\end{aligned}$$

Plugging all of these bounds in we obtain

$$\begin{aligned}
\|\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} \otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} \mathbf{V}_1\|_{2,\infty} \leq{}& \mu_0^2 \frac{\sqrt{r_2 r_3}}{p_j} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right); \\
&+ \mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_j p_3}} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) + \mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_j p_2}} \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \\
&+ 6\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_2 p_3}} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) + 3\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_2 p_3}} \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \\
&+ \mu_0 \sqrt{\frac{r_1}{p_{-1}}}.
\end{aligned}$$

This shows that the first part of the event in $\widetilde{\mathcal{E}}_{j-m}^{t_0,1}$ must hold. For the second part of the

event, we note that

$$\begin{aligned}
\|\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}\otimes\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}}\|_{2,\infty} &\leq \|\left(\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}}\right)\otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}}\|_{2,\infty} \\
&\quad + \|\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}}\otimes \mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}}\|_{2,\infty} \\
&\leq \|\left(\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}}\right)\otimes \left(\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}}\right)\|_{2,\infty} \\
&\quad + \|\left(\mathcal{P}_{\widetilde{\mathbf{U}}_2^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}}\right)\otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}}\|_{2,\infty} \\
&\quad + \|\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}}\otimes \left(\mathcal{P}_{\widetilde{\mathbf{U}}_3^{(t_0-1,j-m)}} - \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}}\right)\|_{2,\infty} \\
&\quad + \|\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}}\otimes \mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}}\|_{2,\infty} \\
&\leq \mu_0^2\frac{\sqrt{r_2r_3}}{p_j}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right) \\
&\quad + 2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right) + 2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right) \\
&\quad + 2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}.
\end{aligned}$$

where we used the fact that $\|\widehat{\mathbf{U}}_3^{(t_0-1)}\|_{2,\infty} \leq 2\mu_0\sqrt{\frac{r_3}{p_3}}$ on the events in question, and similarly for $\widehat{\mathbf{U}}_2^{(t_0-1)}$. This shows the second part of the event must hold, which completes the proof.

□

Lemma 39 (Proximity of the Leave-one-out Sequence on a good event)**.** *Let* $1 \leq j \leq 3$ *and* $1 \leq m \leq p_j$ *be fixed. Then*

$$\mathbb{P}\left\{\left\{\|\sin\Theta(\widehat{\mathbf{U}}_k^{t_0}, \widetilde{\mathbf{U}}_k^{t_0,j-m})\| \geq \left(\frac{\delta_{\mathrm{L}}^{(k)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_k}{p_j}}\right\}\bigcap\mathcal{E}_{\mathrm{Good}}\bigcap\mathcal{E}_{\mathrm{main}}^{t_0-1,k}\right\} \leq p^{-29}.$$

Proof. On the event $\mathcal{E}_{\mathrm{Good}}$ it holds that $\tau_k \leq C\sqrt{pr} \ll \lambda$ by assumption, since $r \leq Cp_{\min}^{1/2}$ and $\lambda \gtrsim \kappa p/p_{\min}^{1/4}\sqrt{\log(p)}$. Therefore, the eigengap assumption in Lemma 37 is met, so on the event $\mathcal{E}_{\mathrm{Good}}$ it holds that

$$\begin{aligned}
\|\sin\Theta(\widehat{\mathbf{U}}_k^{(t)}, \widetilde{\mathbf{U}}_k^{t,j-m})\| &\leq \frac{16\kappa}{\lambda}\tau_k\left(\eta_k^{(t_0,j-m)}\right) + \frac{16\kappa}{\lambda}\xi_k^{t_0,j-m}\left(\eta_k^{(t_0,j-m)}\right) \\
&\quad + \frac{8\kappa}{\lambda}\widetilde{\xi}_k^{t_0,j-m} + \frac{16}{\lambda^2}\tau_k^2\left(\eta_k^{(t_0,j-m)}\right) + \frac{8}{\lambda^2}\tau_k\xi_k^{t_0,j-m} + \frac{4}{\lambda^2}(\xi_k^{t_0,j-m})^2,
\end{aligned}$$

where we recall the notation

$$\eta_k^{(t,j-m)} \coloneqq \begin{cases} \left\|\sin\Theta\left(\widetilde{\mathbf{U}}_{k+1}^{t-1,j-m}, \widehat{\mathbf{U}}_{k+1}^{(t-1)}\right)\right\| + \left\|\sin\Theta\left(\widetilde{\mathbf{U}}_{k+2}^{t-1,j-m}, \widehat{\mathbf{U}}_{k+2}^{(t-1)}\right)\right\| & k=1; \\ \left\|\sin\Theta\left(\widetilde{\mathbf{U}}_{k+1}^{t-1,j-m}, \widehat{\mathbf{U}}_{k+1}^{(t-1)}\right)\right\| + \left\|\sin\Theta\left(\widetilde{\mathbf{U}}_{k+2}^{t,j-m}, \widehat{\mathbf{U}}_{k+2}^{(t)}\right)\right\| & k=2 \\ \left\|\sin\Theta\left(\widetilde{\mathbf{U}}_{k+1}^{t,j-m}, \widehat{\mathbf{U}}_{k+1}^{(t)}\right)\right\| + \left\|\sin\Theta\left(\widetilde{\mathbf{U}}_{k+2}^{t,j-m}, \widehat{\mathbf{U}}_{k+2}^{(t)}\right)\right\| & k=3; \end{cases};$$

$$\xi_k^{(t,j-m)} \coloneqq \left\|\left(\mathbf{Z}_k^{j-m} - \mathbf{Z}_k\right)\widetilde{\mathcal{P}}_k^{t,j-m}\right\|;$$

$$\widetilde{\xi}_k^{(t,j-m)} \coloneqq \left\|\left(\mathbf{Z}_k^{j-m} - \mathbf{Z}_k\right)\widetilde{\mathcal{P}}_k^{t,j-m}\mathbf{V}_k\right\|.$$

Similar to Lemma 38 we now complete the proof for $k=1$ without loss of generality (if $k=2$ or 3, the proof is similar since slightly stronger bounds hold, but this again is not needed for our analysis). On the event $\mathcal{E}_{\text{Good}} \cap \mathcal{E}_{\text{main}}^{t_0-1,1}$, we have the additional bounds

$$\begin{aligned} \tau_1 &\le C_1\sqrt{pr}; \\ \eta_1^{(t_0,j-m)} &\le \mu_0\sqrt{\frac{r_2}{p_j}}\left(\frac{\delta_{\text{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + \mu_0\sqrt{\frac{r_3}{p_j}}\left(\frac{\delta_{\text{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right). \end{aligned}$$

Plugging this in to the deterministic bound for $\|\sin\Theta(\widehat{\mathbf{U}}_k^{(t)}, \widetilde{\mathbf{U}}_k^{t,j-m})\|$ above yields

$$\begin{aligned} \|\sin\Theta(\widehat{\mathbf{U}}_1^{(t)}, \widetilde{\mathbf{U}}_1^{t,j-m})\| \le{}& \frac{16C_1\kappa}{\lambda}\sqrt{pr}\left(\mu_0\sqrt{\frac{r_2}{p_j}}\left(\frac{\delta_{\text{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + \mu_0\sqrt{\frac{r_3}{p_j}}\left(\frac{\delta_{\text{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\right) \\ &+ \frac{16\kappa}{\lambda}\xi_k^{t_0,j-m}\left(\mu_0\sqrt{\frac{r_2}{p_j}}\left(\frac{\delta_{\text{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + \mu_0\sqrt{\frac{r_3}{p_j}}\left(\frac{\delta_{\text{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\right) \\ &+ \frac{8\kappa}{\lambda}\widetilde{\xi}_k^{t_0,j-m} + \frac{16C_1^2pr}{\lambda^2}\left(\mu_0\sqrt{\frac{r_2}{p_j}}\left(\frac{\delta_{\text{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + \mu_0\sqrt{\frac{r_3}{p_j}}\left(\frac{\delta_{\text{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\right) \\ &+ \frac{8C_1\sqrt{pr}}{\lambda^2}\xi_k^{t_0,j-m} + \frac{4}{\lambda^2}(\xi_k^{t_0,j-m})^2. \end{aligned}$$

Observe that

$$\begin{aligned}
&\frac{16C_1\kappa}{\lambda}\sqrt{pr}\bigg(\mu_0\sqrt{\frac{r_2}{p_j}}\bigg(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\bigg)+\mu_0\sqrt{\frac{r_3}{p_j}}\bigg(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\bigg)\bigg)\\
&\quad=\mu_0\sqrt{\frac{r_2}{p_j}}\frac{16C_1\kappa\sqrt{pr}}{\lambda}\frac{C_0\kappa\sqrt{p_2\log(p)}}{\lambda}+\mu_0\sqrt{\frac{r_3}{p_j}}\frac{16C_1\kappa\sqrt{pr}}{\lambda}\frac{C_0\kappa\sqrt{p_3\log(p)}}{\lambda}\\
&\qquad+\frac{16C_1\kappa\sqrt{pr}}{\lambda}\bigg(\mu_0\sqrt{\frac{r_3}{p_j}}+\mu_0\sqrt{\frac{r_2}{p_j}}\bigg)\frac{1}{2^{t_0-1}}\\
&\quad=\mu_0\sqrt{\frac{r_1}{p_j}}\frac{C_0\kappa\sqrt{p_1\log(p)}}{\lambda}\bigg(\frac{16C_1\kappa\sqrt{pr}\sqrt{\frac{p_2}{p_1}}\sqrt{\frac{r_2}{r_1}}}{\lambda}+\frac{16C_1\kappa\sqrt{pr}\sqrt{\frac{p_3}{p_1}}\sqrt{\frac{r_3}{r_1}}}{\lambda}\bigg)\\
&\qquad+\mu_0\sqrt{\frac{r_1}{p_j}}\frac{1}{2^{t_0-1}}\bigg(\frac{16C_1\kappa\sqrt{pr}\sqrt{\frac{r_3}{r_1}}}{\lambda}+\frac{16C_1\kappa\sqrt{pr}\sqrt{\frac{r_2}{r_1}}}{\lambda}\bigg)\\
&\quad\le\mu_0\sqrt{\frac{r_1}{p_j}}\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}\bigg(\frac{32C_1\kappa\sqrt{pr}\sqrt{\frac{p}{p_{\min}}}}{\lambda}\max\bigg\{\sqrt{\frac{r_3}{r_1}},\sqrt{\frac{r_2}{r_1}}\bigg\}\bigg)\\
&\qquad+\mu_0\sqrt{\frac{r_1}{p_j}}\frac{1}{2^{t_0-1}}\bigg(\frac{32C_1\kappa\sqrt{pr}}{\lambda}\max\bigg\{\sqrt{\frac{r_3}{r_1}},\sqrt{\frac{r_2}{r_1}}\bigg\}\bigg)\\
&\quad\le\mu_0\sqrt{\frac{r_1}{p_j}}\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}\bigg(\frac{32C_1C_2\kappa p/p_{\min}^{1/4}}{\lambda}\bigg\}\bigg)+\mu_0\sqrt{\frac{r_1}{p_j}}\frac{1}{2^{t_0-1}}\bigg(\frac{32C_1C_2\kappa\sqrt{pr}}{\lambda}\bigg)\\
&\quad\le\frac{1}{8}\mu_0\sqrt{\frac{r_1}{p_j}}\bigg(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\bigg),
\end{aligned}$$

which holds under the assumption $\lambda \gtrsim \kappa\sqrt{\log(p)}p/p_{\min}^{1/4}$, $r_k/r_j \le C$, and $\mu_0^2 r \lesssim p_{\min}^{1/2}$. By a similar argument,

$$\frac{16C_1^2pr}{\lambda^2}\bigg(\mu_0\sqrt{\frac{r_2}{p_j}}\bigg(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\bigg)\le\frac{1}{8}\mu_0\sqrt{\frac{r_1}{p_j}}\bigg(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\bigg).$$

Therefore,

$$\begin{aligned}
\|\sin\Theta(\widehat{\mathbf{U}}_1^{(t)},\widetilde{\mathbf{U}}_1^{t,j-m})\|\le{}&\frac{1}{8}\mu_0\sqrt{\frac{r_1}{p_j}}\bigg(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\bigg)+\frac{8\kappa}{\lambda}\widetilde{\xi}_1^{t_0,j-m}\\
&+\xi_1^{t_0,j-m}\frac{16\kappa}{\lambda}\bigg(\mu_0\sqrt{\frac{r_2}{p_j}}\bigg(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\bigg)+\mu_0\sqrt{\frac{r_3}{p_j}}\bigg(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\bigg)\bigg)\\
&+\frac{8C_1\sqrt{pr}}{\lambda^2}\xi_1^{t_0,j-m}+\frac{4}{\lambda^2}(\xi_1^{t_0,j-m})^2.
\end{aligned}$$

The bound above depends only on $\xi_1^{t_0,j-m}$ and $\widetilde{\xi}_1^{t_0,j-m}$. Define

$$(I) \coloneqq \xi_1^{t_0,j-m}\left\{\frac{16\kappa}{\lambda}\left(\mu_0\sqrt{\frac{r_2}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0\sqrt{\frac{r_3}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\right)+\frac{8C_1\sqrt{pr}}{\lambda^2}\right\};$$
$$(II) \coloneqq \frac{4}{\lambda^2}(\xi_1^{t_0,j-m})^2;$$
$$(III) \coloneqq \frac{8\kappa}{\lambda}\widetilde{\xi}_1^{t_0,j-m}.$$

Then

$$\begin{aligned}
\mathbb{P}\left\{\left\{\|\sin\Theta(\widehat{\mathbf{U}}_1^{t_0},\widetilde{\mathbf{U}}_1^{t_0,j-m})\| \geq \left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_j}}\right\}\bigcap\mathcal{E}_{\mathrm{Good}}\bigcap\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\right\} \\
\leq \mathbb{P}\left\{\left\{(I)\geq\frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_j}}\right\}\bigcap\mathcal{E}_{\mathrm{Good}}\bigcap\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\right\} \\
+\mathbb{P}\left\{\left\{(II)\geq\frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_j}}\right\}\bigcap\mathcal{E}_{\mathrm{Good}}\bigcap\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\right\} \\
+\mathbb{P}\left\{\left\{(III)\geq\frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_j}}\right\}\bigcap\mathcal{E}_{\mathrm{Good}}\bigcap\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\right\}.
\end{aligned}$$

We now will derive probabilistic bounds for each of the terms above on the event $\mathcal{E}_{\mathrm{Good}}\cap\mathcal{E}_{\mathrm{main}}^{t_0-1,1}$. We will consider each term separately, though the strategy for each will remain the same: since there is nontrivial dependence between the events above and the random variable ξ_1^{j-m}, we use the auxiliary event $\widetilde{\mathcal{E}}_{j-m}^{t_0,1}$, which is independent of the nonzero entries in the random matrix $\mathbf{Z}_1^{j-m}-\mathbf{Z}_1$. We then use Lemma 38 to show that the intersection of this event with other events is empty.

The term (I): We note that

$$\begin{aligned}
&\mathbb{P}\Bigg\{\Bigg\{(I) \geq \frac{1}{4}\Bigg(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\Bigg)\mu_0\sqrt{\frac{r_1}{p_j}}\Bigg\}\bigcap \mathcal{E}_{\mathrm{Good}}\bigcap \mathcal{E}_{\mathrm{main}}^{t_0-1,1}\Bigg\} \\
&\quad\leq \mathbb{P}\Bigg\{\Bigg\{(I) \geq \frac{1}{4}\Bigg(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\Bigg)\mu_0\sqrt{\frac{r_1}{p_j}}\Bigg\}\bigcap \mathcal{E}_{\mathrm{Good}}\bigcap \mathcal{E}_{\mathrm{main}}^{t_0-1,1}\bigcap \widetilde{\mathcal{E}}_{j-m}^{t_0,1}\Bigg\} \\
&\qquad + \mathbb{P}\Bigg\{\mathcal{E}_{\mathrm{Good}}\bigcap \mathcal{E}_{\mathrm{main}}^{t_0-1,1}\bigcap (\widetilde{\mathcal{E}}_{j-m}^{t_0,1})^c\Bigg\} \\
&\quad\leq \mathbb{P}\Bigg\{\Bigg\{(I) \geq \frac{1}{4}\Bigg(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\Bigg)\mu_0\sqrt{\frac{r_1}{p_j}}\Bigg\}\bigcap \widetilde{\mathcal{E}}_{j-m}^{t_0,1}\Bigg\},
\end{aligned}$$

where we have used Lemma 38 to show that the intesection of the complement $\widetilde{\mathcal{E}}_{j-m}^{t_0,1}$ with the other events is zero.

Now we simply observe that $\widetilde{\mathcal{E}}_{j-m}^{t_0,k}$ does not depend on any of the random variables in the matrix $\mathbf{Z}_k^{j-m} - \mathbf{Z}_k$, so we are free to condition on this event. Recall that

$$\xi_k^{t_0,j-m} = \Bigg\|\Bigg(\mathbf{Z}_k - \mathbf{Z}_k^{j-m}\Bigg)\widetilde{\mathcal{P}}_1^{t_0,j-m}\Bigg\|.$$

By Lemma 47, it holds that

$$\xi_1^{t_0,j-m} \leq C\sqrt{p_{-j}\log(p)}\Bigg\|\widetilde{\mathcal{P}}_1^{t_0,j-m}\Bigg\|_{2,\infty}$$

with probability at least $1 - O(p^{-30})$. On the event $\widetilde{\mathcal{E}}_{j-m}^{t_0,k}$ we have that

$$\begin{aligned}
\Bigg\|\widetilde{\mathcal{P}}_1^{t_0,j-m}\Bigg\|_{2,\infty} &\leq \mu_0^2\frac{\sqrt{r_2r_3}}{p_j}\Bigg(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\Bigg)\Bigg(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\Bigg) \\
&\quad + 2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3}}\Bigg(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\Bigg) + 2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2}}\Bigg(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\Bigg) \\
&\quad + 2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}.
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&\xi_1^{t_0,j-m} \\
&\le C'\sqrt{p_{-j}\log(p)}\left(\frac{\mu_0^2\sqrt{r_2r_3}}{p_j}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\mu_0^2\sqrt{r_2r_3}}{p_j}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\right) \\
&\quad + C'\sqrt{p_{-j}\log(p)}\left(\mu_0^2\frac{\sqrt{r_2r_3}}{p_j}\frac{1}{2^{t_0-1}}\frac{1}{2^{t_0-1}}+2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3}}\frac{1}{2^{t_0-1}}+2\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2}}\frac{1}{2^{t_0-1}}\right) \\
&\le C''\mu_0\sqrt{\frac{r_1}{p_j}}\sqrt{p_1\log(p)}\sqrt{p_{-1}}\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right)+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\right\} \\
&\quad + C''\mu_0\sqrt{\frac{r_1}{p_j}}\sqrt{p_1\log(p)}\sqrt{p_{-1}}\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\frac{1}{2^{t_0-1}}+2\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{1}{2^{t_0-1}}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{1}{2^{t_0-1}}\right\},
\end{aligned}
$$

where we have absorbed the constants in each term. Therefore, with probability at least $1-O(p^{-30})$ it holds that

$$
\begin{aligned}
(I) &\le C''\mu_0\sqrt{\frac{r_1}{p_j}}\sqrt{p_1\log(p)}\sqrt{p_{-1}}\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right)+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\right\} \\
&\quad\times\left\{\frac{16\kappa}{\lambda}\left(\mu_0\sqrt{\frac{r_2}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0\sqrt{\frac{r_3}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\right)+\frac{8C_1\sqrt{pr}}{\lambda^2}\right\} \\
&\quad + C''\mu_0\sqrt{\frac{r_1}{p_j}}\sqrt{p_1\log(p)}\sqrt{p_{-1}}\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\frac{1}{2^{t_0-1}}+2\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{1}{2^{t_0-1}}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{1}{2^{t_0-1}}\right\} \\
&\quad\times\left\{\frac{16\kappa}{\lambda}\left(\mu_0\sqrt{\frac{r_2}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0\sqrt{\frac{r_3}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\right)+\frac{8C_1\sqrt{pr}}{\lambda^2}\right\}.
\end{aligned}
$$

We now show the first term is less than $\frac{1}{8}\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_j}}$ and the second term is less than $\frac{1}{8}\frac{1}{2^{t_0}}\mu_0\sqrt{\frac{r_1}{p_j}}$. The first term will be less than this provided that

$$
\begin{aligned}
&\frac{8}{C_0C''}\sqrt{p_{-1}}\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right)+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\right\} \\
&\qquad\times\left\{16\left(\mu_0\sqrt{\frac{r_2}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0\sqrt{\frac{r_3}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\right)+\frac{8C_1\sqrt{pr}}{\lambda\kappa}\right\}
\end{aligned}
$$

is less than one. This follows from basic algebra and the assumptions $\lambda \gtrsim \kappa\sqrt{\log(p)}p/p_{\min}^{1/4}$, that $\mu_0^2 r \lesssim p_{\min}^{1/2}$, and that $r_k \asymp r$. A similar argument shows that the second term is smaller than $\frac{1}{8}\frac{1}{2^{t_0}}\mu_0\sqrt{\frac{r_1}{p_j}}$. Therefore, on the event $\widetilde{\mathcal{E}}_{j-m}^{t_0,1}$, with probability at least $1-O(p^{-30})$ it

holds that

$$(I) \leq \frac{1}{8}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_j}}.$$

The term (II): By a similar argument, we note that

$$\begin{aligned}
&\mathbb{P}\left\{\left\{(II) \geq \frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_k}{p_j}}\right\}\bigcap \mathcal{E}_{\mathrm{Good}} \bigcap \mathcal{E}_{\mathrm{main}}^{t_0-1,1}\right\} \\
&\leq \mathbb{P}\left\{\left\{(II) \geq \frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_k}{p_j}}\right\}\bigcap \mathcal{E}_{\mathrm{Good}} \bigcap \mathcal{E}_{\mathrm{main}}^{t_0-1,1}\bigcap \widetilde{\mathcal{E}}_{j-m}^{t_0,1}\right\} \\
&\quad + \mathbb{P}\left\{\mathcal{E}_{\mathrm{Good}} \bigcap \mathcal{E}_{\mathrm{main}}^{t_0-1,1}\bigcap (\widetilde{\mathcal{E}}_{j-m}^{t_0,1})^c\right\} \\
&\leq \mathbb{P}\left\{\left\{(II) \geq \frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_k}{p_j}}\right\}\bigcap \widetilde{\mathcal{E}}_{j-m}^{t_0,1}\right\},
\end{aligned}$$

where again we used Lemma 38. Conditioning on the event $\widetilde{\mathcal{E}}_{j-m}^{t_0,1}$, by the same argument as in Term (I), with probability at least $1 - O(p^{-30})$ one has

$$\begin{aligned}
&\xi_1^{j-m} \\
&\leq \widetilde{C}\mu_0\sqrt{\frac{r_1}{p_j}}\sqrt{p_1\log(p)}\sqrt{p_{-1}}\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right) + \mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\right\} \\
&\quad + \widetilde{C}\mu_0\sqrt{\frac{r_1}{p_j}}\sqrt{p_1\log(p)}\sqrt{p_{-1}}\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\frac{1}{2^{t_0-1}} + 2\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{1}{2^{t_0-1}} + \mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{1}{2^{t_0-1}}\right\},
\end{aligned}$$

where we have once again absorbed the constant. Therefore, with probability at least $1 -$

$O(p^{-30})$,

$$
\begin{aligned}
(II) &= \frac{4}{\lambda^2}\big(\xi_1^{j-m}\big)^2 \\
&\le \frac{4}{\lambda^2}\widetilde{C}\mu_0\sqrt{\frac{r_1}{p_j}}\sqrt{p_1\log(p)}\sqrt{p_{-1}}\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right)+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\right\} \\
&\quad+\frac{4}{\lambda^2}\widetilde{C}\mu_0\sqrt{\frac{r_1}{p_j}}\sqrt{p_1\log(p)}\sqrt{p_{-1}}\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\frac{1}{2^{t_0-1}}+2\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{1}{2^{t_0-1}}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{1}{2^{t_0-1}}\right\}, \\
&= \frac{1}{8}\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_j}}\left(\frac{4\widetilde{C}\sqrt{p_{-1}}}{C_0\lambda}\right)\times\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right)+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\right\} \\
&\quad+\frac{1}{8}\frac{1}{2^{t_0}}\mu_0\sqrt{\frac{r_1}{p_j}}\left(\frac{4\widetilde{C}\sqrt{p_{-1}}}{C_0\lambda}\right)\left\{\frac{\mu_0\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right)+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\right\} \\
&\le \frac{1}{8}\left(\frac{\delta_{\mathrm{L}}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_j}},
\end{aligned}
$$

where the final inequality holds when the additional terms are smaller than one, which holds via basic algebra as long as $C_0 \ge 4C\widetilde{C}$, $\lambda \gtrsim \kappa\sqrt{\log(p)}p/p_{\min}^{1/4}$, $r_k \asymp r$ and $\mu_0^2 r \lesssim p_{\min}^{1/2}$.

The term (III): Proceeding similarly again,

$$
\begin{aligned}
&\mathbb{P}\left\{\left\{(III) \ge \frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_j}}\right\}\bigcap\mathcal{E}_{\mathrm{Good}}\bigcap\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\right\} \\
&\le \mathbb{P}\left\{\left\{(III) \ge \frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_j}}\right\}\bigcap\mathcal{E}_{\mathrm{Good}}\bigcap\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\bigcap\widetilde{\mathcal{E}}_{j-m}^{t_0,1}\right\} \\
&\quad+\mathbb{P}\left\{\mathcal{E}_{\mathrm{Good}}\bigcap\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\bigcap(\widetilde{\mathcal{E}}_{j-m}^{t_0,1})^c\right\} \\
&\le \mathbb{P}\left\{\left\{(III) \ge \frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_j}}\right\}\bigcap\widetilde{\mathcal{E}}_{j-m}^{t_0,1}\right\},
\end{aligned}
$$

where again we used Lemma 38. Conditioning on the event $\widetilde{\mathcal{E}}_{j-m}^{t_0,1}$, by Lemma 47, with probability at least $1-O(p^{-30})$ one has

$$
\begin{aligned}
\widetilde{\xi}_1^{j-m} &= \left\|\left(\mathbf{Z}_1^{j-m}-\mathbf{Z}_1\right)\widetilde{\mathcal{P}}_1^{t_0,j-m}\mathbf{V}_1\right\| \\
&\le C\sqrt{p_{-j}\log(p)}\left\|\widetilde{\mathcal{P}}_1^{t_0,j-m}\mathbf{V}_k\right\|_{2,\infty}.
\end{aligned}
$$

On the event $\widetilde{\mathcal{E}}_{j-m}^{t_0,1}$ it holds that

$$\begin{aligned}\|\widetilde{\mathcal{P}}_k^{t_0,j-m}\mathbf{V}_1\|_{2,\infty} \leq \mu_0^2\frac{\sqrt{r_2r_3}}{p_j}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right);\\ +\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\\ +6\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+3\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\\ +\mu_0\sqrt{\frac{r_1}{p_{-1}}}.\end{aligned}$$

Therefore, with probability at least $1-O(p^{-30})$,

$$\begin{aligned}(III) &\coloneqq \frac{8\kappa}{\lambda}\widetilde{\xi}_1^{t_0,j-m}\\ &\leq \frac{8C\kappa}{\lambda}\sqrt{p_{-j}\log(p)}\\ &\quad\times\left\{\mu_0^2\frac{\sqrt{r_2r_3}}{p_j}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\right.\\ &\quad+\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\\ &\quad\left.+6\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+3\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0\sqrt{\frac{r_1}{p_{-1}}}\right\}\\ &= \frac{1}{8}\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_j}}\left(\frac{64C}{C_0}\sqrt{p_{-1}}\right)\\ &\quad\times\left\{\mu_0\frac{\sqrt{r_2r_3}}{p_j\sqrt{r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\right.\\ &\quad+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_3r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_jp_2r_1}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\\ &\quad\left.+6\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+3\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3r_1}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\sqrt{\frac{1}{p_{-1}}}\right\}\\ &\leq \frac{1}{8}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_j}},\end{aligned}$$

where the final inequality holds by basic algebra as long as $C_0 \geq 64CC'$ for some other constant C', as well as the assumptions $\lambda \gtrsim \kappa\sqrt{\log(p)}p/p_{\min}^{1/4}$, $r_k \asymp r$ and $\mu_0^2 r \lesssim p_{\min}^{1/2}$.

Consequently, we have shown that the desired bounds on the terms (I), (II), and (III)

hold with probability at most $O(p^{-30}) \leq p^{-29}$ as desired. $\square$

Lemma 40 (Bounding the linear term on a good event)**.** *Let t_0 and k be fixed, and let m be such that $1 \leq m \leq p_k$. Then*

$$\mathbb{P}\left\{\left\{\|e_m^\top \mathbf{L}_k^{t_0}\| \geq \frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(k)}}{\lambda} + \frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_k}{p_k}}\right\} \bigcap \mathcal{E}_{\mathrm{Good}} \bigcap \mathcal{E}_{\mathrm{main}}^{t_0-1,k}\right\} \leq p^{-29}.$$

Proof of Lemma 40. The proof of this is similar to the proof of Lemma 39, only using the deterministic bound in Lemma 34 instead of the deterministic bound in Lemma 37. Once again without loss of generality we prove the result for $k = 1$; the bounds for $k = 2$ and $k = 3$ are similar.

First, on the event $\mathcal{E}_{\mathrm{Good}}$, it holds that $\lambda/2 \leq \lambda_{r_1}(\widehat{\mathbf{\Lambda}}_k^{(t_0-1)})$ for $t_0 \geq 1$. By Lemma 34, it holds that

$$\|e_m^\top \mathbf{L}_k^{t_0}\| \leq \frac{8\kappa}{\lambda}\|\mathbf{U}_1\|_{2,\infty}\left(\tau_1\eta_1^{(t)} + \|\mathbf{U}_1^\top \mathbf{Z}_k \mathbf{V}_1\|\right) + \frac{8\kappa}{\lambda}\left(\tau_k \eta_k^{(t,k-m)}\right) + \frac{4\kappa}{\lambda}\widetilde{\xi}_k^{t_0,k-m}.$$

On the event $\mathcal{E}_{\mathrm{main}}^{t_0-1,k} \cap \mathcal{E}_{\mathrm{Good}}$, we have the following bounds:

$$\begin{aligned}
\tau_1 &\leq C_1\sqrt{pr};\\
\|\mathbf{U}_1^\top \mathbf{Z}_1 \mathbf{V}_1\| &\leq C_1\left(\sqrt{r} + \sqrt{\log(p)}\right);\\
\eta_1^{(t)} &\leq \frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{2}{2^{t_0-1}};\\
\eta_1^{(t,1-m)} &\leq \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\mu_0\sqrt{\frac{r_2}{p_1}} + \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\mu_0\sqrt{\frac{r_3}{p_1}}.
\end{aligned}$$

Plugging in these bounds yields

$$\begin{aligned}
\|e_m^\top \mathbf{L}_k^{t_0}\| &\leq \frac{8C\kappa}{\lambda}\mu_0\sqrt{\frac{r_1}{p_1}}\left(\sqrt{pr}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{2}{2^{t_0-1}}\right)+\sqrt{r}+\sqrt{\log(p)}\right) \\
&\quad + \frac{8C\sqrt{pr}\kappa}{\lambda}\left(\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\mu_0\sqrt{\frac{r_2}{p_1}}+\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\mu_0\sqrt{\frac{r_3}{p_1}}\right)+\frac{4\kappa}{\lambda}\widetilde{\xi}_k^{t_0,k-m} \\
&\leq \mu_0\sqrt{\frac{r_1}{p_1}}\frac{8C\kappa}{\lambda}\left[\sqrt{pr}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right)+\sqrt{r}+\sqrt{\log(p)}\right]+\mu_0\sqrt{\frac{r_1}{p_1}}\frac{8C\kappa\sqrt{pr}}{\lambda}\frac{2}{2^{t_0-1}} \\
&\quad + \mu_0\sqrt{\frac{r_1}{p_1}}\frac{8C\sqrt{pr}\kappa}{\lambda}\sqrt{\frac{r_2}{r_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0\sqrt{\frac{r_1}{p_1}}\frac{8C\sqrt{pr}\kappa}{\lambda}\sqrt{\frac{r_3}{r_1}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right) \\
&\quad + \frac{4\kappa}{\lambda}\widetilde{\xi}_k^{t_0,k-m} \\
&\leq \mu_0\sqrt{\frac{r_1}{p_1}}\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}\left(\frac{8C}{C_0\sqrt{p_1\log(p)}}\right)\left[\sqrt{pr}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right)+\sqrt{r}+\sqrt{\log(p)}+\sqrt{\frac{r_2}{r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\sqrt{\frac{r_3}{r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right] \\
&\quad + \mu_0\sqrt{\frac{r_1}{p_1}}\frac{1}{2^{t_0}}\frac{32C\kappa\sqrt{pr}}{\lambda}+\mu_0\sqrt{\frac{r_1}{p_1}}\frac{1}{2^{t_0}}\frac{32C\sqrt{pr}\kappa}{\lambda}\sqrt{\frac{r}{r_1}} \\
&\quad + \frac{4\kappa}{\lambda}\widetilde{\xi}_1^{t_0,1-m} \\
&\leq \frac{1}{8}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)+\frac{4\kappa}{\lambda}\widetilde{\xi}_1^{t_0,1-m},
\end{aligned}$$

where the final inequality holds as long as

$$\left(\frac{8C}{C_0\sqrt{p_1\log(p)}}\right)\left[\sqrt{pr}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right)+\sqrt{r}+\sqrt{\log(p)}+\sqrt{\frac{r_2}{r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\sqrt{\frac{r_3}{r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right]\leq\frac{1}{8}$$

and

$$\frac{32C\kappa\sqrt{pr}}{\lambda}+\frac{32C\sqrt{pr}\kappa}{\lambda}\sqrt{\frac{r}{r_1}}\leq\frac{1}{8}.$$

These two inequalities hold as long as C_0 is larger than some fixed constant and the assump-

tions $\lambda \gtrsim \kappa\sqrt{\log(p)}p/p_{\min}^{1/4}$, $r_k \asymp r$ and $\mu_0^2 r \lesssim p_{\min}^{1/2}$. Consequently,

$$\begin{aligned}
&\mathbb{P}\Bigg\{\Big\{\|e_m^\top \mathbf{L}_1^{t_0}\| \geq \frac{1}{4}\Big(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\Big)\mu_0\sqrt{\frac{r_1}{p_1}}\Big\}\bigcap \mathcal{E}_{\mathrm{Good}}\bigcap \mathcal{E}_{\mathrm{main}}^{t_0,1}\Bigg\} \\
&\quad\leq \mathbb{P}\Bigg\{\Big\{\frac{4\kappa}{\lambda}\widetilde{\xi}_1^{t_0,1-m} \geq \frac{1}{8}\Big(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\Big)\mu_0\sqrt{\frac{r_1}{p_1}}\Big\}\bigcap \mathcal{E}_{\mathrm{Good}}\bigcap \mathcal{E}_{\mathrm{main}}^{t_0,1}\Bigg\} \\
&\quad\leq \mathbb{P}\Bigg\{\Big\{\frac{4\kappa}{\lambda}\widetilde{\xi}_1^{t_0,1-m} \geq \frac{1}{8}\Big(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\Big)\mu_0\sqrt{\frac{r_1}{p_1}}\Big\}\bigcap \mathcal{E}_{\mathrm{Good}}\bigcap \mathcal{E}_{\mathrm{main}}^{t_0,1}\bigcap \widetilde{\mathcal{E}}_{1-m}^{t_0,1}\Bigg\} \\
&\qquad+ \mathbb{P}\Bigg\{\bigcap \mathcal{E}_{\mathrm{Good}}\bigcap \mathcal{E}_{\mathrm{main}}^{t_0,1}\bigcap (\widetilde{\mathcal{E}}_{1-m}^{t_0,1})^c\Bigg\} \\
&\quad\leq \mathbb{P}\Bigg\{\Big\{\frac{4\kappa}{\lambda}\widetilde{\xi}_1^{t_0,1-m} \geq \frac{1}{8}\Big(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\Big)\mu_0\sqrt{\frac{r_1}{p_1}}\Big\}\bigcap \widetilde{\mathcal{E}}_{1-m}^{t_0,1}\Bigg\}
\end{aligned}$$

where we have used Lemma 38 to conclude that the event in the penultimate line is empty. Therefore, it suffices to bound $\widetilde{\xi}_1^{t_0,1-m}$ on the event $\widetilde{\mathcal{E}}_{1-m}^{t_0,1}$. Since this event is independent from the random variables belonging to $e_m^\top \mathbf{Z}_1$, by Lemma 46, it holds that with probability at least $1 - O(p^{-30})$ that

$$\begin{aligned}
\widetilde{\xi}_1^{t_0,1-m} &= \Big\|\Big(\mathbf{Z}_1^{1-m} - \mathbf{Z}_1\Big)\widetilde{\mathcal{P}}_1^{t_0,1-m}\mathbf{V}_1\Big\| \\
&\leq C\sqrt{p_{-1}\log(p)}\Big\|\widetilde{\mathcal{P}}_1^{t_0,j-m}\mathbf{V}_1\Big\|_{2,\infty}.
\end{aligned}$$

On the event $\widetilde{\mathcal{E}}_{1-m}^{t_0,1}$ it holds that

$$\begin{aligned}
\|\widetilde{\mathcal{P}}_1^{t_0,1-m}\mathbf{V}_1\|_{2,\infty} \leq\ & \mu_0^2\frac{\sqrt{r_2r_3}}{p_1}\Big(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\Big)\Big(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\Big); \\
&+ \mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_1p_3}}\Big(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\Big) + \mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_1p_2}}\Big(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\Big) \\
&+ 6\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\Big(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\Big) + 3\mu_0^2\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3}}\Big(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\Big) \\
&+ \mu_0\sqrt{\frac{r_1}{p_{-1}}}.
\end{aligned}$$

Therefore with probability at least $1 - O(p^{-30})$, one has

$$\begin{aligned}
\frac{4\kappa}{\lambda}\widetilde{\xi}_1^{t_0,1-m} &\leq \frac{4C\kappa\sqrt{\log(p)}}{\lambda}\sqrt{p_{-1}} \\
&\times \Bigg\{ \mu_0^2 \frac{\sqrt{r_2 r_3}}{p_1}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right); \\
&\quad + \mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_3}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + \mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_2}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) \\
&\quad + 6\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_2 p_3}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + 3\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_2 p_3}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + \mu_0\sqrt{\frac{r_1}{p_{-1}}} \Bigg\} \\
&\leq \frac{C_0\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{4C\sqrt{p_{-1}}}{C_0}\right) \\
&\times \Bigg\{ \mu_0 \frac{\sqrt{r_2 r_3}}{\sqrt{r_1 p_1}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right); \\
&\quad + \mu_0 \frac{\sqrt{r_2 r_3}}{\sqrt{r_1 p_1 p_3}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + \mu_0 \frac{\sqrt{r_2 r_3}}{\sqrt{r_1 p_1 p_2}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) \\
&\quad + 6\mu_0 \frac{\sqrt{r_2 r_3}}{\sqrt{r_1 p_2 p_3}}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + 3\mu_0 \frac{\sqrt{r_2 r_3}}{\sqrt{r_1 p_2 p_3}}\left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + \sqrt{\frac{1}{p_{-1}}} \Bigg\} \\
&\leq \frac{1}{8}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_1}},
\end{aligned}$$

where the final inequality holds by similar algebraic manipulations as in the previous part of this proof provided that C_0 is larger than some fixed constant together with the assumptions $\lambda \gtrsim \kappa\sqrt{\log(p)}p/p_{\min}^{1/4}$, $r_k \asymp r$ and $\mu_0^2 r \lesssim p_{\min}^{1/2}$. This completes the proof. □

Lemma 41 (Bounding the quadratic term on a good event)**.** *The quadratic term satisfies*

$$\mathbb{P}\left\{ \left\{ \|e_m^\top \mathbf{Q}_k^{(t)}\| \geq \frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(k)}}{\lambda} + \frac{1}{2^t}\right)\mu_0\sqrt{\frac{r_k}{p_k}} \right\} \bigcap \mathcal{E}_{\mathrm{main}}^{t_0-1,1} \bigcap \mathcal{E}_{\mathrm{Good}} \right\} \leq p^{-29}.$$

Proof. Again without loss of generality we prove the result for $k = 1$; the bounds for $k = 2$ and 3 are similar. First, on the event $\mathcal{E}_{\mathrm{Good}}$ it holds that $\lambda/2 \leq \lambda_{r_1}(\widehat{\mathbf{\Lambda}}^{(t_0-1)})$, and hence by

Lemma 35 it holds that

$$\|e_m^\top \mathbf{Q}_1^{t_0}\| \le \frac{4}{\lambda^2}\|\mathbf{U}_1\|_{2,\infty}\left(\tau_1\eta_1^{(t)} + \left\|\mathbf{U}_1^\top \mathbf{Z}_1\left[\mathcal{P}_{\mathbf{U}_2}\otimes\mathcal{P}_{\mathbf{U}_3}\right]\right\|\right) + \frac{16}{\lambda^2}\tau_1^2\left(\eta_1^{(t,1-m)}\right)$$
$$+ \frac{4}{\lambda^2}\xi_1^{t,1-m}\left(\tau_1\|\sin\Theta(\widehat{\mathbf{U}}_1^{(t)},\mathbf{U}_1)\| + \tau_1\eta_1^{(t-1)} + \left\|\mathbf{U}_1\mathbf{U}_1^\top\mathbf{Z}_1\mathcal{P}_{\mathbf{U}_2}\otimes\mathcal{P}_{\mathbf{U}_3}\right\|\right).$$

On the event $\mathcal{E}_{\text{main}}^{t_0-1,1}\cap\mathcal{E}_{\text{Good}}$, one has the following bounds:

$$\begin{aligned}
\tau_1 &\le C\sqrt{pr};\\
\eta_1^{(t_0,1-m)} &\le \mu_0\sqrt{\frac{r_2}{p_1}}\left(\frac{\delta_{\text{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0\sqrt{\frac{r_3}{p_1}}\left(\frac{\delta_{\text{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right);\\
\eta_1^{(t_0)} &\le \frac{\delta_{\text{L}}^{(2)}}{\lambda}+\frac{\delta_{\text{L}}^{(3)}}{\lambda}+\frac{2}{2^{t_0-1}};\\
\|\sin\Theta(\widehat{\mathbf{U}}_1^{(t_0-1)},\mathbf{U}_1^{(t_0-1)})\| &\le \frac{\delta_{\text{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0-1}};\\
\|\mathbf{U}_1^\top\mathbf{Z}_1\mathcal{P}_{\mathbf{U}_2}\otimes\mathcal{P}_{\mathbf{U}_3}\| &\le C(r+\sqrt{\log(p)}).
\end{aligned}$$

Plugging these in yields

$$\begin{aligned}
\|e_m^\top\mathbf{Q}_1^{t_0}\| &\le \frac{4C}{\lambda^2}\mu_0\sqrt{\frac{r_1}{p_1}}\left[\sqrt{pr}\left(\frac{\delta_{\text{L}}^{(2)}}{\lambda}+\frac{\delta_{\text{L}}^{(3)}}{\lambda}+\frac{2}{2^{t_0-1}}\right)+r+\sqrt{\log(p)}\right]\\
&\quad+\frac{16C^2pr}{\lambda^2}\left[\mu_0\sqrt{\frac{r_2}{p_1}}\left(\frac{\delta_{\text{L}}^{(2)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)+\mu_0\sqrt{\frac{r_3}{p_1}}\left(\frac{\delta_{\text{L}}^{(3)}}{\lambda}+\frac{1}{2^{t_0-1}}\right)\right]\\
&\quad+\frac{4C}{\lambda^2}\xi_1^{t,1-m}\left[\sqrt{pr}\left(\frac{\delta_{\text{L}}^{(1)}}{\lambda}+\frac{\delta_{\text{L}}^{(2)}}{\lambda}+\frac{\delta_{\text{L}}^{(3)}}{\lambda}+\frac{3}{2^{t_0-1}}\right)+r+\sqrt{\log(p)}\right]\\
&\le \frac{C_0\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{4C}{C_0}\frac{1}{\lambda\sqrt{p_1\log(p)}}\right)\left[\sqrt{pr}\left(\frac{\delta_{\text{L}}^{(2)}}{\lambda}+\frac{\delta_{\text{L}}^{(3)}}{\lambda}\right)+r+\sqrt{\log(p)}\right]\\
&\quad+\frac{C_0\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{16C^2pr}{C_0\lambda\sqrt{p_1\log(p)}}\right)\left[\sqrt{\frac{r_2}{r_1}}\frac{\delta_{\text{L}}^{(2)}}{\lambda}+\sqrt{\frac{r_3}{r_1}}\frac{\delta_{\text{L}}^{(3)}}{\lambda}\right]\\
&\quad+\frac{1}{2^{t_0}}\mu_0\sqrt{\frac{r_1}{p_1}}\left[\frac{32C\sqrt{pr}}{\lambda}+\sqrt{\frac{r_2}{r_1}}\frac{32C^2pr}{\lambda^2}+\sqrt{\frac{r_3}{r_1}}\frac{32C^2pr}{\lambda^2}\right]\\
&\quad+\xi_1^{t_0,1-m}\frac{4C}{\lambda^2}\left[\sqrt{pr}\left(\frac{\delta_{\text{L}}^{(1)}}{\lambda}+\frac{\delta_{\text{L}}^{(2)}}{\lambda}+\frac{\delta_{\text{L}}^{(3)}}{\lambda}+\frac{3}{2^{t_0-1}}\right)+r+\sqrt{\log(p)}\right]\\
&\le \frac{1}{8}\left(\frac{\delta_{\text{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_1}}\\
&\quad+\xi_1^{t_0,1-m}\frac{4C}{\lambda^2}\left[\sqrt{pr}\left(\frac{\delta_{\text{L}}^{(1)}}{\lambda}+\frac{\delta_{\text{L}}^{(2)}}{\lambda}+\frac{\delta_{\text{L}}^{(3)}}{\lambda}+\frac{3}{2^{t_0-1}}\right)+r+\sqrt{\log(p)}\right],
\end{aligned}$$

where the final inequality holds as long as

$$\left(\frac{4C}{C_0}\frac{1}{\lambda\sqrt{p_1\log(p)}}\right)\left[\sqrt{pr}\left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right)+r+\sqrt{\log(p)}\right]+\left(\frac{16C^2pr}{C_0\lambda\sqrt{p_1\log(p)}}\right)\left[\sqrt{\frac{r_2}{r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\sqrt{\frac{r_3}{r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}\right]\leq\frac{1}{8}$$

and

$$\frac{32C\sqrt{pr}}{\lambda}+\sqrt{\frac{r_2}{r_1}}\frac{32C^2pr}{\lambda^2}+\sqrt{\frac{r_3}{r_1}}\frac{32C^2pr}{\lambda^2}\leq\frac{1}{8}.$$

Both of these conditions hold when $\lambda\gtrsim\kappa\sqrt{\log(p)}p/p_{\min}^{1/4}$, $r_k\asymp r$ and $r\leq Cp_{\min}^{1/2}$ provided the constant C_0 is larger than some fixed constant. Finally, we note that

$$\begin{aligned}
&\xi_1^{t_0,1-m}\frac{4C}{\lambda^2}\left[\sqrt{pr}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{3}{2^{t_0-1}}\right)+r+\sqrt{\log(p)}\right]\\
&\quad\leq\frac{\xi_1^{t_0,1-m}}{\lambda}\left(\frac{4C}{\lambda}\right)\left[\sqrt{pr}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+\frac{3}{2^{t_0-1}}\right)+r+\sqrt{\log(p)}\right]\\
&\quad\leq\frac{\xi_1^{t_0,1-m}}{\lambda}\left(\frac{\widetilde{C}\big(\sqrt{pr}+r+\sqrt{\log(p)}\big)}{\lambda}\right).
\end{aligned}$$

Define

$$B_1:=\frac{\xi_1^{t_0,1-m}}{\lambda}\left(\frac{\widetilde{C}\big(\sqrt{pr}+r+\sqrt{\log(p)}\big)}{\lambda}\right).$$

Then

$$\begin{aligned}
&\mathbb{P}\left\{\left\{\|e_m^\top\mathbf{Q}_1^{t_0}\|\geq\frac{1}{4}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_1}}\right\}\bigcap\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\bigcap\mathcal{E}_{\mathrm{Good}}\right\}\\
&\quad\leq\mathbb{P}\left\{\left\{B_1\geq\frac{1}{8}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_1}}\right\}\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\bigcap\mathcal{E}_{\mathrm{Good}}\right\}\\
&\quad\leq\mathbb{P}\left\{\left\{B_1\geq\frac{1}{8}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_1}}\right\}\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\bigcap\mathcal{E}_{\mathrm{Good}}\bigcap\widetilde{\mathcal{E}}_{1-m}^{t_0,1}\right\}\\
&\qquad+\mathbb{P}\left\{\mathcal{E}_{\mathrm{main}}^{t_0-1,1}\bigcap\mathcal{E}_{\mathrm{Good}}\bigcap(\widetilde{\mathcal{E}}_{1-m}^{t_0,1})^c\right\}\\
&\quad\leq\mathbb{P}\left\{\left\{B_1\geq\frac{1}{8}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}+\frac{1}{2^{t_0}}\right)\mu_0\sqrt{\frac{r_1}{p_1}}\right\}\bigcap\widetilde{\mathcal{E}}_{1-m}^{t_0,1}\right\},
\end{aligned}$$

where we have used Lemma 38 to conclude that the event in the penultimate line is empty.

Since the event $\widetilde{\mathcal{E}}_{1-m}^{t_0,1}$ is independent from the random variables belonging to $e_m^\top \mathbf{Z}_1$, by Lemma 46, it holds that with probability at least $1 - O(p^{-30})$ that

$$\begin{aligned}\xi_1^{t_0,1-m} &:= \left\| \left(\mathbf{Z}_1^{1-m} - \mathbf{Z}_1 \right) \widetilde{\mathcal{P}}_1^{t_0,1-m} \right\| \\ &\leq C\sqrt{p_{-1}\log(p)} \left\| \widetilde{\mathcal{P}}_1^{t_0,1-m} \right\|_{2,\infty}.\end{aligned}$$

On the event $\widetilde{\mathcal{E}}_{1-m}^{t_0,1}$, we have that

$$\begin{aligned}\left\| \widetilde{\mathcal{P}}_1^{t_0,1-m} \right\|_{2,\infty} &\leq \mu_0^2 \frac{\sqrt{r_2 r_3}}{p_1} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \\ &\quad + 2\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_3}} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) + 2\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_2}} \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \\ &\quad + 2\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_2 p_3}}.\end{aligned}$$

Therefore, with probability at least $1 - O(p^{-30})$, it holds that

$$\begin{aligned}B_1 &= \frac{\xi_1^{t_0,1-m}}{\lambda} \left(\frac{\widetilde{C}\left(\sqrt{pr} + r + \sqrt{\log(p)}\right)}{\lambda} \right) \\ &\leq \frac{C'\sqrt{p_{-1}\log(p)}}{\lambda} \left(\frac{\sqrt{pr} + r + \sqrt{\log(p)}}{\lambda} \right) \\ &\quad \times \left\{ \mu_0^2 \frac{\sqrt{r_2 r_3}}{p_1} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \right. \\ &\quad \left. + 2\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_3}} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) + 2\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_2}} \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) + 2\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_2 p_3}} \right\} \\ &\leq \frac{C'\sqrt{p_{-1}\log(p)}}{\lambda} \left(\frac{\sqrt{pr} + r + \sqrt{\log(p)}}{\lambda} \right) \\ &\quad \times \left\{ \mu_0^2 \frac{\sqrt{r_2 r_3}}{p_1} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) + 2\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_3}} \left(\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) \right. \\ &\quad \left. + 2\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_2}} \left(\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + \frac{1}{2^{t_0-1}} \right) + 2\mu_0^2 \frac{\sqrt{r_2 r_3}}{\sqrt{p_2 p_3}} \right\} \\ &\leq \frac{C_0\sqrt{p_1\log(p)}}{\lambda} \mu_0 \sqrt{\frac{r_1}{p_1}} \left(\sqrt{p_{-1}} \frac{C_0}{C'} \frac{\sqrt{pr}}{\lambda} \right) \left(\mu_0 \frac{\sqrt{r_2 r_3}}{p_1 \sqrt{r_1}} \frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + 2\mu_0 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_3 r_1}} \frac{\delta_{\mathrm{L}}^{(2)}}{\lambda} + 2\mu_0 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_2 r_1}} \frac{\delta_{\mathrm{L}}^{(3)}}{\lambda} + 2\mu_0 \frac{\sqrt{r_2 r_3}}{\sqrt{p_2 p_3 r_1}} \right) \\ &\quad + \frac{1}{2^{t_0}} \mu_0 \sqrt{\frac{r_1}{p_1}} \left(\frac{C''\sqrt{p_1}\sqrt{p_{-1}\log(p)}}{\lambda} \frac{\sqrt{pr}}{\lambda} \right) \left[\mu_0 \frac{\sqrt{r_2 r_3}}{p_1 \sqrt{r_1}} + \mu_0 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_3 r_1}} + \mu_0 \frac{\sqrt{r_2 r_3}}{\sqrt{p_1 p_2 r_1}} \right] \\ &\leq \frac{1}{8} \left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}} \right) \mu_0 \sqrt{\frac{r_1}{p_1}},\end{aligned}$$

where the final inequality holds as long as

$$\left(\sqrt{p_{-1}}\frac{C_0}{C'}\frac{\sqrt{pr}}{\lambda}\right)\left(\mu_0\frac{\sqrt{r_2r_3}}{p_1\sqrt{r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+2\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_1p_3r_1}}\frac{\delta_{\mathrm{L}}^{(2)}}{\lambda}+2\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_1p_2r_1}}\frac{\delta_{\mathrm{L}}^{(3)}}{\lambda}+2\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_2p_3r_1}}\right)\leq\frac{1}{8}$$

and

$$\left(\frac{C''\sqrt{p_1}\sqrt{p_{-1}\log(p)}}{\lambda}\frac{\sqrt{pr}}{\lambda}\right)\left[\mu_0\frac{\sqrt{r_2r_3}}{p_1\sqrt{r_1}}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_1p_3r_1}}+\mu_0\frac{\sqrt{r_2r_3}}{\sqrt{p_1p_2r_1}}\right]\leq\frac{1}{8},$$

both of which hold when C_0 is larger than some fixed constant and $\lambda \gtrsim \kappa\sqrt{\log(p)}p/p_{\min}^{1/4}$, $r_k \asymp r$ and $r \leq Cp_{\min}^{1/2}$. This completes the proof. □

D.1.4 Putting it all together: Proof of Theorem 11

Recall we define

$$
\begin{aligned}
\mathcal{E}_{\text{Good}} &:= \left\{ \max_k \tau_k \le C\sqrt{pr} \right\} \bigcap \left\{ \|\sin\Theta(\widehat{\mathbf{U}}_k^{(t)}, \mathbf{U}_k)\| \le \frac{\delta_{\text{L}}^{(k)}}{\lambda} + \frac{1}{2^t} \text{ for all } t \le t_{\max} \text{ and } 1 \le k \le 3 \right\} \\
&\quad \bigcap \left\{ \max_k \left\| \mathbf{U}_k^\top \mathbf{Z}_k \mathbf{V}_k \right\| \le C\left(\sqrt{r} + \sqrt{\log(p)}\right) \right\}; \\
&\quad \bigcap \left\{ \max_k \left\| \mathbf{U}_k^\top \mathbf{Z}_k \mathcal{P}_{\mathbf{U}_{k+1}} \otimes \mathcal{P}_{\mathbf{U}_{k+2}} \right\| \le C\left(r + \sqrt{\log(p)}\right) \right\}; \\
&\quad \bigcap \left\{ \max_k \left\| \mathbf{Z}_k \mathbf{V}_k \right\| \le C\sqrt{p_k} \right\}; \\
\mathcal{E}_{2,\infty}^{t,k} &:= \left\{ \|\widehat{\mathbf{U}}_k^{(t)} - \mathbf{U}_k \mathbf{W}_k^{(t)}\|_{2,\infty} \le \left(\frac{\delta_{\text{L}}^{(k)}}{\lambda} + \frac{1}{2^t}\right)\mu_0\sqrt{\frac{r_k}{p_k}} \right\}; \\
\mathcal{E}_{j-m}^{t,k} &:= \left\{ \|\sin\Theta(\widetilde{\mathbf{U}}_k^{t,j-m}, \widehat{\mathbf{U}}_k^{(t)})\| \le \left(\frac{\delta_{\text{L}}^{(k)}}{\lambda} + \frac{1}{2^t}\right)\mu_0\sqrt{\frac{r_k}{p_j}} \right\}; \\
\widetilde{\mathcal{E}}_{j-m}^{t,k} &:= \Bigg\{ \|\widetilde{\mathcal{P}}_k^{t_0,j-m}\mathbf{V}_k\|_{2,\infty} \le \mu_0^2 \frac{\sqrt{r_{-k}}}{p_j}\left(\frac{\delta_{\text{L}}^{(k+1)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\left(\frac{\delta_{\text{L}}^{(k+2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right); \\
&\qquad + \mu_0^2 \frac{\sqrt{r_{-k}}}{\sqrt{p_j p_{k+2}}}\left(\frac{\delta_{\text{L}}^{(k+1)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + \mu_0^2 \frac{\sqrt{r_{-k}}}{\sqrt{p_j p_{k+1}}}\left(\frac{\delta_{\text{L}}^{(k+2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) \\
&\qquad + 6\mu_0^2 \frac{\sqrt{r_{-k}}}{\sqrt{p_{-k}}}\left(\frac{\delta_{\text{L}}^{(k+1)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + 3\mu_0^2 \frac{\sqrt{r_{-k}}}{\sqrt{p_{-k}}}\left(\frac{\delta_{\text{L}}^{(k+2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + \mu_0\sqrt{\frac{r_k}{p_{-k}}}. \Bigg\} \\
&\quad \bigcap \Bigg\{ \|\widetilde{\mathcal{P}}_k^{t_0,j-m}\|_{2,\infty} \le \mu_0^2 \frac{\sqrt{r_{-k}}}{p_j}\left(\frac{\delta_{\text{L}}^{(k+1)}}{\lambda} + \frac{1}{2^{t_0-1}}\right)\left(\frac{\delta_{\text{L}}^{(k+2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) \\
&\qquad + 2\mu_0^2 \frac{\sqrt{r_{-k}}}{\sqrt{p_j p_{k+2}}}\left(\frac{\delta_{\text{L}}^{(k+1)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + 2\mu_0^2 \frac{\sqrt{r_{-k}}}{\sqrt{p_j p_{k+1}}}\left(\frac{\delta_{\text{L}}^{(k+2)}}{\lambda} + \frac{1}{2^{t_0-1}}\right) + 2\mu_0^2 \frac{\sqrt{r_{-k}}}{\sqrt{p_{-k}}}. \Bigg\}; \\
\mathcal{E}_{\text{main}}^{t_0-1,1} &:= \bigcap_{t=1}^{(t_0-1)} \left\{ \bigcap_{k=1}^{3} \mathcal{E}_{2,\infty}^{t,k} \cap \bigcap_{j=1}^{3} \bigcap_{m=1}^{p_j} \mathcal{E}_{k-m}^{t,j} \right\}; \\
\mathcal{E}_{\text{main}}^{t_0-1,2} &:= \mathcal{E}_{\text{main}}^{t_0-1,1} \cap \left\{ \bigcap_{k=1}^{3} \bigcap_{m=1}^{p_k} \mathcal{E}_{k-m}^{t_0,1} \right\} \cap \mathcal{E}_{2,\infty}^{t_0,1} \\
\mathcal{E}_{\text{main}}^{t_0-1,3} &:= \mathcal{E}_{\text{main}}^{t_0-1,2} \cap \left\{ \bigcap_{k=1}^{3} \bigcap_{m=1}^{p_k} \mathcal{E}_{k-m}^{t_0,2} \right\} \cap \mathcal{E}_{2,\infty}^{t_0,2}.
\end{aligned}
$$

Proof of Theorem 11. We will show that by induction that with probability at least $1 -$

$3(t_0+1)p^{-15}$ that simultaneously for all $t \leq t_0$ and each k

$$\|\widehat{\mathbf{U}}_k^{t_0} - \mathbf{U}_k \mathbf{W}_k^{t_0}\|_{2,\infty} \leq \frac{\delta_{\mathrm{L}}^{(k)}}{\lambda}\mu_0\sqrt{\frac{r_k}{p_k}} + \frac{1}{2^t}\mu_0\sqrt{\frac{r_k}{p_k}};$$
$$\max_{1\leq m\leq p_k}\max_{1\leq j\leq 3}\|\sin\Theta(\widehat{\mathbf{U}}_j^{t_0}, \widetilde{\mathbf{U}}_j^{t_0,k-m})\| \leq \frac{\delta_{\mathrm{L}}^{(k)}}{\lambda}\mu_0\sqrt{\frac{r_k}{p_j}} + \frac{1}{2^t}\mu_0\sqrt{\frac{r_k}{p_j}}.$$

Assuming that for the moment, suppose the algorithm is run for at most $C\log(\kappa p/\lambda)$ iterations. Then since $\lambda \gtrsim \kappa p\sqrt{\log(p)}/p_{\min}^{1/4}$ it holds that

$$t \geq \max\{C\log(p), 1\}$$

and hence that

$$\begin{aligned}\frac{1}{2^t} &\leq \frac{1}{2^{C\log(\kappa p/\lambda)}} \\ &\leq \frac{C_0\sqrt{p_k\log(p)}}{\lambda} \\ &\leq \frac{\delta_{\mathrm{L}}^{(k)}}{\lambda}.\end{aligned}$$

Moreover, the probability holds with at least

$$1-(t-1)p^{-15} \geq 1 - C(\log(p)-1)p^{-15} \geq 1-p^{-10}.$$

Therefore, it remains to show that the result holds by induction.

Step 1: Base Case

By Theorem 25 it holds that with probability at least $1-O(p^{-20})$ that

$$\begin{aligned}\|\widehat{\mathbf{U}}_k^S - \mathbf{U}_k\mathbf{W}_k^S\|_{2,\infty} &\lesssim \frac{\kappa\mu_0\sqrt{r_1\log(p)}}{\lambda} + \frac{\mu_0\sqrt{r_k p_{-k}}\log(p)}{\lambda^2} + \kappa^2\mu_0^2\frac{r_k}{p_k} \\ &\leq \left(\frac{C\kappa\sqrt{p_k\log(p)}}{\lambda} + \frac{1}{2}\right)\mu_0\sqrt{\frac{r_k}{p_k}},\end{aligned}$$

where the final inequality holds since $\lambda \gtrsim \kappa p\sqrt{\log(p)}p_{\min}^{1/4}$ and $\mu_0^2 r \leq C p_{\min}^{1/2}$ and that $\kappa^2 \leq c p_{\min}^{1/4}$ as long as $c \times \sqrt{C} \leq \frac{1}{4}$. In addition, by Lemma 32 we have the initial bound for each k via

$$\begin{aligned}\max_j \max_m \|\sin\Theta(\widehat{\mathbf{U}}_k^S, \widetilde{\mathbf{U}}_k^{j-m})\| &\lesssim \frac{\kappa\sqrt{p_k\log(p)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_j}} + \frac{(p_1p_2p_3)^{1/2}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r_1}{p_j}} \\ &\leq \left(\frac{C\kappa\sqrt{p_k\log(p)}}{\lambda} + \frac{1}{2}\right)\mu_0\sqrt{\frac{r_k}{p_j}},\end{aligned}$$

which holds with probability at least $1 - O(p^{-19})$. Therefore, we have established the base case, which holds with probability $1 - O(p^{-19}) \geq 1 - 3p^{-15}$, as long as C_0 in the definition of δ_{L} satisfies $C_0 \geq C$, with C as above.

Step 2: Induction Step

Suppose that for all $t \leq t_0 - 1$ it holds that with probability at least $1 - 3t_0p^{-15}$ that

$$\begin{aligned}\max_k \|\widehat{\mathbf{U}}_k^{(t)} - \mathbf{U}_k\mathbf{W}_k^{(t)}\|_{2,\infty} &\leq \frac{\delta_{\mathrm{L}}^{(k)}}{\lambda}\mu_0\sqrt{\frac{r_k}{p_k}} + \frac{1}{2^t}\mu_0\sqrt{\frac{r_k}{p_k}}; \\ \max_k \max_m \max_j \|\sin\Theta(\widehat{\mathbf{U}}_j^{(t)}, \widetilde{\mathbf{U}}_j^{t,k-m})\| &\leq \frac{\delta_{\mathrm{L}}^{(k)}}{\lambda}\mu_0\sqrt{\frac{r_k}{p_j}} + \frac{1}{2^t}\mu_0\sqrt{\frac{r_k}{p_j}}.\end{aligned}$$

Observe that the induction hypothesis is equivalent to stating that $\mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}$ holds with probability at least $1-3t_0p^{-15}$. We will now show that with probability at least $1-3t_0p^{-15}-p^{-15}$ that $\mathcal{E}_{2,\mathrm{main}}^{(t_0-1)}$ holds, which is equivalent to showing that

$$\begin{aligned}\|\widehat{\mathbf{U}}_1^{t_0} - \mathbf{U}_1\mathbf{W}_1^{t_0}\|_{2,\infty} &\leq \frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_1}} + \frac{1}{2^{t_0}}\mu_0\sqrt{\frac{r_1}{p_1}}; \\ \max_m \max_j \|\sin\Theta(\widehat{\mathbf{U}}_j^{t_0}, \widetilde{\mathbf{U}}_j^{t_0,1-m})\| &\leq \frac{\delta_{\mathrm{L}}^{(1)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_j}} + \frac{1}{2^{t_0}}\mu_0\sqrt{\frac{r_1}{p_j}}.\end{aligned}$$

In other words, we will show that all of the bounds for the first mode hold. Note that

$$\begin{aligned}\widehat{\mathbf{U}}_1^{t_0} - \mathbf{U}_1\mathbf{U}_1^\top\widehat{\mathbf{U}}_1^{t_0} &= \mathbf{U}_{1\perp}\mathbf{U}_{1\perp}^\top\mathbf{Z}_1\left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}}\right]\mathbf{Z}_1^\top\widehat{\mathbf{U}}_1^{t_0}(\widehat{\mathbf{\Lambda}}_1^{(t_0)})^{-2} \\ &\quad + \mathbf{U}_{1\perp}\mathbf{U}_{1\perp}^\top\mathbf{Z}_1\left[\mathcal{P}_{\widehat{\mathbf{U}}_2^{(t_0-1)}}\otimes\mathcal{P}_{\widehat{\mathbf{U}}_3^{(t_0-1)}}\right]\mathbf{T}_1^\top\widehat{\mathbf{U}}_1^{t_0}(\widehat{\mathbf{\Lambda}}_1^{(t_0)})^{-2} \\ &= \mathbf{Q}_1^{t_0} + \mathbf{L}_1^{t_0}.\end{aligned}$$

Therefore,

$$\begin{aligned}\widehat{\mathbf{U}}_1^{t_0} - \mathbf{U}_1\mathbf{W}_1^{t_0} &= \widehat{\mathbf{U}}_1^{t_0} - \mathbf{U}_1\mathbf{U}_1\widehat{\mathbf{U}}_1^{t_0} + \mathbf{U}_1(\mathbf{U}_1\widehat{\mathbf{U}}_1^{t_0} - \mathbf{W}_1^{t_0}) \\ &= \mathbf{Q}_1^{t_0} + \mathbf{L}_1^{t_0} + \mathbf{U}_1(\mathbf{U}_1\widehat{\mathbf{U}}_1^{t_0} - \mathbf{W}_1^{t_0}).\end{aligned}$$

Consequently,

$$e_m^\top\left(\widehat{\mathbf{U}}_1^{t_0} - \mathbf{U}_1\mathbf{W}_1^{t_0}\right) = e_m^\top\mathbf{Q}_1^{t_0} + e_m^\top\mathbf{L}_1^{t_0} + e_m^\top\mathbf{U}_1(\mathbf{U}_1\widehat{\mathbf{U}}_1^{t_0} - \mathbf{W}_1^{t_0}).$$

We now proceed by bounding probabilistically. Observe that

$$
\begin{aligned}
&\mathbb{P}\Bigg\{\|\widehat{\mathbf{U}}_1^{t_0} - \mathbf{U}_1\mathbf{W}_1^{t_0}\|_{2,\infty} \geq \mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right) \\
&\qquad \bigcup \max_m \max_j \|\sin\Theta(\widehat{\mathbf{U}}_j^{t_0}, \widetilde{\mathbf{U}}_j^{t_0,1-m}\| \geq \mu_0\sqrt{\frac{r_1}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\Bigg\} \\
&\leq \mathbb{P}\Bigg\{\|\widehat{\mathbf{U}}_1^{t_0} - \mathbf{U}_1\mathbf{W}_1^{t_0}\|_{2,\infty} \geq \mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right) \\
&\qquad \bigcup \max_m \max_j \|\sin\Theta(\widehat{\mathbf{U}}_j^{t_0}, \widetilde{\mathbf{U}}_j^{t_0,1-m}\| \geq \mu_0\sqrt{\frac{r_1}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\Bigg\} + 3t_0p^{-15} \\
&\leq \mathbb{P}\Bigg\{\|\widehat{\mathbf{U}}_1^{t_0} - \mathbf{U}_1\mathbf{W}_1^{t_0}\|_{2,\infty} \geq \mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\Bigg\} \\
&\qquad + \mathbb{P}\Bigg\{\max_m \max_j \|\sin\Theta(\widehat{\mathbf{U}}_j^{t_0}, \widetilde{\mathbf{U}}_j^{t_0,1-m}\| \geq \mu_0\sqrt{\frac{r_1}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\Bigg\} + 3t_0p^{-15} \\
&\leq \mathbb{P}\Bigg\{\|\widehat{\mathbf{U}}_1^{t_0} - \mathbf{U}_1\mathbf{W}_1^{t_0}\|_{2,\infty} \geq \mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\bigcap \mathcal{E}_{\mathrm{Good}}\Bigg\} \\
&\qquad + \mathbb{P}\Bigg\{\max_m \max_j \|\sin\Theta(\widehat{\mathbf{U}}_j^{t_0}, \widetilde{\mathbf{U}}_j^{t_0,1-m}\| \geq \mu_0\sqrt{\frac{r_1}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\bigcap \mathcal{E}_{\mathrm{Good}}\Bigg\} \\
&\qquad + O(p^{-30}) + 3t_0p^{-15} \\
&\leq p\max_m \mathbb{P}\Bigg\{\|e_m^\top(\widehat{\mathbf{U}}_1^{t_0} - \mathbf{U}_1\mathbf{W}_1^{t_0})\| \geq \mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\bigcap \mathcal{E}_{\mathrm{Good}}\Bigg\} \\
&\qquad + p\max_m \mathbb{P}\Bigg\{\max_j \|\sin\Theta(\widehat{\mathbf{U}}_j^{t_0}, \widetilde{\mathbf{U}}_j^{t_0,1-m}\| \geq \mu_0\sqrt{\frac{r_1}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\bigcap \mathcal{E}_{\mathrm{Good}}\Bigg\} \\
&\qquad + O(p^{-30}) + 3t_0p^{-15} \\
&\leq p\max_m \Bigg[\mathbb{P}\Bigg\{\|e_m^\top\mathbf{L}_1^{t_0}\| \geq \frac{1}{4}\mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\bigcap \mathcal{E}_{\mathrm{Good}}\Bigg\} \\
&\qquad + \mathbb{P}\Bigg\{\|e_m^\top\mathbf{Q}_1^{t_0}\| \geq \frac{1}{4}\mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\bigcap \mathcal{E}_{\mathrm{Good}}\Bigg\} \\
&\qquad + \mathbb{P}\Bigg\{\|e_m^\top\mathbf{U}_1(\mathbf{U}_1^\top\widehat{\mathbf{U}}_1^{t_0} - \mathbf{W}_1^{t_0})\| \geq \frac{1}{4}\mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\bigcap \mathcal{E}_{\mathrm{Good}}\Bigg\}\Bigg] \\
&\qquad + p\max_m \mathbb{P}\Bigg\{\max_j \|\sin\Theta(\widehat{\mathbf{U}}_j^{t_0}, \widetilde{\mathbf{U}}_j^{t_0,k-m}\| \geq \mu_0\sqrt{\frac{r_1}{p_j}}\left(\frac{\delta_{\mathrm{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}}\right)\bigcap \mathcal{E}_{1,\mathrm{main}}^{(t_0-1)}\bigcap \mathcal{E}_{\mathrm{Good}}\Bigg\} \\
&\qquad + O(p^{-30}) + 3t_0p^{-15} \\
&\leq 3p^{-28} + p^{-28} + O(p^{-30}) + 3t_0p^{-15} \\
&\leq (3t_0+1)p^{-15},
\end{aligned}
$$

for p sufficiently large, where the penultimate inequality holds by Lemmas 39, 40, and 41, and the fact that on $\mathcal{E}_{\text{Good}}$,

$$\|e_m^\top \mathbf{U}_1 \left(\mathbf{U}_1^\top \widehat{\mathbf{U}}_1^{t_0} - \mathbf{W}_1^{t_0} \right) \| \leq \frac{1}{4}\mu_0 \sqrt{\frac{r_1}{p_1}} \left(\frac{\delta_{\text{L}}^{(1)}}{\lambda} + \frac{1}{2^{t_0}} \right)$$

by Lemma 33 since $t_0 \leq C \log(p)$ by assumption. Therefore, we have shown that the bound holds for $k = 1$. For $k = 2$, we proceed similarly, only now on the hypothesis that $\mathcal{E}_{\text{main}}^{t_0-1,2}$ holds with probability at least $1 - 3t_0 p^{-15} - p^{-15}$. The exact same argument goes through,accumulating an additional factor of p^{-15}. Finally, for $k = 3$, we proceed again, only now assuming that $\mathcal{E}_{\text{main}}^{t_0-1,3}$ holds with probability at least $1 - 3t_0 p^{-15} - 2p^{-15}$. This accumulates a final factor of p^{-15}. Therefore, since this accumulates three factors of p^{-15}, it holds that $\mathcal{E}_{\text{main}}^{t_0,1}$ holds with probability at least $1 - 3(t_0 + 1)p^{-15}$ as desired, which completes the proof. □

D.1.5 Initialization Bounds

This section contains the proof the initialization bounds. Appendix D.1.5 contains preliminary lemmas and their proofs, Appendix D.1.5 contains the proof of Theorem 25, and Appendix D.1.5 contains the proof of Lemma 32.

Preliminary Lemmas

The following result establishes concentration inequalities for the spectral norm of the noise matrices, needed in order to establish sufficient eigengap conditions.

Lemma 42. *The following bounds hold simultaneously with probability at least* $1 - O(p^{-30})$:

1. $\|\text{diag}(\mathbf{Z}_1 \mathbf{T}_1^\top)\| \lesssim \lambda_1 \mu_0^2 r \sqrt{\frac{\log(p)}{p_1}}$;
2. $\|\Gamma(\mathbf{Z}_1 \mathbf{Z}_1^\top)\| \lesssim (p_1 p_2 p_3)^{1/2}$
3. $\|\Gamma(\mathbf{T}_1 \mathbf{Z}_1^\top)\| \lesssim \lambda_1 \sqrt{p_1}$;
4. $\|\mathbf{U}_1 \mathbf{Z}_1 \mathbf{V}_1\| \lesssim \sqrt{r}$.

Proof. Part one follows since

$$\begin{aligned}
\|\mathrm{diag}(\mathbf{Z}_1\mathbf{T}_1^\top)\| &= \max_i \left|e_i^\top \mathbf{Z}_1\mathbf{T}_1^\top e_i\right| \\
&\leq \max_i \|e_i^\top \mathbf{Z}_1\mathbf{V}_1\|\|e_i^\top \mathbf{U}_1\mathbf{\Lambda}_1\| \\
&\leq \|\mathbf{Z}_1\mathbf{V}_1\|_{2,\infty}\mu_0\sqrt{\frac{r_1}{p_1}}\lambda_1 \\
&\leq C\sqrt{p_{-1}\log(p)}\mu_0\sqrt{\frac{r_1}{p_1}}\lambda_1\|\mathbf{V}_1\|_{2,\infty} \\
&\lesssim \lambda_1\mu_0^2 r\sqrt{\frac{\log(p)}{p_1}},
\end{aligned}$$

where the final inequality holds with probability at least $1 - O(p^{-30})$ by Lemma 47.

Part two follows by a slight modification of Lemma 1 of Agterberg and Sulam (2022) (with M in the statement therein taken to be 0), where the higher probability holds by adjusting the constant in the definition of δ in the proof therein. We omit the detailed proof for brevity.

Part three follows since

$$\begin{aligned}
\|\Gamma(\mathbf{Z}_1\mathbf{T}_1^\top)\| &\leq \|\mathbf{Z}_1\mathbf{T}_1^\top\| + \|\mathrm{diag}(\mathbf{Z}_1\mathbf{T}_1^\top)\| \\
&\lesssim \|\mathbf{Z}_1\mathbf{V}_1\|\lambda_1 + \lambda_1\mu_0^2 r\sqrt{\frac{\log(p)}{p_1}} \\
&\lesssim \sqrt{p_1}\lambda_1 + \lambda_1\mu_0^2 r\sqrt{\frac{\log(p)}{p_1}} \\
&\lesssim \sqrt{p_1}\lambda_1,
\end{aligned}$$

where the penultimate inequality $\|\mathbf{Z}_1\mathbf{V}_1\| \lesssim \sqrt{p_1}$ holds by a standard $\varepsilon-$net argument, and the final inequality holds since $\mu_0^2 r \lesssim \sqrt{p_{\min}}$ by assumption.

Part four follows via a standard ε-net argument. □

We also have the following result, needed in establishing the concentration of the leave-one-out sequences.

Lemma 43. *The following bounds hold with probability at least* $1 - O(p^{-30})$*:*

1. $\|\Gamma(\mathbf{T}_1(\mathbf{Z}_1 - \mathbf{Z}_1^{1-m})^\top)\| \lesssim \lambda_1\mu_0\sqrt{r\log(p)}$;
2. $\|\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top)\| \lesssim p^{3/2}$
3. $\|\Gamma(\mathbf{T}_1(\mathbf{Z}_1^{1-m} - \mathbf{Z}_1^{1-m,1-l})^\top)\| \lesssim \mu_0\lambda_1\sqrt{r\log(p)}\sqrt{\frac{p_1}{p_{-1}}}$;
4. $\|\Gamma(\mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top - \mathbf{Z}_1^{1-m,1-l}(\mathbf{Z}_1^{1-m,1-l})^\top)\| \lesssim p$.
5. $\|\Gamma(\mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m}) - \mathbf{Z}_1^{1-m,1-l}(\mathbf{Z}_1^{1-m,1-l})^\top)\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\| \lesssim \sqrt{p_1}\log(p)\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty}$

Here $\mathbf{Z}_1^{1-m,1-l}$ *is the matrix* $\mathbf{Z}_1$ *with its* m*'th row and* l*'th column removed, and* $\widetilde{\mathbf{U}}_1^{S,1-m,1-l}$ *is matrix of leading eigenvectors obtained by initializing with the noise matrix* $\mathbf{Z}_1$ *replaced with* $\mathbf{Z}_1^{1-m,1-l}$.

Proof of Lemma 43. For part one, we observe that $\mathbf{Z}_1 - \mathbf{Z}_1^{1-m}$ is a zero matrix with only its m'th row nonzero. Therefore,

$$\begin{aligned}
\|\Gamma\big(\mathbf{Z}_1 - \mathbf{Z}_1^{1-m})\mathbf{T}_1^\top\big)\| &\le \|(\mathbf{Z}_1 - \mathbf{Z}_1^{1-m})\mathbf{T}_1^\top\| + \|\mathrm{diag}\big((\mathbf{Z}_1 - \mathbf{Z}_1^{1-m})\mathbf{T}_1^\top\big)\| \\
&= \|e_m^\top\mathbf{Z}_1\mathbf{T}_1^\top\| + \max_i |e_i^\top\big(\mathbf{Z}_1 - \mathbf{Z}_1^{1-m}\big)\mathbf{T}_1^\top e_i| \\
&\le \|\mathbf{Z}_1\mathbf{V}_1\|_{2,\infty}\lambda_1 + \|\mathbf{Z}_1\mathbf{V}_1\|_{2,\infty}\lambda_1\mu_0\sqrt{\frac{r}{p_1}} \\
&\lesssim \|\mathbf{Z}_1\mathbf{V}_1\|_{2,\infty}\lambda_1 \\
&\lesssim \lambda_1\sqrt{p_{-1}\log(p)}\|\mathbf{V}_1\|_{2,\infty} \\
&\lesssim \lambda_1\mu_0\sqrt{r\log(p)},
\end{aligned}$$

with probability at least $1 - O(p^{-30})$, where the penultimate line follows from Lemma 46.

For part 2, we first observe that $\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top\big)$ is a matrix with i,j entry equal to

$$\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top\big)_{ij} = \begin{cases} \langle e_m^\top\mathbf{Z}_1, e_j^\top\mathbf{Z}_1\rangle & i = m, j \neq m \\ \langle e_m^\top\mathbf{Z}_1, e_i^\top\mathbf{Z}_1\rangle & j = m, i \neq m \\ 0 & \text{else.} \end{cases}$$

Therefore, we can decompose this matrix via

$$\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top\big) = \mathbf{G}_{\text{row}} + \mathbf{G}_{\text{col}},$$

where $\mathbf{G}_{\text{row}}$ is the matrix whose only nonzero row is its m'th row, in which case it the m, j entry is $\langle e_m^\top \mathbf{Z}_1, e_j^\top \mathbf{Z}_1\rangle$ for $j \neq m$, and $\mathbf{G}_{\text{col}}$ is defined as the transpose of this matrix. We then observe that with high probability

$$\begin{aligned}
\|\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top\big)\| &\leq \|\mathbf{G}_{\text{row}}\| + \|\mathbf{G}_{\text{col}}\| \\
&\leq 2\|\mathbf{G}_{\text{row}}\| \\
&= 2\|e_m^\top \Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\| \\
&\leq 2\|\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\| \\
&\lesssim p^{3/2},
\end{aligned}$$

where the final inequality follows from Lemma 42.

For part three, we first observe that $\mathbf{Z}_1^{1-m} - \mathbf{Z}_1^{1-m,1-l}$ is a matrix with only its l'th column nonzero. Then we note

$$\begin{aligned}
\|\Gamma\big((\mathbf{Z}_1^{1-m} - \mathbf{Z}_1^{1-m,1-l})\mathbf{T}_1^\top\big)\| &\leq \sqrt{p_1}\|\Gamma\big((\mathbf{Z}_1^{1-m} - \mathbf{Z}_1^{1-m,1-l})\mathbf{T}_1^\top\big)\|_{2,\infty} \\
&\leq \sqrt{p_1}\Big[\|\big(\mathbf{Z}_1^{1-m} - \mathbf{Z}_1^{1-m,1-l}\big)\mathbf{T}_1^\top\|_{2,\infty} \\
&\qquad + \|\text{diag}\Big(\big(\mathbf{Z}_1^{1-m} - \mathbf{Z}_1^{1-m,1-l}\big)\mathbf{T}_1^\top\Big)\|_{2,\infty}\Big] \\
&\leq \sqrt{p_1}\Big[\max_i \|\big(\mathbf{Z}_1\big)_{il}\big(\mathbf{T}_1^\top\big)_{l\cdot}\| + \max_i |(\mathbf{Z}_1)_{il}(\mathbf{T}_1^\top)_{li}|\Big] \\
&\lesssim \sqrt{p_1}\Big[\sqrt{\log(p)}\|\mathbf{T}_1^\top\|_{2,\infty} + \sqrt{\log(p)}\|\mathbf{T}_1^\top\|_{\max}\Big] \\
&\lesssim \sqrt{p_1\log(p)}\|\mathbf{T}_1^\top\|_{2,\infty} \\
&\lesssim \sqrt{p_1\log(p)}\|\mathbf{V}_1\|_{2,\infty}\lambda_1 \\
&\lesssim \mu_0\lambda_1\sqrt{r\log(p)}\sqrt{\frac{p_1}{p_{-1}}},
\end{aligned}$$

where we used the fact that $\max_{i,l}|(\mathbf{Z}_1)_{il}| \lesssim \sqrt{\log(p)}$ with probability at least $1 - O(p^{-30})$.

We next note that $\Gamma\big(\mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top - \mathbf{Z}_1^{1-m,1-l}(\mathbf{Z}_1^{1-m,1-l})^\top\big)$ is a matrix with entries equal to $(\mathbf{Z}_1)_{il}(\mathbf{Z}_1)_{jl}$ for $i \neq j$ and $i, j \neq m$. In particular, it is a the $p_1 - 1 \times p_1 - 1$ dimensional submatrix of the matrix whose entries are simply $(\mathbf{Z}_1)_{il}(\mathbf{Z}_1)_{jl}$ for $i \neq j$. This is a sample Gram matrix, so by Lemma 1 of Agterberg and Sulam (2022), it holds that

$$\|\Gamma\big(\mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top - \mathbf{Z}_1^{1-m,1-l}(\mathbf{Z}_1^{1-m,1-l})^\top\big)\| \lesssim p$$

with probability at least $1 - O(p^{-30})$ (where as in the proof of Lemma 42 the result holds by taking $M = 0$, $d = 1$, and modifying the constant on δ in the proof of Lemma 1 of Agterberg and Sulam (2022)).

For the final term, we note that

$$\begin{aligned}
&\|\Gamma(\mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m}) - \mathbf{Z}_1^{1-m,1-l}(\mathbf{Z}_1^{1-m,1-l})^\top)\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty} \\
&= \max_a \Big\| \sum_{j\neq a} (\mathbf{Z}_1)_{al}(\mathbf{Z}_1)_{jl}\big(\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\big)_{j\cdot}\Big\| \\
&\leq \max_a |(\mathbf{Z}_1)|_{al}\Big\| \sum_{j\neq a} (\mathbf{Z}_1)_{jl}\big(\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\big)_{j\cdot}\Big\| \\
&\lesssim \sqrt{\log(p)} \max_a \Big\| \sum_{j\neq a} (\mathbf{Z}_1)_{jl}\big(\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\big)_{j\cdot}\Big\| \\
&\lesssim \sqrt{p_1}\log(p)\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty},
\end{aligned}$$

where the final inequality follows from Lemma 46 applied to $\mathbf{Z}_1^\top$. □

The following result verifies the eigengap conditions that we use repeatedly throughout the proof. We adopt similar notation to Cai et al. (2021a).

Lemma 44. *Define the matrices*

$$\begin{aligned}
\mathbf{G} &:= \Gamma\big(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{T}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{T}_1^\top + \mathbf{Z}_1\mathbf{Z}_1^\top\big); \\
\mathbf{G}^{(m)} &:= \Gamma\big(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{T}_1(\mathbf{Z}_1^{-m})^\top + \mathbf{Z}_1^{-m}\mathbf{T}_1^\top + \mathbf{Z}_1^{-m}(\mathbf{Z}_1^{-m})^\top\big); \\
\mathbf{G}^{(m,l)} &:= \Gamma\big(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{T}_1(\mathbf{Z}_1^{-m-l})^\top + \mathbf{Z}_1^{-m-l}\mathbf{T}_1^\top + \mathbf{Z}_1^{-m-l}(\mathbf{Z}_1^{-m-l})^\top\big).
\end{aligned}$$

Then on the events in Lemma 42 and Lemma 43, it holds that

$$\begin{aligned}
\lambda_r^2 - \|\mathbf{G} - \mathbf{T}_1\mathbf{T}_1^\top\| &\gtrsim \lambda^2;\\
\lambda_r(\mathbf{G}) - \lambda_{r+1}(\mathbf{G}) - \|\mathbf{G} - \mathbf{G}^{-m}\| &\gtrsim \lambda^2;\\
\lambda_r(\mathbf{G}^{-m}) - \lambda_{r+1}(\mathbf{G}^{-m}) - \|\mathbf{G}^{-m} - \mathbf{G}^{-m-l}\| &\gtrsim \lambda^2;\\
\lambda_r(\mathbf{G}) &\gtrsim \lambda^2.
\end{aligned}$$

Proof of Lemma 44. First, we note that on the event in Lemma 42,

$$\begin{aligned}
\|\mathbf{G} - \mathbf{T}_1\mathbf{T}_1^\top\| &\le \|\mathrm{diag}(\mathbf{T}_1\mathbf{T}_1^\top)\| + 2\|\Gamma(\mathbf{T}_1\mathbf{Z}_1^\top)\| + \|\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\|\\
&\lesssim \lambda_1^2\mu_0^2\frac{r}{p_1} + \lambda_1\sqrt{p_1} + p^{3/2}\\
&\ll \lambda^2;\\
\|\mathbf{G} - \mathbf{G}^{-m}\| &\le 2\|\Gamma(\mathbf{T}_1(\mathbf{Z}_1 - \mathbf{Z}_1^{1-m})^\top)\| + \|\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top\|\\
&\lesssim \lambda_1\mu_0\sqrt{r\log(p)} + p^{3/2}\\
&\ll \lambda^2;\\
\|\mathbf{G}^{-m} - \mathbf{G}^{-m-l}\| &\le \|\Gamma\big(\mathbf{T}_1(\mathbf{Z}_1^{1-m} - \mathbf{Z}_1^{1-m,1-l})^\top\big)\| + \left\|\Gamma\left(\mathbf{Z}_1^{1-m}\left(\mathbf{Z}_1^{1-m}\right)^\top - \mathbf{Z}_1^{1-m,1-l}\left(\mathbf{Z}_1^{1-m,1-l}\right)^\top\right)\right\|\\
&\lesssim \mu_0\lambda_1\sqrt{r\log(p)}\sqrt{\frac{p_1}{p_{-1}}} + p\\
&\ll \lambda^2.
\end{aligned}$$

Therefore, by Weyl's inequality,

$$\begin{aligned}
\lambda_r(\mathbf{G}) - \lambda_{r+1}(\mathbf{G}) - \|\mathbf{G} - \mathbf{T}_1\mathbf{T}_1^\top\| &\ge \lambda_r^2 - 3\|\mathbf{G} - \mathbf{T}_1\mathbf{T}_1^\top\|\\
&\gtrsim \lambda^2.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\lambda_r(\mathbf{G}^{-m}) - \lambda_{r+1}(\mathbf{G}^{-m}) - \|\mathbf{G} - \mathbf{G}^{-m}\| &\gtrsim \lambda_r(\mathbf{G}) - \lambda_{r+1}(\mathbf{G}) - 2\|\mathbf{G} - \mathbf{G}^{-m}\|\\
&\gtrsim \lambda^2.
\end{aligned}$$

Finally, we note that

$$\begin{aligned}\lambda_r(\mathbf{G}) &\geq \lambda_r^2 - \|\mathbf{G} - \mathbf{T}_1\mathbf{T}_1^\top\| \\ &\geq \lambda^2 - \|\mathbf{G} - \mathbf{T}_1\mathbf{T}_1^\top\| \\ &\gtrsim \lambda^2.\end{aligned}$$

This completes the proof. □

With these spectral norm concentration and eigengap conditions fixed, we now consider the $\ell_{2,\infty}$ analysis. The first step is to show that several terms are negligible with respect to the main bound. For the remainder of the analysis, we implicitly use the eigenvalue bounds in Lemma 44, which hold under the events in Lemma 42 and Lemma 43.

Lemma 45. *The following bounds hold with probability at least* $1 - O(p^{-30})$*:*

1. $\|\mathbf{U}_1\mathbf{U}_1^\top\mathbf{Z}_1\mathbf{T}_1^\top\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\frac{\sqrt{r}\kappa}{\lambda}$
2. $\|\mathbf{U}_1\mathbf{U}_1^\top\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\frac{(p_1p_2p_3)^{1/2}}{\lambda^2}$
3. $\|\mathbf{U}_1\mathbf{U}_1^\top\mathrm{diag}\Big(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{T}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{T}_1^\top\Big)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\Big(\kappa^2\mu_0^2\frac{r}{p_1} + \frac{\kappa\mu_0^2 r\sqrt{\log(p)}}{\lambda\sqrt{p_1}}\Big)$
4. $\|\mathrm{diag}(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{T}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{T}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\|_{2,\infty} \lesssim \kappa^2\mu_0^2\frac{r_1}{p_1} + \frac{\kappa\mu_0^2 r\sqrt{\log(p)}}{\lambda\sqrt{p_1}}$
5. $\|\mathbf{U}_1(\mathbf{W}_1^S - \mathbf{U}_1^\top\widehat{\mathbf{U}}_1^S)\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\Big(\kappa^2\mu_0^2\frac{r}{p_1} + \frac{\kappa\sqrt{p_1}}{\lambda} + \frac{(p_1p_2p_3)^{1/2}}{\lambda^2}\Big)^2$.

Moreover, all of these terms are upper bounded by the quantity

$$\Big(\frac{\kappa\sqrt{p_1\log(p)}}{\lambda} + \frac{(p_1p_2p_3)^{1/2}\log(p)}{\lambda^2} + \kappa^2\mu_0\sqrt{\frac{r}{p_1}}\Big)\mu_0\sqrt{\frac{r}{p_1}}.$$

Proof of Lemma 45. For part one, we note that

$$\begin{aligned}\|\mathbf{U}_1\mathbf{U}_1^\top\mathbf{Z}_1\mathbf{T}_1^\top\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\|_{2,\infty} &\lesssim \|\mathbf{U}_1\|_{2,\infty}\frac{\|\mathbf{U}_1^\top\mathbf{Z}_1\mathbf{V}_1\|\lambda_1}{\lambda^2} \\ &\lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\frac{\sqrt{r}\kappa}{\lambda},\end{aligned}$$

since $\|\mathbf{U}_1^\top\mathbf{Z}_1\mathbf{V}_1\| \lesssim \sqrt{r}$ by Lemma 42.

For part 2, we note

$$\begin{aligned}\|\mathbf{U}_1\mathbf{U}_1^\top\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\|_{2,\infty} &\lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\frac{\|\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\|}{\lambda^2}\\ &\lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\frac{(p_1p_2p_3)^{1/2}}{\lambda^2}.\end{aligned}$$

by Lemma 42.

For part 3,

$$\begin{aligned}\|\mathbf{U}_1\mathbf{U}_1^\top\mathrm{diag}\Big(\mathbf{T}_1\mathbf{T}_1^\top+\mathbf{T}_1\mathbf{Z}_1^\top+\mathbf{Z}_1\mathbf{T}_1^\top\Big)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\|_{2,\infty} &\lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\frac{\|\mathrm{diag}\Big(\mathbf{T}_1\mathbf{T}_1^\top+\mathbf{T}_1\mathbf{Z}_1^\top+\mathbf{Z}_1\mathbf{T}_1^\top\Big)\|}{\lambda^2}\\ &\lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\left(\frac{1}{\lambda_2}\max_i|e_i^\top\mathbf{U}_1\mathbf{\Lambda}_1^2\mathbf{U}_1^\top e_i|+\frac{2\|\mathrm{diag}(\mathbf{Z}_1\mathbf{T}_1^\top)\|}{\lambda^2}\right)\\ &\lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\left(\kappa^2\mu_0^2\frac{r}{p_1}+\frac{2\|\mathrm{diag}(\mathbf{Z}_1\mathbf{T}_1^\top)\|}{\lambda^2}\right)\\ &\lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\left(\kappa^2\mu_0^2\frac{r}{p_1}+\frac{\kappa\mu_0^2r\sqrt{\log(p)}}{\lambda\sqrt{p_1}}\right),\end{aligned}$$

by Lemma 42.

For part 4,

$$\begin{aligned}\|\mathrm{diag}(\mathbf{T}_1\mathbf{T}_1^\top+\mathbf{T}_1\mathbf{Z}_1^\top+\mathbf{Z}_1\mathbf{T}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\|_{2,\infty} &\lesssim \frac{1}{\lambda^2}\Big(\|\mathrm{diag}(\mathbf{T}_1\mathbf{T}_1^\top)\|+2\|\mathrm{diag}(\mathbf{Z}_1\mathbf{T}_1^\top)\|\Big)\\ &\lesssim \kappa^2\mu_0^2\frac{r_1}{p_1}+\frac{\|\mathrm{diag}(\mathbf{Z}_1\mathbf{T}_1^\top)\|}{\lambda^2}\\ &\lesssim \kappa^2\mu_0^2\frac{r_1}{p_1}+\frac{\kappa\mu_0^2r\sqrt{\log(p)}\|}{\lambda\sqrt{p_1}}\end{aligned}$$

by Lemma 42.

Finally, by Lemma 33,

$$\|\mathbf{U}_1(\mathbf{W}_1^S-\mathbf{U}_1^\top\widehat{\mathbf{U}}_1^S)\|_{2,\infty}\le\mu_0\sqrt{\frac{r_1}{p_1}}\|\sin\Theta(\widehat{\mathbf{U}}_1^S,\mathbf{U}_1)\|^2.$$

Note that $\mathbf{U}_1$ are the eigenvectors of $\mathbf{T}_1\mathbf{T}_1^\top$ and $\widehat{\mathbf{U}}_1^S$ are the eigenvectors of the matrix

$$\Gamma\Big(\mathbf{T}_1\mathbf{T}_1^\top+\mathbf{Z}_1\mathbf{T}_1^\top+\mathbf{T}_1\mathbf{Z}_1^\top+\mathbf{Z_1}\mathbf{Z_1}^\top\Big).$$

We note that

$$\begin{aligned}\left\|\mathbf{T}_1\mathbf{T}_1^\top - \Gamma\left(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{Z}_1\mathbf{T}_1^\top + \mathbf{T}_1\mathbf{Z}_1^\top + \mathbf{Z_1}\mathbf{Z_1}^\top\right)\right\| &\leq \|\text{diag}(\mathbf{T}_1\mathbf{T}_1^\top)\| + 2\|\Gamma(\mathbf{Z}_1\mathbf{T}_1^\top)\| + \|\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\| \\ &\lesssim \lambda_1^2\mu_0^2\frac{r}{p_1} + \lambda_1\sqrt{p_1} + p^{3/2} \\ &\ll \lambda^2,\end{aligned}$$

where we note that we used the fact that $\lambda_r \gtrsim \kappa p^{3/4}\sqrt{\log(p)}$, the fact that $\mu_0^2 r \lesssim \sqrt{p}$, and the assumption $\kappa \lesssim p^{1/4}$. Therefore, by the Davis-Kahan Theorem,

$$\begin{aligned}\|\sin\Theta(\widehat{\mathbf{U}}_1^S, \mathbf{U}_1)\| &\lesssim \frac{\lambda_1^2\mu_0^2\frac{r}{p_1} + \lambda_1\sqrt{p_1} + (p_1p_2p_3)^{1/2}}{\lambda^2} \\ &\lesssim \kappa^2\mu_0^2\frac{r}{p_1} + \frac{\kappa\sqrt{p_1}}{\lambda} + \frac{(p_1p_2p_3)^{1/2}}{\lambda^2}.\end{aligned}$$

Therefore,

$$\|\mathbf{U}_1(\mathbf{W}_1^S - \mathbf{U}_1^\top\widehat{\mathbf{U}}_1^S)\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r_1}{p_1}}\left(\kappa^2\mu_0^2\frac{r}{p_1} + \frac{\kappa\sqrt{p_1}}{\lambda} + \frac{(p_1p_2p_3)^{1/2}}{\lambda^2}\right)^2.$$

□

Proof of Theorem 25

Proof of Theorem 25. Without loss of generality, we consider $k = 1$. We simply decompose

$$\begin{aligned}\widehat{\mathbf{U}}_1^S - \mathbf{U}_1\mathbf{W}_1^S &= \mathbf{Z}_1\mathbf{T}_1^\top\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2} + \Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2} - \mathbf{U}_1\mathbf{U}_1^\top\mathbf{Z}_1\mathbf{T}_1^\top\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2} - \mathbf{U}_1\mathbf{U}_1^\top\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2} \\ &\quad + \mathbf{U}_1\mathbf{U}_1^\top\text{diag}\left(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{T}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{T}_1^\top\right)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2} \\ &\quad - \text{diag}(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{T}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{T}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2} + \mathbf{U}_1(\mathbf{W}_1^S - \mathbf{U}_1^\top\widehat{\mathbf{U}}_1^S) \\ &= (I) + (II) + (III) + (IV) + (V) + (VI),\end{aligned}$$

where

$$
\begin{aligned}
(I) &\coloneqq \mathbf{Z}_1\mathbf{T}_1^\top\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2};\\
(II) &\coloneqq \Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\\
(III) &= -\mathbf{U}_1\mathbf{U}_1^\top\mathbf{Z}_1\mathbf{T}_1^\top\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2};\\
(IV) &= -\mathbf{U}_1\mathbf{U}_1^\top\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\\
(V) &= \mathbf{U}_1\mathbf{U}_1^\top\mathrm{diag}\Big(\mathbf{T}_1\mathbf{T}_1^\top+\mathbf{T}_1\mathbf{Z}_1^\top+\mathbf{Z}_1\mathbf{T}_1^\top\Big)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\\
(VI) &= -\mathrm{diag}(\mathbf{T}_1\mathbf{T}_1^\top+\mathbf{T}_1\mathbf{Z}_1^\top+\mathbf{Z}_1\mathbf{T}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\\
(VII) &= \mathbf{U}_1(\mathbf{W}_1^S-\mathbf{U}_1^\top\widehat{\mathbf{U}}_1^S)
\end{aligned}
$$

We note that terms $(III)-(VII)$ are all of smaller order than the bound we desire by Lemma 45 (with high probability). With these bounds out of the way, we now turn our attention to terms (I) and (II). For Term (I), we simply note that by Lemma 46

$$
\begin{aligned}
\|\mathbf{Z}_1\mathbf{T}_1^\top\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}_1^{-2}\|_{2,\infty} &\lesssim \frac{\kappa\|\mathbf{Z}_1\mathbf{V}_1\|_{2,\infty}}{\lambda}\\
&\lesssim \frac{\kappa\mu_0\sqrt{r\log(p)}}{\lambda}.
\end{aligned}
$$

It remains to show that the final term is of smaller order than the bound we desire, which will require the leave-one-out sequences. Note that

$$
\begin{aligned}
\|e_m^\top\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widehat{\mathbf{U}}_1^S\widehat{\mathbf{\Lambda}}^{-2}\| &\lesssim \|e_m^\top\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widehat{\mathbf{U}}_1^S-\widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top\widehat{\mathbf{U}}_1^S\|\lambda^{-2}+\|e_m^\top\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widetilde{\mathbf{U}}_1^{S,1-m}\|\lambda^{-2}\\
&\lesssim \frac{\|\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\|}{\lambda^2}\|\widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top-\widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top\|+\|e_m^\top\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widetilde{\mathbf{U}}_1^{S,1-m}\|\lambda^{-2}\\
&\coloneqq A+B.
\end{aligned}
$$

The term B: For this term, we note that

$$e_m^\top \Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widetilde{\mathbf{U}}_1^{S,1-m} = \sum_{a\neq m}\langle \mathbf{Z}_{m\cdot}, \mathbf{Z}_{a\cdot}\rangle(\widetilde{\mathbf{U}}_1^{S,1-m})_{a\cdot}$$
$$= \sum_l \mathbf{Z}_{ml}\Big(\sum_{a\neq m}\mathbf{Z}_{al}(\widetilde{\mathbf{U}}_1^{S,1-m})_{a\cdot}\Big)$$

is a sum of p_{-1} independent random variables (over l), and hence satisfies

$$\|\sum_l \mathbf{Z}_{ml}\Big(\sum_{a\neq m}\mathbf{Z}_{al}(\widetilde{\mathbf{U}}_1^{S,1-m})_{a\cdot}\Big)\| \lesssim \sqrt{p_{-1}\log(p)}\max_l \|\sum_{a\neq m}(\mathbf{Z}_{al}\widetilde{\mathbf{U}}_1^{S,1-m})_{a\cdot}\|.$$

However $\widetilde{\mathbf{U}}_1^{S,1-m}$ is still dependent on the a'th column of $\mathbf{Z}$, so we introduce a leave-two-out estimator $\widetilde{\mathbf{U}}_1^{S,1-m,1-l}$, obtained by initializing (with diagonal deletion) with the noise matrix $\mathbf{Z}_1$ replaced with $\mathbf{Z}_1^{1-m,1-l}$. For fixed l, we observe that

$$\begin{aligned}\|\sum_{a\neq m}(\mathbf{Z}_{al}\widetilde{\mathbf{U}}_1^{S,1-m})_{a\cdot}\| &\leq \|\sum_{a\neq m}(\mathbf{Z}_{al})\big(\widetilde{\mathbf{U}}_1^{S,1-m})_{a\cdot} - \big(\widetilde{\mathbf{U}}_1^{S,1-m,1-l}(\widetilde{\mathbf{U}}_1^{S,1-m,1-l})^\top\widetilde{\mathbf{U}}_1^{S,1-m}\big)_{a\cdot}\| \\ &\quad + \|\sum_{a\neq m}(\mathbf{Z}_{al})\big(\widetilde{\mathbf{U}}_1^{S,1-m,1-l}(\widetilde{\mathbf{U}}_1^{S,1-m,1-l})^\top\widetilde{\mathbf{U}}_1^{S,1-m}\big)_{a\cdot}\| \\ &\leq \|(\mathbf{Z}^{-m})^\top\|\|\widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top - \widetilde{\mathbf{U}}_1^{S,1-m,1-l}(\widetilde{\mathbf{U}}_1^{S,1-m,1-l})^\top\| \\ &\quad + \|e_l^\top(\mathbf{Z}^{-m})^\top\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|.\end{aligned}$$

Note that by Lemma 47, it holds that

$$\|e_l^\top(\mathbf{Z}^{-m})^\top\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\| \lesssim \sqrt{p_1\log(p)}\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty}. \tag{D.9}$$

In addition, by the Davis-Kahan Theorem (using the eigengap condition in Lemma 44),

$$\begin{aligned}
&\|\widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top - \widetilde{\mathbf{U}}_1^{S,1-m,1-l}(\widetilde{\mathbf{U}}_1^{S,1-m,1-l})^\top\| \\
&\lesssim \frac{1}{\lambda^2}\Big(\|\Gamma(\mathbf{Z}_1^{1-m} - \mathbf{Z}_1^{1-m,1-l})\mathbf{T}_1^\top\| \\
&\qquad + \|\Gamma(\mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m}) - \mathbf{Z}_1^{1-m,1-l}(\mathbf{Z}_1^{1-m,1-l})^\top)\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|\Big) \\
&\lesssim \frac{1}{\lambda^2}\Big(\mu_0\lambda_1\sqrt{r\log(p)}\sqrt{\frac{p_1}{p_{-1}}} + \sqrt{p_1}\log(p)\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty}\Big), \qquad \text{(D.10)}
\end{aligned}$$

where the final inequality holds by Lemma 43. Consequently, plugging this and (D.9) into the bound for B, we obtain

$$\begin{aligned}
B &\lesssim \frac{\sqrt{p_{-1}\log(p)}}{\lambda^2}\Big\{\|\mathbf{Z}^{-m}\|\frac{1}{\lambda^2}\Big(\lambda_1\mu_0\sqrt{r\log(p)}\sqrt{\frac{p_1}{p_{-1}}} + \sqrt{p_1}\log(p)\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty}\Big) \\
&\quad + \sqrt{p_1\log(p)}\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty}\Big\}. \\
&\lesssim \frac{\sqrt{p_{-1}\log(p)}}{\lambda^2}\Big\{\frac{p}{\lambda^2}\Big(\lambda_1\mu_0\sqrt{r\log(p)}\sqrt{\frac{p_1}{p_{-1}}} + \sqrt{p_1}\log(p)\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty}\Big) \\
&\quad + \sqrt{p_1\log(p)}\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty}\Big\}, \qquad \text{(D.11)}
\end{aligned}$$

where we used the fact that $\|\mathbf{Z}^{-m}\| \leq \|\mathbf{Z}\| \lesssim p$ with high probability. The bound (D.11) can be improved so as not to depend on $\widetilde{\mathbf{U}}_1^{S,1-m,1-l}$. By the bound in (D.10), it holds that

$$\begin{aligned}
\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty} &\leq \|\widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty} \\
&\quad + \|\widetilde{\mathbf{U}}_1^{S,1-m,1-l} - \widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty} \\
&\leq \|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} + \frac{1}{\lambda^2}\Big(\mu_0\lambda_1\sqrt{r\log(p)}\sqrt{\frac{p_1}{p_{-1}}} + \sqrt{p_1}\log(p)\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty}\Big), \\
&\leq \|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} + \mu_0\sqrt{\frac{r}{p_1}} + o(1)\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty},
\end{aligned}$$

where we have implicitly observed that

$$\frac{\kappa p_1\sqrt{\log(p)}}{\lambda\sqrt{p_{-1}}}\mu_0\sqrt{\frac{r}{p_1}} \leq \mu_0\sqrt{\frac{r}{p_1}},$$

which holds since $\lambda \gtrsim \kappa\sqrt{\log(p)}p/p_{\min}^{1/4}$. By rearranging, we therefore have that

$$\|\widetilde{\mathbf{U}}_1^{S,1-m,1-l}\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r}{p_1}} + \|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty}.$$

Plugging this into (D.11), we obtain

$$\begin{aligned}
B &\lesssim \frac{\sqrt{p_{-1}\log(p)}}{\lambda^2}\Bigg\{\frac{p}{\lambda^2}\Bigg(\lambda_1\mu_0\sqrt{r\log(p)}\sqrt{\frac{p_1}{p_{-1}}} + \sqrt{p_1}\log(p)\|\Bigg[\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} + \mu_0\sqrt{\frac{r}{p_1}}\Bigg]\Bigg) \\
&\quad + \sqrt{p_1\log(p)}\Bigg[\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} + \mu_0\sqrt{\frac{r}{p_1}}\Bigg]\Bigg\} \\
&\lesssim \frac{p\sqrt{p_1 r}\kappa\mu_0\log(p)}{\lambda^3} + \frac{p\sqrt{p_1p_2p_3}\log^{3/2}(p)}{\lambda^4}\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} + \frac{p\sqrt{p_1p_2p_3}\log^{3/2}(p)}{\lambda^4}\mu_0\sqrt{\frac{r}{p_1}} \\
&\quad + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}} \\
&\asymp \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}}\Bigg(\frac{p\sqrt{p_1\log(p)}}{\lambda^2}\Bigg) + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty}\Bigg(1 + \frac{p\sqrt{\log(p)}}{\lambda^2}\Bigg) \\
&\quad + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}}\Bigg(1 + \frac{p\sqrt{\log(p)}}{\lambda^2}\Bigg) \\
&\lesssim \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}},
\end{aligned}$$

which holds whenever $\lambda^2 \gtrsim p\sqrt{p_1\log(p)}$. This bound still depends on the leave-one-out sequence, but we will obtain a bound independent of this sequence shortly upon analyzing term A.

The term A: Note that by Lemma 42 we have that $\|\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\| \lesssim (p_1p_2p_3)^{1/2}$, which yields

$$A \lesssim \frac{(p_1p_2p_3)^{1/2}}{\lambda^2}\|\widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top - \widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top\|.$$

Therefore it suffices to bound the term on the right. By the Davis-Kahan Theorem, it holds that

$$\begin{aligned}
&\|\widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top - \widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top\| \\
&\lesssim \frac{\|\Gamma((\mathbf{T}_1\mathbf{Z}_1 - \mathbf{Z}_1^{1-m})^\top + (\mathbf{Z}_1 - \mathbf{Z}_1^{1-m})\mathbf{T}_1^\top + \mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top)\widetilde{\mathbf{U}}_1^{S,1-m}\|}{\lambda^2} \\
&\lesssim \frac{1}{\lambda^2}\Big\{\|\Gamma\big((\mathbf{Z}_1 - \mathbf{Z}_1^{1-m})\mathbf{T}_1^\top\big)\| + \|\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top\big)\widetilde{\mathbf{U}}_1^{S,1-m}\|\Big\}.
\end{aligned}$$

By Lemma 43, it holds that

$$\|\Gamma\big((\mathbf{Z}_1 - \mathbf{Z}_1^{1-m})\mathbf{T}_1^\top\big)\| \lesssim \lambda_1\mu_0\sqrt{r\log(p)}, \tag{D.12}$$

so it suffices to consider the second term. Note that the matrix

$$\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top\big)$$

is rank one symmetric matrix whose (m,l) entry is simply $\langle e_m^\top\mathbf{Z}_1, e_l^\top\mathbf{Z}_1\rangle$ for $l \neq m$. Therefore, define the matrices $\mathbf{G}_{\text{col}}$ and $\mathbf{G}_{\text{row}}$ with $\mathbf{G}_{\text{col}}$ the matrix whose only nonzero entries are in the m'th column, in which case they are $\langle e_m^\top\mathbf{Z}_1, e_l\mathbf{Z}_1\rangle$ for $l \neq m$, and $\mathbf{G}_{\text{row}}$ the matrix whose only nonzero entries are in the m'th row, with entries defined similarly. Then

$$\big\|\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{1-m}(\mathbf{Z}_1^{1-m})^\top\big)\widetilde{\mathbf{U}}_1^{S,1-m}\| \leq \|\mathbf{G}_{\text{row}}\widetilde{\mathbf{U}}_1^{S,1-m}\| + \|\mathbf{G}_{\text{col}}\widetilde{\mathbf{U}}_1^{S,1-m}\|.$$

We consider each term separately. First, note that

$$\|\mathbf{G}_{\text{row}}\widetilde{\mathbf{U}}_1^{S,1-m}\| = \|e_m^\top\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widetilde{\mathbf{U}}_1^{S,1-m}\|.$$

This was already bounded en route to the analysis for term B. In fact, we already have the upper bound

$$\begin{aligned}&\|e_m^\top \Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\widetilde{\mathbf{U}}^{S,1-m}\| \\ &\quad\lesssim \kappa\sqrt{p_1\log(p)}\mu_0\sqrt{\frac{r}{p_1}} + \sqrt{p_1p_2p_3}\log(p)\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} + \sqrt{p_1p_2p_3}\log(p)\mu_0\sqrt{\frac{r}{p_1}}.\end{aligned} \tag{D.13}$$

Next, we argue similarly to the proof of Lemma 4 of Cai et al. (2021a). We have

$$\begin{aligned}\|\mathbf{G}_{\text{col}}\widetilde{\mathbf{U}}_1^{S,1-m}\| &\le \|\mathbf{G}_{\text{col}}\widetilde{\mathbf{U}}_1^{S,1-m}\|_F \\ &= \left(\sum_{j\neq m}\|\langle e_m^\top\mathbf{Z}_1, e_j^\top\mathbf{Z}_1\rangle(\widetilde{\mathbf{U}}_1^{S,1-m})_{m\cdot}\|^2\right)^{1/2} \\ &\le \left(\sum_{j\neq m}|\langle e_m^\top\mathbf{Z}_1, e_j^\top\mathbf{Z}_1\rangle|^2\|(\widetilde{\mathbf{U}}_1^{S,1-m})_{m\cdot}\|^2\right)^{1/2} \\ &\le \|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty}\|\Gamma(\mathbf{Z}_1\mathbf{Z}_1^\top)\| \\ &\lesssim (p_1p_2p_3)^{1/2}\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty}.\end{aligned}$$

Therefore,

$$\begin{aligned}&\|\widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top - \widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top\| \\ &\lesssim \frac{1}{\lambda^2}\Big\{\lambda_1\mu_0\sqrt{r\log(p)} + \kappa\sqrt{p_1\log(p)}\mu_0\sqrt{\frac{r}{p_1}} + \sqrt{p_1p_2p_3}\log(p)\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} \\ &\quad + \sqrt{p_1p_2p_3}\log(p)\mu_0\sqrt{\frac{r}{p_1}} + (p_1p_2p_3)^{1/2}\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty}\Big\} \\ &\lesssim \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty}.\end{aligned}$$

Therefore, it holds that

$$\begin{aligned}\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} &\le \|\widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} + \|\widetilde{\mathbf{U}}_1^{S,1-m} - \widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} \\ &\le \|\widehat{\mathbf{U}}_1^S\|_{2,\infty} + \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} \\ &\le \|\widehat{\mathbf{U}}_1^S\|_{2,\infty} + \mu_0\sqrt{\frac{r}{p}} + o(1)\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty},\end{aligned}$$

so by rearranging we arrive at

$$\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} \lesssim \|\widehat{\mathbf{U}}_1^S\|_{2,\infty} + \mu_0\sqrt{\frac{r}{p}}.$$

Consequently,

$$\begin{aligned}
&\|\widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top - \widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top\| \\
&\quad\lesssim \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widehat{\mathbf{U}}_1^S\|_{2,\infty} \quad \text{(D.14)}
\end{aligned}$$

Therefore, we have that

$$\begin{aligned}
A &\lesssim \frac{(p_1p_2p_3)^{1/2}}{\lambda^2}\|\widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top - \widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top\| \\
&\lesssim \frac{(p_1p_2p_3)^{1/2}}{\lambda^2}\left\{\frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widehat{\mathbf{U}}_1^S\|_{2,\infty}\right\} \\
&\ll \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widehat{\mathbf{U}}_1^S\|_{2,\infty}.
\end{aligned}$$

In addition,

$$\begin{aligned}
B &\lesssim \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widetilde{\mathbf{U}}_1^{S,1-m}\|_{2,\infty} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}} \\
&\lesssim \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widehat{\mathbf{U}}_1^S\|_{2,\infty} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}},.
\end{aligned}$$

Combining both of these with the initial bounds in Lemma 45, we arrive at

$$\|e_m^\top(\widehat{\mathbf{U}}_1^S - \mathbf{U}_1\mathbf{W}_1^S)\| \lesssim \left(\frac{\kappa\sqrt{p_1\log(p)}}{\lambda} + \frac{(p_1p_2p_3)^{1/2}\log(p)}{\lambda^2} + \kappa^2\mu_0\sqrt{\frac{r}{p_1}}\right)\mu_0\sqrt{\frac{r}{p_1}} + \frac{(p_1p_2p_3)^{1/2}}{\lambda^2}\|\widehat{\mathbf{U}}_1^S\|_{2,\infty}.$$

By taking a union bound over all the rows, we have that with probability at least $1-O(p^{-29})$ that

$$\|\widehat{\mathbf{U}}_1^S - \mathbf{U}_1\mathbf{W}_1^S\|_{2,\infty} \lesssim \left(\frac{\kappa\sqrt{p_1\log(p)}}{\lambda} + \frac{(p_1p_2p_3)^{1/2}\log(p)}{\lambda^2} + \kappa^2\mu_0\sqrt{\frac{r}{p_1}}\right)\mu_0\sqrt{\frac{r}{p_1}} + \frac{(p_1p_2p_3)^{1/2}}{\lambda^2}\|\widehat{\mathbf{U}}_1^S\|_{2,\infty}.$$

Therefore,

$$\begin{aligned}\|\widehat{\mathbf{U}}\|_{2,\infty} &\leq \|\mathbf{U}_1\|_{2,\infty} + \|\widehat{\mathbf{U}}_1^S - \mathbf{U}_1\mathbf{W}_1^S\|_{2,\infty} \\ &\leq \mu_0\sqrt{\frac{r}{p_1}} + o(1)\|\widehat{\mathbf{U}}_1^S\|_{2,\infty},\end{aligned}$$

which, by rearranging, yields

$$\|\widehat{\mathbf{U}}_1^S\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r}{p_1}}.$$

Therefore,

$$\|\widehat{\mathbf{U}}_1^S - \mathbf{U}_1\mathbf{W}_1^S\|_{2\infty} \lesssim \left(\frac{\kappa\sqrt{p_1\log(p)}}{\lambda} + \frac{(p_1p_2p_3)^{1/2}\log(p)}{\lambda^2} + \kappa^2\mu_0\sqrt{\frac{r}{p_1}}\right)\mu_0\sqrt{\frac{r}{p_1}}.$$

In addition, (D.14) together with the bound above shows that with high probability,

$$\begin{aligned}&\|\widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top - \widetilde{\mathbf{U}}_1^{S,1-m}(\widetilde{\mathbf{U}}_1^{S,1-m})^\top\| \\ &\quad\lesssim \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\|\widehat{\mathbf{U}}_1^S\|_{2,\infty} \\ &\quad\lesssim \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r}{p_1}} + \frac{\sqrt{p_1p_2p_3}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r}{p_1}}.\end{aligned}$$

Taking another union bound over m shows that this bound holds for all m with probability at least $1-O(p^{-29})$. Both of these bounds therefore hold with probability at least $1-p^{-20}$. □

Proof of Lemma 32

In this section we prove Lemma 32, which controls the remaining two leave-one-out sequences not bounded in Theorem 25.

Proof of Lemma 32. First we provide concentration guarantees, similar to Lemma 43. We will bound the following terms:

- $\|\Gamma\big(\mathbf{T}_1(\mathbf{Z}_1 - \mathbf{Z}_1^{j-m})^\top\big)\|$;
- $\|\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{j-m}(\mathbf{Z}_1^{j-m})^\top\big)\|$

- $\|\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{j-m}(\mathbf{Z}_1^{j-m})^\top\big)\widetilde{\mathbf{U}}_1^{S,j-m}\|$.

First, note that by Lemma 47, with probability at least $1 - O(p^{-30})$ it holds that

$$\begin{aligned}\|\Gamma\big((\mathbf{Z}_1 - \mathbf{Z}_1^{j-m})\mathbf{T}_1^\top\big)\| &\leq \|(\mathbf{Z}_1 - \mathbf{Z}_1^{j-m})\mathbf{T}_1^\top\| + \|\mathrm{diag}\big((\mathbf{Z}_1 - \mathbf{Z}_1^{j-m})\mathbf{T}_1^\top\big)\| \\ &\leq 2\|(\mathbf{Z}_1 - \mathbf{Z}_1^{j-m})\mathbf{T}_1^\top\| \\ &\lesssim \sqrt{p_{-j}\log(p)}\|\mathbf{T}_1^\top\|_{2,\infty} \\ &\lesssim \lambda_1\mu_0\sqrt{r_1\frac{p_1}{p_j}\log(p)}.\end{aligned}$$

Next, we consider the matrix $\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{j-m}(\mathbf{Z}_1^{j-m})^\top\big)$. First, observe that for $i \neq k$ this matrix has entries of the form

$$\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{j-m}(\mathbf{Z}_1^{j-m})^\top\big)_{ik} = \sum_{l\in\Omega}(\mathbf{Z}_1)_{il}(\mathbf{Z}_1)_{kl},$$

where Ω is the set of indices such that the l'th column of $\mathbf{Z}_1$ corresponds to elements belonging to the m'th row of $\mathbf{Z}_j$. A general formula is possible, but not needed for our purposes here; the cardinality of Ω is equal to the number of nonzero columns of $\mathbf{Z}_1 - \mathbf{Z}_1^{j-m}$, which is p_{-1-j}. Since this matrix is a sample gram matrix, by Lemma 1 of Agterberg and Sulam (2022) it holds that with probability at least $1 - O(p^{-30})$ (where as in the proof of Lemma 42 the higher probability holds by modifying the constant on δ in the proof of Lemma 1 of Agterberg and Sulam (2022)) that

$$\begin{aligned}\|\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{j-m}(\mathbf{Z}_1^{j-m})^\top\big)\| &\lesssim \sqrt{p_1} + \sqrt{p_1 p_{-1-j}} \\ &\lesssim \sqrt{p_1} + \sqrt{p_{-j}} \\ &\ll p^2/p_{\min}^{1/2}\end{aligned}$$

For the remaining term, we note that

$$\begin{aligned}\|\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{j-m}(\mathbf{Z}_1^{j-m})^\top\big)\widetilde{\mathbf{U}}_1^{S,j-m}\| &\leq \sqrt{p_1}\|\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{j-m}(\mathbf{Z}_1^{j-m})^\top\big)\widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty} \\ &= \sqrt{p_1}\max_i \|\sum_{k\neq i,k=1}^{p_1}\sum_{l\in\Omega}(\mathbf{Z}_1)_{il}(\mathbf{Z}_1)_{kl}\big(\widetilde{\mathbf{U}}_1^{S,j-m}\big)_{k\cdot}\| \\ &\lesssim \sqrt{p_1}\sqrt{p_{-1-j}\log(p)}\max_l \|\sum_{k\neq i,k=1}^{p_1}(\mathbf{Z}_1)_{kl}\big(\widetilde{\mathbf{U}}_1^{S,j-m}\big)_{k\cdot}\| \\ &\lesssim p_1\sqrt{p_{-1-j}}\log(p)\|\widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty} \\ &\lesssim \sqrt{p_1}\sqrt{p_{-j}}\log(p)\|\widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty},\end{aligned}$$

where we have implicitly used the matrix Hoeffding's inequality twice: once over the summation over l conditional on the collection $(\mathbf{Z}_1)_{kl}$ for $k \neq i$, and then again over the summation in k.

Note that $\widehat{\mathbf{U}}_1^S$ are the eigenvectors of the matrix $\Gamma\big(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{Z}_1\mathbf{Z}_1^\top + \mathbf{T}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{T}_1^\top\big)$ and $\widetilde{\mathbf{U}}_1^{S,j-m}$ are the eigenvectors of the matrix $\Gamma\big(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{Z}_1^{j-m}(\mathbf{Z}_1^{j-m})^\top + \mathbf{T}_1(\mathbf{Z}_1^{j-m})^\top + \mathbf{Z}_1^{j-m}\mathbf{T}_1^\top\big)$. Therefore, the spectral norm of the difference is upper bounded by

$$\begin{aligned}2\|\Gamma\big(\mathbf{T}_1(\mathbf{Z}_1 - \mathbf{Z}_1^{j-m})^\top\big)\| + \|\Gamma\big(\mathbf{Z}_1\mathbf{Z}_1^\top - \mathbf{Z}_1^{j-m}(\mathbf{Z}_1^{j-m})^\top\big)\| &\lesssim \lambda_1\mu_0\sqrt{r_1\frac{p_1}{p_j}\log(p)} + p^2/p_{\min}^{1/2} \\ &\ll \lambda^2.\end{aligned}$$

Moreover, Lemma 44 shows that

$$\lambda_r\Big(\Gamma\big(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{Z}_1\mathbf{Z}_1^\top + \mathbf{T}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{T}_1^\top\big)\Big) - \lambda_{r+1}\Big(\Gamma\big(\mathbf{T}_1\mathbf{T}_1^\top + \mathbf{Z}_1\mathbf{Z}_1^\top + \mathbf{T}_1\mathbf{Z}_1^\top + \mathbf{Z}_1\mathbf{T}_1^\top\big)\Big) \gtrsim \lambda^2,$$

so by the Davis-Kahan Theorem,

$$\begin{aligned}\|\widetilde{\mathbf{U}}_1^{j-m}(\widetilde{\mathbf{U}}_1^{S,j-m})^\top - \widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top\| &\lesssim \frac{1}{\lambda^2}\Big(\lambda_1\mu_0\sqrt{r_1\frac{p_1}{p_j}\log(p)} + \sqrt{p_1}\sqrt{p_{-j}}\log(p)\|\widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty}\Big) \\ &\lesssim \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_j}} + \frac{\sqrt{p_1p_{-j}}\log(p)}{\lambda^2}\|\widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty}\end{aligned} \tag{D.15}$$

In addition, we note that by Theorem 25, with probability at least $1 - O(p^{-20})$ it holds that

$$\begin{aligned}
\|\widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty} &\leq \|\widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top \widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty} + \|\widetilde{\mathbf{U}}_1^{S,j-m} - \widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top \widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty} \\
&\leq \|\widehat{\mathbf{U}}_1^S\|_{2,\infty} + \|\widetilde{\mathbf{U}}_1^{S,j-m}(\widetilde{\mathbf{U}}_1^{S,j-m})^\top - \widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top\| \\
&\leq \|\widehat{\mathbf{U}}_1^S\|_{2,\infty} + \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_1}} + \frac{(p_1p_2p_3)^{1/2}\log(p)}{\lambda^2}\|\widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty} \\
&\leq \mu_0\sqrt{\frac{r_1}{p_1}} + o(1)\|\widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty},
\end{aligned}$$

so by rearranging we obtain that

$$\|\widetilde{\mathbf{U}}_1^{S,j-m}\|_{2,\infty} \lesssim \mu_0\sqrt{\frac{r_1}{p_1}}.$$

Plugging this into (D.15) yields

$$\begin{aligned}
\|\widetilde{\mathbf{U}}_1^{j-m}(\widetilde{\mathbf{U}}_1^{S,j-m})^\top - \widehat{\mathbf{U}}_1^S(\widehat{\mathbf{U}}_1^S)^\top\| &\lesssim \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_j}} + \frac{\sqrt{p_{-j}}\log(p)}{\lambda^2}\mu_0\sqrt{r_1} \\
&\lesssim \frac{\kappa\sqrt{p_1\log(p)}}{\lambda}\mu_0\sqrt{\frac{r_1}{p_j}} + \frac{(p_1p_2p_3)^{1/2}\log(p)}{\lambda^2}\mu_0\sqrt{\frac{r_1}{p_j}}
\end{aligned}$$

with probability at least $1 - O(p^{-20})$. The proof is then completed by taking a union bound over all p_1 rows. □

D.2 Proofs of Tensor Mixed-Membership Blockmodel Identifiability and Estimation

In this section we prove our main results concerning the mixed-membership identifiability and estimation. First we establish Proposition 2 as well as Lemma 6 relating the properties of the tensor mixed-membership blockmodel to the tensor denoising model. We then prove our estimation guarantees Theorem 10. Throughout we let $\mathbf{S}_k = \mathcal{M}_k(\mathcal{S})$ and $\mathbf{T}_k$ defined similarly.

D.2.1 Proofs of Proposition 2, Proposition 3, and Lemma 6

First we prove Proposition 3 and Lemma 6 simultaneously as we will require part of the proof in the proof of Proposition 2.

Proof of Proposition 3 and Lemma 6. For the first part, we follow the proof of Lemma 2.3 of Mao et al. (2021). Without loss of generality we prove the result for mode 1. Let $\mathbf{T}_1$ have singular value decomposition $\mathbf{T}_1 = \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{V}_1^\top$ Then since $\mathbf{U}_1\mathbf{\Lambda}_1\mathbf{V}_1^\top = \mathbf{\Pi}_1\mathbf{S}_1(\mathbf{\Pi_2}\otimes\mathbf{\Pi}_3)^\top$, without loss of generality we may assume the first r_1 rows of $\mathbf{T}_1$ correspond to pure nodes. We note that therefore

$$\mathbf{U}_1^{(\text{pure})}\mathbf{\Lambda}_1^2(\mathbf{U}_1^{(\text{pure})})^\top = \mathbf{S}_1(\mathbf{\Pi}_2\otimes\mathbf{\Pi}_3)^\top(\mathbf{\Pi}_2\otimes\mathbf{\Pi}_3)\mathbf{S}_1^\top.$$

Since the rank of $\mathbf{\Pi}_2$ and $\mathbf{\Pi}_3$ are r_2 and r_3 respectively, it holds that the matrix above is rank r_1 as long as $r_1 \leq r_2r_3$ since $\mathbf{S}_1$ is rank r_1, which shows that $\mathbf{U}_1^{(\text{pure})}$ is rank r_1. Furthermore, we have that $\mathbf{T}_1[1:r_1,\cdot] = \mathbf{U}_1^{(\text{pure})}\mathbf{\Lambda}_1\mathbf{V}_1^\top = \mathbf{S}_1(\mathbf{\Pi_2}\otimes\mathbf{\Pi}_3)^\top$ which shows that $\mathbf{U}_1^{(\text{pure})} = \mathbf{S}_1(\mathbf{\Pi_2}\otimes\mathbf{\Pi}_3)^\top\mathbf{V}_1\mathbf{\Lambda}_1^{-1}$. Therefore,

$$\begin{aligned}\mathbf{U}_1 &= \mathbf{T}_1\mathbf{V}_1\mathbf{\Lambda}_1^{-1}\\ &= \mathbf{\Pi}_1\mathbf{S}_1(\mathbf{\Pi}_2\otimes\mathbf{\Pi}_3)^\top\mathbf{V}_1\mathbf{\Lambda}_1^{-1}\\ &= \mathbf{\Pi}_1\mathbf{U}_1^{(\text{pure})}.\end{aligned}$$

Next, we observe that

$$\begin{aligned}\lambda^2 &= \min_k \lambda_{\min}\big(\mathbf{T}_k\mathbf{T}_k^\top\big)\\ &= \min_k \lambda_{\min}\big(\mathbf{\Pi}_k\mathbf{S}_k\big(\mathbf{\Pi}_{k+1}\otimes\mathbf{\Pi}_{k+2}\big)^\top\big(\mathbf{\Pi}_{k+1}\otimes\mathbf{\Pi}_{k+2}\big)\mathbf{S}_k^\top\mathbf{\Pi}_k^\top\big)\\ &\geq \min_k \lambda_{\min}(\mathbf{\Pi}_k^\top\mathbf{\Pi}_k)\lambda_{\min}\big(\mathbf{S}_k\big(\mathbf{\Pi}_{k+1}\otimes\mathbf{\Pi}_{k+2}\big)^\top\big(\mathbf{\Pi}_{k+1}\otimes\mathbf{\Pi}_{k+2}\big)\mathbf{S}_k^\top\big)\\ &\geq \min_k \frac{p_k}{r_k}\lambda_{\min}\big(\mathbf{S}_k\big(\mathbf{\Pi}_{k+1}\otimes\mathbf{\Pi}_{k+2}\big)^\top\big(\mathbf{\Pi}_{k+1}\otimes\mathbf{\Pi}_{k+2}\big)\mathbf{S}_k^\top\big)\\ &\gtrsim \Delta^2\min_k \frac{p_k}{r_k}\lambda_{\min}\Big(\big(\mathbf{\Pi}_{k+1}\otimes\mathbf{\Pi}_{k+2}\big)^\top\big(\mathbf{\Pi}_{k+1}\otimes\mathbf{\Pi}_{k+2}\big)\Big)\\ &\gtrsim \Delta^2\frac{p_1p_2p_3}{r_1r_2r_3},\end{aligned}$$

where the penultimate line follows from the fact that $\big(\mathbf{\Pi}_{k+1} \otimes \mathbf{\Pi}_{k+2}\big)$ has full column rank. Therefore, $\lambda \gtrsim \Delta \frac{(p_1 p_2 p_3)^{1/2}}{(r_1 r_2 r_3)^{1/2}}$. For the reverse direction, by a similar argument,

$$\begin{aligned}
\lambda^2 &= \min_k \lambda_{\min}(\mathbf{T}_k \mathbf{T}_k^\top) \\
&= \min_k \lambda_{\min}\big(\mathbf{\Pi}_k \mathbf{S}_k \big(\mathbf{\Pi}_{k+1} \otimes \mathbf{\Pi}_{k+2}\big)^\top \big(\mathbf{\Pi}_{k+1} \otimes \mathbf{\Pi}_{k+2}\big) \mathbf{S}_k^\top \mathbf{\Pi}_k^\top\big) \\
&\leq \min_k \lambda_{\max}(\mathbf{\Pi}_k^\top \mathbf{\Pi}_k) \lambda_{\min}\big(\mathbf{S}_k \big(\mathbf{\Pi}_{k+1} \otimes \mathbf{\Pi}_{k+2}\big)^\top \big(\mathbf{\Pi}_{k+1} \otimes \mathbf{\Pi}_{k+2}\big) \mathbf{S}_k^\top\big) \\
&\leq \min_k \frac{p_k}{r_k} \lambda_{\min}\big(\mathbf{S}_k \big(\mathbf{\Pi}_{k+1} \otimes \mathbf{\Pi}_{k+2}\big)^\top \big(\mathbf{\Pi}_{k+1} \otimes \mathbf{\Pi}_{k+2}\big) \mathbf{S}_k^\top\big) \\
&\leq \min_k \frac{p_k}{r_k} \lambda_{\max}\big(\mathbf{\Pi}_{k+1} \otimes \mathbf{\Pi}_{k+2}\big)^\top \big(\mathbf{\Pi}_{k+1} \otimes \mathbf{\Pi}_{k+2}\big) \lambda_{\min}(\mathbf{S}_k^\top \mathbf{S}_k) \\
&\leq \min_k \frac{p_k p_{k+1} p_{k+2}}{r_1 r_2 r_3} \lambda_{\min}(\mathbf{S}_k^\top \mathbf{S}_k) \\
&\leq \Delta^2 \frac{p_1 p_2 p_3}{r_1 r_2 r_3},
\end{aligned}$$

where we have used the assumption that $\lambda_{\max}(\mathbf{\Pi}_k^\top \mathbf{\Pi}_k) \leq \frac{p_k}{r_k}$.

For the remaining part, we note that by the previous argument, we have

$$\mathbf{U}_k = \mathbf{\Pi}_k \mathbf{U}_k^{(\text{pure})}.$$

Since $\mathbf{U}_k^\top \mathbf{U}_k = \mathbf{I}_{r_1}$, it holds that

$$\mathbf{U}_k^{(\text{pure})} (\mathbf{U}_k^{(\text{pure})})^\top \mathbf{\Pi}_k^\top \mathbf{\Pi}_k \mathbf{U}_k^{(\text{pure})} (\mathbf{U}_k^{(\text{pure})})^\top = \mathbf{U}_k^{(\text{pure})} \mathbf{U}_k^\top \mathbf{U}_k (\mathbf{U}_k^{(\text{pure})})^\top = \mathbf{U}_k^{(\text{pure})} (\mathbf{U}_k^{(\text{pure})})^\top,$$

which demonstrates that

$$\mathbf{U}_k^{(\text{pure})} (\mathbf{U}_k^{(\text{pure})})^\top = (\mathbf{\Pi}_k^\top \mathbf{\Pi}_k)^{-1}.$$

Since $\mathbf{U}_k^{(\text{pure})}$ is an $r_1 \times r_1$ matrix and $\lambda_{r_1}(\mathbf{\Pi}_k^\top \mathbf{\Pi}_k) \gtrsim \frac{p_k}{r_k}$, it holds that

$$\begin{aligned}\|\mathbf{U}_k^{(\text{pure})}\|_{2,\infty}^2 &= \max_i \langle (\mathbf{U}_k^{(\text{pure})})_{i\cdot}, (\mathbf{U}_k^{(\text{pure})})_{i\cdot} \rangle \\ &\leq \lambda_{\max}\big(\mathbf{U}_k^{(\text{pure})}(\mathbf{U}_k^{(\text{pure})})^\top\big) \\ &\leq \lambda_{\max}\big(\mathbf{\Pi}_k^\top \mathbf{\Pi}_k\big) \\ &\lesssim \frac{r_k}{p_k}.\end{aligned}$$

Since $\mathbf{U}_k = \mathbf{\Pi}_k \mathbf{U}_k^{(\text{pure})}$ has rows that are convex combinations of $\mathbf{U}_k^{(\text{pure})}$, it holds that

$$\|\mathbf{U}_k\|_{2,\infty} \lesssim \sqrt{\frac{r_k}{p_k}},$$

which demonstrates that $\mu_0 = O(1)$. This completes the proof. □

Proof of Proposition 2. The proof of the first part is similar to Theorem 2.1 of Mao et al. (2021). Suppose that $\mathbf{T}_k$ has SVD $\mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^\top$. By Proposition 3 (which only relies on the assumptions in Proposition 2) it holds that there an invertible matrix $\mathbf{U}_k^{(\text{pure})}$ such that $\mathbf{U}_k = \mathbf{\Pi}_k \mathbf{U}_k^{(\text{pure})}$, where $\mathbf{U}_k^{(\text{pure})}$ consists of the rows of $\mathbf{U}_k$ corresponding to pure nodes. Therefore, for each i it holds that $(\mathbf{U}_k)_{i\cdot}$ is in the convex hull of $\mathbf{U}_k^{(\text{pure})}$.

Now suppose that there exists other parameters $\mathcal{S}', \mathbf{\Pi}_1'$, $\mathbf{\Pi}_2'$ and $\mathbf{\Pi}_3'$ such that $\mathcal{T} = \mathcal{S}' \times_1 \mathbf{\Pi}_1' \times_2 \mathbf{\Pi}_2' \times_3 \mathbf{\Pi}_3'$, where each $\mathbf{\Pi}_k'$ may have different pure nodes. Note that since $\mathcal{T}$ is the same regardless of $\mathbf{\Pi}_k$ and $\mathbf{\Pi}_k'$, its singular value decomposition is fixed (where we arbitrarily specify a choice of sign for unique singular values or basis for repeated singular values). By the previous argument we have that $\widetilde{\mathbf{U}}_k^{(\text{pure})}$ must belong to the convex hull of $\mathbf{U}_k^{(\text{pure})}$, where $\widetilde{\mathbf{U}}_k^{(\text{pure})}$ corresponds to the pure nodes associated to $\mathbf{\Pi}_k'$. By applying Proposition 3 again to the new decomposition, it must hold that $\mathbf{U}_k = \mathbf{\Pi}_k' \widetilde{\mathbf{U}}_k^{(\text{pure})}$, which shows that $\mathbf{U}_k^{(\text{pure})}$ belongs to the convex hull of $\widetilde{\mathbf{U}}_k^{(\text{pure})}$. Since both convex hulls are subsets of each other, it holds that the convex hulls of $\mathbf{U}_k^{(\text{pure})}$ and $\widetilde{\mathbf{U}}_k^{(\text{pure})}$ are the same. Consequently, it must hold that $\mathbf{U}_k^{(\text{pure})} = \mathcal{P}_k \widetilde{\mathbf{U}}_k^{(\text{pure})}$ for some permutation matrix $\mathcal{P}_k$.

Now we note that by the identity $\mathbf{U}_k = \mathbf{\Pi}_k \mathbf{U}_k^{(\text{pure})} = \mathbf{\Pi}_k' \widetilde{\mathbf{U}}_k^{(\text{pure})}$, it holds that $\mathbf{\Pi}_k \mathbf{U}_k^{(\text{pure})} =$

$\mathbf{\Pi}_k' \mathcal{P}_k \mathbf{U}_k^{(\text{pure})}$, which demonstrates that

$$(\mathbf{\Pi}_k - \mathbf{\Pi}_k' \mathcal{P}_k)\mathbf{U}_k^{(\text{pure})} = 0.$$

Since $\mathbf{U}_k^{(\text{pure})}$ is full rank, it must therefore hold that $\mathbf{\Pi}_k = \mathbf{\Pi}_k' \mathcal{P}_k$. Consequently,

$$\mathcal{T} = \mathcal{S}' \times_1 \left(\mathbf{\Pi}_1 \mathcal{P}_1\right) \times_2 \left(\mathbf{\Pi}_2 \mathcal{P}_2\right) \times_3 \left(\mathbf{\Pi}_3 \mathcal{P}_3\right),$$

which shows that $\mathcal{S} = \mathcal{S}' \times_1 \mathcal{P}_1 \times_2 \mathcal{P}_2 \times_3 \mathcal{P}_3$, which completes the proof of the first part of the result.

The second part of the result essentially follows the proof of Theorem 2.2 of Mao et al. (2021). Without loss of generality we prove the result for mode 1. Assume for contradiction that there is a community without any pure nodes; without loss of generality let it be the first community. Then there is some $\delta > 0$ such that $(\mathbf{\Pi}_1)_{i1} \leq 1 - \delta$ for all i. Define

$$\mathbf{H} := \left[\begin{array}{c|c} 1 + (r_1 - 1)\varepsilon^2 & -\varepsilon^2 \mathbf{1}_{r_1 - 1}^\top \\ \hline 0 & \varepsilon \mathbf{1}_{r_1 - 1}\mathbf{1}_{r_1 - 1}^\top + (1 - (r_1 - 1)\varepsilon)\mathbf{I}_{r_1 - 1} \end{array}\right],$$

where $0 < \varepsilon < \delta$. For ε sufficiently small, $\mathbf{H}$ is full rank, and the rows of $\mathbf{H}$ sum to one. Consequently, $\widetilde{\mathbf{\Pi}}_1 := \mathbf{\Pi}_1 \mathbf{H}$ also has rows that sum to one. Moreover, for any i, $(\widetilde{\mathbf{\Pi}}_1)_{i1} = (\mathbf{\Pi}_1)_{i1}(1 + (K-1)\varepsilon^2) \geq 0$, and for any $2 \leq l \leq r_1$,

$$\begin{aligned}
(\widetilde{\mathbf{\Pi}}_1)_{il} &= -(\mathbf{\Pi}_1)_{i1}\varepsilon^2 + \sum_{l'=1}^{r_1} (\mathbf{\Pi}_1)_{il'} \mathbf{H}_{l'l} \\
&= -(\mathbf{\Pi}_1)_{i1}\varepsilon^2 + (\mathbf{\Pi}_1)_{il}(1 - (K-1)\varepsilon) + \sum_{l'=1}^{r_1} (\mathbf{\Pi}_1)_{il'}\varepsilon \\
&\geq -(\mathbf{\Pi}_1)_{i1}\varepsilon^2 + \sum_{l'=1}^{r_1} (\mathbf{\Pi}_1)_{il'}\varepsilon \\
&\geq (1 - \delta)\varepsilon^2 + \varepsilon\delta \\
&> 0,
\end{aligned}$$

and hence $\widetilde{\mathbf{\Pi}}_1$ has positive entries. Therefore, for ε sufficiently small $\widetilde{\mathbf{\Pi}}_1$ is a valid member-

ship matrix. In addition, we have that

$$\begin{aligned}\mathcal{M}_1(\mathcal{T}) &= \mathbf{\Pi}_1 \mathcal{M}_1(\mathcal{S})\big(\mathbf{\Pi}_2 \otimes \mathbf{\Pi}_3\big)^\top \\ &= \widetilde{\mathbf{\Pi}}_1 \mathbf{H}^{-1} \mathcal{M}_1(\mathcal{S})\big(\mathbf{\Pi}_2 \otimes \mathbf{\Pi}_3\big)^\top \\ &= \widetilde{\mathbf{\Pi}}_1 \mathcal{M}_1(\mathcal{S} \times_1 \mathbf{H}^{-1})\big(\mathbf{\Pi}_2 \otimes \mathbf{\Pi}_3\big)^\top,\end{aligned}$$

which shows that $\mathcal{T} = \widetilde{\mathcal{S}} \times_1 \widetilde{\mathbf{\Pi}}_1 \times_2 \mathbf{\Pi}_2 \times_3 \mathbf{\Pi}_3$ is another representation of $\mathcal{T}$, where $\widetilde{\mathcal{S}} = \mathcal{S} \times_1 \mathbf{H}^{-1}$. Since $\mathbf{H}$ is not a permutation matrix, we see that we have a contradiction, which completes the proof. □

D.2.2 Proof of Theorem 10

Proof of Theorem 10. Our proof is similar to the proof of the main result in Mao et al. (2021) as well as the proof of Theorem 4.9 in Xie (2022), where we will apply Theorem 3 of Gillis and Vavasis (2014). We first prove the result assuming that $\lambda/\sigma \gtrsim \kappa p\sqrt{\log(p)}/p_{\min}^{1/4}$, that $r \leq p_{\min}^{1/4}$ and that $\mu_0 = O(1)$; the result will then follow by applying Lemma 6. Without loss of generality, we prove the result for $k = 1$.

First, observe that by Theorem 11, with probability at least $1 - p^{-10}$ it holds that there is an orthogonal matrix $\mathbf{W}$ such that for t iterations with t as in Theorem 11 it holds that the output $\widehat{\mathbf{U}}$ of HOOI satisfies

$$\widehat{\mathbf{U}} = \mathbf{U}\mathbf{W} + \text{error},$$

with

$$\|\text{error}\|_{2,\infty} \lesssim \frac{\kappa\sqrt{r_k \log(p)}}{\lambda/\sigma}.$$

Since by Lemma 6 $\mathbf{U} = \mathbf{\Pi}\mathbf{U}^{(\text{pure})}$, it holds that

$$\widehat{\mathbf{U}}^\top = \mathbf{W}^\top (\mathbf{U}^{(\text{pure})})^\top \mathbf{\Pi}^\top + \text{error}^\top.$$

We will apply Theorem 3 of Gillis and Vavasis (2014), with $\mathbf{M}$, $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{N}$ therein equal to

$\mathbf{U}$, $\mathbf{W}^\top(\mathbf{U}^{(\text{pure})})^\top$, $\mathbf{\Pi}^\top$ and $\text{error}^\top$ respectively. Define, for some sufficiently large constant C,

$$\varepsilon := C\frac{\kappa\sqrt{r_k\log(p)}}{\lambda/\sigma}.$$

It then holds that $\|\text{error}_{i\cdot}\| \leq \varepsilon$ on the event in Theorem 11. We also need to check the bound

$$\varepsilon < \lambda_{\min}(\mathbf{U}^{(\text{pure})})\min\left(\frac{1}{2\sqrt{r_1-1}}, \frac{1}{4}\right)\left(1+80\frac{\sigma_1^2(\mathbf{U}^{(\text{pure})})}{\sigma_r^2(\mathbf{U}^{(\text{pure})})}\right)^{-1}.$$

First we note that by the proof of Lemma 6, we have that $\lambda_{\min}^2(\mathbf{U}^{(\text{pure})}) = \lambda_{\min}\big(\mathbf{U}^{(\text{pure})}(\mathbf{U}^{(\text{pure})})^\top\big) = \lambda_{\min}\big((\mathbf{\Pi}^\top\mathbf{\Pi})^{-1}\big)$. Since $\lambda_{\max}\big(\mathbf{\Pi}^\top\mathbf{\Pi}\big) \lesssim \frac{p_1}{r_1}$, we have that $\lambda_{\min}(\mathbf{U}^{(\text{pure})}) \gtrsim \frac{\sqrt{r_1}}{\sqrt{p_1}}$.

We note that

$$\begin{aligned}\frac{\|\mathbf{U}^{(\text{pure})}\mathbf{W}\|_{2,\infty}^2}{\lambda_r^2(\mathbf{U}^{(\text{pure})})} &\leq \frac{\lambda_{\max}^2(\mathbf{U}^{(\text{pure})})}{\lambda_r^2(\mathbf{U}^{(\text{pure})})} \\ &= \frac{\lambda_{\max}^2(\mathbf{U}^{(\text{pure})})}{\lambda_r^2(\mathbf{U}^{(\text{pure})})} \\ &= \frac{\lambda_{\max}(\mathbf{\Pi}^\top\mathbf{\Pi})}{\lambda_{\min}(\mathbf{\Pi}^\top\mathbf{\Pi})} \\ &\asymp C,\end{aligned}$$

since $\lambda_{\min}(\mathbf{\Pi}^\top\mathbf{\Pi}) \gtrsim p_1$ by assumption. Consequently, plugging in these estimates, it suffices to show that

$$\varepsilon < c\frac{\sqrt{r_1}}{\sqrt{p_1}}\frac{1}{\sqrt{r_1}} = \frac{c}{\sqrt{p_1}},$$

where c is some sufficiently small constant. Plugging in the definition of ε, we see that we require that

$$C\frac{\kappa\sqrt{r_1\log(p)}}{\lambda/\sigma} \leq \frac{c}{\sqrt{p_1}},$$

which is equivalent to the condition

$$\lambda/\sigma \gtrsim \kappa\sqrt{p_1 r_1 \log(p)},$$

which holds under the condition $\lambda/\sigma \gtrsim \kappa p\sqrt{\log(p)}/p_{\min}^{1/4}$ and $r \leq p_{\min}^{1/4}$. Therefore, we may apply Theorem 3 of Gillis and Vavasis (2014) to find that there exists a permutation $\mathcal{P}$ such that

$$\|\widehat{\mathbf{U}}^{(\text{pure})} - \mathcal{P}^\top \mathbf{U}^{(\text{pure})}\mathbf{W}\|_{2,\infty} \leq C\varepsilon.$$

We now use this bound to provide our final bound. First, since $\varepsilon \lesssim \frac{1}{\sqrt{p}}$, by Weyl's inequality it holds that

$$\begin{aligned}\lambda_{\min}(\widehat{\mathbf{U}}^{(\text{pure})}) &\geq \lambda_{\min}(\mathbf{U}_k^{(\text{pure})}) - \sqrt{r_1}\varepsilon \\ &\geq C\frac{\sqrt{r_1}}{\sqrt{p_1}} - \frac{c\sqrt{r_1}}{\sqrt{p_1}} \\ &\gtrsim \frac{\sqrt{r_1}}{\sqrt{p_1}},\end{aligned}$$

as long as c is sufficiently small. Consequently, $\|(\widehat{\mathbf{U}}^{(\text{pure})})^{-1}\| \lesssim \sqrt{\frac{p_1}{r_1}}$. Therefore,

$$\begin{aligned}
\|\widehat{\mathbf{\Pi}} - \mathbf{\Pi}\mathcal{P}\|_{2,\infty} &= \|\widehat{\mathbf{U}}(\widehat{\mathbf{U}}^{(\text{pure})})^{-1} - \mathbf{U}(\mathbf{U}^{(\text{pure})})^{-1}\mathcal{P}\|_{2,\infty} \\
&= \|\widehat{\mathbf{U}}(\widehat{\mathbf{U}}^{(\text{pure})})^{-1} - \mathbf{U}\mathbf{W}(\mathcal{P}^\top \mathbf{U}^{(\text{pure})}\mathbf{W})^{-1}\|_{2,\infty} \\
&\le \|\big(\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}\big)(\widehat{\mathbf{U}}^{(\text{pure})})^{-1}\|_{2,\infty} + \|\mathbf{U}\mathbf{W}\big((\widehat{\mathbf{U}}^{(\text{pure})})^{-1} - (\mathcal{P}^\top \mathbf{U}^{(\text{pure})}\mathbf{W})^{-1}\big)\|_{2,\infty} \\
&\le \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}\|_{2,\infty}\|(\widehat{\mathbf{U}}^{(\text{pure})})^{-1}\| + \|\mathbf{\Pi}\mathbf{U}^{(\text{pure})}\mathbf{W}\big((\widehat{\mathbf{U}}^{(\text{pure})})^{-1} - (\mathcal{P}^\top \mathbf{U}^{(\text{pure})}\mathbf{W})^{-1}\big)\|_{2,\infty} \\
&\le \varepsilon\|(\widehat{\mathbf{U}}^{(\text{pure})})^{-1}\| + \|\mathbf{\Pi}\|_{\infty\to\infty}\|\mathbf{U}^{(\text{pure})}\mathbf{W}\big((\widehat{\mathbf{U}}^{(\text{pure})})^{-1} - (\mathcal{P}^\top \mathbf{U}^{(\text{pure})}\mathbf{W})^{-1}\big)\|_{2,\infty} \\
&\le \varepsilon\|(\widehat{\mathbf{U}}^{(\text{pure})})^{-1}\| + \|\mathbf{U}^{(\text{pure})}\mathbf{W}\big((\mathcal{P}^\top \widehat{\mathbf{U}}^{(\text{pure})})^{-1} - (\mathbf{U}^{(\text{pure})}\mathbf{W})^{-1}\big)\mathcal{P}\|_{2,\infty} \\
&\le \varepsilon\|(\widehat{\mathbf{U}}^{(\text{pure})})^{-1}\| + \|\big(\mathbf{U}^{(\text{pure})}\mathbf{W}(\mathcal{P}\widehat{\mathbf{U}}^{(\text{pure})})^{-1} - \mathbf{I}_{r_k}\big)\mathcal{P}\|_{2,\infty} \\
&\le \varepsilon\|(\widehat{\mathbf{U}}^{(\text{pure})})^{-1}\| + \|\big(\mathbf{U}^{(\text{pure})}\mathbf{W} - \mathcal{P}\widehat{\mathbf{U}}^{(\text{pure})}\big)(\mathcal{P}\widehat{\mathbf{U}}^{(\text{pure})})^{-1}\mathcal{P}\|_{2,\infty} \\
&\le \varepsilon\|(\widehat{\mathbf{U}}^{(\text{pure})})^{-1}\| + \|\mathcal{P}^\top \mathbf{U}^{(\text{pure})}\mathbf{W} - \widehat{\mathbf{U}}^{(\text{pure})}\|_{2,\infty}\|(\mathcal{P}\widehat{\mathbf{U}}^{(\text{pure})})^{-1}\|_{2,\infty} \\
&\le 2\varepsilon\sqrt{p_1/r_1} \\
&\lesssim \frac{\kappa\sqrt{r_1\log(p)}}{\lambda/\sigma}\sqrt{\frac{p_1}{r_1}} \\
&\asymp \frac{\kappa\sqrt{p_1\log(p)}}{\lambda/\sigma}.
\end{aligned}$$

Therefore, all that remains is to apply Lemma 6. First, we need to check that the condition

$$\lambda/\sigma \gtrsim \kappa p\sqrt{\log(p)}/p_{\min}^{1/4}$$

holds; by Lemma 6 this is equivalent to the condition

$$\Delta/\sigma \gtrsim \frac{\kappa p\sqrt{\log(p)}}{p_{\min}^{1/4}}\frac{\sqrt{r_1r_2r_3}}{\sqrt{p_1p_2p_3}},$$

which is in Assumption 4.2. Finally, by Lemma 6, we obtain the final upper bound

$$\|\widehat{\mathbf{\Pi}} - \mathbf{\Pi}\mathcal{P}\|_{2,\infty} \lesssim \frac{\kappa\sigma\sqrt{r_1r_2r_3\log(p)}}{\Delta(p_{-1})^{1/2}},$$

as desired. □

D.2.3 Proof of Corollary 4

Proof of Corollary 4. Fix an index k, and let $\mathcal{P}$ denote the permutation matrix from Theorem 10. Then it holds that

$$\begin{aligned}\inf_{\text{Permutations } \mathcal{P}} \|\big(\widehat{\mathbf{\Pi}}_k - \mathbf{\Pi}_k\mathcal{P}\big)_{i\cdot})_{i\cdot}\|_1 &\leq \sqrt{r_k}\|\big(\widehat{\mathbf{\Pi}}_k - \mathbf{\Pi}_k\mathcal{P}\big)_{i\cdot}\|_2 \\ &\leq \frac{r^2\kappa\sqrt{\log(p)}}{(\Delta/\sigma)(p_{-k})^{1/2}}.\end{aligned}$$

Averaging over the rows completes the proof. □

D.3 Auxiliary Probabilistic Lemmas

Lemma 46. *Let* $\mathbf{A}$ *be any fixed matrix independent from* $e_m^\top \mathbf{Z}_k$. *Then there exists an absolute constant* $C > 0$ *such that with probability at least* $1 - O(p_{\max}^{-20})$,

$$\|e_m^\top \mathbf{Z}_k \mathbf{A}\| \leq C\sigma\sqrt{p_{-k}\log(p_{\max})}\|\mathbf{A}\|_{2,\infty}.$$

Proof. This follows from Cai et al. (2021a), Lemma 12. □

Lemma 47. *Let* $\mathbf{A}$ *be a matrix independent from* $\mathbf{Z}_k - \mathbf{Z}_k^{j-m}$, *where* $\mathbf{Z}_k^{j-m}$ *is defined in Appendix D.1. Then there exists an absolute constant* $C > 0$ *such that with probability at least* $1 - O(p_{\max}^{-30})$,

$$\|\left(\mathbf{Z}_k - \mathbf{Z}_k^{j-m}\right)\mathbf{A}\| \leq C\sigma\sqrt{p_{-j}\log(p_{\max})}\|\mathbf{A}\|_{2,\infty}.$$

Proof. If $j = k$, the result follows by Lemma 46. Therefore, we restrict our attention to when $j \neq k$. First, note that

$$\|\left(\mathbf{Z}_k - \mathbf{Z}_k^{j-m}\right)\mathbf{A}\| \leq \sqrt{p_k}\|\left(\mathbf{Z}_k - \mathbf{Z}_k^{j-m}\right)\mathbf{A}\|_{2,\infty}.$$

Next, consider any fixed row of $\mathbf{Z}_k - \mathbf{Z}_k^{j-m}$. Observe that the q'th row can be written as

$$\sum_{\Omega} (\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot},$$

where the set Ω consists of the p_{-k-j} random variables in the q'th row of $\mathbf{Z}_k - \mathbf{Z}_k^{j-m}$. Note that this is a sum of independent random matrices. By the matrix Bernstein inequality (Proposition 2 of Koltchinskii et al. (2011)), it holds that with probability at least $1 - p_{\max}^{-31}$ that

$$\| \sum_{\Omega} (\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot} \| \leq C \max \left\{ \sigma_Z \sqrt{p_{-k-j} \log(p_{\max})}, U_Z \log(p) \right\}$$

where

$$\sigma_Z^2 \coloneqq \max_l \max \left\{ \left\| \mathbb{E}\big((\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot} \big) \big((\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot} \big)^\top \right\|, \left\| \mathbb{E}\big((\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot} \big)^\top \big((\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot} \big) \right\| \right\};$$
$$U_Z \coloneqq \max_l \| (\mathbf{Z}_k)_{ql} A_{l\cdot} \|_{\psi_2}$$

(Note that Proposition 2 of Koltchinskii et al. (2011) holds for IID random matrices, but the proof works equally as well if uniform bounds on σ_Z and U_Z are obtained). Observe that

$$\begin{aligned} \left\| \mathbb{E}\left[(\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot} \right] \left[(\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot} \right]^\top \right\| &\leq \sigma^2 \| \mathbf{A}_{l\cdot} \mathbf{A}_{l\cdot}^\top \| \\ &\leq \sigma^2 \| \mathbf{A} \|_{2,\infty}^2; \\ \left\| \mathbb{E}\left[(\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot} \right]^\top \left[(\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot} \right] \right\| &\leq \sigma^2 \ \| \mathbf{A}_{l\cdot}^\top \mathbf{A}_{l\cdot} \| \\ &\leq \sigma^2 \| \mathbf{A} \|_{2,\infty}^2. \end{aligned}$$

Similarly, by subgaussianity of the entries of $\mathbf{Z}_k$,

$$\max_l \| (\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot} \|_{\psi_2} \leq C\sigma \| \mathbf{A} \|_{2,\infty}.$$

Therefore, with probability at least $1 - p_{\max}^{-31}$, it holds that

$$\begin{aligned}\|\sum_{\Omega} (\mathbf{Z}_k)_{ql} \mathbf{A}_{l\cdot}\| &\leq C \max\left\{\sigma_Z \sqrt{p_{-k-j}\log(p_{\max})}, U_Z \log(p_{\max})\right\} \\ &\leq C\sigma \|\mathbf{A}\|_{2,\infty} \max\left\{\sqrt{p_{-k-j}\log(p_{\max})}, \log(p_{\max})\right\} \\ &\leq C\sigma \|\mathbf{A}\|_{2,\infty} \sqrt{p_{-k-j}\log(p_{\max})}.\end{aligned}$$

Taking a union bound over all p_k rows shows that this holds uniformly with probability at least $1 - O(p_{\max}^{-30})$. Therefore,

$$\begin{aligned}\left\|\left(\mathbf{Z}_k - \mathbf{Z}_k^{j-m}\right)\mathbf{A}\right\| &\leq \sqrt{p_k}\left\|\left(\mathbf{Z}_k - \mathbf{Z}_k^{j-m}\right)\mathbf{A}\right\|_{2,\infty} \\ &\leq C\sigma \|\mathbf{A}\|_{2,\infty}\sqrt{p_{-j}\log(p_{\max})}\end{aligned}$$

as desired. □

Lemma 48. *Suppose $\mathcal{Z} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ is a tensor with mean-zero subgaussian entries, each with ψ_2 norm bounded by 1. Suppose that $r_2 r_3 \leq p_1 r_1$. Then for some universal constant C, the following holds with probability at least $1 - c\exp(-cp_{\max})$:*

$$\sup_{\substack{\|\mathbf{U}_1\|=1, \text{rank}(\mathbf{U}_1)\leq 2r_1 \\ ,\|\mathbf{U}_2\|=1, \text{rank}(\mathbf{U}_2)\leq 2r_2}} \|\mathbf{Z}\left(\mathcal{P}_{\mathbf{U}_1} \otimes \mathcal{P}_{\mathbf{U}_2}\right)\| \leq C\sqrt{p_{\max} r_{\max}}.$$

Proof. See Lemma 8 of Han et al. (2021) or Lemma 3 of Zhang and Han (2019). □

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix E

Proofs from Chapter 5

E.1 Proofs of Main Results

In this section, we first prove Proposition 4, Theorems 13 and 14 and Corollaries 5 and 6. We then prove Propositions 5, 6, 7, and 8. Finally, we prove the more technical results. Our proofs require careful tabulation of the various alignment matrices orthogonal matrices. We remark that all orthogonal matrices appearing in the following proofs are written with the letter $\mathbf{W}$ and all indefinite orthogonal matrices are written with the letter $\mathbf{Q}$, and we allow the constants implicit in the notation $O(\cdot)$ to depend on d in an arbitrary manner. Finally, in our proofs, we will provide bounds with a constant C that may change from line to line.

Before proving Proposition 4, we include some important related results that we will require. First, Theorem 7 in Solanki et al. (2019) says that when we have a (p, q)-admissible distribution, the support is bounded.

Theorem 26 (Theorem 7 of Solanki et al. (2019))**.** *Suppose F is a (p, q)-admissible distribution; that is for all $\mathbf{x}, \mathbf{y} \in \operatorname{supp}(F)$, $\mathbf{x}^\top \mathbf{I}_{p,q} \mathbf{y} \in [0, 1]$. Then $\operatorname{supp}(F)$ is bounded.*

If needed, we can assume without loss of generality that the support $\mathbf{\Omega}$ is compact by extending it to its closure if necessary. We will also need an adaptation of Theorem 1 from Agterberg et al. (2020b) for the two graph setting. The proof is straightforward and included in Section E.1.5.

Lemma 49. *Suppose $F_X \simeq F_Y$, and that $\mathbf{\Delta_X} \mathbf{I}_{p,q}$ and $\mathbf{\Delta_Y} \mathbf{I}_{p,q}$ have distinct eigenvalues.*

Let $\mathbf{Q_X}$ *be the matrix such that* $\mathbf{U_X}|\mathbf{\Lambda}|^{\mathbf{1/2}}\mathbf{Q_X} = \mathbf{X}$. *and similarly for* $\mathbf{Q_Y}$ *and* $\mathbf{Y}$. *Then there exists a fixed matrix* $\widetilde{\mathbf{Q}}$ *such that both* $||\mathbf{Q_X} - \widetilde{\mathbf{Q}}|| \to 0$ *and* $||\mathbf{Q_Y}\mathbf{T^{-1}} - \widetilde{\mathbf{Q}}|| \to 0$ *almost surely, where* $\mathbf{T} \in \mathbb{O}(p,q)$ *is the matrix such that* $F_Y = F_X \circ \mathbf{T}$.

Finally, we need the following restatement of Theorem 5 of Rubin-Delanchy et al. (2020). Given Lemma 49 and the concentration results in this section (c.f. Lemma 50) the proof is straightforward by adapting the proof in Rubin-Delanchy et al. (2020) and is thus omitted. Note that $\mathbf{Q_X^{-1}} = \mathbf{I}_{p,q}\mathbf{Q_X}^\top\mathbf{I}_{p,q}$ from the equation $\mathbf{Q_X}\mathbf{I}_{p,q}\mathbf{Q_X}^\top = \mathbf{I}_{p,q}$, which will be useful in the sequel.

Theorem 27. *Let* $(\mathbf{A}, \mathbf{X}) \sim GRDPG(F_X, n, \alpha_n)$ *for* $n\alpha_n = \omega(\log^4(n))$. *Let* $\mathbf{Q_X}$ *be the matrix such that* $\mathbf{U_X}|\mathbf{\Lambda}|^{1/2}\mathbf{Q_X} = \mathbf{X}$. *Then there exists an orthogonal matrix* $\mathbf{W}_* \in \mathbb{O}(d) \cap \mathbb{O}(p,q)$ *depending on* n *such that with probability at least* $1 - n^{-2}$

$$\|\widehat{\mathbf{X}} - \alpha_n^{1/2}\mathbf{X}\mathbf{Q_X^{-1}}\mathbf{W}_*\|_{2,\infty} = O\left(\frac{\log(n)}{n^{1/2}}\right).$$

Furthermore, as $n \to \infty$, *if* $\mathbf{\Delta}\mathbf{I}_{p,q}$ *has distinct eigenvalues, then* $\|\mathbf{W}_* - \mathbf{I}\| = O\left((n\alpha_n)^{-1}\right)$ *with probability at least* $1 - n^{-2}$. *In this case, we have the bound*

$$\|\widehat{\mathbf{X}} - \alpha_n^{1/2}\mathbf{X}\mathbf{Q_X^{-1}}\|_{2,\infty} = O\left(\frac{\log(n)}{n^{1/2}}\right).$$

Proof of Proposition 4. We first show that $|U_{n,m}(\widehat{\mathbf{X}}/\alpha_n^{1/2}, \widehat{\mathbf{Y}}/\beta_m^{1/2}) - U_{n,m}(\mathbf{X}\mathbf{Q_X^{-1}}\mathbf{W}_*^{\mathbf{X}}, \mathbf{Y}\mathbf{Q_Y^{-1}}\mathbf{W}_*^{\mathbf{Y}})| \to 0$ almost surely. By continuity of κ and the fact that the supports of F_X and F_Y are bounded by Theorem 26, we have that

$$|\kappa(\widehat{X}_i/\alpha_n^{1/2}, \widehat{X}_j/\alpha_n^{1/2}) - \kappa((\mathbf{W}_*^{\mathbf{X}}\mathbf{Q_X^{-1}})^\top X_i, (\mathbf{W}_*^{\mathbf{X}}\mathbf{Q_X^{-1}})^\top X_j)| \leq C||\alpha_n^{-1/2}\widehat{\mathbf{X}} - \mathbf{X}\mathbf{Q_X^{-1}}\mathbf{W}_*^{\mathbf{X}}||_{2,\infty};$$

$$|\kappa(\widehat{Y}_i/\beta_m^{1/2}, \widehat{Y}_j/\beta_m^{1/2}) - \kappa((\mathbf{W}_*^{\mathbf{Y}}\mathbf{Q_Y^{-1}})^\top Y_i, (\mathbf{W}_*^{\mathbf{Y}}\mathbf{Q_Y^{-1}})^\top Y_j)| \leq C||\beta_m^{-1/2}\widehat{\mathbf{Y}} - \mathbf{Y}\mathbf{Q_Y^{-1}}\mathbf{W}_*^{\mathbf{Y}}||_{2,\infty};$$

$$|\kappa(\widehat{X}_i/\alpha_n^{1/2}, \widehat{Y}_j/\beta_m^{1/2}) - \kappa((\mathbf{W}_*^{\mathbf{X}}\mathbf{Q_X^{-1}})^\top X_i, (\mathbf{W}_*^{\mathbf{Y}}\mathbf{Q_Y^{-1}})^\top Y_j)| \leq C \max\Big(||\alpha_n^{-1/2}\widehat{\mathbf{X}} - \mathbf{X}\mathbf{Q_X^{-1}}\mathbf{W}_*^{\mathbf{X}}||_{2,\infty},$$
$$||\beta_m^{-1/2}\widehat{\mathbf{Y}} - \mathbf{Y}\mathbf{Q_Y^{-1}}\mathbf{W}_*^{\mathbf{Y}}||_{2,\infty}\Big).$$

Each term tends to zero almost surely by Theorem 27. Hence,

$$|U_{n,m}(\widehat{\mathbf{X}}/\alpha_n^{1/2}, \widehat{\mathbf{Y}}/\beta_m^{1/2}) - U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_*^{\mathbf{X}}, \mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_*^{\mathbf{Y}})| \to 0$$

almost surely. Furthermore, we see that since $\mathbf{\Delta I}_{p,q}$ has distinct eigenvalues, the matrices $\mathbf{W}_*^{\mathbf{X}}$ and $\mathbf{W}_*^{\mathbf{Y}}$ are converging to the identity as $n \to \infty$.

Hence, it suffices to consider what the term $U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}, \mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}})$ is converging to under the null and alternative respectively. Note that

$$\begin{aligned}|U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}, \mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}) - U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}^{-\mathbf{1}}, \mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}})| &\leq C\left(\|\mathbf{X}(\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}} - \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}})\|_{2,\infty} + \|\mathbf{Y}(\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}} - \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}})\|_{2,\infty}\right) \\ &\leq C\left(\|\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}} - \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\| + \|\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}} - \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\|\right),\end{aligned}$$

where we have implicitly used Theorem 26 and the fact that κ is twice continuously differentiable and hence Lipschitz on the closure of the support of $F_X \circ \widetilde{\mathbf{Q}}^{-\mathbf{1}}$. The right hand side tends to zero almost surely by Lemma 49 and Theorem 2 in Agterberg et al. (2020b).

Hence, it suffices to analyze the convergence of $U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}, \mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}})$ under the null and alternative respectively. Note that $\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}$ and $\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}$ are deterministic matrices. Under the null hypothesis, the matrix $\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}$ can be replaced with the limiting matrix $\mathbf{T}^{-\mathbf{1}}\widetilde{\mathbf{Q}}^{-\mathbf{1}}$ by Lemma 49. Therefore, under the null hypothesis, $F_X \circ \mathbf{T} = F_Y$, so $\mu[F_X \circ \widetilde{\mathbf{Q}}^{-\mathbf{1}}] = \mu[F_Y \circ \mathbf{T}^{-\mathbf{1}}\widetilde{\mathbf{Q}}^{-\mathbf{1}}]$ since κ is assumed to be a characteristic kernel. By Gretton et al. (2012), as $n, m \to \infty$ and $n/(n+m) \to \rho \in (0,1)$, we have that

$$U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}^{-\mathbf{1}}, \mathbf{Y}\mathbf{T}^{-\mathbf{1}}\widetilde{\mathbf{Q}}^{-\mathbf{1}}) \to \|\mu[F_{\mathbf{X}} \circ \widetilde{\mathbf{Q}}^{-\mathbf{1}}] - \mu[F_{\mathbf{Y}} \circ \mathbf{T}^{-\mathbf{1}} \circ \widetilde{\mathbf{Q}}^{-\mathbf{1}}]\|_{\mathcal{H}}^2 = 0. \tag{E.1}$$

Hence, $U_{n,m}(\widehat{\mathbf{X}}/\alpha_n^{1/2}, \widehat{\mathbf{Y}}/\beta_m^{1/2})$ converges to zero under the null hypothesis.

Under the alternative hypothesis, the matrix $\mathbf{T}^{-\mathbf{1}}\widetilde{\mathbf{Q}}^{-\mathbf{1}}$ is replaced with a matrix $\check{\mathbf{Q}}$, and the term

$$\|\mu[F_X \circ \widetilde{\mathbf{Q}}^{-1}] - \mu[F_Y \circ \check{\mathbf{Q}}^{-\mathbf{1}}]\|_{\mathcal{H}}^2 = c > 0$$

or otherwise the null hypothesis would be true. Therefore, under the alternative, the term

$U_{n,m}(\widehat{\mathbf{X}}/\alpha_n^{1/2}, \widehat{\mathbf{Y}}/\beta_m^{1/2})$ converges to some positive constant which completes the proof. □

E.1.1 Proof of Theorems 13 and 14 and Corollaries 5 and 54

We will require a few supporting lemmas. The first is on the rate of approximation of the limiting matrix $\widetilde{\mathbf{Q}}$ from Theorem 2 of Agterberg et al. (2020b). We note that this improves on a bound of order $\left(\frac{\log(n)}{n}\right)^{1/8}$ given in Solanki et al. (2019). The proof is in Section E.1.5.

Lemma 50. *Define* $\mathbf{Q}_{\mathbf{X}}$ *as the matrix such that* $\mathbf{U}_{\mathbf{X}}|\mathbf{\Lambda}_{\mathbf{X}}|^{1/2} = \alpha_n^{1/2}\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}$*, and let* $\widetilde{\mathbf{Q}}$ *be its corresponding limit. Then with probability at least* $1 - n^{-2}$ *that there exists an orthogonal matrix* $\mathbf{W}_{\mathbf{X}} \in \mathbb{O}(d) \cap \mathbb{O}(p, q)$ *such that*

$$\|\mathbf{W}_{\mathbf{X}}\mathbf{Q}_{\mathbf{X}} - \widetilde{\mathbf{Q}}\| = O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right). \tag{E.2}$$

Moreover, $\mathbf{W}_{\mathbf{X}}$ *has blocks corresponding to repeated eigenvalues of* $\mathbf{\Delta}\mathbf{I}_{p,q}$*. If* $\mathbf{\Delta}\mathbf{I}_{p,q}$ *has no repeated eigenvalues, then* $\mathbf{W}_{\mathbf{X}}$ *is a sign matrix. In addition, the matrices* $\mathbf{Q}_{\mathbf{X}}$ *and* $\widetilde{\mathbf{Q}}$ *do not depend on the sparsity factor.*

The next lemma is a technical result concerning the Frobenius norm concentration of $\widehat{\mathbf{X}}$ to $\mathbf{X}$ and is needed to guarantee the existence of the specific matrices $\mathbf{W}_*^{\mathbf{X}}$ and $\mathbf{W}_*^{\mathbf{Y}}$, though it is also used in the proof of Lemma 52. We note that similar results were proven in Tang et al. (2017a) and Tang and Priebe (2018) in the setting of random dot product graphs, and in Athreya et al. (2020) in the setting of numerical linear algebra for random matrices. The asymptotic normality of a related Frobenius norm error for stochastic blockmodels was proven in Li and Li (2018). The proof is in Section E.1.3. Throughout this section, recall that we define

$$\widetilde{\mathbf{X}} := \mathbf{U}_{\mathbf{X}}|\mathbf{\Lambda}_{\mathbf{X}}|^{1/2}; \qquad \mathbf{P}^{(1)} = \mathbf{U}_{\mathbf{X}}\mathbf{\Lambda}_{\mathbf{X}}\mathbf{U}_{\mathbf{X}}^{\top} = \alpha_n\mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^{\top}$$

with similar notation for $\widetilde{\mathbf{Y}}$ and $\mathbf{P}^{(2)}$. We therefore have the identity

$$\widetilde{\mathbf{X}} = \alpha_n^{1/2}\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}. \tag{E.3}$$

Note that $\widetilde{\mathbf{X}}$ and $\widehat{\mathbf{X}}$ depend on the sparsity α_n, but the matrix $\mathbf{X}$ does not (since its rows are i.i.d. F_X), and that the matrix $\widetilde{\mathbf{X}}$ can be thought of as the adjacency spectral embedding of the probability generating matrix $\mathbf{P} = \alpha_n \mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top$. See also Table E.2 below.

Lemma 51. *Let $(\mathbf{A}, \mathbf{X}) \sim GRDPG(F_X, n, \alpha_n)$ for some α_n satisfying $n\alpha_n \geq \omega(\log^4(n))$. Asymptotically almost surely, for the sequence of block-orthogonal matrices $\mathbf{W}_*$ from Theorem 27, the matrix $\widehat{\mathbf{X}}\mathbf{W}_*^\top - \widetilde{\mathbf{X}}$ admits the decomposition*

$$\widehat{\mathbf{X}}\mathbf{W}_*^\top - \widetilde{\mathbf{X}} = (\mathbf{A} - \mathbf{P})\mathbf{U}_{\mathbf{X}}|\mathbf{\Lambda}_{\mathbf{X}}|^{-1/2}\mathbf{I}_{p,q} + \mathbf{R}$$

where the matrix $\mathbf{R}$ satisfies

$$\|\mathbf{R}\|_F = O\left(\sqrt{\frac{\log(n)}{n\alpha_n}}\right)$$

with high probability. Furthermore, also with high probability,

$$\left|\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{W}_*\|_F^2 - C(\widetilde{\mathbf{X}})^2\right| = O\left(\sqrt{\frac{\log(n)}{n\alpha_n}}\right),$$

where

$$C^2(\mathbf{X}) := \mathbb{E}\|(\mathbf{A} - \mathbf{P})\mathbf{U}|\mathbf{S}|^{-\mathbf{1/2}}\|_F^2.$$

where the expectation is with respect to the randomness in $\mathbf{A}$. Finally, as $n \to \infty$, we have that

$$\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{W}_*\|_F^2 \to Tr\left(\widetilde{\mathbf{Q}}^{-\mathbf{1}}\mathbf{\Delta}^{-\mathbf{1}}\mathbf{\Gamma}\mathbf{\Delta}^{-\mathbf{1}}\widetilde{\mathbf{Q}}^{-\top}\right).$$

almost surely, where Γ is defined via

$$\Gamma := \begin{cases} \mathbb{E}[XX^\top(X^\top\mathbf{I}_{p,q}\mu - X^\top\mathbf{I}_{p,q}\Delta\mathbf{I}_{p,q}X)] & \alpha_n \equiv 1 \\ \mathbb{E}[XX^\top(X^\top\mathbf{I}_{p,q}\mu)] & \alpha_n \to 0. \end{cases}$$

Finally, we present the following functional central limit theorem for the approximation of $\widehat{\mathbf{X}}$ to $\widetilde{\mathbf{X}}$ under sparsity. The proof can be found in Section E.1.4. The result is similar to Theorem 5 in Tang et al. (2017b), but requires a number of different technical considerations.

Lemma 52. *Let* $(\mathbf{A}, \mathbf{X}) \sim GRDPG(F_X, n, \alpha_n)$ *where* $n\alpha_n \gg \log^4(n)$. *Suppose* $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}$ *is a collection of twice continuously differentiable functions. Let* $\widetilde{\mathbf{X}} := \mathbf{U}_{\mathbf{X}}|\mathbf{\Lambda}_{\mathbf{X}}|^{1/2}$, *and let* $\mathbf{W}_*$ *be the matrix guaranteed by Lemma 51. Then, as* $n \to \infty$, *the empirical process*

$$f \in \mathcal{F} \mapsto \widehat{\mathbb{G}}_n f := \sqrt{\frac{\alpha_n}{n}} \sum_{i=1}^{n} \left[f\left(\frac{\mathbf{W}_* \widehat{X}_i}{\sqrt{\alpha_n}} \right) - f\left(\frac{\widetilde{X}_i}{\sqrt{\alpha_n}} \right) \right] \to 0 \tag{E.4}$$

almost surely as $n \to \infty$. *Moreover, with probability at least* $1 - O(n^{-2})$,

$$\sup_{f \in \mathcal{F}} \left| \sqrt{\frac{\alpha_n}{n}} \sum_{i=1}^{n} \left[f\left(\frac{\mathbf{W}_* \widehat{X}_i}{\sqrt{\alpha_n}} \right) - f\left(\frac{\widetilde{X}_i}{\sqrt{\alpha_n}} \right) \right] \right| = O\left(\sqrt{\frac{\log(n)}{n\alpha_n}} \right). \tag{E.5}$$

In addition, the results hold with the replacement $\mathbf{W}_* \widehat{X}_i$ *and* $\widetilde{X}_i$ *replaced with* $\widehat{X}_i$ *and* $\mathbf{W}_*^\top \widetilde{X}_i$ *respectively. Finally, if* $\sqrt{n}\alpha_n = \omega(n^{1/2} \log^{1+\eta}(n))$ *for some* $\eta > 0$ *then the result in Equation E.4 still holds under the scaling* $\frac{1}{\sqrt{n}}$ *instead of* $\frac{\sqrt{\alpha_n}}{\sqrt{n}}$, *and in Equation (E.5) the right hand side bound is of the form* $O\left(\frac{\log^{1/2}(n)}{n^{1/2}\alpha_n} \right)$.

Notation	Definition
$\mathbf{Q_X}, \mathbf{Q_Y} \in \mathbb{O}(p,q)$	The matrices such that $\mathbf{U_X}\lvert\mathbf{\Lambda_X}\rvert^{\mathbf{1/2}}\mathbf{Q_X} = \alpha_n^{1/2}\mathbf{X}$ and $\mathbf{U_Y}\lvert\mathbf{\Lambda_Y}\rvert^{\mathbf{1/2}}\mathbf{Q_Y} = \beta_m^{1/2}\mathbf{Y}$ (Equation E.3)
$\widetilde{\mathbf{Q}}_\mathbf{X}, \widetilde{\mathbf{Q}}_\mathbf{Y} \in \mathbb{O}(p,q)$	The limiting matrices for $\mathbf{Q_X}$ and $\mathbf{Q_Y}$ from Lemma 50 (Equation E.3)
$\mathbf{W}_*^\mathbf{X}, \mathbf{W}_*^\mathbf{Y} \in \mathbb{O}(d) \cap \mathbb{O}(p,q)$	The block-orthogonal matrices aligning $\widehat{\mathbf{X}}$ and $\widetilde{\mathbf{X}}$ (Lemma 51 and Theorem 27)
$\mathbf{W_X}$, $\mathbf{W_Y} \in \mathbb{O}(d) \cap \mathbb{O}(p,q)$	The block-orthogonal matrices aligning $\mathbf{Q_X}$ and $\widetilde{\mathbf{Q}}_\mathbf{X}$ from Lemma 50.

Table E.1: Table of Notation

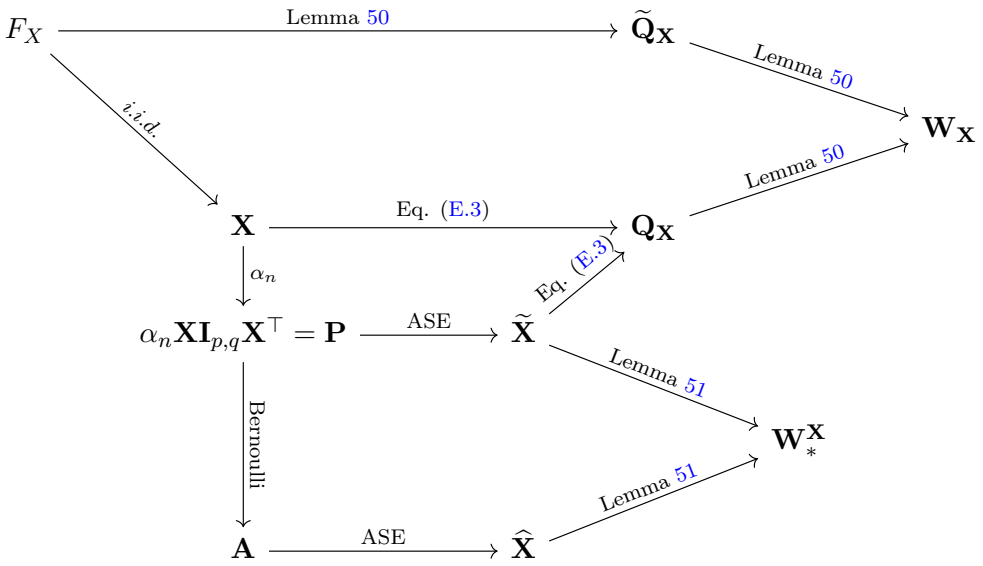


Table E.2: Diagram of the alignment matrices and where they come from. Both $\widetilde{\mathbf{Q}}_\mathbf{X}$ and $\mathbf{W_X}$ come from Lemma 50, whereas the matrix $\mathbf{W}_*^\mathbf{X}$ comes from Lemma 51 (or Theorem 27).

Armed with these technical results, we are now ready to prove Theorems 13 and 14. To make the proof more straightforward, we have compiled the notation for all the alignment matrices in Table E.1. Although the proof mirrors that in Tang et al. (2017b), the steps require careful tabulation of sparsity parameters, orthogonal transformations, and indefinite orthogonal transformations, all of which require novel technical analyses. Table E.2 also shows how to find the various alignment matrices. Essentially, we have the approximations

$$\widehat{\mathbf{X}}\mathbf{W}_*^{\mathbf{X}}\mathbf{W}_{\mathbf{X}} \approx \widetilde{\mathbf{X}}\mathbf{W}_{\mathbf{X}} = \alpha_n^{1/2}\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}} \approx \alpha_n^{1/2}\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}},$$

using Lemmas 50 and 51. Similar approximations hold for $\widehat{\mathbf{Y}}$ and $\widetilde{\mathbf{Y}}$.

Proof of Theorems 13 and 14. Define the matrices $\mathbf{W}_*^{\mathbf{X}}$ and $\mathbf{W}_{\mathbf{X}}$ where $(\mathbf{W}_*^{\mathbf{X}})^\top$ is the orthogonal matrix from Lemma 51, and $\mathbf{W}_{\mathbf{X}}$ is the matrix from Lemma 50. Define $\mathbf{W}_*^{\mathbf{Y}}$ and $\mathbf{W}_{\mathbf{Y}}$ similarly.

First, define $\widetilde{V}_{n,m} := V_{n,m}\left(\frac{\widetilde{\mathbf{X}}\mathbf{W}_{\mathbf{X}}}{\sqrt{\alpha_n}}, \frac{\widetilde{\mathbf{Y}}\mathbf{W}_{\mathbf{Y}}}{\sqrt{\beta_m}}\right)$ via

$$\begin{aligned}
\widetilde{V}_{n,m}\left(\frac{\widetilde{\mathbf{X}}\mathbf{W}_{\mathbf{X}}}{\sqrt{\alpha_n}}, \frac{\widetilde{\mathbf{Y}}\mathbf{W}_{\mathbf{Y}}}{\sqrt{\beta_m}}\right) &= \left\| \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{\mathbf{W}_{\mathbf{X}}^\top \widetilde{X}_i}{\sqrt{\alpha_n}}\right) - \frac{1}{m}\sum_{k=1}^{m}\Phi\left(\frac{\mathbf{W}_{\mathbf{Y}}^\top \widetilde{Y}_k}{\sqrt{\beta_m}}\right)\right\|_{\mathcal{H}}^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\kappa\left(\frac{\mathbf{W}_{\mathbf{X}}^\top \widetilde{X}_i}{\sqrt{\alpha_n}}, \frac{\mathbf{W}_{\mathbf{X}}^\top \widetilde{X}_j}{\sqrt{\alpha_n}}\right) \\
&\quad - \frac{2}{mn}\sum_{i=1}^{n}\sum_{k=1}^{m}\kappa\left(\frac{\mathbf{W}_{\mathbf{X}}^\top \widetilde{X}_i}{\sqrt{\alpha_n}}, \frac{\mathbf{W}_{\mathbf{Y}}^\top \widetilde{Y}_k}{\sqrt{\beta_m}}\right) \\
&\quad + \frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}\kappa\left(\frac{\mathbf{W}_{\mathbf{Y}}^\top \widetilde{Y}_k}{\sqrt{\beta_m}}, \frac{\mathbf{W}_{\mathbf{Y}}^\top \widetilde{Y}_l}{\sqrt{\beta_m}}\right)
\end{aligned}$$

and analogously for $\widehat{V}_{n,m}$. We have the decomposition

$$\begin{aligned}
&(m\beta_m + n\alpha_n)\left(V_{n,m}\left(\frac{\widetilde{\mathbf{X}}\mathbf{W}_{\mathbf{X}}}{\sqrt{\alpha_n}}, \frac{\widetilde{\mathbf{Y}}\mathbf{W}_{\mathbf{Y}}}{\sqrt{\beta_m}}\right) - V_{n,m}\left(\frac{\widehat{\mathbf{X}}\mathbf{W}_*^{\mathbf{X}}\mathbf{W}_{\mathbf{X}}}{\sqrt{\alpha_n}}, \frac{\widehat{\mathbf{Y}}\mathbf{W}_*^{\mathbf{Y}}\mathbf{W}_{\mathbf{Y}}}{\sqrt{\beta_m}}\right)\right) \\
&\qquad = (m\beta_m + n\alpha_n)\left(U_{n,m}\left(\frac{\widetilde{\mathbf{X}}\mathbf{W}_{\mathbf{X}}}{\sqrt{\alpha_n}}, \frac{\widetilde{\mathbf{Y}}\mathbf{W}_{\mathbf{Y}}}{\sqrt{\beta_m}}\right) - U_{n,m}\left(\frac{\widehat{\mathbf{X}}\mathbf{W}_*^{\mathbf{X}}\mathbf{W}_{\mathbf{X}}}{\sqrt{\alpha_n}}, \frac{\widehat{\mathbf{Y}}\mathbf{W}_*^{\mathbf{Y}}\mathbf{W}_{\mathbf{Y}}}{\sqrt{\beta_m}}\right)\right) + r_1 + r_2,
\end{aligned}$$

where

$$r_1 = \frac{(m\beta_m + n\alpha_n)}{n(n-1)} \sum_{i=1}^{n} \left[\kappa\left(\frac{\mathbf{W}_{\mathbf{X}}^\top \widetilde{X}_i}{\sqrt{\alpha_n}}, \frac{\mathbf{W}_{\mathbf{X}}^\top \widetilde{X}_i}{\sqrt{\alpha_n}} \right) - \kappa\left(\frac{(\mathbf{W}_*^{\mathbf{X}} \mathbf{W}_{\mathbf{X}})^\top \widehat{X}_i}{\sqrt{\alpha_n}}, \frac{(\mathbf{W}_*^{\mathbf{X}} \mathbf{W}_{\mathbf{X}})^\top \widehat{X}_i}{\sqrt{\alpha_n}} \right) \right]$$
$$+ \frac{(m\beta_m + n\alpha_n)}{n(n-1)} \sum_{k=1}^{m} \left[\kappa\left(\frac{\mathbf{W}_{\mathbf{Y}}^\top \widetilde{Y}_k}{\sqrt{\beta_m}}, \frac{\mathbf{W}_{\mathbf{Y}}^\top \widetilde{Y}_k}{\sqrt{\beta_m}} \right) - \kappa\left(\frac{(\mathbf{W}_*^{\mathbf{Y}} \mathbf{W}_{\mathbf{Y}})^\top \widehat{Y}_k}{\sqrt{\beta_m}}, \frac{(\mathbf{W}_*^{\mathbf{Y}} \mathbf{W}_{\mathbf{Y}})^\top \widehat{Y}_k}{\sqrt{\beta_m}} \right) \right];$$

$$r_2 = \frac{(m\beta_m + n\alpha_n)}{n^2(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\kappa\left(\frac{\mathbf{W}_{\mathbf{X}}^\top \widetilde{X}_i}{\sqrt{\alpha_n}}, \frac{\mathbf{W}_{\mathbf{X}}^\top \widetilde{X}_j}{\sqrt{\alpha_n}} \right) - \kappa\left(\frac{(\mathbf{W}_*^{\mathbf{X}} \mathbf{W}_{\mathbf{X}})^\top \widehat{X}_i}{\sqrt{\alpha_n}}, \frac{(\mathbf{W}_*^{\mathbf{X}} \mathbf{W}_{\mathbf{X}})^\top \widehat{X}_j}{\sqrt{\alpha_n}} \right) \right]$$
$$+ \frac{(m\beta_m + n\alpha_n)}{m^2(m-1)} \sum_{k=1}^{m} \sum_{l=1}^{m} \left[\kappa\left(\frac{\mathbf{W}_{\mathbf{Y}}^\top \widetilde{Y}_k}{\sqrt{\beta_m}}, \frac{\mathbf{W}_{\mathbf{Y}}^\top \widetilde{Y}_l}{\sqrt{\beta_m}} \right) - \kappa\left(\frac{(\mathbf{W}_*^{\mathbf{Y}} \mathbf{W}_{\mathbf{Y}})^\top \widehat{Y}_k}{\sqrt{\beta_m}}, \frac{(\mathbf{W}_*^{\mathbf{Y}} \mathbf{W}_{\mathbf{Y}})^\top \widehat{Y}_l}{\sqrt{\beta_m}} \right) \right].$$

By Theorem 26, $\bar{\mathbf{\Omega}}$ is compact, so by the fact that κ is twice continuously differentiable, κ is Lipschitz on $\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}} \bar{\mathbf{\Omega}}$ since $\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}} \bar{\mathbf{\Omega}}$ is compact. In particular, for some positive K, we have that

$$\begin{aligned}
\|r_1\| &\leq K \frac{m\beta_m + n\alpha_n}{n-1} \max_i \left\| \frac{\mathbf{W}_{\mathbf{X}}^\top \widetilde{X}_i}{\sqrt{\alpha_n}} - \frac{(\mathbf{W}_*^{\mathbf{X}} \mathbf{W}_{\mathbf{X}})^\top \widehat{X}_i}{\sqrt{\alpha_n}} \right\| \\
&\quad + K \frac{m\beta_m + n\alpha_n}{m-1} \max_i \left\| \frac{\mathbf{W}_{\mathbf{Y}}^\top \widetilde{Y}_i}{\sqrt{\alpha_n}} - \frac{(\mathbf{W}_*^{\mathbf{Y}} \mathbf{W}_{\mathbf{Y}})^\top \widehat{Y}_i}{\sqrt{\alpha_n}} \right\| \\
&\leq 2K \frac{m\beta_m + n\alpha_n}{n\alpha_n} \sqrt{\alpha_n} \frac{\log(n)}{\sqrt{n\alpha_n}} + 2K \frac{m\beta_m + n\alpha_n}{m\beta_m} \sqrt{\beta_m} \frac{\log(m)}{\sqrt{m\beta_m}} \\
&\leq C\left(\frac{\log(n)}{\sqrt{n}} + \frac{\log(m)}{\sqrt{m}} \right),
\end{aligned}$$

by the assumption that $\frac{m\beta_m}{m\beta_m + n\alpha_n} \to \rho \in (0,1)$ and the $2,\infty$ bound in Theorem 27. By a similar argument,

$$\begin{aligned}
\|r_2\| &\leq K \frac{m\beta_m + n\alpha_n}{(n-1)} \max_i \left\| \frac{\mathbf{W}_{\mathbf{X}}^\top \widetilde{X}_i}{\sqrt{\alpha_n}} - \frac{(\mathbf{W}_*^{\mathbf{X}} \mathbf{W}_{\mathbf{X}})^\top \widehat{X}_i}{\sqrt{\alpha_n}} \right\| \\
&\quad + K \frac{m\beta_m + n\alpha_n}{m-1} \max_i \left\| \frac{\mathbf{W}_{\mathbf{Y}}^\top \widetilde{Y}_i}{\sqrt{\alpha_n}} - \frac{(\mathbf{W}_*^{\mathbf{Y}} \mathbf{W}_{\mathbf{Y}})^\top \widehat{Y}_i}{\sqrt{\alpha_n}} \right\| \\
&\leq C\left(\frac{\log(n)}{\sqrt{n}} + \frac{\log(m)}{\sqrt{m}} \right).
\end{aligned}$$

Both of these bounds are independent of α_n and β_m. Define

$$\xi := \frac{\sqrt{m\beta_m + n\alpha_n}}{n} \sum_{i=1}^{n} \kappa(\mathbf{W}_{\mathbf{X}}^{\top} \widetilde{X}_i / \sqrt{\alpha_m}, \cdot) - \frac{\sqrt{m\beta_m + n\alpha_n}}{m} \sum_{k=1}^{m} \kappa(\mathbf{W}_{\mathbf{Y}}^{\top} \widetilde{Y}_k / \sqrt{\beta_m}, \cdot)$$
$$\widehat{\xi} := \frac{\sqrt{m\beta_m + n\alpha_n}}{n} \sum_{i=1}^{n} \kappa((\mathbf{W}_*^{\mathbf{X}} \mathbf{W}_{\mathbf{X}})^{\top} \widehat{X}_i / \sqrt{\alpha_m}, \cdot) - \frac{\sqrt{m\beta_m + n\alpha_n}}{m} \sum_{k=1}^{m} \kappa((\mathbf{W}_*^{\mathbf{Y}} \mathbf{W}_{\mathbf{Y}})^{\top} \widehat{Y}_k / \sqrt{\beta_m}, \cdot).$$

We now have that

$$\begin{aligned}
&\left| (m\beta_m + n\alpha_n) \left(V_{n,m} \left(\frac{\widetilde{\mathbf{X}} \mathbf{W}_*^{\mathbf{X}} \mathbf{W}_{\mathbf{X}}}{\sqrt{\alpha_n}}, \frac{\widetilde{\mathbf{Y}} \mathbf{W}_*^{\mathbf{Y}} \mathbf{W}_{\mathbf{Y}}}{\sqrt{\beta_m}} \right) - V_{n,m} \left(\frac{\widehat{\mathbf{X}} \mathbf{W}_{\mathbf{X}}}{\sqrt{\alpha_n}}, \frac{\widehat{\mathbf{Y}} \mathbf{W}_{\mathbf{Y}}}{\sqrt{\beta_m}} \right) \right) \right| \\
&\qquad = \left| \|\xi\|_{\mathcal{H}}^2 - \|\widehat{\xi}\|_{\mathcal{H}}^2 \right| \\
&\qquad \leq \|\xi - \widehat{\xi}\|_{\mathcal{H}} \left(2\|\xi\|_{\mathcal{H}} + \|\xi - \widehat{\xi}\|_{\mathcal{H}} \right).
\end{aligned}$$

We have that

$$\begin{aligned}
\xi - \widehat{\xi} &= \sqrt{\frac{m\beta_m + n\alpha_n}{n}} \sum_{i=1}^{n} \frac{\kappa(\mathbf{W}_{\mathbf{X}}^{\top} (\mathbf{W}_*^{\mathbf{X}})^{\top} \widehat{X}_i / \sqrt{\alpha_n}, \cdot) - \kappa(\mathbf{W}_{\mathbf{X}}^{\top} \widetilde{X}_i / \sqrt{\alpha_n}, \cdot)}{\sqrt{n}} \\
&\quad - \sqrt{\frac{m\beta_m + n\alpha_n}{n}} \sum_{i=1}^{n} \frac{\kappa(\mathbf{W}_{\mathbf{Y}}^{\top} (\mathbf{W}_*^{\mathbf{Y}})^{\top} \widehat{Y}_i / \sqrt{\beta_m}, \cdot) - \kappa((\mathbf{W}_{\mathbf{Y}}^{\top} \widetilde{Y}_i / \sqrt{\beta_m}, \cdot)}{\sqrt{n}} \\
&:= \zeta_X - \zeta_Y.
\end{aligned}$$

We note that since κ is radial, we can disregard $\mathbf{W}_{\mathbf{X}}^{\top}$ and $\mathbf{W}_{\mathbf{Y}}^{\top}$. Moreover,

$$\begin{aligned}
\|\zeta_X\| &= \left\| \sqrt{\frac{m\beta_m + n\alpha_n}{n}} \sum_{i=1}^{n} \frac{\kappa((\mathbf{W}_*^{\mathbf{X}})^{\top} \widehat{X}_i / \sqrt{\alpha_n}, \cdot) - \kappa(\widetilde{X}_i / \sqrt{\alpha_n}, \cdot)}{\sqrt{n}} \right\| \\
&= \left\| \sqrt{\frac{m\beta_m + n\alpha_n}{n\alpha_n}} \sqrt{\alpha_n} \sum_{i=1}^{n} \frac{\kappa((\mathbf{W}_*^{\mathbf{X}})^{\top} \widehat{X}_i / \sqrt{\alpha_n}, \cdot) - \kappa(\widetilde{X}_i / \sqrt{\alpha_n}, \cdot)}{\sqrt{n}} \right\| \\
&= \sqrt{\frac{m\beta_m + n\alpha_n}{n\alpha_n}} \left\| \sqrt{\alpha_n} \sum_{i=1}^{n} \frac{\kappa((\mathbf{W}_*^{\mathbf{X}})^{\top} \widehat{X}_i / \sqrt{\alpha_n}, \cdot) - \kappa(\widetilde{X}_i / \sqrt{\alpha_n}, \cdot)}{\sqrt{n}} \right\|.
\end{aligned}$$

Since $m\beta_m/(m\beta_m + n\alpha_n) \to \rho \in (0,1)$ the term outside of the norm is of order $O(1)$. Lemma 52 then implies that the term inside of the norm tends to zero, and the same argument holds

for ζ_Y. In particular, we see that by Lemma 52,

$$\|\xi - \widehat{\xi}\|_{\mathcal{H}} = O\bigg(\sqrt{\frac{\log(n)}{n\alpha_n}} + \sqrt{\frac{\log(m)}{m\beta_m}}\bigg) \tag{E.6}$$

with probability at least $1 - O(n^{-2} + m^{-2})$

We now bound $\|\xi\|_{\mathcal{H}}$ under the null and alternative respectively. Recall that $\widetilde{\mathbf{Q}}_{\mathbf{X}}$ is the limiting matrix guaranteed by Lemma 50. Note further that $\widetilde{\mathbf{X}}/\sqrt{\alpha_n} = \mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}$ so that $\widetilde{\mathbf{X}}\mathbf{W}_{\mathbf{X}}/\sqrt{\alpha_n} = \mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}}$, where $\mathbf{W}_{\mathbf{X}}$ was the matrix such that $\mathbf{Q}_{\mathbf{X}} - \mathbf{W}_{\mathbf{X}}\widetilde{\mathbf{Q}}_{\mathbf{X}}$ is of order $\sqrt{\log(n)/n}$ with probability at least $1 - n^{-2}$. Hence, from the equation

$$\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}} = \mathbf{I}_{p,q}\mathbf{Q}_{\mathbf{X}}^{\top}\mathbf{I}_{p,q},$$

we see that with probability at least $1 - n^{-2}$ that

$$\begin{aligned}
\|\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}} - \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\| &= \|\mathbf{I}_{p,q}\mathbf{Q}_{\mathbf{X}}^{\top}\mathbf{I}_{p,q}\mathbf{W}_{\mathbf{X}} - \mathbf{I}_{p,q}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{\top}\mathbf{I}_{p,q}\| \\
&= \|\mathbf{Q}_{\mathbf{X}}^{\top}\mathbf{I}_{p,q}\mathbf{W}_{\mathbf{X}} - \widetilde{\mathbf{Q}}_{\mathbf{X}}^{\top}\mathbf{I}_{p,q}\| \\
&= \|\mathbf{Q}_{\mathbf{X}}^{\top}\mathbf{W}_{\mathbf{X}}\mathbf{I}_{p,q} - \widetilde{\mathbf{Q}}_{\mathbf{X}}^{\top}\mathbf{I}_{p,q}\| \\
&= \|\mathbf{Q}_{\mathbf{X}}^{\top}\mathbf{W}_{\mathbf{X}} - \widetilde{\mathbf{Q}}_{\mathbf{X}}^{\top}\| \\
&= \|\mathbf{W}_{\mathbf{X}}^{\top}\mathbf{Q}_{\mathbf{X}} - \widetilde{\mathbf{Q}}_{\mathbf{X}}\| \\
&= \|\mathbf{Q}_{\mathbf{X}} - \mathbf{W}_{\mathbf{X}}\widetilde{\mathbf{Q}}_{\mathbf{X}}\| \\
&= O\bigg(\frac{\sqrt{\log(n)}}{\sqrt{n}}\bigg).
\end{aligned}$$

Hence, we have that

$$
\begin{aligned}
\xi &= \frac{\sqrt{m\beta_m + n\alpha_n}}{n}\sum_{i=1}^{n}\kappa(\mathbf{W}_{\mathbf{X}}^{\top}\widetilde{X}_i/\sqrt{\alpha_m},\cdot) - \frac{\sqrt{m\beta_m + n\alpha_n}}{m}\sum_{k=1}^{m}\kappa(\mathbf{W}_{\mathbf{Y}}^{\top}\widetilde{Y}_k/\sqrt{\beta_m},\cdot) \\
&= \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{n}}\sum_{i=1}^{n}\frac{\kappa((\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}})^{\top}X_i,\cdot)}{\sqrt{n}} - \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{m}}\sum_{k=1}^{m}\frac{\kappa((\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{Y}})^{\top}Y_i,\cdot)}{\sqrt{m}} \\
&= \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{n}}\sum_{i=1}^{n}\frac{\kappa((\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}})^{\top}X_i,\cdot) - \kappa(\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\top}X_i,\cdot)}{\sqrt{n}} \\
&\quad - \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{m}}\sum_{k=1}^{m}\frac{\kappa((\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{Y}})^{\top}Y_i,\cdot) - \kappa(\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\top}X_Y,\cdot)}{\sqrt{m}} \\
&\quad + \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{n}}\sum_{i=1}^{n}\frac{\kappa(\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\top}X_i,\cdot)}{\sqrt{n}} - \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{m}}\sum_{k=1}^{m}\frac{\kappa(\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\top}Y_k,\cdot)}{\sqrt{m}} \\
&= \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{n}}\sum_{i=1}^{n}\frac{\kappa((\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}})^{\top}X_i,\cdot) - \kappa(\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\top}X_i,\cdot)}{\sqrt{n}} \\
&\quad - \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{m}}\sum_{k=1}^{m}\frac{\kappa((\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{Y}})^{\top}Y_i,\cdot) - \kappa(\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\top}X_Y,\cdot)}{\sqrt{m}} \\
&\quad + \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{n}}\sum_{i=1}^{n}\frac{\kappa(\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\top}X_i,\cdot) - \mu[F_X\circ\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\top}]}{\sqrt{n}} - \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{m}}\sum_{k=1}^{m}\frac{\kappa(\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\top}Y_k,\cdot) - \mu[F_Y\circ\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\top}]}{\sqrt{m}} \\
&\quad + \sqrt{m\beta_m + n\alpha_n}\mu[F_X\circ\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\top}] - \sqrt{m\beta_m + n\alpha_n}\mu[F_Y\circ\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\top}]
\end{aligned}
$$

For the first two terms, by the Lipschitz property of κ and the fact that $\bar{\mathbf{\Omega}}$ is bounded by Theorem 26, we have that

$$
\begin{aligned}
\frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{n}}\sum_{i=1}^{n}\frac{\kappa((\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}})^{\top}X_i,\cdot) - \kappa(\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\top}X_i,\cdot)}{\sqrt{n}} &\leq \frac{\sqrt{m\beta_m + n\alpha_n}}{\sqrt{n}}\sum_{i=1}^{n}\frac{\|(\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}})^{\top}X_i - \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\top}X_i\|}{\sqrt{n}} \\
&\leq \sqrt{m\beta_m + n\alpha_n}\|\max_i\|X_i(\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}}) - \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}})\| \\
&\leq \|\mathbf{X}\|_{2,\infty}\sqrt{m\beta_m + n\alpha_n}\|\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}} - \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\| \\
&\leq C\sqrt{m\beta_m + n\alpha_n}\frac{\sqrt{\log(n)}}{\sqrt{n}} \\
&= O\left(\sqrt{\alpha_n\log(n)}\right)
\end{aligned}
$$

with probability at least $1 - O(n^{-2})$ by Lemma 50. Hence the first term is of order

$$O\bigg(\sqrt{\alpha_n \log(n)} + \sqrt{\beta_m \log(m)}\bigg)$$

with probability at least $1 - O(n^{-2} + m^{-2})$.

Now, define

$$\psi_X := \sum_{i=1}^n \frac{\kappa(\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}} X_i, \cdot) - \mu(F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}})}{n}$$
$$\psi_Y := \sum_{i=1}^n \frac{\kappa(\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}} Y_i, \cdot) - \mu(F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}})}{m}$$

By Remark 1 in Schneider (2016), we see that

$$\mathbb{P}(\|\psi_X\|^2 > \varepsilon^2) \leq 2 \exp\left[\frac{-n\varepsilon^2}{64}\right]$$

which in particular shows that $\|\psi_X\| \leq \frac{C\sqrt{\log(n)}}{\sqrt{n}}$ with probability at least $1 - O(n^{-2})$, and similarly for ψ_Y. Thus far, we have shown with probability at least $1 - O(n^{-2} + m^{-2})$ that

$$\|\xi\|_{\mathcal{H}} = O\bigg(\sqrt{\alpha_n \log(n)} + \sqrt{\beta_m \log(m)}\bigg) + \sqrt{m\beta_m + n\alpha_n}\bigg(\|\mu[F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\top}] - \mu[F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\top}]\|_{\mathcal{H}}\bigg).$$

Under the null hypothesis, the term in the parentheses is zero as $\widetilde{\mathbf{Q}}_{\mathbf{Y}}$ can be chosen to be $\mathbf{T}^{-\mathbf{1}}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}$ by Lemma 50 and the fact that $\widetilde{\mathbf{Q}}_{\mathbf{Y}}$ and $\widetilde{\mathbf{Q}}_{\mathbf{X}}$ do not depend on the sparsity factors by Lemma 50. Hence, we see that

$$\|\xi - \widehat{\xi}\|\bigg(2\|\xi\| + \|\xi - \widehat{\xi}\|\bigg) = O\bigg(\sqrt{\frac{\log(n)}{n\alpha_n}} + \sqrt{\frac{\log(m)}{m\beta_m}}\bigg)\bigg(\sqrt{\alpha_n \log(n) + \beta_m \log(m)}\bigg).$$

Under the alternative the term $\|\mu[F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\top}] - \mu[F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\top}]\|_{\mathcal{H}}$ is not necessarily zero, so we have that

$$\begin{aligned}\|\xi - \widehat{\xi}\|\bigg(2\|\xi\| + \|\xi - \widehat{\xi}\|\bigg) &= O\bigg(\sqrt{\frac{\log(n)}{n\alpha_n}} + \sqrt{\frac{\log(m)}{m\beta_m}}\bigg)\bigg(\sqrt{\alpha_n \log(n) + \beta_m \log(m)} + \sqrt{n\alpha_n + m\beta_m}\bigg) \\ &= O\bigg(\sqrt{\log(n)} + \sqrt{\log(m)}\bigg).\end{aligned}$$

Hence, dividing by $\log(n)$ yields the result.

□

We immediately derive the proof of Corollary 5.

Proof of Corollary 5. We will highlight where the previous proof changes. Examining the proof of Theorems 13 and 14, we see that we have (under the new scaling $(m+n)$) the residual bounds

$$r_1 = O\left(\frac{\log(n)}{\sqrt{n\alpha_n}} + \frac{\log(m)}{\sqrt{m\beta_m}}\right).$$

and similarly for r_2. Furthermore, by the final statement in Lemma 52, we have that

$$\|\xi - \widehat{\xi}\|_{\mathcal{H}} \leq O\left(\frac{\sqrt{\log(n)}}{\sqrt{n\alpha_n}} + \frac{\sqrt{\log(m)}}{\sqrt{m\beta_m}}\right).$$

Under the null hypothesis, through a similar analysis, we have that

$$\|\xi\|_{\mathcal{H}} = O\left(\sqrt{\log(n) + \log(m)}\right).$$

Therefore, with probability $1 - O(n^{-2} + m^{-2})$, we have the bound

$$\begin{aligned}\|\xi - \widehat{\xi}\|\left(2\|\xi\| + \|\xi - \widehat{\xi}\|\right) &= O\left(\frac{\sqrt{\log(n)}}{\sqrt{n\alpha_n}} + \frac{\sqrt{\log(m)}}{\sqrt{m\beta_m}}\right)\left(\sqrt{\log(n)} + \sqrt{\log(m)}\right) \\ &= O\left(\frac{\log(n)}{\sqrt{n\alpha_n}} + \frac{\log(m)}{\sqrt{m\beta_m}}\right)\end{aligned}$$

since $m/(n+m) \to \rho \in (0,1)$. Therefore, since $\min(\alpha_n, \beta_m) \geq n^{-1/2}\log^{1+\eta}(n)$ for some $\eta > 0$, the right hand side tends to zero. □

In order to prove Corollary 6, we will need the following additional lemmas. The first is straightforward and included in Section E.1.5 for completeness.

Lemma 53. *When $\mathbb{E}(X_1^\top \mathbf{I}_{p,q} X_2) = 1$, we have with probability at least $1 - O(n^{-2})$ that*

$$\frac{1}{\sqrt{\widehat{\alpha}_n}} - \frac{1}{\sqrt{\alpha_n}} = O\left(\frac{\sqrt{\log(n)}}{n^{1/2}\alpha_n}\right).$$

The next lemma shows we can replace α_n with $\widehat{\alpha}_n$ in the appropriate place and still maintain convergence in probability. The proof is in Section E.1.3 and is simply a modification of the proof of Lemma 52.

Lemma 54. *Under the setting of Lemma 52, the limiting result holds with $\widehat{X}_i/\alpha_n^{1/2}$ replaced with $\widehat{X}_i/\widehat{\alpha}_n^{1/2}$ with the almost sure convergence replaced with convergence in probability.*

Proof of Corollary 6. Now, the result follows by simply noting that the functional CLT still holds in probability, and hence the result holds with $\widehat{\mathbf{X}}/\alpha_n^{1/2}$ and $\widehat{\mathbf{Y}}/\beta_m^{1/2}$ replaced with $\widehat{\mathbf{X}}/\widehat{\alpha}_n^{1/2}$ and $\widehat{\mathbf{Y}}/\widehat{\beta}_m^{1/2}$.

□

E.1.2 Proofs of Propositions

In this section we prove Proposisions 5, 6, and 8.

Proof of Proposition 5

We need the following generalization of Lemma 49 to the repeated eigenvalues setting. The proof is also straightforward and postponed to Section E.1.5.

Lemma 55. *Let $\mathbf{Q_X}$ be defined as above, and $\mathbf{Q_Y}$ similarly. Then there exist deterministic matrices $\widetilde{\mathbf{Q}}_\mathbf{X}$ and $\widetilde{\mathbf{Q}}_\mathbf{Y}$ and sequences of orthogonal matrices $\mathbf{W_X}$ and $\mathbf{W_Y}$ such that $\mathbf{Q_X W_X} - \widetilde{\mathbf{Q}}_\mathbf{X} \to 0$ and similarly for $\mathbf{Q_Y}$ and $\mathbf{W_Y}$. In particular, under the null hypothesis there exists some $\mathbf{T}$ such that $F_X \circ \mathbf{T} = F_Y$, in which case $\widetilde{\mathbf{Q}}_\mathbf{Y}$ can be chosen such that $\widetilde{\mathbf{Q}}_\mathbf{Y} = \widetilde{\mathbf{Q}}_\mathbf{X}\mathbf{T}$.*

Proof of Proposition 5. We will show the result holds for any two sequences of block-orthogonal matrices $\mathbf{W}_n^1$ and $\mathbf{W}_n^2$, since the result follows by taking $\mathbf{W}_n := \mathbf{W}_n^1(\mathbf{W}_n^2)^\top$. First, by a similar argument to the proof of Proposition 4, we note that it suffices to prove the result with the replacement $\mathbf{XQ_X^{-1}}$ and $\mathbf{YQ_Y^{-1}}$ instead of $\widehat{\mathbf{X}}/\alpha_n^{1/2}$ and $\widehat{\mathbf{Y}}/\beta_m^{1/2}$, since the $2 \to \infty$ result in Theorem 27 and the Lipschitz property of κ shows that

$$|U_{n,m}(\widehat{\mathbf{X}}\mathbf{W}_*^{\mathbf{X}}\mathbf{W}_n^1/\alpha_n^{1/2}, \widehat{\mathbf{Y}}\mathbf{W}_*^{\mathbf{Y}}\mathbf{W}_n^2/\beta_m^{1/2}) - U_{n,m}(\mathbf{XQ_X^{-1}W}_n^1, \mathbf{YQ_Y^{-1}W}_n^2)| \to 0$$

for any sequence of block-orthogonal matrices $\mathbf{W}_n^1$ and $\mathbf{W}_n^2$.

We prove under the alternative first; that is, suppose $F_X \not\approx F_Y$. Suppose that

$$\liminf U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_n^1, \mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_n^2) = 0$$

for some sequence of block-orthogonal matrices $\mathbf{W}_n^1, \mathbf{W}_n^2$. By passing to a convergent subsequence, we may assume the limit exists. Let $\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}$ and $\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}$ be the limiting matrices given by Lemma 55, and similarly for the sequences of block orthogonal matrices $\mathbf{W}^{\mathbf{X}}$ and $\mathbf{W}^{\mathbf{Y}}$. We have that

$$\begin{aligned}
|U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_n^1, \mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_n^2)| &= |U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}}(\mathbf{W}_{\mathbf{X}})^\top\mathbf{W}_n^1, \mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{Y}}(\mathbf{W}_{\mathbf{Y}})^\top\mathbf{W}_n^2)| \\
&\leq \Big|U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}}(\mathbf{W}_{\mathbf{X}})^\top\mathbf{W}_n^1, \mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{Y}}(\mathbf{W}_{\mathbf{Y}})^\top\mathbf{W}_n^2) \\
&\qquad - U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{X}})^\top\mathbf{W}_n^1, \mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{Y}})^\top\mathbf{W}_n^2)\Big| \\
&\qquad + |U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{X}})^\top\mathbf{W}_n^1, \mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{Y}})^\top\mathbf{W}_n^2)|. \quad \text{(E.7)}
\end{aligned}$$

We note that

$$\Big|U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{X}}(\mathbf{W}_{\mathbf{X}})^\top\mathbf{W}_n^1, \mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_{\mathbf{Y}}(\mathbf{W}_{\mathbf{Y}})^\top\mathbf{W}_n^2) - U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{X}})^\top\mathbf{W}_n^1, \mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{Y}})^\top\mathbf{W}_n^2)\Big|$$

tends to zero by Lemma 50. Therefore, we see that

$$\limsup |U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_n^1, \mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_n^2)| \leq \liminf |U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{X}})^\top\mathbf{W}_n^1, \mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{Y}})^\top\mathbf{W}_n^2)|$$

where by assumption the right hand side is presumed to exist. Note that the only terms on the left hand side that are random are the matrices $\mathbf{X}$ and $\mathbf{Y}$ whose rows are drawn i.i.d. F_X and F_Y respectively. If the term on the right hand side tends to zero (which it does by assumption), then that implies that

$$\limsup |U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_n^1, \mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_n^2)| \to 0.$$

This implies that

$$\|\mu[F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_n^1] - \mu[F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_n^2]\|_{\mathcal{H}}^2 \to 0,$$

by, e.g. Remark 1 in Schneider (2016) (as in the proof of Theorems 13 and 14). The only quantities that are changing in n and m are $\mathbf{W}_n^1$ and $\mathbf{W}_n^2$. Define the map

$$\mathbf{W} \mapsto \|\mu[F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}] - \mu[F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}]\|_{\mathcal{H}}^2. \tag{E.8}$$

By Lemma 6 in Gretton et al. (2012), we have that for any two distributions $X \sim F$ and $Y \sim G$ that

$$\|\mu[F] - \mu[G \circ \mathbf{W}_1]\|_{\mathcal{H}}^2 - \|\mu[F] - \mu[G \circ \mathbf{W}_2]\|_{\mathcal{H}}^2 = 2\mathbb{E}_{F,G}\left[\kappa(X, \mathbf{W}_1^\top Y) - \kappa(X, \mathbf{W_2}^\top Y)\right] + \mathbb{E}_G\left[\kappa(\mathbf{W}_1^\top Y, \mathbf{W}_1^\top Y') - \kappa(\mathbf{W}_2^\top Y, \mathbf{W_2}^\top Y')\right]$$

by the definition that $Y \sim G \circ \mathbf{W}$ if $\mathbf{W}^\top Y \sim G$. Hence, continuity of the map in (E.8) follows from continuity of κ. Therefore, by the assumption that κ is radial, we see that since

$$\|\mu[F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}] - \mu[F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_n^2(\mathbf{W}_n^1)^\top]\|_{\mathcal{H}}^2 \to 0,$$

we must have that some subsequence of $\mathbf{W}_n^2(\mathbf{W}_n^1)^\top$ is converging (since the set $\mathbb{O}(p,q) \cap \mathbb{O}(d)$ is compact). Let $\widetilde{\mathbf{W}}$ be this subsequential limit. Then this implies that

$$\|\mu[F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}] - \mu[F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\widetilde{\mathbf{W}}]\|_{\mathcal{H}}^2 = 0.$$

However, under the alternative, since κ is characteristic, we have that $\mu[F_X] \neq \mu[F_Y \circ \mathbf{T}]$ for any $\mathbf{T} \in \mathbb{O}(p,q)$. But then the above equation is a contradiction. Furthermore, working

backwards, we have the chain of inequalities

$$\begin{aligned}\inf_{\mathbf{W}\in\mathbb{O}(d)\cap\mathbb{O}(p,q)} \|\mu[F_X\circ\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}]-\mu[F_Y\circ\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}]\|_{\mathcal{H}}^2 &\leq \liminf |U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_n^1,\mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_n^2)|\\ &\leq \limsup |U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}\mathbf{W}_n^1,\mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}_n^2)|\\ &\leq \liminf |U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{X}})^\top\mathbf{W}_n^1,\mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{Y}})^\top\mathbf{W}_n^2)|,\end{aligned}$$

which shows that

$$C := \inf_{\mathbf{W}\in\mathbb{O}(d)\cap\mathbb{O}(p,q)} \|\mu[F_X\circ\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}]-\mu[F_Y\circ\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}\mathbf{W}]\|_{\mathcal{H}}^2$$

is a lower bound independent of the particular sequence $\mathbf{W}_n^1,\mathbf{W}_n^2$. This proves the second assertion.

Now, suppose the null hypothesis holds. Then, let $\mathbf{W}_n^1=\mathbf{W}_n^2=\mathbf{I}$ above. By similar manipulations as in (E.7), we have that

$$\begin{aligned}|U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{X}})^\top,\mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{Y}})^\top)| &\leq |U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}},\mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}})-U_{n,m}(\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{X}})^\top,\mathbf{Y}\mathbf{Q}_{\mathbf{Y}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{Y}})^\top)|\\ &\quad+|U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}},\mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}})|.\end{aligned}$$

Again, the first term tends to zero by Lemma 50. We argue the second term tends to zero, and hence the result follows. We note that by Lemma 55, we have that $\widetilde{\mathbf{Q}}_{\mathbf{Y}}=\widetilde{\mathbf{Q}}_{\mathbf{X}}\mathbf{T}$, where $\mathbf{T}$ is such that $F_X\circ\mathbf{T}=F_Y$. But then

$$|U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}},\mathbf{Y}\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}})| = |U_{n,m}(\mathbf{X}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}},\mathbf{Y}\mathbf{T}^{-\mathbf{1}}\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}})|,$$

and since $F_X\circ\mathbf{T}=F_Y$, we also have that $F_X\circ\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-1}=F_Y\circ\mathbf{T}^{-1}{\mathbf{Q}_{\mathbf{X}}}^{-1}$, and hence the left hand side tends to zero.

□

Proofs of Propositions 6 and 7

The proofs of these propositions are similar to Propositions 4 and 5 but require analysis of Wasserstein distances.

Proof of Proposition 6. Suppose first that the sparsity factors α_n and β_m are known. By Lemmas 55 and 50, we have that under the null hypothesis there exist sequences of orthogonal matrices $\mathbf{W_X}$ and $\mathbf{W_Y}$ such that with probability at least $1 - n^{-2} - m^{-2}$

$$\|\mathbf{Q_X} - \mathbf{W_X}\widetilde{\mathbf{Q}}\| = O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right); \qquad \|\mathbf{Q_Y T^{-1}} - \mathbf{W_Y}\widetilde{\mathbf{Q}}\| = O\left(\frac{\sqrt{\log(m)}}{\sqrt{m}}\right). \tag{E.9}$$

Furthermore, by Theorem 27, there exists block-orthogonal $\mathbf{W}_*^{\mathbf{X}}$ and $\mathbf{W}_*^{\mathbf{Y}}$ (depending on n and m) such that with probability $1 - n^{-2} - m^{-2}$

$$\|\widehat{\mathbf{X}} - \alpha_n^{1/2}\mathbf{X Q_X^{-1}} \mathbf{W}_*^{\mathbf{X}}\|_{2\to\infty} = O\Big(\frac{\log(n)}{\sqrt{n}}\Big); \qquad \|\widehat{\mathbf{Y}} - \beta_m^{1/2}\mathbf{Y Q_Y^{-1}} \mathbf{W}_*^{\mathbf{Y}}\|_{2\to\infty} = O\Big(\frac{\log(m)}{\sqrt{m}}\Big) \tag{E.10}$$

Define the event $\mathcal{A} := \{$(E.9) and (E.10) hold.$\}$ Note that $\mathbb{P}(\mathcal{A}) \geq 1 - 2n^{-2} - 2m^{-2}$. Note that on $\mathcal{A}$ we have that

$$\|\widehat{\mathbf{X}} - \alpha_n^{1/2}\mathbf{X}\widetilde{\mathbf{Q}}^{\mathbf{-1}}\mathbf{W}_{\mathbf{X}}^{\top}\mathbf{W}_*^{\mathbf{X}}\|_{2,\infty} = O\Big(\frac{\log(n)}{\sqrt{n}}\Big); \qquad \|\widehat{\mathbf{Y}} - \beta_m^{1/2}\mathbf{Y T^{-1}}\widetilde{\mathbf{Q}}^{\mathbf{-1}}\mathbf{W}_{\mathbf{Y}}^{\top}\mathbf{W}_*^{\mathbf{Y}}\|_{2\to\infty} = O\Big(\frac{\log(m)}{\sqrt{m}}\Big).$$

Define $\Gamma_{\widehat{X},\widehat{Y}}$ to be the set of couplings of $\widehat{F}_{\widehat{X}/\alpha_n^{1/2}}$ and $\widehat{F}_{\widehat{Y}/\beta_m^{1/2}}$. We have that for any block-orthogonal $\mathbf{W}_n$, the minimizer $\widehat{\mathbf{W}}_n$ satisfies

$$d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}}, \widehat{F}_{\widehat{Y}/\beta_m^{1/2}} \circ \widehat{\mathbf{W}}_n) \leq d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}}, \widehat{F}_{\widehat{Y}/\beta_m^{1/2}} \circ \mathbf{W}_n)$$

To show this tends to zero, we choose an appropriate block-orthogonal matrix $\mathbf{W}_n$. Define

$$\mathbf{W}_n := (\mathbf{W}_*^{\mathbf{X}}\mathbf{W_X})^{\top}(\mathbf{W_Y}\mathbf{W}_*^{\mathbf{Y}}).$$

Note that under the null hypothesis $F_X \simeq F_Y$ we have that $F_X \circ \widetilde{\mathbf{Q}}^{\mathbf{-1}} = F_Y \circ \mathbf{T^{-1}} \circ \widetilde{\mathbf{Q}}^{-1}$ for

some $\mathbf{T} \in \mathbb{O}(p, q)$. Then, by the rotational invariance of the Euclidean norm, we have that,

$$\begin{aligned} d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}} \circ \mathbf{W}_n, \widehat{F}_{\widehat{Y}/\beta_m^{1/2}}) &= d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}} \circ (\mathbf{W}_{\mathbf{X}}^\top \mathbf{W}_*^{\mathbf{X}})^\top, \widehat{F}_{\widehat{Y}/\beta_m^{1/2}} \circ (\mathbf{W}_{\mathbf{Y}} \mathbf{W}_*^{\mathbf{Y}})^\top) \\ &\leq d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}} \circ (\mathbf{W}_{\mathbf{X}}^\top \mathbf{W}_*^{\mathbf{X}})^\top, \widehat{F}_X \circ \widetilde{\mathbf{Q}}^{-1}) \\ &\quad + d_2(\widehat{F}_{\widehat{Y}/\beta_m^{1/2}} \circ (\mathbf{W}_{\mathbf{Y}} \mathbf{W}_*^{\mathbf{Y}})^\top, \widehat{F}_Y \circ \mathbf{T}^{-\mathbf{1}} \widetilde{\mathbf{Q}}^{-1}) \\ &\quad + d_2(\widehat{F}_X \circ \widetilde{\mathbf{Q}}^{-1}, F_X \circ \widetilde{\mathbf{Q}}^{-1}) + d_2(\widehat{F}_Y \circ \mathbf{T}^{-\mathbf{1}} \widetilde{\mathbf{Q}}^{-1}, F_Y \circ \mathbf{T}^{-\mathbf{1}} \widetilde{\mathbf{Q}}^{-1}) \end{aligned}$$

We show each term is small. For the first two terms, note that these are the empirical CDFs of the points $\widehat{X}_i$ and X_i, where the X_i aare suitably transformed but fixed. Consider the coupling γ which places mass of $\frac{1}{n}$ at the joint observation $(\widehat{X}_i, X_i)$. Then on the event $\mathcal{A}$,

$$\begin{aligned} d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}} \circ (\mathbf{W}_{\mathbf{X}}^\top \mathbf{W}_*^{\mathbf{X}})^\top), \widehat{F}_X \circ \widetilde{\mathbf{Q}}^{-1}) &\leq \left(\frac{1}{n} \|\widehat{\mathbf{X}} (\mathbf{W}_{\mathbf{X}}^\top \mathbf{W}_*^{\mathbf{X}})^\top / \sqrt{\alpha_n} - \mathbf{X} \widetilde{\mathbf{Q}}^{-\mathbf{1}} \|_F^2 \right)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{n}} \|\widehat{\mathbf{X}} / \sqrt{\alpha_n} - \mathbf{X} \widetilde{\mathbf{Q}}^{-\mathbf{1}} \mathbf{W}_{\mathbf{X}}^\top \mathbf{W}_*^{\mathbf{X}} \|_F \\ &\leq \|\widehat{\mathbf{X}} / \sqrt{\alpha_n} - \mathbf{X} \widetilde{\mathbf{Q}}^{-\mathbf{1}} \mathbf{W}_{\mathbf{X}}^\top \mathbf{W}_*^{\mathbf{X}} \|_{2,\infty} \\ &= O\left(\frac{\log(n)}{(n\alpha_n)^{1/2}} \right) \end{aligned}$$

Similarly,

$$d_2(\widehat{F}_{\widehat{Y}/\beta_m^{1/2}} \circ (\mathbf{W}_*^{\mathbf{Y}} \mathbf{W}_{\mathbf{Y}}), \widehat{F}_Y \circ \mathbf{T}^{-\mathbf{1}} \widetilde{\mathbf{Q}}^{-1}) = O\left(\frac{\log(m)}{(m\beta_m)^{1/2}} \right).$$

Finally, we bound the final two terms. By Theorem 26, $\mathrm{supp}(F)$ is bounded and hence so is any fixed invertible linear transformation of $\mathrm{supp}(F)$. Hence, since $\|X\|_\infty \leq M$ almost surely, we can apply Theorem 2 of Fournier and Guillin (2015) to see that with probability $1 - n^{-2}$ that

$$d_2(\widehat{F}_X \circ \widetilde{\mathbf{Q}}^{-1}, F_X \circ \widetilde{\mathbf{Q}}^{-1}) = O\left(\frac{\log^{1/d}(n)}{n^{1/d}} \right);$$

see also Levin and Levina (2019); Lei (2020a,b) for a related problem. Similarly,

$$d_2(\widehat{F}_Y \circ \mathbf{T^{-1}}\widetilde{\mathbf{Q}}^{-1}, F_Y \circ \mathbf{T^{-1}}\widetilde{\mathbf{Q}}^{-1}) = O\left(\frac{\log^{1/d}(m)}{m^{1/d}}\right)$$

with probability $1 - m^{-2}$. Therefore, putting it all together, we see that with probability $1 - O(n^{-2} + m^{-2})$,

$$d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}}, \widehat{F}_{\widehat{Y}/\beta_m^{1/2}} \circ \widehat{\mathbf{W}}_n) = O\left(\frac{\log^{1/d}(n)}{n^{1/d}} + \frac{\log^{1/d}(m)}{m^{1/d}} + \frac{\log(n)}{(n\alpha_n)^{1/2}} + \frac{\log(m)}{(m\beta_m)^{1/2}}\right).$$

Finally, if the sparsity is not known, we have that by Lemma 53, $\frac{1}{\sqrt{\widehat{\alpha}_n}} = \frac{1}{\sqrt{\alpha_n}}\left(1 + O\left(\frac{\sqrt{\log(n)}}{n^{1/2}\alpha_n}\right)\right)$ with probability at least $1 - O(n^{-2})$. Hence, we see that

$$\begin{aligned}
\|\widehat{\mathbf{X}}(\widehat{\alpha}_n^{-1/2} - \alpha_n^{-1/2})\|_{2,\infty} &\leq \|\widehat{\mathbf{X}}\|_{2,\infty} O\left(\frac{\sqrt{\log(n)}}{n^{1/2}\alpha_n}\right) \\
&= O\left(\frac{\sqrt{\log(n)}}{n^{1/2}\alpha_n}\right)\left(\|\widehat{\mathbf{X}} - \sqrt{\alpha_n}\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-1}\mathbf{W}_*^{\mathbf{X}}\|_{2,\infty} + \|\sqrt{\alpha_n}\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{-1}\mathbf{W}_*^{\mathbf{X}}\|_{2,\infty}\right) \\
&= O\left(\frac{\sqrt{\log(n)}}{n^{1/2}\alpha_n}\right)\left(\frac{\log(n)}{\sqrt{n}} + O(\sqrt{\alpha_n})\right) \\
&= O\left(\frac{\log(n)^{3/2}}{n\alpha_n}\right) + O\left(\sqrt{\frac{\log(n)}{n\alpha_n}}\right) \\
&= O\left(\sqrt{\frac{\log(n)}{n\alpha_n}}\right).
\end{aligned}$$

Therefore, by analogous arguments as in the setting with the sparsity factors known, replacing $\widehat{\alpha}_n^{-1/2}$ and $\widehat{\beta}_m^{-1/2}$ with $\alpha_n^{-1/2}$ and $\beta_m^{-1/2}$ is negligible compared to the $2, \infty$ bound. Hence, we have that with probability at least $1 - O(n^{-2} + m^{-2})$

$$d_2(\widehat{F}_{\widehat{X}/\widehat{\alpha}_n^{1/2}}, \widehat{F}_{\widehat{Y}/\widehat{\beta}_m^{1/2}} \circ \widehat{\mathbf{W}}_n) = O\left(\frac{\log^{1/d}(n)}{n^{1/d}} + \frac{\log^{1/d}(m)}{m^{1/d}} + \frac{\log(n)}{(n\alpha_n)^{1/2}} + \frac{\log(m)}{(m\beta_m)^{1/2}}\right).$$

□

Proof of Proposition 7. Like the previous proof, we assume first that the sparsity factors α_n and β_m are known. Let $\mathbf{W_X}, \mathbf{W_Y}, \mathbf{W}_*^{\mathbf{X}}$ and $\mathbf{W}_*^{\mathbf{Y}}$ be as in the proof of Proposition 6, and let $\widetilde{\mathbf{Q}}_{\mathbf{X}}$ be the limit guaranteed by Lemma 50 and similarly for $\widetilde{\mathbf{Q}}_{\mathbf{Y}}$. Note in this setting

$\widetilde{\mathbf{Q}}_{\mathbf{X}}$ is not necessarily equal to $\widetilde{\mathbf{Q}}_{\mathbf{Y}}$. Let $\mathcal{A}$ be the same event as in the previous proof. By the reverse triangle inequality, we have that

$$\begin{aligned} d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}}, \widehat{F}_{\widehat{Y}/\beta_m^{1/2}} \circ \widehat{\mathbf{W}}_n) \geq & -d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}}, F_X \circ (\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{X}}\mathbf{W}_*^{\mathbf{X}})) \\ & - d_2(\widehat{F}_{\widehat{Y}/\beta_m^{1/2}} \circ \widehat{\mathbf{W}}_n, F_Y \circ (\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{Y}}\mathbf{W}_*^{\mathbf{Y}}) \circ \widehat{\mathbf{W}}_n) \\ & + d_2(F_X \circ (\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{X}}\mathbf{W}_*^{\mathbf{X}}), F_Y \circ (\widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{Y}}\mathbf{W}_*^{\mathbf{Y}}) \circ \widehat{\mathbf{W}}_n)) \end{aligned}$$

From the proof of Proposition 6, with probability $1 - O(n^{-2})$ it holds that

$$d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}}, F_X \circ (\widetilde{\mathbf{Q}}_{\mathbf{X}}^{-\mathbf{1}}(\mathbf{W}_{\mathbf{X}}\mathbf{W}_*^{\mathbf{X}})) = O\bigg(\frac{\log(n)}{(n\alpha_n)^{1/2}} + \frac{\log^{1/d}(n)}{n^{1/d}}\bigg),$$

and analogously for the term depending on F_Y.

Consider the sequence $\widetilde{\mathbf{W}}_n := (\mathbf{W}_{\mathbf{Y}}\mathbf{W}_*^{\mathbf{Y}})^\top \widehat{\mathbf{W}}_n (\mathbf{W}_{\mathbf{X}}\mathbf{W}_*^{\mathbf{X}})$. We note as a product of block-orthogonal matrices $\widetilde{\mathbf{W}}_n$ is block-orthogonal and hence an element of $\mathbb{O}(p, q)$. Therefore, through the above argument and the invariance of the Frobenius norm to orthogonal transformations, we have that with probability $1 - O(n^{-2} + m^{-2})$ that

$$d_2(\widehat{F}_{\widehat{X}/\alpha_n^{1/2}}, \widehat{F}_{\widehat{Y}/\beta_m^{1/2}} \circ \widehat{\mathbf{W}}_n) \geq c_n - \varepsilon_n > 0,$$

where $\varepsilon_n \to 0$, and

$$c_n := d_2(F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-1}, F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}} \circ \widetilde{\mathbf{W}}_n).$$

Note that the only dependence on n in the above comes from $\widetilde{\mathbf{W}}_n$. Furthermore, $\mathbb{O}(p, q) \cap \mathbb{O}(d)$ is a closed subgroup of $\mathbb{O}(d)$, and hence compact. Also, the mapping

$$\widetilde{\mathbf{W}} \mapsto d_2(F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-1}, F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}} \circ \widetilde{\mathbf{W}})$$

is continuous since for any fixed $\mu_0 \in \mathbb{R}^d$, we have that

$$\int \|\mu_0 - \widetilde{\mathbf{W}}^\top Y\|^2 dF(Y) = \mathbb{E}\|\mu_0 - \widetilde{\mathbf{W}}^\top Y\|^2,$$

is continuous. Hence, consider a convergent subsequence $\widetilde{\mathbf{W}}_{k_n}$ obtaining $\liminf c_n$ and let $\widetilde{\mathbf{W}}$ be its associated subsequential limit. Hence, on the sequence of events $\mathcal{A}_n$,

$$\liminf d_2(F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-1}, F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}} \circ \widetilde{\mathbf{W}}_{k_n}) = d_2(F_X \circ \widetilde{\mathbf{Q}}_{\mathbf{X}}^{-1}, F_Y \circ \widetilde{\mathbf{Q}}_{\mathbf{Y}}^{-\mathbf{1}} \circ \widetilde{\mathbf{W}}) \geq C$$

for some constant $C > 0$ or else we would have $F_X \simeq F_Y$ since $\widetilde{\mathbf{W}} \in \mathbb{O}(p, q)$. Therefore, by Borel-Cantelli, there exists a constant $C > 0$ such that

$$\liminf d_2(F_{\widehat{X}/\alpha_n^{1/2}}, F_{\widehat{Y}/\beta_m^{1/2}} \circ \widehat{\mathbf{W}}_n) \geq C$$

almost surely.

To replace the α_n and β_m with their respective estimated counterparts, the same argument as in the proof of Proposition 6 goes through.

□

Proof of Proposition 8

Proof of Proposition 8. Let $\mathbf{R}$ have block diagonal components $\mathbf{R}_p$ and $\mathbf{R}_q$, and let $\mathbf{W}_p$ and $\mathbf{W}_q$ be the top $p \times p$ and bottom $q \times q$ block of $\mathbf{W}$ respectively. Expanding out the Frobenius norm, we have that

$$\begin{aligned}
\operatorname*{arg\,min}_{\mathbf{R} \in \mathbb{O}(d) \cap \mathbb{O}(p,q)} \|\mathbf{R} - \mathbf{W}\|_F &= \operatorname*{arg\,min}_{\mathbf{R} \in \mathbb{O}(d) \cap \mathbb{O}(p,q)} \|\mathbf{R} - \mathbf{W}\|_F^2 \\
&= \operatorname*{arg\,min}_{\mathbf{R} \in \mathbb{O}(d) \cap \mathbb{O}(p,q)} \operatorname{Tr}\Big((\mathbf{R} - \mathbf{W})^\top (\mathbf{R} - \mathbf{W}) \Big) \\
&= \operatorname*{arg\,min}_{\mathbf{R} \in \mathbb{O}(d) \cap \mathbb{O}(p,q)} \operatorname{Tr}\Big(\mathbf{R}^\top \mathbf{R} - \mathbf{2R}^\top \mathbf{W} + \mathbf{W}^\top \mathbf{W} \Big) \\
&= \operatorname*{arg\,max}_{\mathbf{R} \in \mathbb{O}(d) \cap \mathbb{O}(p,q)} \operatorname{Tr}(\mathbf{R}^\top \mathbf{W}) \\
&= \operatorname*{arg\,max}_{\mathbf{R} \in \mathbb{O}(d) \cap \mathbb{O}(p,q)} \operatorname{Tr}(\mathbf{R}_\mathbf{p}^\top \mathbf{W}_\mathbf{p}) + \operatorname{Tr}(\mathbf{R}_\mathbf{q}^\top \mathbf{W}_\mathbf{q}) \\
&= \operatorname*{arg\,max}_{\mathbf{R}_\mathbf{p} \in \mathbb{O}(p)} \operatorname{Tr}(\mathbf{R}_\mathbf{p}^\top \mathbf{W}_\mathbf{p}) + \operatorname*{arg\,max}_{\mathbf{R}_\mathbf{q} \in \mathbb{O}(q)} \operatorname{Tr}(\mathbf{R}_\mathbf{q}^\top \mathbf{W}_\mathbf{q}),
\end{aligned}$$

since $\mathbf{R}^\top\mathbf{R} = \mathbf{W}^\top\mathbf{W} = \mathbf{I}$. Let $\mathbf{W_p}$ have singular value decomposition $\mathbf{U_p}\mathbf{\Sigma_p}\mathbf{V_p}^\top$ and similarly for $\mathbf{W_q}$. Then the maximum for each of the above is achieved by setting $\mathbf{R_p} = \mathbf{U_p}\mathbf{V_p}^\top$ and $\mathbf{R_q} = \mathbf{U_q}\mathbf{V_q}^\top$. □

E.1.3 Proof of the Frobenius Concentration (Lemma 51)

In this section we present the proof of Lemma 51. We will need the following Lemma, adapted from Lemma A.4 Athreya et al. (2020).

Lemma 56. *Let* $\mathbf{A}$ *be a matrix whose entries are generated via* $\mathbf{A}_{ij} \sim Bernoulli(\mathbf{P}_{ij})$ *for* $i \leq j$*, and let* $\mathbf{V} = \mathbf{U_X}|\mathbf{\Lambda_X}|^{-\mathbf{1/2}}$ *. Then with probability at least* $1 - O(n^{-2})$*,*

$$\left| \|(\mathbf{A}-\mathbf{P})\mathbf{V}\|_F^2 - \mathbb{E}(\|(\mathbf{A}-\mathbf{P})\mathbf{V}\|_F^2) \right| = O\left(\sqrt{\frac{\log(n)}{n\alpha_n}}\right).$$

Proof of Lemma 56. We follow the proof in Athreya et al. (2020), though we have a slightly different argument for the inclusion of the sparsity factor. Let $\mathbf{A}' \sim \mathbf{P}$ be independent from $\mathbf{A}$. For $1 \leq r \leq s \leq n$ define the term

$$Z_{rs} := \|(\mathbf{A}^{(r,s)} - \mathbf{P})\mathbf{V}\|_F^2,$$

where the matrix $\mathbf{A}^{(r,s)}$ agrees with $\mathbf{A}$ in every entry except the (r,s) and (s,r)ones, where it equals $\mathbf{A}'$. Defining $Z := \|(\mathbf{A}-\mathbf{P})\mathbf{V}\|_F^2$, we see that for $r \neq s$

$$Z - Z_{rs} = 2\left(\mathbf{A}-\mathbf{A}'\right)_{rs}\left[\left[(\mathbf{A}-\mathbf{P})\mathbf{V}\mathbf{V}^\top\right]_{rs} + \left[(\mathbf{A}-\mathbf{P})\mathbf{V}\mathbf{V}^\top\right]_{sr} + \left(\mathbf{A}'-\mathbf{P}\right)_{rs}\left(\left(\mathbf{V}\mathbf{V}^\top\right)_{ss} + \left(\mathbf{V}\mathbf{V}^\top\right)_{rr}\right)\right]$$

Furthermore, we have that

$$(Z - Z_{rs})^2 \leq \begin{cases} 16\left([(\mathbf{A}-\mathbf{P})\mathbf{V}\mathbf{V}^\top]_{rs}^2 + [(\mathbf{A}-\mathbf{P})\mathbf{V}\mathbf{V}^\top]_{sr}^2 + (\mathbf{A}'-\mathbf{P})_{rs}^2\left[(\mathbf{V}\mathbf{V}^\top)_{ss}^2 + \mathbf{V}\mathbf{V}^\top)_{rr}^2\right]\right) & r \neq s \\ 8\left([(\mathbf{A}-\mathbf{P})\mathbf{V}\mathbf{V}^\top]_{rr}^2 + (\mathbf{A}'-\mathbf{P})_{rs}^2\left[\mathbf{V}\mathbf{V}^\top\right]_{rr}^2\right) & r = s \end{cases}$$

Hence,

$$\sum_{r\leq s}(Z-Z_{rs})^2 \leq 16Z\|\mathbf{V}\|_2^2 + 8\sum_r(\mathbf{V}\mathbf{V}^\top)_{rr}^2 + 16\sum_{r<s}(\mathbf{A}'-\mathbf{P})_{rs}^2[(\mathbf{V}\mathbf{V}^{\top 2}_{ss} + (\mathbf{V}\mathbf{V}^\top)_{rr}^2]$$
$$= 16Z\|\mathbf{V}\|_2^2 + 8\|\mathrm{diag}(\mathbf{V}\mathbf{V}^\top)\|_F^2 + 16\sum_{s=1}^n\sum_{r=1}^{s-1}(\mathbf{A}'-\mathbf{P})_{rs}^2[(\mathbf{V}\mathbf{V}^\top)_{ss}^2 + (\mathbf{V}\mathbf{V}^\top)_{rr}^2].$$

For the final term above, we have that

$$\mathbb{E}_{\mathbf{A}'}\left[\sum_{s=1}^n\sum_{r=1}^{s-1}(\mathbf{A}'-\mathbf{P})_{rs}^2[(\mathbf{V}\mathbf{V}^\top)_{ss}^2 + (\mathbf{V}\mathbf{V}^\top)_{rr}^2]\right] = \sum_{s=1}^n\sum_{r=1}^{s-1}\mathbb{E}_{\mathbf{A}'}(\mathbf{A}'-\mathbf{P})_{rs}^2[(\mathbf{V}\mathbf{V}^\top)_{ss}^2 + (\mathbf{V}\mathbf{V}^\top)_{rr}^2]$$
$$\leq 2n\alpha_n\|\mathrm{diag}(\mathbf{V}\mathbf{V}^\top)\|_F^2.$$

Moreover, from the definitions of $\mathbf{V}$, we have that $\|\mathrm{diag}(\mathbf{V}\mathbf{V}^\top)\|_F^2 \leq d\lambda_d^{-2} \leq Cd(n\alpha_n)^{-2}$, and $\|\mathbf{V}\|_2^2 = |\lambda_d|^{-1} \leq C(n\alpha_n)^{-1}$. Hence, we see that

$$\mathbb{E}_{\mathbf{A}'}\sum_{r\leq s}(Z-Z_{rs})^2 \leq \frac{C_1}{n\alpha_n}Z + \frac{C_2 d}{(n\alpha_n)^2} + \frac{C_3 d}{n\alpha_n},$$

Define $a := \frac{C_1}{n\alpha_n}$ and $b := \frac{C_2 d}{(n\alpha_n)^2} + \frac{C_3 d}{n\alpha_n}$. By Theorems 5 and 6 in Boucheron et al. (2003), we have that

$$\mathbb{P}(|Z-\mathbb{E}Z| > t) \leq 2\exp\left(\frac{-t^2}{4a\mathbb{E}(Z) + 4b + 2at}\right)$$
$$\leq 2\exp\left(\frac{-t^2 n\alpha_n}{4C_1\mathbb{E}(Z) + 4C_2 d(n\alpha_n)^{-1} + 4C_3 d + 2C_1 t}\right)$$
$$\leq 2\exp\left(\frac{-t^2 n\alpha_n}{\widetilde{C}_1 + \widetilde{C}_2 t}\right)$$

for some constants $\widetilde{C}_1$ and $\widetilde{C}_2$ (depending on d) and for $n\alpha_n$ sufficiently large. Note that we

implicitly used that $\mathbb{E}Z = O(1)$, which can be seen from the fact that

$$\begin{aligned}
\mathbb{E}(Z) &= \mathbb{E}\|(\mathbf{A}-\mathbf{P})\mathbf{V}\|_F^2 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{d}\mathbb{E}\Big(\sum_j (\mathbf{A}_{ij}-\mathbf{P}_{ij})\mathbf{V}_{jk}\Big)^2 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{d}\sum_{j=1}^{n}\mathbf{V}_{jk}^2\mathbb{E}((\mathbf{A}_{ij}-\mathbf{P}_{ij}))^2 + \sum_{j\neq l}^{n}\mathbf{V}_{jk}\mathbf{V}_{lk}\mathbb{E}((\mathbf{A}_{ij}-\mathbf{P}_{ij})(\mathbf{A}_{il}-\mathbf{P}_{il})) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{d}\sum_{j=1}^{n}\mathbf{V}_{jk}^2\mathbf{P}_{ij}(1-\mathbf{P}_{ij}) \\
&\leq n\alpha_n\sum_{k=1}^{d}\sum_{j=1}^{n}\mathbf{V}_{jk}^2 \\
&\leq n\alpha_n\|\mathbf{U}_{\mathbf{X}}|\mathbf{\Lambda}_{\mathbf{X}}|^{-1/2}\|_F^2 \\
&\leq C
\end{aligned}$$

for some constant C depending on d and λ_d. Hence, with the choice $t = \widetilde{C}\sqrt{\frac{\log(n)}{n\alpha_n}}$ for some constant $\widetilde{C}$ depending on d, this is bounded above by $2n^{-2}$. □

We are now ready to prove Lemma 51.

Proof of Lemma 51. First, by the proof of Theorem 5 in Rubin-Delanchy et al. (2020), we note that there exists an orthogonal matrix $\mathbf{W}_* \in \mathbb{O}(d)\cap\mathbb{O}(p,q)$ (see equations 5 and 6 in Rubin-Delanchy et al. (2020)) such that

$$\widehat{\mathbf{U}}|\widehat{\mathbf{\Lambda}}|^{\mathbf{1/2}} - \mathbf{U}|\mathbf{\Lambda}|^{\mathbf{1/2}}\mathbf{W}_*^\top = (\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda}|^{\mathbf{-1/2}}\mathbf{W}_*^\top\mathbf{I}_{p,q} + \mathbf{R},$$

where the matrix $\mathbf{R}$ satisfies

$$\|\mathbf{R}\|_{2,\infty} = O\Big(\frac{d^{1/2}\log^{1/2}(n)}{\sqrt{n}(n\alpha_n)^{1/2}}\Big).$$

Passing to the Frobenius norm, we see that

$$\|\mathbf{R}\|_F = O\Big(\frac{d^{1/2}\log^{1/2}(n)}{(n\alpha_n)^{1/2}}\Big).$$

This proves the first claim. Hence,

$$\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{W}_*^\top\|_F = \|(\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda}|^{-\mathbf{1/2}}\|_F + O\left(\sqrt{\frac{d\log(n)}{n\alpha_n}}\right).$$

We then can apply Lemma 56 to see that

$$\mathbb{P}\left(|\|(\mathbf{A}-\mathbf{P})\mathbf{V}\|_F^2 - C(\mathbf{P})^2| > C\sqrt{\log(n)/n\alpha_n}\right) = O(n^{-2}),$$

where $\mathbf{V} = \mathbf{U}|\mathbf{\Lambda}|^{-\mathbf{1/2}}$ and $C^2(\mathbf{P}) = \mathbb{E}\|(\mathbf{A}-\mathbf{P})\mathbf{V}\|_F^2$. The rest of the proof is similar to Tang and Priebe (2018). By similar manipulations as those leading to Equation 18 in Rubin-Delanchy et al. (2020), we have that

$$(\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda}|^{-\mathbf{1/2}}\mathbf{W}_*^\top\mathbf{I}_{p,q} = \alpha_n^{-1/2}(\mathbf{A}-\mathbf{P})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{I}_{p,q}{\mathbf{Q}_\mathbf{X}}^{-1}.$$

By Lemma 50, we have that there exists a sequence of block-orthogonal matrices such that $\mathbf{W}_\mathbf{n}^\top\mathbf{Q}_\mathbf{X}^{-\mathbf{1}} \to \widetilde{\mathbf{Q}}^{-\mathbf{1}}$ almost surely. Hence, we have that

$$\begin{aligned}
\|(\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda}|^{-\mathbf{1/2}}\|_F^2 &= \frac{1}{\alpha_n}\|(\mathbf{A}-\mathbf{P})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{I}_{p,q}{\mathbf{Q}_\mathbf{X}}^{-1}\|_F^2 \\
&= \frac{1}{\alpha_n}\mathrm{Tr}\left(\mathbf{Q}_\mathbf{X}^{-\mathbf{1}}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbb{E}(\mathbf{A}-\mathbf{P})^{\mathbf{2}}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{Q}_\mathbf{X}^{-\top}\right) \\
&= \frac{1}{\alpha_n}\mathrm{Tr}\left(\mathbf{W}_\mathbf{n}^\top\mathbf{Q}_\mathbf{X}^{-\mathbf{1}}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\top\mathbb{E}(\mathbf{A}-\mathbf{P})^{\mathbf{2}}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}\mathbf{Q}_\mathbf{X}^{-\top}\mathbf{W}_\mathbf{n}\right) \\
&= \mathrm{Tr}\left(\mathbf{W}_\mathbf{n}^\top\mathbf{Q}_\mathbf{X}^{-\mathbf{1}}(n(\mathbf{X}^\top\mathbf{X})^{-1})\left[\frac{\mathbf{X}^\top\mathbb{E}(\mathbf{A}-\mathbf{P})^{\mathbf{2}}\mathbf{X}}{n^2\alpha_n}\right](n(\mathbf{X}^\top\mathbf{X})^{-1})\mathbf{Q}_\mathbf{X}^{-\top}\mathbf{W}_\mathbf{n}\right).
\end{aligned}$$

By the strong law of large numbers the term $\mathbf{X}^\top\mathbf{X}/n \to \Delta$ almost surely, so $n(\mathbf{X}^\top\mathbf{X})^{-1} \to \Delta^{-1}$ almost surely by the continuous mapping theorem. In addition, we have that

$$\begin{aligned}
\frac{\mathbf{X}^\top\mathbb{E}(\mathbf{A}-\mathbf{P})^{\mathbf{2}}\mathbf{X}}{n^2\alpha_n} &= \frac{1}{n^2\alpha_n}\sum_{i=1}^n\sum_k X_iX_i^\top(p_{ik}(1-p_{ik})) \\
&= \frac{1}{n^2}\sum_{i=1}^n\sum_k X_iX_i^\top(X_i^\top\mathbf{I}_{p,q}X_k - \alpha_n X_i^\top\mathbf{I}_{p,q}X_kX_k^\top\mathbf{I}_{p,q}X_i).
\end{aligned}$$

As $n \to \infty$, this is tending to the matrix Γ, where Γ is defined via

$$\Gamma := \begin{cases} \mathbb{E}(XX^\top(X^\top \mathbf{I}_{p,q}\mu - X^\top \mathbf{I}_{p,q}\Delta \mathbf{I}_{p,q} X)) & \alpha \equiv 1 \\ \mathbb{E}(XX^\top(X^\top \mathbf{I}_{p,q}\mu)) & \alpha \to 0, \end{cases}$$

where $\mu = \mathbb{E}(X)$. Hence, putting it together, we have almost surely,

$$\|\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{W}_*^\top\|_F^2 \to \mathrm{Tr}\left(\widetilde{\mathbf{Q}}^{-1}\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}\boldsymbol{\Delta}^{-1}\widetilde{\mathbf{Q}}^{-\top}\right).$$

□

E.1.4 Proof of the Functional CLT (Lemma 52) and Related Lemmas

In this section, we prove Lemma 52 and Lemma 54.

Proof of Lemma 52. We follow the proof by analogy to the proof Lemma 3 of Tang et al. (2017b), though we use the decomposition from Lemma 51. As $\mathcal{F}$ is twice continuously differentiable, for $f \in \mathcal{F}$ we Taylor expand to note that

$$\frac{\sqrt{\alpha_n}}{\sqrt{n}}\sum_{i=1}^n (f(\widehat{X}_i/\sqrt{\alpha_n}) - f(\mathbf{W}_*\widetilde{X}_i)/\sqrt{\alpha_n}) = \frac{\sqrt{\alpha_n}}{\sqrt{n}}\sum_{i=1}^n (\partial f)(\widetilde{X}_i)\frac{(\widehat{X}_i - \mathbf{W}_*\widetilde{X}_i)}{\sqrt{\alpha_n}} + \frac{\sqrt{\alpha_n}}{2\sqrt{n}}\sum_i \frac{(\widehat{X}_i - \mathbf{W}_*\widetilde{X}_i)^\top(\partial^2 f)(X_i^*)(\widehat{X}_i - \mathbf{W}_*\widetilde{X}_i)}{\alpha_n},$$

for some X_i^*. The second order term is straightforwardly bounded by noting that by Theorem 26 $\bar{\Omega}$ is compact, and hence we can apply Lemma 51 to see that there exists some constant C such that

$$\sup_{f\in\mathcal{F}}\sum_{i=1}^n \frac{(\widehat{X}_i - \mathbf{W}_*\widetilde{X}_i)^\top(\partial^2 f)(X_i^*)(\widehat{X}_i - \mathbf{W}_*\widetilde{X}_i)}{\sqrt{n\alpha_n}} \le \sup_{f\in\mathcal{F}, X\in\bar{\Omega}} \frac{\|\partial^2(f)(X)\|\|\widehat{X}_i - \mathbf{W}_*\widetilde{X}_i\|_F^2}{\sqrt{n\alpha_n}} \le \frac{C}{\sqrt{n\alpha_n}},$$

which converges to zero almost surely.

We now bound the linear terms. Let $\mathbf{M}(\partial f) = \mathbf{M}(\partial f; \widetilde{X}_1, \ldots, \widetilde{X}_n) \in \mathbb{R}^{n\times d}$ be the

matrix whose rows are the vectors $\partial(f)(\widetilde{X}_1)$. Then

$$\begin{aligned}\zeta(f) :&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\partial f)(\widetilde{X}_i)^\top(\widehat{X}_i - \mathbf{W}_*\widetilde{X}_i)\\ &= \frac{1}{\sqrt{n}}\mathrm{Tr}\big((\widehat{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{W}_*)[\mathbf{M}(\partial f)]^\top\big)\\ &= \frac{1}{\sqrt{n}}\mathrm{Tr}\big((\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda}|^{-\mathbf{1/2}}\mathbf{I}_{p,q})[\mathbf{M}(\partial f)]^\top\big) + \frac{1}{\sqrt{n}}\mathrm{Tr}\big(\mathbf{R}[\mathbf{M}(\partial f)]^\top\big),\end{aligned}$$

where $\mathbf{R}$ is the residual matrix in 51. Recall the second term satisfies

$$\begin{aligned}\frac{1}{\sqrt{n}}\mathrm{Tr}\big(\mathbf{R}[\mathbf{M}(\partial f)]^\top\big) &= \frac{1}{\sqrt{n}}\langle \mathbf{R}, \mathbf{M}(\partial f)\rangle\\ &\le \sup_{f\in\mathcal{F}, X\in\bar{\Omega}}\frac{\sqrt{n}\|\partial f(X)\|}{\sqrt{n}}\|\mathbf{R}\|_F\\ &\le \frac{C\sqrt{n}}{\sqrt{n}}\|\mathbf{R}\|_F\\ &\le \frac{C\sqrt{\log(n)}}{\sqrt{n\alpha_n}},\end{aligned}$$

where the penultimate inequality comes from the fact that $\bar{\Omega}$ can be taken to be compact by Theorem 26, and since $\mathcal{F}$ is twice-continuously differentiable, the gradient is Lipschitz on any fixed transformation of the support.

Now, we show that the final term converges to zero. The rest of the proof is largely the same as in Tang et al. (2017b). Define the set of derivatives of $\partial\mathcal{F} := \{\partial f : f \in \mathcal{F}\}$. Let $\|\partial f\|_\infty$ be the maximum Euclidean norm attained by f on $\bar{\Omega}$. Note that by enlarging it if necessary, $\bar{\Omega}$ can be taken to be compact and contain the $\widehat{X}_i$'s by the fact that κ is twice continuously differentiable on $\mathbb{R}^d$ and by virtue of Theorems 27 and 26 and Lemma 50. Therefore the set of derivatives is totally bounded; define $M := \sup_{\partial\mathcal{F}}\|\partial f\|_\infty$.

Then for any j there exists a finite subset S_j of covering functions such that for any $g \in \partial\mathcal{F}$, we have that $\|g - f_j\|_\infty \le 2^{-j}M$. Define the mapping $\mathcal{P}_j$ as the mapping that

assigns the function $g \in \partial\mathcal{F}$ to its closest function $f_j \in S_j$. Then we have that

$$\begin{aligned}
&\sup_f \left| \frac{1}{\sqrt{n}} \mathrm{Tr}\big((\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda}^{\mathbf{-1/2}}\mathbf{I}_{p,q})[\mathbf{M}(\partial f)]^\top\big) \right| \\
&\qquad = \sup_f \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{j=0}^\infty (\mathcal{P}_{j+1}\partial f - \mathcal{P}_j \partial f)(\widetilde{X}_i)^\top \big[(\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda}|^{\mathbf{-1/2}}\mathbf{I}_{p,q}\big]_i \right| \\
&\qquad \le \sum_{j=0}^\infty \sup_f \left| \sum_{i=1}^n (\mathcal{P}_{j+1}\partial f - \mathcal{P}_j \partial f)(\widetilde{X}_i)^\top \big[(\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda}|^{\mathbf{-1/2}}\mathbf{I}_{p,q}\big]_i \right| \\
&\qquad + \frac{1}{\sqrt{n}} \left| \sum_{i=1}^n (\mathcal{P}_0 \partial f)(\widetilde{X}_i)^\top \big[(\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda}|^{\mathbf{-1/2}}\mathbf{I}_{p,q}\big]_i \right|.
\end{aligned}$$

We note that for fixed j, defining the term $\mathfrak{P}_j^f$ as the $n \times d$ matrix whose rows are $(\mathcal{P}_{j+1}\partial f - \mathcal{P}_j\partial f)(\widetilde{X}_i)^\top$, we have that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left| (\mathcal{P}_{j+1}\partial f - \mathcal{P}_j \partial f)(\widetilde{X}_i)^\top \big[(\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda}|^{\mathbf{-1/2}}\big]_i \right| = \frac{1}{\sqrt{n}} \left| \sum_{s=1}^d ((\mathfrak{P}_j^f)^\top \big((\mathbf{A}-\mathbf{P})\mathbf{U}_s \frac{-\mathbb{I}_{s>p} + \mathbb{I}_{s\le p}}{\sqrt{|\lambda_s|^{1/2}}}\big) \right|.$$

We note that

$$\|(\mathfrak{P}_f^j)_s\| \le \frac{3}{2} 2^{-j} M \sqrt{n}.$$

Hence, for fixed $s \in \{1 \dots, d\}$ this is a linear combination of mean-zero random variables. Therefore we have by Hoeffding's inequality that

$$\mathbb{P}\bigg(((\mathfrak{P}_j^f)_s^\top \big((\mathbf{A}-\mathbf{P})\mathbf{U}_s \frac{-\mathbb{I}_{s>p} + \mathbb{I}_{s\le p}}{\sqrt{|\lambda_s|^{1/2}}}\big) > t \bigg) \le 2 \exp\bigg(- \frac{t^2}{C 2^{-2j} \lambda_d^{-1}} \bigg),$$

for some constant C depending on M. Hence, by the union bound, we have that

$$\mathbb{P}\bigg[\frac{1}{\sqrt{n}} \left| \sum_{s=1}^d ((\mathfrak{P}_j^f)_s^\top \big((\mathbf{A}-\mathbf{P})\mathbf{U}_s \frac{-\mathbb{I}_{s>p} + \mathbb{I}_{s\le p}}{\sqrt{|\lambda_s|^{1/2}}}\big) \right| > dt \bigg] \le 2d \exp\bigg(- \frac{t^2}{C 2^{-2j} |\lambda_d|^{-1}} \bigg),$$

provided that t is chosen appropriately. By another union bound over the set S_j, we have that

$$\mathbb{P}\bigg[\sup_f \frac{1}{\sqrt{n}} \left| \sum_{i=1}^n ((\mathfrak{P}_j^f)^\top \big((\mathbf{A}-\mathbf{P})\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}\big)_i \right| > dt \bigg] \le 2d|S_j| \exp\bigg(- \frac{t^2}{C 2^{-2j} |\lambda_d|^{-1}} \bigg).$$

We note that $|S_j| \leq (C2^j)^d$ by using the bound on the covering number (e.g. Lemma 2.5 in van de Geer (2009)). Following the steps from equation A.5 to A.6 in Tang et al. (2017b), by rearranging the equation above, we have that for any $t_j > 0$

$$\mathbb{P}\left[\sup_f \left|\frac{1}{\sqrt{n}}\right| \sum_{i=1}^n ((\mathfrak{P}_j^f)^\top \big((\mathbf{A}-\mathbf{P})\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}\big)_i\right| > \eta_j\right] \leq 2d\exp(-t_j^2),$$

where $\eta_j = d\sqrt{C2^{-2j}\lambda_d^{-1}(t_j^2 + \log|S_{j+1}|^2)}$. Summing over j and bounding the zeroth order term similarly, we have

$$\mathbb{P}\left[\sup_f |\frac{1}{\sqrt{n}}\mathrm{Tr}\big(\big((\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda^{-1/2}}\mathbf{I}_{p,q})[\mathbf{M}(\partial f)]^\top\big) \geq \sum_{j=0}^\infty \widetilde{C}\eta_j\right] \leq 2d\sum_{j=0}^\infty \exp\left(-t_j^2\right).$$

Taking $t_j^2 = 2(\log(j) + \log(n))$, we have that

$$\mathbb{P}\left[\sup_f |\frac{1}{\sqrt{n}}\mathrm{Tr}\big(\big((\mathbf{A}-\mathbf{P})\mathbf{U}|\mathbf{\Lambda^{-1/2}}\mathbf{I}_{p,q})[\mathbf{M}(\partial f)]^\top\big) \geq d\lambda_d^{-1/2}(C_1\sqrt{\log(n)} + C_2)\right] \leq \frac{2dC_3}{n^2},$$

for some constants C_1, C_2, and C_3. Hence, combining all this, we see the linear term satisfies

$$\sup_f |\zeta(f)| \leq C\frac{\sqrt{\log(n)}}{\sqrt{n\alpha_n}}$$

with probability at least $1 - O(n^{-2})$. We note that the hidden constants and the constants in the bound depend on the diameter of the set $\partial\mathcal{F}$ and the dimension d.

Finally, if $\sqrt{n}\alpha_n \geq \log^{1+\eta}(n)$, then the right hand side still tends to zero with an additional factor of $\sqrt{\alpha_n}$ in the denominator, which is the final assertion of Lemma 52. □

Proof of Lemma 54. First, by Lemma 53, we have that $\sqrt{\alpha_n}/\sqrt{\widehat{\alpha}_n} \to 1$ in probability (and almost surely). In the proof of Lemma 52, we have shown that

$$\sup_f \frac{\sqrt{\alpha_n}}{\sqrt{n}}\sum_{i=1}^n (\partial f)(\widetilde{X}_i)\frac{(\widehat{X}_i - \mathbf{W}_*\widetilde{X}_i)}{\sqrt{\alpha_n}} \to 0;$$
$$\sup_f \frac{\sqrt{\alpha_n}}{2\sqrt{n}}\sum_i \frac{(\widehat{X}_i - \mathbf{W}_*\widetilde{X}_i)^\top(\partial^2 f)(X_i^*)(\widehat{X}_i - \mathbf{W}_*\widetilde{X}_i)}{\alpha_n} \to 0$$

almost surely. Therefore, we can replace $\sqrt{\alpha_n}$ by $\sqrt{\widehat{\alpha}_n}$ and apply Slutsky's Theorem to conclude that

$$\sup_f \frac{\sqrt{\alpha_n}}{\sqrt{n}} \sum_{i=1}^{n} (\partial f)(\widetilde{X}_i) \frac{(\widehat{X}_i}{\sqrt{\widehat{\alpha}_n}} - \frac{\mathbf{W}_* \widetilde{X}_i)}{\sqrt{\alpha_n}} \to 0;$$
$$\sup_f \frac{\sqrt{\alpha_n}}{2\sqrt{n}} \sum_i \frac{(\widehat{X}_i}{\sqrt{\widehat{\alpha}_n}} - \frac{\mathbf{W}_* \widetilde{X}_i)^\top}{\sqrt{\alpha_n}} (\partial^2 f)(X_i^*) \frac{(\widehat{X}_i}{\sqrt{\widehat{\alpha}_n}} - \frac{\mathbf{W}_* \widetilde{X}_i)}{\sqrt{\alpha_n}} \to 0$$

in probability. □

E.1.5 Proofs of Auxiliary Lemmas

In this section we prove the additional technical lemmas; namely Lemmas 49, 55, 50, and 53.

Proofs of Lemmas 49, 55, and 50

This section contains various results associated to the approximation of the matrix $\mathbf{Q_X}$ to its limiting value.

Proof of Lemma 49. By Agterberg et al. (2020b), we see that $\mathbf{Q_X} \xrightarrow{a.s.} \widetilde{\mathbf{Q}}_\mathbf{X}$, where $\widetilde{\mathbf{Q}}$ is a fixed indefinite orthogonal matrix, and similarly for $\mathbf{Q_Y}$, so that $\mathbf{Q_Y} \to \widetilde{\mathbf{Q}}'$ for some fixed matrix $\widetilde{\mathbf{Q}}'$. Define the matrix $\widetilde{\mathbf{X}} := \mathbf{Y}\mathbf{T}^{-\mathbf{1}}$. Note that the rows of $\widetilde{\mathbf{X}}$ are distributed iid F_X. Suppose that $\mathbf{Q}_{\widetilde{\mathbf{X}}}$ is the indefinite orthogonal matrix such that

$$\mathbf{U_Y}|\mathbf{S_Y}|^{\mathbf{1/2}}\mathbf{Q}_{\widetilde{\mathbf{X}}} = \widetilde{\mathbf{X}}.$$

Then we apply Theorem 1 from Agterberg et al. (2020b) again, which implies that $||\mathbf{Q}_{\widetilde{\mathbf{X}}} - \mathbf{Q_X}|| \to 0$ almost surely. However, noting $\widetilde{\mathbf{X}} = \mathbf{Y}\mathbf{T}^{-\mathbf{1}}$ gives that $\mathbf{Q}_{\widetilde{\mathbf{X}}} = \mathbf{Q_Y}\mathbf{T}^{-\mathbf{1}}$, which gives the result. □

Proof of Lemma 55. The proof mostly is similar to that of Lemma 49, only instead we apply Corollary 2 from Agterberg et al. (2020b). We have that there exists a sequence of block-orthogonal matrices $\mathbf{W_X}$ such that $||\mathbf{Q_X} - \mathbf{W_X}\widetilde{\mathbf{Q}}|| \to 0$. Again, we are free to repeat the

argument in the case $F_X = F_Y \circ \mathbf{T}$, only replacing $\mathbf{Q_X}$ with $\mathbf{Q_Y T^{-1}}$, to see that

$$||\mathbf{Q_Y T^{-1}} - \mathbf{W_Y}\widetilde{\mathbf{Q}}|| \to 0.$$

Note that the (block)-orthogonal matrix above need not necessarily be the same due to nonidentifiability of the eigenvectors. Hence,

$$\mathbf{W_Y^\top Q_Y T^{-1}} - \mathbf{W_X^\top Q_X} \to 0,$$

which in particular implies that both terms are tending towards $\widetilde{\mathbf{Q}}$ almost surely. □

Proof of Lemma 50. Define $\mathbf{\Delta} = \mathbb{E}(XX^\top)$, and let $\widetilde{\mathbf{V}}$ and $\mathbf{V}$ be the orthogonal matrices in the eigendecomposition of $\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right)^{1/2}\mathbf{I}_{p,q}\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right)^{1/2}$ and $\mathbf{\Delta}^{1/2}\mathbf{I}_{p,q}\mathbf{\Delta}^{1/2}$ respectively. Let $\mathbf{\Lambda}$ be the eigenvalues of $\mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top$ and let $\widetilde{\mathbf{\Lambda}}$ be the eigenvalues of $\mathbf{\Delta}^{1/2}\mathbf{I}_{p,q}\mathbf{\Delta}^{1/2}$. We will first show that with probability at least $1 - n^{-2}$, that the following hold simultaneously:

$$\|\mathbf{\Delta}^{1/2} - \left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right)^{1/2}\| = O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right); \tag{E.11}$$

$$\|\left(\frac{|\mathbf{\Lambda}|}{n}\right)^{-1/2} - |\widetilde{\mathbf{\Lambda}}|^{-1/2}\| = O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right); \tag{E.12}$$

$$\|\mathbf{V} - \widetilde{\mathbf{V}}\mathbf{W}_n^\top\| = O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right);, \tag{E.13}$$

where $\mathbf{W}_n \in \mathbb{O}(p,q) \cap \mathbb{O}(d)$ will be defined later.

First, by Theorem 6.2 in Higham (2008), for $\mathbf{A}$ and $\mathbf{B}$ positive-definite matrices, we have that

$$\|\mathbf{A}^{1/2} - \mathbf{B}^{1/2}\| \leq \frac{1}{\lambda_{\min}(\mathbf{A})^{1/2} + \lambda_{\min}(\mathbf{B})^{1/2}}\|\mathbf{A} - \mathbf{B}\|. \tag{E.14}$$

The result above is stated in Higham (2008) for matrices with distinct eigenvalues. However, if $\mathbf{A}$ and $\mathbf{B}$ do not have distinct eigenvalues, the result holds by adding small values of ε to each of the repeated eigenvalues, applying the result for the new slightly perturbed matrices, and then taking the limit as $\varepsilon \to 0$.

By the Law of Large Numbers, $\frac{\mathbf{X}^\top\mathbf{X}}{n}$ is almost surely positive definite whenever $\mathbf{\Delta}$ is. Applying the above inequality to $\mathbf{\Delta}$ and $\frac{\mathbf{X}^\top\mathbf{X}}{n}$, we see that

$$\begin{aligned}\|\mathbf{\Delta}^{1/2} - (\frac{\mathbf{X}^\top\mathbf{X}}{n})^{1/2}\| &\leq \frac{1}{\lambda_{\min}(\mathbf{\Delta})^{1/2} + \lambda_{\min}(\frac{\mathbf{X}^\top\mathbf{X}}{\mathbf{n}})^{1/2}}\|\mathbf{\Delta} - \frac{\mathbf{X}^\top\mathbf{X}}{n}\| \\ &\leq \frac{1}{\lambda_{\min}(\mathbf{\Delta})^{1/2}}\|\mathbf{\Delta} - \frac{\mathbf{X}^\top\mathbf{X}}{n}\|.\end{aligned}$$

By Theorem 6.5 in Wainwright (2019) (which applies to second moment matrices by shifting by the mean), we have that for some constants c_1, c_2 and c_3,

$$\|\mathbf{\Delta} - \frac{\mathbf{X}^\top\mathbf{X}}{n}\| \leq \|\mathbf{\Delta}\|c_1\left\{\sqrt{\frac{d}{n}} + \frac{d}{n}\right\} + \delta$$

with probability at least $1 - c_3\exp(-c_2 n\min(\delta, \delta^2))$. Taking $\delta = \sqrt{\frac{2\log(n)}{c_3 n}}$, we see that

$$\|\mathbf{\Delta} - \frac{\mathbf{X}^\top\mathbf{X}}{n}\| = O\left(\|\mathbf{\Delta}\|\sqrt{\frac{d}{n}}\sqrt{\log(n)}\right)$$

with probability at least $1 - O(n^{-2})$. Putting it all together and noting $\|\mathbf{\Delta}\|$ and d are constants in n, we arrive at

$$\|\mathbf{\Delta}^{1/2} - \frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2}\| \leq O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right).$$

which proves (E.11).

For (E.12), we note that $\mathbf{\Lambda}$ and $\widetilde{\mathbf{\Lambda}}/n$ are the eigenvalues of the matrix $\Delta^{1/2}\mathbf{I}_{p,q}\Delta^{1/2}$ and $\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right)^{1/2}\mathbf{I}_{p,q}\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right)^{1/2}$ respectively. To see the latter, we note that $\mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^\top$ has the same nonzero eigenvalues as $(\mathbf{X}^\top\mathbf{X})^{1/2}\mathbf{I}_{p,q}(\mathbf{X}^\top\mathbf{X})^{1/2}$ by similarity. By Weyl's inequality, we

have that

$$\begin{aligned}|\lambda_i - \widetilde{\lambda}_i| &\leq \|\frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2}\mathbf{I}_{p,q}\frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2} - \Delta^{1/2}\mathbf{I}_{p,q}\Delta^{1/2}\| \\ &\leq \|\frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2}\mathbf{I}_{p,q}\Delta^{1/2} - \Delta^{1/2}\mathbf{I}_{p,q}\Delta^{1/2}\| + \|\frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2}\mathbf{I}_{p,q}\Delta^{1/2} - \frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2}\mathbf{I}_{p,q}\frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2}\| \\ &\leq \|\frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2} - \Delta^{1/2}\|\|\Delta^{1/2}\| + \|\frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2}\|\|\Delta^{1/2} - \frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2}\| \\ &= O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right)\end{aligned}$$

by (E.11) and the fact that $\frac{\mathbf{X}^\top\mathbf{X}}{n}$ can be bounded above by a constant. Now, define the function $g(\lambda) := \frac{1}{|\lambda|^{1/2}}$. Since Δ is full-rank and $\mathbf{I}_{p,q}$ is orthogonal, so is $\Delta\mathbf{I}_{p,q}$ and hence $\Delta^{1/2}\mathbf{I}_{p,q}\Delta^{1/2}$ by similarity. Therefore, there exists an $\varepsilon > 0$ such that all eigenvalues of $\Delta^{1/2}\mathbf{I}_{p,q}\Delta^{1/2}$ are outside the range $(-\varepsilon, \varepsilon)$, and hence so are those of $(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{1/2}\mathbf{I}_{p,q}(\frac{\mathbf{X}^\top\mathbf{X}}{n})^{1/2}$ for n sufficiently large with probability at least $1 - n^{-2}$. The function $g(\lambda)$ is differentiable outside of $(-\varepsilon, \varepsilon)$, and hence by the Delta method applied to each eigenvalue individually, $g(\lambda_i) - g(\widetilde{\lambda}_i) = O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right)$, for n sufficiently large with probability $1 - O(n^{-2})$, where the hidden constant depends on $g'(\widetilde{\lambda}_i)$. This proves (E.12).

For (E.13), we simply apply the Davis-Kahan Theorem to the eigenvectors associated to each eigenvalue. First, consider i such that $\widetilde{\lambda}_i$ is unique. By (E.12), for n sufficiently large, the eigenvalues outside of i are separated from each other and we can apply the Davis-Kahan Theorem. We see that

$$\begin{aligned}\|\mathbf{v}_i - \widetilde{\mathbf{v}}_i\| &\leq C\frac{\|\frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2}\mathbf{I}_{p,q}\frac{\mathbf{X}^\top\mathbf{X}}{n}^{1/2} - \Delta^{1/2}\mathbf{I}_{p,q}\Delta^{1/2}\|}{\delta_{gap}^{(i)}} \\ &= O\left(\frac{\sqrt{\log(n)}}{\sqrt{n}}\right),\end{aligned}$$

where the hidden constant depends on $\delta_{gap}^{(i)} := \min(\lambda_{i+1} - \lambda_i, \lambda_i - \lambda_{i-1})$, though this is a deterministic value depending only on Δ. For non-unique eigenvalues, we apply the same argument, only with an orthogonal matrix attached to the $\mathbf{V}_i$. We note that the orthogonal matrix is chosen only for each group of repeated eigenvalues, and hence the combined matrix is block-orthogonal.

We are now ready to prove the result in Equation E.2. By Agterberg et al. (2020b), we can write

$$\mathbf{Q}_{\mathbf{X}} = \left(\frac{|\mathbf{\Lambda}|}{n}\right)^{-1/2} \mathbf{V}^{\top} \left(\frac{\mathbf{X}^{\top}\mathbf{X}}{n}\right)^{1/2}$$
$$\widetilde{\mathbf{Q}} = \left(|\widetilde{\mathbf{\Lambda}}|\right)^{-1/2} \widetilde{\mathbf{V}}^{\top} (\mathbf{\Delta})^{1/2}.$$

The result follows by using (E.11), (E.12), and (E.13) together and adding and subtracting terms and noting that by construction the orthogonal matrix from (E.13) commutes with $\left(|\widetilde{\mathbf{\Lambda}}|\right)^{-1/2}$.

We note that the matrix $\mathbf{Q}_{\mathbf{X}}$ is invariant to the sparsity factor, since the eigenvectors of $\alpha_n \mathbf{P} = \alpha_n \mathbf{X}\mathbf{I}_{p,q}\mathbf{X}^{\top}$ are the same as those of $\mathbf{P}$ and the eigenvalues are scaled by α_n so that if $\mathbf{D}$ are the eigenvalues of $\alpha_n \mathbf{P}$ then

$$\begin{aligned}\mathbf{U}_{\mathbf{X}}|\mathbf{D}|^{\mathbf{1/2}} &= \sqrt{\alpha_n}\mathbf{U}_{\mathbf{X}}|\mathbf{\Lambda}_{\mathbf{X}}|^{\mathbf{1/2}} \\ &= \sqrt{\alpha_n}\mathbf{X}\mathbf{Q}_{\mathbf{X}}^{\mathbf{-1}},\end{aligned}$$

showing that the matrix $\mathbf{Q}_{\mathbf{X}}$ depends only on the matrix $\mathbf{P}$ and not the sparsity component α_n. □

Proof of Lemma 53

Proof of Lemma 53. First, for any fixed matrix $\mathbf{X}$, we have that the $\mathbf{A}_{ij}$'s are independent random variables, and recall that

$$\widehat{\alpha}_n = \frac{1}{\binom{n}{2}} \sum_{i<j} A_{ij}.$$

Define $\theta_n := \frac{\alpha_n}{\binom{n}{2}} \sum_{i<j} \mathbf{P}_{ij}$. Then $\mathbb{E}(\widehat{\alpha}_n|\mathbf{X}) = \theta_n$. Therefore, we have that

$$\mathbb{P}\Big(|\widehat{\alpha}_n - \alpha_n| > 2t\Big) \leq \mathbb{P}(|\widehat{\alpha}_n - \theta_n| > t) + \mathbb{P}(|\theta_n/\alpha_n - 1| > t/\alpha_n).$$

For the first, term, we note that by applying Hoeffding's inequality, we see that

$$\begin{aligned}\mathbb{P}\bigg(|\widehat{\alpha}_n - \theta_n| \geq t\bigg) &\leq 2\exp\bigg(-2\binom{n}{2}t^2\bigg) \\ &\leq 2\exp\bigg(-(n^2+n)t^2\bigg) \\ &\leq 2\exp\bigg(-\frac{n^2t^2}{2}\bigg).\end{aligned}$$

For the second term, note that $\frac{1}{\binom{n}{2}}\sum_{i<j} X_i^\top \mathbf{I}_{p,q} X_j$ is a U-statistic with expected value 1. Hoeffding's inequality for U-statistics (e.g. Example 2.23 in Wainwright (2019)) shows that

$$\mathbb{P}\bigg(|\frac{1}{\binom{n}{2}}\sum_{i<j} X_i^\top \mathbf{I}_{p,q} X_j - 1| \geq t\bigg) \leq 2\exp(-nt^2/8).$$

Hence, we see that

$$\mathbb{P}\bigg(|\widehat{\alpha}_n - \alpha_n| > 2t\bigg) \leq 2\exp\bigg(-\frac{n^2t^2}{2}\bigg) + 2\exp(-\frac{nt^2}{8\alpha_n^2})$$

Set $t = 4\sqrt{\frac{\alpha_n \log(n)}{n}}$. Then recalling that for some $C > 0$ $1 \geq \alpha_n \geq C\log^4(n)/n$, we have

$$\begin{aligned}\mathbb{P}\bigg(|\widehat{\alpha}_n - \alpha_n| > 8\sqrt{\frac{\alpha_n \log(n)}{n}}\bigg) &\leq 2\exp\bigg(-8n\alpha_n\log(n)\bigg) + 2\exp\bigg(-2\log(n)\alpha_n^{-1}\bigg) \\ &\leq 2\exp\bigg(-8C\log^5(n)\bigg) + 2n^{-2} \\ &\leq 4n^{-2}.\end{aligned}$$

Now, define the event $\mathcal{A} := \{|\widehat{\alpha}_n - \alpha_n| \leq 4\sqrt{\frac{\alpha_n\log(n)}{n}}\}$. On $\mathcal{A}$, since $\alpha_n \geq C\frac{\log^4(n)}{n}$,

$$\frac{|\widehat{\alpha}_n - \alpha_n|}{\alpha_n} \leq \frac{4}{\log^{1.5}(n)},$$

which is small for n sufficiently large. Hence, by Taylor expansion, we have that

$$\begin{aligned}\frac{1}{\sqrt{\widehat{\alpha}_n}} &= \frac{1}{\sqrt{\alpha_n + (\widehat{\alpha}_n - \alpha_n)}} \\ &= \frac{1}{\sqrt{\alpha_n}}\left(\sqrt{1 + \frac{\widehat{\alpha}_n - \alpha_n}{\alpha_n}}\right)^{-1} \\ &= \frac{1}{\sqrt{\alpha_n}}\left(1 + \frac{1}{2}\frac{\widehat{\alpha}_n - \alpha_n}{\alpha_n} + O\left(\frac{\widehat{\alpha}_n - \alpha_n}{\alpha_n}\right)^2\right).\end{aligned}$$

By the previous observations, we have that this is equal to

$$\frac{1}{\sqrt{\alpha_n}}\left(1 + O\left[\sqrt{\frac{\log(n)}{n\alpha_n}}\right]\right).$$

for n sufficiently large. □

E.2 More on the Discussion in Section 5.3.2

In this section, for f and g two functions of n, we write $f(n) \ll g(n)$ if $f(n)/g(n) \to 0$ as n tends to infinity.

The eigenvalues of $\alpha_n \mathbf{X}\mathbf{X}^\top$ are the same as those of $\alpha_n \mathbf{X}^\top \mathbf{X}$, and as $n \to \infty$, the matrix $\frac{1}{n}\mathbf{X}^\top \mathbf{X}$ is converging almost surely to $\mathbb{E}(XX^\top)$. Therefore, as $n \to \infty$,

$$\frac{1}{n\alpha_n}\big(\lambda_i(\mathbf{P}) - \lambda_{i+1}(\mathbf{P})\big) \to \delta_i,$$

where $\delta_i = \lambda_i(\mathbb{E}(XX^\top)) - \lambda_{i+1}(\mathbb{E}(XX^\top))$. We have

$$\|\mathbf{W}_* - \mathbf{I}\|_F \leq \|\mathbf{W}_* - \mathbf{U}_\mathbf{A}^\top \mathbf{U}_\mathbf{P}\|_F + \|\mathbf{U}_\mathbf{A}^\top \mathbf{U}_\mathbf{P} - \mathbf{I}\|_F.$$

The first term can be bounded directly using the Davis-Kahan Theorem as

$$\|\mathbf{W}_* - \mathbf{U}_\mathbf{A}^\top \mathbf{U}_\mathbf{P}\|_F = O\left(\frac{\|\mathbf{A} - \mathbf{P}\|}{\lambda_d(\mathbf{P})}\right)^2 = O((n\alpha_n)^{-1}).$$

For the second term, following the analysis on Page 24 of Rubin-Delanchy et al. (2020), we

have that for $i \neq j$,

$$(\mathbf{U}_{\mathbf{A}}^\top \mathbf{U}_{\mathbf{P}})_{ij} = -\frac{(\mathbf{U}_{\mathbf{A}})_i^\top (\mathbf{A}-\mathbf{P})(\mathbf{U}_{\mathbf{P}})_j}{\lambda_j(\mathbf{P}) - \lambda_i(\mathbf{A})}.$$

By previous results on eigenvalue concentration (e.g. Eldridge et al. (2018); O'Rourke et al. (2018); Cape et al. (2017)), we have that

$$|\lambda_i(\mathbf{A}) - \lambda_i(\mathbf{P})| \leq C\log(n)$$

with high probability. Hence, the above bound can be written as

$$\frac{(\mathbf{U}_{\mathbf{A}})_i^\top (\mathbf{A}-\mathbf{P})(\mathbf{U}_{\mathbf{P}})_j}{\lambda_j(\mathbf{P}) - \lambda_i(\mathbf{A})} = \frac{(\mathbf{U}_{\mathbf{A}})_i^\top (\mathbf{A}-\mathbf{P})(\mathbf{U}_{\mathbf{P}})_j}{\lambda_j(\mathbf{P}) - \lambda_i(\mathbf{P}) \pm C\log(n)}$$

Moreover, by Koltchinskii and Gine (2000), we have that $\frac{\lambda_i(\mathbf{P})}{n\alpha_n} - \lambda_i(\mathbb{E}(XX^\top)) = O_{\mathbb{P}}(n^{-1/2})$. Hence, we can further simplify the bound to when $i = j+1$

$$\begin{aligned}\frac{(\mathbf{U}_{\mathbf{A}})_i^\top (\mathbf{A}-\mathbf{P})(\mathbf{U}_{\mathbf{P}})_j}{\lambda_j(\mathbf{P}) - \lambda_i(\mathbf{P}) \pm C\log(n)} &= \frac{(\mathbf{U}_{\mathbf{A}})_i^\top (\mathbf{A}-\mathbf{P})(\mathbf{U}_{\mathbf{P}})_j}{n\alpha_n\lambda_j(\mathbb{E}(XX^\top)) - n\alpha_n\lambda_i(\mathbb{E}(XX^\top)) \pm O_{\mathbb{P}}(\sqrt{n}\alpha_n)} \\ &= \frac{(\mathbf{U}_{\mathbf{A}})_i^\top (\mathbf{A}-\mathbf{P})(\mathbf{U}_{\mathbf{P}})_j}{n\alpha_n\delta \pm O_{\mathbb{P}}(\sqrt{n}\alpha_n) \pm O(\log(n))}.\end{aligned}$$

Expanding the numerator, we see that we can write this via

$$\begin{aligned}&\Big(n\alpha_n\delta \pm O_{\mathbb{P}}(\sqrt{n}\alpha_n) \pm O(\log(n))\Big)^{-1}\Big((\mathbf{U}_{\mathbf{A}})_i^\top (\mathbf{A}-\mathbf{P})(\mathbf{U}_{\mathbf{P}})_j\Big) \\ &\quad= \Big(n\alpha_n\delta \pm O_{\mathbb{P}}(\sqrt{n}\alpha_n) \pm O(\log(n))\Big)^{-1}\Big[(\mathbf{U}_{\mathbf{A}})_i^\top (\mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^\top)(\mathbf{A}-\mathbf{P})(\mathbf{U}_{\mathbf{P}})_j\Big] \\ &\qquad+ \Big(n\alpha_n\delta \pm O_{\mathbb{P}}(\sqrt{n}\alpha_n) \pm O(\log(n))\Big)^{-1}\Big[(\mathbf{U}_{\mathbf{A}})_i^\top (\mathbf{I}-\mathbf{U}_{\mathbf{P}}\mathbf{U}_{\mathbf{P}}^\top)(\mathbf{A}-\mathbf{P})(\mathbf{U}_{\mathbf{P}})_j\Big].\end{aligned}$$

The first term $\mathbf{U}_{\mathbf{P}}^\top(\mathbf{A}-\mathbf{P})(\mathbf{U}_{\mathbf{P}})_j$ is a sum of independent random variables, so Hoeffding's inequality reveals that it is of order at most $\log(n)$ with high probability. The second term

can be bounded via the Davis-Kahan theorem as

$$\begin{aligned}\left[(\mathbf{U_A})_i^\top(\mathbf{I}-\mathbf{U_P}\mathbf{U_P}^\top)(\mathbf{A}-\mathbf{P})(\mathbf{U_P})_j\right] &\leq \|(\mathbf{U_A})_i^\top(\mathbf{I}-\mathbf{U_P}\mathbf{U_P}^\top)\|\|(\mathbf{A}-\mathbf{P})\| \\ &\leq \frac{C\|\mathbf{A}-\mathbf{P}\|}{\lambda_d(\mathbf{P})}\|(\mathbf{A}-\mathbf{P})\| \\ &= O(1).\end{aligned}$$

Putting it together, we arrive at

$$\|\mathbf{U_A}^\top\mathbf{U_P}-\mathbf{I}\|_F = O\left(\frac{\log(n)}{n\alpha_n\delta}\right),$$

where $\delta = \min_i \delta_i$, provided $\sqrt{n}\alpha_n \ll n\alpha_n\delta$ and $\log(n) \ll n\alpha_n\delta$.

From an asymptotic standpoint, the term δ in the denominator makes little difference as it is a constant, and $n\alpha_n \to \infty$. However, for finite n, depending on α_n and n, even if $n\alpha_n \gg \log^4(n)$, the constant may depend on δ. Therefore, for any fixed model, though the rate of convergence depends on $n\alpha_n$, for all practical purposes it also depends on the eigengap δ.

Appendix F

Proofs from Chapter 6

F.1 Proofs of Identifiability, Algorithm Recovery Results, and Theorem 18

In this section we prove Theorem 15 and Proposition 9, as well as Lemma 7 and Lemma 8.

F.1.1 Proof of Theorem 15

Proof of Theorem 15. We first prove the "if" direction. Suppose for contradiction that there is another block membership matrix $\widetilde{\mathbf{Z}} \in \{0,1\}^{n \times K'}$ with at least one vertex assigned to each community, and positive diagonal matrices $\{\widetilde{\mathbf{\Theta}}^{(l)}\}_{l=1}^{L}$ and symmetric matrices $\{\widetilde{\mathbf{B}}^{(l)}\}_{l=1}^{L}$ such that

$$\mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}\mathbf{Z}^\top\mathbf{\Theta}^{(l)} = \widetilde{\mathbf{\Theta}}^{(l)}\widetilde{\mathbf{Z}}\widetilde{\mathbf{B}}^{(l)}\widetilde{\mathbf{Z}}^\top\widetilde{\mathbf{\Theta}} \qquad \text{for each } l \in [L].$$

Equivalently, since the matrices $\widetilde{\mathbf{\Theta}}^{(l)}$ have positive diagonal, for all $l \in [L]$ it holds that

$$\begin{aligned}\widetilde{\mathbf{Z}}\widetilde{\mathbf{B}}^{(l)}\widetilde{\mathbf{Z}}^\top =& [\widetilde{\mathbf{\Theta}}^{(l)}]^{-1}\mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}\mathbf{Z}^\top[\widetilde{\mathbf{\Theta}}^{(l)}]^{-1}\mathbf{\Theta}^{(l)} \\ :=& \mathbf{\Gamma}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}(\mathbf{\Gamma}^{(l)}\mathbf{Z})^\top \\ =& \mathbf{\Gamma}^{(l)}\mathbf{Z}\mathbf{V}^{(l)}\mathbf{D}^{(l)}(\mathbf{\Gamma}^{(l)}\mathbf{Z}\mathbf{V}^{(l)})^\top. \end{aligned} \tag{F.1}$$

For any vertex index $i \in [n]$, denote by $z(i)$ and $\widetilde{z}(i)$ the community memberships according to $\mathbf{Z}$ and $\widetilde{\mathbf{Z}}$. We will show that $K' \geq K$ and if $K' = K$ then $z(i) = z(j)$ if and only if

$\widetilde{z}(i) = \widetilde{z}(j)$.

By the RHS of Eq. (F.1), the column space of $\mathbf{\Gamma}^{(l)}\mathbf{Z}\mathbf{V}^{(l)}$ should be contained within the column space of $\widetilde{\mathbf{Z}}$ (as these two matrices are full rank by construction), and hence, there is a matrix $\widetilde{\mathbf{M}}^{(l)} \in \mathbb{R}^{K \times K_l}$ such that

$$\mathbf{\Gamma}^{(l)}\mathbf{Z}\mathbf{V}^{(l)} = \widetilde{\mathbf{Z}}\widetilde{\mathbf{M}}^{(l)}, \qquad \text{for all } l \in [L]. \tag{F.2}$$

In particular, this implies that for any $i \in [n]$,

$$\mathbf{\Gamma}^{(l)}_{ii}\mathbf{V}^{(l)}_{z(i)\cdot} = \widetilde{\mathbf{M}}^{(l)}_{\widetilde{z}(i)\cdot} \qquad \text{for all } l \in [L]. \tag{F.3}$$

If $\widetilde{z}(i) = \widetilde{z}(j)$ then

$$\mathbf{\Gamma}^{(l)}_{ii}\mathbf{V}^{(l)}_{z(i)\cdot} = \mathbf{\Gamma}^{(l)}_{jj}\mathbf{V}^{(l)}_{z(j)\cdot} \qquad \text{for all } l \in [L].$$

This equation implies that the normalized rows are the same, i.e., $\mathbf{Q}^{(l)}_{z(i)\cdot} = \mathbf{Q}^{(l)}_{z(j)\cdot}$ for all $l \in [L]$, and hence $\mathbf{Q}_{z(i)\cdot} = \mathbf{Q}_{z(j)\cdot}$, which is only possible if $z(i) = z(j)$ according to the condition in the proposition.

Now, take a set of vertices $\mathcal{T} \subset [n]$ such that each vertex is in a different community according to $\widetilde{\mathbf{Z}}$. Without loss of generality, suppose that $\widetilde{\mathbf{Z}}_{\mathcal{T}\cdot} = I$, and hence, Eq. (F.2) implies

$$(\mathbf{\Gamma}^{(l)}\mathbf{Z}\mathbf{V}^{(l)})_{\mathcal{T}\cdot} = \mathbf{\Gamma}^{(l)}_{\mathcal{T}\cdot}\mathbf{V}^{(l)}_{z(\mathcal{T})\cdot} = \mathbf{M}^{(l)} \qquad \text{for all } l \in [L].$$

If there are two indexes $i, j \in \mathcal{T}$ such that $z(i) = z(j)$, then the corresponding rows of $\mathbf{M}^{(l)}$ are proportional, that is $\mathbf{M}_{z(i)\cdot} = \mathbf{\Gamma}^{(l)}_{ii}\mathbf{V}^{(l)}_{z(i)\cdot}$ and $\mathbf{M}_{z(j)\cdot} = \mathbf{\Gamma}^{(l)}_{jj}\mathbf{V}^{(l)}_{z(i)\cdot}$ for all $l \in [L]$. If $K' = K$, this implies that $\mathbf{M}^{(l)}$ can only have at most $K - 1$ different rows that are not proportional, and these are the same for all $l \in [L]$. Hence, by Eq. (F.3) the matrix $\mathbf{Q}$ has at most $K - 1$ different rows, which contradicts the assumption. Note that this is also the case if $K' < K$. If $K' > K$, then it is still possible to have $z(i) = z(j)$, but then $\mathbf{Z}$ can fit the same model with fewer communities.

We now prove the "only if" direction. Suppose for contradiction that $\mathbf{Q}$ has repeated rows; we will construct $\widetilde{\mathbf{Z}}$ and $\mathbf{B}^{(l)}$ that yield the same $\mathbf{P}^{(l)}$ matrices. Without loss of generality

we may assume that rows one and two are repeated, since communities are identifiable up to permutation. Furthermore, without loss of generality we can have $\mathbf{Q}^{(l)} = \mathbf{V}^{(l)}$. Indeed, for $i \in \mathcal{C}(r)$, we can rescale $\theta_i^{(l)}$ via $\theta_i^{(l)} \mapsto \theta_i^{(l)} \|\mathbf{V}_{r\cdot}^{(l)}\|$, which still yields the same matrix $\mathbf{P}^{(l)}$ since

$$\mathbf{P}_{ij}^{(l)} = \theta_i^{(l)}\theta_j^{(l)}(\mathbf{V}^{(l)}\mathbf{D}^{(l)}\mathbf{V}^{(l)})_{z(i)z(j)}^{\top} = \big(\theta_i^{(l)}\|\mathbf{V}_{z(i)\cdot}^{(l)}\|\big)\big(\theta_j^{(l)}\|\mathbf{V}_{z(j)\cdot}^{(l)}\|\big)\frac{(\mathbf{V}^{(l)}\mathbf{D}^{(l)}\mathbf{V}^{(l)})_{z(i)z(j)}^{\top}}{\|\mathbf{V}_{z(i)\cdot}^{(l)}\|\|\mathbf{V}_{z(j)\cdot}^{(l)}\|}.$$

Therefore, the first two rows of $\mathbf{V}^{(l)}$ are repeated for all l. However, this implies that

$$\mathbf{B}_{1r}^{(l)} = \big(\mathbf{V}^{(l)}\mathbf{D}^{(l)}(\mathbf{V}^{(l)})^{\top}\big)_{1r} = \sum_{s=1}^{K_l}\mathbf{V}_{1s}^{(l)}\mathbf{D}_{ss}^{(l)}\mathbf{V}_{rs}^{(l)} = \sum_{s=1}^{K_l}\mathbf{V}_{2s}^{(l)}\mathbf{D}_{ss}^{(l)}\mathbf{V}_{rs}^{(l)} = \mathbf{B}_{2r}^{(l)},$$

which shows that the first row and column of $\mathbf{B}^{(l)}$ is repeated. Therefore, we can collapse the first two communities into one community, creating a new matrix $\widetilde{\mathbf{B}}^{(l)}$ with $K-1$ communities (with the first two communities merged). Then we have that

$$\mathbf{P}_{ij}^{(l)} = \theta_i^{(l)}\theta_j^{(l)}\mathbf{B}_{z(i)z(j)}^{(l)} = \theta_i^{(l)}\theta_j^{(l)}\mathbf{B}_{\widetilde{z}(i)\widetilde{z}(j)}^{(l)},$$

which shows that $\mathbf{Z}$ is not identifiable unless $\mathbf{Q}$ has no repeated rows. □

F.1.2 Proof of Proposition 9

Proof of Proposition 9. We will demonstrate that the left singular vectors obtained immediately before clustering contain exactly K unique rows, for which the final result follows. We will analyze each stage separately.

First Stage (individual network embedding): First, suppose that $\mathbf{P}^{(l)} = \mathbf{U}^{(l)}\Lambda^{(l)}(\mathbf{U}^{(l)})^{\top}$ with $\mathbf{U} \in \mathbb{R}^{n\times K_l}$ is the eigendecomposition of $\mathbf{P}^{(l)}$, and let $\mathbf{B}^{(l)} = \mathbf{V}^{(l)}\mathbf{D}^{(l)}(\mathbf{V}^{(l)})^{\top}$ be the eigendecomposition of $\mathbf{B}^{(l)}$, with $\mathbf{V} \in \mathbb{R}^{K\times K_l}$ a matrix with orthogonal columns and $\mathbf{D}^{(l)} \in \mathbb{R}^{K_l\times K_l}$ a diagonal matrix with non-zero elements in the diagonal. From this factor-

ization it is evident that

$$\mathbf{P}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}\mathbf{Z}^\top\mathbf{\Theta}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{V}^{(l)}\mathbf{D}^{(l)}(\mathbf{V}^{(l)})^\top\mathbf{Z}^\top\mathbf{\Theta}^{(l)}.$$

Since $\mathbf{U}^{(l)}$ and $\mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{V}^{(l)}$ are full rank matrices, they have the same column space, so

$$\mathbf{U}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{V}^{(l)}\mathbf{H}^{(l)},$$

where $\mathbf{H}^{(l)} \in \mathbb{R}^{K_l \times K_l}$ is a full rank matrix. From this decomposition it is immediate that $\mathbf{U}^{(l)}$ consists of rows of $\mathbf{V}^{(l)}\mathbf{H}^{(l)}$ with each row of $\mathbf{U}^{(l)}$ scaled by $\theta_i^{(l)}$. Let $\xi_r^{(l)}$ denote the r'th row of $\mathbf{V}\mathbf{H}|\Lambda^{(l)}|^{1/2}$. Then if $z(i) = r$,

$$(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{1/2})_{i\cdot} = \theta_i^{(l)}\xi_r^{(l)} = \theta_i^{(l)}\mathbf{V}_{r\cdot}^{(l)}\mathbf{H}^{(l)},$$

and hence

$$\mathbf{Y}_{i\cdot}^{(l)} = \frac{\theta_i^{(l)}\xi_r^{(l)}}{\|\theta_i^{(l)}\xi_r^{(l)}\|} = \frac{1}{\|\xi_r^{(l)}\|}\xi_r^{(l)} = \frac{1}{\|\mathbf{V}_{r\cdot}^{(l)}\mathbf{H}^{(l)}\|}\mathbf{V}_{r\cdot}^{(l)}\mathbf{H}^{(l)},$$

which does not depend on $\theta_i^{(l)}$.

Second Stage (joint network embedding): We now consider the left singular vectors of the matrix $\mathcal{Y}$ defined as

$$\mathcal{Y} = [\mathbf{Y}^{(1)}, \cdots, \mathbf{Y}^{(L)}].$$

Observe that the leading K left singular vectors $\mathbf{U}$ of $\mathcal{Y}$ are given by the leading K eigenvectors of the matrix $\mathcal{Y}\mathcal{Y}^\top$, which can equivalently be written as

$$\mathcal{Y}\mathcal{Y}^\top = \sum_{l=1}^{L}\mathbf{Y}^{(l)}(\mathbf{Y}^{(l)})^\top.$$

Consider i and j in community r and s respectively. Then from the analysis in the previous

step,

$$(\mathcal{Y}\mathcal{Y}^\top)_{ij} = \sum_{l=1}^{L} \frac{\langle \xi_r^{(l)}, \xi_s^{(l)} \rangle}{\|\xi_r^{(l)}\| \|\xi_s^{(l)}\|}.$$

Consequently, this shows that $\mathcal{Y}\mathcal{Y}^\top$ is a matrix of the form

$$\mathcal{Y}\mathcal{Y}^\top = \mathbf{Z}\Big(\sum_{l=1}^{L} (\Xi^{(l)})(\Xi^{(l)})^\top \Big) \mathbf{Z}^\top = (\mathbf{Z}\boldsymbol{\Xi})(\mathbf{Z}\boldsymbol{\Xi})^\top,$$

where $\Xi^{(l)}$ is the matrix whose rows are $\xi_r^{(l)}/\|\xi_r^{(l)}\|$ and $\boldsymbol{\Xi} = [\Xi^{(1)} \cdots \Xi^{(l)}]$. Next observe that

$$\Xi^{(l)} = \mathbf{D}^{(l)}\mathbf{Q}^{(l)}\mathbf{H}^{(l)} = \mathbf{Q}^{(l)}\mathbf{M}^{(l)}$$

for some matrix $\mathbf{M}^{(l)}$ that is full rank, where $\mathbf{Q}^{(l)}$ is as in Theorem 15. Since $\mathbf{Q}$ has K different rows (by assumption), $\boldsymbol{\Xi}$ has K different rows, and hence $\boldsymbol{\Xi}\boldsymbol{\Xi}^\top$ is a $K \times K$ block matrix. Let $\mathbf{U}$ denote the leading $\widetilde{K}$ eigenvectors of $\mathcal{Y}\mathcal{Y}^\top$, where $\widetilde{K}$ is the rank of $\mathcal{Y}$. Let $\mathbf{V}\boldsymbol{\Gamma}\mathbf{V}^\top$ denote the eigendecomposiion of $(\mathbf{Z}^\top\mathbf{Z})^{1/2}\boldsymbol{\Xi}\boldsymbol{\Xi}^\top(\mathbf{Z}^\top\mathbf{Z})^{1/2}$. Then it is straightforward to see that $\mathbf{U} = \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1/2}\mathbf{V}$ since they both have orthonormal columns. It suffices to argue that $\mathbf{V}$ does not have repeated rows. Assuming this for the moment, by taking $\mathbf{M} = (\mathbf{Z}^\top\mathbf{Z})\mathbf{V}$, it holds that $\mathbf{U} = \mathbf{Z}\mathbf{M}$, with $\mathbf{M}$ having no repeated rows, whence the result is proven.

It remains to argue that $\mathbf{V}$ does not have repeated rows. Under the conditions of Theorem 15 we have already shown that $\boldsymbol{\Xi}$ does not have repeated rows. Hence $\boldsymbol{\Xi}\boldsymbol{\Xi}^\top$ is a block matrix with no repeated rows and columns, and hence $(\mathbf{Z}^\top\mathbf{Z})^{1/2}\boldsymbol{\Xi}\boldsymbol{\Xi}^\top(\mathbf{Z}^\top\mathbf{Z})^{1/2}$ is also a block matrix with no repeated rows and columns. Now assume for contradiction that $\mathbf{V}$ has repeated rows. This implies that $\mathbf{V} = \widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}$ for some matrices $\widetilde{\mathbf{Z}} \in \{0,1\}^{K \times \widetilde{K}}$ and $\widetilde{\mathbf{V}} \in \mathbb{R}^{\widetilde{K} \times \widetilde{K}}$ a full rank matrix. Suppose that $\mathbf{V}$ has rows r and r' repeated, and without loss of generality suppose that row is the first row of $\widetilde{\mathbf{V}}$ (or else permute $\widetilde{\mathbf{V}}$), so that $\widetilde{\mathbf{Z}}_{r1} = \widetilde{\mathbf{Z}}_{r'1} = 1$. Then

from the equation $(\mathbf{Z}^\top\mathbf{Z})^{1/2}\mathbf{\Xi}\mathbf{\Xi}^\top(\mathbf{Z}^\top\mathbf{Z})^{1/2} = \widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}\Gamma\widetilde{\mathbf{V}}^\top\widetilde{\mathbf{Z}}^\top$, it holds that for all $1 \le s \le K$,

$$\begin{aligned}\frac{1}{\sqrt{n_r n_s}}\langle \xi_{r\cdot}, \xi_{s\cdot}\rangle &= \langle \left(\widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}\Gamma^{1/2}\right)_{r\cdot}, \left(\widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}\Gamma^{1/2}\right)_{s\cdot}\rangle \\ &= \langle \left(\widetilde{\mathbf{V}}\Gamma^{1/2}\right)_{1\cdot}, \left(\widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}\Gamma^{1/2}\right)_{s\cdot}\rangle \\ &= \langle \left(\widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}\Gamma^{1/2}\right)_{r'\cdot}, \left(\widetilde{\mathbf{Z}}\widetilde{\mathbf{V}}\Gamma^{1/2}\right)_{s\cdot}\rangle \\ &= \frac{1}{\sqrt{n_{r'} n_s}}\langle \xi_{r'\cdot}, \xi_{s\cdot}\rangle.\end{aligned}$$

Consequently, since the above identity holds for all s, this shows that the r and r''th rows and columns of $(\mathbf{Z}^\top\mathbf{Z})^{1/2}\mathbf{\Xi}\mathbf{\Xi}^\top(\mathbf{Z}^\top\mathbf{Z})^{1/2}$ are identical. However, this is a contradiction, which completes the proof. □

F.1.3 Proof of Lemma 7

We will restate Lemma 7 for convenience.

Lemma 7 (Population Properties: Stage I)**.** *Suppose Assumption 6.1 holds, and let* $\lambda_r^{(l)}$ *denote the eigenvalues of* $\mathbf{P}^{(l)}$ *and let* $\lambda_r(\mathbf{B}^{(l)})$ *denote the eigenvalues of* $\mathbf{B}^{(l)}$*. Then for all* $1 \le r \le K$*,*

$$\begin{aligned}\theta_i^{(l)} \lesssim \|\mathbf{X}_{i\cdot}^{(l)}\| &\lesssim \theta_i^{(l)}; \\ \|\mathbf{U}_{i\cdot}^{(l)}\| &\lesssim \sqrt{K}\frac{\theta_i^{(l)}}{\|\theta^{(l)}\|}; \\ \lambda_r^{(l)} &\asymp \frac{\|\theta^{(l)}\|^2}{K}\lambda_r(\mathbf{B}^{(l)}).\end{aligned}$$

Proof of Lemma 7. Define the matrix

$$\mathbf{G}^{(l)} := K\|\theta^{(l)}\|^{-2}\mathrm{diag}(\|\theta^{(l)}_{\mathcal{C}(1)}\|, \ldots, \|\theta^{(l)}_{\mathcal{C}(K)}\|)\mathbf{B}^{(l)}\mathrm{diag}(\|\theta^{(l)}_{\mathcal{C}(1)}\|, \ldots, \|\theta^{(l)}_{\mathcal{C}(K)}\|).$$

Letting $\lambda_r(\cdot)$ denote the eigenvalues of a matrix, by Ostrowski's Theorem (Theorem 4.5.9 of Horn and Johnson (2012) and Assumption 6.1, the eigenvalues of $\mathbf{G}^{(l)}$ satisfy $\lambda_r(\mathbf{G}^{(l)}) \asymp \lambda_r(\mathbf{B}^{(l)})$. Since the eigenvalues of $\mathbf{P}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}\mathbf{Z}^\top\mathbf{\Theta}^{(l)}$ are the same as the eigenvalues

of the matrix

$$\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\mathbf{B}^{(l)}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2},$$

we have

$$\lambda_r\bigg(\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\mathbf{B}^{(l)}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\bigg) \asymp \frac{\|\theta^{(l)}\|^2}{K}\lambda_r\big(\mathbf{G}^{(l)}\big) \asymp \frac{\|\theta^{(l)}\|^2}{K}\lambda_r(\mathbf{B}^{(l)}).$$

To prove the other two assertions, we first observe that

$$\mathbf{P}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{-1/2}\bigg[\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\mathbf{B}^{(l)}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\bigg]\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{-1/2}\mathbf{Z}^\top\mathbf{\Theta}^{(l)}.$$

Suppose that the matrix $\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\mathbf{B}^{(l)}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}$ has singular value decomposition $\widetilde{\mathbf{U}}\widetilde{\Sigma}\widetilde{\mathbf{V}}^\top$. Then it holds that

$$\mathbf{P}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{-1/2}\widetilde{\mathbf{U}}\widetilde{\Sigma}\widetilde{\mathbf{V}}^\top\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{-1/2}\mathbf{Z}\mathbf{\Theta}^{(l)}.$$

However, since the columns of the matrix $\mathbf{\Theta}^{(l)}\mathbf{Z}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{-1/2}\widetilde{\mathbf{U}}$ are orthonormal, the decomposition above is a valid singular value decomposition for $\mathbf{P}^{(l)}$. Since $\mathbf{P}^{(l)}$ is symmetric, its eigenvectors coincide with its singular vectors up to an orthogonal transformation, which in particular shows that there exists an orthogonal transformation $\mathbf{W}$ such that

$$\mathbf{U}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{-1/2}\widetilde{\mathbf{U}}\mathbf{W}.$$

We immediately obtain the bound

$$\|\mathbf{U}^{(l)}_{i\cdot}\| = \theta_i^{(l)}\frac{1}{\|\theta^{(l)}_{\mathcal{C}(z(i))}\|} \lesssim \theta_i^{(l)}\frac{\sqrt{K}}{\|\theta^{(l)}\|},$$

where we have used the fact that $\|\theta_{\mathcal{C}(r)}\|^2 \asymp \frac{\|\theta^{(l)}\|^2}{K}$ for all r. Similarly, it holds that

$$\|\mathbf{X}^{(l)}_{i\cdot}\| \le \|\mathbf{U}^{(l)}_{i\cdot}\|\||\Lambda^{(l)}|^{1/2}\| \lesssim \theta_i^{(l)}.$$

It remains to provide a lower bound on $\mathbf{X}_{i\cdot}$. By the fact that the singular values of a

symmetric matrix are the absolute value of its eigenvalues, it holds that

$$\begin{aligned}\mathbf{U}^{(l)}|\Lambda^{(l)}|^{1/2} &= \mathbf{U}^{(l)}\widetilde{\Sigma}^{1/2} \\ &= \mathbf{\Theta}^{(l)}\mathbf{Z}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{-1/2}\widetilde{\mathbf{U}}\mathbf{W}\widetilde{\Sigma}^{1/2}.\end{aligned}$$

Observe that $\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{-1/2}\widetilde{\mathbf{U}} = \mathbf{B}^{(l)}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\widetilde{\mathbf{V}}\widetilde{\Sigma}^{-1}$, which shows that

$$\begin{aligned}\mathbf{U}^{(l)}|\Lambda^{(l)}|^{1/2} &= \mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\widetilde{\mathbf{V}}\widetilde{\Sigma}^{-1}\mathbf{W}\widetilde{\Sigma}^{1/2} \\ &= \mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{B}^{(l)}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\widetilde{\mathbf{V}}|\Lambda^{(l)}|^{-1}\mathbf{W}|\Lambda^{(l)}|^{1/2}.\end{aligned}$$

Consider a given row i and suppose that $z(i) = r$. Then by Ostrowki's Theorem again,

$$\begin{aligned}\|\mathbf{X}_{i\cdot}^{(l)}\| &\geq \theta_i^{(l)}\|\mathbf{B}_{r\cdot}^{(l)}\|\sigma_{\min}\bigg(\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\widetilde{\mathbf{V}}|\Lambda^{(l)}|^{-1}\mathbf{W}|\Lambda^{(l)}|^{1/2}\bigg) \\ &\geq \theta_i^{(l)}\|\mathbf{B}_{r\cdot}^{(l)}\|\sigma_{\min}\big(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z}\big)^{1/2}\sigma_{\min}\big(|\Lambda^{(l)}|^{-1}\mathbf{W}|\Lambda^{(l)}|^{1/2}\big) \\ &\geq \theta_i^{(l)}\|\mathbf{B}_{r\cdot}^{(l)}\|\min_r\|\theta_{\mathcal{C}(r)}\|\sigma_{\min}\big(|\Lambda^{(l)}|^{-1/2}\big) \\ &\geq \theta_i^{(l)}\|\mathbf{B}_{r\cdot}^{(l)}\|\min_r\|\theta_{\mathcal{C}(r)}^{(l)}\|\frac{\sqrt{K}}{\|\theta^{(l)}\|} \\ &\gtrsim \theta_i^{(l)}\|\mathbf{B}_{r\cdot}^{(l)}\| \\ &\gtrsim \theta_i^{(l)},\end{aligned}$$

where the final line follows from the assumption that $\mathbf{B}^{(l)}$ has unit diagonals. This completes the proof. □

F.1.4 Proof of Lemma 8

We restate Lemma 8 for convenience.

Lemma 8 (Population Properties: Stage II)**.** *Suppose that $\mathcal{Y}$ is rank K, and let $\mathcal{Y} = \mathbf{U}\Sigma\mathbf{V}^\top$ be its (rank K) singular value decomposition. Then it holds that*

$$\mathbf{U} = \mathbf{Z}\mathbf{M},$$

where $\mathbf{M} \in \mathbb{R}^{K \times K}$ *is some invertible matrix satisfying*

$$\|\mathbf{M}_{r\cdot} - \mathbf{M}_{s\cdot}\| = \sqrt{n_r^{-1} + n_s^{-1}}.$$

In addition, when $n_{\min} \asymp n_{\max}$*, it holds that*

$$\lambda_Y^2 := \lambda_{\min}\left(\sum_l \mathbf{Y}^{(l)}(\mathbf{Y}^{(l)})^\top\right) \gtrsim \frac{n}{K} L\bar{\lambda}.$$

Proof of Lemma 8. The first part of the proof holds by Lemma 2.1 of Lei and Rinaldo (2015) applied to the matrix $\mathcal{Y}\mathcal{Y}^\top$, which is a block matrix. See also the proof of Proposition 9.

For the second part we proceed as follows. First recall by the proof of Proposition 9 that we can write the matrix $\mathcal{Y}\mathcal{Y}^\top$ as the matrix

$$\mathcal{Y}\mathcal{Y}^\top = \sum_{l=1}^{L} \mathbf{\Xi}^{(l)}(\mathbf{\Xi}^{(l)})^\top,$$

where the matrix $\mathbf{\Xi}^{(l)}$ is defined as follows. First, let $\mathbf{Q}^{(l)}$ be the matrix such that

$$\mathbf{U}^{(l)} = \mathbf{\Theta}^{(l)}\mathbf{Z}\mathbf{Q}^{(l)}.$$

Then the rows of $\Xi^{(l)}$ are equal to the rows of $\mathbf{Q}^{(l)}|\Lambda^{(l)}|^{1/2}$ normalized by their magnitude. It was discussed in the proof of Proposition 9 that the entries of $\mathbf{Q}^{(l)}$ are of order $\frac{1}{\|\theta^{(l)}\|}$. Observe that we can write $\mathbf{Y}^{(l)} = \mathbf{Z}(\widetilde{\mathbf{D}}^{(l)})^{-1}\mathbf{Q}^{(l)}|\Lambda^{(l)}|^{1/2}$, where $\widetilde{\mathbf{D}}^{(l)}$ is the $K \times K$ diagonal matrix of row norms of $\mathbf{Q}^{(l)}|\Lambda^{(l)}|^{1/2}$. Observe that

$$\begin{aligned}
\|\widetilde{\mathbf{D}}^{(l)}\|^2 &= \max_i \|\left(\mathbf{Q}^{(l)}|\Lambda^{(l)}|^{1/2}\right)_{i\cdot}\|^2 \\
&= \max_i \sum_{r=1}^{K} (\mathbf{Q}_{ir}^{(l)})^2 |\lambda_r| \\
&= \max_i \sum_{r=2}^{K} \frac{C}{\|\theta^{(l)}\|^2} \frac{\|\theta^{(l)}\|^2}{K} |\lambda_r(\mathbf{B}^{(l)})| + \frac{C}{\|\theta^{(l)}\|^2}\|\theta^{(l)}\|^2 \\
&\lesssim 1,
\end{aligned}$$

where we have applied Lemma 7 to observe that $\lambda_r \asymp \frac{\|\theta^{(l)}\|^2}{K}\lambda_r(\mathbf{B}^{(l)})$ for $2 \leq r \leq K$ and $\lambda_1 \asymp \|\theta^{(l)}\|^2$, since by Assumption 6.1 that the largest eigenvalue of $\mathbf{B}^{(l)}$ is upper bounded by a constant. Therefore, we have that

$$\begin{aligned}
\lambda_{\min}\Big(\sum_l \mathbf{Y}^{(l)}(\mathbf{Y}^{(l)})^\top\Big) &= \lambda_{\min}\Big(\sum_l \mathbf{Z}(\widetilde{\mathbf{D}}^{(l)})^{-1}\mathbf{Q}^{(l)}|\Lambda^{(l)}|(\mathbf{Q}^{(l)})^\top(\widetilde{\mathbf{D}}^{(l)})^{-1}\mathbf{Z}^\top\Big) \\
&= \lambda_{\min}\Big(\mathbf{Z}^\top\mathbf{Z}\big(\sum_l(\widetilde{\mathbf{D}}^{(l)})^{-1}(\mathbf{Q}^{(l)}|\Lambda^{(l)}|(\mathbf{Q}^{(l)})^\top(\widetilde{\mathbf{D}}^{(l)})^{-1}\big)\Big) \\
&\geq \lambda_{\min}(\mathbf{Z}^\top\mathbf{Z})\lambda_{\min}\Big(\sum_l(\widetilde{\mathbf{D}}^{(l)})^{-1}\mathbf{Q}^{(l)}|\Lambda^{(l)}|(\mathbf{Q}^{(l)})^\top(\widetilde{\mathbf{D}}^{(l)})^{-1}\Big) \\
&\gtrsim \frac{n}{K}L\Big(\frac{1}{L}\sum_l \lambda_{\min}\Big[(\widetilde{\mathbf{D}}^{(l)})^{-1}\mathbf{Q}^{(l)}|\Lambda^{(l)}|(\mathbf{Q}^{(l)})^\top(\widetilde{\mathbf{D}}^{(l)})^{-1}\Big]\Big)
\end{aligned}$$

where we have used the fact that $\lambda_{\min}(\mathbf{Z}\top\mathbf{Z}) = n_{\min} \asymp n/K$ and that the term inside the sum is rank K and hence invertible. Consequently, it suffices to show that

$$\lambda_{\min}\Big[(\widetilde{\mathbf{D}}^{(l)})^{-1}\mathbf{Q}^{(l)}|\Lambda^{(l)}|(\mathbf{Q}^{(l)})^\top(\widetilde{\mathbf{D}}^{(l)})^{-1}\Big] \gtrsim \lambda^{(l)}_{\min}.$$

However, by the argument in Lemma 7, it holds that $\mathbf{Q}^{(l)}(\mathbf{Q}^{(l)})^\top = (\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z})^{-1}$. Set $\mathbf{G}^{(l)} := K^{-1}\|\theta^{(l)}\|^2(\mathbf{Z}^\top(\mathbf{\Theta}^{(l)})^2\mathbf{Z})$. By Assumption 6.1, $\|(\mathbf{G}^{(l)})^{-1}\| \leq C$ and $\mathbf{Q}^{(l)}(\mathbf{Q}^{(l)})^\top = K\|\theta^{(l)}\|^{-2}(\mathbf{G}^{(l)})^{-1}$. Consequently,

$$\begin{aligned}
\lambda_{\min}\Big[(\widetilde{\mathbf{D}}^{(l)})^{-1}\mathbf{Q}^{(l)}|\Lambda^{(l)}|(\mathbf{Q}^{(l)})^\top(\widetilde{\mathbf{D}}^{(l)})^{-1}\Big] &\gtrsim \lambda_{\min}((\widetilde{\mathbf{D}}^{(l)})^{-1})^2\lambda_{\min}\big(\mathbf{Q}^{(l)}|\Lambda^{(l)}|(\mathbf{Q}^{(l)})^\top\big) \\
&\gtrsim \lambda_{\min}\big((\mathbf{Q}^{(l)})^\top\mathbf{Q}^{(l)}|\Lambda^{(l)}|\big) \\
&\gtrsim \lambda_{\min}\big((\mathbf{Q}^{(l)})^\top\mathbf{Q}^{(l)}\big)\lambda_{\min}(|\Lambda^{(l)}|) \\
&\gtrsim K\|\theta^{(l)}\|^{-2}\lambda_{\min}(\mathbf{G}^{(l)})\frac{\|\theta^{(l)}\|^2}{K}\lambda^{(l)}_{\min} \\
&\gtrsim \lambda^{(l)}_{\min},
\end{aligned}$$

where we have used Assumption 6.1 and Lemma 7 implicitly. This completes the proof. □

F.1.5 Proof of Theorem 18

Proof of Theorem 18. The proof of this result is similar to the proof of the main result. First we demonstrate the initial error implies that each community contains at least $\frac{3}{4}$ of its true members, whereupon we study the empirical centroids and show that they are closer to their true cluster centroid than they are to each other. Finally, instead of applying Theorem 21 to obtain the exponential error rate we apply Theorem 19. As this result only involves a single network, we suppress the dependency on l for ease of notation.

Step 1: Initial Hamming Error

Observe that $\mathbf{Y} = \mathbf{Z}\mathbf{M}_{\mathbf{Y}}$, where it straightforward to check that

$$\sqrt{\lambda_{\min}} \leq \|(\mathbf{M}_{\mathbf{Y}})_{r\cdot} - (\mathbf{M}_{\mathbf{Y}})_{s\cdot}\| \leq 2.$$

The upper bound is immediate; as for the lower bound, we may apply the same argument as in the proof of Lemma 8. Let the matrix $\widehat{\mathbf{X}} := \widehat{\mathbf{Z}}\widehat{\mathbf{M}}_{\mathbf{Y}}$, where $\widehat{\mathbf{Z}}$ and $\widehat{\mathbf{M}}_{\mathbf{Y}}$ are the outputs of $(1+\varepsilon)$ K-means on the rows of $\widehat{\mathbf{Y}}$, and let $S_r := \{i \in \mathcal{C}(r) : \|\mathbf{W}_*\widehat{\mathbf{X}}_{i\cdot} - \mathbf{Y}_{i\cdot}\| \geq \delta_r/2\}$, where $\delta_r = \sqrt{\lambda_{\min}}$. By Lemma 5.3 of Lei and Rinaldo (2015) and a similar argument as in the proof of Theorem 16, it holds that

$$\begin{aligned}\inf_{\mathcal{P}} \sum_{i=1}^{n} \mathbb{I}\{\widehat{z}(i) \neq \mathcal{P}(z(i))\} &\leq \frac{C_\varepsilon}{\lambda_{\min}} \|\widehat{\mathbf{Y}}\mathbf{W}_*^\top - \mathbf{Y}\|_F^2 \\ &\leq \frac{C_\varepsilon n}{\lambda_{\min}} \|\widehat{\mathbf{Y}}\mathbf{W}_*^\top - \mathbf{Y}\|_{2,\infty}^2.\end{aligned}$$

By Corollary 9, with probability at least $1 - O(n^{-15})$ it holds that

$$\begin{aligned}\|\widehat{\mathbf{Y}}\mathbf{W}_*^\top - \mathbf{Y}\|_{2,\infty} &\lesssim \left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2} \frac{K\sqrt{\log(n)}}{\|\theta\|\lambda_{\min}^{1/2}} \\ &\quad + \frac{K^2\theta_{\max}^{(l)}\|\theta^{(l)}\|_1}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^4}\left(\log(n)\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} + \frac{1}{\lambda_{\min}^{(l)}} + \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\right)^{1/2}\frac{K^{5/2}\log(n)}{(\lambda_{\min}^{(l)})^{1/2}}\right) \\ &\leq \frac{\beta}{8\sqrt{C_\varepsilon K}}\lambda_{\min},\end{aligned}$$

where $\beta \in (0,1]$ is such that $n_{\min} \geq \beta n_{\max}$, and where the final bound holds under the conditions of Theorem 18. Let this event be denoted $\mathcal{E}$. By squaring the above bound we arrive at

$$\begin{aligned}\inf_{\mathcal{P}} \sum_{i=1}^{n} \mathbb{I}\{\widehat{z}(i) \neq \mathcal{P}(z(i))\} &\leq n\frac{\beta^2}{64K}\lambda_{\min} \\ &\leq \frac{\beta}{64}\lambda_{\min} n_{\min}.\end{aligned}$$

Therefore, each cluster is associated to a true cluster, denoted as $\widehat{\mathcal{C}}(r)$, where $|\widehat{\mathcal{C}}(r)| \geq (1-\beta\lambda_{\min}/64)n_{\min}$ and $|\widehat{\mathcal{C}}(r)\backslash\mathcal{C}(r)| \leq \frac{\beta\lambda_{\min}}{64} n_{\min}$. Note that since $\beta \in (0,1)$ and $\lambda_{\min} \in (0,1)$, then $\beta\lambda_{\min}/64 < 1$ this is a well-defined fraction.

Step 2: Properties of Empirical Centroids

Recall that we denote $(\widehat{\mathbf{M}}_{\mathbf{Y}})_{r\cdot}$ and $(\mathbf{M}_{\mathbf{Y}})_{r\cdot}$ as the cluster centroids for $\widehat{\mathcal{C}}(r)$ and $\mathcal{C}(r)$ respectively. Then by a similar argument as in the proof of Theorem 16, we have that

$$\begin{aligned}\|\mathbf{W}_*(\widehat{\mathbf{M}}_{\mathbf{Y}})_{r\cdot} - (\mathbf{M}_{\mathbf{Y}})_{r\cdot}\| &\leq \frac{1}{|\widehat{\mathcal{C}}(r)|^{1/2}}\|\widehat{\mathbf{Y}}\mathbf{W}_*^\top - \mathbf{Y}\|_F + 2\frac{|\widehat{\mathcal{C}}(r) \setminus \mathcal{C}(r)|}{|\widehat{\mathcal{C}}(r)|} \\ &\leq \frac{1}{\sqrt{n_{\min}(1-\beta\lambda_{\min}/64)}}\sqrt{n}\|\widehat{\mathbf{Y}}\mathbf{W}_*^\top - \mathbf{Y}\|_{2,\infty} + 2\frac{\beta n_{\min}\lambda_{\min}}{64(1-\beta\lambda_{\min}/64)n_{\min}} \\ &\leq \frac{1}{\sqrt{n_{\min}\beta(1-\beta\lambda_{\min}/64)}}\sqrt{n}\frac{\beta}{8\sqrt{C_\varepsilon K}}\lambda_{\min} + \frac{\lambda_{\min}}{32(1-\beta\lambda_{\min}/64)} \\ &\leq \frac{1}{\sqrt{n_{\min}\beta(1-\beta\lambda_{\min}/64)}}\sqrt{\frac{Kn_{\min}}{\beta}}\frac{\beta}{8\sqrt{C_\varepsilon K}}\lambda_{\min} + \frac{\lambda_{\min}}{32(1-\beta\lambda_{\min}/64)} \\ &\leq \frac{1}{8}\sqrt{\lambda_{\min}},\end{aligned}$$

since $\lambda_{\min} \in (0,1)$ by assumption. The above bound holds on the event $\mathcal{E}$.

Step 3: Applying The Asymptotic Expansion

Arguing similarly as in the proof of Theorem 16, it holds that

$$\mathbb{E}\ell(\widehat{z}, z) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\big(\mathbf{Z}_{i\cdot} \neq \widehat{\mathbf{Z}}_{i\cdot}, \mathcal{E}\big) + O(n^{-15}).$$

On the event $\mathcal{E}$, it holds that

$$\|\mathcal{R}_{\text{Stage I}}\|_{2,\infty} \leq \frac{1}{8}\sqrt{\lambda_{\min}},$$

and hence by repeating the arguments in the proof of Theorem 16,

$$\begin{aligned}\mathbb{P}(\mathbf{Z}_{i\cdot} \neq \widehat{\mathbf{Z}}_{i\cdot}, \mathcal{E}) &\leq \mathbb{P}\big(\|(\widehat{\mathbf{Y}}\mathbf{W}_*^\top)_{i\cdot} - \mathbf{Y}_{i\cdot}\| \geq \frac{1}{4}\sqrt{\lambda_{\min}}, \mathcal{E}\big) \\ &\leq K \max_k \mathbb{P}\bigg\{|e_i^\top \mathcal{L}(\mathbf{E})\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}e_k| \geq \frac{1}{4}\sqrt{\lambda_{\min}/K}\bigg\}.\end{aligned}$$

Here $\mathcal{L}(\mathbf{E})$ is the linear term from Theorem 19 with $\mathbf{E} = \mathbf{A} - \mathbf{P}$. We now apply Bernstein's inequality. The variance v is upper bounded by

$$\begin{aligned}v &\leq \sum_j \theta_i\theta_j \|e_j^\top \mathbf{U}^{(l)}\|^2 \||\Lambda^{(l)}|^{-1/2}\|^2 \|\mathbf{J}(\mathbf{X}_{i\cdot})\|^2 \\ &\lesssim \sum_j \theta_i\theta_j \frac{\theta_j^2 K}{\|\theta\|^2}\frac{K}{\|\theta\|^2\lambda_{\min}}\frac{1}{\theta_i^2} \\ &\lesssim \frac{K^2\|\theta\|_3^3}{\|\theta\|^4\lambda_{\min}\theta_i}.\end{aligned}$$

Similarly,

$$\begin{aligned}\max_j \|e_j^\top \mathbf{U}^{(l)}\|\||\Lambda^{(l)}|^{-1/2}\|\|\mathbf{J}(\mathbf{X}_{i\cdot})\| &\lesssim \theta_j \frac{K}{\|\theta\|^2\lambda_{\min}^{1/2}}\frac{1}{\theta_i} \\ &\lesssim \frac{K\theta_{\max}}{\|\theta\|^2\lambda_{\min}^{1/2}\theta_i}.\end{aligned}$$

By Bernstein's inequality,

$$\begin{aligned}\mathbb{P}\bigg\{|e_i^\top \mathcal{L}(\mathbf{E})\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}e_k| \geq \frac{1}{8}\sqrt{\lambda_{\min}/K}\bigg\} &\leq 2\exp\bigg(-\frac{\frac{1}{128}\frac{\lambda_{\min}}{K}}{C\frac{K^2\|\theta\|_3^3}{\|\theta\|^4\lambda_{\min}\theta_i} + \frac{\lambda_{\min}^{1/2}}{24\sqrt{K}}C\frac{K\theta_{\max}}{\|\theta\|^2\lambda_{\min}^{1/2}\theta_i}}\bigg) \\ &\leq 2\exp\bigg(-c\theta_i \min\bigg\{\frac{\|\theta\|^4\lambda_{\min}^2}{K^3\|\theta\|_3^3}, \frac{\|\theta\|^2\lambda_{\min}}{K^{3/2}\theta_{\max}}\bigg\}\bigg).\end{aligned}$$

Assembling everything together completes the proof. □

F.2 Proof of First Stage Characterization (Theorem 19)

This section contains the full proof of Theorem 19. First, we will restate Theorem 19 here for convenience.

Theorem 19 (Asymptotic Expansion: Stage I)**.** *Suppose that Assumption 6.1 and Assumption 6.2 hold. Fix a given $l \in [L]$. Let $\mathbf{W}_*^{(l)}$ denote the orthogonal matrix satisfing*

$$\mathbf{W}_*^{(l)} := \underset{\mathbf{W} \in \mathbb{O}(K)}{\arg\min} \|\widehat{\mathbf{U}}^{(l)} - \mathbf{U}^{(l)}\mathbf{W}_*^{(l)}\|_F.$$

Then there is an event $\mathcal{E}^{(l)}_{\text{Stage I}}$ with $\mathbb{P}(\mathcal{E}^{(l)}_{\text{Stage I}}) \geq 1 - O(n^{-15})$ such that the following expansion holds:

$$\widehat{\mathbf{Y}}^{(l)}(\mathbf{W}_*^{(l)})^\top - \mathbf{Y}^{(l)} = \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)}) + \mathcal{R}^{(l)}_{\text{Stage I}},$$

where the matrix $\mathcal{R}^{(l)}_{\text{Stage I}}$ satisfies

$$\|\mathcal{R}^{(l)}_{\text{Stage I}}\|_{2,\infty} \lesssim \frac{K^2\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\lambda^{(l)}_{\min}\|\theta^{(l)}\|^4}\left(\log(n)\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}} + \frac{1}{\lambda^{(l)}_{\min}} + \left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\frac{K^{5/2}\log(n)}{(\lambda^{(l)}_{\min})^{1/2}}\right),$$

and the matrix $\mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})$ has rows given by

$$\mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})_{i\cdot} = \frac{1}{\|\mathbf{X}^{(l)}_{i\cdot}\|}\left(\mathbf{I} - \frac{\mathbf{X}^{(l)}_{i\cdot}(\mathbf{X}^{(l)}_{i\cdot})^\top}{\|\mathbf{X}^{(l)}_{i\cdot}\|^2}\right)\left((\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}^{(l)}_{p,q}\right)_{i\cdot}.$$

As an immediate application of Theorem 19, we can obtain a spectral norm concentration bound for the residual, which will be useful in subsequent steps.

Lemma 57. *The residual term $\mathcal{R}^{(l)}_{\text{Stage I}}$ satisfies*

$$\|\mathcal{R}^{(l)}_{\text{Stage I}}\| \lesssim \sqrt{n}\frac{K^2\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\lambda^{(l)}_{\min}\|\theta^{(l)}\|^4}\left(\log(n)\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}} + \frac{1}{\lambda^{(l)}_{\min}} + \left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\frac{K^{5/2}\log(n)}{(\lambda^{(l)}_{\min})^{1/2}}\right).$$

with probability at least $1 - O(n^{-15})$.

The proof of this result follows immediately by noting that $\|\cdot\| \leq \sqrt{n}\|\cdot\|_{2,\infty}$ and the bound in Theorem 19.

We will also use an $\ell_{2,\infty}$ bound for the linear term appearing in Theorem 19 in the proof of Theorem 21.

Lemma 58. *The linear term in Theorem 16 satisfies, with probability at least* $1 - O(n^{-15})$,

$$\|\mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\|_{2,\infty} \lesssim \left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2} \frac{K\sqrt{\log(n)}}{(\lambda^{(l)}_{\min})^{1/2}\|\theta^{(l)}\|}.$$

Proof of Lemma 58. Throughout this proof we suppress the dependence of $\mathbf{\Theta}^{(l)}, \Lambda^{(l)}$ and $\mathbf{U}^{(l)}$ on the index l, and we denote λ via $\frac{1}{\lambda} = \|(\Lambda^{(l)})^{-1}\|$. Define $\mathbf{E} := \mathbf{A}^{(l)} - \mathbf{P}^{(l)}$, so that $\mathbf{E}$ is a mean-zero random matrix.

We will apply the Matrix Bernstein inequality to each row separately. To wit, by Corollary 3.3 of Chen et al. (2021c), we have that with probability at least $1 - O(n^{-16})$ it holds that

$$\|e_i^\top \mathcal{L}(\mathbf{E})\| \lesssim \sqrt{v\log(n)} + w\log(n),$$

where

$$\begin{aligned} v = \max\Bigg\{ &\Bigg\| \sum_j \left(\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}\mathbf{J}(\mathbf{X}_{i\cdot})\right)_{j\cdot}^\top \mathbb{E}\mathbf{E}_{ij}^2 \left(\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}\mathbf{J}(\mathbf{X}_{i\cdot})\right)_{j\cdot} \Bigg\|, \\ &\Bigg\| \sum_j \mathbb{E}\mathbf{E}_{ij} \left(\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}\mathbf{J}(\mathbf{X}_{i\cdot})\right)_{j\cdot} \left(\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}\mathbf{J}(\mathbf{X}_{i\cdot})\right)_{j\cdot}^\top \mathbf{E}_{ij} \Bigg\| \Bigg\}; \\ w = \max_j &\left\| \mathbf{E}_{ij} \left(\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}\mathbf{J}(\mathbf{X}_{i\cdot})\right)_{j\cdot} \right\|. \end{aligned}$$

Since $\mathbb{E}(\mathbf{E}_{ij}^2)$ is a scalar, we have that by Lemma 7,

$$\begin{aligned}
v &\leq \sum_j \mathbb{E}\mathbf{E}_{ij}^2 \left\| \left(\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}\mathbf{J}(\mathbf{X}_{i\cdot}) \right)_{j\cdot} \right\|^2 \\
&\leq \sum_j \theta_i\theta_j \|e_j^\top \mathbf{U}\|^2 \frac{1}{\lambda}\|\mathbf{J}(\mathbf{X}_{i\cdot})\|^2 \\
&\lesssim \sum_j \theta_i\theta_j \frac{\theta_j^2 K}{\|\theta\|^2} \frac{K}{\|\theta\|^2\lambda_{\min}} \frac{1}{\|\mathbf{X}_{i\cdot}\|^2} \\
&\lesssim \sum_j \theta_i\theta_j \frac{\theta_j^2 K}{\|\theta\|^2} \frac{K}{\|\theta\|^2\lambda_{\min}} \frac{1}{\theta_i^2} \\
&\lesssim \sum_j \frac{K^2}{\lambda_{\min}\|\theta\|^4} \left(\frac{\theta_j}{\theta_i}\right) \theta_j^2 \\
&\lesssim \frac{K^2}{\lambda_{\min}\|\theta\|^4} \left(\frac{\theta_{\max}}{\theta_{\min}}\right) \|\theta\|^2 \\
&\lesssim \frac{K^2}{\lambda_{\min}\|\theta\|^2} \left(\frac{\theta_{\max}}{\theta_{\min}}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
w &\leq \max_j \left\| \left(\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}\mathbf{J}(\mathbf{X}_{i\cdot}) \right)_{j\cdot} \right\| \\
&\lesssim \frac{\theta_{\max}\sqrt{K}}{\|\theta\|} \frac{\sqrt{K}}{\|\theta\|\lambda_{\min}^{1/2}} \frac{1}{\theta_i} \\
&\lesssim \left(\frac{\theta_{\max}}{\theta_{\min}}\right) \frac{K}{\|\theta\|^2\lambda_{\min}^{1/2}}.
\end{aligned}$$

Therefore, with probability at least $1 - O(n^{-16})$, we have that

$$\begin{aligned}
\|e_i^\top \mathcal{L}(\mathbf{E})\| &\lesssim \left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2} \frac{K\sqrt{\log(n)}}{\lambda_{\min}^{1/2}\|\theta\|} + \left(\frac{\theta_{\max}}{\theta_{\min}}\right) \frac{K\log(n)}{\lambda_{\min}^{1/2}\|\theta\|^2} \\
&\lesssim \left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2} \frac{K\sqrt{\log(n)}}{\lambda_{\min}^{1/2}\|\theta\|} \max\left\{1, \left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2} \frac{\sqrt{\log(n)}}{\|\theta\|}\right\}.
\end{aligned}$$

We now show that Assumption 6.2 implies that 1 is the maximum above. Assumption 6.2 states that

$$C\frac{\theta_{\max}}{\theta_{\min}} \frac{K^8\theta_{\max}\|\theta\|_1\log(n)}{\|\theta\|^4(\lambda_{\min})^2} \leq \bar{\lambda}.$$

Since $\bar{\lambda} \leq 1$ by assumption and $\|\theta\|^2 \leq \theta_{\max}\|\theta\|_1$, it is straightforward to verify that Assumption 6.2 implies that

$$\left(\frac{\theta_{\max}}{\theta_{\min}}\right)\frac{\log(n)}{\|\theta^{(l)}\|^2} \lesssim 1.$$

Taking square roots reveals that

$$\left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2}\frac{\sqrt{\log(n)}}{\|\theta^{(l)}\|} \lesssim 1,$$

which shows that one is the dominant term in the maximum, as long as C is larger than some universal constant. Taking a union bound over all the rows completes the proof. □

F.2.1 Preliminary Lemmas

Throughout this section and its proof we suppress the dependence on l in all terms. We also let λ denote the absolute value of the smallest nonzero eigenvalue of $\mathbf{P}$. In what follows, we will assume that $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$, which by Lemma 7 holds under Assumption 6.2. We will verify this explicitly at the beginning of the proof of Theorem 19.

The following result shows a form of spectral norm concentration.

Lemma 59 (Spectral Norm Concentration for One Graph)**.** *When* $\theta_{\max}\|\theta\|_1 \geq \log(n)$*, it holds that*

$$\|\mathbf{A} - \mathbf{P}\| \lesssim \sqrt{\theta_{\max}\|\theta\|_1};$$
$$\|\mathbf{U}^\top(\mathbf{A} - \mathbf{P})\mathbf{U}\| \lesssim \sqrt{K} + \sqrt{\log(n)},$$

with probability at least $1 - O(n^{-20})$.

Proof. See Lemma C.1 of Jin et al. (2019), or directly apply Remark 3.13 from Bandeira and Handel (2016). The other part follows from a straightforward ε-net argument. □

The following lemma demonstrates good concentration for several residual terms, showing that several terms "approximately commute."

Lemma 60 (Approximate Commutation). *When* $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$ *and* $\min_i \theta_i \|\theta\|_1 \gtrsim \log(n)$*, the following bounds hold with probability at least* $1 - O(n^{-20})$:

$$\|\mathbf{W}_* - \mathbf{U}^\top \widehat{\mathbf{U}}\| \lesssim \frac{\theta_{\max}\|\theta\|_1}{\lambda^2} \tag{F.4}$$

$$\|\widehat{\mathbf{U}}^\top \mathbf{U}|\Lambda|^{1/2} - |\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top \mathbf{U}\| \lesssim \frac{K^2}{\lambda^{1/2}}\left(\sqrt{K\log(n)} + \frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \tag{F.5}$$

$$\|\widehat{\mathbf{U}}^\top \mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q} - |\widehat{\Lambda}|^{-1/2}\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top \mathbf{U}\| \lesssim \frac{K^2}{\lambda^{3/2}}\left(\sqrt{K\log(n)} + \frac{\theta_{\max}\|\theta\|_1}{\lambda}\right). \tag{F.6}$$

Proof of Lemma 60. For (F.4), the argument follows since $\mathbf{W}_*$ is the product of the orthogonal matrices in the singular value decomposition of $\mathbf{U}^\top \widehat{\mathbf{U}}$ and hence

$$\begin{aligned}
\|\mathbf{W}_* - \mathbf{U}^\top \widehat{\mathbf{U}}\| &= \|\mathbf{I} - \cos\boldsymbol{\Theta}\| \\
&\leq \|\sin\boldsymbol{\Theta}(\mathbf{U}, \widehat{\mathbf{U}})\|^2 \\
&\lesssim \frac{\|\mathbf{A} - \mathbf{P}\|^2}{\lambda^2} \\
&\lesssim \frac{\theta_{\max}\|\theta\|_1}{\lambda^2}
\end{aligned}$$

which holds with probability at least $1 - O(n^{-20})$ by Lemma 59.

For all the following terms, we first show that $|\widehat{\Lambda}|(\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top \mathbf{U} - \widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q})$ is sufficiently small by modifying a similar argument to Rubin-Delanchy et al. (2022). Observe that

$$\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top \mathbf{U} - \widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q} = \begin{pmatrix} 0 & 2\widehat{\mathbf{U}}_+^\top \mathbf{U}_- \\ -2\widehat{\mathbf{U}}_-^\top \mathbf{U}_+ & 0 \end{pmatrix},$$

where $\mathbf{U}_+$ denotes the eigenvectors of $\mathbf{U}$ corresponding to the positive eigenvectors (and similarly for $\mathbf{U}_-$, $\widehat{\mathbf{U}}_+$, and $\widehat{\mathbf{U}}_-$ respectively). Let $\mathbf{u}_j^+$ and $\widehat{\mathbf{u}}_j^-$ denote the j'th columns of $\widehat{\mathbf{U}}_+$ and $\widehat{\mathbf{U}}_-$ respectively. Then the j, i entry of $\widehat{\mathbf{U}}_-^\top \mathbf{U}_+$ is simply $(\mathbf{u}_i^+)^\top \widehat{\mathbf{u}}_j^-$, and hence by the eigenvector-eigenvalue equation,

$$\begin{aligned}
(\mathbf{u}_i^+)^\top \widehat{\mathbf{u}}_j^- &= \frac{(\mathbf{u}_i^+)^\top (\mathbf{A} - \mathbf{P})\widehat{\mathbf{u}}_{j,-}}{\widehat{\lambda}_{j,-} - \lambda_{i,+}} \\
&= \frac{(\mathbf{u}_i^+)^\top (\mathbf{A} - \mathbf{P})\mathbf{U}_-\mathbf{U}_-^\top \widehat{\mathbf{u}}_{j,-}}{\widehat{\lambda}_{j,-} - \lambda_{i,+}} + \frac{(\mathbf{u}_i^+)^\top (\mathbf{A} - \mathbf{P})\big(\mathbf{I} - \mathbf{U}_-\mathbf{U}_-^\top\big)\widehat{\mathbf{u}}_{j,-}}{\widehat{\lambda}_{j,-} - \lambda_{i,+}},
\end{aligned}$$

where $\lambda_{i,+}$ denotes the i'th largest in magnitude eigenvalue of $\mathbf{P}$ (and similarly for $\widehat{\lambda}_{j,-}$ for the negative eigenvalues of $\mathbf{A}$). It is straightforward to check that the j,i entry of the matrix $|\widehat{\Lambda}_+|(\mathbf{I}_{p,q}\widehat{\mathbf{U}}_+^\top\mathbf{U}_- - \widehat{\mathbf{U}}_+^\top\mathbf{U}_-\mathbf{I}_{p,q})$ is given by

$$\frac{|\widehat{\lambda}_{j+}|}{\widehat{\lambda}_{j,+}-\lambda_{i,-}}(\mathbf{u}_i^+)^\top(\mathbf{A}-\mathbf{P})\widehat{\mathbf{u}}_{j,-} = \frac{1}{1-\frac{\lambda_{i,-}}{\widehat{\lambda}_{j,+}}}(\mathbf{u}_i^+)^\top(\mathbf{A}-\mathbf{P})\mathbf{U}_+\mathbf{U}_+^\top\widehat{\mathbf{u}}_{j,-} + \frac{1}{1-\frac{\lambda_{i,-}}{\widehat{\lambda}_{j,+}}}(\mathbf{u}_i^+)^\top(\mathbf{A}-\mathbf{P})(\mathbf{I}-\mathbf{U}_+\mathbf{U}_+^\top)\widehat{\mathbf{u}}_{j,-}$$

Since $\lambda_{i,-}$ is negative and $\widehat{\lambda}_{j,+}$ is positive with high probability, $1-\frac{\lambda_{i,-}}{\widehat{\lambda}_{j,+}}$ is strictly larger than one. A similar argument holds for the entries with the "+" changed to a "−".

Without loss of generality consider the term corresponding to the negative eigenvalues. We can write the matrix as follows. Denote $\mathbf{M}$ as the matrix whose i,j entry is $\frac{|\widehat{\lambda}_{j,-}|}{\widehat{\lambda}_{j,-}-\lambda_{i,+}}$. Then we have the equality

$$\mathbf{U}_+^\top\widehat{\mathbf{U}}_-|\widehat{\Lambda}_-| = \mathbf{M}\circ\Big(\mathbf{U}_+^\top(\mathbf{A}-\mathbf{P})\mathbf{U}_-\mathbf{U}_-^\top\widehat{\mathbf{U}}_- + \mathbf{U}_+^\top(\mathbf{A}-\mathbf{P})\big(\mathbf{I}-\mathbf{U}_-\mathbf{U}_-^\top\big)\widehat{\mathbf{U}}_-\Big).$$

Therefore,

$$\begin{aligned}\|\mathbf{U}_+^\top\widehat{\mathbf{U}}_-|\widehat{\Lambda}_-|\| &\le \|\mathbf{M}\|\Big(\|\mathbf{U}_+^\top(\mathbf{A}-\mathbf{P})\mathbf{U}_-\| + \|\mathbf{U}_+^\top(\mathbf{A}-\mathbf{P})\big(\mathbf{I}-\mathbf{U}_-\mathbf{U}_-^\top\big)\widehat{\mathbf{U}}_-\|\Big)\\ &\le \|\mathbf{M}\|\Big(\|\mathbf{U}^\top(\mathbf{A}-\mathbf{P})\mathbf{U}\| + \|\mathbf{A}-\mathbf{P}\|\|\big(\mathbf{I}-\mathbf{U}_-\mathbf{U}_-^\top\big)\widehat{\mathbf{U}}_-\|\Big),\end{aligned} \tag{F.7}$$

where we have used the fact that $\mathbf{U}_+^\top(\mathbf{A}-\mathbf{P})\mathbf{U}_-$ is a submatrix of $\mathbf{U}^\top(\mathbf{A}-\mathbf{P})\mathbf{U}$. We now note that

$$\|\big(\mathbf{I}-\mathbf{U}_-\mathbf{U}_-^\top\big)\widehat{\mathbf{U}}_-\| = \|\sin\boldsymbol{\Theta}(\mathbf{U}_-,\widehat{\mathbf{U}}_-)\|.$$

In addition, the eigenvalues corresponding to $\widehat{\mathbf{U}}_-$ are all negative, and the eigengap condition is satisfied since the eigenvalues corresponding to $(\mathbf{I}-\mathbf{U}_-\mathbf{U}_-)\mathbf{P}$ are either all zero or positive.

Consequently, the eigengap satisfies

$$\min_{\lambda_i>0} \lambda_i - \max_{p+1\le i\le n} \widehat{\lambda}_i \gtrsim \lambda$$

by applying Weyl's inequality to the negative eigenvalues and the bottom $n-K$ eigenvalues separately. We can therefore apply the Davis-Kahan Theorem to obtain

$$\begin{aligned} \|\sin\boldsymbol{\Theta}(\mathbf{U}_-,\widehat{\mathbf{U}}_-)\| &\lesssim \frac{\|\mathbf{A}-\mathbf{P}\|}{\lambda} \\ &\lesssim \frac{\sqrt{\theta_{\max}\|\theta\|_1}}{\lambda} \end{aligned} \tag{F.8}$$

with probability at least $1-O(n^{-20})$. In addition, observe that the matrix $\mathbf{M}$ satisfies

$$\|\mathbf{M}\| \lesssim K. \tag{F.9}$$

Finally, by Lemma 59, we have that $\|\mathbf{U}^\top(\mathbf{A}-\mathbf{P})\mathbf{U}\| \lesssim \sqrt{K}+\sqrt{\log(n)}$ with high probability. Plugging in this estimate, (F.9), and (F.8) into (F.7) yields

$$\|\mathbf{U}_+^\top\widehat{\mathbf{U}}_-\| \lesssim K\left(\sqrt{K}+\sqrt{\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right).$$

Therefore, by applying a similar argument to $\mathbf{U}_-^\top\widehat{\mathbf{U}}_+$, we obtain

$$\begin{aligned} \||\widehat{\Lambda}|(\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top\mathbf{U}-\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q})\| &\lesssim K\Big(\|\mathbf{U}^\top(\mathbf{A}-\mathbf{P})\mathbf{U}\|+\|\mathbf{A}-\mathbf{P}\|\|\sin\boldsymbol{\Theta}(\mathbf{U}_-,\widehat{\mathbf{U}}_-)\| \\ &\qquad +\|\mathbf{A}-\mathbf{P}\|\|\sin\boldsymbol{\Theta}(\mathbf{U}_+,\widehat{\mathbf{U}}_+)\|\Big) \\ &\lesssim K\left(\sqrt{K}+\sqrt{\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right), \end{aligned} \tag{F.10}$$

which holds with probability at least $1-O(n^{-20})$.

We now bound (F.5). First, note that we have

$$\begin{aligned} \|\widehat{\mathbf{U}}^\top\mathbf{U}|\Lambda|^{1/2}-|\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top\mathbf{U}\| &= \|\widehat{\mathbf{U}}^\top\mathbf{U}|\Lambda|^{1/2}\mathbf{I}_{p,q}-|\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q}\| \\ &= \|\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{1/2}-|\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q}\|, \end{aligned}$$

where the first line follows since $\mathbf{I}_{p,q}$ is orthogonal and the second line follows since diagonal matrices commute. We observe that the k,l entry of the matrix above can be written as

$$\begin{aligned}\left(\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{1/2} - |\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}\right)_{kl} &= \langle \widehat{\mathbf{U}}_{\cdot k}, \mathbf{U}_{\cdot l}(\mathbf{I}_{p,q})_{ll}\rangle \left(|\lambda_l|^{1/2} - |\widehat{\lambda}_k|^{1/2}\right) \\ &= \langle \widehat{\mathbf{U}}_{\cdot k}, \mathbf{U}_{\cdot l}\rangle (\mathbf{I}_{p,q})_{ll} \frac{|\lambda_l| - |\widehat{\lambda}_k|}{|\lambda_l|^{1/2} + |\widehat{\lambda}_k|^{1/2}}.\end{aligned}$$

Define the matrix $\mathbf{H}$ via $\mathbf{H}_{kl} := \frac{1}{|\lambda_l|^{1/2} + |\widehat{\lambda}_k|^{1/2}}$. Then the matrix $\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{1/2} - |\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}$ can be written as

$$\begin{aligned}\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{1/2} - |\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q} &= \mathbf{H} \circ \left(\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}|\Lambda| - |\widehat{\Lambda}|\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}\right) \\ &= \mathbf{H} \circ \left(\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}|\Lambda| - |\widehat{\Lambda}|\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top \mathbf{U}\right) \\ &\quad + \mathbf{H} \circ \left(|\widehat{\Lambda}|\left(\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top \mathbf{U} - \widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}\right)\right) \\ &= \mathbf{H} \circ \left(\widehat{\mathbf{U}}^\top \mathbf{U}\Lambda - \widehat{\Lambda}\widehat{\mathbf{U}}^\top \mathbf{U}\right) \\ &\quad + \mathbf{H} \circ \left(|\widehat{\Lambda}|\left(\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top \mathbf{U} - \widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}\right)\right).\end{aligned}$$

where $\circ$ denotes the hadamard product. It is straightforward to observe that $\|\mathbf{H}\| \lesssim \frac{K}{\lambda^{1/2}}$. Consequently, we have that

$$\begin{aligned}&\|\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{1/2} - |\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}\| \\ &\qquad \le \|\mathbf{H}\|\|\widehat{\mathbf{U}}^\top \mathbf{U}\Lambda - \widehat{\Lambda}\widehat{\mathbf{U}}^\top \mathbf{U}\| + \|\mathbf{H}\|\||\widehat{\Lambda}|\big(\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top \mathbf{U} - \widehat{\mathbf{U}}^\top \mathbf{U}\mathbf{I}_{p,q}\big)\| \\ &\qquad \lesssim \frac{K}{\lambda^{1/2}}\|\widehat{\mathbf{U}}^\top \mathbf{U}\Lambda - \widehat{\Lambda}\widehat{\mathbf{U}}^\top \mathbf{U}\| + \frac{K^2}{\lambda^{1/2}}\left(\sqrt{K} + \sqrt{\log(n)} + \frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \\ &\qquad \lesssim \frac{K}{\lambda^{1/2}}\|\widehat{\mathbf{U}}^\top \mathbf{P}\mathbf{U} - \widehat{\mathbf{U}}^\top \mathbf{A}\mathbf{U}\| + \frac{K^2}{\lambda^{1/2}}\left(\sqrt{K} + \sqrt{\log(n)} + \frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \\ &\qquad \lesssim \frac{K}{\lambda^{1/2}}\|\widehat{\mathbf{U}}^\top (\mathbf{P} - \mathbf{A})\mathbf{U}\| + \frac{K^2}{\lambda^{1/2}}\left(\sqrt{K} + \sqrt{\log(n)} + \frac{\theta_{\max}\|\theta\|_1}{\lambda}\right). \quad \text{(F.11)}\end{aligned}$$

We note that

$$\begin{aligned}\|\widehat{\mathbf{U}}^\top(\mathbf{P}-\mathbf{A})\mathbf{U}\| &\lesssim \|\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{U}^\top(\mathbf{P}-\mathbf{A})\mathbf{U}\| + \|\widehat{\mathbf{U}}^\top(\mathbf{I}-\mathbf{U}\mathbf{U}^\top)(\mathbf{P}-\mathbf{A})\mathbf{U}\| \\ &\lesssim \sqrt{K}+\sqrt{\log(n)} + \|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\|\|\mathbf{A}-\mathbf{P}\| \\ &\lesssim \sqrt{K}+\sqrt{\log(n)} + \frac{\theta_{\max}\|\theta\|_1}{\lambda}.\end{aligned}$$

Plugging this into our bound (F.11), we obtain that

$$\begin{aligned}\|\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{1/2} - |\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q}\| &\lesssim \frac{K}{\lambda^{1/2}}\left(\sqrt{K}+\sqrt{\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \\ &\quad + \frac{K^2}{\lambda^{1/2}}\left(\sqrt{K}+\sqrt{\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \\ &\lesssim \frac{K^2}{\lambda^{1/2}}\left(\sqrt{K\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right).\end{aligned}$$

This proves (F.5).

We now consider the term (F.6). Since diagonal matrices commute,

$$\begin{aligned}&\|\widehat{\mathbf{U}}^\top\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q} - |\widehat{\Lambda}|^{-1/2}\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top\mathbf{U}\| \\ &\quad= \|\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{-1/2} - |\widehat{\Lambda}|^{-1/2}\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top\mathbf{U}\| \\ &\quad= \||\widehat{\Lambda}|^{-1/2}\mathbf{I}_{p,q}\left(|\widehat{\Lambda}|^{1/2}\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{1/2}\right)\mathbf{I}_{p,q}|\Lambda|^{-1/2}\| \\ &\quad\lesssim \frac{1}{\lambda}\||\widehat{\Lambda}|^{1/2}\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{1/2}\| \\ &\quad\lesssim \frac{1}{\lambda}\left(\||\widehat{\Lambda}|^{1/2}(\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q})\| + \||\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q} - \widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{1/2}\|\right) \\ &\quad\lesssim \frac{1}{\lambda^{3/2}}\||\widehat{\Lambda}|(\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q})\| + \frac{1}{\lambda}\||\widehat{\Lambda}|^{1/2}\widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q} - \widehat{\mathbf{U}}^\top\mathbf{U}\mathbf{I}_{p,q}|\Lambda|^{1/2}\| \\ &\quad\lesssim \frac{K}{\lambda^{3/2}}\left(\sqrt{K}+\sqrt{\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \\ &\qquad+ \frac{K^2}{\lambda^{3/2}}\left(\sqrt{K\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \\ &\quad\lesssim \frac{K^2}{\lambda^{3/2}}\left(\sqrt{K\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right).\end{aligned}$$

where we have implicitly used the bound (F.5) and (F.11). This bound holds cumulatively with probability at least $1 - O(n^{-20})$, which completes the proof.

□

The following lemma characterizes the row-wise concentration of terms that involve $\widehat{\mathbf{U}}$. However, this proof requires the use of leave-one-out sequences, so we defer its proof to Appendix F.2.3 after the proof of Theorem 19.

Lemma 61 (Row-Wise Concentration I). *When* $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$ *and* $\min_i \theta_i \|\theta\|_1 \gtrsim \log(n)$*, it holds that*

$$\|e_i^\top (\mathbf{A} - \mathbb{E}\mathbf{A})\widehat{\mathbf{U}}\| \lesssim \sqrt{\theta_i \|\theta\|_1 \log(n)} \|\widehat{\mathbf{U}}\|_{2,\infty}$$

The following result demonstrates that $\widehat{\mathbf{U}}$ is sufficiently close to $\mathbf{U}\mathbf{U}^\top \widehat{\mathbf{U}}$ in $\|\cdot\|_{2,\infty}$.

Lemma 62 (Closeness of $\widehat{\mathbf{U}}$ to $\mathbf{U}$). *When* $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$ *and* $\min_i \theta_i \|\theta\|_1 \gtrsim \log(n)$*, the following bounds holds with probability at least* $1 - O(n^{-19})$:

$$\begin{aligned}
\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top \widehat{\mathbf{U}}\|_{2,\infty} &\lesssim \frac{\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda} \|\mathbf{U}\|_{2,\infty} \\
\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} &\lesssim \frac{\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda} \|\mathbf{U}\|_{2,\infty} \\
\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U} - \mathbf{U}\|_{2,\infty} &\lesssim \frac{\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda} \|\mathbf{U}\|_{2,\infty}.
\end{aligned}$$

The bound above matches the bound in Jin et al. (2019), Lemma 2.1.

Proof of Lemma 62. Observe that since $\mathbf{U}$ are the eigenvectors of $\mathbf{P}$ and $\mathbf{P}$ is rank K,

$$\begin{aligned}
e_i^\top \big(\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top \widehat{\mathbf{U}}\big) &= e_i^\top (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\widehat{\mathbf{U}} \\
&= e_i^\top (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathbf{A}\widehat{\mathbf{U}}\widehat{\Lambda}^{-1} \\
&= e_i^\top (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)(\mathbf{A} - \mathbf{P})\widehat{\mathbf{U}}\widehat{\Lambda}^{-1} \\
&= e_i^\top (\mathbf{A} - \mathbf{P})\widehat{\mathbf{U}}\widehat{\Lambda}^{-1} - e_i^\top \mathbf{U}\mathbf{U}^\top (\mathbf{A} - \mathbf{P})\widehat{\mathbf{U}}\widehat{\Lambda}^{-1}.
\end{aligned}$$

Taking norms reveals that

$$\|e_i^\top \big(\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top \widehat{\mathbf{U}}\big)\| \le \|e_i^\top \big(\mathbf{A} - \mathbf{P}\big)\widehat{\mathbf{U}}\|\|\widehat{\Lambda}^{-1}\| + \|e_i^\top \mathbf{U}\|\|\mathbf{A} - \mathbf{P}\|\|\widehat{\Lambda}^{-1}\|.$$

By Lemma 59, we have that $\|\mathbf{A} - \mathbf{P}\| \lesssim \sqrt{\theta_{\max}\|\theta\|_1}$. In addition, Weyl's inequality implies that $\|\widehat{\Lambda}^{-1}\| \lesssim \lambda^{-1}$. Therefore, combining these bounds with Lemma 61, we see that with probability at least $1 - O(n^{-20})$ that

$$\begin{aligned}\|e_i^\top (\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top \widehat{\mathbf{U}})\| &\lesssim \frac{\sqrt{\theta_i \|\theta\|_1 \log(n)}}{\lambda} \|\widehat{\mathbf{U}}\|_{2,\infty} + \|e_i^\top \mathbf{U}\| \frac{\sqrt{\theta_{\max}\|\theta\|_1}}{\lambda} \\ &\lesssim \frac{\sqrt{\theta_{\max} \|\theta\|_1 \log(n)}}{\lambda} \|\widehat{\mathbf{U}}\|_{2,\infty} + \|\mathbf{U}\|_{2,\infty} \frac{\sqrt{\theta_{\max}\|\theta\|_1}}{\lambda}.\end{aligned}$$

This bound is independent of row i, so taking a union bound reveals that with probability at least $1 - O(n^{-19})$ that

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top \widehat{\mathbf{U}}\|_{2,\infty} \lesssim \frac{\sqrt{\theta_{\max} \|\theta\|_1 \log(n)}}{\lambda} \|\widehat{\mathbf{U}}\|_{2,\infty} + \|\mathbf{U}\|_{2,\infty} \frac{\sqrt{\theta_{\max}\|\theta\|_1}}{\lambda}.$$

By Lemma 60, it holds that

$$\|\mathbf{W}_* - \mathbf{U}^\top \widehat{\mathbf{U}}\| \lesssim \frac{\theta_{\max}\|\theta\|_1}{\lambda^2}.$$

Therefore,

$$\begin{aligned}\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} &\le \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top \widehat{\mathbf{U}}\|_{2,\infty} + \|\mathbf{U}\|_{2,\infty} \|\mathbf{W}_* - \mathbf{U}^\top \widehat{\mathbf{U}}\| \\ &\lesssim \frac{\sqrt{\theta_{\max} \|\theta\|_1 \log(n)}}{\lambda} \|\widehat{\mathbf{U}}\|_{2,\infty} + \|\mathbf{U}\|_{2,\infty} \frac{\sqrt{\theta_{\max}\|\theta\|_1}}{\lambda} + \frac{\theta_{\max}\|\theta\|_1}{\lambda^2} \|\mathbf{U}\|_{2,\infty} \\ &\lesssim \frac{\sqrt{\theta_{\max} \|\theta\|_1 \log(n)}}{\lambda} \|\widehat{\mathbf{U}}\|_{2,\infty} + \|\mathbf{U}\|_{2,\infty} \frac{\sqrt{\theta_{\max}\|\theta\|_1}}{\lambda}.\end{aligned}$$

As a byproduct, this also reveals that

$$\begin{aligned}\|\widehat{\mathbf{U}}\|_{2,\infty} &\le \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} + \|\mathbf{U}\|_{2,\infty} \\ &\le \frac{1}{2}\|\widehat{\mathbf{U}}\|_{2,\infty} + \frac{3}{2}\|\mathbf{U}\|_{2,\infty},\end{aligned}$$

as long as $\lambda \ge C\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$ for some sufficiently large constant C (which we verify at the beginning of the proof of Theorem 19, and which holds under Assumption 6.2). By

rearranging, it holds that $\|\widehat{\mathbf{U}}\|_{2,\infty} \lesssim \|\mathbf{U}\|_{2,\infty}$. Plugging this in yields

$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} \lesssim \frac{\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda}\|\mathbf{U}\|_{2,\infty};$$
$$\|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\|_{2,\infty} \lesssim \frac{\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda}\|\mathbf{U}\|_{2,\infty}.$$

The final inequality holds since

$$\begin{aligned}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{U}\|_{2,\infty} &\le \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \widehat{\mathbf{U}}\mathbf{W}_*^\top\|_{2,\infty} + \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} \\ &\le \|\widehat{\mathbf{U}}\|_{2,\infty}\|\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{W}_*^\top\| + \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty} \\ &\lesssim \|\mathbf{U}\|_{2,\infty}\|\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}\| + \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_*\|_{2,\infty}.\end{aligned}$$

The proof is completed by plugging in the previous bounds. □

The following result establishes finer control over the rows of the estimated eigenvectors. We relegate the proof of his result to Appendix F.2.3, since it requires the use of leave-one-out sequences.

Lemma 63 (Row-wise Concentration II)**.** *When* $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$ *and* $\min_i \theta_i\|\theta\|_1 \gtrsim \log(n)$*, with probability at least* $1 - O(n^{-19})$*, it holds that*

$$\|e_i^\top(\mathbf{A} - \mathbb{E}\mathbf{A})(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{U})\| \lesssim \sqrt{\theta_i\|\theta\|_1 \log(n)}\|\frac{\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda}\|\mathbf{U}\|_{2,\infty}.$$

F.2.2 Proof of Theorem 19

Proof of Theorem 19. Throughout the proof we suppress the dependence of these terms on the index l. Our proof proceeds in several steps: first, we express $\widehat{\mathbf{X}}\mathbf{W}_* - \mathbf{X}$ as a linear term plus a residual term, where the residual term obeys a strong row-wise concentration bound. Next, we demonstrate that the rows of $\widehat{\mathbf{Y}}$ (i.e. the normalized rows of $\widehat{\mathbf{X}}$) concentrate about the corresponding rows of $\mathbf{Y}$. Before embarking on the proof, we make note of several

preliminary facts. By Lemma 7, we have that

$$\begin{aligned}\lambda &\gtrsim \frac{\|\theta\|^2 \lambda_{\min}}{K};\\ \|e_i^\top \mathbf{U}\| &\lesssim \frac{\sqrt{K}\theta_i}{\|\theta\|};\\ \theta_i \lesssim \|e_i^\top \mathbf{X}\| &\leq \theta_i.\end{aligned}$$

We will use these bounds repeatedly without reference when simplifying our results.

In addition, many of the previous lemmas require that $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$. We verify that this condition holds under Assumption 6.2 now. Assumption 6.2 requires that

$$C\left(\frac{\theta_{\max}}{\theta_{\min}}\right)\frac{K^8 \theta_{\max}\|\theta\|_1 \log(n)}{\|\theta\|^4 \lambda_{\min}^2} \leq \bar{\lambda}. \tag{F.12}$$

By Lemma 7 it holds that

$$\lambda \gtrsim \frac{\|\theta\|^2}{K}\lambda_{\min}.$$

Consequently, it suffices to argue that (F.12) implies the condition

$$\frac{\|\theta\|^2}{K}\lambda_{\min} \gtrsim \sqrt{\theta_{\max}\|\theta\|_1 \log(n)},$$

or equivalently,

$$\frac{K\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\|\theta\|^2 \lambda_{\min}} \lesssim 1.$$

Squaring both sides yields the condition

$$\frac{K^2 \theta_{\max}\|\theta\|_1 \log(n)}{\|\theta\|^4 \lambda_{\min}^2} \lesssim 1.$$

This is weaker than (F.12) as $\lambda_{\min}, \bar{\lambda} \in (0,1)$ by assumption and $K \geq 1$, as long as C is larger than some universal constant.

Step 1: First-Order Approximation of $\widehat{\mathbf{X}}$:

At the outset we recall that $\mathbf{W}_*$ is the Frobenius-optimal matrix aligning $\widehat{\mathbf{U}}$ and $\mathbf{U}$. Moreover, by the concentration inequality in Lemma 59 and the assumption on the eigenvalue λ above, we have that $\|\widehat{\Lambda}^{-1}\| \lesssim \lambda^{-1}$ with probability at least $1 - O(n^{-20})$. We now expand via:

$$\begin{aligned}
\widehat{\mathbf{X}}\mathbf{W}_*^\top - \mathbf{X} &= (\mathbf{A} - \mathbb{E}\mathbf{A})\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q} + \mathbf{R}_1\mathbf{W}_*^\top + \mathbf{R}_2\mathbf{W}_*^\top + \mathbf{R}_3\mathbf{W}_*^\top + \mathbf{R}_4 + \mathbf{R}_5 + \mathbf{R}_6; \\
\mathbf{R}_1 &:= -\mathbf{U}\mathbf{U}^\top(\mathbf{A} - \mathbb{E}\mathbf{A})\widehat{\mathbf{U}}|\widehat{\Lambda}|^{-1/2}\mathbf{I}_{p,q}; \\
\mathbf{R}_2 &:= \mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}}|\widehat{\Lambda}|^{1/2} - |\Lambda|^{1/2}\mathbf{U}^\top\widehat{\mathbf{U}}); \\
\mathbf{R}_3 &:= \mathbf{U}|\Lambda|^{1/2}(\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*); \\
\mathbf{R}_4 &:= (\mathbf{A} - \mathbb{E}\mathbf{A})(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U} - \mathbf{U})|\Lambda|^{-1/2}\mathbf{I}_{p,q}; \\
\mathbf{R}_5 &:= -(\mathbf{A} - \mathbb{E}\mathbf{A})\widehat{\mathbf{U}}(\widehat{\mathbf{U}}^\top\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q} - |\widehat{\Lambda}|^{-1/2}\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top\mathbf{U}); \\
\mathbf{R}_6 &:= (\mathbf{A} - \mathbb{E}\mathbf{A})\widehat{\mathbf{U}}|\widehat{\Lambda}|^{-1/2}\mathbf{I}_{p,q}(\mathbf{W}_*^\top - \widehat{\mathbf{U}}^\top\mathbf{U}).
\end{aligned}$$

We now bound each residual in turn. We will also use Lemma 60, Lemma 61, Lemma 62, Lemma 63 repeatedly without reference; the cumulative probability will be at least $1 - O(n^{-18})$.

The term $\mathbf{R}_1$:

First, we note that

$$\begin{aligned}
\|e_i^\top\mathbf{R}_1\| &\le \|e_i^\top\mathbf{U}\|\|\mathbf{U}^\top(\mathbf{A} - \mathbf{P})\widehat{\mathbf{U}}|\widehat{\Lambda}|^{-1/2}\| \\
&\lesssim \frac{\|e_i^\top\mathbf{U}\|}{\lambda^{1/2}}\left(\|\mathbf{U}^\top(\mathbf{A} - \mathbf{P})\mathbf{U}\| + \|\mathbf{A} - \mathbf{P}\|\|\mathbf{U}_\perp^\top\widehat{\mathbf{U}}\|\right) \\
&\lesssim \frac{\|e_i^\top\mathbf{U}\|}{\lambda^{1/2}}\left(\|\mathbf{U}^\top(\mathbf{A} - \mathbf{P})\mathbf{U}\| + \frac{\|\mathbf{A} - \mathbf{P}\|^2}{\lambda}\right).
\end{aligned}$$

By Lemma 59, we have that $\|\mathbf{U}^\top(\mathbf{A} - \mathbf{P})\mathbf{U}\| \lesssim \sqrt{K} + \sqrt{\log(n)}$ with probability at least $1 - O(n^{-20})$. Consequently,

$$\|e_i^\top\mathbf{R}_1\| \lesssim \frac{\|e_i^\top\mathbf{U}\|}{\lambda^{1/2}}\left(\sqrt{K} + \sqrt{\log(n)} + \frac{\theta_{\max}\|\theta\|_1}{\lambda}\right).$$

By Lemma 7, we have that $\|e_i^\top \mathbf{U}\| \lesssim \frac{\sqrt{K}\theta_i}{\|\theta\|}$ and that $\lambda \gtrsim \frac{\|\theta\|^2}{K}\lambda_{\min}$. Putting it together, we arrive at the bound

$$\begin{aligned}\|e_i^\top \mathbf{R}_1\| &\lesssim \frac{K\theta_i}{\|\theta\|^2\lambda_{\min}^{1/2}}\left(\sqrt{K} + \sqrt{\log(n)} + \frac{K\theta_{\max}\|\theta\|_1}{\|\theta\|^2\lambda_{\min}}\right) \\ &\lesssim \frac{K\theta_i}{\|\theta\|^2\lambda_{\min}^{1/2}}\left(\sqrt{K\log(n)} + \frac{K\theta_{\max}\|\theta\|_1}{\|\theta\|^2\lambda_{\min}}\right). \end{aligned} \tag{F.13}$$

The term $\mathbf{R}_2$:

We have

$$\begin{aligned}\|e_i^\top \mathbf{R}_2\| &\lesssim \|e_i^\top \mathbf{U}\|\|\mathbf{U}^\top\widehat{\mathbf{U}}|\widehat{\Lambda}|^{1/2} - |\Lambda|^{1/2}\mathbf{U}^\top\widehat{\mathbf{U}}\| \\ &\lesssim \frac{\sqrt{K}\theta_i}{\|\theta\|}\|\mathbf{U}^\top\widehat{\mathbf{U}}|\widehat{\Lambda}|^{1/2} - |\Lambda|^{1/2}\mathbf{U}^\top\widehat{\mathbf{U}}\| \\ &\lesssim \frac{\sqrt{K}\theta_i}{\|\theta\|}\frac{K^2}{\lambda^{1/2}}\left(\sqrt{K\log(n)} + \frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \\ &\lesssim \frac{\sqrt{K}\theta_i}{\|\theta\|}\frac{K^{5/2}}{\lambda_{\min}^{1/2}\|\theta\|}\left(\sqrt{K\log(n)} + \frac{K\theta_{\max}\|\theta\|_1}{\|\theta\|^2\lambda_{\min}}\right) \\ &\asymp \frac{K^3\theta_i}{\|\theta\|^2\lambda_{\min}^{1/2}}\left(\sqrt{K\log(n)} + \frac{K\theta_{\max}\|\theta\|_1}{\|\theta\|^2\lambda_{\min}}\right). \end{aligned} \tag{F.14}$$

The term $\mathbf{R}_3$:

Following similarly as the previous step, we have that

$$\begin{aligned}\|e_i^\top \mathbf{R}_3\| &\lesssim \|e_i^\top \mathbf{X}\|\|\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*\| \\ &\lesssim \theta_i\frac{\theta_{\max}\|\theta\|_1}{\lambda^2} \\ &\lesssim \theta_i\frac{K^2\theta_{\max}\|\theta\|_1}{\|\theta\|^4\lambda_{\min}^2}. \end{aligned} \tag{F.15}$$

The term $\mathbf{R}_4$:

By Lemma 63, we have

$$\begin{aligned}\|e_i^\top \mathbf{R}_4\| &\lesssim \|e_i^\top(\mathbf{A}-\mathbf{P})(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}-\mathbf{U})\|\|\Lambda|^{-1/2}\| \\ &\lesssim \frac{\sqrt{\theta_i\|\theta\|_1\log(n)}}{\lambda^{1/2}}\frac{\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\lambda}\|\mathbf{U}\|_{2,\infty} \\ &\lesssim \frac{\sqrt{K\theta_i\|\theta\|_1\log(n)}}{\lambda_{\min}^{1/2}\|\theta\|}\frac{K\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\|\theta\|^2\lambda_{\min}}\frac{\sqrt{K}\theta_{\max}}{\|\theta\|} \\ &\asymp \frac{\theta_i^{1/2}K^2\theta_{\max}^{3/2}\|\theta\|_1\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4}. \end{aligned} \tag{F.16}$$

The term $\mathbf{R}_5$:

By Lemma 61 and Lemma 60, we have that

$$\begin{aligned}\|e_i^\top \mathbf{R}_5\| &\lesssim \|e_i^\top(\mathbf{A}-\mathbf{P})\widehat{\mathbf{U}}\|\|\widehat{\mathbf{U}}^\top\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}-|\widehat{\Lambda}|^{-1/2}\mathbf{I}_{p,q}\widehat{\mathbf{U}}^\top\mathbf{U}\| \\ &\lesssim \|e_i^\top(\mathbf{A}-\mathbf{P})\widehat{\mathbf{U}}\|\frac{K^2}{\lambda^{3/2}}\left(\sqrt{K\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \\ &\lesssim \sqrt{\theta_i\|\theta\|_1\log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty}\frac{K^2}{\lambda^{3/2}}\left(\sqrt{K\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right).\end{aligned}$$

By Lemma 62, we have that $\|\widehat{\mathbf{U}}\|_{2,\infty}\lesssim\|\mathbf{U}\|_{2,\infty}$ as long as $\lambda\gtrsim\sqrt{\theta_{\max}\|\theta\|_1\log(n)}$, which is true by Assumption 6.2. Therefore,

$$\begin{aligned}\|e_i^\top \mathbf{R}_5\| &\lesssim \sqrt{\theta_i\|\theta\|_1\log(n)}\|\mathbf{U}\|_{2,\infty}\frac{K^2}{\lambda^{3/2}}\left(\sqrt{K\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \\ &\lesssim \sqrt{\theta_i\|\theta\|_1\log(n)}\frac{\sqrt{K}\theta_{\max}}{\|\theta\|}\frac{K^2}{\lambda^{3/2}}\left(\sqrt{K\log(n)}+\frac{\theta_{\max}\|\theta\|_1}{\lambda}\right) \\ &\asymp \sqrt{\theta_i\|\theta\|_1\log(n)}\frac{\sqrt{K}\theta_{\max}}{\|\theta\|}\frac{K^{7/2}}{\lambda_{\min}^{3/2}\|\theta\|^3}\left(\sqrt{K\log(n)}+\frac{K\theta_{\max}\|\theta\|_1}{\lambda_{\min}\|\theta\|^2}\right) \\ &\asymp \frac{\theta_i^{1/2}\sqrt{\|\theta\|_1\log(n)}\theta_{\max}K^4}{\lambda_{\min}^{3/2}\|\theta\|^4}\left(\sqrt{K\log(n)}+\frac{K\theta_{\max}\|\theta\|_1}{\lambda_{\min}\|\theta\|^2}\right).\end{aligned} \tag{F.17}$$

The term $\mathbf{R}_6$:

Similarly to the previous term, we obtain

$$\begin{aligned}\|e_i^\top \mathbf{R}_6\| &\lesssim \frac{\|e_i^\top(\mathbf{A}-\mathbf{P})\widehat{\mathbf{U}}\|}{\lambda^{1/2}}\|\widehat{\mathbf{U}}^\top\mathbf{U}-\mathbf{W}_*\| \\ &\lesssim \frac{\sqrt{\theta_i\|\theta\|_1\log(n)}\|\mathbf{U}\|_{2,\infty}}{\lambda^{1/2}}\frac{\theta_{\max}\|\theta\|_1}{\lambda^2} \\ &\lesssim \frac{\sqrt{\theta_i\|\theta\|_1\log(n)}\sqrt{K}\theta_{\max}}{\lambda^{1/2}\|\theta\|}\frac{\theta_{\max}\|\theta\|_1}{\lambda^2} \\ &\asymp \frac{\theta_i^{1/2}\|\theta\|_1^{3/2}\theta_{\max}^2\sqrt{\log(n)}K^3}{\lambda_{\min}^{5/2}\|\theta\|^6} \end{aligned} \tag{F.18}$$

Putting it together:

By (F.13), (F.14), (F.15), (F.16), (F.17), and (F.18), we obtain that

$$\begin{aligned}
\|e_i^\top \mathbf{R}_1\| &\lesssim \theta_i\frac{K^{3/2}}{\|\theta\|^2\lambda_{\min}^{1/2}}\left(\sqrt{K\log(n)}+\frac{K\theta_{\max}\|\theta\|_1}{\|\theta\|^2\lambda_{\min}}\right); \\
\|e_i^\top \mathbf{R}_2\| &\lesssim \frac{K^3\theta_i}{\|\theta\|^2\lambda_{\min}^{1/2}}\left(\sqrt{K\log(n)}+\frac{K\theta_{\max}\|\theta\|_1}{\|\theta\|^2\lambda_{\min}}\right); \\
\|e_i^\top \mathbf{R}_3\| &\lesssim \theta_i\frac{K^2\theta_{\max}\|\theta\|_1}{\|\theta\|^4\lambda_{\min}^2}; \\
\|e_i^\top \mathbf{R}_4\| &\lesssim \theta_i^{1/2}\frac{K^2\theta_{\max}^{3/2}\|\theta\|_1\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4}; \\
\|e_i^\top \mathbf{R}_5\| &\lesssim \frac{\theta_i^{1/2}\sqrt{\|\theta\|_1\log(n)}\theta_{\max}K^4}{\lambda_{\min}^{3/2}\|\theta\|^4}\left(\sqrt{K\log(n)}+\frac{K\theta_{\max}\|\theta\|_1}{\lambda_{\min}\|\theta\|^2}\right); \\
\|e_i^\top \mathbf{R}_6\| &\lesssim \theta_i^{1/2}\frac{K^3\|\theta\|_1^{3/2}\theta_{\max}^2\sqrt{\log(n)}}{\lambda_{\min}^{5/2}\|\theta\|^6}.
\end{aligned}$$

We now group these terms for simplicity. First, observe that the bound for $\|e_i^\top \mathbf{R}_1\|$ is no more than the bound for $\|e_i^\top \mathbf{R}_2\|$ since $\lambda_{\min} < 1$ and $K \geq 2$. Therefore,

$$\|e_i^\top \mathbf{R}_1\| + \|e_i^\top \mathbf{R}_2\| + \|e_i^\top \mathbf{R}_3\| \lesssim \theta_i\left(\frac{K^{7/2}\sqrt{\log(n)}}{\|\theta\|^2\lambda_{\min}^{1/2}} + \frac{K^4\theta_{\max}\|\theta\|_1}{\|\theta\|^4\lambda_{\min}^{3/2}} + \frac{K^2\theta_{\max}\|\theta\|_1}{\|\theta\|^4\lambda_{\min}^2}\right)$$

We now simplify the remaining terms; i.e. the terms $\mathbf{R}_4$ through $\mathbf{R}_6$. We observe that

$$\begin{aligned}
&\|e_i^\top \mathbf{R}_4\| + \|e_i^\top \mathbf{R}_5\| + \|e_i^\top \mathbf{R}_6\| \\
&\quad\lesssim \theta_i^{1/2}\frac{K^2\theta_{\max}^{3/2}\|\theta\|_1\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4} + \theta_i^{1/2}\frac{K^4\sqrt{\|\theta\|_1\log(n)}\theta_{\max}}{\lambda_{\min}^{3/2}\|\theta\|^4}\left(\sqrt{K\log(n)} + \frac{K\theta_{\max}\|\theta\|_1}{\lambda_{\min}\|\theta\|^2}\right) \\
&\qquad + \theta_i^{1/2}\frac{K^3\|\theta\|_1^{3/2}\theta_{\max}^2\sqrt{\log(n)}}{\lambda_{\min}^{5/2}\|\theta\|^6} \\
&\quad\lesssim \theta_i^{1/2}\frac{K^2\theta_{\max}^{3/2}\|\theta\|_1\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4} + \theta_i^{1/2}\frac{K^{9/2}\sqrt{\|\theta\|_1}\theta_{\max}\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4} \\
&\qquad + \theta_i^{1/2}\frac{K^5\|\theta\|_1^{3/2}\theta_{\max}^2\sqrt{\log(n)}}{\lambda_{\min}^{5/2}\|\theta\|^6} + \theta_i^{1/2}\frac{K^3\|\theta\|_1^{3/2}\theta_{\max}^2\sqrt{\log(n)}}{\lambda_{\min}^{5/2}\|\theta\|^6} \\
&\quad\lesssim \theta_i^{1/2}\frac{K^2\theta_{\max}^{3/2}\|\theta\|_1\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4} + \theta_i^{1/2}\frac{K^{9/2}\sqrt{\|\theta\|_1}\theta_{\max}\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4} + \theta_i^{1/2}\frac{K^5\|\theta\|_1^{3/2}\theta_{\max}^2\sqrt{\log(n)}}{\lambda_{\min}^{5/2}\|\theta\|^6} \\
&\quad\lesssim (\theta_i\theta_{\max})^{1/2}\left(\frac{K^2\theta_{\max}\|\theta\|_1\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4} + \frac{K^{9/2}\sqrt{\|\theta\|_1\theta_{\max}}\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4} + \frac{K^5\|\theta\|_1^{3/2}\theta_{\max}^{3/2}\sqrt{\log(n)}}{\lambda_{\min}^{5/2}\|\theta\|^6}\right) \\
&\quad\lesssim \theta_i\left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2}\left(\frac{K^2\theta_{\max}\|\theta\|_1\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4} + \frac{K^{9/2}\sqrt{\|\theta\|_1\theta_{\max}}\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4} + \frac{K^5\|\theta\|_1^{3/2}\theta_{\max}^{3/2}\sqrt{\log(n)}}{\lambda_{\min}^{5/2}\|\theta\|^6}\right) \\
&\quad\lesssim \theta_i\left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2}\left(\frac{K^{9/2}\theta_{\max}\|\theta\|_1\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4}\right),
\end{aligned}$$

where we have used the fact that $\lambda_{\min}\|\theta\|^2 \gtrsim K\sqrt{\theta_{\max}\|\theta\|_1\log(n)}$ and $\theta_{\max}\|\theta\|_1 \gtrsim \log(n)$, the first of which we verified at the beginning of this proof and the second by Assumption 6.2. Putting these together, we arrive at

$$\begin{aligned}
\|e_i^\top \mathbf{R}\| &\lesssim \theta_i\left(\frac{K^{7/2}\sqrt{\log(n)}}{\|\theta\|^2\lambda_{\min}^{1/2}} + \frac{K^4\theta_{\max}\|\theta\|_1}{\|\theta\|^4\lambda_{\min}^{3/2}} + \frac{K^2\theta_{\max}\|\theta\|_1}{\|\theta\|^4\lambda_{\min}^2}\right) \\
&\qquad + \theta_i\left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2}\left(\frac{K^{9/2}\theta_{\max}\|\theta\|_1\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4}\right) \\
&\lesssim \theta_i\frac{K^{7/2}\sqrt{\log(n)}}{\|\theta\|^2\lambda_{\min}^{1/2}} + \theta_i\frac{K^2\theta_{\max}\|\theta\|_1}{\|\theta\|^4\lambda_{\min}^2} + \theta_i\left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2}\left(\frac{K^{9/2}\theta_{\max}\|\theta\|_1\log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4}\right)
\end{aligned}$$

Consequently, we see that with probability at least $1 - O(n^{-18})$, each row i of $\widehat{\mathbf{X}}$ satisfies

$$\widehat{\mathbf{X}}\mathbf{W}_*^\top - \mathbf{X} = e_i^\top(\mathbf{A}-\mathbf{P})\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q} + e_i^\top\mathbf{R},$$

where $\mathbf{R}$ satisfies

$$\begin{aligned}\|e_i^\top \mathbf{R}\| &\lesssim \theta_i \frac{K^{7/2}\sqrt{\log(n)}}{\|\theta\|^2 \lambda_{\min}^{1/2}} + \theta_i \frac{K^2 \theta_{\max}\|\theta\|_1}{\|\theta\|^4 \lambda_{\min}^2} + \theta_i \left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2} \left(\frac{K^{9/2}\theta_{\max}\|\theta\|_1 \log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4}\right) \\ &\lesssim \theta_i \frac{K^2 \theta_{\max}\|\theta\|_1}{\|\theta\|^4 \lambda_{\min}^2} + \theta_i \left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2} \left(\frac{K^{9/2}\theta_{\max}\|\theta\|_1 \log(n)}{\lambda_{\min}^{3/2}\|\theta\|^4}\right) \end{aligned} \tag{F.19}$$

In what follows, denote

$$\alpha_{\mathbf{R}} := \frac{K^2 \theta_{\max}\|\theta\|_1}{\|\theta\|^4 \lambda_{\min}^2} + \left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2} \left(\frac{K^{9/2}\theta_{\max}\|\theta\|_1 \log(n)}{\|\theta\|^4 \lambda_{\min}^{3/2}}\right), \tag{F.20}$$

so that $\|e_i^\top \mathbf{R}\| \lesssim \theta_i \alpha_{\mathbf{R}}$.

Step 2: First Order Approximation of $\widehat{\mathbf{Y}}$:

Now, we note that

$$e_i^\top (\mathbf{A} - \mathbf{P})\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q} = \sum_{j=1}^n (\mathbf{A}_{ij} - \mathbf{P}_{ij})(\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q})_j.$$

is a sum of n independent random matrices. Bernstein's inequality shows that this is less than or equal to

$$\begin{aligned}\frac{\sqrt{\theta_i\|\theta\|_1 \log(n)}}{\lambda^{1/2}}\|\mathbf{U}\|_{2,\infty} &\lesssim \frac{K\sqrt{\theta_i\|\theta\|_1\log(n)}}{\lambda_{\min}^{1/2}\|\theta\|}\frac{\theta_{\max}}{\|\theta\|} \\ &\asymp \frac{K\sqrt{\theta_i\|\theta\|_1\log(n)}\theta_{\max}}{\|\theta\|^2\lambda_{\min}^{1/2}} \\ &\asymp \theta_i\left(\frac{\theta_{\max}}{\theta_i}\right)^{1/2}\left[\frac{K\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\|\theta\|^2\lambda_{\min}^{1/2}}\right].\end{aligned}$$

Consequently, we obtain that

$$\begin{aligned}\|e_i^\top \widehat{\mathbf{X}}\mathbf{W}_*^\top - e_i^\top \mathbf{X}\| &\lesssim \theta_i\left(\frac{\theta_{\max}}{\theta_i}\right)^{1/2}\left[\frac{K\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\|\theta\|^2\lambda_{\min}^{1/2}}\right] + \theta_i\lambda_{\min}^{1/2}\alpha_{\mathbf{R}} \\ &= \theta_i\left\{\left(\frac{\theta_{\max}}{\theta_i}\right)^{1/2}\left[\frac{K\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\lambda_{\min}^{1/2}\|\theta\|^2}\right] + \alpha_{\mathbf{R}}\right\} \\ &\le \frac{1}{64}\|\mathbf{X}_i\|,\end{aligned}$$

since $\|\mathbf{X}_i\| \gtrsim \theta_i$, as long as $\alpha_{\mathbf{R}} \lesssim 1$ and that

$$\left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2} \frac{K\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda_{\min}^{1/2}\|\theta\|^2} \lesssim 1, \tag{F.21}$$

both of which are guaranteed Assumption 6.2, which we will verify now. First, a direct comparison of $\alpha_{\mathbf{R}}$ with Assumption 6.2 shows that $\alpha_{\mathbf{R}} \leq \frac{\bar{\lambda}}{C\sqrt{K}}$, which is strictly less than one. In addition, by squaring (F.21), we see that we require that

$$\frac{\theta_{\max}}{\theta_{\min}} \frac{K^2\theta_{\max}\|\theta\|_1 \log(n)}{\lambda_{\min}\|\theta\|^4} \lesssim 1,$$

but this is of smaller order than the first term in $\alpha_{\mathbf{R}}$. Consequently, we are free to apply Taylor's Theorem to the function $x \mapsto x/\|x\|$ in a neighborhood of at most constant radius of $\mathbf{X}_{i\cdot}$ not containing zero to obtain

$$\begin{aligned}\big(\widehat{\mathbf{Y}}\mathbf{W}_*^\top\big)_{i\cdot} - \mathbf{Y}_{i\cdot} &= \frac{\big(\widehat{\mathbf{X}}\mathbf{W}_*^\top\big)_{i\cdot}}{\|\widehat{\mathbf{X}}_{i\cdot}\|} - \frac{\mathbf{X}_{i\cdot}}{\|\mathbf{X}_{i\cdot}\|} \\ &= \mathbf{J}(\mathbf{X}_{i\cdot})\big((\widehat{\mathbf{X}}\mathbf{W}_*^\top)_{i\cdot} - \mathbf{X}_{i\cdot}\big) + \big(\widetilde{\mathbf{R}}_Y\big)_{i\cdot},\end{aligned}$$

where

$$\|e_i^\top \widetilde{\mathbf{R}}_Y\| \lesssim r^2 \max_{|\alpha|=2} \sup_{\|c-\mathbf{X}_i\|\leq r} \|\mathbf{D}^\alpha(c)\|,$$

where $\mathbf{D}^\alpha$ denotes the partial derivatives of the function $x \mapsto \frac{x}{\|x\|}$, and r satisfies

$$r \leq C\theta_i \left\{ \left(\frac{\theta_{\max}}{\theta_i}\right)^{1/2} \left[\frac{K\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda_{\min}^{1/2}\|\theta\|^2}\right] + \alpha_{\mathbf{R}} \right\}, \tag{F.22}$$

for some constant $C > 0$. We also have used the notation

$$\mathbf{J}(\mathbf{X}_{i\cdot}) = \frac{1}{\|\mathbf{X}_{i\cdot}\|}\left(\mathbf{I} - \frac{\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^\top}{\|\mathbf{X}_{i\cdot}\|^2}\right),$$

which is the Jacobian of the mapping $x \mapsto \frac{x}{\|x\|}$. Expanding further, we have that

$$\begin{aligned}\left(\widehat{\mathbf{Y}}\mathbf{W}_*^\top\right)_{i\cdot} - \mathbf{Y}_{i\cdot} &= \frac{1}{\|\mathbf{X}_{i\cdot}\|}\left(\mathbf{I} - \frac{\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^\top}{\|\mathbf{X}_{i\cdot}\|^2}\right)\left((\widehat{\mathbf{X}}\mathbf{W}_*)_{i\cdot} - \mathbf{X}_{i\cdot}\right) + \left(\widetilde{\mathbf{R}}_Y\right)_{i\cdot} \\ &= \frac{1}{\|\mathbf{X}_{i\cdot}\|}\left(\mathbf{I} - \frac{\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^\top}{\|\mathbf{X}_{i\cdot}\|^2}\right)\left((\mathbf{A}-\mathbf{P})\mathbf{U}|\Lambda|^{-1/2}\mathbf{I}_{p,q}\right)_{i\cdot} + \frac{1}{\|\mathbf{X}_{i\cdot}\|}\left(\mathbf{I} - \frac{\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^\top}{\|\mathbf{X}_{i\cdot}\|^2}\right)(\mathbf{R})_{i\cdot} + \left(\widetilde{\mathbf{R}}_Y\right)_{i\cdot}.\end{aligned}$$

This justifies the linear part of the expansion, where we define

$$\left(\mathcal{R}_{\text{Stage I}}\right)_{i\cdot} := \frac{1}{\|\mathbf{X}_{i\cdot}\|}\left(\mathbf{I} - \frac{\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^\top}{\|\mathbf{X}_{i\cdot}\|^2}\right)(\mathbf{R})_{i\cdot} + \left(\widetilde{\mathbf{R}}_Y\right)_{i\cdot}.$$

Therefore, it remains to bound this residual. Recall that we already have the bound

$$\|e_i^\top \mathbf{R}\| \lesssim \theta_i \alpha_{\mathbf{R}}$$

with probability at least $1 - O(n^{-18})$ by (F.20). Consequently, with this same probability, we note that $\|\mathbf{X}_i\| \gtrsim \theta_i$, so that

$$\begin{aligned}\left\|\frac{1}{\|\mathbf{X}_{i\cdot}\|}\left(\mathbf{I} - \frac{\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^\top}{\|\mathbf{X}_{i\cdot}\|^2}\right)(\mathbf{R})_{i\cdot}\right\| &\lesssim \frac{1}{\theta_i}\left\|\left(\mathbf{I} - \frac{\mathbf{X}_i\mathbf{X}_i^\top}{\|\mathbf{X}_i\|^2}\right)(\mathbf{R})_{i\cdot}\right\| \\ &\lesssim \alpha_{\mathbf{R}},\end{aligned}$$

since the term $\mathbf{I} - \frac{\mathbf{X}_{i\cdot}\mathbf{X}_{i\cdot}^\top}{\|\mathbf{X}_{i\cdot}\|^2}$ is a projection matrix. We therefore need only bound the term $e_i^\top \widetilde{\mathbf{R}}_Y$ which satisfies

$$\|e_i^\top \widetilde{\mathbf{R}}_Y\| \lesssim r^2 \max_{|\alpha|=2} \sup_{\|c-\mathbf{X}_{i\cdot}\|} \|\mathbf{D}^\alpha(c)\|.$$

We now note that the mixed partials of the mapping $x \mapsto \frac{x}{\|x\|}$ are given by

$$\frac{\partial^2}{\partial x_i \partial x_j}\frac{x_k}{\|x\|} = \frac{3x_i x_j x_k}{\|x\|^5} - \frac{\delta_{ik}x_j + \delta_{ij}x_k + \delta_{jk}x_i}{\|x\|^3}.$$

We evaluate this in a neighborhood of $\mathbf{X}_{i\cdot}$ of radius at most r where r satisfies the inequality

in (F.22). It is straightforward to observe that since $r \lesssim \|\mathbf{X}_{i\cdot}\|$, we have

$$\max_{|\alpha|=2} \sup_{\|c-\mathbf{X}_i\| \le r} \|\mathbf{D}^\alpha(c)\| \lesssim \frac{1}{\|\mathbf{X}_i\|^2}.$$

Therefore,

$$\begin{aligned}
\|e_i^\top \widetilde{\mathbf{R}}_Y\| &\lesssim \frac{r^2}{\|\mathbf{X}_i\|^2} \\
&\lesssim \frac{\theta_i^2}{\|\mathbf{X}_{i\cdot}\|^2}\left\{\left(\frac{\theta_{\max}}{\theta_i}\right)^{1/2}\left[\frac{K\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\lambda_{\min}^{1/2}\|\theta\|^2}\right] + \alpha_{\mathbf{R}}\right\}^2 \\
&\lesssim \left\{\left(\frac{\theta_{\max}}{\theta_i}\right)^{1/2}\left[\frac{K\sqrt{\theta_{\max}\|\theta\|_1\log(n)}}{\lambda_{\min}^{1/2}\|\theta\|^2}\right] + \alpha_{\mathbf{R}}\right\}^2 \\
&\lesssim \left(\frac{\theta_{\max}}{\theta_i}\right)\frac{K^2\theta_{\max}\|\theta\|_1\log(n)}{\lambda_{\min}\|\theta\|^4} + \alpha_{\mathbf{R}},
\end{aligned}$$

which holds as long as C in Assumption 6.2 is larger than t he universal constants above, and hence both terms will be smaller than one. Therefore, we obtain that

$$\begin{aligned}
\|e_i^\top \mathcal{R}_{\text{Stage I}}\| &\lesssim \left(\frac{\theta_{\max}}{\theta_{\min}}\right)\frac{K^2\theta_{\max}\|\theta\|_1\log(n)}{\lambda_{\min}\|\theta\|^4} + \alpha_{\mathbf{R}} \\
&\asymp \left(\frac{\theta_{\max}}{\theta_{\min}}\right)\frac{K^2\theta_{\max}\|\theta\|_1\log(n)}{\lambda_{\min}\|\theta\|^4} + \frac{K^2\theta_{\max}\|\theta\|_1}{\|\theta\|^4\lambda_{\min}^2} + \left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2}\left(\frac{K^{9/2}\theta_{\max}\|\theta\|_1\log(n)}{\|\theta\|^4\lambda_{\min}^{3/2}}\right) \\
&\lesssim \frac{K^2\theta_{\max}\|\theta\|_1}{\lambda_{\min}\|\theta\|^4}\left(\log(n)\frac{\theta_{\max}}{\theta_{\min}} + \frac{1}{\lambda_{\min}} + \left(\frac{\theta_{\max}}{\theta_{\min}}\right)^{1/2}\frac{K^{5/2}\log(n)}{\lambda_{\min}^{1/2}}\right)
\end{aligned}$$

which holds with probability at least $1 - O(n^{-18})$. This is the advertised bound, which completes the proof. □

F.2.3 Proofs of Lemmas 61 and 63

To prove these lemmas we require leave-one-out sequences, similar to Abbe et al. (2020). First we state the following lemma concerning the leave-one-out sequences. The proof is deferred to Appendix F.2.3.

Lemma 64 (Good properties of Leave-one-out sequences)**.** *Let $\mathbf{A}^{(l,-i)}$ denote the matrix $\mathbf{A}^{(l)}$ with its i'th row and column replaced with $\mathbf{P}^{(l)}$. Let $\widehat{\mathbf{U}}^{(-i)}$ denote the leading K eigenvectors*

of $\mathbf{A}^{(l,-i)}$. *Suppose that* $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$ *and* $\min_i \theta_i \|\theta\|_1 \gtrsim \log(n)$. *Then the following hold with probability at least* $1 - O(n^{-20})$*:*

$$\begin{aligned}
|\lambda_K(\mathbf{A}^{(l)}) - \lambda_{K+1}(\mathbf{A}^{(l,-i)})| &\gtrsim \lambda^{(l)};\\
\|e_i^\top(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\widehat{\mathbf{U}}^{(-i)}\| &\lesssim \sqrt{\theta_i\|\theta\|_1 \log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty};\\
\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\| &\lesssim \frac{\sqrt{\theta_i\|\theta\|_1 \log(n)}}{\lambda}\|\widehat{\mathbf{U}}\|_{2,\infty}.
\end{aligned}$$

We now prove Lemma 61. The statement is repeated for convenience.

Lemma 61 (Row-Wise Concentration I)**.** *When* $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$ *and* $\min_i \theta_i\|\theta\|_1 \gtrsim \log(n)$*, it holds that*

$$\|e_i^\top(\mathbf{A} - \mathbb{E}\mathbf{A})\widehat{\mathbf{U}}\| \lesssim \sqrt{\theta_i\|\theta\|_1 \log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty}$$

Proof of Lemma 61. First, let $\widehat{\mathbf{U}}^{(-i)}$ denote the eigenvectors of $\mathbf{A}^{(l)}$ with the i'th row and column replaced with the corresponding row and column of $\mathbf{P}^{(l)}$. Observe that

$$\begin{aligned}
\|e_i^\top(\mathbf{A} - \mathbb{E}\mathbf{A})\widehat{\mathbf{U}}\| &= \|e_i^\top(\mathbf{A} - \mathbb{E}\mathbf{A})\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\|\\
&\le \|e_i^\top(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\| + \|e_i^\top(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\big(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\big)\|\\
&\le \|e_i^\top(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\widehat{\mathbf{U}}^{(-i)}\| + \|e_i^\top(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\|\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\|\\
&\le \|e_i^\top(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\widehat{\mathbf{U}}^{(-i)}\| + \|(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\|\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\|\\
&\lesssim \sqrt{\theta_i\|\theta\|_1 \log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty} + \sqrt{\theta_{\max}\|\theta\|_1}\frac{\sqrt{\theta_i\|\theta\|_1 \log(n)}}{\lambda}\|\widehat{\mathbf{U}}\|_{2,\infty},
\end{aligned}$$

where the final inequality holds with probability at least $1 - O(n^{-20})$ by Lemma 64 and Lemma 59. Consequently, since $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1}$, we obtain that

$$\|e_i^\top(\mathbf{A} - \mathbb{E}\mathbf{A})\widehat{\mathbf{U}}\| \lesssim \sqrt{\theta_i\|\theta\|_1 \log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty}$$

with probability at least $1 - O(n^{-20})$ which completes the proof. □

We now restate Lemma 63 for convenience.

Lemma 63 (Row-wise Concentration II). *When* $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1 \log(n)}$ *and* $\min_i \theta_i \|\theta\|_1 \gtrsim \log(n)$*, with probability at least* $1 - O(n^{-19})$*, it holds that*

$$\|e_i^\top (\mathbf{A} - \mathbb{E}\mathbf{A})(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U} - \mathbf{U})\| \lesssim \sqrt{\theta_i \|\theta\|_1 \log(n)} \|\frac{\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda} \|\mathbf{U}\|_{2,\infty}.$$

Proof. First we will argue that

$$\begin{aligned}\|e_i^\top (\mathbf{A} - \mathbb{E}\mathbf{A})(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U} - \mathbf{U})\| \lesssim \sqrt{\theta_i \|\theta\|_1 \log(n)} \frac{\sqrt{\theta_i \|\theta\|_1 \log(n)}}{\lambda} \|\widehat{\mathbf{U}}\|_{2,\infty} \\ + \sqrt{\theta_i \|\theta\|_1 \log(n)} \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U} - \mathbf{U}\|_{2,\infty} \qquad \text{(F.23)}\end{aligned}$$

with probability at least $1 - O(n^{-20})$. Provided this is true, by Lemma 62, we have that

$$\begin{aligned}\|\widehat{\mathbf{U}}\|_{2,\infty} &\lesssim \|\mathbf{U}\|_{2,\infty}; \\ \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U} - \mathbf{U}\|_{2,\infty} &\lesssim \frac{\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda} \|\mathbf{U}\|_{2,\infty},\end{aligned}$$

with probability at least $1 - O(n^{-19})$. Plugging these in yields

$$\begin{aligned}\|e_i^\top (\mathbf{A} - \mathbb{E}\mathbf{A})(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U} - \mathbf{U})\| &\lesssim \sqrt{\theta_i \|\theta\|_1 \log(n)} \frac{\sqrt{\theta_i \|\theta\|_1 \log(n)}}{\lambda} \|\mathbf{U}\|_{2,\infty} \\ &\quad + \sqrt{\theta_i \|\theta\|_1 \log(n)} \|\frac{\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda} \|\mathbf{U}\|_{2,\infty} \\ &\lesssim \sqrt{\theta_i \|\theta\|_1 \log(n)} \|\frac{\sqrt{\theta_{\max}\|\theta\|_1 \log(n)}}{\lambda} \|\mathbf{U}\|_{2,\infty},\end{aligned}$$

which is the desired bound. Therefore, it remains to prove the claim (F.23).

Proceeding similarly to the proof of Lemma 61,

$$\begin{aligned}&\|e_i^\top (\mathbf{A} - \mathbb{E}\mathbf{A})(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U} - \mathbf{U})\| \\ &\quad \le \|e_i^\top \big(\mathbf{A} - \mathbf{P}\big)(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{U} - \widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U})\| + \|e_i^\top \big(\mathbf{A} - \mathbf{P}\big)\big(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\big)\| \\ &\quad \le \|e_i^\top \big(\mathbf{A} - \mathbf{P}\big)\|\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\| + \|e_i^\top \big(\mathbf{A} - \mathbf{P}\big)\big(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\big)\|.\end{aligned}$$

First, we note that the matrix $\big(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\big)$ is independent from the i'th row of

$\mathbf{A} - \mathbf{P}$. The matrix Bernstein inequality (Corollary 3.3 of Chen et al. (2021c)) shows that

$$\|e_i^\top (\mathbf{A} - \mathbf{P})\big(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\big)\| \le \sqrt{42v\log(n)} + \frac{42}{3} w \log(n)$$

with probability at least $1 - 2n^{-20}$, where we have defined

$$\begin{aligned} v :&= \max\Bigg\{ \Bigg\| \sum_{j=1}^n \mathbb{E}\big[(\mathbf{A}_{ij} - \mathbf{P}_{ij})\big(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\big)_{j\cdot}\big(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\big)_{j\cdot}^\top (\mathbf{A}_{ij} - \mathbf{P}_{ij})\Bigg\|, \\ &\qquad \Bigg\| \sum_{j=1}^n \mathbb{E}\Big[\big(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\big)_{j\cdot}^\top (\mathbf{A}_{ij} - \mathbf{P}_{ij})^2 \big(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\big)_{j\cdot}\Big]\Bigg\|\Bigg\}; \\ w :&= \max_{1\le j\le n} \|(\mathbf{A}_{ij} - \mathbf{P}_{ij})\big(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\big)_{j\cdot}\| \\ &\le \|\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\|_{2,\infty}. \end{aligned}$$

For the term v, we recognize that $\mathbf{A}_{ij} - \mathbf{P}_{ij}$ is a scalar, yielding

$$v \le \theta_i \|\theta\|_1 \|\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\|_{2,\infty}^2$$

(for details on this calculation, see the proof of Lemma 64). Consequently,

$$\begin{aligned} \|e_i^\top (\mathbf{A} - \mathbf{P})\big(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\big)\| &\le \sqrt{42v\log(n)} + \frac{42}{3} w \log(n) \\ &\lesssim \sqrt{\theta_i \|\theta\|_1 \log(n)} \|\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top \mathbf{U} - \mathbf{U}\|_{2,\infty}, \end{aligned}$$

as long as $\min_i \theta_i \|\theta\|_1 \gtrsim \log(n)$. Moreover, a straightforward Bernstein inequality argument shows that $\|e_i^\top (\mathbf{A} - \mathbf{P})\| \lesssim \sqrt{\theta_i \|\theta\|_1 \log(n)}$ with probability at least $1 - O(n^{-20})$. Consequently, by Lemma 64 and Lemma 59, with probability at least $1 - O(n^{-20})$ it holds

that

$$\begin{aligned}
&\|e_i^\top(\mathbf{A}-\mathbb{E}\mathbf{A})(\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}-\mathbf{U})\| \\
&\quad\le \|e_i^\top(\mathbf{A}-\mathbf{P})\|\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top-\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\| \\
&\qquad+\|e_i^\top(\mathbf{A}-\mathbf{P})(\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\mathbf{U}-\mathbf{U})\| \\
&\quad\lesssim \sqrt{\theta_i\|\theta\|_1\log(n)}\frac{\sqrt{\theta_i\|\theta\|_1\log(n)}}{\lambda}\|\widehat{\mathbf{U}}\|_{2,\infty} \\
&\qquad+\sqrt{\theta_i\|\theta\|_1\log(n)}\|\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\mathbf{U}-\mathbf{U}\|_{2,\infty} \\
&\quad\lesssim \sqrt{\theta_i\|\theta\|_1\log(n)}\frac{\sqrt{\theta_i\|\theta\|_1\log(n)}}{\lambda}\|\widehat{\mathbf{U}}\|_{2,\infty} \\
&\qquad+\sqrt{\theta_i\|\theta\|_1\log(n)}\Big(\|\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\mathbf{U}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}\|_{2,\infty}+\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}-\mathbf{U}\|_{2,\infty}\Big) \\
&\quad\lesssim \sqrt{\theta_i\|\theta\|_1\log(n)}\frac{\sqrt{\theta_i\|\theta\|_1\log(n)}}{\lambda}\|\widehat{\mathbf{U}}\|_{2,\infty}+\sqrt{\theta_i\|\theta\|_1\log(n)}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{U}-\mathbf{U}\|_{2,\infty}.
\end{aligned}$$

□

Proof of Lemma 64

We restate Lemma 64 for convenience.

Lemma 64 (Good properties of Leave-one-out sequences)**.** *Let $\mathbf{A}^{(l,-i)}$ denote the matrix $\mathbf{A}^{(l)}$ with its i'th row and column replaced with $\mathbf{P}^{(l)}$. Let $\widehat{\mathbf{U}}^{(-i)}$ denote the leading K eigenvectors of $\mathbf{A}^{(l,-i)}$. Suppose that $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1\log(n)}$ and $\min_i \theta_i\|\theta\|_1 \gtrsim \log(n)$. Then the following hold with probability at least $1-O(n^{-20})$:*

$$\begin{aligned}
|\lambda_K(\mathbf{A}^{(l)})-\lambda_{K+1}(\mathbf{A}^{(l,-i)})| &\gtrsim \lambda^{(l)}; \\
\|e_i^\top(\mathbf{A}^{(l)}-\mathbf{P}^{(l)})\widehat{\mathbf{U}}^{(-i)}\| &\lesssim \sqrt{\theta_i\|\theta\|_1\log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty}; \\
\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top-\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\| &\lesssim \frac{\sqrt{\theta_i\|\theta\|_1\log(n)}}{\lambda}\|\widehat{\mathbf{U}}\|_{2,\infty}.
\end{aligned}$$

Proof of Lemma 64. First, by Lemma 59, it holds that

$$\begin{aligned}
\|\mathbf{A}^{(l)}-\mathbf{P}^{(l)}\| &\lesssim \sqrt{\theta_{\max}\|\theta\|_1} \\
&\le \lambda/\sqrt{\log(n)}.
\end{aligned}$$

Therefore, Weyl's inequality shows that

$$\begin{aligned}|\lambda_K(\mathbf{A}^{(l)})| &\geq |\lambda_K| - \|\mathbf{A}^{(l)} - \mathbf{P}^{(l)}\| \\ &\geq \lambda - \lambda/\sqrt{\log(n)} \\ &\geq \lambda/2 \gtrsim \lambda,\end{aligned}$$

and that $|\lambda_{K+1}(\mathbf{A}^{(l)})| \leq \|\mathbf{A}^{(l)} - \mathbf{P}^{(l)}\| \leq |\lambda_K|/\sqrt{\log(n)}$. Therefore, $|\lambda_K(\mathbf{A}^{(l)})| - |\lambda_{K+1}(\mathbf{A}^{(l)})| \gtrsim \lambda$. Furthermore,

$$\|e_i^\top \big(\mathbf{A}^{(l)} - \mathbf{P}^{(l)}\big)\| \leq \|\mathbf{A}^{(l)} - \mathbf{P}^{(l)}\|.$$

Observe that $\mathbf{A}^{(l)} = \mathbf{A}^{(l,-i)} + e_i e_i^\top \big(\mathbf{A}^{(l)} - \mathbf{P}^{(l)}\big) + \big(\mathbf{A}^{(l)} - \mathbf{P}^{(l)}\big) e_i e_i^\top - e_i e_i^\top \big(\mathbf{A}^{(l)} - \mathbf{P}^{(l)}\big) e_i e_i^\top$. Consequently, by Weyl's inequality,

$$\begin{aligned}|\lambda_K(\mathbf{A}^{(l,-i)}) - \lambda_{K+1}(\mathbf{A}^{(l)})| &\geq |\lambda_K(\mathbf{A}^{(l)})| - |\lambda_{K+1}(\mathbf{A}^{(l)})| \\ &\quad - \left\| e_i e_i^\top \big(\mathbf{A}^{(l)} - \mathbf{P}^{(l)}\big) + \big(\mathbf{A}^{(l)} - \mathbf{P}^{(l)}\big) e_i e_i^\top - e_i e_i^\top \big(\mathbf{A}^{(l)} - \mathbf{P}^{(l)}\big) e_i e_i^\top \right\| \\ &\gtrsim |\lambda_K|.\end{aligned}$$

This proves the first assertion. As a byproduct, we are free to apply the Davis-Kahan Theorem to $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{U}}^{(-i)}$ to observe that

$$\begin{aligned}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\| &\lesssim \frac{\|e_i^\top \big(\mathbf{A} - \mathbf{P}^{(l)}\big)\widehat{\mathbf{U}}^{(-i)}\| + \|\big(\mathbf{A} - \mathbf{P}\big) e_i e_i^\top \widehat{\mathbf{U}}^{(-i)}\|}{\lambda} \\ &\lesssim \frac{\|e_i^\top \big(\mathbf{A} - \mathbf{P}^{(l)}\big)\widehat{\mathbf{U}}^{(-i)}\|}{\lambda} + \frac{\|e_i^\top \big(\mathbf{A} - \mathbf{P}\big)\| \|e_i^\top \widehat{\mathbf{U}}^{(-i)}\|}{\lambda}.\end{aligned}$$

Consequently, we need only bound the numerators above; however, a bound on the first term will also prove the second assertion of this lemma. Note that

$$e_i^\top \big(\mathbf{A} - \mathbf{P}\big)\widehat{\mathbf{U}}^{(-i)} = \sum_{j=1}^n \big(\mathbf{A}_{ij} - \mathbf{P}_{ij}\big)\widehat{\mathbf{U}}_{j\cdot}^{(-i)}.$$

Since $\widehat{\mathbf{U}}^{(-i)}$ is independent from the i'th row of $\mathbf{A}_{ij}$, this is a sum of n independent random

matrices condition on $\widehat{\mathbf{U}}^{(-i)}$. Therefore, the matrix Bernstein inequality (Corollary 3.3 of Chen et al. (2021c)) reveals that

$$\|e_i^\top (\mathbf{A}-\mathbf{P})\widehat{\mathbf{U}}^{(-i)}\| \le \sqrt{42v\log(n)} + \frac{42}{3}w\log(n)$$

with probability at least $1-2n^{-20}$. Here we note that

$$\begin{aligned} v &:= \max\left\{ \left\| \sum_{j=1}^n \mathbb{E}\big[(\mathbf{A}_{ij}-\mathbf{P}_{ij})\widehat{\mathbf{U}}_{j\cdot}^{(-i)}(\widehat{\mathbf{U}}_{j\cdot}^{(-i)})^\top(\mathbf{A}_{ij}-\mathbf{P}_{ij})\big]\right\|, \left\| \sum_{j=1}^n \mathbb{E}\big[(\widehat{\mathbf{U}}_{j\cdot}^{(-i)})^\top(\mathbf{A}_{ij}-\mathbf{P}_{ij})^2\widehat{\mathbf{U}}_{j\cdot}^{(-i)}\big]\right\|\right\}; \\ w &:= \max_{1\le j\le n} \|(\mathbf{A}_{ij}-\mathbf{P}_{ij})\widehat{\mathbf{U}}_{j\cdot}^{(-i)}\| \\ &\le \|\widehat{\mathbf{U}}^{(-i)}\|_{2,\infty}, \end{aligned}$$

where the expectation in the first term is conditional on $\widehat{\mathbf{U}}^{(-i)}$. Observing that $\mathbf{A}_{ij}-\mathbf{P}_{ij}$ is a scalar reveals that

$$\begin{aligned} v &\le \max\left\{ \left\| \sum_{j=1}^n \widehat{\mathbf{U}}_{j\cdot}^{(-i)}(\widehat{\mathbf{U}}_{j\cdot}^{(-i)})^\top \mathbb{E}(\mathbf{A}_{ij}-\mathbf{P}_{ij})^2\right\|, \left\| \sum_{j=1}^n (\widehat{\mathbf{U}}_{j\cdot}^{(-i)})^\top\widehat{\mathbf{U}}_{j\cdot}^{(-i)}\mathbb{E}(\mathbf{A}_{ij}-\mathbf{P}_{ij})^2\right\|\right\} \\ &\le \sum_{j=1}^n \|\widehat{\mathbf{U}}_{j\cdot}^{(-i)}\|^2\theta_i\theta_j \\ &\le \|\widehat{\mathbf{U}}^{(-i)}\|_{2,\infty}^2\theta_i\|\theta\|_1. \end{aligned}$$

Therefore, it holds that

$$\begin{aligned} \|e_i^\top (\mathbf{A}-\mathbf{P})\widehat{\mathbf{U}}^{(-i)}\| &\le \sqrt{42v\log(n)} + \frac{42}{3}L\log(n) \\ &\le \sqrt{42\theta_i\|\theta\|_1\log(n)}\|\widehat{\mathbf{U}}^{(-i)}\|_{2,\infty} + \frac{42}{3}\log(n)\|\widehat{\mathbf{U}}^{(-i)}\|_{2,\infty} \\ &\lesssim \sqrt{\theta_i\|\theta\|_1\log(n)}\|\widehat{\mathbf{U}}^{(-i)}\|_{2,\infty}, \end{aligned}$$

which holds as long as $\min_i \theta_i\|\theta\|_1 \gtrsim \log(n)$. Moreover, we have that $\|e_i^\top(\mathbf{A}-\mathbf{P})\| \lesssim \sqrt{\theta_i\|\theta\|_1\log(n)}$ by a direct application of matrix Bernstein again. Consequently, applying

these bounds yields that

$$\begin{aligned}\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\| &\lesssim \frac{\|e_i^\top(\mathbf{A}-\mathbf{P}^{(l)})\widehat{\mathbf{U}}^{(-i)}\|}{\lambda} + \frac{\|e_i^\top(\mathbf{A}-\mathbf{P})\|\|e_i^\top\widehat{\mathbf{U}}^{(-i)}\|}{\lambda} \\ &\lesssim \frac{\sqrt{\theta_i\|\theta\|_1\log(n)}}{\lambda}\|\widehat{\mathbf{U}}^{(-i)}\|_{2,\infty}.\end{aligned}$$

As a byproduct, we also have that

$$\begin{aligned}\|\widehat{\mathbf{U}}^{(-i)}\|_{2,\infty} &= \|\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\|_{2,\infty} \\ &\le \|\widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\|_{2,\infty} + \|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\|_{2,\infty} \\ &\le \frac{1}{2}\|\widehat{\mathbf{U}}^{(-i)}\|_{2,\infty} + \|\widehat{\mathbf{U}}\|_{2,\infty},\end{aligned}$$

which holds as long as $\lambda \gtrsim \sqrt{\theta_{\max}\|\theta\|_1\log(n)}$. Consequently, by rearranging, we have that $\|\widehat{\mathbf{U}}^{(-i)}\|_{2,\infty} \lesssim \|\widehat{\mathbf{U}}\|_{2,\infty}$ which yields the inequality

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \widehat{\mathbf{U}}^{(-i)}(\widehat{\mathbf{U}}^{(-i)})^\top\| \lesssim \frac{\sqrt{\theta_i\|\theta\|_1\log(n)}}{\lambda}\|\widehat{\mathbf{U}}\|_{2,\infty},$$

which holds with probability at least $1 - O(n^{-20})$. Moreover, with this same probability, we have that

$$\|e_i^\top(\mathbf{A}-\mathbf{P})\widehat{\mathbf{U}}^{(-i)}\| \lesssim \sqrt{\theta_i\|\theta\|_1\log(n)}\|\widehat{\mathbf{U}}\|_{2,\infty}.$$

This completes the proof. □

F.3 Proof of Second Stage $\sin\Theta$ Bound (Theorem 20)

First we will restate Theorem 20.

Theorem 20 ($\sin\Theta$ Perturbation Bound)**.** *Suppose the conditions in Theorem 16 hold. Define*

$$\alpha_{\max} = \frac{K^2\theta_{\max}^{(l)}\|\theta^{(l)}\|_1}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^4}\left(\log(n)\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} + \frac{1}{\lambda_{\min}^{(l)}} + \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\right)^{1/2}\frac{K^{5/2}\log(n)}{(\lambda_{\min}^{(l)})^{1/2}}\right);$$

i.e., $\alpha_{\max}$ *is the residual upper bound from Theorem 19. Then with probability at least* $1 - O(n^{-10})$*, it holds that*

$$\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}}, \mathbf{U})\| \lesssim K^2\sqrt{\log(n)}\frac{\left(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\right)^{1/2}}{\sqrt{L\bar{\lambda}}} + K^3\log(n)\frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} + K^2\sqrt{\log(n)}\frac{\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K\alpha_{\max}}{\bar{\lambda}}.$$

In particular, under the conditions of Theorem 16 it holds that

$$\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}}, \mathbf{U})\| \lesssim \frac{1}{K}.$$

In what follows we give a high-level overview of the proof. Define the matrix $\mathcal{Y} := [\mathbf{Y}^{(1)}, \cdots, \mathbf{Y}^{(L)}] \in \mathbb{R}^{n\times LK}$, and let $\widehat{\mathcal{Y}}$ be defined similarly. Since we consider the singular vectors of $\mathcal{Y}$ and $\widehat{\mathcal{Y}}$, we will examine the *eigenvectors* of their associated $n\times n$ Gram matrices, or the matrices $\mathcal{Y}\mathcal{Y}^\top$ and $\widehat{\mathcal{Y}}\widehat{\mathcal{Y}}^\top$ respectively. Therefore, we will view $\widehat{\mathcal{Y}}\widehat{\mathcal{Y}}^\top$ as a perturbation of matrix $\mathcal{Y}\mathcal{Y}^\top$. We expand via

$$\mathcal{Y}\mathcal{Y}^\top - \widehat{\mathcal{Y}}\widehat{\mathcal{Y}}^\top = \mathcal{L}(\mathcal{E})\mathcal{Y}^\top + \mathcal{Y}\mathcal{L}(\mathcal{E})^\top + \mathcal{R}_{\mathrm{all}},$$

where we define

$$\mathcal{R}_{\mathrm{all}} := \sum_l \mathcal{L}(\mathbf{E}^{(l)})\mathcal{L}(\mathbf{E}^{(l)})^\top + \mathcal{L}(\mathbf{E}^{(l)})(\mathcal{R}^{(l)}_{\text{Stage I}})^\top + \mathcal{R}^{(l)}_{\text{Stage I}}\mathcal{L}(\mathbf{E}^{(l)})^\top + \mathcal{R}^{(l)}_{\text{Stage I}}(\mathcal{R}^{(l)}_{\text{Stage I}})^\top + \sum_l \mathcal{R}^{(l)}_{\text{Stage I}}(\mathbf{Y}^{(l)})^\top + \mathbf{Y}^{(l)}(\mathcal{R}^{(l)}_{\text{Stage I}})^\top,$$

and

$$\mathcal{L}(\mathcal{E}) := \left[\mathcal{L}(\mathbf{E}^{(1)}), \cdots, \mathcal{L}(\mathbf{E}^{(L)})\right],$$

where we have defined $\mathbf{E}^{(l)}$ as the mean-zero random matrix $\mathbf{E}^{(l)} := \mathbf{A}^{(l)} - \mathbf{P}^{(l)}$. Hence,

$$\mathcal{L}(\mathcal{E})\mathcal{Y}^\top + \mathcal{Y}\mathcal{L}(\mathcal{E})^\top = \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top + \mathbf{Y}^{(l)}\mathcal{L}(\mathbf{E}^{(l)})^\top.$$

By virtue of the tight characterization for each $\mathcal{R}_Y^{(l)}$ in Theorem 19, we can see that $\mathcal{Y}\mathcal{Y}^\top$ is *nearly* a linear perturbation of $\widehat{\mathcal{Y}}\widehat{\mathcal{Y}}^\top$. The proof of Theorem 20 makes this rigorous.

F.3.1 Preliminary Lemmas: Spectral Norm Concentration Bounds

Throughout this section we use the notation $\mathbf{E}^{(l)} := \mathbf{A}^{(l)} - \mathbf{P}^{(l)}$. The following lemma bounds several terms involving $\mathcal{L}(\mathbf{E}^{(l)})$ in spectral norm.

Lemma 65. *It holds that*

$$\|\mathcal{L}(\mathbf{E}^{(l)})\| \lesssim \frac{K\sqrt{n\log(n)}}{(\lambda_{\min}^{(l)})^{1/2}\|\theta^{(l)}\|}\left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\right)^{1/2};$$
$$\|\sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\| \lesssim Kn\sqrt{L\log(n)}\left[\frac{1}{L}\sum_l\left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\right)\frac{1}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^2}\right]^{1/2}.$$

with probability at least $1 - O(n^{-15})$.

Proof of Lemma 65. We recall that

$$\mathcal{L}(\mathbf{E}^{(l)})_{i\cdot} = \mathbf{J}(\mathbf{X}_{i\cdot})\left((\mathbf{A}^{(l)} - \mathbf{P}^{(l)})\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\right)_{i\cdot}.$$

Therefore, we can write this matrix via

$$\begin{aligned}\mathcal{L}(\mathbf{E}^{(l)}) &= \left(\sum_{i,j}\mathbf{E}_{ij}^{(l)}e_ie_j^\top\right)\left(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\right)\\ &= \sum_{i\le j}\mathbf{E}_{ij}^{(l)}e_ie_j^\top\left(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\right) + \sum_{j<i}\mathbf{E}_{ij}^{(l)}e_je_i^\top\left(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{j\cdot})\right),\end{aligned}$$

both of which are a sum of independent random matrices. Without loss of generality we bound the first term; the second is similar. We will apply the matrix Bernstein inequality

(Chen et al. (2021c), Corollary 3.3). We need to bound:

$$v := \max\Bigg\{\Bigg\|\sum_{i\leq j}\mathbb{E}(\mathbf{E}_{ij}^{(l)})^2\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\Big)^\top e_ie_j^\top e_je_i^\top\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\Big)^\top\Bigg\|,$$
$$\Bigg\|\sum_{i\leq j}\mathbb{E}(\mathbf{E}_{ij}^{(l)})^2e_ie_j^\top\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\Big)\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\Big)^\top e_je_i^\top\Bigg\|\Bigg\};$$
$$w := \max_{i,j}\|\mathbf{E}_{ij}^{(l)}e_ie_j^\top\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\Big)\|.$$

For v, we note that

$$\begin{aligned}
&\Bigg\|\sum_{i\leq j}\mathbb{E}(\mathbf{E}_{ij}^{(l)})^2\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\Big)^\top e_ie_j^\top e_je_i^\top\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\Big)^\top\Bigg\|\\
&\quad\leq\sum_{i\leq j}\mathbb{E}(\mathbf{E}_{ij}^{(l)})^2\Bigg\|\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\Big)^\top e_ie_j^\top e_je_i^\top\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\Big)^\top\Bigg\|\\
&\quad\leq\sum_{i\leq j}\theta_i^{(l)}\theta_j^{(l)}\|e_j^\top\mathbf{U}^{(l)}\|^2\||\Lambda^{(l)}|^{-1/2}\|^2\|\mathbf{J}(\mathbf{X}_{i\cdot})\|^2\\
&\quad\leq\frac{K}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^2}\sum_{i\leq j}\theta_i^{(l)}\theta_j^{(l)}\|e_j^\top\mathbf{U}^{(l)}\|^2\|\mathbf{J}(\mathbf{X}_{i\cdot})\|^2\\
&\quad\leq\frac{K}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^2}\sum_{i\leq j}\theta_i^{(l)}\theta_j^{(l)}\frac{(\theta_j^{(l)})^2K}{\|\theta^{(l)}\|^2}\frac{1}{(\theta_i^{(l)})^2}\\
&\quad\leq\frac{K^2}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^4}\sum_{i\leq j}\frac{\theta_j^{(l)}}{\theta_i^{(l)}}(\theta_j^{(l)})^2\\
&\quad\leq\frac{K^2n}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^4}\Big(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\Big)\sum_j\theta_j^2\\
&\quad\leq\frac{K^2n}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^4}\Big(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\Big)\|\theta^{(l)}\|^2\\
&\quad\leq\frac{K^2n}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^2}\Big(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\Big).
\end{aligned}$$

The other term satisfies the same upper bound. In addition,

$$\begin{aligned} w &= \max_{i,j} \|\mathbf{E}_{ij}^{(l)} e_i e_j^\top \left(\mathbf{U}^{(l)} |\Lambda^{(l)}|^{-1/2} \mathbf{I}_{p,q}^{(l)} \mathbf{J}(\mathbf{X}_{i\cdot}) \right) \| \\ &\leq \|\mathbf{U}^{(l)}\|_{2,\infty} \| |\Lambda^{(l)}|^{-1/2} \| \max_i \|\mathbf{J}(\mathbf{X}_{i\cdot})\| \\ &\leq \frac{K}{\|\theta\|^2 \lambda_{\min}^{(l)}} \left(\frac{\theta_{\max}^{(l)}}{(\theta_{\min}^{(l)})^{1/2}} \right). \end{aligned}$$

Therefore, by the Matrix Bernstein inequality, with probability at least $1 - O(n^{-20})$ it holds that

$$\begin{aligned} \|\mathcal{L}(\mathbf{E}^{(l)})\| &\lesssim \sqrt{v \log(n)} + w \log(n) \\ &\lesssim \frac{K\sqrt{n \log(n)}}{(\lambda_{\min}^{(l)})^{1/2} \|\theta^{(l)}\|} \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right)^{1/2} + \frac{K \log(n)}{\|\theta^{(l)}\|^2 (\lambda_{\min}^{(l)})^{1/2}} \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right) \\ &\leq \frac{K\sqrt{\log(n)}}{(\lambda_{\min}^{(l)})^{1/2} \|\theta^{(l)}\|} \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right)^{1/2} \max \left\{ \sqrt{n}, \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right)^{1/2} \frac{\sqrt{\log(n)}}{\|\theta^{(l)}\|} \right\}. \end{aligned}$$

Finally, we note that by Assumption 6.2, it holds that $\frac{\theta_{\max}}{\theta_{\min}} \lesssim \sqrt{n}$, which implies that $\sqrt{n}$ is the maximum of the term above. Therefore,

$$\|\mathcal{L}(\mathbf{E}^{(l)})\| \lesssim \frac{K\sqrt{n \log(n)}}{(\lambda_{\min}^{(l)})^{1/2} \|\theta^{(l)}\|} \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right)^{1/2},$$

which completes the proof of the first statement.

For the next statement, we proceed similarly, only now streamlining the analysis. Representing the sum similarly, we have that

$$\begin{aligned} \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top = \sum_l \Bigg(\sum_{i \leq j} \mathbf{E}_{ij}^{(l)} e_i e_j^\top \left(\mathbf{U}^{(l)} |\Lambda^{(l)}|^{-1/2} \mathbf{I}_{p,q}^{(l)} \mathbf{J}(\mathbf{X}_{i\cdot}) \right) (\mathbf{Y}^{(l)})^\top \\ + \sum_l \Bigg(\sum_{j < i} \mathbf{E}_{ij}^{(l)} e_j e_i^\top \left(\mathbf{U}^{(l)} |\Lambda^{(l)}|^{-1/2} \mathbf{I}_{p,q}^{(l)} \mathbf{J}(\mathbf{X}_{j\cdot}) \right) (\mathbf{Y}^{(l)})^\top. \end{aligned}$$

We focus again on the first term. Since it holds that $\|\mathbf{Y}^{(l)}\| \leq \|\mathbf{Y}^{(l)}\|_F = \sqrt{n}$, we have that

$$\begin{aligned} v &\leq \sum_l \frac{K^2 n^2}{\lambda_{\min}^{(l)} \|\theta^{(l)}\|^2} \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right) \\ &= K^2 n^2 \sum_l \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right) \frac{1}{\lambda_{\min}^{(l)} \|\theta^{(l)}\|^2} \end{aligned}$$

and

$$\begin{aligned} w &= \max_{i,j,m} \|\mathbf{E}_{ij}^{(l)} e_i e_j^\top \left(\mathbf{U}^{(l)} |\Lambda^{(l)}|^{-1/2} \mathbf{I}_{p,q}^{(l)} \mathbf{J}(\mathbf{X}_{i\cdot}) \right) (\mathbf{Y}^{(l)})^\top \| \\ &\leq \max_l \frac{K\sqrt{n}}{\|\theta^{(l)}\|^2 (\lambda_{\min}^{(l)})^{1/2}} \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right). \end{aligned}$$

Therefore, with probability at least $1 - O(n^{-15})$,

$$\begin{aligned} \| \sum_l \mathcal{L}(\mathbf{E}^{(l)}) (\mathbf{Y}^{(l)})^\top \| \lesssim Kn\sqrt{\log(n)} \left[\sum_l \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right) \frac{1}{\lambda_{\min}^{(l)} \|\theta^{(l)}\|^2} \right]^{1/2} \\ + K\sqrt{n} \log(n) \max_l \frac{1}{\|\theta^{(l)}\|^2 (\lambda_{\min}^{(l)})^{1/2}} \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right). \end{aligned}$$

Finally, we note that as long as $\frac{\theta_{\max}}{\theta_{\min}} \lesssim \sqrt{n}$, the first term dominates. Therefore,

$$\| \sum_l \mathcal{L}(\mathbf{E}^{(l)}) (\mathbf{Y}^{(l)})^\top \| \lesssim Kn\sqrt{L\log(n)} \left[\frac{1}{L} \sum_l \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \right) \frac{1}{\lambda_{\min}^{(l)} \|\theta^{(l)}\|^2} \right]^{1/2}.$$

□

Next, we bound residual term $\mathcal{R}_{\text{all}}$ in spectral norm.

Lemma 66. *The residual term $\mathcal{R}_{\text{all}}$ satisfies*

$$\|\mathcal{R}_{\text{all}}\| \lesssim LK^2 n \log(n) \|\text{SNR}^{-1}\|_\infty^2 + KLn\sqrt{\log(n)} \alpha_{\max} \|\text{SNR}^{-1}\|_\infty + nL\alpha_{\max}$$

with probability at least $1 - O(n^{-15})$.

Proof of Lemma 66. Recall that

$$\begin{aligned}\mathcal{R}_{\text{all}} :&= \sum_l \mathcal{L}(\mathbf{E}^{(l)})\mathcal{L}(\mathbf{E}^{(l)})^\top + \mathcal{L}(\mathbf{E}^{(l)})(\mathcal{R}^{(l)}_{\text{Stage I}})^\top + \mathcal{R}^{(l)}_{\text{Stage I}}\mathcal{L}(\mathbf{E}^{(l)})^\top + \mathcal{R}^{(l)}_{\text{Stage I}}(\mathcal{R}^{(l)}_{\text{Stage I}})^\top \\ &\quad + \sum_l \mathcal{R}^{(l)}_{\text{Stage I}}(\mathbf{Y}^{(l)})^\top + \mathbf{Y}^{(l)}(\mathcal{R}^{(l)}_{\text{Stage I}})^\top \\ &:= (I) + (II) + (III) + (IV),\end{aligned}$$

where

$$\begin{aligned}(I) &:= \sum_l \mathcal{L}(\mathbf{E}^{(l)})\mathcal{L}(\mathbf{E}^{(l)})^\top; \\ (II) &:= \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathcal{R}^{(l)}_{\text{Stage I}})^\top + (\mathcal{R}^{(l)}_{\text{Stage I}})\mathcal{L}(\mathbf{E}^{(l)})^\top; \\ (III) &:= \sum_l \mathcal{R}^{(l)}_{\text{Stage I}}(\mathcal{R}^{(l)}_{\text{Stage I}})^\top; \\ (IV) &:= \sum_l \mathcal{R}^{(l)}_{\text{Stage I}}(\mathbf{Y}^{(l)})^\top + \mathbf{Y}^{(l)}(\mathcal{R}^{(l)}_{\text{Stage I}})^\top.\end{aligned}$$

We bound each term separately.

The Term (I): We note that by Lemma 65 we have the bound

$$\|\mathcal{L}(\mathbf{E}^{(l)})\| \lesssim \frac{K\sqrt{n\log(n)}}{(\lambda^{(l)}_{\min})^{1/2}\left\|\theta^{(l)}\right\|}\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}.$$

Therefore,

$$\begin{aligned}\left\|\sum_l \mathcal{L}(\mathbf{E}^{(l)})\mathcal{L}(\mathbf{E}^{(l)})\right\| &\lesssim L\max_l\left(\frac{K\sqrt{n\log(n)}}{(\lambda^{(l)}_{\min})^{1/2}\left\|\theta^{(l)}\right\|}\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\right)^2 \\ &= LK^2 n\log(n)\max_l\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)\frac{1}{\lambda^{(l)}_{\min}\|\theta^{(l)}\|^2} \\ &\asymp LK^2 n\log(n)\|\text{SNR}^{-1}\|^2_\infty. \qquad \text{(F.24)}\end{aligned}$$

The term (II) : without loss of generality we consider the first term. By Lemma 57, it

holds that

$$\|\mathcal{R}^{(l)}_{\text{Stage I}}\| \lesssim \sqrt{n}\alpha^{(l)},$$

where $\alpha^{(l)}$ is the residual bound from Theorem 19. Therefore,

$$\begin{aligned}\sum_l \|\mathcal{L}(\mathbf{E}^{(l)})\mathcal{R}^{(l)}_{\text{Stage I}}\| &\lesssim L\sqrt{n}\max_l \alpha^{(l)} \max_l \|\mathcal{L}(\mathbf{E}^{(l)})\| \\ &\asymp KLn\sqrt{\log(n)}\alpha_{\max}\|\text{SNR}^{-1}\|_\infty, \end{aligned} \tag{F.25}$$

where we set $\alpha_{\max} := \max_l \alpha^{(l)}$.

The Term (III)**:** By a similar argument,

$$\begin{aligned}(III) &\lesssim nL\max_l(\alpha^{(l)})^2 \\ &\lesssim nL\alpha^2_{\max}. \end{aligned} \tag{F.26}$$

The term (IV)**:** Finally, it holds that

$$\begin{aligned}\sum_l \|\mathcal{R}^{(l)}_{\text{Stage I}}\|\|\mathbf{Y}^{(l)}\| &\lesssim L\sqrt{n}\alpha_{\max}\max_l \|\mathbf{Y}^{(l)}\| \\ &\lesssim Ln\alpha_{\max}. \end{aligned} \tag{F.27}$$

Putting it all together: Combining (F.24), (F.25), (F.26), and (F.27), we have that

$$\begin{aligned}\|\mathcal{R}_{\text{all}}\| &\lesssim LK^2n\log(n)\|\text{SNR}^{-1}\|^2_\infty + KLn\sqrt{\log(n)}\alpha_{\max}\|\text{SNR}^{-1}\|_\infty + nL\alpha^2_{\max} + nL\alpha_{\max} \\ &\asymp LK^2n\log(n)\|\text{SNR}^{-1}\|^2_\infty + KLn\sqrt{\log(n)}\alpha_{\max}\|\text{SNR}^{-1}\|_\infty + nL\alpha_{\max},\end{aligned}$$

since $\alpha_{\max} < 1$ by Assumption 6.2 (as shown in the proof of Theorem 19). □

F.3.2 Proof of Theorem 20

Proof of Theorem 20. First, by Lemma 65, we have the bound

$$\|\sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\| \lesssim Kn\sqrt{L\log(n)}\left[\frac{1}{L}\sum_l \left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)\frac{1}{\lambda^{(l)}_{\min}\|\theta^{(l)}\|^2}\right]^{1/2}.$$

Recall we define

$$(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2) := \frac{1}{L}\sum_l \left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)\frac{1}{\lambda^{(l)}_{\min}\|\theta^{(l)}\|^2}.$$

Then the bound can be concisely written as

$$\|\sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\| \lesssim Kn\sqrt{L\log(n)}(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2)^{1/2}.$$

In addition, by Lemma 66, we have that

$$\mathcal{R}_{\mathrm{all}} \lesssim LK^2 n\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2 + KLn\sqrt{\log(n)}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty + nL\alpha_{\max}.$$

Therefore, it holds that

$$\begin{aligned}\|\mathcal{Y}\mathcal{Y}^\top - \widehat{\mathcal{Y}}\widehat{\mathcal{Y}}^\top\| \lesssim{}& Kn\sqrt{L\log(n)}(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2)^{1/2} + LK^2 n\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2 \\ &+ KLn\sqrt{\log(n)}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty + nL\alpha_{\max}.\end{aligned}$$

Recall that $\lambda_Y^2 \gtrsim \frac{n}{K}L\bar{\lambda}$ by Lemma 8. Therefore, as long as

$$\begin{aligned}nL\bar{\lambda} \gtrsim{}& K^2 n\sqrt{L\log(n)}(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2)^{1/2} + LK^3 n\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2 \\ &+ K^2 Ln\sqrt{\log(n)}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty + nLK\alpha_{\max} \qquad \text{(F.28)}\end{aligned}$$

it holds that

$$\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\| \lesssim K^2\sqrt{\log(n)}\frac{\left(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\right)^{1/2}}{\sqrt{L\bar{\lambda}}} + K^3\log(n)\frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} + K^2\sqrt{\log(n)}\frac{\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K\alpha_{\max}}{\bar{\lambda}}. \tag{F.29}$$

Since the events listed above hold together with probability at least $1 - O(Ln^{-15})$, we see that the whole event holds with probability at least $1 - O(n^{-10})$ by the assumption that $L \lesssim n^5$.

We now verify (F.28). It is sufficient to check that the $\sin\boldsymbol{\Theta}$ bound in (F.29) is less than one (which is equivalent to checking (F.28)). In fact, we will show that each term is less than (in order) $\frac{1}{K}$, which is the second statement of the result.

Assumption 6.2 requires that

$$C\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)\frac{K^8\theta^{(l)}_{\max}\|\theta^{(l)}\|_1\log(n)}{\|\theta^{(l)}\|^4(\lambda^{(l)}_{\min})^2} \le \bar{\lambda}.$$

This immediately implies that $\frac{\alpha_{\max}}{\bar{\lambda}} \lesssim \frac{1}{K}$ from the definition of $\alpha_{\max}$. By plugging in the definition of SNR_l^{-1}, we see that we require

$$CX\frac{K^8\theta^{(l)}_{\max}\|\theta^{(l)}\|_1\log(n)}{\|\theta^{(l)}\|^2\mathrm{SNR}_l^2} \le \bar{\lambda}\lambda^{(l)}_{\min}.$$

Therefore the final three terms being are less than $\frac{1}{K}$ since $\frac{\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|^2}$ is always larger than one. For the remaining term, we observe that by averaging the above equation over L, we require that

$$C\frac{K^8\log(n)}{L}\sum_l \frac{\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|^2\mathrm{SNR}_l^2} \le \bar{\lambda}^2. \tag{F.30}$$

By squaring the first term, we see that we need the first term to satisfy

$$\frac{K^2\log(n)}{L^2\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_2^2 \lesssim \frac{1}{K^2}.$$

This is weaker than the condition (F.30). The proof is now complete. □

F.4 Proof of Second Stage Asymptotic Expansion (Theorem 21)

First we will restate Theorem 21.

Theorem 21 (Asymptotic Expansion: Stage II)**.** *Suppose the conditions of Theorem 16 hold. Define*

$$\mathbf{W}_* := \operatorname*{arg\,min}_{\mathbf{W}\in\mathbb{O}(K)} \|\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}\|_F.$$

There is an event $\mathcal{E}_{\text{Stage II}}$ satisfying $\mathbb{P}\big(\mathcal{E}_{\text{Stage II}}\big) \geq 1 - O(n^{-10})$ such that on this event, we have the asymptotic expansion

$$\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U} = \sum_l \mathcal{L}(\mathbf{A}^{(l)} - \mathbf{P}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2} + \mathcal{R}_{\text{Stage II}},$$

where $\mathcal{L}(\cdot)$ is the operator from Theorem 19 and the residual satisfies

$$\begin{aligned}\|\mathcal{R}_{\text{Stage II}}\|_{2,\infty} \lesssim \frac{K^3\sqrt{\log(n)}}{nL\bar{\lambda}}\|\text{SNR}^{-1}\|_2 + \frac{K^4\log(n)}{L^2\sqrt{n}\bar{\lambda}^2}\|\text{SNR}^{-1}\|_2^2 \\ + \frac{K^{7/2}\log(n)}{\sqrt{n}\bar{\lambda}}\|\text{SNR}^{-1}\|_\infty^2 + \frac{\alpha_{\max}}{\sqrt{n}\bar{\lambda}}.\end{aligned}$$

Here $\alpha_{\max}$ is as Theorem 20. In particular, under the assumptions of Theorem 16, it holds that

$$\|\mathcal{R}_{\text{Stage II}}\|_{2,\infty} \leq \frac{1}{16\sqrt{n_{\max}}}.$$

To prove Theorem 21 we first state and prove several $\|\cdot\|_{2,\infty}$ concentration results for the residual terms that arise in the asymptotic expansion, and we prove Theorem 21 in Appendix F.4.2.

F.4.1 Preliminary Lemmas: $\ell_{2,\infty}$ Residual Concentration Bounds

The following lemma bounds each of these residual terms in $\|\cdot\|_{2,\infty}$.

Lemma 67 (Second Stage Residual Bounds)**.** *The following bounds hold with probability at*

least $1 - O(n^{-10})$*:*

$$\|\mathbf{U}\mathbf{U}^\top \mathcal{L}(\mathcal{E})\mathcal{Y}^\top \widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\|_{2,\infty} \lesssim \frac{K^3\sqrt{\log(n)}}{n\sqrt{L\bar{\lambda}}}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2};$$
$$\|(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathcal{R}_{\mathrm{all}}\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\|_{2,\infty} \lesssim \frac{K^{3/2}\alpha_{\max}}{\sqrt{n\bar{\lambda}}} + \frac{K^{7/2}\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2}{\sqrt{n\bar{\lambda}}} + \frac{K^{5/2}\sqrt{\log(n)}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\sqrt{n\bar{\lambda}}}.$$

Proof of Lemma 67. At the outset, we note that Weyl's inequality and the condition in Theorem 20 implies that $\|\widehat{\Sigma}^{-2}\| \lesssim K(\bar{\lambda}nL)^{-1}$ with high probability.

We analyze each term separately. First, we observe that

$$\begin{aligned}\|\mathbf{U}\mathbf{U}^\top \mathcal{L}(\mathcal{E})\mathcal{Y}^\top \widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\|_{2,\infty} &\lesssim \|\mathbf{U}\|_{2,\infty}\|\|\mathbf{U}^\top \mathcal{L}(\mathcal{E})\mathcal{Y}^\top\|\|\widehat{\Sigma}^{-2}\| \\ &\lesssim \frac{K^{3/2}}{\sqrt{n}}\frac{1}{nL\bar{\lambda}}\|\mathbf{U}^\top \mathcal{L}(\mathcal{E})\mathcal{Y}^\top\|.\end{aligned}$$

We now establish a concentration inequality for the term $\mathbf{U}^\top \mathcal{L}(\mathcal{E})\mathcal{Y}^\top$. The result is similar to the proof of Lemma 65, so we postpone it to the end. For now, we simply state that with probability at least $1 - O(n^{-20})$,

$$\|\mathbf{U}^\top \mathcal{L}(\mathcal{E})\mathcal{Y}^\top\| \lesssim K^{3/2}\sqrt{nL\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2} + K^{3/2}\log(n)\max_l\left(\frac{\theta_{\max}}{\theta_{\min}}\right)\frac{1}{\|\theta^{(l)}\|^2(\lambda_{\min}^{(l)})^{1/2}} \tag{F.31}$$
$$\lesssim K^{3/2}\sqrt{nL\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2},$$

as long as $\max_l \frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}} \lesssim \sqrt{n/\log(n)}$, which holds under Assumption 6.2. Putting it together, we obtain

$$\begin{aligned}\|\mathbf{U}\mathbf{U}^\top \mathcal{L}(\mathcal{E})\mathcal{Y}^\top \widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\|_{2,\infty} &\lesssim \frac{K^{3/2}}{\sqrt{n}}\frac{1}{nL\bar{\lambda}}K^{3/2}\sqrt{nL\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2} \\ &\asymp \frac{K^3\sqrt{\log(n)}}{n\sqrt{L\bar{\lambda}}}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}.\end{aligned}$$

For the next term, we note that

$$\begin{aligned}\|(\mathbf{I}-\mathbf{U}\mathbf{U}^\top)\mathcal{R}_{\text{all}}\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\|_{2,\infty} &\le \|\mathcal{R}_{\text{all}}\|_{2,\infty}\|\widehat{\Sigma}^{-2}\| + \|\mathbf{U}\|_{2,\infty}\|\mathcal{R}_{\text{all}}\|\|\widehat{\Sigma}^{-2}\| \\ &\lesssim K\frac{\|\mathcal{R}_{\text{all}}\|_{2,\infty}}{nL\bar{\lambda}} + \frac{K^{3/2}}{n^{3/2}L\bar{\lambda}}\|\mathcal{R}_{\text{all}}\|. \end{aligned} \tag{F.32}$$

By Lemma 58, Lemma 65, and Lemma 57, we have the bounds

$$\begin{aligned}\|\mathcal{L}(\mathbf{E}^{(l)})\|_{2,\infty} &\lesssim \left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\frac{K\sqrt{\log(n)}}{(\lambda^{(l)}_{\min})^{1/2}\|\theta^{(l)}\|};\\ \|\mathcal{L}(\mathbf{E}^{(l)})\| &\le \frac{K\sqrt{n\log(n)}}{(\lambda^{(l)}_{\min})^{1/2}\left\|\theta^{(l)}\right\|}\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2};\\ \|\mathcal{R}_{\text{all}}\| &\lesssim LK^2 n\log(n)\|\text{SNR}^{-1}\|^2_\infty + KLn\sqrt{\log(n)}\|\text{SNR}^{-1}\|_\infty\alpha_{\max} + nL\alpha_{\max};\\ \|\mathcal{R}^{(l)}_{\text{Stage I}}\|_{2,\infty} &\lesssim \alpha_{\max};\\ \|\mathcal{R}^{(l)}_{\text{Stage I}}\| &\lesssim \sqrt{n}\alpha_{\max}\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}\|\mathcal{R}_{\text{all}}\|_{2,\infty} &\lesssim L\max_l\|\mathcal{L}(\mathbf{E}^{(l)})\|_{2,\infty}\|\mathcal{L}(\mathbf{E}^{(l)})\| + L\max_l\|\mathcal{L}(\mathbf{E}^{(l)})\|_{2,\infty}\|\mathcal{R}^{(l)}_Y\| \\ &\quad + L\max_l\|\mathcal{R}^{(l)}_{\text{Stage I}}\|_{2,\infty}\|\mathcal{L}(\mathbf{E}^{(l)})\| + L\max_l\|\mathcal{R}^{(l)}_{\text{Stage I}}\|_{2,\infty}\|\mathcal{R}^{(l)}_{\text{Stage I}}\| \\ &\quad + L\max_l\|\mathcal{R}^{(l)}_{\text{Stage I}}\|_{2,\infty}\|\mathbf{Y}^{(l)}\| + L\max_l\|\mathbf{Y}^{(l)}\|_{2,\infty}\|\mathcal{R}^{(l)}_{\text{Stage I}}\| \\ &\lesssim L\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\frac{K\sqrt{\log(n)}}{(\lambda^{(l)}_{\min})^{1/2}\|\theta^{(l)}\|}\frac{K\sqrt{n\log(n)}}{(\lambda^{(l)}_{\min})^{1/2}\left\|\theta^{(m)}\right\|}\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2} \\ &\quad + L\max_l\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\frac{K\sqrt{\log(n)}}{(\lambda^{(l)}_{\min})^{1/2}\|\theta^{(l)}\|^2}\sqrt{n}\alpha_{\max} \\ &\quad + L\alpha_{\max}\max_l\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\frac{K\sqrt{n\log(n)}}{(\lambda^{(l)}_{\min})^{1/2}\|\theta^{(l)}\|} + L\sqrt{n}\alpha^2_{\max} + \sqrt{n}L\alpha_{\max} \\ &\asymp L\sqrt{n}\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)\frac{K^2\log(n)}{\lambda^{(l)}_{\min}\|\theta^{(l)}\|^2} + L\sqrt{n}\alpha_{\max},\end{aligned}$$

where we have used the assumption that

$$\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2} \frac{K\sqrt{\log(n)}}{(\lambda^{(l)}_{\min})^{1/2}\|\theta\|} \lesssim 1. \tag{F.33}$$

We will verify this momentarily. Plugging this into (F.32), we obtain that

$$\begin{aligned}
&\|(\mathbf{I}-\mathbf{U}\mathbf{U}^\top)\mathcal{R}_{\text{all}}\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\|_{2,\infty} \\
&\quad\lesssim K\frac{\|\mathcal{R}_{\text{all}}\|_{2,\infty}}{nL\bar{\lambda}} + \frac{K^{3/2}}{n^{3/2}L\bar{\lambda}}\|\mathcal{R}_{\text{all}}\| \\
&\quad\lesssim \frac{K}{nL\bar{\lambda}}\left(L\sqrt{n}\max_l\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)\frac{K^2\log(n)}{\lambda^{(l)}_{\min}\|\theta^{(l)}\|^2} + L\sqrt{n}\alpha_{\max}\right) \\
&\qquad + \frac{K^{3/2}}{n^{3/2}L\bar{\lambda}}\left\{LK^2n\log(n)\|\text{SNR}^{-1}\|_\infty^2 + KLn\sqrt{\log(n)}\|\text{SNR}^{-1}\|_\infty\alpha_{\max} + nL\alpha_{\max}\right\} \\
&\quad\asymp \frac{K^{3/2}\alpha_{\max}}{\sqrt{n\bar{\lambda}}} + \frac{K^{7/2}\log(n)\|\text{SNR}^{-1}\|_\infty^2}{\sqrt{n\bar{\lambda}}} + \frac{K^{5/2}\sqrt{\log(n)}\alpha_{\max}\|\text{SNR}^{-1}\|_\infty}{\sqrt{n\bar{\lambda}}},
\end{aligned}$$

since

$$\|\text{SNR}^{-1}\|_\infty = \max_l\left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\frac{1}{(\lambda^{(l)}_{\min})^{1/2}\|\theta^{(l)}\|}.$$

which holds with probability at least $1-O(n^{-15})$.

We now verify (F.33). By Assumption 6.2, the definition of SNR, and the fact that $\frac{\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|^2} \geq 1$, it holds that $\bar{\lambda} \geq K^5\log(n)\|\text{SNR}^{-1}\|_\infty^2$, which in particular implies that $K^2\log(n)\|\text{SNR}\|_\infty^2 \leq 1$ since $\bar{\lambda} \leq 1$. This verifies (F.33).

Therefore, we will have completed the proof provided we can establish the bound (F.31).

Observe that

$$\begin{aligned}
\mathbf{U}^\top \mathcal{L}(\mathcal{E})\mathcal{Y}^\top &= \sum_l \mathbf{U}^\top \left[\sum_{i\le j} \mathbf{E}_{ij}^{(l)} e_i e_j^\top\right] \big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\big)(\mathbf{Y}^{(l)})^\top \\
&\quad + \sum_l \mathbf{U}^\top \left[\sum_{j<i} \mathbf{E}_{ij}^{(l)} e_j e_i^\top\right] \big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{j\cdot})\big)(\mathbf{Y}^{(l)})^\top \\
&= \sum_l \sum_{i\le j} \mathbf{E}_{ij}^{(l)} \mathbf{U}^\top e_i e_j^\top \big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\big)(\mathbf{Y}^{(l)})^\top \\
&\quad + \sum_l \sum_{j<i} \mathbf{E}_{ij}^{(l)} \mathbf{U}^\top e_j e_i^\top \big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{j\cdot})\big)(\mathbf{Y}^{(l)})^\top,
\end{aligned}$$

both of which are a sum of independent random matrices. We bound the first term now; the second is similar. We will apply Matrix Bernstein (Corollary 3.3 of Chen et al. (2021c)). To wit, we need to bound

$$\begin{aligned}
v &:= \sum_l \sum_{i\le j} \mathbb{E}\big(\mathbf{E}_{ij}^{(l)}\big)^2 \|\mathbf{U}^\top e_i e_j^\top \big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\big)(\mathbf{Y}^{(l)})^\top\|^2; \\
w &:= \max_{m,i,j} \|\mathbf{U}^\top e_i e_j^\top \big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\big)(\mathbf{Y}^{(l)})^\top\|.
\end{aligned}$$

We observe that

$$\begin{aligned}
v &\le \sum_l \sum_{i\le j} \theta_i^{(l)}\theta_j^{(l)} \|\mathbf{U}^\top e_i e_j^\top \big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\big)(\mathbf{Y}^{(l)})^\top\|^2 \\
&\le \sum_l \sum_{i\le j} \theta_i^{(l)}\theta_j^{(l)} \|\mathbf{U}\|_{2,\infty}^2 \|e_j^\top \big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})\big)(\mathbf{Y}^{(l)})^\top\|^2 \\
&\lesssim \frac{K}{n} \sum_l \sum_{i\le j} \theta_i^{(l)}\theta_j^{(l)} \|e_j^\top \mathbf{U}^{(l)}\|^2 \||\Lambda^{(l)}|^{-1/2}\|^2 \|\mathbf{J}(\mathbf{X}_{i\cdot})\|^2 \|\mathbf{Y}^{(l)}\|^2 \\
&\lesssim K^3 \sum_l \sum_{i\le j} \theta_i^{(l)}\theta_j^{(l)} \frac{(\theta_j^{(l)})^2}{\|\theta^{(l)}\|^2} \frac{1}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^2} \frac{1}{(\theta_i^{(l)})^2} \\
&\lesssim K^3 \sum_l \frac{1}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^4} \sum_{i,j} \frac{\theta_j^{(l)}}{\theta_i^{(l)}} (\theta_j^{(l)})^2 \\
&\lesssim K^3 n \sum_l \frac{1}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^2} \left(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\right) \\
&= K^3 n L\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big),
\end{aligned}$$

where we have implicitly used Lemma 7. In addition, via similar arguments,

$$w \lesssim K^{3/2} \max_l \left(\frac{\theta_{\max}}{\theta_{\min}}\right) \frac{1}{\|\theta^{(l)}\|^2 (\lambda_{\min}^{(l)})^{1/2}}.$$

Therefore, the result is completed by applying Matrix Bernstein. This completes the proof.

□

The following result bounds several additional "approximate commutation" terms, analogous to Lemma 60 for Stage 1.

Lemma 68 (Second Stage Approximate Commutation)**.** *The following bounds hold with probability at least* $1 - O(n^{-10})$:

$$\begin{aligned}
\|\mathbf{U}^\top \widehat{\mathbf{U}} - \mathbf{W}_*\| &\lesssim \left(K^2 \sqrt{\log(n)} \frac{\left(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\right)^{1/2}}{\sqrt{L}\bar{\lambda}} + K^3 \log(n) \frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} \right. \\
&\qquad \left. + K^2 \sqrt{\log(n)} \frac{\alpha_{\max} \|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K \alpha_{\max}}{\bar{\lambda}} \right)^2; \\
\|\Sigma^{-2} \mathbf{U}^\top \widehat{\mathbf{U}} - \mathbf{U}^\top \widehat{\mathbf{U}} \widehat{\Sigma}^{-2}\| &\lesssim \frac{K^3 \sqrt{\log(n)} \left(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\right)^{1/2}}{nL^{3/2}\bar{\lambda}^2} + \frac{K^4 \log(n) \|\mathrm{SNR}^{-1}\|_\infty^2}{nL\bar{\lambda}^2} \\
&\qquad + \frac{K^3 \sqrt{\log(n)} \alpha_{\max} \|\mathrm{SNR}^{-1}\|_\infty}{nL\bar{\lambda}^2} + \frac{K^2 \alpha_{\max}}{nL\bar{\lambda}^2}, ; \\
\| \sum_l \mathcal{L}(\mathbf{E}^{(l)}) (\mathbf{Y}^{(l)})^\top \|_{2,\infty} &\lesssim K \sqrt{Ln \log(n)} \left(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\right)^{1/2};
\end{aligned}$$

Proof. For the first bound, we observe that

$$\begin{aligned}
\|\mathbf{U}^\top \widehat{\mathbf{U}} - \mathbf{W}_*\| &\lesssim \|\sin \boldsymbol{\Theta}(\widehat{\mathbf{U}}, \mathbf{U})\|^2 \\
&\lesssim \left(K^2 \sqrt{\log(n)} \frac{\left(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\right)^{1/2}}{\sqrt{L}\bar{\lambda}} + K^3 \log(n) \frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} \right. \\
&\qquad \left. + K^2 \sqrt{\log(n)} \frac{\alpha_{\max} \|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K \alpha_{\max}}{\bar{\lambda}} \right)^2,
\end{aligned}$$

where the final inequality holds by Theorem 20, with probability at least $1 - O(n^{-10})$.

For the second bound, we observe that

$$
\begin{aligned}
&\|\Sigma^{-2}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\| \\
&\quad= \|\Sigma^{-2}\big(\mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^2 - \Sigma^2\mathbf{U}^\top\widehat{\mathbf{U}}\big)\widehat{\Sigma}^{-2}\| \\
&\quad\lesssim \frac{K^2}{n^2L^2\bar{\lambda}^2}\|\mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^2 - \Sigma^2\mathbf{U}^\top\widehat{\mathbf{U}}\| \\
&\quad\lesssim \frac{K^2}{n^2L^2\bar{\lambda}^2}\|\mathbf{U}^\top(\widehat{\mathcal{Y}}\widehat{\mathcal{Y}}^\top - \mathcal{Y}\mathcal{Y}^\top)\widehat{\mathbf{U}}\| \\
&\quad\lesssim \frac{K^2}{n^2L^2\bar{\lambda}^2}\Big\{\|\mathcal{L}(\mathcal{E})\mathcal{Y}^\top\| + \|\mathcal{R}_{\text{all}}\|\Big\} \\
&\quad\lesssim \frac{K^2}{n^2L^2\bar{\lambda}^2}\Big\{Kn\sqrt{L\log(n)}\big(\frac{1}{L}\|\text{SNR}^{-1}\|_2^2\big)^{1/2} + LK^2n\log(n)\|\text{SNR}^{-1}\|_\infty^2 \\
&\qquad\qquad + KLn\sqrt{\log(n)}\alpha_{\max}\|\text{SNR}^{-1}\|_\infty + nL\alpha_{\max}\Big\} \\
&\quad\asymp \frac{K^3\sqrt{\log(n)}\big(\frac{1}{L}\|\text{SNR}^{-1}\|_2^2\big)^{1/2}}{nL^{3/2}\bar{\lambda}^2} + \frac{K^4\log(n)\|\text{SNR}^{-1}\|_\infty^2}{nL\bar{\lambda}^2} \\
&\qquad + \frac{K^3\sqrt{\log(n)}\alpha_{\max}\|\text{SNR}^{-1}\|_\infty}{nL\bar{\lambda}^2} + \frac{K^2\alpha_{\max}}{nL\bar{\lambda}^2},
\end{aligned}
$$

which holds with probability at least $1 - O(n^{-10})$ by Lemma 65 and Lemma 66.

For the third term, we note that we can write the i'th row of the matrix in question via

$$
\Big(\sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\Big)_{i\cdot} = \sum_l\sum_j \mathbf{E}_{ij}^{(l)}\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})(\mathbf{Y}^{(l)})^\top\Big)_{j\cdot},
$$

which is a sum of independent random matrices. To wit, we bound via the Matrix Bernstein inequality (Corollary 3.3 of Chen et al. (2021c)). The proof is similar to Lemma 65 (amongst others), so we omit the detailed proof for brevity. Matrix Bernstein then implies that with probability at least $1 - O(n^{-11})$ that

$$
\begin{aligned}
&\Big\|\sum_l\sum_j \mathbf{E}_{ij}^{(l)}\Big(\mathbf{U}^{(l)}|\Lambda^{(l)}|^{-1/2}\mathbf{I}_{p,q}^{(l)}\mathbf{J}(\mathbf{X}_{i\cdot})(\mathbf{Y}^{(l)})^\top\Big)_{j\cdot}\Big\| \\
&\qquad\lesssim K\sqrt{\log(n)}\max_l\|(\mathbf{Y}^{(l)})^\top\|\Big(\sum_l\Big(\frac{\theta_{\max}^{(l)}}{\theta_{\min}^{(l)}}\Big)\frac{1}{\lambda_{\min}^{(l)}\|\theta^{(l)}\|^2}\Big)^{1/2} \\
&\qquad\lesssim K\sqrt{L\log(n)}\big(\frac{1}{L}\|\text{SNR}^{-1}\|_2^2\big)^{1/2}\max_l\|(\mathbf{Y}^{(l)})^\top\| \\
&\qquad\lesssim K\sqrt{Ln\log(n)}\big(\frac{1}{L}\|\text{SNR}^{-1}\|_2^2\big)^{1/2}.
\end{aligned}
$$

Taking a union bound over all n rows completes the proof of this bound. □

F.4.2 Proof of Theorem 21

Proof of Theorem 21. First, recall we have the expansion

$$\begin{aligned}\widehat{\mathcal{Y}}\widehat{\mathcal{Y}}^\top - \mathcal{Y}\mathcal{Y}^\top &= \sum_{l=1}^{L} \widehat{\mathbf{Y}}^{(l)}(\widehat{\mathbf{Y}}^{(l)})^\top - (\mathbf{Y}^{(l)})(\mathbf{Y}^{(l)})^\top \\ &:= \mathcal{L}(\mathcal{E})\mathcal{Y}^\top + \mathcal{Y}\mathcal{L}(\mathcal{E})^\top + \mathcal{R}_{\text{all}},\end{aligned}$$

where recall we define

$$\begin{aligned}\mathcal{R}_{\text{all}} := &\sum_l \mathcal{L}(\mathbf{E}^{(l)})\mathcal{L}(\mathbf{E}^{(l)})^\top + \mathcal{L}(\mathbf{E}^{(l)})(\mathcal{R}^{(l)})^\top + \mathcal{R}^{(m)})\mathcal{L}(\mathbf{E}^{(l)})^\top + \mathcal{R}^{(l)}(\mathcal{R}^{(l)})^\top \\ &+ \sum_l \mathcal{R}^{(l)}(\mathbf{Y}^{(l)})^\top + \mathbf{Y}^{(l)}(\mathcal{R}^{(l)})^\top,\end{aligned}$$

and

$$\mathcal{L}(\mathcal{E}) := \big[\mathcal{L}(\mathbf{E}^{(1)}), \cdots, \mathcal{L}(\mathbf{E}^{(L)})\big],$$

and hence that

$$\mathcal{L}(\mathcal{E})\mathcal{Y}^\top + \mathcal{Y}\mathcal{L}(\mathcal{E})^\top = \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top + \mathbf{Y}^{(l)}\mathcal{L}(\mathbf{E}^{(l)})^\top.$$

We now study how well $\widehat{\mathbf{U}}$ approximates $\mathbf{U}$ in an entrywise sense. We start with the expansion:

$$\begin{aligned}
\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_* &= (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\big(\mathcal{Y}\mathcal{Y}^\top - \widehat{\mathcal{Y}}\widehat{\mathcal{Y}}^\top\big)\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} + \mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*) \\
&= (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\big(\mathcal{L}(\mathcal{E})\mathcal{Y}^\top + \mathcal{Y}\mathcal{L}(\mathcal{E})^\top + \mathcal{R}_{\text{all}}\big)\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} + \mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*) \\
&= \mathcal{L}(\mathcal{E})\mathcal{Y}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} - \mathbf{U}\mathbf{U}^\top\mathcal{L}(\mathcal{E})\mathcal{Y}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} \\
&\quad + (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathcal{Y}\mathcal{L}(\mathcal{E})^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} + (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathcal{R}_{\text{all}}\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} + \mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*) \\
&= \mathcal{L}(\mathcal{E})\mathcal{Y}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} - \mathbf{U}\mathbf{U}^\top\mathcal{L}(\mathcal{E})\mathcal{Y}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} \\
&\quad + (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathcal{R}_{\text{all}}\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} + \mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*),
\end{aligned} \tag{F.34}$$

where we have observed that the term

$$(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathcal{Y}\mathcal{L}(\mathcal{E})^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} \equiv 0,$$

since $\mathcal{Y}$ has left singular vectors $\mathbf{U}$. We now expand the first-order term out further. Observe that

$$\begin{aligned}
\sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} &= \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\mathbf{U}\Sigma^{-2}\mathbf{W}_* + \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\mathbf{U}\Sigma^{-2}\big(\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}\big) \\
&\quad + \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\mathbf{U}\big(\Sigma^{-2}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\big) \\
&\quad + \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\big(\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big)\widehat{\Sigma}^{-2}.
\end{aligned} \tag{F.35}$$

Plugging (F.35) into (F.34) yields the full expansion

$$\begin{aligned}\widehat{\mathbf{U}} - \mathbf{U}\mathbf{W}_* &= \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}\mathbf{W}_* + \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}\big(\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}\big) \\ &\quad + \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\big(\Sigma^{-2}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\big) \\ &\quad + \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \big(\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big)\widehat{\Sigma}^{-2} \\ &\quad - \mathbf{U}\mathbf{U}^\top\mathcal{L}(\mathcal{E})\mathcal{Y}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} \\ &\quad + (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathcal{R}_{\text{all}}\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} + \mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*).\end{aligned}$$

Multiplying through by $\mathbf{W}_*^\top$ yields

$$\begin{aligned}\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U} &= \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2} + \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}\big(\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}\big)\mathbf{W}_*^\top \\ &\quad + \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\big(\Sigma^{-2}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\big)\mathbf{W}_*^\top \\ &\quad + \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \big(\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big)\widehat{\Sigma}^{-2}\mathbf{W}_*^\top \\ &\quad - \mathbf{U}\mathbf{U}^\top\mathcal{L}(\mathcal{E})\mathcal{Y}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\mathbf{W}_*^\top \\ &\quad + (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathcal{R}_{\text{all}}\widehat{\mathbf{U}}\widehat{\Sigma}^{-2} + \mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*)\mathbf{W}_*^\top \\ &:= \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2} + \mathbf{R}_1 + \mathbf{R}_2 + \mathbf{R}_3 + \mathbf{R}_4 + \mathbf{R}_5 + \mathbf{R}_6,\end{aligned}$$

where

$$\begin{aligned}\mathbf{R}_1 &:= \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\Sigma^{-2}\big(\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}\big)\mathbf{W}_*^\top; \\ \mathbf{R}_2 &:= \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \mathbf{U}\big(\Sigma^{-2}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\big)\mathbf{W}_*^\top; \\ \mathbf{R}_3 &:= \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top \big(\widehat{\mathbf{U}} - \mathbf{U}\mathbf{U}^\top\widehat{\mathbf{U}}\big)\widehat{\Sigma}^{-2}\mathbf{W}_*^\top; \\ \mathbf{R}_4 &:= -\mathbf{U}\mathbf{U}^\top\mathcal{L}(\mathcal{E})\mathcal{Y}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\mathbf{W}_*^\top; \\ \mathbf{R}_5 &:= (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathcal{R}_{\text{all}}\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}; \\ \mathbf{R}_6 &:= \mathbf{U}(\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*)\mathbf{W}_*^\top.\end{aligned}$$

By Lemma 67, we have the bounds

$$\|\mathbf{R}_4\|_{2,\infty} \lesssim \frac{K^3\sqrt{\log(n)}}{n\sqrt{L\bar{\lambda}}}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2};$$
$$\|\mathbf{R}_5\|_{2,\infty} \lesssim \frac{K^{3/2}\alpha_{\max}}{\sqrt{n\bar{\lambda}}} + \frac{K^{7/2}\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2}{\sqrt{n\bar{\lambda}}} + \frac{K^{5/2}\sqrt{\log(n)}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\sqrt{n\bar{\lambda}}}.$$

In addition, by properties of the $\ell_{2,\infty}$ norm and Lemma 68, it holds that

$$\begin{aligned}
\|\mathbf{R}_6\|_{2,\infty} &\le \|\mathbf{U}\|_{2,\infty}\|\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*\| \\
&\lesssim \sqrt{\frac{K}{n}}\bigg(K^2\sqrt{\log(n)}\frac{\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{L\bar{\lambda}}} + K^3\log(n)\frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} \\
&\qquad + K^2\sqrt{\log(n)}\frac{\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K\alpha_{\max}}{\bar{\lambda}}\bigg)^2 \\
&\lesssim \frac{K^{9/2}\log(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)}{\sqrt{n}L\bar{\lambda}^2} + \frac{K^{7/2}\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2}{\sqrt{n\bar{\lambda}}} \\
&\quad + \frac{K^{5/2}\sqrt{\log(n)}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\sqrt{n\bar{\lambda}}} + \frac{K^{3/2}\alpha_{\max}^2}{\sqrt{n}\bar{\lambda}^2},
\end{aligned}$$

where we have used the fact that each of the terms inside of the parentheses on the bound for $\mathbf{R}_6$ is less than one, which was verified in the proof of Theorem 20 (note that these terms in parentheses are simply the $\sin\boldsymbol{\Theta}$ upper bound).

Combining these, we obtain that with probability at least $1 - O(n^{-10})$,

$$\begin{aligned}
\|\mathbf{R}_4\|_{2,\infty} + \|\mathbf{R}_5\|_{2,\infty} + \|\mathbf{R}_6\|_{2,\infty} \lesssim{}& \frac{K^3\sqrt{\log(n)}}{n\sqrt{L\bar{\lambda}}}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2} + \frac{K^{7/2}\log(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)}{\sqrt{n}L\bar{\lambda}^2} \\
&+ \frac{K^{7/2}\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2}{\sqrt{n\bar{\lambda}}} + \frac{K^{5/2}\sqrt{\log(n)}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\sqrt{n\bar{\lambda}}} + \frac{\alpha_{\max}}{\sqrt{n\bar{\lambda}}},
\end{aligned}$$

where we have used the fact that $\frac{K^{3/2}\alpha_{\max}}{\bar{\lambda}} \lesssim 1$.

For the terms $\mathbf{R}_1$ through $\mathbf{R}_3$, we observe that

$$\|\mathbf{R}_1\|_{2,\infty} \lesssim \frac{K}{nL\bar{\lambda}}\|\sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\|_{2,\infty}\|\mathbf{W}_* - \mathbf{U}^\top\widehat{\mathbf{U}}\|;$$
$$\|\mathbf{R}_2\|_{2,\infty} \lesssim \|\sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\|_{2,\infty}\|\Sigma^{-2}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\|;$$
$$\|\mathbf{R}_3\|_{2,\infty} \lesssim \frac{K}{nL\bar{\lambda}}\|\sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\|_{2,\infty}\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\|.$$

Lemma 68 shows that with probability at least $1 - O(n^{-10})$ that

$$\|\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{W}_*\| \lesssim \Big(K^2\sqrt{\log(n)}\frac{\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{L\bar{\lambda}}} + K^3\log(n)\frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} + K^2\sqrt{\log(n)}\frac{\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K\alpha_{\max}}{\bar{\lambda}}\Big)^2;$$
$$\|\Sigma^{-2}\mathbf{U}^\top\widehat{\mathbf{U}} - \mathbf{U}^\top\widehat{\mathbf{U}}\widehat{\Sigma}^{-2}\| \lesssim \frac{K^3\sqrt{\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{nL^{3/2}\bar{\lambda}^2} + \frac{K^4\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2}{nL\bar{\lambda}^2} + \frac{K^3\sqrt{\log(n)}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{nL\bar{\lambda}^2} + \frac{K^2\alpha_{\max}}{nL\bar{\lambda}^2};$$
$$\|\sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\|_{2,\infty} \lesssim K\sqrt{Ln\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}.$$

In addition, by Theorem 20, we have that

$$\|\sin\boldsymbol{\Theta}(\widehat{\mathbf{U}},\mathbf{U})\| \lesssim K^2\sqrt{\log(n)}\frac{\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{L\bar{\lambda}}} + K^3\log(n)\frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} + K^2\sqrt{\log(n)}\frac{\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K\alpha_{\max}}{\bar{\lambda}}.$$

Plugging these bounds in yields that

$$
\begin{aligned}
\|\mathbf{R}_1\|_{2,\infty} &\lesssim \frac{K}{nL\bar{\lambda}} K\sqrt{Ln\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2} \\
&\quad\times\Bigg(K^2\sqrt{\log(n)}\frac{\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{L}\bar{\lambda}} + K^3\log(n)\frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} \\
&\qquad + K^2\sqrt{\log(n)}\frac{\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K\alpha_{\max}}{\bar{\lambda}}\Bigg)^2 \\
&\asymp \frac{K^4\log(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)}{L\sqrt{n}\bar{\lambda}^2} + \frac{K^5\log^{3/2}(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}\|\mathrm{SNR}^{-1}\|_\infty^2}{\sqrt{nL}\bar{\lambda}^2} \\
&\quad + \frac{K^4\log(n)\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{nL}\bar{\lambda}^2} + \frac{K^3\sqrt{\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}\alpha_{\max}}{\sqrt{nL}\bar{\lambda}^2}; \\
\|\mathbf{R}_2\|_{2,\infty} &\lesssim K\sqrt{Ln\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2} \\
&\quad\times\Bigg(\frac{K^3\sqrt{\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{nL^{3/2}\bar{\lambda}^2} + \frac{K^4\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2}{nL\bar{\lambda}^2} \\
&\qquad + \frac{K^3\sqrt{\log(n)}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{nL\bar{\lambda}^2} + \frac{K^2\alpha_{\max}}{nL\bar{\lambda}^2}\Bigg) \\
&\asymp \frac{K^4\log(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)}{\sqrt{n}L\bar{\lambda}^2} + \frac{K^5\log^{3/2}(n)\|\mathrm{SNR}^{-1}\|_\infty^2\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{nL}\bar{\lambda}^2} \\
&\quad + \frac{K^4\log(n)\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{nL}\bar{\lambda}^2} + \frac{K^3\sqrt{\log(n)}\alpha_{\max}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{nL}\bar{\lambda}^2}; \\
\|\mathbf{R}_3\|_{2,\infty} &\lesssim \frac{K}{nL\bar{\lambda}} K\sqrt{Ln\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2} \\
&\quad\times\Bigg(K^2\sqrt{\log(n)}\frac{\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{L}\bar{\lambda}} + K^3\log(n)\frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} \\
&\qquad + K^2\sqrt{\log(n)}\frac{\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K\alpha_{\max}}{\bar{\lambda}}\Bigg) \\
&\asymp \frac{K^4\log(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)}{L\sqrt{n}\bar{\lambda}^2} + \frac{K^5\log^{3/2}(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}\|\mathrm{SNR}^{-1}\|_\infty^2}{\sqrt{nL}\bar{\lambda}^2} \\
&\quad + \frac{K^4\log(n)\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{nL}\bar{\lambda}^2} + \frac{K^3\sqrt{\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}\alpha_{\max}}{\sqrt{nL}\bar{\lambda}^2}.
\end{aligned}
$$

We note that we have used the fact that

$$
K^2\sqrt{\log(n)}\frac{\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{L}\bar{\lambda}} + K^3\log(n)\frac{\|\mathrm{SNR}^{-1}\|_\infty^2}{\bar{\lambda}} + K^2\sqrt{\log(n)}\frac{\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\bar{\lambda}} + \frac{K\alpha_{\max}}{\bar{\lambda}} \lesssim 1,
$$

as was verified in the proof of Theorem 20 (observe that this term matches the $\sin\boldsymbol{\Theta}$ upper

bound, and hence is less than one by assumption). Consequently, since each term is the same, we obtain

$$\begin{aligned}
&\|\mathbf{R}_1\|_{2,\infty}+\|\mathbf{R}_2\|_{2,\infty} + \|\mathbf{R}_3\|_{2,\infty} \\
&\quad\lesssim \frac{K^4\log(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)}{L\sqrt{n}\bar{\lambda}^2} + \frac{K^5\log^{3/2}(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}\|\mathrm{SNR}^{-1}\|_\infty^2}{\sqrt{nL}\bar{\lambda}^2} \\
&\qquad+ \frac{K^4\log(n)\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{nL}\bar{\lambda}^2} + \frac{K^3\sqrt{\log(n)}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}\alpha_{\max}}{\sqrt{nL}\bar{\lambda}^2} \\
&\quad\asymp \frac{K^4\log(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)}{L\sqrt{n}\bar{\lambda}^2} + \frac{K^5\log^{3/2}(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}\|\mathrm{SNR}^{-1}\|_\infty^2}{\sqrt{nL}\bar{\lambda}^2} \\
&\qquad+ \frac{K^3\sqrt{\log(n)}\alpha_{\max}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{nL}\bar{\lambda}^2},
\end{aligned}$$

where we have used the assumption that that $K\sqrt{\log(n)}\|\mathrm{SNR}^{-1}\|_\infty \lesssim 1$, which follows immediately the fact that $\frac{\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|^2\lambda^{(l)}_{\min}} \geq 1$ and from Assumption 6.2, which requires that $K^8\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2\frac{\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|^2\lambda^{(l)}_{\min}} \lesssim \bar{\lambda}$. Therefore, we have shown that

$$\widehat{\mathbf{U}}\mathbf{W}_*^\top - \mathbf{U} = \sum_l \mathcal{L}(\mathbf{E}^{(l)})(\mathbf{Y}^{(l)})^\top\mathbf{U}\Sigma^{-2} + \mathcal{R}_{\text{Stage II}},$$

with

$$\begin{aligned}
\|\mathcal{R}_{\text{Stage II}}\|_{2,\infty} &\lesssim \frac{K^3\sqrt{\log(n)}}{n\sqrt{L}\bar{\lambda}}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2} + \frac{K^{7/2}\log(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)}{\sqrt{n}L\bar{\lambda}^2} \\
&\quad + \frac{K^{7/2}\log(n)\|\mathrm{SNR}^{-1}\|_\infty^2}{\sqrt{n}\bar{\lambda}} + \frac{K^{5/2}\sqrt{\log(n)}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty}{\sqrt{n}\bar{\lambda}} + \frac{\alpha_{\max}}{\sqrt{n}\bar{\lambda}} \\
&\quad + \frac{K^4\log(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)}{L\sqrt{n}\bar{\lambda}^2} + \frac{K^5\log^{3/2}(n)\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}\|\mathrm{SNR}^{-1}\|_\infty^2}{\sqrt{nL}\bar{\lambda}^2} \\
&\quad + \frac{K^3\sqrt{\log(n)}\alpha_{\max}\big(\frac{1}{L}\|\mathrm{SNR}^{-1}\|_2^2\big)^{1/2}}{\sqrt{nL}\bar{\lambda}^2} \\
&\asymp \frac{K^3\sqrt{\log(n)}}{nL\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_2 + \frac{K^4\log(n)}{L^2\sqrt{n}\bar{\lambda}^2}\|\mathrm{SNR}^{-1}\|_2^2 + \frac{K^{7/2}\log(n)}{\sqrt{n}\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_\infty^2 \\
&\quad + \frac{K^{5/2}\sqrt{\log(n)}}{\sqrt{n}\bar{\lambda}}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_\infty + \frac{\alpha_{\max}}{\sqrt{n}\bar{\lambda}} \\
&\quad + \frac{K^5\log^{3/2}(n)}{L\sqrt{n}\bar{\lambda}^2}\|\mathrm{SNR}^{-1}\|_2\|\mathrm{SNR}^{-1}\|_\infty^2 + \frac{K^3\sqrt{\log(n)}}{L\sqrt{n}\bar{\lambda}^2}\alpha_{\max}\|\mathrm{SNR}^{-1}\|_2^2 \\
&= \frac{K^3\sqrt{\log(n)}}{nL\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_2 + \frac{K^4\log(n)}{L^2\sqrt{n}\bar{\lambda}^2}\|\mathrm{SNR}^{-1}\|_2^2 \\
&\quad + \frac{K^{7/2}\log(n)}{\sqrt{n}\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_\infty^2\bigg(1 + \frac{K^{3/2}\sqrt{\log(n)}}{L\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_2\bigg) \\
&\quad + \frac{\alpha_{\max}}{\sqrt{n}\bar{\lambda}}\bigg(1 + K^{5/2}\sqrt{\log(n)}\|\mathrm{SNR}^{-1}\|_\infty + \frac{K^3\sqrt{\log(n)}}{L\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_2^2\bigg) \\
&\lesssim \frac{K^3\sqrt{\log(n)}}{nL\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_2 + \frac{K^4\log(n)}{L^2\sqrt{n}\bar{\lambda}^2}\|\mathrm{SNR}^{-1}\|_2^2 \\
&\quad + \frac{K^{7/2}\log(n)}{\sqrt{n}\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_\infty^2 + \frac{\alpha_{\max}}{\sqrt{n}\bar{\lambda}},
\end{aligned}$$

where the final inequality holds as long as

$$\frac{K^{3/2}\sqrt{\log(n)}}{L\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_2 \lesssim 1; \tag{F.36}$$

$$K^{5/2}\|\mathrm{SNR}^{-1}\|_\infty \lesssim 1 \tag{F.37}$$

$$\frac{K^3\sqrt{\log(n)}}{L\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_2^2 \lesssim 1. \tag{F.38}$$

We will verify these bounds now. First, Assumption 6.2 implies that

$$\frac{K^8\log(n)\theta_{\max}^{(l)}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|_2^2}(\mathrm{SNR}_l^{-1})^2 \lesssim \bar{\lambda}\lambda_{\min}^{(l)},$$

as long as C in the assumption is sufficiently large. Observe that this immediately implies equation (F.37) since $\frac{\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\|\theta^{(l)}\|_2^2} \geq 1$ and $\lambda^{(l)}_{\min} \in (0,1)$ by assumption. For the other two terms, by averaging this condition over l, we see that Assumption 6.2 implies

$$\frac{K^8 \log(n)}{L}\|\mathrm{SNR}^{-1}\|_2^2 \lesssim \bar{\lambda}^2.$$

This implies (F.36) and (F.38).

Hence, we have shown so far that

$$\|\mathcal{R}_{\text{Stage II}}\|_{2,\infty} \lesssim \frac{K^3\sqrt{\log(n)}}{nL\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_2 + \frac{K^4\log(n)}{L^2\sqrt{n}\bar{\lambda}^2}\|\mathrm{SNR}^{-1}\|_2^2 + \frac{K^{7/2}\log(n)}{\sqrt{n}\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_\infty^2 + \frac{\alpha_{\max}}{\sqrt{n}\bar{\lambda}}.$$

This holds cumulatively with probability at least $1 - O(n^{-10})$. We now verify that the sum of these terms is less than $\frac{1}{16\sqrt{n_{\max}}}$. Since $n_{\max} \leq n$, it suffices to show that this upper bound is at most $\frac{1}{16\sqrt{n}}$. By pulling out a factor of $1/\sqrt{n}$ it suffices to show that

$$\frac{K^3\sqrt{\log(n)}}{\sqrt{n}L\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_2 + \frac{K^4\log(n)}{L^2\bar{\lambda}^2}\|\mathrm{SNR}^{-1}\|_2^2 + \frac{K^{7/2}\log(n)}{\bar{\lambda}}\|\mathrm{SNR}^{-1}\|_\infty^2 + \frac{\alpha_{\max}}{\bar{\lambda}} \lesssim 1.$$

By similar manipulations as in verifying the bounds (F.36), (F.37), and (F.38), it is straightforward to check the condition above holds, except for the condition $\frac{\alpha_{\max}}{\bar{\lambda}} \lesssim 1$. Plugging in the definition for $\alpha_{\max}$, we see that we require

$$\frac{1}{\bar{\lambda}}\max_l \frac{K^2\theta^{(l)}_{\max}\|\theta^{(l)}\|_1}{\lambda^{(l)}_{\min}\|\theta^{(l)}\|^4}\left(\log(n)\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}} + \frac{1}{\lambda^{(l)}_{\min}} + \left(\frac{\theta^{(l)}_{\max}}{\theta^{(l)}_{\min}}\right)^{1/2}\frac{K^{5/2}\log(n)}{(\lambda^{(l)}_{\min})^{1/2}}\right) \lesssim 1.$$

This is covered by Assumption 6.2. Therefore, this completes the proof. □

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017. 150, 152

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics*, 48(3): 1452–1474, June 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1854. 15, 31, 32, 60, 82, 94, 107, 110, 150, 152, 153, 194, 270, 286, 465

Emmanuel Abbe, Jianqing Fan, and Kaizheng Wang. An ℓ_p theory of PCA and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385, August 2022. ISSN 0090-5364, 2168-8966. doi: 10.1214/22-AOS2196. 5, 15, 31, 32, 33, 35, 44, 60, 82, 95, 107, 153, 270

Joshua Agterberg and Jeremias Sulam. Entrywise Recovery Guarantees for Sparse PCA via Sparsistent Algorithms. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 6591–6629. PMLR, May 2022. ISSN: 2640-3498. ii, 82, 94, 153, 270, 360, 363, 376

Joshua Agterberg and Anru Zhang. Estimating higher-order mixed memberships via $\ell_{2,\infty}$ tensor perturbation bounds. *In preparation*, 2022. ii

Joshua Agterberg, Minh Tang, and Carey Priebe. Nonparametric Two-Sample Hypothesis Testing for Random Graphs with Negative and Repeated Eigenvalues. *arXiv:2012.09828 [math, stat]*, December 2020a. ii, 31, 531

Joshua Agterberg, Minh Tang, and Carey E. Priebe. On Two Distinct Sources of Nonidenti-

fiability in Latent Position Random Graph Models. *arXiv:2003.14250 [math, stat]*, March 2020b. 128, 391, 393, 394, 422, 426

Joshua Agterberg, Zachary Lubberts, and Jesús Arroyo. Joint Spectral Clustering in Multilayer Degree-Corrected Stochastic Blockmodels, December 2022a. arXiv:2212.05053 [math, stat]. ii, 101

Joshua Agterberg, Zachary Lubberts, and Carey E. Priebe. Entrywise Estimation of Singular Vectors of Low-Rank Matrices With Heteroskedasticity and Dependence. *IEEE Transactions on Information Theory*, 68(7):4618–4650, July 2022b. ISSN 1557-9654. doi: 10.1109/TIT.2022.3159085. ii, 60, 82, 94, 96, 97, 153

Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9(65):1981–2014, 2008. ISSN 1533-7928. 81, 152

David Alvarez-Melis, Stefanie Jegelka, and Tommi S. Jaakkola. Towards Optimal Transport with Global Invariances. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1870–1879. PMLR, April 2019. 136, 141

Arash A. Amini and Zahra S. Razaee. Concentration of kernel matrices with application to kernel spectral clustering. *The Annals of Statistics*, 49(1):531–556, February 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS1967. 5, 31, 32, 52, 230, 231, 233

Arash A. Amini and Martin J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. *Annals of Statistics*, 37(5B):2877–2921, October 2009. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS664. 60, 63, 65, 72, 278

N. H. Anderson, P. Hall, and D. M. Titterington. Two-Sample Test Statistics for Measuring Discrepancies Between Two Multivariate Probability Density Functions Using Kernel-Based Density Estimates. *Journal of Multivariate Analysis*, 50(1):41–54, July 1994. ISSN 0047-259X. doi: 10.1006/jmva.1994.1033. 116, 117

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, July 2003. ISBN 978-0-471-36091-9. Google-Books-ID: Cmm9QgAACAAJ. 59

Jesús Arroyo, Avanti Athreya, Joshua Cape, Guodong Chen, Carey E. Priebe, and Joshua T. Vogelstein. Inference for Multiple Heterogeneous Networks with a Common Invariant Subspace. *Journal of Machine Learning Research*, 22(142):1–49, 2021. ISSN 1533-7928. 150, 153, 159, 161, 170

Jesús D. Arroyo-Relión, Daniel Kessler, Elizaveta Levina, and Stephan F. Taylor. Network classification with applications to brain connectomics. *Annals of Applied Statistics*, 13 (3):1648–1677, September 2019. ISSN 1932-6157, 1941-7330. doi: 10.1214/19-AOAS1252. 115

Avanti Athreya, Carey E Priebe, Minh Tang, Vince Lyzinski, David J Marchette, and Daniel L Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78(1):1–18, 2016. 162

Avanti Athreya, Donniell E. Fishkind, Minh Tang, Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L. Sussman. Statistical Inference on Random Dot Product Graphs: a Survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018. ISSN 1533-7928. 31, 116, 122, 152, 158

Avanti Athreya, Michael Kane, Bryan Lewis, Zachary Lubberts, Vince Lyzinski, Youngser Park, Carey E. Priebe, and Minh Tang. Numerical tolerance for spectral decompositions of random matrices. *arXiv:1608.00451 [cs, math, stat]*, January 2020. 394, 414

Arnab Auddy and Ming Yuan. On Estimating Rank-One Spiked Tensors in the Presence of Heavy Tailed Errors. *IEEE Transactions on Information Theory*, pages 1–1, 2022a. ISSN 1557-9654. doi: 10.1109/TIT.2022.3191883. 82

Arnab Auddy and Ming Yuan. Perturbation Bounds for (Nearly) Orthogonally Decomposable Tensors, January 2022b. arXiv:2007.09024 [cs, math, stat]. 81

Afonso S. Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of

random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, July 2016. ISSN 0091-1798, 2168-894X. doi: 10.1214/15-AOP1025. 12, 110, 447

Jess Banks, Cristopher Moore, Roman Vershynin, Nicolas Verzelen, and Jiaming Xu. Information-Theoretic Bounds and Phase Transitions in Clustering, Sparse PCA, and Submatrix Localization. *IEEE Transactions on Information Theory*, 64(7):4872–4894, July 2018. ISSN 1557-9654. doi: 10.1109/TIT.2018.2810020. 60

Zhigang Bao, Xiucai Ding, Jingming Wang, and Ke Wang. Statistical inference for principal components of spiked covariance matrices. *arXiv:2008.11903 [math, stat]*, September 2020. 60, 66

Zhigang Bao, Xiucai Ding, and and Ke Wang. Singular vector and singular subspace distribution for the matrix denoising model. *The Annals of Statistics*, 49(1), February 2021. ISSN 0090-5364. doi: 10.1214/20-AOS1960. 32, 33, 41

Marya Bazzi, Lucas GS Jeub, Alex Arenas, Sam D Howison, and Mason A Porter. A framework for the construction of generative models for mesoscale structure in multilayer networks. *Physical Review Research*, 2(2):023100, 2020. 150, 153, 155

Andrew C. Berry. The Accuracy of the Gaussian Approximation to the Sum of Independent Variates. *Transactions of the American Mathematical Society*, 49(1):122–136, 1941. ISSN 0002-9947. doi: 10.2307/1990053. 56, 219

Quentin Berthet and Philippe Rigollet. Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 41(4):1780–1815, August 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1127. 60

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, 1997. ISBN 0-387-94846-5. 10, 11, 32, 34, 61, 68, 74, 269, 286, 294

Sharmodeep Bhattacharyya and Shirshendu Chatterjee. Spectral Clustering for Multiple Sparse Networks: I. *arXiv:1805.10594 [cs, math, stat]*, May 2018. 153

Sharmodeep Bhattacharyya and Shirshendu Chatterjee. Consistent recovery of communities from sparse multi-relational networks: A scalable algorithm with optimal recovery conditions. In *Complex Networks XI*, pages 92–103. Springer, 2020. 153, 155, 170

Peter J. Bickel and Purnamrita Sarkar. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):253–273, 2016. ISSN 1467-9868. doi: 10.1111/rssb.12117. 116

Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. *arXiv:1711.08947 [math, stat]*, November 2019. 141

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, July 2003. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1055425791. 415

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience*, 10(3):186–198, March 2009. ISSN 1471-0048. doi: 10.1038/nrn2575. 115

Edward T. Bullmore and Danielle S. Bassett. Brain Graphs: Graphical Models of the Human Brain Connectome. *Annual Review of Clinical Psychology*, 7(1):113–140, 2011. doi: 10.1146/annurev-clinpsy-040510-143934. 115

Bureau of Transportation Statistics. Air Carrier Statistics (Form 41 Traffic)- All Carriers. available at https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EEE&DB_URL=, 2022. 173

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-20192-9. 61

Changxiao Cai, Gen Li, Yuejie Chi, H. Vincent Poor, and Yuxin Chen. Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *The Annals of Statistics*, 49(2):944–967, April 2021a. ISSN 0090-5364, 2168-8966. doi:

10.1214/20-AOS1986. 15, 31, 32, 33, 35, 43, 44, 60, 70, 82, 83, 94, 95, 110, 153, 154, 270, 313, 363, 373, 387

Changxiao Cai, Gen Li, H. Vincent Poor, and Yuxin Chen. Nonconvex Low-Rank Tensor Completion from Noisy Data. *Operations Research*, 70(2):1219–1237, March 2022. ISSN 0030-364X. doi: 10.1287/opre.2021.2106. 81, 107

T. Tony Cai and Xiaodong Li. Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059, June 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/14-AOS1290. 116

T. Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Annals of Statistics*, 46(1):60–89, February 2018. ISSN 0090-5364, 2168-8966. doi: 10.1214/17-AOS1541. 10, 11, 13, 31, 37, 71, 312, 335

T. Tony Cai, Zongming Ma, and Yihong Wu. Sparse PCA: Optimal rates and adaptive estimation. *Annals of Statistics*, 41(6):3074–3110, December 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1178. 60, 70

T. Tony Cai, Xiao Han, and Guangming Pan. Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. *Annals of Statistics*, 48 (3):1255–1280, June 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/18-AOS1798. 60

T. Tony Cai, Hongzhe Li, and Rong Ma. Optimal Structured Principal Subspace Estimation: Metric Entropy and Minimax Rates. *Journal of machine learning research*, 22, January 2021b. 31

E. J. Candes and Y. Plan. Matrix Completion With Noise. *Proceedings of the IEEE*, 98(6): 925–936, June 2010. ISSN 0018-9219. doi: 10.1109/JPROC.2009.2035722. 67

Emmanuel J. Candes and Terence Tao. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, May 2010. ISSN 1557-9654. doi: 10.1109/TIT.2010.2044061. 67

Joshua Cape. Orthogonal Procrustes and norm-dependent optimality. *The Electronic Journal of Linear Algebra*, 36(36):158–168, March 2020. ISSN 1081-3810. doi: 10.13001/ela.2020.5009. 34, 35

Joshua Cape, Minh Tang, and Carey E. Priebe. The Kato–Temple inequality and eigenvalue concentration with applications to graph inference. *Electronic Journal of Statistics*, 11(2): 3954–3978, 2017. ISSN 1935-7524. doi: 10.1214/17-EJS1328. 429

Joshua Cape, Minh Tang, and Carey E. Priebe. Signal-plus-noise matrix models: eigenvector deviations and fluctuations. *Biometrika*, 106(1):243–250, March 2019a. ISSN 0006-3444. doi: 10.1093/biomet/asy070. 32, 33, 43, 60, 74, 82, 153, 270, 273, 286, 294

Joshua Cape, Minh Tang, and Carey E. Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *Annals of Statistics*, 47(5):2405–2439, October 2019b. ISSN 0090-5364, 2168-8966. doi: 10.1214/18-AOS1752. 14, 15, 32, 34, 60, 67, 70, 82, 94, 153, 270

P. J. Carrington, J. Scott, and S. Wasserman. *Models and methods in social network analysis*, volume 28. Cambridge university press, 2005. 115

Vasileios Charisopoulos, Austin R Benson, and Anil Damle. Entrywise convergence of iterative methods for eigenproblems. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5644–5655. Curran Associates, Inc., 2020. 60

Fan Chen and Karl Rohe. A New Basis for Sparse PCA. *arXiv:2007.00596 [cs, stat]*, July 2020. 60, 63

Hao Chen and Jerome H. Friedman. A New Graph-Based Two-Sample Test for Multivariate and Object Data. *Journal of the American Statistical Association*, 112(517):397–409, January 2017. ISSN 0162-1459. doi: 10.1080/01621459.2016.1147356. 116

Pinhan Chen, Chao Gao, and Anderson Y. Zhang. Partial recovery for top-k ranking: Optimality of MLE and SubOptimality of the spectral method. *The Annals of Statistics*, 50(3):1618–1652, June 2022. ISSN 0090-5364, 2168-8966. doi: 10.1214/21-AOS2166. 154

Shuxiao Chen, Sifan Liu, and Zongming Ma. Global and Individualized Community Detection in Inhomogeneous Multilayer Networks. *arXiv:2012.00933 [cs, math, stat]*, January 2021a. 153

Xiaohui Chen and Yun Yang. Hanson-Wright inequality in Hilbert spaces with application to K-means clustering for non-Euclidean data. *arXiv:1810.11180 [math, stat]*, July 2020. 256

Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized MLE are both optimal for top-K ranking. *Annals of Statistics*, 47(4):2204–2235, August 2019a. ISSN 0090-5364, 2168-8966. doi: 10.1214/18-AOS1745. 154

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Inference and Uncertainty Quantification for Noisy Matrix Completion. *Proceedings of the National Academy of Sciences*, 116(46):22931–22937, November 2019b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1910053116. 31, 67

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, and Yuling Yan. Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, January 2020. ISSN 1052-6234. doi: 10.1137/19M1290000. 67, 82

Yuxin Chen, Chen Cheng, and Jianqing Fan. Asymmetry helps: Eigenvalue and eigenvector analyses of asymmetrically perturbed low-rank matrices. *The Annals of Statistics*, 49(1): 435–458, February 2021b. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS1963. 32

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral Methods for Data Science: A Statistical Perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021c. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000079. 15, 17, 31, 32, 34, 35, 38, 53, 60, 61, 80, 82, 107, 152, 153, 201, 202, 270, 317, 445, 468, 471, 475, 486, 488

Yuxin Chen, Jianqing Fan, Cong Ma, and Yuling Yan. Bridging convex and nonconvex optimization in robust PCA: Noise, outliers and missing data. *The Annals of Statistics*,

49(5):2948–2971, October 2021d. ISSN 0090-5364, 2168-8966. doi: 10.1214/21-AOS2066. 82, 107

Yuxin Chen, Jianqing Fan, Bingyan Wang, and Yuling Yan. Convex and Nonconvex Optimization Are Both Minimax-Optimal for Noisy Blind Deconvolution Under Random Designs. *Journal of the American Statistical Association*, 0(0):1–11, July 2021e. ISSN 0162-1459. doi: 10.1080/01621459.2021.1956501. 82, 107

Chen Cheng, Yuting Wei, and Yuxin Chen. Tackling Small Eigen-Gaps: Fine-Grained Eigenvector Estimation and Inference Under Heteroscedastic Noise. *IEEE Transactions on Information Theory*, 67(11):7380–7419, November 2021. ISSN 1557-9654. doi: 10.1109/TIT.2021.3111828. 32

Eric C. Chi, Brian R. Gaines, Will Wei Sun, Hua Zhou, and Jian Yang. Provable convex co-clustering of tensors. *The Journal of Machine Learning Research*, 21(1):214:8792–214:8849, January 2020. ISSN 1532-4435. 78, 80

Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, October 2019. ISSN 1941-0476. doi: 10.1109/TSP.2019.2937282. 31, 38, 80, 82

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 5(1), pages 89–96, 2011. 149

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. *arXiv:1306.0895 [stat]*, June 2013. 137

Anil Damle and Yuekai Sun. Uniform Bounds for Invariant Subspace Perturbations. *SIAM Journal on Matrix Analysis and Applications*, 41(3):1208–1236, January 2020. ISSN 0895-4798. doi: 10.1137/19M1262760. 32, 60

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I. Jordan, and Gert R. G. Lanckriet. A Direct Formulation for Sparse PCA Using Semidefinite Programming. *SIAM Review*, 49(3):434–448, January 2007. ISSN 0036-1445. doi: 10.1137/050645506. 60, 63

Caterina De Bacco, Eleanor A Power, Daniel B Larremore, and Cristopher Moore. Community detection, link prediction, and layer interdependence in multilayer networks. *Physical Review E*, 95(4):042317, 2017. 153

Manlio De Domenico, Vincenzo Nicosia, Alexandre Arenas, and Vito Latora. Structural reducibility of multilayer networks. *Nature Communications*, 6(1):6864, April 2015. ISSN 2041-1723. doi: 10.1038/ncomms7864. 99, 104

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000. 87

Lijun Ding and Yudong Chen. Leave-one-out Approach for Matrix Completion: Primal and Dual Analysis. *arXiv:1803.07554 [cs, math, stat]*, June 2020. 82, 107

Xiucai Ding. High dimensional deformed rectangular matrices with applications in matrix denoising. *Bernoulli*, 26(1):387–417, February 2020. ISSN 1350-7265. doi: 10.3150/19-BEJ1129. 32, 33

Xiucai Ding and Qiang Sun. Modified Multidimensional Scaling and High Dimensional Clustering. *arXiv:1810.10172 [cs, math, stat]*, January 2019. 5, 32

Xiucai Ding and Fan Yang. Spiked separable covariance matrices and principal components. *The Annals of Statistics*, 49(2):1113–1138, April 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS1995. 66

Benjamin Draves and Daniel L. Sussman. Bias-Variance Tradeoffs in Joint Spectral Embeddings. *arXiv:2005.02511 [math, stat]*, May 2020. 116, 138, 139

Benjamin Draves and Daniel L. Sussman. Bias-Variance Tradeoffs in Joint Spectral Embeddings, December 2021. arXiv:2005.02511 [math, stat]. 153

Xinjie Du and Minh Tang. Hypothesis Testing for Equality of Latent Positions in Random Graphs. Technical Report arXiv:2105.10838, arXiv, March 2022. arXiv:2105.10838 [stat] type: article. 22, 24, 26, 52, 179, 181

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, February 2010. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS648. 66

Noureddine El Karoui, Derek Bean, Peter J. Bickel, Chinghway Lim, and Bin Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110(36):14557–14562, September 2013. doi: 10.1073/pnas.1307842110. 82

Justin Eldridge, Mikhail Belkin, and Yusu Wang. Unperturbed: spectral analysis beyond Davis-Kahan. In *Algorithmic Learning Theory*, pages 321–358, April 2018. 32, 429

Andreas Elsener and Sara van de Geer. Sparse spectral estimation with missing and corrupted measurements. *Stat*, 8(1):e229, 2019. ISSN 2049-1573. doi: 10.1002/sta4.229. 65

Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, December 2001. ISSN 0162-1459. doi: 10.1198/016214501753382273. 61

Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An ℓ_{∞} Eigenvector Perturbation Bound and Its Application. *Journal of Machine Learning Research*, 18(207):1–42, 2018. ISSN 1533-7928. 32, 60

Jianqing Fan, Yingying Fan, Xiao Han, and Jinchi Lv. Asymptotic Theory of Eigenvectors for Random Matrices With Diverging Spikes. *Journal of the American Statistical Association*, 0(0):1–14, October 2020. ISSN 0162-1459. doi: 10.1080/01621459.2020.1840990. 32, 33, 43, 60

Jianqing Fan, Yingying Fan, Xiao Han, and Jinchi Lv. SIMPLE: Statistical inference on membership profiles in large networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2):630–653, 2022. ISSN 1467-9868. doi: 10.1111/rssb.12505. 22, 24, 26, 31, 52, 116, 153, 179, 181

Emily S. Finn, Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. Functional connectome

fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11):1664–1671, November 2015. ISSN 1546-1726. doi: 10.1038/nn.4135. 115

Laura Florescu and Will Perkins. Spectral thresholds in the bipartite stochastic block model. In *Conference on Learning Theory*, pages 943–959, June 2016. ISSN: 1938-7228 Section: Machine Learning. 33

Santo Fortunato and Mark EJ Newman. 20 years of network community detection. *Nature Physics*, 18(8):848–850, 2022. 152

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, August 2015. ISSN 1432-2064. doi: 10.1007/s00440-014-0583-7. 410

G. W. Stewart and J.-G. Sun. *Matrix pertubation theory*. Academic Press, 1990. 61, 68, 286

Aditya Gangrade, Praveen Venkatesh, Bobak Nazer, and Venkatesh Saligrama. Efficient Near-Optimal Testing of Community Changes in Balanced Stochastic Block Models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 10364–10375. Curran Associates, Inc., 2019. 141

Chao Gao, Zongming Ma, and Harrison H. Zhou. Sparse CCA: Adaptive estimation and computational barriers. *The Annals of Statistics*, 45(5):2074–2101, October 2017. ISSN 0090-5364, 2168-8966. doi: 10.1214/16-AOS1519. 60

Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou. Community detection in degree-corrected block models. *The Annals of Statistics*, 46(5):2153–2185, October 2018. ISSN 0090-5364, 2168-8966. doi: 10.1214/17-AOS1615. 150, 152

Milana Gataric, Tengyao Wang, and Richard J. Samworth. Sparse principal component analysis via axis-aligned random projections. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):329–359, 2020. ISSN 1467-9868. doi: 10.1111/rssb.12360. 60, 63

Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike von Luxburg. Two-Sample Tests for Large Random Graphs Using Network Statistics. *arXiv:1705.06168 [stat]*, May 2017. 138, 139

Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike von Luxburg. Two-sample hypothesis testing for inhomogeneous random graphs. *Annals of Statistics*, 48(4):2208–2229, August 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1884. 116, 139

Nicolas Gillis and Stephen A. Vavasis. Fast and Robust Recursive Algorithmsfor Separable Nonnegative Matrix Factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714, April 2014. ISSN 1939-3539. doi: 10.1109/TPAMI.2013.226. 79, 81, 87, 88, 97, 383, 385

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised Alignment of Embeddings with Wasserstein Procrustes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1880–1890, April 2019. ISSN: 1938-7228 Section: Machine Learning. 141

Robert Everist Greene and Steven George Krantz. *Function Theory of One Complex Variable*. American Mathematical Soc., 2006. ISBN 978-0-8218-3962-1. Google-Books-ID: u5vhseYCcqkC. 73, 289

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-sample Test. *J. Mach. Learn. Res.*, 13:723–773, March 2012. ISSN 1532-4435. 25, 116, 117, 125, 132, 133, 393, 407

Quanquan Gu, Zhaoran Wang, and Han Liu. Sparse PCA with Oracle Property. *Advances in neural information processing systems*, 2014:1529–1537, 2014. ISSN 1049-5258. 60, 63, 65, 66, 69

Qiuyi Han, Kevin Xu, and Edoardo Airoldi. Consistent estimation of dynamic and multi-layer block models. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1511–1520. PMLR, June 2015. ISSN: 1938-7228. 150, 153, 159, 170

Rungang Han, Yuetian Luo, Miaoyan Wang, and Anru R. Zhang. Exact Clustering in Tensor Block Model: Statistical Optimality and Computational Limit. *arXiv:2012.09996 [math, stat]*, March 2021. 78, 80, 90, 91, 99, 100, 101, 178, 389

Xiao Han, Qing Yang, and Yingying Fan. Universal Rank Inference via Residual Subsampling with Application to Large Networks. *arXiv:1912.11583 [math, stat]*, July 2020. 52, 163

Nicholas J. Higham. *Functions of Matrices*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, January 2008. ISBN 978-0-89871-646-7. doi: 10.1137/1.9780898717778. 423

Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent Space Approaches to Social Network Analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002. doi: 10.1198/016214502388618906. 116

P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983. 116, 149, 150

Victoria Hore, Ana Viñuela, Alfonso Buil, Julian Knight, Mark I. McCarthy, Kerrin Small, and Jonathan Marchini. Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics*, 48(9):1094–1100, September 2016. ISSN 1546-1718. doi: 10.1038/ng.3624. 80

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012. 61, 73, 436

Jiaxin Hu and Miaoyan Wang. Multiway Spherical Clustering via Degree-Corrected Tensor Block Models. *arXiv:2201.07401 [math, stat]*, January 2022. 90, 114

Ningyuan Huang, David W. Hogg, and Soledad Villar. Dimensionality reduction, regularization, and generalization in overparameterized regressions. *arXiv:2011.11477 [cs, stat]*, November 2020a. 71

Sihan Huang, Haolei Weng, and Yang Feng. Spectral clustering via adaptive layer aggregation for multi-layer networks. *arXiv:2012.04646 [cs, math, stat]*, December 2020b. 153

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1): 193–218, 1985. 170

Jana Janková and Sara van de Geer. De-Biased Sparse PCA: Inference for Eigenstructure of Large Covariance Matrices. *IEEE Transactions on Information Theory*, 67(4):2507–2527, April 2021. ISSN 1557-9654. doi: 10.1109/TIT.2021.3059765. 26, 63, 65, 72

Jiashun Jin. Fast community detection by SCORE. *The Annals of Statistics*, 43(1), February 2015. ISSN 0090-5364. doi: 10.1214/14-AOS1265. 150, 152, 159, 162

Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Estimating network memberships by simplex vertex hunting. *arXiv:1708.07852 [stat]*, September 2019. 15, 31, 60, 114, 153, 154, 447, 453

Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Improvements on SCORE, Especially for Weak Signals. *Sankhya A*, March 2021. ISSN 0976-836X, 0976-8378. doi: 10.1007/s13171-020-00240-1. 152, 153, 154, 165, 167, 168, 169, 181, 185

Jiashun Jin, Zheng Tracy Ke, Shengming Luo, and Minzhe Wang. Optimal Estimation of the Number of Communities. *arXiv:2009.09177 [math, stat]*, January 2022. 157, 158, 163, 164

Bing-Yi Jing, Ting Li, Zhongyuan Lyu, and Dong Xia. Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics*, 49(6): 3181–3205, December 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/21-AOS2079. 17, 96, 104, 105, 110, 150, 153, 177

Iain M. Johnstone and Arthur Yu Lu. On Consistency and Sparsity for Principal Components Analysis in High Dimensions. *Journal of the American Statistical Association*, 104 (486):682–693, June 2009. ISSN 0162-1459. doi: 10.1198/jasa.2009.0121. 31, 59, 60

Andrew Jones and Patrick Rubin-Delanchy. The multilayer random dot product graph, January 2021. arXiv:2007.10455 [cs, stat]. 127, 153

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011. 150, 152, 155, 156, 173

Tosio Kato. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Springer-Verlag, Berlin Heidelberg, 2 edition, 1995. ISBN 978-3-540-58661-6. doi: 10.1007/978-3-642-66282-9. 12

Zheng Tracy Ke and Jingming Wang. Optimal Network Membership Estimation Under Severe Degree Heterogeneity, April 2022. arXiv:2204.12087 [math, stat]. 153, 154, 178

Zheng Tracy Ke, Feng Shi, and Dong Xia. Community Detection for Hypergraph Networks via Regularized Tensor Power Iteration. *arXiv:1909.06503 [math, stat]*, January 2020. 96, 110

Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. *Journal of complex networks*, 2(3):203–271, 2014. 150

Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1):257–352, October 2017. ISSN 1432-2064. doi: 10.1007/s00440-016-0730-4. 66

Tamara G. Kolda and Brett W. Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3):455–500, August 2009. ISSN 0036-1445. doi: 10.1137/07070111X. 83

Tamara Gibson Kolda. *Multilinear operators for higher-order decompositions*, volume 2. United States. Department of Energy, 2006. 84

Vladimir Koltchinskii and Evarist Gine. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, February 2000. ISSN 1350-7265. 33, 429

Vladimir Koltchinskii and Karim Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(4):1976–2013, November 2016. ISSN 0246-0203. doi: 10.1214/15-AIHP705. 31, 42, 43

Vladimir Koltchinskii and Karim Lounici. Normal approximation and concentration of spectral projectors of sample covariance. *The Annals of Statistics*, 45(1):121–157, February 2017. ISSN 0090-5364, 2168-8966. doi: 10.1214/16-AOS1437. 31, 32

Vladimir Koltchinskii and Dong Xia. Perturbation of linear forms of singular vectors under gaussian noise. *High Dimensional Probability VII, the Cargèse Volume*, 2015. 82

Vladimir Koltchinskii and Dong Xia. Perturbation of Linear Forms of Singular Vectors Under Gaussian Noise. In Christian Houdré, David M. Mason, Patricia Reynaud-Bouret, and Jan Rosiński, editors, *High Dimensional Probability VII*, Progress in Probability, pages 397–423, Cham, 2016. Springer International Publishing. ISBN 978-3-319-40519-3. doi: 10.1007/978-3-319-40519-3_18. 31, 32, 288

Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39 (5):2302–2329, 2011. 388

Vladimir Koltchinskii, Matthias Löffler, and Richard Nickl. Efficient estimation of linear functionals of principal components. *The Annals of Statistics*, 48(1):464–490, February 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1816. 31, 32, 42

Vladimir I. Koltchinskii. Asymptotics of Spectral Projections of Some Random Matrices Approximating Integral Operators. In Ernst Eberlein, Marjorie Hahn, and Michel Talagrand, editors, *High Dimensional Probability*, Progress in Probability, pages 191–227, Basel, 1998. Birkhäuser. ISBN 978-3-0348-8829-5. doi: 10.1007/978-3-0348-8829-5_11. 32

Piotr Koniusz and Anoop Cherian. Sparse Coding for Third-Order Super-Symmetric Tensor Descriptors with Application to Texture Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5395–5403, June 2016. doi: 10.1109/CVPR.2016.582. ISSN: 1063-6919. 80

Robert Krauthgamer, Boaz Nadler, and Dan Vilenchik. Do semidefinite relaxations solve sparse PCA up to the information limit? *Annals of Statistics*, 43(3):1300–1322, June 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1310. 60

Can M Le and Elizaveta Levina. Estimating the number of communities by spectral methods. *Electronic Journal of Statistics*, 16(1):3315–3342, 2022. 163

William Leeb and Elad Romanov. Optimal Spectral Shrinkage and PCA With Heteroscedastic Noise. *IEEE Transactions on Information Theory*, 67(5):3009–3037, May 2021. ISSN 1557-9654. doi: 10.1109/TIT.2021.3055075. 33

Jing Lei. A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, 44(1): 401–424, February 2016. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1370. 116

Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, February 2020a. ISSN 1350-7265. doi: 10.3150/19-BEJ1151. 411

Jing Lei. Network Representation Using Graph Root Distributions. *Annals of Statistics*, 2020b. 121, 122, 135, 140, 411

Jing Lei and Kevin Z. Lin. Bias-Adjusted Spectral Clustering in Multi-Layer Stochastic Block Models. *Journal of the American Statistical Association*, 0(0):1–13, March 2022. ISSN 0162-1459. doi: 10.1080/01621459.2022.2054817. 33, 150, 153, 159, 170, 177

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Annals of Statistics*, 43(1):215–237, February 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/14-AOS1274. 13, 31, 45, 46, 150, 152, 159, 162, 168, 170, 183, 185, 186, 220, 224, 439, 441

Jing Lei and Vincent Q. Vu. Sparsistency and agnostic inference in sparse PCA. *Annals of Statistics*, 43(1):299–322, February 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/14-AOS1273. 60, 63, 64, 65, 66, 69

Jing Lei, Kehui Chen, and Brian Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73, March 2020. ISSN 0006-3444. doi: 10.1093/biomet/asz068. 153, 177

Lihua Lei. Unified $\ell_{2\rightarrow\infty}$ Eigenspace Perturbation Theory for Symmetric Random Matrices. *arXiv:1909.04798 [math, stat]*, September 2019. 31, 32, 60, 150, 286, 288, 289, 291

Keith Levin and Elizaveta Levina. Bootstrapping Networks with Latent Space Structure. *arXiv:1907.10821 [math, stat]*, July 2019. 117, 127, 135, 138, 139, 140, 411

Keith Levin, Avanti Athreya, Minh Tang, Vince Lyzinski, Youngser Park, and Carey E. Priebe. A central limit theorem for an omnibus embedding of multiple random graphs and implications for multiscale network inference. *arXiv:1705.09355 [stat]*, June 2019. 116, 138, 153

Gen Li, Changxiao Cai, Yuantao Gu, H. Vincent Poor, and Yuxin Chen. Minimax Estimation of Linear Functions of Eigenvectors in the Face of Small Eigen-Gaps. *arXiv:2104.03298 [cs, math, stat]*, April 2021. 32

Gongkai Li, Minh Tang, Nicolas Charon, and Carey Priebe. Central limit theorems for classical multidimensional scaling. *Electronic Journal of Statistics*, 14(1):2362–2394, January 2020a. ISSN 1935-7524, 1935-7524. doi: 10.1214/20-EJS1720. 5, 32

Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, 2020b. 163

Yezheng Li and Hongzhe Li. Two-sample Test of Community Memberships of Weighted Stochastic Block Models. *arXiv:1811.12593 [math, stat]*, November 2018. 32, 33, 116, 138, 139, 141, 394

Qiaohui Lin, Robert Lunde, and Purnamrita Sarkar. Higher-Order Correct Multiplier Bootstraps for Count Functionals of Networks. *arXiv:2009.06170 [math, stat]*, September 2020a. 117, 127, 139

Qiaohui Lin, Robert Lunde, and Purnamrita Sarkar. On the Theoretical Properties of the Network Jackknife. *arXiv:2004.08935 [math, stat]*, April 2020b. 117, 127, 139

Anna Little, Yuying Xie, and Qiang Sun. Exact Cluster Recovery via Classical Multidimensional Scaling. *arXiv:1812.11954 [math, stat]*, July 2020. 5, 32

Karim Lounici. Sparse Principal Component Analysis with Missing Observations. In Christian Houdré, David M. Mason, Jan Rosiński, and Jon A. Wellner, editors, *High Dimen-*

sional Probability VI, Progress in Probability, pages 327–356, Basel, 2013. Springer. ISBN 978-3-0348-0490-5. doi: 10.1007/978-3-0348-0490-5_20. 31

Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, August 2014. ISSN 1350-7265. doi: 10.3150/12-BEJ487. 31, 33

László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012. 116

Yue M. Lu and Gen Li. Spectral initialization for nonconvex estimation: High-dimensional limit and phase transitions. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3015–3019, June 2017. doi: 10.1109/ISIT.2017.8007083. ISSN: 2157-8117. 31

Robert Lunde and Purnamrita Sarkar. Subsampling Sparse Graphons Under Minimal Assumptions. *arXiv:1907.12528 [math, stat]*, August 2019. 117, 127, 132, 139

Feng Luo, Yunfeng Yang, Chin-Fu Chen, Roger Chang, Jizhong Zhou, and Richard H Scheuermann. Modular organization of protein interaction networks. *Bioinformatics*, 23(2):207–214, 2007. 149

Yuetian Luo and Anru R. Zhang. Tensor clustering with planted structures: Statistical optimality and computational limits. *The Annals of Statistics*, 50(1):584–613, February 2022. ISSN 0090-5364, 2168-8966. doi: 10.1214/21-AOS2123. 80

Yuetian Luo, Rungang Han, and Anru R. Zhang. A Schatten-q Matrix Perturbation Theory via Perturbation Projection Error Bound. *arXiv:2008.01312 [cs, math, stat]*, November 2020. 32

Yuetian Luo, Garvesh Raskutti, Ming Yuan, and Anru R. Zhang. A Sharp Blockwise Tensor Perturbation Bound for Orthogonal Iteration. *arXiv:2008.02437 [cs, math, stat]*, June 2021. 81, 94, 330

Vince Lyzinski, Daniel L Sussman, Minh Tang, Avanti Athreya, and Carey E Priebe. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic journal of statistics*, 8(2):2905–2922, 2014. 150, 154, 159, 168

Matthias Löffler, Anderson Y. Zhang, and Harrison H. Zhou. Optimality of spectral clustering in the Gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530, October 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS2044. 5, 31, 32, 90, 178

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval, Matrix Completion, and Blind Deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632, June 2020. ISSN 1615-3383. doi: 10.1007/s10208-019-09429-9. 31, 82, 107

Shujie Ma, Liangjun Su, and Yichong Zhang. Determining the number of communities in degree-corrected stochastic block models. *Journal of machine learning research*, 22(69), 2021. 163

Zongming Ma. Sparse principal component analysis and iterative thresholding. *Annals of Statistics*, 41(2):772–801, April 2013. ISSN 0090-5364, 2168-8966. doi: 10.1214/13-AOS1097. 60, 63

P W MacDonald, E Levina, and J Zhu. Latent space models for multiplex networks with shared structure. *Biometrika*, 109(3):683–706, September 2022. ISSN 1464-3510. doi: 10.1093/biomet/asab058. 153

Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Estimating Mixed Memberships With Sharp Eigenvector Deviations. *Journal of the American Statistical Association*, 0(0):1–13, April 2020. ISSN 0162-1459. doi: 10.1080/01621459.2020.1751645. 31, 32, 60, 288

Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. Estimating Mixed Memberships With Sharp Eigenvector Deviations. *Journal of the American Statistical Association*,

116(536):1928–1940, October 2021. ISSN 0162-1459. doi: 10.1080/01621459.2020.1751645. 81, 86, 87, 92, 152, 153, 379, 381, 382, 383

A. M. Mood. On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests. *The Annals of Mathematical Statistics*, 25(3):514–522, 1954. ISSN 0003-4851. 115

M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. 115

M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, February 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.012582999. 115

Majid Noroozi and Marianna Pensky. Sparse Subspace Clustering in Diverse Multiplex Network Model, June 2022. arXiv:2206.07602 [cs, stat]. 153

Sean O'Rourke, Van Vu, and Ke Wang. Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, 540:26–59, March 2018. ISSN 0024-3795. doi: 10.1016/j.laa.2017.11.014. 32, 429

Konstantinos Pantazis, Avanti Athreya, Jesus Arroyo, William N. Frost, Evan S. Hill, and Vince Lyzinski. The Importance of Being Correlated: Implications of Dependence in Joint Spectral Inference across Multiple Networks. *Journal of Machine Learning Research*, 23 (141):1–77, 2022. ISSN 1533-7928. 153

Subhadeep Paul and Yuguo Chen. Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *The Annals of Statistics*, 48(1):230–250, February 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/18-AOS1800. 153, 159, 170, 177

Tiago P Peixoto. Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E*, 89(1):012804, 2014a. 170

Tiago P. Peixoto. The graph-tool python library. *figshare*, 2014b. doi: 10.6084/m9.figshare.1164194. 170

Tiago P Peixoto. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. *Physical Review E*, 92(4):042807, 2015. 153, 155, 170

Marianna Pensky and Yaxuan Wang. Clustering of Diverse Multiplex Networks. *arXiv:2110.05308 [stat]*, October 2021. 150, 153

Marianna Pensky and Teng Zhang. Spectral clustering in the dynamic stochastic block model. *Electronic Journal of Statistics*, 13(1):678–709, January 2019. ISSN 1935-7524, 1935-7524. doi: 10.1214/19-EJS1533. 150

Carey E. Priebe, Youngser Park, Joshua T. Vogelstein, John M. Conroy, Vince Lyzinski, Minh Tang, Avanti Athreya, Joshua Cape, and Eric Bridgeford. On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences*, 116(13):5995–6000, March 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1814462116. 115

T. Qin and K. Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. *Advances in Neural Information Processing Systems*, 2013. 157, 159, 168

Emile Richard and Andrea Montanari. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014. 82

Karl Rohe and Muzhe Zeng. Vintage Factor Analysis with Varimax Performs Statistical Inference. *arXiv:2004.05387 [math, stat]*, April 2020. 32, 60

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, August 2011. ISSN 0090-5364. doi: 10.1214/11-AOS887. 150, 152

Patrick Rubin-Delanchy. Manifold structure in graph embeddings. *arXiv:2006.05168 [cs, stat]*, June 2020. 122

Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E. Priebe. A statistical interpretation of spectral embedding: the generalised random dot product graph. *arXiv:1709.05506 [cs, stat]*, January 2020. 31, 116, 121, 122, 124, 392, 416, 417, 428

Patrick Rubin-Delanchy, Joshua Cape, Minh Tang, and Carey E. Priebe. A statistical interpretation of spectral embedding: The generalised random dot product graph. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a(n/a), 2022. ISSN 1467-9868. doi: 10.1111/rssb.12509. 25, 152, 158, 162, 448

Geoffrey Schiebinger, Martin J. Wainwright, and Bin Yu. The geometry of kernelized spectral clustering. *The Annals of Statistics*, 43(2):819–846, April 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/14-AOS1283. 5, 31, 32

Markus Schneider. Probability inequalities for kernel embeddings in sampling without replacement. In *Artificial Intelligence and Statistics*, pages 66–74, 2016. 403, 407

Evan Schwab, Benjamin D. Haeffele, René Vidal, and Nicolas Charon. Global Optimality in Separable Dictionary Learning with Applications to the Analysis of Diffusion MRI. *SIAM Journal on Imaging Sciences*, 12(4):1967–2008, January 2019. doi: 10.1137/18M121976X. 80

R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series. Wiley, 1980. ISBN 978-0-471-02403-3. 117, 125

Vinesh Solanki, Patrick Rubin-Delanchy, and Ian Gallagher. Persistent Homology of Graph Embeddings. *arXiv:1912.10238 [math, stat]*, December 2019. 391, 394

Olaf Sporns and Richard F Betzel. Modular brain networks. *Annual review of psychology*, 67:613, 2016. 149

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011. ISSN 1533-7928. 125

Prateek R. Srivastava, Purnamrita Sarkar, and Grani A. Hanasusanto. A Robust Spectral Clustering Algorithm for Sub-Gaussian Mixture Models with Outliers. *arXiv:1912.07546 [cs, math, stat]*, January 2021. 31

Liangjun Su, Wuyi Wang, and Yichong Zhang. Strong Consistency of Spectral Clustering for Stochastic Block Models. *IEEE Transactions on Information Theory*, 66(1):324–338, January 2020. ISSN 1557-9654. doi: 10.1109/TIT.2019.2934157. 150, 153, 154, 159, 164, 168

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29): 14516–14525, July 2019. doi: 10.1073/pnas.1810420116. 82

Pragya Sur, Yuxin Chen, and Emmanuel J. Candès. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled Chi-square. *Probability Theory and Related Fields*, 175(1):487–558, October 2019. ISSN 1432-2064. doi: 10.1007/s00440-018-00896-9. 82

Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012. 162

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, August 2013. ISSN 0378-3758. doi: 10.1016/j.jspi.2013.03.018. 116

Minh Tang. The eigenvalues of stochastic blockmodel graphs. *arXiv:1803.11551 [cs, stat]*, March 2018. 74, 273, 294

Minh Tang and Carey E. Priebe. Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415, October 2018. ISSN 0090-5364, 2168-8966. doi: 10.1214/17-AOS1623. 394, 417

Minh Tang, Daniel L. Sussman, and Carey E. Priebe. Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430, June 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1112. 122

Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, Youngser Park, and Carey E. Priebe. A Semiparametric Two-Sample Hypothesis Testing Problem for Random

Graphs. *Journal of Computational and Graphical Statistics*, 26(2):344–354, April 2017a. ISSN 1061-8600. doi: 10.1080/10618600.2016.1193505. 116, 138, 394

Minh Tang, Avanti Athreya, Daniel L. Sussman, Vince Lyzinski, and Carey E. Priebe. A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli*, 23 (3):1599–1630, August 2017b. ISSN 1350-7265. doi: 10.3150/15-BEJ789. 139, 140, 147, 396, 398, 418, 419, 421

Minh Tang, Joshua Cape, and Carey E. Priebe. Asymptotically efficient estimators for stochastic blockmodels: the naive MLE, the rank-constrained MLE, and the spectral. *arXiv:1710.10936 [stat]*, October 2017c. 74, 127, 270, 273, 294

Wei Tang, Zhengdong Lu, and Inderjit S Dhillon. Clustering with multiple graphs. In *2009 Ninth IEEE International Conference on Data Mining*, pages 1016–1021. IEEE, 2009. 159

Joel A. Tropp. An Introduction to Matrix Concentration Inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, May 2015. ISSN 1935-8237. doi: 10.1561/2200000048. 202, 205

M. Udell and A. Townsend. Why Are Big Data Matrices Approximately Low Rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, January 2019. doi: 10.1137/18M1183480. 122

S.A. van de Geer. *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2009. ISBN 978-0-521-12325-9. 421

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, June 2000. ISBN 978-0-521-78450-4. Google-Books-ID: SYlmEAAAQBAJ. 2

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, June 2004. ISSN 0022-0000. doi: 10.1016/j.jcss.2003.11.008. 5, 32

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge

University Press, 2018. doi: 10.1017/9781108231596. 1, 2, 34, 74, 85, 193, 200, 237, 244, 257, 311

Roman Vershynin. Concentration inequalities for random tensors. *arXiv:1905.00802 [math]*, June 2020. 256

J. T. Vogelstein, W. G. Roncal, R. J. Vogelstein, and C. E. Priebe. Graph Classification Using Signal-Subgraphs: Applications in Statistical Connectomics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(7):1539–1551, 2013. 115

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, December 2007. ISSN 1573-1375. doi: 10.1007/s11222-007-9033-z. 5, 32, 150

Vincent Q. Vu and Jing Lei. Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics*, 41(6):2905–2947, December 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1151. 60, 62, 70

Vincent Q Vu, Juhee Cho, Jing Lei, and Karl Rohe. Fantope Projection and Selection: A near-optimal convex relaxation of sparse PCA. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2670–2678. Curran Associates, Inc., 2013. 60, 63, 70

Martin Wahl. A note on the prediction error of principal component regression. *arXiv:1811.02998 [math, stat]*, April 2019a. 288

Martin Wahl. On the perturbation series for eigenvalues and eigenprojections. *arXiv:1910.08460 [math, stat]*, October 2019b. 288

M. J. Wainwright. Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using $_1$ -Constrained Quadratic Programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009. doi: 10.1109/TIT.2009.2016018. 61

Martin J. Wainwright. High-Dimensional Statistics: A Non-Asymptotic Viewpoint, February 2019. ISBN: 9781108627771 9781108498029 Library Catalog: www.cambridge.org Publisher: Cambridge University Press. 2, 72, 74, 278, 279, 424, 427

Haifeng Wang, Jinchi Chen, and Ke Wei. Implicit Regularization and Entrywise Convergence of Riemannian Optimization for Low Tucker-Rank Tensor Completion. Technical Report arXiv:2108.07899, arXiv, November 2021. arXiv:2108.07899 [math] type: article. 81

Miaoyan Wang and Yuchen Zeng. Multiway clustering via tensor block models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 78, 100

Tengyao Wang, Quentin Berthet, and Richard J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5): 1896–1930, October 2016. ISSN 0090-5364. doi: 10.1214/15-AOS1369. 60

Weichen Wang and Jianqing Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45(3):1342–1374, June 2017. ISSN 0090-5364, 2168-8966. doi: 10.1214/16-AOS1487. 31

Y. X. Rachel Wang and Peter J. Bickel. Likelihood-based model selection for stochastic block models. *Annals of Statistics*, 45(2):500–528, April 2017. ISSN 0090-5364, 2168-8966. doi: 10.1214/16-AOS1457. 163

Zhaoran Wang, Huanran Lu, and Han Liu. Tighten after Relax: Minimax-Optimal Sparse PCA in Polynomial Time. *Advances in neural information processing systems*, 2014:3383–3391, 2014. ISSN 1049-5258. 63

Tao Wu, Austin R Benson, and David F Gleich. General Tensor Spectral Co-clustering for Higher-Order Data. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 78, 80

Yihong Wu and Pengkun Yang. Optimal estimation of Gaussian mixtures via denoised method of moments. *The Annals of Statistics*, 48(4):1981 – 2007, 2020. doi: 10.1214/19-AOS1873. 90

Dong Xia. Confidence Region of Singular Subspaces for Low-Rank Matrix Regression. *IEEE Transactions on Information Theory*, 65(11):7437–7459, November 2019. ISSN 1557-9654. doi: 10.1109/TIT.2019.2924900. 32, 210

Dong Xia. Normal approximation and confidence region of singular subspaces. *Electronic Journal of Statistics*, 15(2):3798–3851, January 2021. ISSN 1935-7524, 1935-7524. doi: 10.1214/21-EJS1876. 12, 32, 33, 37, 41, 42, 54, 210, 219, 288

Dong Xia and Ming Yuan. Statistical Inferences of Linear Forms for Noisy Matrix Completion. *arXiv:1909.00116 [cs, math, stat]*, June 2020. 32, 41, 42, 60, 270, 273, 300

Dong Xia and Fan Zhou. The Sup-norm Perturbation of HOSVD and Low Rank Tensor Denoising. *Journal of Machine Learning Research*, 20(61):1–42, 2019. ISSN 1533-7928. 82, 154

Fangzheng Xie. Euclidean Representation of Low-Rank Matrices and Its Statistical Applications. *arXiv:2103.04220 [math, stat]*, March 2021. 31

Fangzheng Xie. Entrywise limit theorems of eigenvectors for signal-plus-noise matrix models with weak signals, March 2022. Number: arXiv:2106.09840 arXiv:2106.09840 [math, stat]. 60, 81, 92, 270, 383

Fangzheng Xie and Yanxun Xu. Efficient Estimation for Random Dot Product Graphs via a One-step Procedure. *arXiv:1910.04333 [math, stat]*, November 2020. 31

Fangzheng Xie, Joshua Cape, Carey E. Priebe, and Yanxun Xu. Bayesian Sparse Spiked Covariance Model with a Continuous Matrix Shrinkage Prior. *Bayesian Analysis*, -1(-1): 1–25, January 2022. ISSN 1936-0975, 1931-6690. doi: 10.1214/21-BA1292. 31, 32, 60, 61, 63, 70, 74, 270, 273, 294

Jiaming Xu. Rates of Convergence of Spectral Methods for Graphon Estimation. In *International Conference on Machine Learning*, pages 5433–5442. PMLR, July 2018. ISSN: 2640-3498. 122

Yuling Yan, Yuxin Chen, and Jianqing Fan. Inference for Heteroskedastic PCA with Missing Data. *arXiv:2107.12365 [cs, math, stat]*, July 2021. 15, 20, 44, 60, 70, 82, 96, 97, 107, 153, 270

Congyuan Yang, Carey E. Priebe, Youngser Park, and David J. Marchette. Simultaneous Dimensionality and Complexity Model Selection for Spectral Graph Clustering. *Journal*

of Computational and Graphical Statistics, 0(0):1–20, September 2020. ISSN 1061-8600. doi: 10.1080/10618600.2020.1824870. 52

Fan Yang. Edge universality of separable covariance matrices. *Electronic Journal of Probability*, 24, 2019. ISSN 1083-6489. doi: 10.1214/19-EJP381. 66

Fan Yang. Linear spectral statistics of eigenvectors of anisotropic sample covariance matrices. *arXiv:2005.00999 [math, stat]*, May 2020. 66

Zhuoran Yang, Zhaoran Wang, Han Liu, Yonina C. Eldar, and Tong Zhang. Sparse Nonlinear Regression: Parameter Estimation and Asymptotic Inference. *arXiv:1511.04514 [cs, math, stat]*, November 2015. 60

Yi Yu, Tengyao Wang, and Richard Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102, April 2014. doi: 10.1093/biomet/asv008. 32, 61, 269

Ming Yuan and Cun-Hui Zhang. Incoherent Tensor Norms and Their Applications in Higher Order Tensor Completion. *IEEE Transactions on Information Theory*, 63(10):6753–6766, October 2017. ISSN 1557-9654. doi: 10.1109/TIT.2017.2724549. 110

Anderson Y. Zhang and Harrison H. Zhou. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, October 2016. ISSN 0090-5364, 2168-8966. doi: 10.1214/15-AOS1428. 149

Anderson Y. Zhang and Harrison H. Zhou. Leave-one-out Singular Subspace Perturbation Analysis for Spectral Clustering, May 2022. Number: arXiv:2205.14855 arXiv:2205.14855 [cs, math, stat]. 5, 82, 154, 178

Anru Zhang and Rungang Han. Optimal Sparse Singular Value Decomposition for High-Dimensional High-Order Data. *Journal of the American Statistical Association*, 114(528): 1708–1725, October 2019. ISSN 0162-1459. doi: 10.1080/01621459.2018.1527227. 82, 389

Anru Zhang and Dong Xia. Tensor SVD: Statistical and Computational Limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, November 2018. ISSN 1557-9654. doi: 10.1109/TIT.2018.2841377. 82, 87, 90, 94, 95, 96, 110, 330

Anru R. Zhang, T. Tony Cai, and Yihong Wu. Heteroskedastic PCA: Algorithm, Optimality, and Applications. *Annals of Statistics, to appear*, April 2021. 251, 253

Anru R. Zhang, T. Tony Cai, and Yihong Wu. Heteroskedastic PCA: Algorithm, optimality, and applications. *The Annals of Statistics*, 50(1):53–80, February 2022. ISSN 0090-5364, 2168-8966. doi: 10.1214/21-AOS2074. 31, 33, 35, 36, 37, 44, 53, 96, 97, 251

Chenyu Zhang, Rungang Han, Anru R. Zhang, and Paul. M. Voyles. Denoising atomic resolution 4D scanning transmission electron microscopy data with tensor singular value decomposition. *Ultramicroscopy*, 219:113123, December 2020a. ISSN 0304-3991. doi: 10.1016/j.ultramic.2020.113123. 80

Yichi Zhang and Minh Tang. Perturbation Analysis of Randomized SVD and its Applications to High-dimensional Statistics, March 2022. Number: arXiv:2203.10262 arXiv:2203.10262 [cs, math, stat]. 154

Yuan Zhang. Unseeded low-rank graph matching by transform-based unsupervised point registration. *arXiv:1807.04680 [cs, stat]*, July 2018. 141

Yuan Zhang and Dong Xia. Edgeworth expansions for network moments. *arXiv:2004.06615 [cs, math, stat]*, April 2020. 117, 127, 139

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Detecting Overlapping Communities in Networks Using Spectral Methods. *SIAM Journal on Mathematics of Data Science*, 2(2):265–283, January 2020b. doi: 10.1137/19M1272238. 31

Peng Zhao and Bin Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research*, 7(90):2541–2563, 2006. ISSN 1533-7928. 61

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4): 2266–2292, 2012. 152

Runbing Zheng and Minh Tang. Limit results for distributed estimation of invariant subspaces in multiple networks inference and PCA. Technical Report arXiv:2206.04306, arXiv, June 2022. arXiv:2206.04306 [math, stat] type: article. 153

Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, January 2018. ISSN 1052-6234, 1095-7189. doi: 10.1137/17M1122025. 32, 82, 107

Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, November 2006. ISSN 0167-9473. doi: 10.1016/j.csda.2005.09.010. 52, 101, 163

Ziwei Zhu, Tengyao Wang, and Richard J. Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv:1906.12125 [math, stat]*, June 2019. 31, 32

Vita

Joshua Agterberg received the Master of Science in Engineering degree in Applied Mathematics and Statistics from Johns Hopkins University in 2018, and he received the Bachelor's of Business Administration degree in Actuarial Science and Mathematics from the University of Wisconsin-Madison in 2017. Joshua's PhD studies have been supported through the MINDS (Mathematical Institute of Data Science) Fellowship (awarded three times) and the Charles and Catherine Counselman Fellowship (Fall 2019 - Spring 2023), and he is a recipient of the Acheson J. Duncan Fund for the Advancement of Research in Statistics Travel Award (twice), the IMS (Institute of Mathematical Statistics) Hannan Graduate Student Travel Award, and the best presentation award at the Joint Statistical Meetings Student Competition in Nonparametric Statistics for his work Agterberg et al. (2020a). He is a Johns Hopkins Applied Mathematics and Statistics Teaching Fellow. In his free time when he is not thinking about low-rank matrix models and high-dimensional statistics he enjoys cycling, hiking, and reading fantasy novels.