Journal of Machine Learning Research 20 (2019) 1-42 Submitted 7/17; Revised 12/18; Published 3/19

The Sup-norm Perturbation of HOSVD and Low Rank Tensor Denoising

Dong Xia * MADXIA@UST.HK
Department of Mathematics
Hong Kong University of Science and Technology
Hong Kong SAR, China

Fan Zhou † FZHOU40@MATH.GATECH.EDU
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332, USA

Editor: Ryota Tomioka

Abstract

The higher order singular value decomposition (HOSVD) of tensors is a generalization of matrix SVD. The perturbation analysis of HOSVD under random noise is more delicate than its matrix counterpart. Recently, polynomial time algorithms have been proposed where statistically optimal estimates of the singular subspaces and the low rank tensors are attainable in the Euclidean norm. In this article, we analyze the sup-norm perturbation bounds of HOSVD and introduce estimators of the singular subspaces with sharp deviation bounds in the sup-norm. We also investigate a low rank tensor denoising estimator and demonstrate its fast convergence rate with respect to the entry-wise errors. The sup-norm perturbation bounds reveal unconventional phase transitions for statistical learning applications such as the exact clustering in high dimensional Gaussian mixture model and the exact support recovery in sub-tensor localizations. In addition, the bounds established for HOSVD also elaborate the one-sided sup-norm perturbation bounds for the singular subspaces of unbalanced (or fat) matrices.

Keywords: HOSVD, Entry-wise perturbation, Gaussian noise, High dimensional clustering.

1. Introduction

A tensor is a multi-array of more than 2 dimensions, which can be viewed as a higher order generalization of matrices. Data of tensor types has been widely available in many fields, such as image and video processing (see Liu et al. (2013), Westin et al. (2002), Hildebrand and Rüegsegger (1997), Li and Li (2010), Vasilescu and Terzopoulos (2002)); latent variable modelling (see Anandkumar et al. (2014), Cichocki et al. (2015), Chaganty and Liang (2013)); genomic signal processing (Omberg et al. (2007), Muralidhara et al.

∗. Most of the work was done when the first author was a Visiting Assistant Professor in Department of Statistics at University of Wisconsin-Madison and later a Post-doctoral Research Scientist in Department of Statistics at Columbia University.

†. Supported in part by NSF Grants DMS-1509739 and CCF-1523768.

(2011) and Ponnapalli et al. (2011)) and references therein. It is demanding to handle these datasets in order to take the most advantages of the tensor structures. The task is challenging due to the highly non-convexity of tensor related optimization problems. For instance, computing the tensor operator norm is generally NP-hard (see, e.g., Hillar and Lim (2013)) while it can be implemented fast for matrices.

The higher order singular value decomposition (HOSVD) is one machinery to deal with tensors which generalizes the matrix SVD to higher order tensors, see Zheng and Tomioka (2015), De Lathauwer et al. (2000b), Bergqvist and Larsson (2010), Chen and Saad (2009) and Kolda and Bader (2009). The conceptual simplicity and computational efficiency make HOSVD popular. It has been successfully applied to various statistical learning tasks, for instance, face recognition (see Vasilescu and Terzopoulos (2002)), genomic signal processing (see Muralidhara et al. (2011)) and more examples in a survey paper (Acar and Yener (2009)). Basically, the HOSVD unfolds a higher order tensor into matrices and treat it with standard matrix techniques to obtain the principal singular subspaces in each dimension (see more details in Section 2). Although the HOSVD shows appealing effectiveness, there are several fundamental theoretic mysteries yet to be uncovered.

One particularly important question is related to the perturbation of HOSVD when a low rank tensor is contaminated by stochastic noise. The difficulty comes from both methodological and theoretical aspects. The computation of HOSVD is essentially reduced to matrix SVD which can be implemented efficiently. This naive estimator is actually statistically sub-optimal. It is well-known that further power iterations can ameliorate the naive spectral initializations and thus deliver statistically optimal estimators, see more details in Richard and Montanari (2014), Zhang and Xia (2018), Hopkins et al. (2015), Liu et al. (2017) and references therein. Another intriguing phenomenon is on the phase transitions of the signal-to-noise ratio (SNR). Actually, the SNR exhibits distinct computational and statistical phase transitions, while the differences do not exist for matrix SVD. In particular, there is a gap on SNR between statistical optimality and computational optimality for HOSVD, see Zhang and Xia (2018). For introductory simplicity [1], we focus on the third-order tensors. Suppose that an unknown tensor $\mathbf{A} \in \mathbb{R}^{d\times d\times d}$ with multilinear ranks (r, r, r) is planted in a noisy observation $\mathbf{Y}$ with

$$\mathbf{Y} = \mathbf{A} + \mathbf{Z} \in \mathbb{R}^{d\times d\times d}. \tag{1}$$

The noise tensor $\mathbf{Z}$ has i.i.d. entries with $Z(i, j, k) \sim \mathcal{N}(0, \sigma^2)$ for $i, j, k \in [d]$ and noise variance $\sigma^2 > 0$. Here, we denote by $[d] := \{1, \ldots, d\}$. The signal strength $\underline{\Lambda}(\mathbf{A})$ is defined as the smallest nonzero singular values of the matrices unfolded from $\mathbf{A}$ (see definitions in Section 3.3). Let $\mathbf{U}, \mathbf{V}, \mathbf{W} \in \mathbb{R}^{d\times r}$ denote the singular vectors of $\mathbf{A}$ in the corresponding dimensions. It was proved (see Zheng and Tomioka (2015), Zhang and Xia (2018) and Liu et al. (2017)) that if the signal strength $\underline{\Lambda}(\mathbf{A}) \geq D_1\sigma d^{3/4}$ for a large enough constant

1. More general results where $\mathbf{A}$ is $d_1 \times d_2 \times d_3$ with multilinear ranks (r_1, r_2, r_3) can be found in Section 3. The results of this article can be easily generalized to higher order tensors.

$D_1 > 0$, the following bound holds

$$r^{-1/2}\max\Big\{\big\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top-\mathbf{U}\mathbf{U}^\top\big\|_{\ell_2},\big\|\hat{\mathbf{V}}\hat{\mathbf{V}}^\top-\mathbf{V}\mathbf{V}^\top\big\|_{\ell_2},\big\|\hat{\mathbf{W}}\hat{\mathbf{W}}^\top-\mathbf{W}\mathbf{W}^\top\big\|_{\ell_2}\Big\}$$
$$=O_p\Big(\frac{\sigma d^{1/2}}{\underline{\Lambda}(\mathbf{A})}+\frac{\sigma d^{3/2}}{\underline{\Lambda}^2(\mathbf{A})}\Big),$$

where $\hat{\mathbf{U}},\hat{\mathbf{V}},\hat{\mathbf{W}}$ represent the naive SVD obtained from noisy tensor $\mathbf{Y}$ and $\|\cdot\|_{\ell_2}$ denotes the Euclidean norm. Power iterations (also called higher order orthogonal iterations, see De Lathauwer et al. (2000a)) can improve the estimate (denoted by $\widetilde{\mathbf{U}},\widetilde{\mathbf{V}},\widetilde{\mathbf{W}}$) to

$$r^{-1/2}\max\Big\{\big\|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top-\mathbf{U}\mathbf{U}^\top\big\|_{\ell_2},\big\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top-\mathbf{V}\mathbf{V}^\top\big\|_{\ell_2},\big\|\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^\top-\mathbf{W}\mathbf{W}^\top\big\|_{\ell_2}\Big\}$$
$$=O_p\Big(\frac{\sigma d^{1/2}}{\underline{\Lambda}(\mathbf{A})}\Big), \qquad (2)$$

which is statistically optimal (see Zhang and Xia (2018)). Moreover, it is demonstrated in Zhang and Xia (2018), built on a hardness conjecture of the hyper-graphical planted clique detection problem, that if $\underline{\Lambda}(\mathbf{A})=o\big(\sigma d^{3/4}\big)$, then all polynomial time algorithms deliver trivial estimates of $\mathbf{U},\mathbf{V},\mathbf{W}$ in general.

One focus of this article is on estimating the linear forms of tensor singular vectors in model (1). More specifically, let $\mathbf{U}=\big(\mathbf{u}_1,\ldots,\mathbf{u}_r\big)\in\mathbb{R}^{d\times r}$ be $\mathbf{A}$'s singular vectors in certain mode, our goal is to estimate $\langle\mathbf{u}_j,\mathbf{x}\rangle$ for fixed $\mathbf{x}\in\mathbb{R}^d$ and $j=1,\ldots,r$. Through choosing $\mathbf{x}$ all over the canonical basis vectors in $\mathbb{R}^d$, we end up with an estimate of $\mathbf{u}_j$ whose component-wise perturbation bound can be attained. Unlike the ℓ_2-norm perturbation bound, the ℓ_∞ bound can characterize the entry-wise sign consistency and entry-wise significance (i.e. entry-wise magnitude) of singular vectors. The component-wise signs of singular vectors are critical in numerous applications such as community detection (see Florescu and Perkins (2015), Newman (2004), Mitra (2009) and Jin (2015)). The entry-wise significance is advantageous in sub-matrix localizations, see Cai et al. (2015), Ma and Wu (2015) and references therein. In Section 4, we show that the sup-norm perturbation bounds reveal unconventional phase transitions for the exact clustering in high dimensional Gaussian mixture model. Put it simply, algorithms based on the sup-norm bounds require weaker SNR conditions than algorithms driven by the ℓ_2-norm bounds to guarantee exact clustering. Furthermore, it enables us to construct a low rank denoising estimator of $\mathbf{A}$ so that entry-wise denoising is fulfilled. To the best of our knowledge, ours is the first result concerning the low rank tensor denoising with sharp entry-wise deviation bounds. In Section 4, we show that a simple algorithm based on the ℓ_∞ bounds can exactly recover the supports for sub-tensor localizations (see Remark 13).

To better highlight our contributions, suppose that $\mathbf{A}$ is an orthogonally decomposable third order tensor with (in particular, the CP decomposition of orthogonally decomposable tensors)

$$\mathbf{A}=\sum_{k=1}^r\lambda_k\big(\mathbf{u}_k\otimes\mathbf{v}_k\otimes\mathbf{w}_k\big),\quad \lambda_1\geq\ldots\geq\lambda_r>0 \qquad (3)$$

where $\mathbf{U}=(\mathbf{u}_1,\ldots,\mathbf{u}_r),\mathbf{V}=(\mathbf{v}_1,\ldots,\mathbf{v}_r)$ and $\mathbf{W}=(\mathbf{w}_1,\ldots,\mathbf{w}_r)$ are $d\times r$ matrices containing orthonormal columns. The k-th eigengap is written as $\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\big)=\bar{g}_k\big(\mathcal{M}_2(\mathbf{A})\big)=$

$\bar{g}_k\big(\mathcal{M}_3(\mathbf{A})\big) = \min\big(\lambda_{k-1}-\lambda_k, \lambda_k-\lambda_{k+1}\big)$ where $\mathcal{M}_j(\mathbf{A})$ represents the matrices from unfoldings of $\mathbf{A}$ (see Section 2). We preset $\lambda_0 = +\infty$ and $\lambda_{r+1} = 0$ for notational consistency. We show that, if $\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big) \geq D_1\big(\sigma\lambda_1 d^{1/2} + \sigma^2 d^{3/2}\big)$ for a large enough absolute constant $D_1 > 0$, then the following bound holds for any $\mathbf{x} \in \mathbb{R}^d$,

$$\Big|\langle \hat{\mathbf{u}}_k, \mathbf{x}\rangle - (1+b_k)^{1/2}\langle \mathbf{u}_k, \mathbf{x}\rangle\Big| = O_p\bigg(\|\mathbf{x}\|_{\ell_2}\frac{\lambda_1\sigma + d\sigma^2}{\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big)}\bigg) = O_p\bigg(\frac{\|\mathbf{x}\|_{\ell_2}}{d^{1/2}}\bigg). \tag{4}$$

where $b_k \in [-1/2, 0]$ is a constant which does not depend on $\mathbf{x}$. The $d \times r$ matrix $\hat{\mathbf{U}} = (\hat{\mathbf{u}}_1, \cdots, \hat{\mathbf{u}}_k)$ represent the empirical left singular vectors of mode-1 unfolding of $\mathbf{Y}$ satisfying model (1).

In the special case that $r = 1$ (rank one spiked tensor PCA model, see Richard and Montanari (2014)) such that $\underline{\Lambda}(\mathbf{A}) = \bar{g}_1\big(\mathcal{M}_1(\mathbf{A})\big) = \lambda_1$, we get from (4) that

$$\Big|\langle \hat{\mathbf{u}}_1, \mathbf{x}\rangle - (1+b_1)^{1/2}\langle \mathbf{u}_1, \mathbf{x}\rangle\Big| = O_p\bigg(\frac{\sigma}{\underline{\Lambda}(\mathbf{A})} + \frac{\sigma^2 d}{\underline{\Lambda}^2(\mathbf{A})}\bigg)\|\mathbf{x}\|_{\ell_2}. \tag{5}$$

By taking $\mathbf{x}$ over the canonical basis vectors in $\mathbb{R}^d$, the bounds in (5) imply that

$$\big\|\hat{\mathbf{u}}_1 - (1+b_1)^{1/2}\mathbf{u}_1\big\|_{\ell_\infty} = O_p\bigg(\Big(\frac{\log d}{d}\Big)^{1/2}\bigg) \tag{6}$$

under the eigen-gap condition $\underline{\Lambda}(\mathbf{A}) = \lambda_1 \gg \sigma d^{3/4}$. It is the standard requirement in tensor PCA. [2] Based on (6), we propose a low rank tensor estimator (denoted by $\hat{\mathbf{A}}$) under the same SNR requirements such that

$$\|\hat{\mathbf{A}} - \mathbf{A}\|_{\ell_\infty} = O_p\bigg(\Big(\frac{\sigma^2 d}{\lambda_1} + \sigma\Big)\big(\|\mathbf{u}_1\|_{\ell_\infty}\|\mathbf{v}_1\|_{\ell_\infty} + \|\mathbf{u}_1\|_{\ell_\infty}\|\mathbf{w}_1\|_{\ell_\infty} + \|\mathbf{v}_1\|_{\ell_\infty}\|\mathbf{w}_1\|_{\ell_\infty}\big)\bigg). \tag{7}$$

Equation (7) shows that the entry-wise denoising bound of the novel estimator $\hat{\mathbf{A}}$ is determined by the coherences of the singular vectors $\mathbf{u}_1, \mathbf{v}_1$ and $\mathbf{w}_1$. In particular, if $\mathbf{u}_1, \mathbf{v}_1, \mathbf{w}_1$ are incoherent so that $\max\{\|\mathbf{u}_1\|_{\ell_\infty}, \|\mathbf{v}_1\|_{\ell_\infty}, \|\mathbf{w}_1\|_\infty\} = O\big(\frac{1}{\sqrt{d}}\big)$, then equation (7) implies that

$$\|\hat{\mathbf{A}} - \mathbf{A}\|_{\ell_\infty} = O_p\Big(\frac{\sigma^2}{\lambda_1} + \frac{\sigma}{d}\Big).$$

Our main contribution is on the theoretical front. The HOSVD is essentially the standard SVD computed on an unbalanced matrix where the column size is much larger than the row size. The perturbation tools, such as Wedin's $\sin\Theta$ theorem (Wedin (1972)), characterize the ℓ_2 bounds through the larger dimension, even when the left singular space lies in a low dimensional space. At the high level, the HOSVD is connected to the one-sided spectral analysis (see, e.g.,Zheng and Tomioka (2015), Wang (2015), Cai and Zhang (2016) and references therein) which provide sharp perturbation bounds in ℓ_2-norm. There are recent bounds (see Fan et al. (2016) and Cape et al. (2017)) in ℓ_∞-norm developed under

2. We shall point out that a similar result on matrix SVD has appeared in Koltchinskii and Xia (2016) which is sub-optimal for tensors or unbalanced matrices. Indeed, the result in Koltchinskii and Xia (2016) is established under the eigengap condition $\lambda_1 \geq D_1\sigma d$.

additional constraint (incoherent singular subspaces) and structural noise (sparse noise). To obtain a sharp ℓ_∞-norm bound, we borrow the instruments invented by Koltchinskii and Lounici (2016) and extensively applied in Koltchinskii and Xia (2016). Our framework starts from a second order method of estimating the singular subspaces, which improves the eigengap condition than the first order method. Similar techniques have been proposed for tensor completion (Xia and Yuan (2019)) and tensor PCA (Zheng and Tomioka (2015) and Liu et al. (2017)). The success of this seemingly natural treatment hinges upon delicate dealing with the correlations among higher order terms. We benefit from these ℓ_∞-norm perturbation bounds by proposing a low rank estimator for tensor denoising where the entry-wise deviation error is guaranteed by the tensor incoherence conditions.

We organize our paper as follows. Tensor notations and preliminaries on HOSVD are explained in Section 2. Our main theoretical contributions are presented in Section 3 which includes the ℓ_∞-norm bound of the singular subspace perturbation and the entry-wise accuracy of a low rank tensor denoising estimator. In Section 4, we apply our theoretical results on applications including high dimensional clustering and sub-tensor localizations to manifest the advantages of utilizing ℓ_∞ bounds, where algorithms driven by the ℓ_∞-norm bounds are designed. Results of numerical experiments are displayed in Section 4.3. The proofs are provided in Section 6.

2. Preliminaries on Tensor and HOSVD

2.1. Notations

We first review some notations that will be used through the paper. We use boldfaced upper-case letters to denote tensors or matrices, and use the same letter in normal font with indices to denote its entries. We use boldfaced lower-case letters to represent vectors, and the same letter in normal font with indices to represent its entries. For notationally simplicity, our main context is focused on third-order tensors, while our results can be easily generalized to higher order tensors.

Given a third-order tensor $\mathbf{A} \in \mathbb{R}^{d_1\times d_2\times d_3}$, define a linear mapping $\mathcal{M}_1 : \mathbb{R}^{d_1\times d_2\times d_3} \mapsto \mathbb{R}^{d_1\times(d_2d_3)}$ such that

$$\mathcal{M}_1(\mathbf{A})\big(i_1, (i_2-1)d_3+i_3\big) = A(i_1,i_2,i_3), \quad i_1\in[d_1], i_2\in[d_3], i_3\in[d_3]$$

which is conventionally called the unfolding (or matricization) of tensor $\mathbf{A}$. It is also called the mode-1 unfolding of $\mathbf{A}$. The columns of matrix $\mathcal{M}_1(\mathbf{A})$ are called the mode-1 fibers of $\mathbf{A}$. The corresponding matricizations $\mathcal{M}_2(\mathbf{A})$ and $\mathcal{M}_3(\mathbf{A})$ can be defined in a similar fashion. The multilinear ranks of $\mathbf{A}$ are then defined by:

$$r_1(\mathbf{A}) := \operatorname{rank}\big(\mathcal{M}_1(\mathbf{A})\big), \quad r_2(\mathbf{A}) := \operatorname{rank}\big(\mathcal{M}_2(\mathbf{A})\big), \quad r_3(\mathbf{A}) := \operatorname{rank}\big(\mathcal{M}_3(\mathbf{A})\big)$$

Note that $r_1(\mathbf{A}), r_2(\mathbf{A}), r_3(\mathbf{A})$ are unnecessarily equal with each other in general. We write $\mathbf{r}(\mathbf{A}) := \big(r_1(\mathbf{A}), r_2(\mathbf{A}), r_3(\mathbf{A})\big)$ which are also called the Tucker ranks of $\mathbf{A}$.

The marginal product $\times_1 : \mathbb{R}^{r_1\times r_2\times r_3} \times \mathbb{R}^{d_1\times r_1} \mapsto \mathbb{R}^{d_1\times r_2\times r_3}$ is given by

$$\mathbf{C}\times_1 \mathbf{U} = \left(\sum_{j_1=1}^{r_1} C(j_1,j_2,j_3)U(i_1,j_1)\right)_{i_1\in[d_1], j_2\in[r_2], j_3\in[r_3]},$$

and $\times_2$ and $\times_3$ are defined similarly. Therefore, we write the multilinear product of tensors $\mathbf{C} \in \mathbb{R}^{r_1 \times r_2 \times r_3}, \mathbf{U} \in \mathbb{R}^{d_1 \times r_1}, \mathbf{V} \in \mathbb{R}^{d_2 \times r_2}$ and $\mathbf{W} \in \mathbb{R}^{d_3 \times r_3}$ as

$$\mathbf{C} \cdot (\mathbf{U}, \mathbf{V}, \mathbf{W}) = \mathbf{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W} \in \mathbb{R}^{d_1 \times d_2 \times d_3}.$$

We use $\|\cdot\|$ to denote the operator norm of matrices and $\|\cdot\|_{\ell_2}$ and $\|\cdot\|_{\ell_\infty}$ to denote ℓ_2 and ℓ_∞ norms of vectors, or vectorized matrices and tensors.

2.2. HOSVD and Eigengaps

For a tensor $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ with multilinear ranks $\mathbf{r}(\mathbf{A}) = \big(r_1(\mathbf{A}), r_2(\mathbf{A}), r_3(\mathbf{A})\big)$, let $\mathbf{U} \in \mathbb{R}^{d_1 \times r_1(\mathbf{A})}, \mathbf{V} \in \mathbb{R}^{d_2 \times r_2(\mathbf{A})}$ and $\mathbf{W} \in \mathbb{R}^{d_3 \times r_3(\mathbf{A})}$ be the left singular vectors of $\mathcal{M}_1(\mathbf{A}), \mathcal{M}_2(\mathbf{A})$ and $\mathcal{M}_3(\mathbf{A})$ respectively, which can be computed efficiently via matricization followed by thin singular value decomposition. The higher order singular value decomposition (HOSVD) refers to the decomposition

$$\mathbf{A} = \mathbf{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W} \tag{8}$$

where the $r_1(\mathbf{A}) \times r_2(\mathbf{A}) \times r_3(\mathbf{A})$ core tensor $\mathbf{C}$ is obtained by $\mathbf{C} := \mathbf{A} \times_1 \mathbf{U}^\top \times_2 \mathbf{V}^\top \times_3 \mathbf{W}^\top$.

Suppose that a noisy version of $\mathbf{A}$ is observed as in model (1) so that

$$\mathbf{Y} = \mathbf{A} + \mathbf{Z}$$

where $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is an unknown noise tensor with i.i.d. entries satisfying $Z(i,j,k) \sim \mathcal{N}(0, \sigma^2)$. By observing $\mathbf{Y}$, our goal is to estimate $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$. An immediate approach is to compute HOSVD of $\mathbf{Y}$. To this end, let $\hat{\mathbf{U}} \in \mathbb{R}^{d_1 \times r_1}, \hat{\mathbf{V}} \in \mathbb{R}^{d_2 \times r_2}, \hat{\mathbf{W}} \in \mathbb{R}^{d_3 \times r_3}$ be the corresponding top singular vectors of $\mathcal{M}_1(\mathbf{Y}), \mathcal{M}_2(\mathbf{Y})$ and $\mathcal{M}_3(\mathbf{Y})$. The key factor characterizing the perturbation bounds of $\hat{\mathbf{U}}, \hat{\mathbf{V}}$ and $\hat{\mathbf{W}}$ is the so-called eigengap.

Since the computing of $\hat{\mathbf{U}}$ is essentially via the matrix SVD on $\mathcal{M}_1(\mathbf{A})$, it suffices to consider the eigengaps of matrices. Given a rank r matrix $\mathbf{M} \in \mathbb{R}^{m_1 \times m_2}$ with SVD:

$$\mathbf{M} = \sum_{k=1}^{r} \lambda_k \big(\mathbf{g}_k \otimes \mathbf{h}_k\big)$$

where $\mathbf{M}$'s singular values are $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ and $\{\mathbf{g}_1, \ldots, \mathbf{g}_r\}$ are the corresponding left singular vectors and $\{\mathbf{h}_1, \ldots, \mathbf{h}_r\}$ are $\mathbf{M}$'s corresponding right singular vectors. We further introduce $\lambda_0 = +\infty$ and $\lambda_{r+1} = 0$. The k-th eigengap of matrix $\mathbf{M}$ is then defined by

$$\bar{g}_k(\mathbf{M}) := \min\big(\lambda_k - \lambda_{k+1}, \lambda_{k-1} - \lambda_k\big), \quad \forall\ 1 \leq k \leq r.$$

Recall that $\mathbf{U}, \hat{\mathbf{U}} \in \mathbb{R}^{d_1 \times r_1}$ are the top-r_1 left singular vectors of $\mathcal{M}_1(\mathbf{A})$ and $\mathcal{M}_1(\mathbf{Y})$ respectively. By Davis-Kahan Theorem (Davis and Kahan (1970)) or Wedin's $\sin\Theta$ theorem (Wedin (1972)), we get

$$\|\hat{\mathbf{U}}\hat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\| = O\bigg(\frac{\|\mathcal{M}_1(\mathbf{Z})\|}{\bar{g}_{r_1}\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big)}\bigg), \tag{9}$$

which is generally sub-optimal especially when $\mathcal{M}_1(\mathbf{Z}) \in \mathbb{R}^{d_1 \times (d_2 d_3)}$ is unbalanced such that $d_2 d_3 \gg d_1$. Sharper bounds in ℓ_2-norm concerning one sided perturbation have been derived in Zheng and Tomioka (2015), Wang (2015) and Cai and Zhang (2016). In this paper, we derive sharp perturbation bounds of $\hat{\mathbf{U}}, \hat{\mathbf{V}}, \hat{\mathbf{W}}$ in ℓ_∞-norm which illustrate unconventional phase transitions for various statistical learning applications. More generally, we will investigate the perturbation bounds of linear forms $\langle \hat{\mathbf{u}}_k, \mathbf{x}\rangle$ for any fixed vector $\mathbf{x} \in \mathbb{R}^{d_1}$. Similar results can be obtained for singular vectors $\hat{\mathbf{V}}$ and $\hat{\mathbf{W}}$.

3. Main Results

3.1. Second Order Method for One-sided Spectral Analysis

The ℓ_∞-norm perturbation bounds for singular subspaces of balanced matrices has been developed in Koltchinskii and Xia (2016). Recall that $\mathbf{u}_k$ denotes the k-th left singular vector of $\mathcal{M}_1(\mathbf{A})$ and $\hat{\mathbf{u}}_k$ denotes the k-th left singular vector of $\mathcal{M}_1(\mathbf{Y})$ where $\mathcal{M}_1(\mathbf{A})$ is of size $d_1 \times (d_2 d_3)$. The operator norm $\|\mathcal{M}_1(\mathbf{Z})\|$ is generally determined by the larger dimension $(d_1 \vee d_2 d_3)$, see Section 6. It turns out that the machinery in Koltchinskii and Xia (2016) is sub-optimal concerning the SNR requirement. Indeed, the eigengap requirement in Koltchinskii and Xia (2016) becomes $\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big) \gg \sigma\big(d_1 \vee d_2 d_3\big)^{1/2}$, which shall is unnecessarily strong in view of the recent results in Zheng and Tomioka (2015), Cai and Zhang (2016), Zhang and Xia (2018) and Liu et al. (2017).

To bridge such gaps, we conduct a second order spectral analysis for $\hat{\mathbf{U}}$. The key observation is that the top left singular vectors of $\mathcal{M}_1(\mathbf{Y})$ are also the top eigenvectors of $\mathcal{M}_1(\mathbf{Y})\mathcal{M}_1^\top(\mathbf{Y})$. The second order method seeks the eigenspace perturbation on $\mathcal{M}_1(\mathbf{Y})\mathcal{M}_1^\top(\mathbf{Y})$ instead of singular space perturbation on $\mathcal{M}_1(\mathbf{Y})$. Clearly, we have

$$\mathcal{M}_1(\mathbf{Y})\mathcal{M}_1^\top(\mathbf{Y}) = \mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A}) + \boldsymbol{\Gamma} \in \mathbb{R}^{d_1 \times d_1}$$

where $\boldsymbol{\Gamma} = \mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{Z}) + \mathcal{M}_1(\mathbf{Z})\mathcal{M}_1^\top(\mathbf{A}) + \mathcal{M}_1(\mathbf{Z})\mathcal{M}_1^\top(\mathbf{Z})$. Note that $\mathbf{U}$ are the leading eigenvectors of $\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})$ and $\hat{\mathbf{U}}$ are the top-r_1 eigenvectors of $\mathcal{M}_1(\mathbf{Y})\mathcal{M}_1^\top(\mathbf{Y})$. Moreover, the following relation on eigengaps is obvious:

$$\bar{g}_{r_1}\Big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\Big) \geq \bar{g}_{r_1}^2\big(\mathcal{M}_1(\mathbf{A})\big).$$

The advantage of second order method comes from the observation that even though $\mathbb{E}\big\|\mathcal{M}_1(\mathbf{Z})\mathcal{M}_1^\top(\mathbf{Z})\big\|$ is of the order $\sigma^2(d_1 \vee d_2 d_3)$, the symmetric matrix $\mathcal{M}_1(\mathbf{Z})\mathcal{M}_1^\top(\mathbf{Z})$ is concentrated at $d_2 d_3 \sigma^2 \mathbf{I}_{d_1}$ such that (see more details in Section 6)

$$\big\|\mathcal{M}_1(\mathbf{Z})\mathcal{M}_1^\top(\mathbf{Z}) - \sigma^2 d_2 d_3 \mathbf{I}_{d_1}\big\| = O_p\Big(\sigma^2\big(d_1 d_2 d_3\big)^{1/2}\Big).$$

Note that subtracting by an identity matrix does not affect the eigen-structure. The second order method introduces the additional term $\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{Z})$ whose operator norm is bounded by $O_p\big(\sigma\sqrt{d_1}\|\mathcal{M}_1(\mathbf{A})\|\big)$, which creates a constraint on the condition number of $\mathcal{M}_1(\mathbf{A})$. However, in order to characterize sharp perturbation bounds of linear forms $\langle \hat{\mathbf{u}}_k, \mathbf{x}\rangle$, we need to pay more attention to dealing with correlations among the higher order terms than the first order method in Koltchinskii and Xia (2016). We note that the idea of second order method is already existing in the literature (see, e.g., Zheng and Tomioka (2015) for the ℓ_2-norm perturbation bounds). The second order moment method is only the starting point of our technical analysis which significantly reduces the SNR requirements.

Our most fundamental contribution is about the sup-norm characterization of the empirical singular vectors. Basically, we observe that the empirical singular vectors are biased and the bias is nicely aligned with the true singular vectors. After subtracting the bias, the empirical singular vectors exhibit the so-called *delocalization* property where all the entry-wise perturbations have comparable magnitudes. Such *delocalization* property is universal meaning that no conditions on the true singular vectors are needed. In Section 4, we show that the sup-norm perturbation bounds indeed reveal unconventional phase transitions in statistical learning applications such as the exact clustering in high dimensional Gaussian mixture models and the exact support recovery in sub-tensor localizations.

3.2. Perturbation of Linear Forms of Singular Vectors

In this section, we present our main theorem characterizing the perturbation of linear forms $\langle \hat{\mathbf{u}}_k, \mathbf{x}\rangle$ for any $\mathbf{x}\in\mathbb{R}^{d_1}$, where $\hat{\mathbf{u}}_k$ is the k-th left singular vector of $\mathcal{M}_1(\mathbf{Y})$. Our results have similar implications as the previous work Koltchinskii and Xia (2016), meaning that the bias $\mathbb{E}\hat{\mathbf{u}}_k\hat{\mathbf{u}}_k^\top - \mathbf{u}_k\mathbf{u}_k^\top$ is well aligned with $\mathbf{u}_k\mathbf{u}_k^\top$. Therefore, after correcting the bias term, we are able to obtain a sharper estimation of linear forms $\langle \mathbf{u}_k, \mathbf{x}\rangle$. To this end, we denote the condition number of the matrix $\mathcal{M}_1(\mathbf{A})$ by

$$\kappa\big(\mathcal{M}_1(\mathbf{A})\big) = \frac{\lambda_{\max}\big(\mathcal{M}_1(\mathbf{A})\big)}{\lambda_{\min}\big(\mathcal{M}_1(\mathbf{A})\big)}$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ return the largest and smallest nonzero singular values, respectively. Since $\hat{\mathbf{u}}_k$ is up to the switch of signs, we choose $\hat{\mathbf{u}}_k$ in the following theorems, remarks and corollaries so that $\big\langle \hat{\mathbf{u}}_k, \mathbf{u}_k\big\rangle > 0$.

Theorem 1 *Let*[3] $\mathbf{M} := \mathcal{M}_1(\mathbf{A})$ *and* $\delta(d_1,d_2,d_3) := \sigma d_1^{1/2}\|\mathbf{M}\| + \sigma^2(d_1d_2d_3)^{1/2}$ *and suppose* $d_2d_3e^{-d_1/2}\le 1$. *There exist absolute constants* $D_1, D_2>0$ *such that the following fact holds. Let* $\mathbf{u}_k$ *be* $\mathbf{M}$*'s* k*-th left singular vector with multiplicity* 1. *If* $\bar{g}_k\big(\mathbf{M}\mathbf{M}^\top\big)\ge D_1\delta(d_1,d_2,d_3)$, *there exist a constant* $b_k\in[-1/2,0]$ *with* $|b_k|\le \frac{\sqrt{2}\delta(d_1,d_2,d_3)}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}$ *such that for any* $\mathbf{x}$, *the following bound holds with probability at least* $1-e^{-t}$,

$$\begin{aligned}&\big|\langle \hat{\mathbf{u}}_k, \mathbf{x}\rangle - (1+b_k)^{1/2}\langle \mathbf{u}_k,\mathbf{x}\rangle\big|\\ \le& D_2\bigg(t^{1/2}\frac{\sigma\|\mathbf{M}\|+\sigma^2(d_2d_3)^{1/2}}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)} + \frac{\sigma^2 d_1}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}\Big(\frac{\delta(d_1,d_2,d_3)}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}\Big)\bigg)\|\mathbf{x}\|_{\ell_2}\end{aligned} \tag{10}$$

for all $\log 8\le t\le d_1$. *In particular, if* $\mathbf{x}=\pm\mathbf{u}_k$, *then with the same probability,*

$$\begin{aligned}\big||\langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle| - 1\big| \le& \big|\sqrt{1+b_k}-1\big|\\ &+D_2\bigg(t^{1/2}\frac{\sigma\|\mathbf{M}\|+\sigma^2(d_2d_3)^{1/2}}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)} + \frac{\sigma^2 d_1}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}\Big(\frac{\delta(d_1,d_2,d_3)}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}\Big)\bigg).\end{aligned}$$

By Theorem 1, it is easy to check that the condition $\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big)\ge D_1\delta(d_1,d_2,d_3)$ holds whenever

$$\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\big)\ge D_1\Big(\sigma(d_1d_2d_3)^{1/4} + \sigma d_1^{1/2}\kappa\big(\mathcal{M}_1(\mathbf{A})\big)\Big).$$

3. Observe that if we set $d_3=1$ and consider the case with $d_1\ll d_2$, then Theorem 1 elaborates the one-sided perturbation bounds in ℓ_∞-norm for singular vectors of unbalanced (or fat) matrices.

If $\kappa\big(\mathcal{M}_1(\mathbf{A})\big) \leq \big(\frac{d_2 d_3}{d_1}\big)^{1/4}$, the above bound becomes $\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\big) \gg \sigma(d_1 d_2 d_3)^{1/4}$ which is a standard requirement in tensor SVD or PCA, see Zheng and Tomioka (2015), Zhang and Xia (2018), Hopkins et al. (2015) and Richard and Montanari (2014). By taking $\mathbf{x}$ over the standard basis vectors in $\mathbb{R}^{d_1}$ and choosing $t \geq D_3 \log d_1$, we end up with a ℓ_∞-norm perturbation bound for empirical singular vector $\hat{\mathbf{u}}_k$.

Corollary 2 *Under the conditions in Theorem 1, there exists a universal constant $D_1 > 0$ such that the following bound holds with probability at least $1 - \frac{1}{d_1}$,*

$$\big\|\hat{\mathbf{u}}_k - (1+b_k)^{1/2}\mathbf{u}_k\big\|_{\ell_\infty} \leq D_1\bigg(\Big(\frac{\log d_1}{d_1}\Big)^{1/2} + \Big(\frac{d_1}{d_2 d_3}\Big)^{1/2}\bigg).$$

If $d_1 \asymp d_2 \asymp d_3 \asymp d$, we obtain

$$\mathbb{P}\Big(\big\|\hat{\mathbf{u}}_k - (1+b_k)^{1/2}\mathbf{u}_k\big\|_{\ell_\infty} \geq D_1\Big(\frac{\log d}{d}\Big)^{1/2}\Big) \leq \frac{1}{d}$$

which has an analogous form to the perturbation bound in Koltchinskii and Xia (2016) implying a famous *delocalization* phenomenon in random matrix theory, see Rudelson and Vershynin (2015) and Vu and Wang (2015) and references therein.

Remark 3 *Let's compare with the ℓ_2-norm bound in Zheng and Tomioka (2015) in the case that rank $r = 1$, $d_1 = d_2 = d_3 = d$ and signal strength $\bar{g}_1(\mathbf{M}\mathbf{M}^\top) = \lambda^2$. By (Zheng and Tomioka, 2015, Theorem 1), if $\lambda \gg \sigma d^{3/4}$, then*

$$\|\hat{\mathbf{u}}_1 - \mathbf{u}_1\|_{\ell_2} = O_p\Big(\frac{d^{1/2}\sigma}{\lambda} + \frac{\sigma^2 d^{3/2}}{\lambda^2}\Big). \tag{11}$$

By Theorem 1, if $\lambda \gg \sigma d^{3/4}$, then we get

$$\big\|\hat{\mathbf{u}}_1 - (1+b_1)^{1/2}\mathbf{u}_1\big\|_{\ell_\infty} = O_p\Big(\frac{\sigma \log^{1/2} d}{\lambda} + \frac{\sigma^2 d \log^{1/2} d}{\lambda^2}\Big) \tag{12}$$

for a constant $b_1 \in [-1/2, 0]$ depending on $\mathbf{u}_1$ and λ only. By (12) and (11), we observe that, after subtracting the bias, the entry-wise deviation of the empirical left singular vector $\hat{\mathbf{u}}_1$ is about $\sqrt{\frac{\log d}{d}}$ of the ℓ_2-norm perturbation bound of $\hat{\mathbf{u}}_1$. It means that, after subtracting the bias, the deviations of all $\hat{\mathbf{u}}_1$'s entries have comparable magnitudes, namely the so-called delocalization property. Interestingly, if $|u_1(j)| \gg \frac{1}{\sqrt{d}}$, then eq. (12) implies that $\hat{u}_1(j)$ has the same sign as $u_1(j)$ as long as $\lambda \gg \sigma d^{3/4}$. This sign consistency is crucial for guaranteeing the exact clustering of high dimensional mixture model, see more details in Section 4.

The bias b_k is usually unknown and we borrow the idea in Koltchinskii and Xia (2016) to estimate b_k based on two independent samples. It happens in the application of tensor decomposition for gene expression data where usually multiple independent copies are available, see more details in Hore et al. (2016).

Suppose that two independent noisy version of $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ are observed with $\mathbf{Y}^{(1)} = \mathbf{A} + \mathbf{Z}^{(1)}$ and $\mathbf{Y}^{(2)} = \mathbf{A} + \mathbf{Z}^{(2)}$ where $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ have i.i.d. centered Gaussian entries with variance σ^2 as in (1). Let $\hat{\mathbf{u}}_k^{(1)}$ and $\hat{\mathbf{u}}_k^{(2)}$ denote the k-th left singular vector of $\mathcal{M}_1\big(\mathbf{Y}^{(1)}\big)$ and $\mathcal{M}_1\big(\mathbf{Y}^{(2)}\big)$, respectively. The signs of $\hat{\mathbf{u}}_k^{(1)}$ and $\hat{\mathbf{u}}_k^{(2)}$ are chosen such that $\langle\hat{\mathbf{u}}_k^{(1)},\hat{\mathbf{u}}_k^{(2)}\rangle\geq 0$. Define the estimator of b_k by

$$\hat{b}_k := \langle\hat{\mathbf{u}}_k^{(1)},\hat{\mathbf{u}}_k^{(2)}\rangle - 1.$$

Define the scaled version of empirical singular vector $\widetilde{\mathbf{u}}_k := \frac{\hat{\mathbf{u}}_k}{(1+\hat{b}_k)^{1/2}}$, which is not necessarily a unit vector.

Theorem 4 *Under the assumptions in Theorem 1, there exists an absolute constant $D_1>0$ such that for any $\mathbf{x}\in\mathbb{R}^{d_1}$, the follow bound holds with probability at least $1-e^{-t}$ for all $\log 8\leq t\leq d_1$,*

$$\big|\hat{b}_k - b_k\big| \leq D_1\bigg(t^{1/2}\frac{\sigma\|\mathbf{M}\|+\sigma^2(d_2d_3)^{1/2}}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)} + \frac{\sigma^2 d_1}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}\Big(\frac{\delta(d_1,d_2,d_3)}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}\Big)\bigg)$$

and

$$\big|\big\langle\widetilde{\mathbf{u}}_k-\mathbf{u}_k,\mathbf{x}\big\rangle\big| \leq D_1\bigg(t^{1/2}\frac{\sigma\|\mathbf{M}\|+\sigma^2(d_2d_3)^{1/2}}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)} + \frac{\sigma^2 d_1}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}\Big(\frac{\delta(d_1,d_2,d_3)}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}\Big)\bigg)\|\mathbf{x}\|_{\ell_2}$$

where $\mathbf{M}=\mathcal{M}_1(\mathbf{A})$.

Remark 5 *By Theorem 4, if $d/2\leq\min_k d_k\leq\max_k d_k\leq 2d$, we get*

$$\mathbb{P}\Big(\|\widetilde{\mathbf{u}}_k-\mathbf{u}_k\|_{\ell_\infty}\geq D_1\Big(\frac{\log d}{d}\Big)^{1/2}\Big)\leq\frac{1}{d}.$$

3.3. Low Rank Tensor Denoising and Entry-wise Deviation Bound

In this section, we study a low rank estimate of $\mathbf{A}$ through the projection of $\mathbf{Y}$. Let $\widetilde{\mathbf{U}}=(\tilde{\mathbf{u}}_1,\ldots,\tilde{\mathbf{u}}_{r_1})\in\mathbb{R}^{d_1\times r_1}$ be scaled singular vectors each of which is computed as in Theorem 4. Similarly, let $\widetilde{\mathbf{V}}\in\mathbb{R}^{d_2\times r_2}$ and $\widetilde{\mathbf{W}}\in\mathbb{R}^{d_3\times r_3}$ be the corresponding scaled singular vectors computed from $\mathcal{M}_2(\mathbf{Y})$ and $\mathcal{M}_3(\mathbf{Y})$. Define the low rank estimate

$$\widetilde{\mathbf{A}} := \mathbf{Y}\times_1\mathbf{P}_{\widetilde{\mathbf{U}}}\times_2\mathbf{P}_{\widetilde{\mathbf{V}}}\times_3\mathbf{P}_{\widetilde{\mathbf{W}}}$$

where $\mathbf{P}_{\widetilde{\mathbf{U}}}$ represents the scaled projector $\mathbf{P}_{\widetilde{\mathbf{U}}}:=\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top$. Clearly, $\mathrm{rank}(\widetilde{\mathbf{A}})=(r_1,r_2,r_3)$ which serves as a low rank estimate of $\mathbf{A}$. We characterize the entry-wise accuracy of $\widetilde{\mathbf{A}}$, namely, the upper bound of $\|\widetilde{\mathbf{A}}-\mathbf{A}\|_{\ell_\infty}$ in terms of the coherence of $\mathbf{U},\mathbf{V}$ and $\mathbf{W}$. Our $\|\widetilde{\mathbf{A}}-\mathbf{A}\|_{\ell_\infty}$ bound relies on the simultaneous ℓ_∞-norm perturbation bounds of $\{\tilde{\mathbf{u}}_{k_1}\}_{k_1=1}^{r_1}$, $\{\tilde{\mathbf{v}}_{k_2}\}_{k_2=1}^{r_2}$ and $\{\tilde{\mathbf{w}}_{k_3}\}_{k_3=1}^{r_3}$. We impose the following conditions on the eigengaps: for a large enough constant $D_1>0$,

$$\bar{g}_{k_1}\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big)\geq D_1\Big(\sigma d_1^{1/2}\overline{\Lambda}(\mathbf{A})+\sigma^2(d_1d_2d_3)^{1/2}\Big),\quad 1\leq k_1\leq r_1, \tag{13}$$

$$\bar{g}_{k_2}\big(\mathcal{M}_2(\mathbf{A})\mathcal{M}_2^\top(\mathbf{A})\big)\geq D_1\Big(\sigma d_2^{1/2}\overline{\Lambda}(\mathbf{A})+\sigma^2(d_1d_2d_3)^{1/2}\Big),\quad 1\leq k_2\leq r_2, \tag{14}$$

$$\bar{g}_{k_3}\big(\mathcal{M}_3(\mathbf{A})\mathcal{M}_3^\top(\mathbf{A})\big) \geq D_1\Big(\sigma d_3^{1/2}\overline{\Lambda}(\mathbf{A}) + \sigma^2(d_1d_2d_3)^{1/2}\Big), \quad 1 \leq k_3 \leq r_3, \tag{15}$$

where we denote by

$$\overline{\Lambda}(\mathbf{A}) := \max\big\{\lambda_{\max}\big(\mathcal{M}_1(\mathbf{A})\big), \lambda_{\max}\big(\mathcal{M}_2(\mathbf{A})\big), \lambda_{\max}\big(\mathcal{M}_3(\mathbf{A})\big)\big\}.$$

Similarly, we define

$$\underline{\Lambda}(\mathbf{A}) := \min\Big\{\lambda_{\min}\big(\mathcal{M}_1(\mathbf{A})\big),\ \lambda_{\min}\big(\mathcal{M}_2(\mathbf{A})\big),\ \lambda_{\min}\big(\mathcal{M}_3(\mathbf{A})\big)\Big\}$$

and the overall eigengap

$$\bar{g}_{\min}(\mathbf{A}) := \min\Big\{\bar{g}_{k_1}^{1/2}\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big), \bar{g}_{k_2}^{1/2}\big(\mathcal{M}_2(\mathbf{A})\mathcal{M}_2^\top(\mathbf{A})\big), \bar{g}_{k_3}^{1/2}\big(\mathcal{M}_3(\mathbf{A})\mathcal{M}_3^\top(\mathbf{A})\big)$$
$$, 1 \leq k_1 \leq r_1, 1 \leq k_2 \leq r_2, 1 \leq k_3 \leq r_3\Big\}.$$

By definition, it is clear that $\underline{\Lambda}(\mathbf{A}) \geq \bar{g}_{\min}(\mathbf{A})$.

Theorem 6 *Suppose conditions (13) (14) (15) hold and assume that for all $i \in [d_1], j \in [d_2], k \in [d_3]$,*

$$\|\mathbf{U}^\top \mathbf{e}_i\|_{\ell_2} \leq \mu_{\mathbf{U}}\sqrt{\frac{r_1}{d_1}}, \quad \|\mathbf{V}^\top \mathbf{e}_j\|_{\ell_2} \leq \mu_{\mathbf{V}}\sqrt{\frac{r_2}{d_2}}, \quad \|\mathbf{W}^\top \mathbf{e}_k\|_{\ell_2} \leq \mu_{\mathbf{W}}\sqrt{\frac{r_3}{d_3}}$$

for some constants $\mu_{\mathbf{U}}, \mu_{\mathbf{V}}, \mu_{\mathbf{W}} \geq 0$. Suppose that $\frac{d}{2} \leq \min_{1\leq k\leq 3} d_k \leq \max_{1\leq k\leq 3} d_k \leq 2d$ and $\frac{r}{2} \leq \min_{1\leq k\leq 3} r_k \leq \max_{1\leq k\leq 3} r_k \leq 2r$. Then, there exists an absolute constant $D_2 > 0$ such that, with probability at least $1 - \frac{1}{d}$,

$$\begin{aligned}&\big\|\tilde{\mathbf{A}} - \mathbf{A}\big\|_{\ell_\infty}\\ &\leq D_2\sigma r^3\bigg(\frac{\widetilde{\kappa}(\mathbf{A})\sigma}{\bar{g}_{\min}(\mathbf{A})} + \frac{\widetilde{\kappa}^2(\mathbf{A})}{d}\bigg)\big(\mu_{\mathbf{U}}\mu_{\mathbf{V}} + \mu_{\mathbf{U}}\mu_{\mathbf{W}} + \mu_{\mathbf{V}}\mu_{\mathbf{W}}\big)\log^{3/2} d\end{aligned}$$

where $\widetilde{\kappa}(\mathbf{A}) = \overline{\Lambda}(\mathbf{A})/\bar{g}_{\min}(\mathbf{A})$.

Remark 7 *To highlight the contribution of Theorem 6, let $r = O(1)$ and $\widetilde{\kappa}(\mathbf{A}) = O(1)$. Note that if the coherence constants $\mu_{\mathbf{U}}, \mu_{\mathbf{V}}, \mu_{\mathbf{W}} = d^{(\frac{3}{4}-\varepsilon)/2}$ for $\varepsilon \in (0, 3/4)$, i.e., $\mathbf{U}, \mathbf{V}, \mathbf{W}$ can be almost spiked, under the minimal eigengap $\bar{g}_{\min}(\mathbf{A}) \gg \sigma d^{3/4}$, we obtain*

$$\|\tilde{\mathbf{A}} - \mathbf{A}\|_{\ell_\infty} = O_p\Big(\frac{\sigma}{d^{\varepsilon}}\log^{3/2} d\Big).$$

It worths to point out that the minimax optimal bound of estimating $\mathbf{A}$ in ℓ_2-norm is $O\big(\sigma d^{1/2}\big)$, see Zhang and Xia (2018). Theorem 6 is more interesting when $\mathbf{A}$ is incoherent such that $\mu_{\mathbf{U}}, \mu_{\mathbf{V}}, \mu_{\mathbf{W}} = O(1)$ where we can conclude that

$$\|\tilde{\mathbf{A}} - \mathbf{A}\|_{\ell_\infty} = O_p\bigg(\Big(\frac{\sigma^2}{\bar{g}_{\min}(\mathbf{A})} + \frac{\sigma}{d}\Big)\log^{3/2} d\bigg) = O_p\Big(\frac{\sigma}{d^{3/4}}\log^{3/2} d\Big). \tag{16}$$

By (16), if the entry $|A(j_1, j_2, j_3)| \gg \frac{\sigma \log^{3/2} d}{d^{3/4}}$, then the entry $\tilde{A}(j_1, j_2, j_3)$ maintains the same sign as $A(j_1, j_2, j_3)$. In Section 4 and Remark 13, we show that the sup-norm bound of $\tilde{\mathbf{A}} - \mathbf{A}$ is useful for the exact support recovery of sub-tensor localizations, under minimal signal strength requirements (that is the support size).

4. Applications

In this section, we review two applications of ℓ_∞-norm perturbation bound. In these applications, we note that it is unnecessary to estimate the bias b_k. We show that the sup-norm perturbation bounds reveal unconventional phase transitions in these statistical learning applications. Meanwhile, novel yet simple statistical algorithms can be designed based on the sup-norm perturbation bounds.

4.1. High Dimensional Clustering

Many statistical and machine learning tasks are associated with clustering high dimensional data, see McCallum et al. (2000), Parsons et al. (2004), Fan and Fan (2008), Hastie et al. (2009), Friedman (1989) and references therein. We consider a two-class Gaussian mixture model such that each data point $\mathbf{y}_i \in \mathbb{R}^p$ can be represented by

$$\mathbf{y}_i = -\ell_i \boldsymbol{\beta} + (1-\ell_i)\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \in \mathbb{R}^p \tag{17}$$

where the associated label $\ell_i \in \{0,1\}$ for $i = 1,2,\ldots,n$ is unknown and the noise vector $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. The vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is unknown with $p \gg n$. We denote the true clusters by

$$\mathcal{N}_0 := \{1 \le i \le n : \ell_i = 0\} \quad \text{and} \quad \mathcal{N}_1 := \{1 \le i \le n : \ell_i = 1\}.$$

Given the data matrix

$$\mathbf{Y} = \big(\mathbf{y}_1, \ldots, \mathbf{y}_n\big)^\top \in \mathbb{R}^{n\times p}$$

, our goal is to cluster the n data points into two disjoint groups. Let $n_{k+1} := \mathrm{Card}\big(\mathcal{N}_k\big)$ for $k = 0,1$ such that $n_1 + n_2 = n$. Observe that $\mathbb{E}\mathbf{Y}$ has rank 1 and its leading left singular vector $\mathbf{u} \in \mathbb{R}^n$ with

$$u(i) = \frac{1-\ell_i}{n^{1/2}} - \frac{\ell_i}{n^{1/2}}, \quad 1 \le i \le n.$$

The signs of $\mathbf{u}$ immediately suggest the cluster memberships of each data points. Moreover, the leading singular value of $\mathbb{E}\mathbf{Y}$ is $n^{1/2}\|\boldsymbol{\beta}\|_{\ell_2}$. Let $\hat{\mathbf{u}}$ denote the leading left singular vector of $\mathbf{Y}$. By Corollary 2, if $\|\boldsymbol{\beta}\|_{\ell_2} \ge D_1\big(1 \vee (p/n)^{1/4}\big)$ such that $|(1+b_k)^{-1/2} - 1| \le 1/2$, then there exists an event $\mathcal{E}$ with $\mathbb{P}(\mathcal{E}) \ge 1 - \frac{1}{n}$ so that on event $\mathcal{E}$,

$$\big\|\hat{\mathbf{u}} - (1+b_k)^{1/2}\mathbf{u}\big\|_{\ell_\infty} \le D_2\Big(\frac{1}{\|\boldsymbol{\beta}\|_{\ell_2}} + \frac{(p/n)^{1/2}}{\|\boldsymbol{\beta}\|_{\ell_2}^2}\Big)\Big(\frac{1}{\|\boldsymbol{\beta}\|_{\ell_2}^2} + \sqrt{\frac{\log n}{n}}\Big).$$

On event $\mathcal{E}$, if $\|\boldsymbol{\beta}\|_{\ell_2} \ge D_1\Big(n^{1/6} \vee p^{1/8} \vee \big(p\log(n)/n\big)^{1/4}\Big)$, then we get

$$\begin{aligned}\|\hat{\mathbf{u}} - \mathbf{u}\|_{\ell_\infty} &\le \|\hat{\mathbf{u}} - (1+b_k)^{1/2}\mathbf{u}\|_{\ell_\infty} + \big|(1+b_k)^{-1/2} - 1\big|\|\mathbf{u}\|_{\ell_\infty} \\ &\le \|\hat{\mathbf{u}} - (1+b_k)^{1/2}\mathbf{u}\|_{\ell_\infty} + \frac{1}{2n^{1/2}} \le \frac{3}{4n^{1/2}}\end{aligned} \tag{18}$$

implying that if $\ell_i = \ell_j$, then $\mathrm{sign}\big(\hat{u}(i)\big) = \mathrm{sign}\big(\hat{u}(j)\big)$ for all $1 \le i,j \le n$. Therefore, we propose a simple clustering algorithm by entry-wise signs of $\hat{\mathbf{u}}$ in Algorithm 1.

By the bound (18), Algorithm 1 can guarantee exact clustering as follows.

Algorithm 1 High dimensional bi-clustering by entry-wise signs.

Input: Data matrix $\mathbf{Y} \in \mathbb{R}^{n\times p}$
Calculate the leading left singular vector of $\mathbf{Y}$, denoted by $\hat{\mathbf{u}} \in \mathbb{R}^n$
Initiate $\hat{\mathcal{N}}_0 = \{\}$ and $\hat{\mathcal{N}}_1 = \{\}$
for $i = 1, \cdots, n$ **do**
  if $\hat{u}(i) \geq 0$ **then**
    $\hat{\mathcal{N}}_0 \leftarrow \hat{\mathcal{N}}_0 \cup \{i\}$
  else
    $\hat{\mathcal{N}}_1 \leftarrow \hat{\mathcal{N}}_1 \cup \{i\}$
  end if
end for
Output: $\hat{\mathcal{N}}_0$ and $\hat{\mathcal{N}}_1$.

Theorem 8 *Suppose model (17) holds with noise vector $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}_p)$. Let $\hat{\mathcal{N}}_0$ and $\hat{\mathcal{N}}_1$ be the output of Algorithm 1. There exists an absolute constant $D_1 > 0$ such that if $\|\boldsymbol{\beta}\|_{\ell_2} \geq D_1\Big(n^{1/6} \vee p^{1/8} \vee \big(p\log(n)/n\big)^{1/4}\Big)$, then with probability at least $1 - \frac{1}{n}$,*

$$\hat{\mathcal{N}}_0 = \mathcal{N}_0 \quad \text{or} \quad \hat{\mathcal{N}}_0 = \mathcal{N}_1.$$

The proof of Theorem 8 is straightforward based on eq. (18). We note that eq. (18) also implies that it is unnecessary to estimate b_k in this application, since scaling switch the entry-wise signs simultaneously and thus maintains the clustering outputs.

Remark 9 *Theorem 8 reveal unconventional phase transition thresholds for the exact clustering of Gaussian mixture model (17). Indeed, by Theorem 8, the sup-norm based clustering algorithm (Algorithm 1) will exactly recover the memberships with high probability when the signal strength satisfies*

$$\|\boldsymbol{\beta}\|_{\ell_2} \gg \Big(n^{1/6} \vee p^{1/8} \vee \big(p\log(n)/n\big)^{1/4}\Big).$$

In comparison, the ℓ_2-norm based clustering algorithm in Cai and Zhang (2016) and Zheng and Tomioka (2015) requires

$$\|\boldsymbol{\beta}\|_{\ell_2} \gg \big(n^{1/2} \vee p^{1/4}\big)$$

for exact clustering. Clearly, with respect to exact recovery, the sup-norm based clustering algorithm requires much weaker SNR conditions.

Remark 10 *The above framework can be directly generalized to Gaussian mixture model with k-clusters. Suppose that the j-th cluster has mean vector $\boldsymbol{\beta}_j$ and size n_j, then without loss of generality, the data matrix $\mathbf{Y} = \mathbf{M} + \mathbf{Z}$*

$$\mathbf{M} = \big(\underbrace{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_1}_{n_1}, \cdots, \underbrace{\boldsymbol{\beta}_j, \cdots, \boldsymbol{\beta}_j}_{n_j}, \cdots, \underbrace{\boldsymbol{\beta}_k, \cdots, \boldsymbol{\beta}_k}_{n_k}\big)^\top \in \mathbb{R}^{N\times p}$$

with $N = \sum_{j=1}^k n_j$ and $\mathbf{Z} \in \mathbb{R}^{N\times p}$ having i.i.d. standard Gaussian entries. Observe that $\text{rank}(\mathbf{M}) \leq k$*, it suffices to consider the top-k left singular vectors of $\mathbf{M}$. However, it* *requires nontrivial effort to investigate the eigengaps of* $\mathbf{M}$ *without further assumptions on* $\{\boldsymbol{\beta}_j\}_{j=1}^k$. *In the case that* $n_j = n$ *and* $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k$ *are mutually orthogonal such that* $\|\boldsymbol{\beta}_1\|_{\ell_2} \geq \ldots \geq \|\boldsymbol{\beta}_k\|_{\ell_2}$, *then* $\mathbf{M}$*'s top-k singular values are* $\lambda_j = \sqrt{n_j}\|\boldsymbol{\beta}_j\|_{\ell_2}, 1 \leq j \leq k$. *Clearly, the non-zero entries of* $\mathbf{M}$*'s top-k left singular vectors provide the cluster membership of each data points. By Theorem 1, if* $\Delta_j \geq C_1\sqrt{k}\|\boldsymbol{\beta}_1\|_{\ell_2} + C_2(kp/n)^{1/2}$ *where* $\Delta_j = \min\{(\|\boldsymbol{\beta}_j\|_{\ell_2}^2 - \|\boldsymbol{\beta}_{j+1}\|_{\ell_2}^2), (\|\boldsymbol{\beta}_{j-1}\|_{\ell_2}^2 - \|\boldsymbol{\beta}_j\|_{\ell_2}^2)\}$, *then*

$$\|\hat{\mathbf{u}}_j - \sqrt{1+b_j}\mathbf{u}_j\|_{\ell_\infty} = O_p\Big(\Big(\frac{\|\boldsymbol{\beta}_1\|_{\ell_2}}{\Delta_j} + \frac{(p/n)^{1/2}}{\Delta_j}\Big)\Big(\frac{k^{3/2}}{\Delta_j} + \sqrt{\frac{k\log n}{n}}\Big)\Big)$$

for all $1 \leq j \leq k$.

4.2. Sub-tensor Localization

In gene expression association analysis (see Hore et al. (2016), Xiong et al. (2012), Kolar et al. (2011) and Ben-Dor et al. (2003)) and planted clique detection (see Brubaker and Vempala (2009), Anandkumar et al. (2013) and Gauvin et al. (2014)), the goal is equivalent to locating a sub-tensor whose entries are statistically more significant than the others. One simple model characterizing this type of tensor data is as

$$\mathbf{Y} = \lambda \mathbf{1}_{C_1} \otimes \mathbf{1}_{C_2} \otimes \mathbf{1}_{C_3} + \mathbf{Z} \in \mathbb{R}^{d_1\times d_2\times d_3} \tag{19}$$

with $C_k = \cup_{j=1}^{s_k} C_k^{(j)} \subset [d_k]$ where $\{C_k^{(1)}, \ldots, C_k^{(s_k)}\}$ are disjoint subsets of $[d_k]$ for $k = 1, 2, 3$, i.e., there are $s_k \geq 1$ dense blocks in the k-th direction. Then, in total, there are $s_1s_2s_3$ dense blocks in $\mathbb{E}\mathbf{Y}$. The vector $\mathbf{1}_{C_k} \in \mathbb{R}^{d_k}$ is a zero-or-one vector whose entry equals 1 only when the index belongs to $\mathbf{C}_k$. The noise tensor $\mathbf{Z}$ has i.i.d. entries such that $Z(i,j,k) \sim \mathcal{N}(0,1)$. Given the noisy observation $\mathbf{Y}$, the goal is to locate the unknown subsets $\{C_1^{(j)}\}_{j=1}^{s_1}, \{C_2^{(j)}\}_{j=1}^{s_2}$ and $\{C_3^{(j)}\}_{j=1}^{s_3}$. The appealing scenario is $\lambda = O(1)$, since otherwise the signal is so strong that the problem can be easily solved by just looking at each entry. The tensor $\mathbb{E}\mathbf{Y}$ has rank 1 with leading singular value $\lambda|C_1|^{1/2}|C_2|^{1/2}|C_3|^{1/2}$ and corresponding singular vectors

$$\mathbf{u} = \frac{1}{|C_1|^{1/2}}\mathbf{1}_{C_1}, \quad \mathbf{v} = \frac{1}{|C_2|^{1/2}}\mathbf{1}_{C_2} \quad \text{and} \quad \mathbf{w} = \frac{1}{|C_3|^{1/2}}\mathbf{1}_{C_3},$$

where $|C|$ denotes the cardinality of C. By Theorem 1, if $\lambda \geq D_1\frac{(d_1d_2d_3)^{1/4}}{|C_1|^{1/2}|C_2|^{1/2}|C_3|^{1/2}}$ for a large enough constant $D_1 > 0$ and $d_{\max} \leq (d_1d_2d_3)^{1/2}$ where $d_{\max} := (d_1 \vee d_2 \vee d_3)$, then with probability at least $1 - \frac{1}{d_{\max}}$, we obtain

$$\begin{aligned}&\|\hat{\mathbf{u}} - (1+b_1)^{1/2}\mathbf{u}\|_{\ell_\infty}\\ &\quad\leq \frac{D_1\log^{1/2} d_{\max}}{\lambda|C_1|^{1/2}|C_2|^{1/2}|C_3|^{1/2}} + \frac{D_1(d_2d_3\log d_{\max})^{1/2}}{\lambda^2|C_1||C_2||C_3|} + \frac{D_1d_1}{\lambda^2|C_1||C_2||C_3|}\Big(\frac{(d_1d_2d_3)^{1/2}}{\lambda^2|C_1||C_2||C_3|}\Big),\end{aligned} \tag{20}$$

where $b_1 \in [-0.5, 0]$ is a constant depending on $\mathbf{u}, \mathbf{v}, \mathbf{w}$ and λ only. Similar bounds can be also derived for $\hat{\mathbf{v}}$ and $\hat{\mathbf{w}}$. By eq. (20), we propose a simple algorithm (Algorithm 2) for the

Algorithm 2 Sub-tensor localizations by entry-wise magnitudes.

Input: Data matrix $\mathbf{Y} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$

Calculate the leading left singular vectors of $\{\mathcal{M}_k(\mathbf{Y})\}_{k=1}^3$, denoted by $\hat{\mathbf{u}} \in \mathbb{R}^{d_1}, \hat{\mathbf{v}} \in \mathbb{R}^{d_2}$ and $\hat{\mathbf{w}} \in \mathbb{R}^{d_3}$, respectively.

Take entry-wise magnitudes $\{|\hat{u}(j_1)|\}_{j_1=1}^{d_1}$ and arrange them in a non-increasing order,

Record the top-$|C_1|$ locations and denote them by $\hat{C}_1$;

Take entry-wise magnitudes $\{|\hat{v}(j_2)|\}_{j_2=1}^{d_2}$ and arrange them in a non-increasing order,

Record the top-$|C_2|$ locations and denote them by $\hat{C}_2$;

Take entry-wise magnitudes $\{|\hat{w}(j_3)|\}_{j_3=1}^{d_3}$ and arrange them in a non-increasing order,

Record the top-$|C_3|$ locations and denote them by $\hat{C}_3$;

Output: $\hat{C}_1$, $\hat{C}_2$ and $\hat{C}_3$.

support recovery of sub-tensor model (19). By bound (20), we can immediately guarantee the exact support recovery by Algorithm 2. The proof is straightforward and is omitted here.

Theorem 11 *Suppose model (19) holds and $(d_1 + d_2 + d_3) \leq 2(d_1 d_2 d_3)^{1/2}$. There exist absolute constants $D_1, D_2 > 0$ such that if $\lambda \geq D_1 \frac{(d_1 d_2 d_3)^{1/4}}{(|C_1||C_2||C_3|)^{1/2}}$ and*

$$\max\left\{\sqrt{\frac{|C_1|}{d_1}}, \sqrt{\frac{|C_2|}{d_2}}, \sqrt{\frac{|C_3|}{d_3}}\right\} \cdot \frac{(d_1 d_2 d_3 \log d_{\max})^{1/2}}{\lambda^2 |C_1||C_2||C_3|} \leq \frac{1}{D_2},$$

then, with probability at least $1 - \frac{1}{d_1+d_2+d_3}$, we get

$$\hat{C}_1 = C_1 \quad \text{and} \quad \hat{C}_2 = C_2 \quad \text{and} \quad \hat{C}_3 = C_3$$

where $\{\hat{C}_k\}_{k=1}^3$ are the output of Algorithm 2.

Note that in Algorithm 2 and Theorem 11, it is also unnecessary to estimate the bias b_1 because we are interested in the top-$|C_1|$ largest entries of $|\hat{\mathbf{u}}|$ and scaling does not affect the ordering of the entry-wise magnitudes.

Remark 12 *The phase transition of Algorithm 2 and model (19) is intriguing. Note that the support localizations are trivial when $\lambda \gg 1$. Therefore, we only focus on the case $\lambda = 1$. Now, let $|C_1| \asymp |C_2| \asymp |C_3| = K$ and $d_1 \asymp d_2 \asymp d_3 = d$. By Theorem 11, we conclude that Algorithm 2 can exactly recover the supports C_1, C_2, C_3 with high probability if the support size $K \gg d^{\frac{1}{2}}$. Meanwhile, by the lower bound arguments in Zhang and Xia (2018), we know that if $K \ll d^{\frac{1}{2}}$, then there exist no polynomial time algorithms which can recover C_1 consistently. Put it differently, phase transition occurs at the threshold $O(d^{\frac{1}{2}})$ such that if $K \ll d^{\frac{1}{2}}$, the problem is unsolvable by polynomial time algorithms; if $K \gg d^{\frac{1}{2}}$, the problem can be perfectly solved by Algorithm 2. In comparison, the ℓ_2-norm based algorithms can only guarantee the consistency of support recovery when $K \gg d^{\frac{1}{2}}$, rather than the exact recovery.*

Remark 13 *We could also investigate the entry-wise denoising of model (19). Suppose that* $|C_1| \asymp |C_2| \asymp |C_3| = K$ *and* $d_1 \asymp d_2 \asymp d_3 = d$. *We denote by* $\mathbf{A} = \mathbf{1}_{C_1} \otimes \mathbf{1}_{C_2} \otimes \mathbf{1}_{C_3}$ *where we fix* $\lambda = 1$ *and we focus only on the support sizes* $\{|C_k|\}_{k=1}^3$. *Let* $\hat{\mathbf{u}}, \hat{\mathbf{v}}$ *and* $\hat{\mathbf{w}}$ *be the empirical singular vectors as in Algorithm 2. Define the projection estimator*

$$\hat{\mathbf{A}} = \mathbf{Y} \times_1 (\hat{\mathbf{u}}\hat{\mathbf{u}}^\top) \times_2 (\hat{\mathbf{v}}\hat{\mathbf{v}}^\top) \times_3 (\hat{\mathbf{w}}\hat{\mathbf{w}}^\top).$$

Similarly as in Theorem 6, we can show that there exists a constant $b \in [\sqrt{2}/4, 1]$ *such that with probability at least* $1 - \frac{1}{d}$,

$$\|\hat{\mathbf{A}} - b \cdot \mathbf{A}\|_{\ell_\infty} \leq D_1 \cdot \Big(\frac{1}{K} + \frac{d}{K^{5/2}}\Big) \log^{3/2} d \tag{21}$$

for some absolute constant $D_1 > 0$. *Recall from model (19) that* $A(j_1, j_2, j_3) = 1$ *if* $(j_1, j_2, j_3) \in C_1 \times C_2 \times C_3$. *From eq. (21), we conclude that if* $K \geq D_2\big(\sqrt{d} + d^{0.4} \log^{0.6} d\big)$ *for a large enough absolute constant* $D_2 > 0$ *(note that the threshold* $\sqrt{d}$ *comes from SNR requirement as in eq. (20)), then*

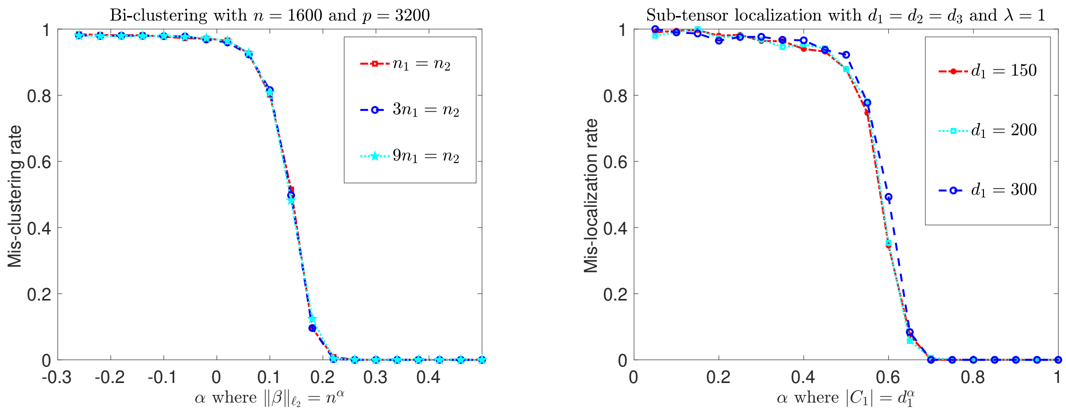
$$\big|\hat{A}(j_1, j_2, j_3)\big| > |\hat{A}(j_1', j_2', j_3')|$$

for all $(j_1, j_2, j_3) \in C_1 \times C_2 \times C_3$ *and* $(j_1', j_2', j_3') \notin C_1 \times C_2 \times C_3$. *As a result, we can choose the locations of* $\hat{\mathbf{A}}$*'s entries with the largest-*$|C_1||C_2||C_3|$ *magnitudes and recover* $\mathbf{A}$*'s supports exactly.*

4.3. Numerical Experiments

We present simulation results of experiments for the applications in Section 4. For high dimensional clustering in model (17), we randomly sample a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ with $p = 3200$. For a fixed $\boldsymbol{\beta}$, we sample $n_1 = n/2 = 800$ random vectors from distribution $\mathcal{N}(\boldsymbol{\beta}, \mathbf{I}_p)$ and $n_2 = n/2 = 800$ random vectors from distribution $\mathcal{N}(-\boldsymbol{\beta}, \mathbf{I}_p)$. Then, we calculate the top left singular vector of $\mathbf{Y}$ as in (17) and apply Algorithm 1 to cluster the 1600 points into two disjoint groups. For each $\boldsymbol{\beta}$, we repeat the experiments for 50 times and the average mis-clustering rate is recorded. The signal strengths are chosen so that $\|\boldsymbol{\beta}\|_{\ell_2} = n^\alpha$ with $\alpha = 0.06 * k - 0.5$ for $1 \leq k \leq 20$. The average mis-clustering rates with respect to signal strengths are displayed in Figure (1a). Moreover, in Figure (1a), we also compare the average mis-clustering rates when two clusters have different sizes such as $3n_1 = n_2 = 1200$ and $9n_1 = n_2 = 1440$. As shown in Figure (1a), there exists a threshold around $\alpha = 0.18$ such that the mis-clustering rates by Algorithm 1 decreases extremely fast when the signal strength exceeds the threshold. Meanwhile, Figure (1a) also shows that the size balances of two clusters does not affect the threshold. Both these numerical observations from Figure (1a) are consistent with the theoretic guarantees from Theorem 8.

For sub-tensor localizations in model (19), we fix $\lambda = 1$ because the support localization task is trivial if $\lambda \gg 1$. Similarly as in Remark 12, it then suffices to investigate the efficiency of Algorithm 2 with respect to the support sizes. For simplicity, we choose $d_1 = d_2 = d_3$ and $C_1 = C_2 = C_3 = [|C_1|]$, that is, the sub-tensor is in the bottom-left-front corner of $\mathbb{E}\mathbf{Y}$. For each $d_1 = 150, d_1 = 200$ and $d_1 = 300$, we show the average mis-localization rates by Algorithm 2 with respect to the support size $|C_1|$. The average mis-localization rates are



(a) *Mis-clustering rates of Algorithm 1* **(b)** *Mis-localization rates of Algorithm 2*

Figure 1: *Simulation results for the performances of Algorithm 1 and Algorithm 2. In Figure (1a), the phase transition happens around the signal strength $\|\boldsymbol{\beta}\|_{\ell_2} \approx n^{0.18}$ which coincides with Theorem 8. Figure (1a) shows that Algorithm 1 can exactly recover the true clusters when signal strength exceeds the aforementioned threshold. Figure (1a) also shows that the efficiency of Algorithm 1 is unaffected when two clusters have unbalanced sizes. In Figure (1b), the phase transition happens when the support C_1 has size around $d_1^{0.6}$. It shows that Algorithm 2 can exactly locate the sub-tensor when the support size exceeds the aforementioned threshold.*

calculated from 50 independent experiments. The support sizes are chosen as $|C_1| = \lceil d_1^{\alpha} \rceil$ with $0.06 \leq \alpha \leq 1$. The results of mis-localization rates are displayed in Figure (1b). Indeed, Figure (1b) shows that the mis-localization rates by Algorithm 2 starts to decrease extremely fast when the support size is around $|C_1| \asymp d_1^{0.6}$. The exponent 0.6 is somewhat larger than the threshold 0.5 claimed in Remark 12. Note that the dimension size d is moderately large (only 300) in our simulations due to the heavy computational cost.

5. Acknowledgements

The authors would like to thank the action editor and three anonymous referees for their insightful suggestions which significantly improve the readability of the article.

6. Proofs

For notational brevity, we write $A \lesssim B$ if there exists an absolute constant D_1 such that $A \leq D_1 B$. A similar notation would be $\gtrsim$ and $A \asymp B$ means that $A \lesssim B$ and $A \gtrsim B$ simultaneously. If the constant D_1 depends on some parameter γ, we shall write $\lesssim_\gamma, \gtrsim_\gamma$ and $\asymp_\gamma$.

Recall that the HOSVD is translated directly from SVD on $\mathcal{M}_1(\mathbf{A})$ and the matrix perturbation model $\mathcal{M}_1(\mathbf{Y}) = \mathcal{M}_1(\mathbf{A}) + \mathcal{M}_1(\mathbf{Z})$. Without loss of generality, it suffices to focus on matrices with unbalanced sizes. In the remaining context, we write $\mathbf{A}, \mathbf{Z}, \mathbf{Y} \in \mathbb{R}^{m_1 \times m_2}$ instead of $\mathcal{M}_1(\mathbf{A}), \mathcal{M}_1(\mathbf{Z}), \mathcal{M}_1(\mathbf{Y}) \in \mathbb{R}^{m_1 \times m_2}$, where $m_1 = d_1$ and $m_2 = d_2 d_3$ such that $m_1 \ll m_2$. The second order spectral analysis begins with

$$\mathbf{Y}\mathbf{Y}^\top = \mathbf{A}\mathbf{A}^\top + \mathbf{\Gamma}, \quad \text{where} \quad \mathbf{\Gamma} = \mathbf{A}\mathbf{Z}^\top + \mathbf{Z}\mathbf{A}^\top + \mathbf{Z}\mathbf{Z}^\top.$$

Suppose that $\mathbf{A}$ has the thin singular value decomposition

$$\mathbf{A} = \sum_{k=1}^{r_1} \lambda_k \big(\mathbf{u}_k \otimes \mathbf{h}_k\big) \in \mathbb{R}^{m_1 \times m_2}$$

where $\{\mathbf{h}_1, \ldots, \mathbf{h}_{r_1}\} \subset \text{span}\big\{\mathbf{v}_j \otimes \mathbf{w}_k^\top : j \in [r_2], k \in [r_3]\big\}$ are the right singular vectors of $\mathbf{A}$. Moreover, $\mathbf{A}\mathbf{A}^\top$ admits the eigen-decomposition:

$$\mathbf{A}\mathbf{A}^\top = \sum_{k=1}^{r_1} \lambda_k^2 \big(\mathbf{u}_k \otimes \mathbf{u}_k\big).$$

In an identical fashion, denote the eigen-decomposition of $\mathbf{Y}\mathbf{Y}^\top$ by

$$\mathbf{Y}\mathbf{Y}^\top = \sum_{k=1}^{m_1} \hat{\lambda}_k^2 \big(\hat{\mathbf{u}}_k \otimes \hat{\mathbf{u}}_k\big).$$

Even though Theorem 1 and Theorem 4 are stated when the singular value λ_k has multiplicity 1, we present more general results in this section. Note that when there are repeated singular values, the singular vectors are not uniquely defined. In this case, let $\mu_1 > \mu_2 > \ldots > \mu_s > 0$ be distinct singular values of $\mathbf{A}$ with $s \le r_1$. Denote $\Delta_k := \{j : \lambda_j = \mu_k\}$ for $1 \le k \le s$ and $\nu_k := \text{Card}(\Delta_k)$ the multiplicity of μ_k. Let $\mu_{s+1} = 0$ which is a trivial eigenvalue of $\mathbf{A}\mathbf{A}^\top$ with multiplicity $m_1 - r_1$. Then, the spectral decomposition of $\mathbf{A}\mathbf{A}^\top$ can be represented as

$$\mathbf{A}\mathbf{A}^\top = \sum_{k=1}^{s+1} \mu_k^2 \mathbf{P}_k^{uu}$$

where the spectral projector $\mathbf{P}_k^{uu} := \sum_{j \in \Delta_k} \mathbf{u}_j \otimes \mathbf{u}_j$ which is uniquely defined. Correspondingly, define the empirical spectral projector based on eigen-decomposition of $\mathbf{Y}\mathbf{Y}^\top$,

$$\hat{\mathbf{P}}_k^{uu} := \sum_{j \in \Delta_k} \hat{\mathbf{u}}_j \otimes \hat{\mathbf{u}}_j.$$

We develop a sharp concentration bound for bilinear forms $\big\langle \hat{\mathbf{P}}_k^{uu} \mathbf{x}, \mathbf{y} \big\rangle$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m_1}$. Observe that $\mathbf{Y}\mathbf{Y}^\top$ has an identical eigen-space as $\mathbf{Y}\mathbf{Y}^\top - m_2\sigma^2 \mathbf{I}_{m_1}$. Let $\hat{\mathbf{\Gamma}} := \mathbf{\Gamma} - m_2\sigma^2\mathbf{I}_{m_1}$ and the spectral analysis shall be realized on $\mathbf{A}\mathbf{A}^\top + \hat{\mathbf{\Gamma}}$.

Several preliminary facts are introduced as follows. It is clear that the k-th eigengap is $\bar{g}_k\big(\mathbf{A}\mathbf{A}^\top\big) := \min\big(\mu_{k-1}^2 - \mu_k^2, \mu_k^2 - \mu_{k+1}^2\big)$ for $1 \le k \le s$, where we set $\mu_0 = +\infty$. The proof of Lemma 14 is provided in the Appendix.

Lemma 14 *For any deterministic matrix $\mathbf{B} \in \mathbb{R}^{m_3 \times m_2}$, the following bounds hold*

$$\mathbb{E}\|\mathbf{B}\mathbf{Z}^\top\| \lesssim \sigma\|\mathbf{B}\|\Big(m_1^{1/2} + m_3^{1/2} + (m_1 m_3)^{1/4}\Big) \tag{22}$$

$$\big\|\mathbb{E}\mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1}\big\| \lesssim \sigma^2(m_1 m_2)^{1/2}.$$

For any $t > 0$, the following inequalities hold with probability at least $1 - e^{-t}$,

$$\begin{aligned} \|\mathbf{B}\mathbf{Z}^\top\| &\lesssim \sigma\|\mathbf{B}\|\Big(m_1^{1/2} + m_3^{1/2} + (m_1 m_3)^{1/4} + t^{1/2} + (m_1 t)^{1/4}\Big) \\ \big\|\mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1}\big\| &\lesssim \sigma^2 m_2^{1/2}\big(m_1^{1/2} + t^{1/2}\big). \end{aligned} \tag{23}$$

6.1. Proof of Theorem 1

To this end, define

$$\mathbf{C}_k^{uu} := \sum_{s\neq k} \frac{1}{\mu_s^2 - \mu_k^2}\mathbf{P}_s^{uu}$$

and

$$\mathbf{P}_k^{hh} := \sum_{j\in\Delta_k} \mathbf{h}_j \otimes \mathbf{h}_j.$$

Theorem 1 is decomposed into two separate parts. Theorem 15 provides the concentration bound for $\big|\langle \mathbf{P}_k\mathbf{x}, \mathbf{y}\rangle - \mathbb{E}\langle \mathbf{P}_k\mathbf{x}, \mathbf{y}\rangle\big|$ by Gaussian isoperimetric inequality and the proof is postponed to the Appendix. In Theorem 17, we characterize the bias $\mathbb{E}\hat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu}$.

Theorem 15 *Let $\delta(m_1, m_2) := \mu_1\sigma m_1^{1/2} + \sigma^2(m_1 m_2)^{1/2}$ and suppose that $\bar{g}_k\big(\mathbf{A}\mathbf{A}^\top\big) \geq D_1\delta(m_1, m_2)$ for a large enough constant $D_1 > 0$. Then, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m_1}$, there exists an absolute constant $D_2 > 0$ such that for all $\log 8 \leq t \lesssim m_1$, the following bound holds with probability at least $1 - e^{-t}$,*

$$\big|\langle \hat{\mathbf{P}}_k^{uu}\mathbf{x}, \mathbf{y}\rangle - \mathbb{E}\langle \hat{\mathbf{P}}_k^{uu}\mathbf{x}, \mathbf{y}\rangle\big| \leq D_2 t^{1/2}\bigg(\frac{\sigma\mu_1 + \sigma^2 m_2^{1/2}}{\bar{g}_k\big(\mathbf{A}\mathbf{A}^\top\big)}\bigg)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.$$

The following spectral representation formula is needed whose proof can be found in Koltchinskii and Lounici (2016).

Lemma 16 *The following bound holds*

$$\|\hat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu}\| \leq \frac{4\|\hat{\mathbf{\Gamma}}\|}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}.$$

Moreover, $\hat{\mathbf{P}}_k^{uu}$ can be represented as

$$\hat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu} = \mathbf{L}_k(\hat{\mathbf{\Gamma}}) + \mathbf{S}_k(\hat{\mathbf{\Gamma}})$$

where $\mathbf{L}_k(\hat{\mathbf{\Gamma}}) = \mathbf{P}_k^{uu}\hat{\mathbf{\Gamma}}\mathbf{C}_k^{uu} + \mathbf{C}_k^{uu}\hat{\mathbf{\Gamma}}\mathbf{P}_k^{uu}$ and

$$\|\mathbf{S}_k(\hat{\mathbf{\Gamma}})\| \leq 14\bigg(\frac{\|\hat{\mathbf{\Gamma}}\|}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\bigg)^2.$$

Theorem 17 *Let $\delta(m_1, m_2) := \mu_1\sigma m_1^{1/2} + \sigma^2(m_1 m_2)^{1/2}$ and suppose that $\bar{g}_k\big(\mathbf{A}\mathbf{A}^\top\big) \geq D_1\delta(m_1, m_2)$ for a large enough constant $D_1 > 0$ and $m_2 e^{-m_1/2} \leq 1$. Then there exists an absolute constant $D_2 > 0$ such that*

$$\big\|\mathbb{E}\hat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu} - \mathbf{P}_k^{uu}\big(\mathbb{E}\hat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu}\big)\mathbf{P}_k^{uu}\big\| \leq D_2\nu_k\frac{\sigma^2 m_1 + \sigma^2 m_2^{1/2} + \sigma\mu_1}{\bar{g}_k\big(\mathbf{A}\mathbf{A}^\top\big)}\bigg(\frac{\delta(m_1, m_2)}{\bar{g}_k\big(\mathbf{A}\mathbf{A}^\top\big)}\bigg).$$

Proof [Proof of Theorem 1] Combining Theorem 15 and Theorem 17, we conclude that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m_1}$ with probability at least $1 - e^{-t}$ for all $\log 8 \le t \le m_1$,

$$\begin{aligned}\big|\langle \hat{\mathbf{P}}_k^{uu}\mathbf{x}, \mathbf{y}\rangle - &\langle \mathbf{P}_k^{uu}\mathbf{x}, \mathbf{y}\rangle - \langle \mathbf{P}_k^{uu}(\mathbb{E}\hat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu})\mathbf{P}_k^{uu}\mathbf{x}, \mathbf{y}\rangle\big| \\ &\lesssim \left(t^{1/2}\frac{\sigma\mu_1 + \sigma^2 m_2^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)} + \frac{\sigma^2 m_1 \delta(m_1, m_2)}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\right)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}\end{aligned}$$

where we used the fact $\frac{\delta(m_1,m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)} \le 1$ and $\nu_k = 1$ (since $\bar{g}_k(\mathbf{M}\mathbf{M}^\top) > 0$). Since $\nu_k = 1$ such that $\mathbf{P}_k^{uu} = \mathbf{u}_k \otimes \mathbf{u}_k$ and $\hat{\mathbf{P}}_k^{uu} = \hat{\mathbf{u}}_k \otimes \hat{\mathbf{u}}_k$, we can write

$$\mathbf{P}_k^{uu}(\mathbb{E}\hat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu})\mathbf{P}_k^{uu} = b_k \mathbf{P}_k^{uu}$$

where

$$b_k = \mathbb{E}\langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle^2 - 1 \in [-1, 0].$$

Moreover, a simple fact is $b_k \le \mathbb{E}\|\hat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu}\| \lesssim \frac{\delta(m_1,m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}$ by Wedin's $\sin\Theta$ theorem (Wedin (1972)). If $\bar{g}_k(\mathbf{A}\mathbf{A}^\top) \ge D\delta(m_1, m_2)$ for a large enough constant $D > 0$, we can ensure $b_k \in [-1/2, 0]$. Then, with probability at least $1 - e^{-t}$,

$$\big|\langle(\hat{\mathbf{P}}_k^{uu} - (1+b_k)\mathbf{P}_k^{uu})\mathbf{x}, \mathbf{y}\rangle\big| \lesssim \left(t^{1/2}\frac{\sigma\mu_1 + \sigma^2 m_2^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)} + \frac{\sigma^2 m_1 \delta(m_1, m_2)}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\right)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.$$

By choosing $\mathbf{x} = \mathbf{y} = \mathbf{u}_k$, we obtain for all $\log 8 \le t \le m_1$,

$$\mathbb{P}\left(\big|\langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle^2 - (1+b_k)\big| \gtrsim t^{1/2}\frac{\sigma\mu_1 + \sigma^2 m_2^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)} + \frac{\sigma^2 m_1 \delta(m_1, m_2)}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\right) \le e^{-t}.$$

Denote this event by $\mathcal{E}_1$. Observe that if the constant $D > 0$ is large enough and $m_1 \ll m_2$, we conclude that on event $\mathcal{E}_1$, $\langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle^2 \ge \frac{1}{4}$. Then, on event $\mathcal{E}_1$,

$$\begin{aligned}\big|\langle \hat{\mathbf{u}}_k, \mathbf{x}\rangle - &\sqrt{1+b_k}\langle \mathbf{u}_k, \mathbf{x}\rangle\big| \\ \le& \left|\frac{1+b_k}{\langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle} - \sqrt{1+b_k}\right||\langle \mathbf{u}_k, \mathbf{x}\rangle| \\ &+ \frac{1}{|\langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle|}\Big|\langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle\langle \hat{\mathbf{u}}_k, \mathbf{x}\rangle - (1+b_k)\langle \mathbf{u}_k, \mathbf{x}\rangle\Big| \\ =& \frac{\sqrt{1+b_k}\big|1+b_k - \langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle^2\big||\langle \mathbf{u}_k, \mathbf{x}\rangle|}{|\langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle|\big(\sqrt{1+b_k} + \langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle\big)} + \frac{1}{|\langle \hat{\mathbf{u}}_k, \mathbf{u}_k\rangle|}\big|\langle(\hat{\mathbf{P}}_k^{uu} - (1+b_k)\mathbf{P}_k^{uu})\mathbf{u}_k, \mathbf{x}\rangle\big| \\ \lesssim& t^{1/2}\frac{\sigma\mu_1 + \sigma^2 m_2^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2} + \frac{\sigma^2 m_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\left(\frac{\delta(m_1, m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\right)\|\mathbf{x}\|_{\ell_2},\end{aligned}$$

which concludes the proof after replacing $\mathbf{A}$ with $\mathcal{M}_1(\mathbf{A})$ and μ_1 with $\|\mathcal{M}_1(\mathbf{A})\|$. ∎

Proof [Proof of Theorem 17] Recall the representation formula of $\hat{\mathbf{P}}_k^{uu}$ in Lemma 16 that

$$\mathbb{E}\hat{\mathbf{P}}_k^{uu} = \mathbf{P}_k^{uu} + \mathbb{E}\mathbf{S}_k(\hat{\mathbf{\Gamma}})$$

where $\hat{\mathbf{\Gamma}} := \mathbf{A}\mathbf{Z}^\top + \mathbf{Z}\mathbf{A}^\top + \mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1}$. To this end, define

$$\widetilde{\mathbf{\Gamma}} := \hat{\mathbf{\Gamma}} - \big(\mathbf{Z}\mathbf{P}_k^{hh}\mathbf{Z}^\top - \nu_k\sigma^2\mathbf{I}_{m_1}\big)$$

such that we can write $\mathbb{E}\hat{\mathbf{P}}_k^{uu} = \mathbf{P}_k^{uu} + \mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) + \Big(\mathbb{E}\mathbf{S}_k(\hat{\mathbf{\Gamma}}) - \mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\Big)$. We derive an upper bound on $\big\|\mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbb{E}\mathbf{S}_k(\hat{\mathbf{\Gamma}})\big\|$ and the proof can be found in the Appendix. Lemma 18 implies that our analysis can be proceeded by replacing $\hat{\mathbf{\Gamma}}$ with $\widetilde{\mathbf{\Gamma}}$.

Lemma 18 *There exists a universal constant $D_1 > 0$ such that if $m_2 e^{-m_1/2} \le 1$, then*

$$\big\|\mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbb{E}\mathbf{S}_k(\hat{\mathbf{\Gamma}})\big\| \le D_1 \frac{\sigma\mu_1 + \sigma^2 m_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)} \bigg(\frac{\delta(m_1, m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\bigg).$$

Let $\delta_t = \mathbb{E}\|\hat{\mathbf{\Gamma}}\| + D_1\sigma\mu_1 t^{1/2} + D_2\sigma^2 m_2^{1/2} t^{1/2}$ for $0 < t \le m_1$ to be determined later and large enough constants $D_1, D_2 > 0$ such that $\mathbb{P}\big(\|\hat{\mathbf{\Gamma}}\| \ge \delta_t\big) \le e^{-t}$. We write

$$\begin{aligned}
&\mathbb{E}\hat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu} - \mathbf{P}_k^{uu}\mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu} \\
&\quad= \mathbb{E}\mathbf{S}_k(\hat{\mathbf{\Gamma}}) - \mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) \\
&\quad+ \mathbb{E}\Big(\mathbf{P}_k^{uu}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})(\mathbf{P}_k^{uu})^\perp + (\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu} + (\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})(\mathbf{P}_k^{uu})^\perp\Big)\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\| \le \delta_t\big) \\
&\quad+ \mathbb{E}\Big(\mathbf{P}_k^{uu}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})(\mathbf{P}_k^{uu})^\perp + (\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu} + (\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})(\mathbf{P}_k^{uu})^\perp\Big)\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\| > \delta_t\big). \quad (24)
\end{aligned}$$

We prove an upper bound for $\mathbb{E}\big\langle\mathbf{x}, (\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu}\mathbf{y}\big\rangle\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\| \le \delta_t\big)$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m_1}$. Similar to the approach in Koltchinskii and Xia (2016), under the assumption $\|\widetilde{\mathbf{\Gamma}}\| \le \delta_t$, $\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})$ is represented in the following analytic form,

$$\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) = -\frac{1}{2\pi i}\oint_{\gamma_k}\sum_{r\ge 2}(-1)^r\Big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\Big)^r\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)d\eta$$

where γ_k is a circle on the complex plane with center μ_k^2 and radius $\frac{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}{2}$, and $\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)$ is the resolvent of the operator $\mathbf{A}\mathbf{A}^\top$ with $\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta) = (\mathbf{A}\mathbf{A}^\top - \eta\mathbf{I}_{m_1})^{-1}$ which can be explicitly written as

$$\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta) := (\mathbf{A}\mathbf{A}^\top - \eta\mathbf{I}_{m_1})^{-1} = \sum_s \frac{1}{\mu_s^2 - \eta}\mathbf{P}_s^{uu}.$$

We also denote

$$\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta) := \mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta) - \frac{1}{\mu_k^2 - \eta}\mathbf{P}_k^{uu} = \sum_{s\ne k}\frac{1}{\mu_s^2 - \eta}\mathbf{P}_s^{uu}.$$

It is easy to check that

$$\begin{aligned}
&(\mathbf{P}_k^{uu})^\perp\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^r\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{P}_k^{uu} \\
&\quad= (\mathbf{P}_k^{uu})^\perp\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^r\frac{1}{\mu_k^2 - \eta}\mathbf{P}_k^{uu} \\
&\quad= \bigg(\frac{1}{(\mu_k^2 - \eta)^2}\sum_{s=2}^r\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-1}\big(\mathbf{P}_k^{uu}\widetilde{\mathbf{\Gamma}}\big)\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\bigg) \\
&\quad+ \frac{1}{\mu_k^2 - \eta}\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^r\mathbf{P}_k^{uu},
\end{aligned}$$

where we used the formula $(a+b)^r = b^r + \sum_{s=1}^r b^{s-1}a(a+b)^{r-s}$. As a result,

$$
\begin{aligned}
&(\mathbf{P}_k^{uu})^{\perp}\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})\mathbf{P}_k^{uu} \\
&= -\sum_{r\geq 2}(-1)^r\frac{1}{2\pi i}\oint_{\gamma_k}\bigg(\frac{1}{(\mu_k^2-\eta)^2}\sum_{s=2}^r\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{s-1}\big(\mathbf{P}_k^{uu}\widetilde{\boldsymbol{\Gamma}}\big)\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu} \\
&\quad+\frac{1}{\mu_k^2-\eta}\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^r\mathbf{P}_k^{uu}\bigg)d\eta.
\end{aligned}
\tag{25}
$$

For any $\mathbf{x},\mathbf{y}\in\mathbb{R}^{m_1}$, we shall derive an upper bound for

$$
\mathbb{E}\Big\langle\mathbf{x},\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{s-1}\big(\mathbf{P}_k^{uu}\widetilde{\boldsymbol{\Gamma}}\big)\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y}\Big\rangle\mathbf{1}\big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\big),\quad s=2,\ldots,r.
$$

Recall that $\text{rank}(\mathbf{P}_k^{uu})=\nu_k$ and $\mathbf{P}_k^{uu}=\sum_{j\in\Delta_k}\mathbf{u}_j\otimes\mathbf{u}_j$. Then,

$$
\begin{aligned}
&\Big\langle\mathbf{x},\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{s-1}\big(\mathbf{P}_k^{uu}\widetilde{\boldsymbol{\Gamma}}\big)\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y}\Big\rangle \\
&\qquad=\sum_{j\in\Delta_k}\Big\langle\mathbf{x},\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{s-1}\big(\mathbf{u}_j\otimes\mathbf{u}_j\widetilde{\boldsymbol{\Gamma}}\big)\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y}\Big\rangle \\
&\qquad=\sum_{j\in\Delta_k}\big\langle\widetilde{\boldsymbol{\Gamma}}\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y},\mathbf{u}_j\big\rangle\big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\mathbf{u}_j,\mathbf{x}\big\rangle.
\end{aligned}
$$

Observe that

$$
\begin{aligned}
\big|\big\langle\widetilde{\boldsymbol{\Gamma}}\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y},\mathbf{u}_j\big\rangle\big| &\leq\|\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\|^{r-s}\|\widetilde{\boldsymbol{\Gamma}}\|^{r-s+1}\|\mathbf{y}\|_{\ell_2} \\
&\leq\Big(\frac{2}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{(r-s)}\|\widetilde{\boldsymbol{\Gamma}}\|^{r-s+1}\|\mathbf{y}\|_{\ell_2}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}\Big\langle\mathbf{x},\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{s-1}\big(\mathbf{P}_k^{uu}\widetilde{\boldsymbol{\Gamma}}\big)\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y}\Big\rangle\mathbf{1}\Big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\Big) \\
&=\sum_{j\in\Delta_k}\mathbb{E}\big\langle\widetilde{\boldsymbol{\Gamma}}\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y},\mathbf{u}_j\big\rangle\big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{s-1}\mathbf{u}_j,\mathbf{x}\big\rangle\mathbf{1}\Big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\Big) \\
&\leq\sum_{j\in\Delta_k}\mathbb{E}^{1/2}\Big|\big\langle\widetilde{\boldsymbol{\Gamma}}\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y},\mathbf{u}_j\big\rangle\mathbf{1}\Big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\Big)\Big|^2 \\
&\qquad\qquad\times\mathbb{E}^{1/2}\Big|\big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{s-1}\mathbf{u}_j,\mathbf{x}\big\rangle\mathbf{1}\Big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\Big)\Big|^2 \\
&\leq\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{r-s}\delta_t\|\mathbf{y}\|_{\ell_2}\sum_{j\in\Delta_k}\mathbb{E}^{1/2}\Big|\big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\mathbf{u}_j,\mathbf{x}\big\rangle\mathbf{1}\Big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\Big)\Big|^2.
\end{aligned}
\tag{26}
$$

It then remains to bound, for each $j\in\Delta_k$,

$$
\mathbb{E}^{1/2}\Big|\Big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\mathbf{u}_j,\mathbf{x}\Big\rangle\Big|^2\mathbf{1}\Big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\Big).
$$

Recall that we can write

$$\widetilde{\mathbf{\Gamma}} = \mathbf{A}\mathbf{Z}^\top + \mathbf{Z}\mathbf{A}^\top + \mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\mathbf{Z}^\top - \sigma^2(m_2-\nu_k)\mathbf{I}_{m_1}$$

and correspondingly

$$\widetilde{\mathbf{\Gamma}}\mathbf{u}_j = \mathbf{A}\mathbf{Z}^\top\mathbf{u}_j + \mathbf{Z}\mathbf{A}^\top\mathbf{u}_j + \mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\mathbf{Z}^\top\mathbf{u}_j - \sigma^2(m_2-\nu_k)\mathbf{u}_j.$$

We write

$$\begin{aligned}
&\Big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\mathbf{u}_j, \mathbf{x}\Big\rangle\\
&\quad=\Big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{Z}\mathbf{A}^\top\mathbf{u}_j, \mathbf{x}\Big\rangle && (27)\\
&\qquad+\Big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{A}\mathbf{Z}^\top\mathbf{u}_j, \mathbf{x}\Big\rangle && (28)\\
&\qquad+\Big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\Big(\mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\mathbf{Z}^\top\mathbf{u}_j - \sigma^2(m_2-\nu_k)\mathbf{u}_j\Big), \mathbf{x}\Big\rangle. && (29)
\end{aligned}$$

The upper bounds of (27), (28), and (29) shall be obtained separately via different representations.

Bound of $\mathbb{E}^{1/2}\big|\big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{Z}\mathbf{A}^\top\mathbf{u}_j, \mathbf{x}\big\rangle\big|^2\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\|\le\delta_t\Big)$**.** Observe that $\mathbf{A}^\top\mathbf{u}_j = \mu_k\mathbf{h}_j\in\mathbb{R}^{m_2}$ for $j\in\Delta_k$ such that

$$\mathbf{Z}\mathbf{A}^\top\mathbf{u}_j = \mu_k\mathbf{Z}\mathbf{h}_j = \mu_k\sum_{i=1}^{m_1}\langle\mathbf{z}_i,\mathbf{h}_j\rangle\mathbf{e}_i$$

where $\{\mathbf{e}_1,\dots,\mathbf{e}_{m_1}\}$ denote the canonical basis vectors in $\mathbb{R}^{m_1}$ and $\{\mathbf{z}_1^\top,\dots,\mathbf{z}_{m_1}^\top\}$ denote the rows of $\mathbf{Z}$. Therefore,

$$\begin{aligned}
&\big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{Z}\mathbf{A}^\top\mathbf{u}_j, \mathbf{x}\big\rangle\\
&\qquad\qquad=\mu_k\sum_{i=1}^{m_1}\langle\mathbf{z}_i,\mathbf{h}_j\rangle\big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{e}_i, \mathbf{x}\big\rangle.
\end{aligned}$$

It is clear that $\langle\mathbf{z}_i,\mathbf{h}_j\rangle, i=1,\dots,m_1$ are i.i.d. and $\langle\mathbf{z}_i,\mathbf{h}_j\rangle\sim\mathcal{N}(0,\sigma^2)$. Recall that $\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)=\sum_{k'\neq k}\frac{\mathbf{P}_{k'}^{uu}}{\mu_{k'}^2-\eta}$, implying that $\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)$ can be viewed as a linear combination of operators

$$(\mathbf{P}_{t_1}^{uu}\widetilde{\mathbf{\Gamma}}\mathbf{P}_{t_2}^{uu})(\mathbf{P}_{t_2}^{uu}\widetilde{\mathbf{\Gamma}}\mathbf{P}_{t_3}^{uu})\dots(\mathbf{P}_{t_{s-2}}^{uu}\widetilde{\mathbf{\Gamma}}\mathbf{P}_{t_{s-1}}^{uu})$$

where $t_1,\dots,t_{s-1}\neq k$. For each $\mathbf{P}_{t_1}^{uu}\widetilde{\mathbf{\Gamma}}\mathbf{P}_{t_2}^{uu}$, we have

$$\mathbf{P}_{t_1}^{uu}\widetilde{\mathbf{\Gamma}}\mathbf{P}_{t_2}^{uu} = \mathbf{P}_{t_1}^{uu}\mathbf{A}\mathbf{Z}^\top\mathbf{P}_{t_2}^{uu} + \mathbf{P}_{t_1}^{uu}\mathbf{Z}\mathbf{A}^\top\mathbf{P}_{t_2}^{uu} + \mathbf{P}_{t_1}^{uu}\big(\mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\mathbf{Z}^\top\big)\mathbf{P}_{t_2}^{uu} - \sigma^2(m_2-\nu_k)\mathbf{P}_{t_1}^{uu}\mathbf{P}_{t_2}^{uu}.$$

Clearly, $\mathbf{P}_{t_1}^{uu}\mathbf{A}\mathbf{Z}^\top$ is a function of random vectors $\mathbf{P}_{t_1}^{uu}\mathbf{A}\mathbf{z}_i, i=1,\ldots,m_1$; $\mathbf{Z}\mathbf{A}^\top\mathbf{P}_{t_2}^{uu}$ is a function of random vectors $\mathbf{P}_{t_2}^{uu}\mathbf{A}\mathbf{z}_i, i=1,\ldots,m_1$; $\mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\mathbf{Z}^\top=\mathbf{Z}\sum_{k'\neq k}(\mathbf{P}_{k'}^{hh})^2\mathbf{Z}^\top$ is a function of random vectors $\mathbf{P}_{k'}^{hh}\mathbf{z}_i, i=1,\ldots,m_1$. The following facts are obvious

$$\mathbb{E}\langle\mathbf{z}_i,\mathbf{h}_j\rangle\mathbf{P}_{t_1}^{uu}\mathbf{A}\mathbf{z}_i=\mathbf{P}_{t_1}^{uu}\mathbf{A}(\mathbb{E}\mathbf{z}_i\otimes\mathbf{z}_i)\mathbf{h}_j=\sigma^2\mathbf{P}_{t_1}^{uu}\mathbf{A}\mathbf{h}_j=\sigma^2\mu_k\mathbf{P}_{t_1}^{uu}\mathbf{u}_j=\mathbf{0},\quad\forall t_1\neq k$$

and

$$\mathbb{E}\langle\mathbf{z}_i,\mathbf{h}_j\rangle\mathbf{P}_{k'}^{hh}\mathbf{z}_i=\mathbf{P}_{k'}^{hh}(\mathbb{E}\mathbf{z}_i\otimes\mathbf{z}_i)\mathbf{h}_j=\sigma^2\mathbf{P}_{k'}^{hh}\mathbf{h}_j=\mathbf{0},\quad\forall k'\neq k.$$

Since $\{\langle\mathbf{z}_i,\mathbf{h}_j\rangle, i=1,\ldots,m_1\}$ are Gaussian random variables and $\{\mathbf{P}_{t_1}^{uu}\mathbf{A}\mathbf{z}_i,\mathbf{P}_{k'}^{hh}\mathbf{z}_i, i=1,\ldots,m_1\}$ are (complex) Gaussian random vectors, uncorrelations indicate that $\{\langle\mathbf{z}_i,\mathbf{h}_j\rangle: i=1,\ldots,m_1\}$ are independent with $\{\mathbf{P}_{t_1}^{uu}\mathbf{A}\mathbf{z}_i,\mathbf{P}_{k'}^{hh}\mathbf{z}_i: t_1\neq k, k'\neq k, i=1,\ldots,m_1\}$. We conclude that $\{\langle\mathbf{z}_i,\mathbf{h}_j\rangle: i=1,\ldots,m_1\}$ are independent with

$$\Big\{\Big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{e}_i,\mathbf{x}\Big\rangle, i=1,\ldots,m_1\Big\}.$$

To this end, define the complex random variables

$$\omega_i(\mathbf{x})=\Big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{e}_i,\mathbf{x}\Big\rangle=\omega_i^{(1)}(\mathbf{x})+\omega_i^{(2)}(\mathbf{x})\mathrm{Im}\in\mathbb{C},\quad i=1,\ldots,m_1$$

where Im denotes the imaginary number. Then,

$$\begin{aligned}&\Big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{Z}\mathbf{A}^\top\mathbf{u}_j,\mathbf{x}\Big\rangle\\&\qquad=\mu_k\sum_{i=1}^{m_1}\langle\mathbf{z}_i,\mathbf{h}_j\rangle\omega_i^{(1)}(\mathbf{x})+\Big(\mu_k\sum_{i=1}^{m_1}\langle\mathbf{z}_i,\mathbf{h}_j\rangle\omega_i^{(2)}(\mathbf{x})\Big)\mathrm{Im}\\&\qquad=:\kappa_1(\mathbf{x})+\kappa_2(\mathbf{x})\mathrm{Im}\in\mathbb{C}.\end{aligned}$$

Conditioned on $\{\mathbf{P}_{t_1}^{uu}\mathbf{A}\mathbf{z}_i,\mathbf{P}_{k'}^{hh}\mathbf{z}_i: t_1\neq k, k'\neq k, i=1,\ldots,m_1\}$, we get

$$\mathbb{E}\kappa_1^2(\mathbf{x})=\mu_k^2\sigma^2\sum_{i=1}^{m_1}\Big(\omega_i^{(1)}(\mathbf{x})\Big)^2$$

and

$$\mathbb{E}\kappa_1(\mathbf{x})\kappa_2(\mathbf{x})=\mu_k^2\sigma^2\sum_{i=1}^{m_1}\omega_i^{(1)}(\mathbf{x})\omega_i^{(2)}(\mathbf{x})$$

implying that the centered Gaussian random vector $(\kappa_1(\mathbf{x}),\kappa_2(\mathbf{x}))$ has covariance matrix:

$$\Big(\mu_k^2\sigma^2\sum_{i=1}^{m_1}\omega_i^{(k_1)}(\mathbf{x})\omega_i^{(k_2)}(\mathbf{x})\Big)_{k_1,k_2=1,2}.$$

Finally,

$$\begin{aligned}&\mathbb{E}^{1/2}\Big|\Big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{Z}\mathbf{A}^\top\mathbf{u}_j,\mathbf{x}\Big\rangle\Big|^2\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\|\leq\delta_t\Big)\\&=\mathbb{E}^{1/2}\big(\kappa_1^2(\mathbf{x})+\kappa_2^2(\mathbf{x})\big)\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\|\leq\delta_t\Big)\\&=\sigma\mu_k\mathbb{E}^{1/2}\Big(\sum_{i=1}^{m_1}\big(\omega_i^{(1)}(\mathbf{x})\big)^2+\big(\omega_i^{(2)}(\mathbf{x})\big)^2\Big)\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\|\leq\delta_t\Big)\\&=\sigma\mu_k\mathbb{E}^{1/2}\sum_{i=1}^{m_1}\big|\omega_i(\mathbf{x})\big|^2\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\|\leq\delta_t\Big).\end{aligned}$$

Moreover,

$$\sum_{i=1}^{m_1}\big|\omega_i(\mathbf{x})\big|^2=\sum_{i=1}^{m_1}\big|\big\langle \widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\mathbf{x},\mathbf{e}_j\big\rangle\big|^2\le \big\|\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\mathbf{x}\big\|_{\ell_2}^2$$
$$\le \|\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\|^{2(s-1)}\|\widetilde{\mathbf{\Gamma}}\|^{2(s-2)}\|\mathbf{x}\|_{\ell_2}^2\le\Big(\frac{2}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{2(s-1)}\|\widetilde{\mathbf{\Gamma}}\|^{2(s-2)}\|\mathbf{x}\|_{\ell_2}^2.$$

As a result,

$$\mathbb{E}^{1/2}\big|\big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{Z}\mathbf{A}^\top\mathbf{u}_j,\mathbf{x}\big\rangle\big|^2\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\|\le\delta_t\Big)$$
$$\le\sigma\mu_k\mathbb{E}^{1/2}\Big(\frac{2}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{2(s-1)}\|\widetilde{\mathbf{\Gamma}}\|^{2(s-2)}\|\mathbf{x}\|_{\ell_2}^2\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\|\le\delta_t\Big)$$
$$\le\frac{\sigma\mu_k}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{s-2}\|\mathbf{x}\|_{\ell_2}.$$

Bound of $\mathbb{E}^{1/2}\big|\big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{A}\mathbf{Z}^\top\mathbf{u}_j,\mathbf{x}\big\rangle\big|^2\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\|\le\delta_t\big)$**.** With a little abuse on the notations, we denote by $\mathbf{z}_1,\ldots,\mathbf{z}_{m_2}\in\mathbb{R}^{m_1}$ the corresponding columns of $\mathbf{Z}$ in this paragraph. Then,

$$\big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{A}\mathbf{Z}^\top\mathbf{u}_j,\mathbf{x}\big\rangle=\sum_{i=1}^{m_2}\langle\mathbf{z}_i,\mathbf{u}_j\rangle\Big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{A}\mathbf{e}_i,\mathbf{x}\Big\rangle.$$

Similarly, $\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)$ can be represented as linear combination of operators

$$\big(\mathbf{P}_{t_1}^{uu}\widetilde{\mathbf{\Gamma}}\mathbf{P}_{t_2}^{uu}\big)\big(\mathbf{P}_{t_2}^{uu}\widetilde{\mathbf{\Gamma}}\mathbf{P}_{t_3}^{uu}\big)\ldots\big(\mathbf{P}_{t_{s-2}}^{uu}\widetilde{\mathbf{\Gamma}}\mathbf{P}_{t_{s-1}}^{uu}\big),\quad t_1,\ldots,t_{s-1}\neq k.$$

To this end, we write

$$\mathbf{P}_{t_1}^{uu}\widetilde{\mathbf{\Gamma}}\mathbf{P}_{t_2}^{uu}=\mathbf{P}_{t_1}^{uu}\mathbf{A}\mathbf{Z}^\top\mathbf{P}_{t_2}^{uu}+\mathbf{P}_{t_1}^{uu}\mathbf{Z}\mathbf{A}^\top\mathbf{P}_{t_2}^{uu}+\mathbf{P}_{t_1}^{uu}\big(\mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\mathbf{Z}^\top\big)\mathbf{P}_{t_2}^{uu}-\sigma^2(m_2-\nu_k)\mathbf{P}_{t_1}^{uu}\mathbf{P}_{t_2}^{uu}.$$

Observe that $\mathbf{P}_{t_1}^{uu}\mathbf{A}\mathbf{Z}^\top\mathbf{P}_{t_2}^{uu}$, $\mathbf{P}_{t_1}^{uu}\mathbf{Z}\mathbf{A}^\top\mathbf{P}_{t_2}^{uu}$ and $\mathbf{P}_{t_1}^{uu}\big(\mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\mathbf{Z}^\top\big)\mathbf{P}_{t_2}^{uu}$ are functions of random vectors $\{\mathbf{P}_{t_1}^{uu}\mathbf{z}_i,\mathbf{P}_{t_2}^{uu}\mathbf{z}_i: t_1,t_2\neq k, i=1,\ldots,m_2\}$. Moreover,

$$\mathbb{E}\langle\mathbf{z}_i,\mathbf{u}_j\rangle\mathbf{P}_{t_1}^{uu}\mathbf{z}_i=\mathbf{P}_{t_1}^{uu}\big(\mathbb{E}\mathbf{z}_i\otimes\mathbf{z}_i\big)\mathbf{u}_j=\sigma^2\mathbf{P}_{t_1}^{uu}\mathbf{u}_j=\mathbf{0},\quad\forall\ t_1\neq k$$

which implies that $\{\langle\mathbf{z}_i,\mathbf{u}_j\rangle: i=1,\ldots,m_2\}$ and $\Big\{\big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{A}\mathbf{e}_i,\mathbf{x}\big\rangle: i=1,\ldots,m_2\Big\}$ are independent. Following an identical analysis as above, we get

$$\mathbb{E}^{1/2}\big|\big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{A}\mathbf{Z}^\top\mathbf{u}_j,\mathbf{x}\big\rangle\big|^2\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\|\le\delta_t\big)\le\frac{\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{s-2}\|\mathbf{x}\|_{\ell_2}.$$

Bound of $\mathbb{E}^{1/2}\Big|\Big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\big(\mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\mathbf{Z}^\top\big)\mathbf{u}_j, \mathbf{x}\Big\rangle\Big|^2\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\|\leq \delta_t\big)$**.** Note that we used the fact $\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{u}_j=\mathbf{0}$ in (29). Again, let $\{\mathbf{z}_1,\ldots,\mathbf{z}_{m_2}\}\subset\mathbb{R}^{m_1}$ denote the corresponding columns of $\mathbf{Z}$. We write

$$\begin{aligned}
&\big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\big(\mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\mathbf{Z}^\top\big)\mathbf{u}_j, \mathbf{x}\big\rangle\\
&\qquad\qquad=\sum_{i=1}^{m_2}\langle \mathbf{z}_i,\mathbf{u}_j\rangle\big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{Z}\big(\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\big)\mathbf{e}_i, \mathbf{x}\big\rangle.
\end{aligned}$$

In a similar fashion, we show that $\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{Z}$ is a function of random vectors $\big\{\mathbf{P}_t^{uu}\mathbf{z}_i: t\neq k, i=1,\ldots,m_2\big\}$ which are independent with $\big\{\langle \mathbf{z}_i,\mathbf{u}_j\rangle: i=1,\ldots,m_2\big\}$. Then,

$$\begin{aligned}
&\mathbb{E}^{1/2}\Big|\Big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\big(\mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\mathbf{Z}^\top\big)\mathbf{u}_j, \mathbf{x}\Big\rangle\Big|^2\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\|\leq\delta\big)\\
&\qquad\leq \mathbb{E}^{1/2}\sigma^2\|\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\|^{2(s-1)}\|\widetilde{\mathbf{\Gamma}}\|^{2(s-2)}\|\mathbf{Z}\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\|^2\|\mathbf{x}\|_{\ell_2}^2\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\|\leq\delta_t\big)\\
&\qquad\lesssim \frac{\sigma^2 m_2^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{s-2}\|\mathbf{x}\|_{\ell_2}.
\end{aligned}$$

where we used the fact $\mathbb{E}^{1/2}\big\|\big(\sum_{k'\neq k}\mathbf{P}_{k'}^{hh}\big)\mathbf{Z}^\top\big\|^2\lesssim \sigma m_2^{1/2}$ from Lemma 14.

Finalize the proof of Theorem. Combining the above bounds into (28), (27) and (29), we conclude that

$$\begin{aligned}
&\mathbb{E}^{1/2}\Big|\big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\mathbf{u}_j, \mathbf{x}\big\rangle\Big|^2\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\|\leq\delta_t\big)\\
&\qquad\lesssim \frac{\sigma^2 m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{s-2}\|\mathbf{x}\|_{\ell_2}.
\end{aligned}$$

Continue from (26) and we end up with

$$\begin{aligned}
&\mathbb{E}\big\langle \mathbf{x}, \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-1}\big(\mathbf{P}_k^{uu}\widetilde{\mathbf{\Gamma}}\big)\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y}\big\rangle\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\|\leq\delta_t\big)\\
&\qquad\lesssim \nu_k\delta_t\frac{\sigma^2 m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{r-2}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.
\end{aligned}$$

Plug the bounds into (25),

$$\begin{aligned}
&\big|\mathbb{E}\big\langle (\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu}\mathbf{y}, \mathbf{x}\big\rangle\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\leq\delta_t\|\big)\big|\\
&\quad\lesssim \sum_{r\geq 2}\frac{\pi\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}{2\pi}\Big(\frac{2}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^2(r-1)\nu_k\delta_t\frac{\sigma^2 m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{r-2}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}\\
&\quad\leq D_1\nu_k\frac{\sigma^2 m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}\sum_{r\geq 2}(r-1)\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{r-1}
\end{aligned}$$

where we used the fact $\oint_{\gamma_k}\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\boldsymbol{\Gamma}}\big)^r\mathbf{P}_k^{uu}d\eta = \mathbf{0}$. By the inequality $\sum_{r\geq 1} rq^r = \frac{q}{(1-q)^2}, \forall q<1$ and the fact $D_1\delta_t \leq \bar{g}_k(\mathbf{A}\mathbf{A}^\top)$ for some large constant $D_1>0$ and $t\leq m_1$, we conclude with

$$\begin{aligned}&\big|\mathbb{E}\big\langle(\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})\mathbf{P}_k^{uu}\mathbf{y},\mathbf{x}\big\rangle\mathbf{1}\big(\|\widetilde{\boldsymbol{\Gamma}}\leq\delta_t\|\big)\big|\\&\qquad\lesssim \nu_k\frac{\sigma^2 m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2},\quad\forall \mathbf{x},\mathbf{y}\in\mathbb{R}^{m_1}\end{aligned}$$

implying that

$$\Big\|\mathbb{E}(\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})\mathbf{P}_k^{uu}\mathbf{1}\big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\big)\Big\|\lesssim\nu_k\frac{\sigma^2 m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big).$$

The same bound holds for

$$\big\|\mathbb{E}\mathbf{P}_k^{uu}\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})(\mathbf{P}_k^{uu})^\perp\mathbf{1}\big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\big)\big\|\quad\text{and}\quad\big\|\mathbb{E}(\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})(\mathbf{P}_k^{uu})^\perp\mathbf{1}\big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\big)\big\|,$$

following the same arguments. As a result,

$$\begin{aligned}&\Big\|\mathbb{E}\Big((\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})\mathbf{P}_k^{uu}+\mathbf{P}_k^{uu}\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})(\mathbf{P}_k^{uu})^\perp+(\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})(\mathbf{P}_k^{uu})^\perp\Big)\mathbf{1}\big(\|\widetilde{\boldsymbol{\Gamma}}\|\leq\delta_t\big)\Big\|\\&\qquad\lesssim\nu_k\frac{\sigma^2 m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big).\end{aligned}\tag{30}$$

By choosing $t=m_1$ such that $\mathbb{P}(\|\widetilde{\boldsymbol{\Gamma}}\|\geq\delta_{m_1})\leq e^{-m_1/2}$, we get

$$\begin{aligned}&\Big\|\mathbb{E}\Big((\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})\mathbf{P}_k^{uu}+\mathbf{P}_k^{uu}\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})(\mathbf{P}_k^{uu})^\perp+(\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})(\mathbf{P}_k^{uu})^\perp\Big)\mathbf{1}\big(\|\widetilde{\boldsymbol{\Gamma}}\|>\delta_{m_1}\big)\Big\|\\&\leq\mathbb{E}\Big\|\Big((\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})\mathbf{P}_k^{uu}+\mathbf{P}_k^{uu}\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})(\mathbf{P}_k^{uu})^\perp+(\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})(\mathbf{P}_k^{uu})^\perp\Big)\Big\|\mathbf{1}\big(\|\widetilde{\boldsymbol{\Gamma}}\|>\delta_{m_1}\big)\\&\leq\mathbb{E}\|\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})\|\mathbf{1}\big(\|\widetilde{\boldsymbol{\Gamma}}\|>\delta_{m_1}\big)\leq\mathbb{E}^{1/2}\|\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})\|^2\mathbb{P}^{1/2}\big(\|\widetilde{\boldsymbol{\Gamma}}\|>\delta_{m_1}\big)\\&\lesssim\Big(\frac{\delta_{m_1}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^2\mathbb{P}^{1/2}\big(\|\widetilde{\boldsymbol{\Gamma}}\|>\delta_{m_1}\big)\lesssim\Big(\frac{\delta_{m_1}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^2e^{-m_1/2},\end{aligned}$$

which is clearly dominated by (30). Substitute the above bounds into (24) and we get

$$\begin{aligned}\Big\|\mathbb{E}\hat{\mathbf{P}}_k^{uu}-\mathbf{P}_k^{uu}-\mathbf{P}_k^{uu}\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})\mathbf{P}_k^{uu}\Big\|&\leq\|\mathbb{E}\mathbf{S}_k(\widetilde{\boldsymbol{\Gamma}})-\mathbf{S}_k(\hat{\boldsymbol{\Gamma}})\|+D_1\nu_k\frac{\sigma^2 m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta(m_1,m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)\\&\leq D_2\nu_k\frac{\sigma^2 m_2^{1/2}+\sigma^2 m_1+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta(m_1,m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big).\end{aligned}$$

■

6.2. Proof of Theorem 4

The proof of Theorem 4 is identical to the proof of Corollary 1.5 in Koltchinskii and Xia (2016) and will be skipped here.

6.3. Proof of Theorem 6

It suffices to prove the upper bound of $\big|\widetilde{A}(i,j,k)-A(i,j,k)\big|$ for $i\in[d_1], j\in[d_2], k\in[d_3]$. To this end, denote by $\mathbf{e}_i$ the i-th canonical basis vectors. Observe that

$$\big\langle \widetilde{\mathbf{A}}-\mathbf{A},\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle = \Big\langle \mathbf{A}\times_1\mathbf{P}_{\widetilde{\mathbf{U}}}\times_2\mathbf{P}_{\widetilde{\mathbf{V}}}\times_3\mathbf{P}_{\widetilde{\mathbf{W}}}-\mathbf{A},\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\Big\rangle + \Big\langle \mathbf{Z}\times_1\mathbf{P}_{\widetilde{\mathbf{U}}}\times_2\mathbf{P}_{\widetilde{\mathbf{V}}}\times_3\mathbf{P}_{\widetilde{\mathbf{W}}},\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\Big\rangle.$$

Some preliminary facts shall be concluded from Theorem 1. By Theorem 4, there exists an event $\mathcal{E}_2$ with $\mathbb{P}(\mathcal{E}_2)\geq 1-\frac{1}{d^2}$ on which

$$\big\|\mathbf{e}_i^\top(\widetilde{\mathbf{U}}-\mathbf{U})\big\|_{\ell_2}\leq r^{1/2}\big\|\mathbf{e}_i^\top(\widetilde{\mathbf{U}}-\mathbf{U})\big\|_{\ell_\infty}\lesssim\frac{\sigma\overline{\Lambda}(\mathbf{A})r^{1/2}+\sigma^2dr^{1/2}}{\bar{g}_{\min}^2(\mathbf{A})}\log^{1/2}d$$

and

$$\big\|\widetilde{\mathbf{U}}^\top\mathbf{U}-\mathbf{I}_{r_1}\big\|\leq\|\widetilde{\mathbf{U}}^\top\mathbf{U}-\mathbf{I}_{r_1}\|_{\rm F}\lesssim r\|\widetilde{\mathbf{U}}^\top\mathbf{U}-\mathbf{I}_{r_1}\|_{\ell_\infty}\lesssim\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2dr}{\bar{g}_{\min}^2(\mathbf{A})}\log^{1/2}d.$$

The following decomposition is straightforward,

$$\begin{aligned}
&\mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}},\mathbf{P}_{\widetilde{\mathbf{V}}},\mathbf{P}_{\widetilde{\mathbf{W}}}\big)-\mathbf{A}\\
&\quad=\mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\mathbf{W}}\big)+\mathbf{A}\cdot\big(\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\widetilde{\mathbf{V}}}-\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\mathbf{W}}\big)\\
&\quad+\mathbf{A}\cdot\big(\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\widetilde{\mathbf{W}}}-\mathbf{P}_{\mathbf{W}}\big)+\mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\widetilde{\mathbf{V}}}-\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\mathbf{W}}\big)\\
&\quad+\mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\widetilde{\mathbf{W}}}-\mathbf{P}_{\mathbf{W}}\big)+\mathbf{A}\cdot\big(\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\widetilde{\mathbf{V}}}-\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\widetilde{\mathbf{W}}}-\mathbf{P}_{\mathbf{W}}\big)\\
&\quad+\mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\widetilde{\mathbf{V}}}-\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\widetilde{\mathbf{W}}}-\mathbf{P}_{\mathbf{W}}\big)
\end{aligned}$$

Recall that $\mathbf{A}=\mathbf{C}\cdot(\mathbf{U},\mathbf{V},\mathbf{W})$ and we get

$$\begin{aligned}
&\Big\langle\mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\mathbf{W}}\big),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\Big\rangle\\
&\qquad=\mathbf{e}_i^\top\Big(\widetilde{\mathbf{U}}\big(\widetilde{\mathbf{U}}^\top\mathbf{U}\big)-\mathbf{U}\Big)\mathcal{M}_1(\mathbf{C})\big(\mathbf{V}\otimes\mathbf{W}\big)^\top(\mathbf{e}_j\otimes\mathbf{e}_k).
\end{aligned}$$

Observe that

$$\mathbf{e}_i^\top\Big(\widetilde{\mathbf{U}}\big(\widetilde{\mathbf{U}}^\top\mathbf{U}\big)-\mathbf{U}\Big)=\mathbf{e}_i^\top\big(\widetilde{\mathbf{U}}-\mathbf{U}\big)\big(\widetilde{\mathbf{U}}^\top\mathbf{U}\big)+\mathbf{e}_i^\top\mathbf{U}\big(\widetilde{\mathbf{U}}^\top\mathbf{U}-\mathbf{I}_{r_1}\big)$$

implying that on event $\mathcal{E}_2$,

$$\begin{aligned}
&\Big\|\mathbf{e}_i^\top\Big(\widetilde{\mathbf{U}}\big(\widetilde{\mathbf{U}}^\top\mathbf{U}\big)-\mathbf{U}\Big)\Big\|_{\ell_2}\\
&\quad\leq\big\|(\widetilde{\mathbf{U}}-\mathbf{U})^\top\mathbf{e}_i\big\|_{\ell_2}\|\widetilde{\mathbf{U}}^\top\mathbf{U}\|+\big\|\widetilde{\mathbf{U}}^\top\mathbf{U}-\mathbf{I}_{r_1}\big\|\|\mathbf{U}^\top\mathbf{e}_i\|_{\ell_2}\\
&\quad\lesssim\frac{\sigma\overline{\Lambda}(\mathbf{A})r^{1/2}+\sigma^2dr^{1/2}}{\bar{g}_{\min}^2(\mathbf{A})}\log^{1/2}d+\|\mathbf{U}^\top\mathbf{e}_i\|_{\ell_2}\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2dr}{\bar{g}_{\min}^2(\mathbf{A})}\log^{1/2}d\\
&\quad\lesssim\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2dr}{\bar{g}_{\min}^2(\mathbf{A})}\log^{1/2}d,
\end{aligned}$$

where we used the facts $\|\widetilde{\mathbf{U}}^\top \mathbf{U}\| \le \|\widetilde{\mathbf{U}}\|\|\mathbf{U}\| \le (1+b_k)^{-1/2} = O(1)$ and

$$\|\mathbf{U}^\top \mathbf{e}_i\|_{\ell_2} = \big\langle \mathbf{U}\mathbf{U}^\top, \mathbf{e}_i \otimes \mathbf{e}_i \big\rangle^{1/2} \le 1.$$

Therefore, on event $\mathcal{E}_2$,

$$\begin{aligned}&\big|\big\langle \mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}} - \mathbf{P}_{\mathbf{U}}, \mathbf{P}_{\mathbf{V}}, \mathbf{P}_{\mathbf{W}}\big), \mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k \big\rangle\big| \\ &\qquad\lesssim \overline{\Lambda}(\mathbf{A})\Big(\frac{\sigma\overline{\Lambda}(\mathbf{A})r + \sigma^2 dr}{\bar{g}^2_{\min}(\mathbf{A})}\log^{1/2} d\Big)\|\mathbf{V}^\top \mathbf{e}_j\|_{\ell_2}\|\mathbf{W}^\top \mathbf{e}_k\|_{\ell_2}.\end{aligned}$$

Similar bounds hold for

$$\big|\big\langle \mathbf{A}\cdot\big(\mathbf{P}_{\mathbf{U}}, \mathbf{P}_{\widetilde{\mathbf{V}}} - \mathbf{P}_{\mathbf{V}}, \mathbf{P}_{\mathbf{W}}\big), \mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k \big\rangle\big| \quad \text{and} \quad \big|\big\langle \mathbf{A}\cdot\big(\mathbf{P}_{\mathbf{U}}, \mathbf{P}_{\mathbf{V}}, \mathbf{P}_{\widetilde{\mathbf{W}}} - \mathbf{P}_{\mathbf{W}}\big), \mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k \big\rangle\big|.$$

Following the same method, we can show that on event $\mathcal{E}_2$,

$$\begin{aligned}&\big|\big\langle \mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}} - \mathbf{P}_{\mathbf{U}}, \mathbf{P}_{\widetilde{\mathbf{V}}} - \mathbf{P}_{\mathbf{V}}, \mathbf{P}_{\mathbf{W}}\big), \mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k \big\rangle\big| \\ &\qquad\lesssim \overline{\Lambda}(\mathbf{A})\Big(\frac{\sigma\overline{\Lambda}(\mathbf{A})r + \sigma^2 dr}{\bar{g}^2_{\min}(\mathbf{A})}\log^{1/2} d\Big)^2\|\mathbf{W}^\top \mathbf{e}_k\|_{\ell_2}\end{aligned}$$

and

$$\begin{aligned}&\big|\big\langle \mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}} - \mathbf{P}_{\mathbf{U}}, \mathbf{P}_{\widetilde{\mathbf{V}}} - \mathbf{P}_{\mathbf{V}}, \mathbf{P}_{\widetilde{\mathbf{W}}} - \mathbf{P}_{\mathbf{W}}\big), \mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k \big\rangle\big| \\ &\qquad\lesssim \overline{\Lambda}(\mathbf{A})\Big(\frac{\sigma\overline{\Lambda}(\mathbf{A})r + \sigma^2 dr}{\bar{g}^2_{\min}(\mathbf{A})}\log^{1/2} d\Big)^3.\end{aligned}$$

We conclude that on event $\mathcal{E}_2$,

$$\begin{aligned}&\big|\big\langle \mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}}, \mathbf{P}_{\widetilde{\mathbf{V}}}, \mathbf{P}_{\widetilde{\mathbf{W}}}\big) - \mathbf{A}, \mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k \big\rangle\big| \\ &\lesssim \overline{\Lambda}(\mathbf{A})\Big(\frac{\sigma\overline{\Lambda}(\mathbf{A})r + \sigma^2 dr}{\bar{g}^2_{\min}(\mathbf{A})}\log^{1/2} d\Big)\Big(\|\mathbf{V}^\top \mathbf{e}_j\|_{\ell_2}\|\mathbf{W}^\top \mathbf{e}_k\|_{\ell_2} \\ &\quad + \|\mathbf{U}^\top \mathbf{e}_i\|_{\ell_2}\|\mathbf{W}^\top \mathbf{e}_k\|_{\ell_2} + \|\mathbf{U}^\top \mathbf{e}_i\|_{\ell_2}\|\mathbf{V}^\top \mathbf{e}_j\|_{\ell_2}\Big) \\ &\quad + \overline{\Lambda}(\mathbf{A})\Big(\frac{\sigma\overline{\Lambda}(\mathbf{A})r + \sigma^2 dr}{\bar{g}^2_{\min}(\mathbf{A})}\log^{1/2} d\Big)^2\Big(\|\mathbf{V}^\top \mathbf{e}_j\|_{\ell_2} + \|\mathbf{U}^\top \mathbf{e}_i\|_{\ell_2} + \|\mathbf{W}^\top \mathbf{e}_k\|_{\ell_2}\Big) \\ &\quad + \overline{\Lambda}(\mathbf{A})\Big(\frac{\sigma\overline{\Lambda}(\mathbf{A})r + \sigma^2 dr}{\bar{g}^2_{\min}(\mathbf{A})}\log^{1/2} d\Big)^3.\end{aligned}$$

Recall that for all $i \in [d_1], j \in [d_2], k \in [d_3]$

$$\|\mathbf{U}^\top \mathbf{e}_i\|_{\ell_2} \le \mu_{\mathbf{U}}\sqrt{\frac{r}{d}}, \quad \|\mathbf{V}^\top \mathbf{e}_j\|_{\ell_2} \le \mu_{\mathbf{V}}\sqrt{\frac{r}{d}}, \quad \|\mathbf{W}^\top \mathbf{e}_k\|_{\ell_2} \le \mu_{\mathbf{W}}\sqrt{\frac{r}{d}}$$

and conditions (13) (14) (15) imply

$$\frac{\sigma\overline{\Lambda}(\mathbf{A})r + \sigma^2 dr}{\bar{g}^2_{\min}(\mathbf{A})}\log^{1/2} d \lesssim r\Big(\frac{\log d}{d}\Big)^{1/2}.$$

We end up with a simpler bound on event $\mathcal{E}_2$,

$$\begin{aligned}\big|\big\langle \mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}},\mathbf{P}_{\widetilde{\mathbf{V}}},\mathbf{P}_{\widetilde{\mathbf{W}}}\big)-\mathbf{A},\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle\big| \\ \lesssim\sigma r^3\bigg(\frac{\sigma\widetilde{\kappa}(\mathbf{A})}{\bar{g}_{\min}(\mathbf{A})}+\frac{\widetilde{\kappa}^2(\mathbf{A})}{d}\bigg)(\mu_{\mathbf{U}}\mu_{\mathbf{V}}+\mu_{\mathbf{U}}\mu_{\mathbf{W}}+\mu_{\mathbf{V}}\mu_{\mathbf{W}})\log^{3/2}d\end{aligned} \tag{31}$$

where $\widetilde{\kappa}(\mathbf{A})=\overline{\Lambda}(\mathbf{A})/\bar{g}_{\min}(\mathbf{A})$.

Next, we prove the upper bound of $\big|\big\langle \mathbf{Z}\cdot(\mathbf{P}_{\widetilde{\mathbf{U}}},\mathbf{P}_{\widetilde{\mathbf{V}}},\mathbf{P}_{\widetilde{\mathbf{W}}}),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle\big|$ and we proceed with the same decomposition. Observe that

$$\begin{aligned}\big\langle \mathbf{Z}\cdot(\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\mathbf{W}}),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle =&\big\langle \mathbf{Z},(\mathbf{P}_{\mathbf{U}}\mathbf{e}_i)\otimes(\mathbf{P}_{\mathbf{V}}\mathbf{e}_j)\otimes(\mathbf{P}_{\mathbf{W}}\mathbf{e}_k)\big\rangle\\ \sim&\mathcal{N}\Big(0,\sigma^2\big\|\mathbf{P}_{\mathbf{U}}\mathbf{e}_i\big\|_{\ell_2}^2\big\|\mathbf{P}_{\mathbf{V}}\mathbf{e}_j\big\|_{\ell_2}^2\big\|\mathbf{P}_{\mathbf{W}}\mathbf{e}_k\big\|_{\ell_2}^2\Big)\end{aligned}$$

The standard concentration inequality of Gaussian random variables yields that with probability at least $1-\frac{1}{d^2}$,

$$\begin{aligned}\big|\big\langle \mathbf{Z}\cdot(\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\mathbf{W}}),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle\big| \lesssim&\sigma\|\mathbf{U}^\top\mathbf{e}_i\|_{\ell_2}\|\mathbf{V}^\top\mathbf{e}_j\|_{\ell_2}\|\mathbf{W}^\top\mathbf{e}_k\|_{\ell_2}\log^{1/2}d\\ \lesssim&\sigma\Big(\frac{r}{d}\Big)^{3/2}\mu_{\mathbf{U}}\mu_{\mathbf{V}}\mu_{\mathbf{W}}\log^{1/2}d.\end{aligned}$$

Similarly, with probability at least $1-\frac{1}{d^2}$,

$$\begin{aligned}\big|\big\langle \mathbf{Z}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\mathbf{W}}\big),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle\big|\\ =&\big|\mathbf{e}_i^\top\big(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}}\big)\mathcal{M}_1(\mathbf{Z})\big(\mathbf{V}\otimes\mathbf{W}\big)\big((\mathbf{V}^\top\mathbf{e}_j)\otimes(\mathbf{W}^\top\mathbf{e}_k)\big)\big|\\ \leq&\|(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}})\mathbf{e}_i\|_{\ell_2}\big\|\mathcal{M}_1(\mathbf{Z})(\mathbf{V}\otimes\mathbf{W})\big\|\|\mathbf{V}^\top\mathbf{e}_j\|_{\ell_2}\|\mathbf{W}^\top\mathbf{e}_k\|_{\ell_2}\\ \lesssim&\sigma d^{1/2}\|(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}})\mathbf{e}_i\|_{\ell_2}\big|\|\mathbf{V}^\top\mathbf{e}_j\|_{\ell_2}\|\mathbf{W}^\top\mathbf{e}_k\|_{\ell_2}\end{aligned}$$

where we used Lemma 14 for the upper bound of $\big\|\mathcal{M}_1(\mathbf{Z})(\mathbf{V}\otimes\mathbf{W})\big\|$. Moreover, since $\mu_{\mathbf{U}}\geq 1$,

$$\begin{aligned}\big\|\big(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}}\big)\mathbf{e}_i\big\|_{\ell_2}\leq&\|(\widetilde{\mathbf{U}}-\mathbf{U})\mathbf{e}_i\|_{\ell_2}+\|\widetilde{\mathbf{U}}-\mathbf{U}\|_{\ell_2}\|\mathbf{U}^\top\mathbf{e}_i\|_{\ell_2}\\ \lesssim&\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2dr}{\bar{g}_{\min}^2(\mathbf{A})}\mu_{\mathbf{U}}\log^{1/2}d.\end{aligned}$$

Denote the above event by $\mathcal{E}_3$. On $\mathcal{E}_2\cap\mathcal{E}_3$,

$$\big|\big\langle \mathbf{Z}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\mathbf{W}}\big),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle\big|\lesssim\frac{\sigma r}{d^{1/2}}\bigg(\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2dr}{\bar{g}_{\min}^2(\mathbf{A})}\bigg)\mu_{\mathbf{U}}\mu_{\mathbf{V}}\mu_{\mathbf{W}}\log^{1/2}d.$$

Similar bounds can be attained for

$$\big|\big\langle \mathbf{Z}\cdot\big(\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\widetilde{\mathbf{V}}}-\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\mathbf{W}}\big),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle\big|\quad\text{and}\quad\big|\big\langle \mathbf{Z}\cdot\big(\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\widetilde{\mathbf{W}}}-\mathbf{P}_{\mathbf{W}}\big),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle\big|.$$

In an identical fashion, on event $\mathcal{E}_2\cap\mathcal{E}_3$,

$$\begin{aligned}\big|\big\langle \mathbf{Z}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\widetilde{\mathbf{V}}}-\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\mathbf{W}}\big),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle\big|\\ \lesssim&\sigma r^{1/2}\bigg(\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2dr}{\bar{g}_{\min}^2(\mathbf{A})}\bigg)^2\mu_{\mathbf{U}}\mu_{\mathbf{V}}\mu_{\mathbf{W}}\log d.\end{aligned}$$

and

$$\begin{aligned}&\left|\left\langle \mathbf{Z}\cdot\left(\mathbf{P}_{\widetilde{\mathbf{U}}}-\mathbf{P}_{\mathbf{U}},\mathbf{P}_{\widetilde{\mathbf{V}}}-\mathbf{P}_{\mathbf{V}},\mathbf{P}_{\widetilde{\mathbf{W}}}-\mathbf{P}_{\mathbf{W}}\right),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\right\rangle\right|\\ &\qquad\lesssim\sigma d^{1/2}\left(\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2 dr}{\bar{g}_{\min}^2(\mathbf{A})}\right)^3\mu_{\mathbf{U}}\mu_{\mathbf{V}}\mu_{\mathbf{W}}\log^{3/2}d.\end{aligned}$$

Observe by conditions (13) (14) (15) that

$$\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2 dr}{\bar{g}_{\min}^2(\mathbf{A})}\lesssim\frac{r}{d^{1/2}}.$$

We conclude on event $\mathcal{E}_2\cap\mathcal{E}_3$ with

$$\left|\left\langle \mathbf{Z}\cdot\left(\mathbf{P}_{\widetilde{\mathbf{U}}},\mathbf{P}_{\widetilde{\mathbf{V}}},\mathbf{P}_{\widetilde{\mathbf{W}}}\right),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\right\rangle\right|\lesssim\frac{\sigma r^2}{d^{1/2}}\left(\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2 dr}{\bar{g}_{\min}^2(\mathbf{A})}\right)\mu_{\mathbf{U}}\mu_{\mathbf{V}}\mu_{\mathbf{W}}\log^{3/2}d. \tag{32}$$

By combining (31) and (32), we get on event $\mathcal{E}_2\cap\mathcal{E}_3$,

$$\begin{aligned}&\left|\left\langle \widetilde{\mathbf{A}}-\mathbf{A},\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\right\rangle\right|\\ &\quad\lesssim\sigma r^3\left(\frac{\sigma\widetilde{\kappa}(\mathbf{A})}{\bar{g}_{\min}(\mathbf{A})}+\frac{\widetilde{\kappa}^2(\mathbf{A})}{d}\right)(\mu_{\mathbf{U}}\mu_{\mathbf{V}}+\mu_{\mathbf{U}}\mu_{\mathbf{W}}+\mu_{\mathbf{V}}\mu_{\mathbf{W}})\log^{3/2}d\\ &\quad+\frac{\sigma r^2}{d^{1/2}}\left(\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2 dr}{\bar{g}_{\min}^2(\mathbf{A})}\right)\mu_{\mathbf{U}}\mu_{\mathbf{V}}\mu_{\mathbf{W}}\log^{3/2}d\\ &\quad\lesssim\sigma r^3\left(\frac{\sigma\widetilde{\kappa}(\mathbf{A})}{\bar{g}_{\min}(\mathbf{A})}+\frac{\widetilde{\kappa}^2(\mathbf{A})}{d}\right)(\mu_{\mathbf{U}}\mu_{\mathbf{V}}+\mu_{\mathbf{U}}\mu_{\mathbf{W}}+\mu_{\mathbf{V}}\mu_{\mathbf{W}})\log^{3/2}d,\end{aligned}$$

where the last inequality is due to fact $\bar{g}_{\min}(\mathbf{A})\gtrsim\sigma d^{3/4}$ and $\max\left\{\mu_{\mathbf{U}},\mu_{\mathbf{V}},\mu_{\mathbf{W}}\right\}\lesssim\sqrt{d}$.

References

Evrim Acar and Bülent Yener. Unsupervised multiway data analysis: A literature survey. *IEEE transactions on knowledge and data engineering*, 21(1):6–20, 2009.

Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham Kakade. A tensor spectral approach to learning mixed membership community models. In *Conference on Learning Theory*, pages 867–881, 2013.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Amir Ben-Dor, Benny Chor, Richard Karp, and Zohar Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of computational biology*, 10(3-4):373–384, 2003.

Göran Bergqvist and Erik G Larsson. The higher-order singular value decomposition: Theory and an application [lecture notes]. *IEEE Signal Processing Magazine*, 27(3):151–154, 2010.

S Charles Brubaker and Santosh S Vempala. Random tensors and planted cliques. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 406–419. Springer, 2009.

T Tony Cai and Anru Zhang. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *arXiv preprint arXiv:1605.00353*, 2016.

T Tony Cai, Tengyuan Liang, and Alexander Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *arXiv preprint arXiv:1502.01988*, 2015.

Joshua Cape, Minh Tang, and Carey E Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *arXiv preprint arXiv:1705.10735*, 2017.

Arun T Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1040–1048, 2013.

Jie Chen and Yousef Saad. On the tensor svd and the optimal low rank orthogonal approximation of tensors. *SIAM Journal on Matrix Analysis and Applications*, 30(4):1709–1734, 2009.

Andrzej Cichocki, Danilo Mandic, Lieven De Lathauwer, Guoxu Zhou, Qibin Zhao, Cesar Caiafa, and Huy Anh Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32 (2):145–163, 2015.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. On the best rank-1 and rank-(r1, r2,..., rn) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000a.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000b.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *arXiv preprint arXiv:1603.03516*, 2016.

Laura Florescu and Will Perkins. Spectral thresholds in the bipartite stochastic block model. *arXiv preprint arXiv:1506.06737*, 2015.

Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

Laetitia Gauvin, André Panisson, and Ciro Cattuto. Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *PloS one*, 9(1):e86028, 2014.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

T Hildebrand and P Rüegsegger. A new method for the model-independent assessment of thickness in three-dimensional images. *Journal of microscopy*, 185(1):67–75, 1997.

Christopher J Hillar and Lek-Heng Lim. Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *COLT*, pages 956–1006, 2015.

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, 2016.

Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

Mladen Kolar, Sivaraman Balakrishnan, Alessandro Rinaldo, and Aarti Singh. Minimax localization of structural information in large noisy matrices. In *Advances in Neural Information Processing Systems*, pages 909–917, 2011.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Vladimir Koltchinskii and Karim Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 52, pages 1976–2013. Institut Henri Poincaré, 2016.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

Vladimir Koltchinskii and Dong Xia. Perturbation of linear forms of singular vectors under gaussian noise. In *High Dimensional Probability VII*, pages 397–423. Springer, 2016.

Nan Li and Baoxin Li. Tensor completion for on-board compression of hyperspectral images. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 517–520. IEEE, 2010.

Ji Liu, Przemyslaw Musialski, Peter Wonka, and Jieping Ye. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):208–220, 2013.

Tianqi Liu, Ming Yuan, and Hongyu Zhao. Characterizing spatiotemporal transcriptome of human brain via low rank tensor decomposition. *arXiv preprint arXiv:1702.07449*, 2017.

Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 2015.

Andrew McCallum, Kamal Nigam, and Lyle H Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178. ACM, 2000.

Pradipta Mitra. Entrywise bounds for eigenvectors of random graphs. *Electronic journal of combinatorics*, 16(1):R131, 2009.

Chaitanya Muralidhara, Andrew M Gross, Robin R Gutell, and Orly Alter. Tensor decomposition reveals concurrent evolutionary convergences and divergences and correlations with structural motifs in ribosomal rna. *PloS one*, 6(4):e18768, 2011.

Mark EJ Newman. Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2):321–330, 2004.

Larsson Omberg, Gene H Golub, and Orly Alter. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proceedings of the National Academy of Sciences*, 104(47):18371–18376, 2007.

Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter*, 6(1):90–105, 2004.

Sri Priya Ponnapalli, Michael A Saunders, Charles F Van Loan, and Orly Alter. A higher-order generalized singular value decomposition for comparison of global mrna expression from multiple organisms. *PloS one*, 6(12):e28072, 2011.

Emile Richard and Andrea Montanari. A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

Mark Rudelson and Roman Vershynin. Delocalization of eigenvectors of random matrices with independent entries. *Duke Mathematical Journal*, 164(13):2507–2538, 2015.

M Vasilescu and Demetri Terzopoulos. Multilinear analysis of image ensembles: Tensorfaces. *Computer Vision?ECCV 2002*, pages 447–460, 2002.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Van Vu and Ke Wang. Random weighted projections, random quadratic forms and random eigenvectors. *Random Structures & Algorithms*, 47(4):792–821, 2015.

Rongrong Wang. Singular vector perturbation under gaussian noise. *SIAM Journal on Matrix Analysis and Applications*, 36(1):158–177, 2015.

Perke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

C-F Westin, Stephan E Maier, Hatsuho Mamata, Arya Nabavi, Ferenc A Jolesz, and Ron Kikinis. Processing and visualization for diffusion tensor mri. *Medical image analysis*, 6 (2):93–108, 2002.

Dong Xia and Ming Yuan. On polynomial time methods for exact low rank tensor completion. *Foundations of Computational Mathematics, to appear*, 2019.

Qing Xiong, Nicola Ancona, Elizabeth R Hauser, Sayan Mukherjee, and Terrence S Furey. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome research*, 22(2):386–397, 2012.

Anru Zhang and Dong Xia. Tensor svd: Statistical and computational limits. *IEEE Transactions on Information Theory*, 64(11):7311–7338, 2018.

Qinqing Zheng and Ryota Tomioka. Interpolating convex and non-convex tensor decompositions via the subspace norm. In *Advances in Neural Information Processing Systems*, pages 3106–3113, 2015.

Appendix A. Proof of Lemma 14

Let $\mathbf{z}_i \in \mathbb{R}^{m_1}, i = 1, \ldots, m_2$ denote the columns of $\mathbf{Z}$. Then, we write

$$\mathbf{Z}\mathbf{Z}^\top - \sigma^2 m_2 \mathbf{I}_{m_1} = \sum_{i=1}^{m_2} \big(\mathbf{z}_i \otimes \mathbf{z}_i - \sigma^2 \mathbf{I}_{m_1}\big).$$

Similarly, let $\tilde{\mathbf{z}}_j^\top \in \mathbb{R}^{m_2}, j = 1, \ldots, m_1$ denote the rows of $\mathbf{Z}$ and observe that $\|\mathbf{B}\mathbf{Z}^\top\| = \|\mathbf{B}\mathbf{Z}^\top\mathbf{Z}\mathbf{B}^\top\|^{1/2}$ and

$$\mathbf{B}\mathbf{Z}^\top\mathbf{Z}\mathbf{B}^\top = \sum_{j=1}^{m_1} \Big(\big(\mathbf{B}\tilde{\mathbf{z}}_j\big) \otimes \big(\mathbf{B}\tilde{\mathbf{z}}_j\big) - \sigma^2 \mathbf{B}\mathbf{B}^\top\Big).$$

The inequalities (28) and (23) are on the concentration of sample covariance operator, where a sharp bound has been derived in Koltchinskii and Lounici (2017) and will be skipped here.

Appendix B. Proof of Theorem 15

Since $\mathbb{E}\hat{\mathbf{\Gamma}} = \mathbf{0}$, we immediately get $\mathbb{E}\mathbf{L}_k(\hat{\mathbf{\Gamma}}) = \mathbf{0}$. Then,

$$\big\langle \mathbf{x}, \hat{\mathbf{P}}_k^{uu}\mathbf{y}\big\rangle - \mathbb{E}\big\langle \mathbf{x}, \hat{\mathbf{P}}_k^{uu}\mathbf{y}\big\rangle = \big\langle \mathbf{x}, \mathbf{L}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\big\rangle + \big\langle \mathbf{x}, \mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\big\rangle - \mathbb{E}\big\langle \mathbf{x}, \mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\big\rangle.$$

Lemma 19 *For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m_1}$, there exists an absolute constant $D_1 > 0$ such that for all $0 \le t \le m_1$, with probability at least $1 - e^{-t}$,*

$$\big|\langle \mathbf{x}, \mathbf{L}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle\big| \le D_1 t^{1/2} \bigg(\frac{\sigma\mu_1 + \sigma^2 m_2^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\bigg) \|\mathbf{x}\|_{\ell_2} \|\mathbf{y}\|_{\ell_2}.$$

Proof Recall that
$$\hat{\mathbf{\Gamma}} = \mathbf{A}\mathbf{Z}^\top + \mathbf{Z}\mathbf{A}^\top + \mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1}.$$
Then, we write $\langle \mathbf{x}, \mathbf{L}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle$ as

$$\begin{aligned}\langle \mathbf{x}, \mathbf{L}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle =&\langle \hat{\mathbf{\Gamma}}\mathbf{P}_k^{uu}\mathbf{x}, \mathbf{C}_k^{uu}\mathbf{y}\rangle + \langle \hat{\mathbf{\Gamma}}\mathbf{C}_k^{uu}\mathbf{x}, \mathbf{P}_k^{uu}\mathbf{y}\rangle\\ =&\langle (\mathbf{A}\mathbf{Z}^\top + \mathbf{Z}\mathbf{A}^\top + \mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1})\mathbf{P}_k^{uu}\mathbf{x}, \mathbf{C}_k^{uu}\mathbf{y}\rangle\\ &+\langle (\mathbf{A}\mathbf{Z}^\top + \mathbf{Z}\mathbf{A}^\top + \mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1})\mathbf{C}_k^{uu}\mathbf{x}, \mathbf{P}_k^{uu}\mathbf{y}\rangle.\end{aligned}$$

It suffices to consider the following terms separately for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m_1}$:

$$\langle \mathbf{Z}\mathbf{A}^\top\mathbf{x}, \mathbf{y}\rangle, \quad \langle \mathbf{A}\mathbf{Z}^\top\mathbf{x}, \mathbf{y}\rangle, \quad \big\langle \big(\mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1}\big)\mathbf{x}, \mathbf{y}\big\rangle.$$

It is straightforward to check that $\langle \mathbf{Z}\mathbf{A}^\top\mathbf{x}, \mathbf{y}\rangle$ is a normal random variable with zero mean and variance

$$\mathbb{E}\langle \mathbf{Z}\mathbf{A}^\top\mathbf{x}, \mathbf{y}\rangle^2 = \mathbb{E}\langle \mathbf{Z}, \mathbf{y}\otimes(\mathbf{A}^\top\mathbf{x})\rangle^2 = \sigma^2\|\mathbf{y}\otimes(\mathbf{A}^\top\mathbf{x})\|_{\ell_2}^2 = \sigma^2\|\mathbf{y}\|_{\ell_2}^2\|\mathbf{A}^\top\mathbf{x}\|_{\ell_2}^2,$$

where we used the fact that $\mathbf{Z}$ is a $m_1\times m_2$ matrix with i.i.d. $\mathcal{N}(0,\sigma^2)$ entries. Therefore,

$$\mathbb{E}\langle \mathbf{Z}\mathbf{A}^\top\mathbf{P}_k^{uu}\mathbf{x}, \mathbf{C}_k^{uu}\mathbf{y}\rangle^2 \le \frac{\sigma^2\mu_k^2}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}^2\|\mathbf{y}\|_{\ell_2}^2,$$

where we used the facts $\|\mathbf{C}_k\| \le \frac{1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}$ and $\|\mathbf{A}^\top\mathbf{P}_k^{uu}\| \le \mu_k$. By the standard concentration inequality of Gaussian random variables, we get for all $t \ge 0$,

$$\mathbb{P}\bigg(\big|\big\langle \mathbf{Z}\mathbf{A}^\top\mathbf{P}_k^{uu}\mathbf{x}, \mathbf{C}_k^{uu}\mathbf{y}\big\rangle\big| \ge 2t^{1/2}\frac{\sigma\mu_k}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}\bigg) \le e^{-t}.$$

Similarly, for all $t \ge 0$,

$$\mathbb{P}\bigg(\big|\big\langle \mathbf{Z}\mathbf{A}^\top\mathbf{C}_k^{uu}\mathbf{x}, \mathbf{P}_k^{uu}\mathbf{y}\big\rangle\big| \ge 2t^{1/2}\frac{\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}\bigg) \le e^{-t}.$$

We next turn to the bound of $\big|\big\langle\big(\mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1}\big)\mathbf{P}_k^{uu}\mathbf{x}, \mathbf{C}_k^{uu}\mathbf{y}\big\rangle\big|$. Recall that $\mathbf{P}_k^{uu}\mathbf{C}_k^{uu} = \mathbf{0}$ implying that it suffices to consider $\big\langle \mathbf{Z}\mathbf{Z}^\top\mathbf{P}_k^{uu}\mathbf{x}, \mathbf{C}_k^{uu}\mathbf{y}\big\rangle$. Let $\mathbf{z}_1, \ldots, \mathbf{z}_{m_2} \in \mathbb{R}^{m_1}$ denote the columns of $\mathbf{Z}$ such that $\mathbf{z}_i \in \mathcal{N}\big(\mathbf{0}, \sigma^2\mathbf{I}_{m_1}\big)$ for $1 \le i \le m_2$. Write

$$\big\langle \mathbf{Z}\mathbf{Z}^\top(\mathbf{P}_k^{uu}\mathbf{x}), \mathbf{C}_k^{uu}\mathbf{y}\big\rangle = \sum_{i=1}^{m_2}\big\langle \mathbf{z}_i, \mathbf{P}_k^{uu}\mathbf{x}\big\rangle\big\langle \mathbf{z}_i, \mathbf{C}_k^{uu}\mathbf{y}\big\rangle.$$

Observe that $\mathbb{E}\big(\mathbf{P}_k^{uu}\mathbf{z}_i\big)\otimes\big(\mathbf{C}_k^{uu}\mathbf{z}_i\big) = \mathbf{0}$ implying that $\big\langle \mathbf{z}_i, \mathbf{P}_k^{uu}\mathbf{x}\big\rangle$ is independent of $\big\langle \mathbf{z}_i, \mathbf{C}_k^{uu}\mathbf{y}\big\rangle$. By concentration inequalities of Gaussian random variables, for all $t \ge 0$,

$$\mathbb{P}\bigg(\big|\big\langle \mathbf{Z}\mathbf{Z}^\top(\mathbf{P}_k^{uu}\mathbf{x}), \mathbf{C}_k^{uu}\mathbf{y}\big\rangle\big| \ge 2t^{1/2}\|\mathbf{y}\|_{\ell_2}\frac{\sigma\big(\sum_{i=1}^{m_2}\langle \mathbf{z}_i, \mathbf{P}_k^{uu}\mathbf{x}\rangle^2\big)^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)} \bigg| \Big\{\langle \mathbf{z}_i, \mathbf{P}_k^{uu}\mathbf{x}\rangle : i = 1, \ldots, m_2\Big\}\bigg) \le e^{-t}.$$

By (Vershynin, 2010, Prop 5.16), the following bound holds with probability at least $1-e^{-t}$,

$$\Big|\sum_{i=1}^{m_2}\langle \mathbf{z}_i, \mathbf{P}_k^{uu}\mathbf{x}\rangle^2-\sigma^2 m_2\|\mathbf{x}\|_{\ell_2}^2\Big|\lesssim \sigma\Big(m_2^{1/2}t^{1/2}+t\Big)\|\mathbf{x}\|_{\ell_2}.$$

If $t\lesssim m_1\le m_2$, we conclude that there exists an absolute constant $D_1>0$ such that

$$\mathbb{P}\bigg(\big|\big\langle \mathbf{Z}\mathbf{Z}^\top(\mathbf{P}_k^{uu}\mathbf{x}),\mathbf{C}_k^{uu}\mathbf{y}\big\rangle\big|\ge D_1\frac{\sigma^2 m_2^{1/2}t^{1/2}}{\bar g_k(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}\bigg)\le e^{-t}.$$

To sum up, for all $0\le t\lesssim m_1$, the following bound holds with probability at least $1-e^{-t}$,

$$\big|\big\langle \mathbf{x},\mathbf{L}_k(\hat{\boldsymbol{\Gamma}})\mathbf{y}\big\rangle\big|\lesssim t^{1/2}\bigg(\frac{\sigma\mu_1+\sigma^2 m_2^{1/2}}{\bar g_k(\mathbf{A}\mathbf{A}^\top)}\bigg)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}$$

which concludes the proof. ∎

It remains to derive the upper bound of $\big|\langle \mathbf{x},\mathbf{S}_k(\hat{\boldsymbol{\Gamma}})\mathbf{y}\rangle-\mathbb{E}\langle \mathbf{x},\mathbf{S}_k(\hat{\boldsymbol{\Gamma}})\mathbf{y}\rangle\big|$.The following lemma is due to Koltchinskii and Lounici (2016).

Lemma 20 *Let* $\delta(m_1,m_2):=\sigma\mu_1 m_1^{1/2}+\sigma^2(m_1m_2)^{1/2}$ *and suppose that* $\delta(m_1,m_2)\le\frac{1-\gamma}{2(1+\gamma)}\bar g_k(\mathbf{A}\mathbf{A}^\top)$ *for some* $\gamma\in(0,1)$. *There exists a constant* $D_\gamma>0$ *such that, for all symmetric* $\widehat{\boldsymbol{\Gamma}}_1,\hat{\boldsymbol{\Gamma}}_2\in\mathbb{R}^{m_1\times m_1}$ *satisfying the condition* $\max\big\{\|\hat{\boldsymbol{\Gamma}}_1\|,\|\hat{\boldsymbol{\Gamma}}_2\|\big\}\le(1+\gamma)\delta(m_1,m_2)$,

$$\|\mathbf{S}_k(\hat{\boldsymbol{\Gamma}}_1)-\mathbf{S}_k(\hat{\boldsymbol{\Gamma}}_2)\|\le D_\gamma\frac{\delta(m_1,m_2)}{\bar g_k^2(\mathbf{A}\mathbf{A}^\top)}\|\hat{\boldsymbol{\Gamma}}_1-\hat{\boldsymbol{\Gamma}}_2\|.$$

Define function $\varphi(\cdot):\mathbb{R}_+\mapsto[0,1]$ such that $\varphi(t)=1$ for $0\le t\le 1$ and $\varphi(t)=0$ for $t\ge(1+\gamma)$ and φ is linear in between. Then, function φ is Lipschitz on $\mathbb{R}_+$ with constant $\frac{1}{\gamma}$. To illustrate the dependence of $\hat{\boldsymbol{\Gamma}}$ on $\mathbf{Z}$, we write $\hat{\boldsymbol{\Gamma}}(\mathbf{Z})$ instead of $\hat{\boldsymbol{\Gamma}}$. To this end, fix $\mathbf{x},\mathbf{y}\in\mathbb{R}^{m_1}$ and constants $\delta_1,\delta_2>0$ and define the function

$$F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}):=\Big\langle \mathbf{x},\mathbf{S}_k\big(\hat{\boldsymbol{\Gamma}}(\mathbf{Z})\big)\mathbf{y}\Big\rangle\varphi\Big(\frac{\|\hat{\boldsymbol{\Gamma}}(\mathbf{Z})\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}\|}{\delta_2}\Big).$$

where we view $\mathbf{Z}$ as a point in $\mathbb{R}^{m_1\times m_2}$ rather than a random matrix.

Lemma 21 *For any* $\delta_1\le\frac{1-\gamma}{2(1+\gamma)}\bar g_k(\mathbf{A}\mathbf{A}^\top)$ *for some* $\gamma\in(0,1)$ *and* $\delta_2>0$, *there exists an absolute constant* $C_\gamma>0$ *such that*

$$\big|F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_1)-F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_2)\big|\le C_\gamma\frac{\delta_1}{\bar g_k^2(\mathbf{A}\mathbf{A}^\top)}\Big(\mu_1+\delta_2+\frac{\delta_1}{\delta_2}\Big)\|\mathbf{Z}_1-\mathbf{Z}_2\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}$$

Proof [Proof of Lemma 21] Since $\varphi(\frac{\|\hat{\boldsymbol{\Gamma}}(\mathbf{Z})\|}{\delta_1})\varphi(\frac{\|\mathbf{Z}\|}{\delta_2})\neq 0$ only if $\|\hat{\boldsymbol{\Gamma}}(\mathbf{Z})\|\le(1+\gamma)\delta_1$ and $\|\mathbf{Z}\|\le(1+\gamma)\delta_2$, Lemma 16 implies that

$$\big|F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z})\big|=\Big|\Big\langle \mathbf{x},\mathbf{S}_k\big(\hat{\boldsymbol{\Gamma}}(\mathbf{Z})\big)\mathbf{y}\Big\rangle\varphi\Big(\frac{\|\hat{\boldsymbol{\Gamma}}(\mathbf{Z})\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}\|}{\delta_2}\Big)\Big|\le 14(1+\gamma)^2\frac{\delta_1^2}{\bar g_k^2(\mathbf{A}\mathbf{A}^\top)}.$$

Case 1. If $\max\{\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\|, \|\hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\|\} \le (1+\gamma)\delta_1$ and $\max\{\|\mathbf{Z}_1\|, \|\mathbf{Z}_2\|\} \le (1+\gamma)\delta_2$.

By the Lipschitzity of function φ, Lemma 20 and definition of $\hat{\mathbf{\Gamma}}(\mathbf{Z})$, it is easy to check

$$
\begin{aligned}
&|F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_1) - F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_2)| \\
&\quad\le \|\mathbf{S}_k\big(\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big) - \mathbf{S}_k\big(\hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\big)\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} \\
&\quad+ \frac{14(1+\gamma)^2\delta_1}{\gamma\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1) - \hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} + \frac{14(1+\gamma)^2\delta_1^2}{\delta_2\gamma\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{Z}_1 - \mathbf{Z}_2\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} \\
&\quad\le D_\gamma \frac{\delta_1}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1) - \hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} + \frac{14(1+\gamma)^2\delta_1^2}{\delta_2\gamma\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{Z}_1 - \mathbf{Z}_2\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} \\
&\quad\le D_\gamma \frac{\delta_1}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\Big(\mu_1 + \delta_2 + \frac{\delta_1}{\delta_2}\Big)\|\mathbf{Z}_1 - \mathbf{Z}_2\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.
\end{aligned}
$$

Case 2. If $\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\| \le (1+\gamma)\delta_1$, $\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\| \ge (1+\gamma)\delta_1$ and $\max\{\|\mathbf{Z}_1\|, \|\mathbf{Z}_2\|\} \le (1+\gamma)\delta_2$. Since $\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\| \ge (1+\gamma)\delta_1$, we have $\varphi\big(\frac{\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\|}{\delta_1}\big) = 0$ and $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_2) = 0$. Then,

$$
\begin{aligned}
&\big|F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_1) - F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_2)\big| \\
&\quad= \Big|\Big\langle \mathbf{x}, \mathbf{S}_k\big(\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\mathbf{y}\Big\rangle \varphi\Big(\frac{\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}_1\|}{\delta_2}\Big)\Big| \\
&\quad= \Big|\Big\langle \mathbf{x}, \mathbf{S}_k\big(\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\mathbf{y}\Big\rangle \varphi\Big(\frac{\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}_1\|}{\delta_2}\Big) - \Big\langle \mathbf{x}, \mathbf{S}_k\big(\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\mathbf{y}\Big\rangle \varphi\Big(\frac{\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}_1\|}{\delta_2}\Big)\Big| \\
&\quad\le \big\|\mathbf{S}_k\big(\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\big\|\frac{1}{\delta_1\gamma}\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1) - \hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} \\
&\quad\le \frac{(1+\gamma)^2\delta_1^2}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)\delta_1\gamma}\big(2\mu_1 + 2(1+\gamma)\delta_2\big)\|\mathbf{Z}_1 - \mathbf{Z}_2\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} \\
&\quad\le D_\gamma \frac{\delta_1}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}(\mu_1 + \delta_2)\|\mathbf{Z}_1 - \mathbf{Z}_2\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.
\end{aligned}
$$

Case 3. If $\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\| \le (1+\gamma)\delta_1$, $\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\| \ge (1+\gamma)\delta_1$, $\|\mathbf{Z}_1\| \le (1+\gamma)\delta_2$, $\|\mathbf{Z}_2\| \ge (1+\gamma)\delta_2$. It can be proved similarly as *Case 2*.

Case 4. If $\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\| \le (1+\gamma)\delta_1$, $\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\| \ge (1+\gamma)\delta_1$, $\|\mathbf{Z}_1\| \ge (1+\gamma)\delta_2$, $\|\mathbf{Z}_2\| \ge (1+\gamma)\delta_2$. It is a trivial case since $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_1) = F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_2) = 0$.

Case 5. If $\max\{\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\|, \|\hat{\mathbf{\Gamma}}(\mathbf{Z}_2)\|\} \le (1+\gamma)\delta_1$, $\|\mathbf{Z}_1\| \le (1+\gamma)\delta_2$, $\|\mathbf{Z}_2\| \ge (1+\gamma)\delta_2$. Again, we have $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_2) = 0$. Then,

$$
\begin{aligned}
&\big|F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_1) - F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_2)\big| \\
&\quad= \Big|\Big\langle \mathbf{x}, \mathbf{S}_k\big(\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\mathbf{y}\Big\rangle \varphi\Big(\frac{\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}_1\|}{\delta_2}\Big)\Big| \\
&\quad= \Big|\Big\langle \mathbf{x}, \mathbf{S}_k\big(\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\mathbf{y}\Big\rangle \varphi\Big(\frac{\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}_1\|}{\delta_2}\Big) - \Big\langle \mathbf{x}, \mathbf{S}_k\big(\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\mathbf{y}\Big\rangle \varphi\Big(\frac{\|\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}_2\|}{\delta_2}\Big)\Big| \\
&\quad\le \big\|\mathbf{S}_k\big(\hat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\big\|\frac{1}{\delta_2\gamma}\|\mathbf{Z}_1 - \mathbf{Z}_2\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} \le \frac{(1+\gamma)^2\delta_1^2}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)\delta_2\gamma}\|\mathbf{Z}_1 - \mathbf{Z}_2\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} \\
&\quad\le D_\gamma \frac{\delta_1}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\frac{\delta_1}{\delta_2}\|\mathbf{Z}_1 - \mathbf{Z}_2\|\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.
\end{aligned}
$$

All the other cases shall be handled similarly and we conclude the proof. ■

Note that $\|\mathbf{Z}_1-\mathbf{Z}_2\|\leq\|\mathbf{Z}_1-\mathbf{Z}_2\|_{\ell_2}$, Lemma 21 indicates that $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z})$ is Lipschitz with constant

$$D_\gamma\frac{\delta_1}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\Big(\mu_1+\delta_2+\frac{\delta_1}{\delta_2}\Big)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.$$

Lemma 22 *Let $\delta(m_1,m_2):=\sigma\mu_1 m_1^{1/2}+\sigma^2(m_1m_2)^{1/2}$ and suppose that $\mathbb{E}\|\hat{\mathbf{\Gamma}}\|\leq\frac{1-\gamma}{2}\bar{g}_k(\mathbf{A}\mathbf{A}^\top)$ for some $\gamma\in(0,1)$. There exists some constant D_γ such that for any $\mathbf{x},\mathbf{y}\in\mathbb{R}^{m_1}$ and all $\log 8\leq t\leq m_1$, the following inequality holds with probability at least $1-e^{-t}$,*

$$\big|\langle\mathbf{x},\mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle-\mathbb{E}\langle\mathbf{x},\mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle\big|\leq D_\gamma t^{1/2}\frac{\sigma\mu_1+\sigma^2m_2^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\bigg(\frac{\delta(m_1,m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\bigg)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.$$

Proof [Proof of Lemma 22] Choose $\delta_1=\delta_1(m_1,m_2)$ and $\delta_2=\delta_2(m_1,m_2)$ as follows where $\log 8\leq t\leq m_1$ is to be determined:

$$\begin{aligned}\delta_1(m_1,m_2):&=\delta_1(m_1,m_2,t):=\mathbb{E}\|\widetilde{\mathbf{\Gamma}}\|+D_1t^{1/2}(\sigma\mu_1+\sigma^2m_2^{1/2})\\ \delta_2(m_1,m_2):&=\delta_2(m_1,m_2,t):=\mathbb{E}\|\mathbf{Z}\|+D_2\sigma t^{1/2}\end{aligned}$$

and the constants $D_1,D_2>0$ are chosen such that $\mathbb{P}\big(\|\hat{\mathbf{\Gamma}}\|\geq\delta_1(m_1,m_2,t)\big)\leq e^{-t}$ and $\mathbb{P}\big(\|\mathbf{Z}\|\geq\delta_2(m_1,m_2,t)\big)\leq e^{-t}$. Let $M:=\mathrm{Med}(\langle\mathbf{x},\mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle)$ denote its median.

Case 1. If $D_1t^{1/2}(\mu_1\sigma+\sigma^2m_2^{1/2})\leq\frac{\gamma}{4}\bar{g}_k(\mathbf{A}\mathbf{A}^\top)$. Then, $\delta_1\leq(1-\frac{\gamma}{2})\frac{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}{2}=\frac{1-2\gamma'}{1+2\gamma'}\frac{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}{2}$ for some $\gamma'\in(0,1/2)$. By Lemma 21, $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\cdot)$ satisfies the Lipschitz condition. By definition of $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z})$, we have $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z})=\langle\mathbf{x},\mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle$ on the event $\{\|\hat{\mathbf{\Gamma}}\|\leq\delta_1,\|\mathbf{Z}\|\leq\delta_2\}$. By Lemma 14 and $t\geq\log 8$,

$$\begin{aligned}&\mathbb{P}\Big\{F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z})\geq M\Big\}\\ &\qquad\geq\mathbb{P}\Big\{F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z})\geq M,\quad\|\hat{\mathbf{\Gamma}}\|\leq\delta_1,\quad\|\mathbf{Z}\|\leq\delta_2\Big\}\\ &\qquad\geq\mathbb{P}\Big\{\langle\mathbf{x},\mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle\geq M\Big\}-\mathbb{P}\{\|\hat{\mathbf{\Gamma}}\|\leq\delta_1,\|\mathbf{Z}\|\leq\delta_2\}\\ &\qquad\geq\mathbb{P}\Big\{\langle\mathbf{x},\mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle\geq M\Big\}-\mathbb{P}\Big\{\|\hat{\mathbf{\Gamma}}\|\leq\delta_1\Big\}-\mathbb{P}\Big\{\|\mathbf{Z}\|\leq\delta_2\Big\}\\ &\qquad\geq\frac{1}{2}-\frac{1}{8}-\frac{1}{8}=1/4,\end{aligned}$$

and similarly,

$$\mathbb{P}\Big\{F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z})\leq M)\Big\}\geq 1/4.$$

It follows from Gaussian isoperimetric inequality (see (Koltchinskii and Xia, 2016, Lemma 2.6)) and Lemma 21 that with some constant $D_\gamma>0$, for all $t\geq\log 8$ with probability at least $1-e^{-t}$,

$$\big|F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z})-M\big|\leq D_\gamma\frac{\sigma\delta_1t^{1/2}}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\Big(\mu_1+\delta_2+\frac{\delta_1}{\delta_2}\Big)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.$$

Since $t \le m_1 \le m_2$, it is easy to check by Lemma 14 that $\delta_1 \asymp \sigma\mu_1 m_1^{1/2} + \sigma^2(m_1m_2)^{1/2}$ and $\delta_2 \asymp \sigma m_2^{1/2}$. Moreover, $\mathbb{P}\big\{\|\hat{\mathbf{\Gamma}}\| \le \delta_1, \|\mathbf{Z}\| \le \delta_2\big\} \ge 1 - 2e^{-t}$. As a result, with probability at least $1 - e^{-3t}$,

$$\big|\langle \mathbf{x}, \mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle - M\big| \le D_\gamma \frac{\sigma\mu_1 t^{1/2} + \sigma^2 m_2^{1/2} t^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\left(\frac{\delta(m_1, m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\right)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}. \tag{33}$$

Case 2. If $D_1 t^{1/2}(\sigma\mu_1 + \sigma^2 m_2^{1/2}) > \frac{\gamma}{4}\bar{g}_k(\mathbf{A}\mathbf{A}^\top)$. It implies that

$$\mathbb{E}\|\hat{\mathbf{\Gamma}}\| \le D_1 \frac{(1-\gamma)}{\gamma} t^{1/2}(\sigma\mu_1 + \sigma^2 m_2^{1/2}),$$

and $\delta_1 \le D_\gamma t^{1/2}(\sigma\mu_1 + \sigma^2 m_2^{1/2})$. By Lemma 14 and Lemma 16, with probability at least $1 - e^{-t}$,

$$|\langle \mathbf{x}, \mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle| \le \|\mathbf{S}_k(\hat{\mathbf{\Gamma}})\| \le D_\gamma t \frac{(\sigma\mu_1 + \sigma^2 m_2^{1/2})^2}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2},$$

which immediately yields that

$$M \le D_\gamma \frac{(\sigma\mu_1 + \sigma^2 m_2^{1/2})^2}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.$$

The above inequalities imply that with probability at least $1 - e^{-t}$ for $\log 8 \le t \le m_1$,

$$\begin{aligned}|\langle \mathbf{x}, \mathbf{S}_k(\hat{\mathbf{\Gamma}})\mathbf{y}\rangle - M| \le& D_\gamma t \frac{(\sigma\mu_1 + \sigma^2 m_2^{1/2})^2}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} \\ \le& D_\gamma \frac{\sigma\mu_1 t^{1/2} + \sigma^2 m_2^{1/2} t^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\left(\frac{\delta(m_1, m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\right)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.\end{aligned} \tag{34}$$

Therefore, bounds (33) and (34) hold in both cases. The rest of the proof is quite standard by integrating the exponential tails and will be skipped here, see Koltchinskii and Xia (2016). ∎

Proof [Proof of Theorem 15] By Lemma 19 and Lemma 22, if $D_1\delta(m_1, m_2) \le \bar{g}_k(\mathbf{A}\mathbf{A}^\top)$ for a large enough constant $D_1 > 0$ such that $\gamma \le 1/2$, we conclude that for all $\log 8 \le t \le m_1$, with probability at least $1 - 2e^{-t}$,

$$\big|\langle \mathbf{x}, \hat{\mathbf{P}}_k\mathbf{y}\rangle\big| \le D t^{1/2} \frac{\sigma\mu_1 + \sigma^2 m_2^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}$$

which concludes the proof after adjusting the constant D accordingly. ∎

Appendix C. Proof of Lemma 18

Observe that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m_1}$ with $\|\mathbf{x}\|_{\ell_2} = \|\mathbf{y}\|_{\ell_2} = 1$ and $\delta_t = \mathbb{E}\|\hat{\mathbf{\Gamma}}\| + D_1\sigma\mu_1 t^{1/2} + D_2\sigma^2 m_2^{1/2} t^{1/2}$ with $t \le m_1$ and some $\gamma \in (0, 1/2]$,

$$\begin{aligned}
\Big|\mathbb{E}\big\langle \mathbf{x}, \big(\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbf{S}_k(\hat{\mathbf{\Gamma}})\big)\mathbf{y}\big\rangle\Big| \le{}& \mathbb{E}\Big\|\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbf{S}_k(\hat{\mathbf{\Gamma}})\Big\| \\
={}& \mathbb{E}\Big\|\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbf{S}_k(\hat{\mathbf{\Gamma}})\Big\|\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| \le (1+\gamma)\delta_t\Big)\mathbf{1}\Big(\|\hat{\mathbf{\Gamma}}\| \le (1+\gamma)\delta_t\Big) \\
&+ \mathbb{E}\Big\|\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbf{S}_k(\hat{\mathbf{\Gamma}})\Big\|\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| \le (1+\gamma)\delta_t\Big)\mathbf{1}\Big(\|\hat{\mathbf{\Gamma}}\| > (1+\gamma)\delta_t\Big) \\
&+ \mathbb{E}\Big\|\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbf{S}_k(\hat{\mathbf{\Gamma}})\Big\|\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| > (1+\gamma)\delta_t\Big)\mathbf{1}\Big(\|\hat{\mathbf{\Gamma}}\| \le (1+\gamma)\delta_t\Big) \\
&+ \mathbb{E}\Big\|\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbf{S}_k(\hat{\mathbf{\Gamma}})\Big\|\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| > (1+\gamma)\delta_t\Big)\mathbf{1}\Big(\|\hat{\mathbf{\Gamma}}\| > (1+\gamma)\delta_t\Big)
\end{aligned}$$

where the constants $D_1, D_2 > 0$ are chosen such that $\max\big\{\mathbb{P}\big(\|\widetilde{\mathbf{\Gamma}}\| \ge \delta_t\big), \mathbb{P}\big(\|\hat{\mathbf{\Gamma}}\| \ge \delta_t\big)\big\} \le e^{-t}$. By Lemma 20,

$$\begin{aligned}
\mathbb{E}\Big\|\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - &\mathbf{S}_k(\hat{\mathbf{\Gamma}})\Big\|\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| \le (1+\gamma)\delta_t\Big)\mathbf{1}\Big(\|\hat{\mathbf{\Gamma}}\| \le (1+\gamma)\delta_t\Big) \\
&\le D_\gamma \frac{\delta_t}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\mathbb{E}\|\widetilde{\mathbf{\Gamma}} - \hat{\mathbf{\Gamma}}\| \le D_\gamma \frac{\delta_t}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\mathbb{E}\|\mathbf{Z}\mathbf{P}_k^{hh}\mathbf{Z}^\top - \nu_k\sigma^2\mathbf{I}_{m_1}\|.
\end{aligned}$$

By writing $\mathbf{P}_k^{hh} := \sum_{j\in\Delta_k} \mathbf{h}_j \otimes \mathbf{h}_j$, we obtain

$$\begin{aligned}
\mathbf{Z}\mathbf{P}_k^{hh}\mathbf{Z}^\top - \sigma^2\nu_k\mathbf{I}_{m_1} ={}& \sum_{j\in\Delta_k} (\mathbf{Z}\mathbf{h}_j) \otimes (\mathbf{Z}\mathbf{h}_j) - \sigma^2\nu_k\mathbf{I}_{m_1} \\
={}& \nu_k\Big(\frac{1}{\nu_k}\sum_{j\in\Delta_k} (\mathbf{Z}\mathbf{h}_j) \otimes (\mathbf{Z}\mathbf{h}_j) - \sigma^2\mathbf{I}_{m_1}\Big).
\end{aligned}$$

where $\nu_k = \mathrm{Card}(\Delta_k)$. The vectors $\mathbf{Z}\mathbf{h}_j \sim \mathcal{N}(0, \sigma^2\mathbf{I}_{m_1})$ and $\{\mathbf{Z}\mathbf{h}_j : \ldots, j \in \Delta_k\}$ are independent. By Koltchinskii and Lounici (2017),

$$\mathbb{E}\Big\|\frac{1}{\nu_k}\sum_{j\in\Delta_k} (\mathbf{Z}\mathbf{h}_j) \otimes (\mathbf{Z}\mathbf{h}_j) - \sigma^2\mathbf{I}_{m_1}\Big\| \lesssim \sigma^2\Big(\sqrt{\frac{m_1}{\nu_k}} \vee \frac{m_1}{\nu_k}\Big).$$

Since $\nu_k \le m_1$, we conclude with

$$\begin{aligned}
\mathbb{E}\Big\|\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbf{S}_k(\hat{\mathbf{\Gamma}})\Big\|&\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| \le (1+\gamma)\delta_t\Big)\mathbf{1}\Big(\|\hat{\mathbf{\Gamma}}\| \le (1+\gamma)\delta_t\Big) \\
&\lesssim_\gamma \frac{\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{m_1\sigma^2}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big).
\end{aligned} \tag{35}$$

Choose $t = m_1$, by Lemma 16 and Lemma 14,

$$\begin{aligned}
&\mathbb{E}\Big\|\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})-\mathbf{S}_k(\hat{\mathbf{\Gamma}})\Big\|\mathbf{1}\Big(\|\hat{\mathbf{\Gamma}}\| \leq (1+\gamma)\delta_{m_1}\Big)\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| > (1+\gamma)\delta_{m_1}\Big)\\
&\qquad\leq D_\gamma \frac{\delta_{m_1}^2}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\mathbb{E}\frac{\|\widetilde{\mathbf{\Gamma}}\|^2}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| > (1+\gamma)\delta_{m_1}\Big)\\
&\qquad\lesssim_\gamma \frac{\delta_{m_1}^2}{\bar{g}_k^4(\mathbf{A}\mathbf{A}^\top)}e^{-m_1/2}\mathbb{E}^{1/2}\|\widetilde{\mathbf{\Gamma}}\|^4 \lesssim \frac{\delta_{m_1}^4}{\bar{g}_k^4(\mathbf{A}\mathbf{A}^\top)}e^{-m_1/2}\\
&\qquad\lesssim \frac{\delta(m_1,m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\left(\frac{\sigma\mu_1+\sigma^2 m_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\right)
\end{aligned}$$

which is clearly dominated by (35) for $t = m_1$ and $m_2 e^{-m_1/2} \leq 1$. The other terms are bounded in a similar fashion. To sum up, we obtain

$$\|\mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbb{E}\mathbf{S}_k(\hat{\mathbf{\Gamma}})\| \lesssim \frac{\sigma\mu_1+\sigma^2 m_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\left(\frac{\delta(m_1,m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\right).$$