STATISTICAL INFERENCE FOR HIGH DIMENSIONAL DATA WITH LOW RANK STRUCTURE

A Dissertation
Presented to
The Academic Faculty

By

Fan Zhou

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Computational Science and Engineering

Georgia Institute of Technology

December 2018

STATISTICAL INFERENCE FOR HIGH DIMENSIONAL DATA WITH LOW RANK STRUCTURE

Approved by:

Dr. Vladimir Koltchinskii, Advisor
School of Mathematics
Georgia Institute of Technology

Dr. Edmond Chow
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Hongyuan Zha
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Sung Ha Kang
School of Mathematics
Georgia Institute of Technology

Dr. Mark A. Davenport
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Mayya Zhilova
School of Mathematics
Georgia Institute of Technology

Date Approved: July 21, 2018

”You can’t connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future. You have to trust in something - your gut, destiny, life, karma, whatever. Because believing that the dots will connect down the road will give you the confidence to follow your heart even when it leads you off the well worn path; and that will make all the difference.”

Steve Jobs

To my parents

ACKNOWLEDGEMENTS

After the most precious five years of my life at Georgia Tech, so many aspects of me have been transformed. I would acknowledge the great teachers I met, the valuable friendship I made, and the unconditional support and love from my family I received that attribute such a transformation to the next level of my life.

First and foremost, I want to express my deepest gratitude to my advisor Prof. Vladimir Koltchinskii, a truly respectful scholar who patiently transformed me from a student to a researcher. Through those uncountable conversations, he not only introduced me lots of beauty in mathematics but also taught me how to be a scholar with integrity and a person with decency. Those qualities are more scarce I realize as I grow older, and they shaped me as who I am today and for sure I will benefit from them for the rest of my life.

I'd also like to thank Prof. Edmond Chow, the first teacher I met at Georgia Tech with whom I would develop a very good friendship later on. It is always nice to talk with Edmond! He introduced me lots of interesting research topics on the computational aspects and real world applications. He broadened my horizon as a researcher.

Most of my research is supported by NSF grants from Prof. Vladimir Koltchinskii's DMS-1509739 and Prof. Arkadi Nemirovski's CCF-1523768.

I want to thank my academic brother Dong Xia who gave me a lot of help and encouragement as a big brother along the path. We found so many shared interests and common attitudes towards research and life which eventually leads to a great friendship. I want to thank my friends Haiyu Zou, Xin Xing, Xin Wang, Haoyan Zhai, Jiangning Chen, Qianli Hu, Junxiong Jia, Jian Zu and Dawei He for those good old days we spent together in Atlanta. I also want to thank Guojing Cong from IBM Research who hosted me for a wonderful summer in New York.

Finally, I want to thank my parents especially my mother for their unconditional support and endless love. My gratitude to them is beyond words.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

As the big data era has come, a lot of machine learning problems involve data with very high dimension. However, the computational power is always limited. Such kind of practical issue motivates the works in this thesis. In the thesis, we study two major topics on statistical inference of high dimensional data with low rank structure occurred in many machine learning and statistics applications.

The first topic is about nonparametric estimation of low rank matrix valued function with applications in building dynamic recommender systems and recovering euclidean distance matrices in molecular biology. We propose an innovative nuclear norm penalized local polynomial estimator and establish an upper bound on its point-wise risk measured by Frobenius norm. Then we extend this estimator globally and prove an upper bound on its integrated risk measured by L_2-norm. We also propose another new estimator based on bias-reducing kernels to study the case when the matrix valued function is not necessarily low rank and establish an upper bound on its risk measured by L_∞-norm. We show that the obtained rates are all optimal up to some logarithmic factor in minimax sense. Finally, we propose an adaptive estimation procedure for practitioners based on Lepski's method and the penalized data splitting technique which is computationally efficient and can be easily implemented and parallelized. Most results in this work is in the paper [1].

The other topic is about spectral perturbation analysis of higher order singular value decomposition (HOSVD) of tensor under Gaussian noise. Given a tensor contaminated with Gaussian noise, we establish sharp upper bounds on the perturbation of linear forms of singular vectors of HOSVD. In particular, sharp upper bounds are proved for the componentwise perturbation of singular vectors. These results can be applied on sub-tensor localization and low rank tensor denoising. This work is a collaboration with Dong Xia and can be found in the paper [2].

CHAPTER 1
INTRODUCTION TO NONPARAMETRIC ESTIMATION OF LOW RANK MATRIX VALUED FUNCTION

1 Introduction

Let $A : [0,1] \to \mathbb{H}_m$ (the space of Hermitian matrices) be a matrix valued function. The goal is to study the problem of statistical estimation of A based on the regression model

$$\mathbb{E}(Y_j|\tau_j, X_j) = \langle A(\tau_j), X_j \rangle, \;\; j = 1, \dots, n, \tag{1.1}$$

where τ_j are i.i.d. time design variables uniformly distributed in $[0,1]$, X_j are i.i.d. matrix completion sampling matrices, Y_j are independent bounded random responses. Sometimes, it will be convenient to write model (1.1) in the form

$$Y_j = \langle A(\tau_j), X_j \rangle + \xi_j, \;\; j = 1, \dots, n, \tag{1.2}$$

where the noise variables $\xi_j = Y_j - \mathbb{E}(Y_j|\tau_j, X_j)$ are independent and have zero means. In particular, we are interested in the case when A is low rank and satisfies certain smoothness condition. When $A(t) = A_0$ for some $A_0 \in \mathbb{H}_m$ and for any $t \in [0,1]$, such problem coincides with the well known matrix completion/recovery problem which has drawn a lot of attention in the statistics community during the past few years, see [3, 4, 5, 6, 7, 8, 9, 10, 11, 12]. The low rank assumption in matrix completion/estimation problems has profound practical background. For instance, when [13] introduced their famous work on matrix factorization techniques for recommender systems, they considered temporal dynamics, see [14]. Another very common example is Euclidean distance matrix (EDM) which contains the distance information of a large set of points like molecules which are in

low dimensional spaces such as $\mathbb{R}^2$ or $\mathbb{R}^3$. To be more specific, given m points $p_1, ..., p_m$ in $\mathbb{R}^d$, the EDM $D \in \mathbb{R}^{m\times m}$ formed by them has entries $D_{ij} = \|p_i - p_j\|_2^2$. Clearly, this matrix has rank at most $d+1$ regardless of its size m. If $m \gg d$, then the recovery problem falls into the low rank realm. Similar topics in cases when points are fixed (see [15]) or in rigid motion (see [16]) have been studied. While points are moving in smooth trajectories, the EDMs are naturally high dimensional low rank matrix valued functions.

An appealing way to address the low rank issue in matrix completion/estimation is through nuclear norm minimization, see [17]. In section 1 of chapter 2, we inherit this idea and propose a local polynomial estimator (see [18]) with nuclear norm penalization:

$$\widehat{S}^h = \arg\min_{S\in\mathbb{D}} \frac{1}{nh}\sum_{j=1}^{n} K\Big(\frac{\tau_j - t_0}{h}\Big)\Big(Y_j - \Big\langle \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau_j - t_0}{h}\Big), X_j\Big\rangle\Big)^2 + \varepsilon\|S\|_1. \quad (1.3)$$

where $\mathbb{D} \subset \mathbb{H}_{(\ell+1)m}$ is a closed subset of block diagonal matrices with $S_j \in \mathbb{H}_m$ on its diagonal, and $\{p_i\}$ is a sequence of orthogonal polynomials with nonnegative weight function K. The solution to the convex optimization problem (1.3) induces a pointwise estimator of $A(t_0)$: $\widehat{S}^h(t_0) := \sum_{i=0}^{\ell} \widehat{S}_i^h p_i(0)$. We prove that under mild conditions, $\widehat{S}^h(t_0)$ achieves a rate of $O\Big(\Big(\frac{mr\log n}{n}\Big)^{2\beta/(2\beta+1)}\Big)$ on the pointwise risk measured by $\frac{1}{m^2}\big\|\widehat{S}^h(t_0) - A(t_0)\big\|_2^2$ over Hölder class $\Sigma(\beta, L)$ with low rank parameter r, where $\|\cdot\|_2$ denotes the Frobenius norm of a matrix. In section 2.1 of chapter 2, we propose a new global estimator $\widehat{A}$ based on the local results and prove that $\widehat{A}$ achieves a rate of $O\Big(\Big(\frac{mr\log n}{n}\Big)^{2\beta/(2\beta+1)}\Big)$ on the integrated risk measured by L_2-norm, i.e. $\frac{1}{m^2}\int_0^1 \big\|\widehat{A}(t) - A(t)\big\|_2^2 dt$. Then we study another naive kernel estimator $\tilde{A}$ which can be used to estimate matrix valued functions which are not necessarily low rank. This estimator is associated with another popular approach to deal with low rank recovery which is called singular value thresholding, see [4, 9, 12]. We prove that $\tilde{A}$ achieves a rate of $O\Big(\Big(\frac{m\log n}{n}\Big)^{2\beta/(2\beta+1)}\Big)$ measured by $\sup_{t\in[h,1-h]} \frac{1}{m^2}\big\|\tilde{A}(t) - A(t)\big\|^2$, where $\|\cdot\|$ denotes the matrix operator norm. Note that those rates coincide with that of classical matrix recovery/estimation setting when the smoothness parameter $\beta \to \infty$.

An immediate question is whether the above rates are optimal. In section 3 of chapter 2, we prove that all the rates we established are optimal up to some logarithmic factor in the minimax sense, which essentially verified the effectiveness of our methodology.

As one may have noticed, there is an adaptation issue involved in (1.3). Namely, one needs to choose a proper bandwidth h and a proper order of degree ℓ of polynomials. Both parameters are closely related to the smoothness of A which is unknown to us in advance. In section 4 of chapter, we propose a model selection procedure based on Lepskii's method ([19]) and the work of [20] and [21]. We prove that this procedure adaptively selects an estimator that achieves a rate on the integrated risk measured by L_2-norm which is the smallest among all candidates plus a negligible term. What is more important, such a procedure is computationally efficient, feasible in high dimensional setting, and can be easily parallelized.

The major contribution of our work is that we generalized the recent developments of matrix completion/estimation theory to low rank matrix valued function setting by proposing a new optimal estimation procedure. To our best knowledge, no one has ever thoroughly studied such problems from a theoretical point of view.

2 Preliminaries

In this section, we introduce some important definitions, basic facts, and notations for the convenience of presentation.

2.1 Notations

For any Hermitian matrices $A, B \in \mathbb{H}_m$, denote $\langle A, B\rangle = tr(AB)$ which is known as the Hilbert-Schmidt inner product. Denote $\langle A, B\rangle_{L_2(\Pi)} = \mathbb{E}\langle A, X\rangle\langle B, X\rangle$, where Π denotes the distribution of X. The corresponding norm $\|A\|^2_{L_2(\Pi)}$ is given by $\|A\|^2_{L_2(\Pi)} = \mathbb{E}\langle A, X\rangle^2$.

We use $\|\cdot\|_2$ to denote the Hilbert-Schimidt norm (Frobenus norm or Schatten 2-norm) generated by the inner product $\langle\cdot,\cdot\rangle$; $\|\cdot\|$ to denote the operator norm (spectral norm) of

a matrix: the largest singular value; $\|\cdot\|_1$ to denote the trace norm (Schatten 1-norm or nuclear norm), i.e. the sum of singular values; $|A|$ to denote the nonnegative matrix with entries $|A_{ij}|$ corresponding to A.

We denote

$$\sigma^2 := \mathbb{E}\xi^2,\ \sigma_X^2 := \left\| n^{-1}\sum_{j=1}^{n}\mathbb{E}X_i^2 \right\|,\ U_X := \left\| \|X\| \right\|_{L_\infty}.$$

2.2 Matrix Completion and statistical learning setting

The matrix completion setting refers to that the random sampling matrices X_j are i.i.d. uniformly distributed on the following orthonormal basis $\mathcal{X}$ of $\mathbb{H}_m$:

$$\mathcal{X} := \{E_{kj} : k, j = 1, ..., m\},$$

where $E_{kk} := e_k \otimes e_k$, $k = 1, ..., m$; $E_{jk} := \frac{1}{\sqrt{2}}(e_k \otimes e_j + e_j \otimes e_k)$, $1 \le k < j \le m$; $E_{kj} := \frac{i}{\sqrt{2}}(e_k \otimes e_j - e_j \otimes e_k)$, $1 \le k < j \le m$ with $\{e_j\}_{j=1}^m$ being the canonical basis of $\mathbb{C}^m$. The following identities are easy to check when the design matrices are under matrix completion setting:

$$\|A\|_{L_2(\Pi)}^2 = \frac{1}{m^2}\|A\|_2^2, \quad \sigma_X^2 \le \frac{2}{m}, \quad U_X = 1. \tag{2.1}$$

The statistical learning setting refers to the bounded response case: there exists a constant a such that

$$\max_{j=1,...n} |Y_j| \le a,\ a.s. \tag{2.2}$$

In this paper, we will consider model (1.1) under both matrix completion and statistical learning setting.

2.3 Matrix Valued Functions

Let $A : [0,1] \to \mathbb{H}_m$ be a matrix valued function. One should notice that we consider the image space to be Hermitian matrix space for the convenience of presentation. Our methods and results can be readily extended to general rectangular matrix space. Now we define the rank of a matrix valued function. Let $\mathrm{rank}_A(t) := \mathrm{rank}(A(t)),\ \forall t \in [0,1]$.

Definition 1. *Let β and L be two positive real numbers. The **Hölder class** $\Sigma(\beta, L)$ on $[0,1]$ is defined as the set of $\ell = \lfloor \beta \rfloor$ times differentiable functions $f : [0,1] \to \mathbb{R}$ with derivative $f^{(\ell)}$ satisfying*

$$|f^{(\ell)}(x) - f^{(\ell)}(x')| \leq L|x - x'|^{\beta-\ell},\ \ \forall x, x' \in [0,1]. \tag{2.3}$$

In particular, we are interested in the following assumptions on matrix valued functions.

A1 Given a measurement matrix X and for some constant a,

$$\sup_{t\in[0,1]} \left|\langle A(t), X\rangle\right| \leq a.$$

A2 Given a measurement matrix X and for some constant a, the derivative matrices $A^{(k)}$ of A satisfy

$$\sup_{t\in(0,1)} \left|\langle A^{(k)}(t), X\rangle\right| \leq a,\ \ k = 1, ..., \ell.$$

A3 The rank of A, A', ...,$A^{(\ell)}$ are uniformly bounded by a constant r,

$$\sup_{t\in[0,1]} \mathrm{rank}_{A^{(k)}}(t) \leq r,\ \ k = 0, 1, ..., \ell.$$

A4 Assume that for $\forall i, j$, a_{ij} is in the Hölder class $\Sigma(\beta, L)$.

CHAPTER 2

OPTIMAL ESTIMATION OF LOW RANK MATRIX VALUED FUNCTION

1 A local polynomial Lasso estimator

In this section, we study the estimation of matrix valued functions that are low rank. The construction of our estimator is inspired by localization of nonparametric least squares and nuclear norm penalization. The intuition of the localization technique originates from classical local polynomial estimators, see [18]. The intuition behind nuclear norm penalization is that whereas rank function counts the number of non-vanishing singular values, the nuclear norm sums their amplitude. The theoretical foundations behind nuclear norm heuristic for the rank minimization was proved by [17]. Instead of using the trivial basis $\{1, t, t^2, ..., t^\ell\}$ to generate an estimator, we use orthogonal polynomials which fits our problem better. Let $\{p_i\}_{i=0}^{\infty}$ be a sequence of orthogonal polynomials with nonnegative weight function K compactly supported on $[-1, 1]$, then

$$\int_{-1}^{1} K(u)p_i(u)p_j(u)du = \delta_{ij}.$$

There exist an invertible linear transformation $T \in \mathbb{R}^{(\ell+1)\times(\ell+1)}$ such that

$$(1, t, t^2/2!, ..., t^\ell/\ell!)^T = T(p_0, p_1, ..., p_\ell)^T.$$

Apparently, T is lower triangular. We denote $R(T) = \max_{1\le j\le \ell+1} \sum_{i=1}^{\ell+1} |T_{ij}|$. Note that in some literature, $R(T)$ is denoted as $\|T\|_1$ as the matrix "column norm". Since we already used $\|\cdot\|_1$ to denote the nuclear norm, $R(T)$ is used to avoid any ambiguity. Denote

$$\mathbb{D} := \left\{ \mathrm{Diag}\begin{bmatrix} S_0 & S_1 & \dots & S_{\ell-1} & S_\ell \end{bmatrix} \right\} \subset \mathbb{H}_{m(\ell+1)}$$

the set of block diagonal matrices with $S_k \in \mathbb{H}_m$ satisfying $|S_{ij}| \le R(T)a$. With observations (τ_j, X_j, Y_j), $j = 1, ..., n$ from model (1.1), define $\widehat{S}^h$ as

$$\widehat{S}^h = \arg\min_{S\in\mathbb{D}} \frac{1}{nh}\sum_{j=1}^{n} K\Big(\frac{\tau_j - t_0}{h}\Big)\Big(Y_j - \Big\langle \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau_j - t_0}{h}\Big), X_j\Big\rangle\Big)^2 + \varepsilon\|S\|_1. \quad (1.1)$$

$\widehat{S}^h$ naturally induces a local polynomial estimator of order ℓ around t_0:

$$\widehat{S}^h(\tau) := \sum_{i=0}^{\ell} \widehat{S}_i^h p_i\Big(\frac{\tau - t_0}{h}\Big)\mathbb{I}\Big\{\Big|\frac{\tau - t_0}{h}\Big| \le 1\Big\}. \quad (1.2)$$

The point estimate at t_0 is given by

$$\widehat{S}^h(t_0) := \sum_{i=0}^{\ell} \widehat{S}_i^h p_i(0). \quad (1.3)$$

In the following theorem, we establish an upper bound on the point-wise risk of $\widehat{S}^h(t_0)$ when $A(t)$ is in the Hölder class $\Sigma(\beta, L)$ with $\ell = \lfloor\beta\rfloor$.

Theorem 1.1. *Under model (1.1), let (τ_j, X_j, Y_j), $j = 1, ..., n$ be i.i.d. copies of the random triplet (τ, X, Y) with X uniformly distributed in $\mathcal{X}$, τ uniformly distributed in $[0, 1]$, X and τ are independent, and $|Y| \le a$, a.s. for some constant $a > 0$. Let A be a matrix valued function satisfying A1, A2, A3, and A4. Denote $\Phi = \max_{i=0,...,\ell} \|\sqrt{K}p_i\|_\infty$, and $\ell = \lfloor\beta\rfloor$. Take*

$$\widehat{h}_n = C_1\Big(\frac{(\ell^3(\ell!)^2\Phi^2 R(T)^2 a^2 mr\log n}{L^2 n}\Big)^{\frac{1}{2\beta+1}},\ \varepsilon = D\ell a\Phi\sqrt{\frac{\log 2m}{nm\widehat{h}_n}},$$

for some numerical constants C_1 and D. Then for any $\widehat{h}_n \le t_0 \le 1 - \widehat{h}_n$, the estimator defined in (1.3) satisfies with probability at least $1 - \frac{1}{n^{mr}}$,

$$\frac{1}{m^2}\Big\|\widehat{S}^h(t_0) - A(t_0)\Big\|_2^2 \le C_1(a, \Phi, \ell, L)\Big(\frac{mr\log n}{n}\Big)^{\frac{2\beta}{2\beta+1}}, \quad (1.4)$$

where $C_1(a, \Phi, \ell, L)$ is a constant depending on a, Φ, ℓ and L.

One should notice that when $\beta \to \infty$, bound (1.4) coincides with similar result in classical matrix completion. In section 3, we prove that bound (1.4) is minimax optimal up to a logarithmic factor.

Proof. Firstly, we introduce a sharp oracle inequality of "locally integrated risk" of estimator (1.2) in the following lemma.

Lemma 1. Assume that the condition of Theorem 1.1 holds. Then there exist a numerical constants $D > 0$ such that for all

$$\varepsilon \geq D(\ell+1)R(T)\Phi a\Big(\sqrt{\frac{\log 2m}{nmh}} \bigvee \frac{(\log 2m)\Phi}{nh}\Big),$$

and for arbitrary $\eta > 0$, the estimator (1.3) satisfies with probability at least $1 - e^{-\eta}$

$$\begin{aligned} &\frac{1}{h}\mathbb{E}K\Big(\frac{\tau - t_0}{h}\Big)\Big\langle A(\tau) - \widehat{S}^h(\tau), X\Big\rangle^2 \\ &\leq \inf_{S\in\mathbb{D}}\Big\{\frac{1}{h}\mathbb{E}K\Big(\frac{\tau - t_0}{h}\Big)\Big\langle A(\tau) - S(\tau), X\Big\rangle^2 \\ &+ \frac{D^2(\ell+1)^2\Phi^2R(T)^2a^2(\mathrm{rank}(S)m\log 2m + \eta)}{nh}\Big\}. \end{aligned} \tag{1.5}$$

where $S(\tau) := \sum_{i=0}^{\ell} S_i p_i\big(\frac{\tau - t_0}{h}\big)$.

The proof of Lemma 1 can be derived from Theorem 19.1 in [22], see Appendix. To be more specific, one just needs to rewrite (1.1) as

$$\widehat{S}^h = \arg\min_{S\in\mathbb{D}} \frac{1}{n}\sum_{j=1}^{n}\Big(\tilde{Y}_j - \Big\langle S, \tilde{X}_j\Big\rangle\Big)^2 + \varepsilon\|S\|_1. \tag{1.6}$$

where

$$\tilde{X}_j = \mathrm{Diag}\Big[\sqrt{\frac{1}{h}K\Big(\frac{\tau_j - t_0}{h}\Big)}p_0\Big(\frac{\tau_j - t_0}{h}\Big)X_j, ..., \sqrt{\frac{1}{h}K\Big(\frac{\tau_j - t_0}{h}\Big)}p_\ell\Big(\frac{\tau_j - t_0}{h}\Big)X_j\Big],$$

and $\tilde{Y}_j = \sqrt{\frac{1}{h}K\Big(\frac{\tau_j - t_0}{h}\Big)}Y_j$. Then Lemma 1 is just application of Theorem 19.1 in [22].

Consider

$$
\begin{aligned}
&\frac{1}{h}\mathbb{E}K\Big(\frac{\tau-t_0}{h}\Big)\Big\langle A(\tau)-\sum_{i=0}^{\ell}\widehat{S}_i^h p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle^2\\
&=\frac{1}{h}\mathbb{E}K\Big(\frac{\tau-t_0}{h}\Big)\Big\langle A(\tau)-\sum_{i=0}^{\ell}S_i p_i\Big(\frac{\tau-t_0}{h}\Big)+\sum_{i=0}^{\ell}(S_i-\widehat{S}_i^h)p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle^2\\
&=\frac{1}{h}\mathbb{E}K\Big(\frac{\tau-t_0}{h}\Big)\Big\langle \sum_{i=0}^{\ell}(S_i-\widehat{S}_i^h)p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle^2\\
&+\frac{1}{h}\mathbb{E}K\Big(\frac{\tau-t_0}{h}\Big)\Big\langle A(\tau)-\sum_{i=0}^{\ell}S_i p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle^2\\
&+\frac{2}{h}\mathbb{E}K\Big(\frac{\tau-t_0}{h}\Big)\Big\langle A(\tau)-\sum_{i=0}^{\ell}S_i p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle\Big\langle \sum_{i=0}^{\ell}(S_i-\widehat{S}_i^h)p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle
\end{aligned}
\tag{1.7}
$$

Therefore, from (1.5) and (1.7), we have for any $S\in\mathbb{D}$

$$
\begin{aligned}
&\frac{1}{h}\mathbb{E}K\Big(\frac{\tau-t_0}{h}\Big)\Big\langle \sum_{i=0}^{\ell}(S_i-\widehat{S}_i^h)p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle^2\\
&\le\frac{2}{h}\mathbb{E}K\Big(\frac{\tau-t_0}{h}\Big)\Big|\Big\langle A(\tau)-\sum_{i=0}^{\ell}S_i p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle\Big\langle \sum_{i=0}^{\ell}(S_i-\widehat{S}_i^h)p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle\Big|\\
&+\frac{D^2(\ell+1)^2\Phi^2R(T)^2a^2(\mathrm{rank}(S)m\log 2m+\eta)}{nh}.\\
&\le\Big(\frac{c^4}{c^2-1}\Big)\frac{1}{h}\mathbb{E}K\Big(\frac{\tau-t_0}{h}\Big)\Big\langle A(\tau)-\sum_{i=0}^{\ell}S_i p_i\Big(\frac{\tau-t_0}{h}\Big),X\Big\rangle^2\\
&+\Big(\frac{c^2}{c^2-1}\Big)\Big\{\frac{D^2(\ell+1)^2\Phi^2R(T)^2a^2(\mathrm{rank}(S)m\log 2m+\eta)}{nh}\Big\},
\end{aligned}
\tag{1.8}
$$

where we used the fact that for any positive constants a and b, $2ab\le\frac{1}{c^2}a^2+c^2b^2$ for some $c>1$. Take S such that

$$
\sum_{i=0}^{\ell}S_i p_i\Big(\frac{\tau-t_0}{h}\Big)=A(t_0)+A'(t_0)h\Big(\frac{\tau-t_0}{h}\Big)+...+\frac{A^{(\ell)}(t_0)h^\ell}{\ell!}\Big(\frac{\tau-t_0}{h}\Big)^\ell. \tag{1.9}
$$

Note that this is possible since the right hand side is a matrix valued polynomial of $\frac{\tau-t_0}{h}$ up to order ℓ, and $\mathrm{span}\{p_0,p_1,...,p_\ell\}=\mathrm{span}\{1,x,...,x^\ell\}$. Under the condition that all

entries of $A^{(k)}(t)$ are bounded by a, then entries of S_k are bounded by $R(T)a$. Thus, the corresponding $S \in \mathbb{D}$. Obviously, $\mathrm{rank}(S_i) \leq (\ell+1-i)r$. Since $A \in \Sigma(\beta, L)$, we consider ℓ-th order Taylor expansion of A at t_0 to get

$$A(\tau) = A(t_0) + A'(t_0)(\tau - t_0) + ... + \frac{\tilde{A}(\tau - t_0)^\ell}{\ell!}, \tag{1.10}$$

where $\tilde{A}$ is the matrix with $\tilde{A}_{ij} = a_{ij}^{(\ell)}(t_0 + \alpha_{ij}(\tau - t_0))$ for some $\alpha_{ij} \in [0, 1]$. Then we apply the Taylor expansion (1.10) and identity (1.9) to get

$$\begin{aligned} &\frac{1}{h}\mathbb{E}K\Big(\frac{\tau - t_0}{h}\Big)\Big\langle A(\tau) - \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau - t_0}{h}\Big), X\Big\rangle^2 \\ &\leq \frac{1}{h}\mathbb{E}K\Big(\frac{\tau - t_0}{h}\Big)\frac{1}{m^2}\Big\|\frac{LU(\tau - t)^\beta}{\ell!}\Big\|_2^2 \leq \frac{L^2 h^{2\beta}}{(\ell!)^2}. \end{aligned} \tag{1.11}$$

where U denotes the matrix with all entries being 1. The first inequality is due to $a_{ij} \in \Sigma(\beta, L)$, and the second is due to $|\tau - t_0| \leq h$. Under the condition that X is uniformly distributed in $\mathcal{X}$, and the orthogonality of $\{p_i\}_{i=0}^{\ell}$, it is easy to check that

$$\frac{1}{h}\mathbb{E}K\Big(\frac{\tau - t_0}{h}\Big)\Big\langle \sum_{i=0}^{\ell}(S_i - \widehat{S}_i^h)p_i\Big(\frac{\tau - t_0}{h}\Big), X\Big\rangle^2 = \frac{1}{m^2}\sum_{i=0}^{\ell}\|\widehat{S}_i^h - S_i\|_2^2 \tag{1.12}$$

Note that

$$\big\|\widehat{S}^h(t_0) - S(t_0)\big\|_2^2 = \big\|\sum_{i=0}^{\ell}(\widehat{S}_i^h - S_i)p_i(0)\big\|_2^2 \leq (\ell+1)\Phi^2\sum_{i=0}^{\ell}\big\|\widehat{S}_i^h - S_i\big\|_2^2, \tag{1.13}$$

where the second inequality is due to Cauchy-Schwarz inequality and p_i are uniformly bounded on [-1,1]. Combining (1.8), (1.11), (1.12), and (1.13), we get with probability at least $1 - e^{-\eta}$

$$\begin{aligned} \frac{1}{m^2}\|\widehat{S}^h(t_0) - A(t_0)\|_2^2 \leq &\Big(\frac{c^4}{c^2-1}\Big)\frac{2L^2h^{2\beta}}{(\ell!)^2} \\ &+ \Big(\frac{c^2}{c^2-1}\Big)\Big\{\frac{D^2(\ell+1)^2\Phi^2R(T)^2a^2(\mathrm{rank}(S)m\log 2m + \eta)}{nh}\Big\}, \end{aligned}$$

By optimizing the right hand side with respect to h and take $\eta = mr\log n$, we take

$$\widehat{h}_n = C\Big(\frac{\ell^3(\ell!)^2\Phi^2 R(T)^2 a^2 mr\log n}{L^2 n}\Big)^{\frac{1}{2\beta+1}}.$$

where C is a numerical constant. This completes the proof of the theorem. □

2 Global estimators and upper bounds on integrated risk

In this section, we propose two global estimators and study their integrated risk measured by L_2-norm and L_∞-norm.

2.1 From localization to globalization

Firstly, we construct a global estimator based on (1.2). Take

$$\widehat{h}_n = C_1\Big(\frac{\ell^3(\ell!)^2\Phi^2 R(T)^2 a^2 mr\log n}{L^2 n}\Big)^{\frac{1}{2\beta+1}},\ M = \lceil 1/\widehat{h}_n\rceil.$$

Without loss of generality, assume that M is even. Denote $\widehat{S}_k^h(t)$ the local polynomial estimator around t_{2k-1} as in (1.2) by using orthogonal polynomials with $K = \mathbb{I}_{[-1,1]}$, where $t_{2k-1} = \frac{2k-1}{M}$, $k = 1, 2, ..., M/2$ and $\mathbb{I}$ is the indicator function. Denote

$$\widehat{A}(t) = \sum_{k=1}^{M/2} \widehat{S}_k^h(t)\mathbb{I}_{(t_{2k-1}-\widehat{h}_n, t_{2k-1}+\widehat{h}_n]},\ \ t\in(0,1). \tag{2.1}$$

Note that the weight function K is not necessary to be $\mathbb{I}_{[-1,1]}$. It can be replaced by any K that satisfies $K \geq K_0 > 0$ on $[-1,1]$. The following result characterizes the global performance of estimator (2.1) under matrix completion setting measured by L_2-norm.

Theorem 2.1. *Assume that the conditions of Theorem 1.1 hold, and let $\widehat{A}$ be an estimator defined as in (2.1). Then with probability at least $1-\frac{1}{n^{mr-1}}$,*

$$\frac{1}{m^2}\int_0^1 \big\|\widehat{A}(t) - A(t)\big\|_2^2 dt \leq C_2(a,\Phi,\ell,L)\Big(\frac{mr\log n}{n}\Big)^{\frac{2\beta}{2\beta+1}}, \tag{2.2}$$

where $C_2(a, \Phi, \ell, L)$ is a constant depending on a, Φ, ℓ, L.

Compared with the integrated risk measured by L_2-norm of real valued functions in Hölder class, the result in (2.2) has an excess $\log n$ term, which is introduced by the matrix Bernstein inequality, see [23]. In section 3, we show that bound (2.2) is minimax optimal up to a logarithmic factor.

Proof. It is easy to see that

$$\int_0^1 \|\widehat{A}(t) - A(t)\|_2^2 dt \leq \sum_{k=1}^{M/2} \int_{t_{2k-1}-\widehat{h}_n}^{t_{2k-1}+\widehat{h}_n} \|\widehat{S}_k^h(t) - A(t)\|_2^2 dt. \tag{2.3}$$

For each k,

$$\frac{1}{m^2} \int_{t_{2k-1}-\widehat{h}_n}^{t_{2k-1}+\widehat{h}_n} \big\|\widehat{S}_k^h(t) - A(t)\big\|_2^2 dt = \mathbb{E}_{\tau,X} \mathbb{I}_{(t_{2k-1}-\widehat{h}_n, t_{2k-1}+\widehat{h}_n]} \Big\langle A(\tau) - \widehat{S}^h(\tau), X \Big\rangle^2$$

By (1.5), (1.11) and arguments used to prove Theorem 1.1, we have with probability at least $1 - \frac{1}{n^{mr}}$,

$$\frac{1}{m^2 \widehat{h}_n} \int_{t_{2k-1}-\widehat{h}_n}^{t_{2k-1}+\widehat{h}_n} \big\|\widehat{S}_k^h(t) - A(t)\big\|_2^2 dt \leq C_1(a, \Phi, \ell, L) \Big(\frac{mr \log n}{n}\Big)^{\frac{2\beta}{2\beta+1}}.$$

We take the union bound over k, from (2.3) we get with probability at least $1 - \frac{1}{n^{mr-1}}$,

$$\frac{1}{m^2} \int_0^1 \|\widehat{A}(t) - A(t)\|_2^2 dt \leq C_2(a, \Phi, \ell, L) \Big(\frac{mr \log n}{n}\Big)^{\frac{2\beta}{2\beta+1}}.$$

where $C_2(a, \Phi, \ell, L)$ is a constant depending on a, Φ, ℓ, L. □

2.2 Bias reduction through higher order kernels

If $A(t)$ is not necessarily low rank, we propose an estimator which is easy to implement and prove an upper bound on its risk measured by L_∞-norm. Such estimators are related to another popular approach parallel to local polynomial estimators for bias reduction, namely,

using high order kernels to reduce bias. They can also be applied to another important technique of low rank estimation or approximation via singular value thresholding, see [4] and [12]. The estimator proposed by [9] is shown to be equivalent to soft singular value thresholding of such type of estimators.

The kernels we are interested in satisfy the following conditions:

K1 $K(\cdot)$ is symmetric, i.e. $K(u) = K(-u)$.

K2 $K(\cdot)$ is compactly supported on $[-1, 1]$.

K3 $R_K = \int_{-\infty}^{\infty} K^2(u)du < \infty$.

K4 $K(\cdot)$ is of order ℓ for some $\ell \in \mathbb{N}^*$.

K5 $K(\cdot)$ is Lipschitz continuous with $0 < L_K < \infty$.

Consider

$$\tilde{A}(t) = \frac{m^2}{nh} \sum_{j=1}^{n} K\Big(\frac{\tau_j - t}{h}\Big) Y_j X_j. \tag{2.4}$$

Note that when $K \geq 0$, (2.4) is the solution to the following optimization problem

$$\tilde{A}(t) = \arg\min_{S \in \mathbb{D}} \frac{1}{nh} \sum_{j=1}^{n} K\Big(\frac{\tau_j - t}{h}\Big) (Y_j - \langle S, X_j \rangle)^2. \tag{2.5}$$

In the following theorem we prove an upper bound on its global performance measured by L_∞-norm over $\Sigma(\beta, L)$ which is much harder to obtain for matrix lasso problems.

Theorem 2.2. *Under model (1.1), let (τ_j, X_j, Y_j), $j = 1, ..., n$ be i.i.d. copies of the random triplet (τ, X, Y) with X uniformly distributed in $\mathcal{X}$, τ uniformly distributed in $[0, 1]$, X and τ are independent, and $|Y| \leq a$ a.s. for some constant $a > 0$; Kernel K satisfies K1-K5; $A(t)$ be any matrix valued function satisfying A1 and A4. Denote $\ell = \lfloor \beta \rfloor$. Take*

$$\tilde{h}_n := c_*(K) \Big(\frac{a^2 (\ell!)^2 m \log n}{2\beta L^2 n} \Big)^{\frac{1}{2\beta+1}}, \tag{2.6}$$

Then with probability at least $1 - n^{-2}$*, the estimator defined in (2.4) satisfies*

$$\sup_{t\in[\tilde{h}_n,1-\tilde{h}_n]} \frac{1}{m^2}\big\|\tilde{A}(t) - A(t)\big\|^2 \le C^*(K)\Big(\frac{a^2(\ell!)^2 m\log n}{2\beta L^2 n}\Big)^{\frac{2\beta}{2\beta+1}}, \tag{2.7}$$

where $C^*(K)$ *and* $c_*(K)$ *are constants depending on* K*.*

When the smoothness parameter β tends to infinity, bound (2.7) coincides with similar bounds in classical matrix completion, which is $O(\frac{m^3\log n}{n})$. When m degenerates to 1, the bound coincides with that of real valued case, which is $O((\frac{\log n}{n})^{2\beta/(2\beta+1)})$. In section 3, we show that this bound is minimax optimal up to a logarithmic factor.

Proof. In this proof, we use $C(K)$ to denote any constant depending on K which may vary from place to place. This simplifies the representation while does no harm to the soundness of our proof. Consider

$$\sup_{t\in[\tilde{h}_n,1-\tilde{h}_n]} \|\tilde{A}(t) - A(t)\| \le \sup_{t\in[\tilde{h}_n,1-\tilde{h}_n]} \|\tilde{A}(t) - \mathbb{E}\tilde{A}(t)\| + \sup_{t\in[\tilde{h}_n,1-\tilde{h}_n]} \|\mathbb{E}\tilde{A}(t) - A(t)\|. \tag{2.8}$$

The first term on the right hand side is recognized as the variance and the second is the bias. Firstly, we deal with the bias term. Denote $B(t_0) := \mathbb{E}\tilde{A}(t_0) - A(t_0)$, $t_0 \in [\tilde{h}_n, 1-\tilde{h}_n]$. Recall from (1.2), $\mathbb{E}(\xi_j|\tau_j, X_j) = 0$ for any $t_0 \in [\tilde{h}_n, 1-\tilde{h}_n]$ we have

$$\mathbb{E}\tilde{A}(t_0) = \mathbb{E}\frac{m^2}{nh}\sum_{j=1}^{n} K\Big(\frac{\tau_j - t_0}{h}\Big)(\langle A(\tau_j), X_j\rangle + \xi_j)X_j = \frac{m^2}{h}\mathbb{E}K\Big(\frac{\tau - t_0}{h}\Big)\langle A(\tau), X\rangle X.$$

By applying the Taylor expansion of $A(\tau)$ as in (1.10) and the fact that K is a kernel of order ℓ, we get

$$\mathbb{E}\tilde{A}(t_0) = \mathbb{E}\frac{m^2}{h}K\Big(\frac{\tau - t_0}{h}\Big)\langle A(t_0), X\rangle X + \mathbb{E}\frac{m^2}{h}K\Big(\frac{\tau - t_0}{h}\Big)\frac{(\tau - t_0)^\ell}{\ell!}\langle \tilde{A}, X\rangle X,$$

where $\tilde{A}$ is the same as in (1.10). It is easy to check that the first term on the right hand side

is $A(t_0)$. Therefore we rewrite $B(t_0)$ as

$$B(t_0) = \mathbb{E}\frac{m^2}{h}K\Big(\frac{\tau - t_0}{h}\Big)\frac{(\tau - t_0)^\ell}{\ell!}\langle \tilde{A}, X\rangle X = \mathbb{E}\frac{m^2}{h}K\Big(\frac{\tau - t_0}{h}\Big)\frac{(\tau - t_0)^\ell}{\ell!}\langle \tilde{A} - A^{(\ell)}(t_0), X\rangle X,$$

where the second equity is due to the fact that each element of $A(t)$ is in $\Sigma(\beta, L)$ and K is a kernel of order ℓ. Then we can bound each element of matrix $B(t_0)$ as

$$\begin{aligned}
|B_{ij}(t_0)| &\leq \int_0^1 \frac{1}{h}K\Big(\frac{\tau - t_0}{h}\Big)\frac{|\tau - t_0|^\ell}{\ell!}|a_{ij}^{(\ell)}(t_0 + \alpha(\tau - t_0)) - a_{ij}^{(\ell)}(t_0)|d\tau \\
&\leq L\int_0^1 |K(u)|\frac{|uh|^\beta}{\ell!}du \\
&\leq C(K)\frac{Lh^\beta}{\ell!}.
\end{aligned}$$

Thus

$$\sup_{t\in[\tilde{h}_n, 1-\tilde{h}_n]} \|B(t)\| \leq C(K)\frac{Lmh^\beta}{\ell!}. \tag{2.9}$$

On the other hand, for the variance term $\sup_{t\in[\tilde{h}_n,1-\tilde{h}_n]} \|\tilde{A}(t) - \mathbb{E}\tilde{A}(t)\|^2$, we construct a $\delta - net$ on the interval $[0, 1]$ with $\delta = 1/M$, and

$$M = n^2, \quad t_j = \frac{2j-1}{2M},\ j = 1, ..., M.$$

Denote $S_n(t) := \tilde{A}(t) - \mathbb{E}\tilde{A}(t)$, then we have

$$\sup_{t\in[\tilde{h}_n,1-\tilde{h}_n]} \|S_n(t)\| \leq \sup_{t\in[0,1]} \|S_n(t)\| \leq \max_i \|S_n(t_i)\| + \sup_{|t-t'|\leq\delta} \|S_n(t) - S_n(t')\|. \tag{2.10}$$

Next, we bound both terms on the right hand side respectively. For each t_i,

$$S_n(t_i) = \frac{m^2}{nh}\sum_{j=1}^n \Big(K\Big(\frac{\tau_j - t_i}{h}\Big)Y_jX_j - \mathbb{E}K\Big(\frac{\tau_j - t_i}{h}\Big)Y_jX_j\Big).$$

The right hand side is a sum of zero mean random matrices, we apply the matrix Bernstein inequality, see [23]. Under the assumption of Theorem 2.2, one can easily check that with

probability at least $1 - e^{-\eta}$,

$$\|S_n(t_i)\| \le C(K)m^2\Big(\sqrt{\frac{a^2(\eta+\log 2m)}{mnh}} \bigvee \frac{a(\eta+\log 2m)}{nh}\Big).$$

Indeed, by setting $\bar{X} = \frac{m^2}{h}K\big(\frac{\tau-t}{h}\big)YX - \mathbb{E}\frac{m^2}{h}K\big(\frac{\tau-t}{h}\big)YX$, it is easy to check that $U_{\bar{X}} \lesssim \|K\|_\infty am^2/h$ and $\sigma^2_{\bar{X}} \lesssim R_K a^2 m^3/h$. By taking the union bound over all i and setting $\eta = 4\log n$, we get with probability at least $1 - n^{-2}$,

$$\max_i \big\|S_n(t_i)\big\|^2 \le C(K)\frac{a^2m^3\log n}{nh},$$

As for the second term on the right hand side of (2.10), by the assumption that K is a Lipschitz function with Lipschitz constant L_K, we have

$$\begin{aligned}\sup_{|t-t'|\le\delta}\|S_n(t) - S_n(t')\| &\le \sup_{|t-t'|\le\delta}\|(\tilde{A}(t) - \tilde{A}(t'))\| + \sup_{|t-t'|\le\delta}\|\mathbb{E}(\tilde{A}(t) - \tilde{A}(t'))\| \\ &\le \frac{L_K am^3}{n^2h^2} + \frac{L_K am}{n^2h^2}.\end{aligned}$$

Thus with probability at least $1 - n^{-2}$,

$$\sup_{t\in[\tilde{h}_n, 1-\tilde{h}_n]}\|S_n(t)\|^2 \le C(K)\frac{a^2m^3\log n}{nh}$$

Together with the upper bound we get on the bias in (2.9), we have with probability at least $1 - n^{-2}$,

$$\sup_{t\in[\tilde{h}_n, 1-\tilde{h}_n]}\frac{1}{m^2}\|\tilde{A}(t) - A(t)\|^2 \le C(K)\Big(\frac{a^2m\log n}{nh} + \frac{L^2h^{2\beta}}{\ell!^2}\Big).$$

Choose

$$\tilde{h}_n = C(K)\Big(\frac{a^2(\ell!)^2m\log n}{2\beta L^2 n}\Big)^{\frac{1}{2\beta+1}},$$

we get

$$\sup_{t\in[\tilde{h}_n, 1-\tilde{h}_n]}\frac{1}{m^2}\|\tilde{A}(t) - A(t)\|^2 \le C(K)\Big(\frac{a^2(\ell!)^2m\log n}{2\beta L^2 n}\Big)^{\frac{2\beta}{2\beta+1}}.$$

□

3 Lower bounds under matrix completion setting

In this section, we prove the minimax lower bound of estimators (1.3), (2.1) and (2.4). In the realm of classical low rank matrix estimation, [11] studied the optimality issue measured by the Frobenius norm on the classes defined in terms of a "spikeness index" of the true matrix; [10] derived optimal rates in noisy matrix completion on different classes of matrices for the empirical prediction error; [9] established that the rates of the estimator they propose under matrix completion setting are optimal up to a logarithmic factor measured by the Frobenius norm. Based on the ideas of [9], standard methods to prove minimax lower bounds in real valued case in [24], and some fundamental results in coding theory, we establish the corresponding minimax lower bounds of (1.4), (2.2) and (2.7) which essentially shows that the upper bounds we get are all optimal up to some logarithmic factor.

For the convenience of representation, we denote by $\inf_{\widehat{A}}$ the infimum over all estimators $\widehat{A}$ of A. We denote by $\mathcal{A}(r,a)$ the set of matrix valued functions satisfying A1, A2, A3, and A4. We denote by $\mathcal{P}(r,a)$ the class of distributions of random triplet (τ,X,Y) that satisfies model (1.1) with any $A\in\mathcal{A}(r,a)$.

Theorem 3.1. *Under model (1.1), let (τ_j,X_j,Y_j), $j=1,...,n$ be i.i.d. copies of the random triplet (τ,X,Y) with X uniformly distributed in $\mathcal{X}$, τ uniformly distributed in $[0,1]$, X and τ are independent, and $|Y|\leq a$, a.s. for some constant $a>0$; let A be any matrix valued function in $\mathcal{A}(r,a)$. Then there is an absolute constant $\eta\in(0,1)$ such that for all $t_0\in[0,1]$*

$$\inf_{\widehat{A}}\sup_{P^A_{\tau,X,Y}\in\mathcal{P}(r,a)}\mathbb{P}_{P^A_{\tau,X,Y}}\Big\{\frac{1}{m^2}\big\|\widehat{A}(t_0)-A(t_0)\big\|_2^2>C(\beta,L,a)\Big(\frac{mr}{n}\Big)^{\frac{2\beta}{2\beta+1}}\Big\}\geq\eta. \tag{3.1}$$

where $C(\beta,L,a)$ is a constant depending on β, L and a.

Note that compared with the upper bound (1.4), the lower bound (3.1) matches it that

up to a logarithmic factor. As a consequence, it shows that the estimator (1.3) achieves a near optimal minimax rate of pointwise estimation. Although, the result of Theorem 3.1 is under bounded response condition, it can be readily extended to the case when the noise in (1.2) is Gaussian.

Proof. Without loss of generality, we assume that both m and r are even numbers. We introduce several notations which are key to construct the hypothesis set. For some constant $\gamma > 0$, denote

$$\mathcal{C} = \Big\{\tilde{A} = (a_{ij}) \in \mathbb{C}^{\frac{m}{2}\times\frac{r}{2}} : a_{ij} \in \{0, \gamma\}, \forall 1 \le i \le m/2, 1 \le j \le r/2\Big\},$$

and consider the set of block matrices

$$\mathcal{B}(\mathcal{C}) = \left\{ \begin{bmatrix} \tilde{A} & \tilde{A} & \dots & \tilde{A} & O \end{bmatrix} \in \mathbb{C}^{\frac{m}{2}\times\frac{m}{2}} : \tilde{A} \in \mathcal{C} \right\}, \tag{3.2}$$

where O denotes the $m/2 \times (m/2 - r\lfloor m/r \rfloor/2)$ zero matrix. Then we consider a subset of Hermitian matrices $\mathcal{S}_m \subset \mathbb{H}_m$,

$$\mathcal{S}_m = \left\{ \begin{bmatrix} \tilde{O} & \widehat{A} \\ \widehat{A}^* & \tilde{O} \end{bmatrix} \in \mathbb{C}^{m\times m} : \widehat{A} \in \mathcal{B}(\mathcal{C}) \right\}. \tag{3.3}$$

An immediate observation is that for any matrix $A \in \mathcal{S}_m$, $\mathrm{rank}(A) \le r$.

Due to the Varshamov-Gilbert bound (see Lemma 2.9 in [24]), there exists a subset $\mathcal{A}^0 \subset \mathcal{S}_m$ with cardinality $\mathrm{Card}(\mathcal{A}^0) \ge 2^{mr/32} + 1$ containing the zero $m \times m$ matrix $\mathbf{0}$ such that for any two distinct elements A_1 and A_2 of $\mathcal{A}^0$,

$$\|A_1 - A_2\|_2^2 \ge \frac{mr}{16}\lfloor\frac{m}{r}\rfloor\gamma^2 \ge \gamma^2\frac{m^2}{32}. \tag{3.4}$$

Let $f_n(t)$ denote the function $f_n(t) := Lh_n^\beta f\big(\frac{t-t_0}{h_n}\big)$, $t \in [0,1]$, where $h_n = c_0\Big(\frac{mr}{n}\Big)^{\frac{1}{2\beta+1}}$,

with some constant $c_0 > 0$, and $f \in \Sigma(\beta, 1/2) \cap C^\infty$ *and* $Supp(f) = [-1/2, 1/2]$. Note that there exist functions f satisfying this condition. For instance, one can take

$$f(t) = \alpha e^{-\frac{1}{1-4u^2}} \mathbb{I}(|u| < 1/2), \tag{3.5}$$

for some sufficient small $\alpha > 0$. It is easy to check that $f_n(t) \in \Sigma(\beta, L)$ on $[0, 1]$.

We consider the following hypotheses of A at t_0:

$$\mathcal{A}_0^\beta := \big\{\widehat{A}(t) = A f_n(t), t \in [0, 1] : A \in \mathcal{A}^0\big\}.$$

The following claims are easy to check: firstly, any element in $\mathcal{A}_0^\beta$ together with its derivative have rank uniformly bounded by r, and the difference of any two elements of $\mathcal{A}_0^\beta$ satisfies the same property for fixed t_0; secondly, the entries of any element of $\mathcal{A}_0^\beta$ together with its derivative are uniformly bounded by some constant for sufficiently small chosen γ; finally, each element of $A(t) \in \mathcal{A}_0^\beta$ belongs to $\Sigma(\beta, L)$. Therefore, $\mathcal{A}_0^\beta \subset \mathcal{A}(r, a)$ with some chosen γ.

According to (3.4), for any two distinct elements $\widehat{A}_1(t)$ and $\widehat{A}_2(t)$ of $\mathcal{A}_0^\beta$, the difference between $\widehat{A}_1(t)$ and $\widehat{A}_2(t)$ at point t_0 is given by

$$\|\widehat{A}_1(t_0) - \widehat{A}_2(t_0)\|_2^2 \geq \frac{\gamma^2 L^2 c_0^{2\beta} f^2(0)}{32} m^2 \Big(\frac{mr}{n}\Big)^{\frac{2\beta}{2\beta+1}}. \tag{3.6}$$

On the other hand, we consider the joint distributions $P^A_{\tau,X,Y}$ such that $\tau \sim U[0, 1]$, $X \sim \Pi_0$ where Π_0 denotes the uniform distribution on $\mathcal{X}$, τ and X are independent, and

$$\mathbb{P}_A(Y|\tau, X) = \begin{cases} \frac{1}{2} + \frac{\langle A(\tau), X\rangle}{4a}, & Y = a, \\ \frac{1}{2} - \frac{\langle A(\tau), X\rangle}{4a}, & Y = -a. \end{cases}$$

One can easily check that as long as $A(\tau) \in \mathcal{A}_0^\beta$, such $P^A_{\tau,X,Y}$ belongs to the distribution

class $\mathcal{P}(r,a)$. We denote the corresponding $n-$product probability measure by $\mathbb{P}_A$. Then for any $A(\tau) \in \mathcal{A}_0^\beta$, the Kullback-Leibler Divergence between $\mathbb{P}_0$ and $\mathbb{P}_A$ is

$$K(\mathbb{P}_0, \mathbb{P}_A) = n\mathbb{E}\Big(p_0(\tau,X)\log\frac{p_0(\tau,X)}{p_A(\tau,X)} + (1-p_0(\tau,X))\log\frac{1-p_0(\tau,X)}{1-p_A(\tau,X)}\Big),$$

where $p_A(\tau,X) = 1/2 + \langle A(\tau), X\rangle/4a$. Note that $\mathbb{P}_A(Y=a|\tau,X) \in [1/4, 3/4]$ is guaranteed provided that $|\langle A(t), X\rangle| \le a$. Thus by the inequality $-\log(1+u) \le -u+u^2/2,\ \forall u > -1$, and the fact that $\mathbb{P}_A(Y=a|\tau,X) \in [1/4,3/4]$, we have

$$K(\mathbb{P}_0,\mathbb{P}_A) \le n\mathbb{E}2(p_0(\tau,X)-p_A(\tau,X))^2 \le \frac{n}{8a^2}\mathbb{E}\langle A(\tau),X\rangle^2.$$

Recall that $A(\tau) = Af_n(\tau) \in \mathcal{A}_0^\beta$, by $\tau \sim U[0,1]$ and $X \sim \Pi_0$, we have

$$K(\mathbb{P}_0,\mathbb{P}_A) \le \frac{n}{8a^2}\frac{1}{m^2}L^2\|f\|_2^2 h_n^{2\beta+1}m^2\gamma^2 \le \frac{L^2\|f\|_2^2 c_0^{2\beta+1}\gamma^2}{8a^2}mr. \tag{3.7}$$

Therefore, provided the fact that $\mathrm{Card}(\mathcal{A}^0) \ge 2^{mr/32}+1$, together with (3.7), we have

$$\frac{1}{\mathrm{Card}(\mathcal{A}_0^\beta)-1}\sum_{A\in\mathcal{A}_0^\beta} K(\mathbb{P}_0,\mathbb{P}_A) \le \alpha\log(\mathrm{Card}(\mathcal{A}_0^\beta)-1) \tag{3.8}$$

is satisfied for any $\alpha > 0$ if γ is chosen as a sufficiently small constant. In view of (3.6) and (3.8), the lower bound (3.1) follows from Theorem 2.5 in [24]. □

Theorem 3.2. *Under model (1.1), let (τ_j, X_j, Y_j), $j = 1, ..., n$ be i.i.d. copies of the random triplet (τ, X, Y) with X uniformly distributed in $\mathcal{X}$, τ uniformly distributed in $[0,1]$, X and τ are independent, and $|Y| \le a$, a.s. for some constant $a > 0$; let A be any matrix valued function in $\mathcal{A}(r,a)$. Then there is an absolute constant $\eta \in (0,1)$ such that*

$$\inf_{\widehat{A}}\sup_{P^A_{\tau,X,Y}\in\mathcal{P}(r,a)} \mathbb{P}_{P^A_{\tau,X,Y}}\Big\{\frac{1}{m^2}\int_0^1 \big\|\widehat{A}(t)-A(t)\big\|_2^2 dt > \tilde{C}(\beta,L,a)\Big(\frac{mr}{n}\Big)^{\frac{2\beta}{2\beta+1}}\Big\} \ge \eta, \tag{3.9}$$

where $\tilde{C}(\beta, L, a)$ is a constant depending on L, β and a.

The lower bound in (3.9) matches the upper bound we get in (2.2) up to a logarithmic factor. Therefore, our estimator (2.1) achieves a near optimal minimax rate on the integrated risk measured by L_2-norm up to a logarithmic factor. The result of Theorem 3.2 can be readily extended to the case when the noise in (1.2) is Gaussian.

Proof. Without loss of generality, we assume that both m and r are even numbers. Take a real number $c_1 > 0$, define

$$M = \Big\lceil c_1 \Big(\frac{n}{mr}\Big)^{\frac{1}{2\beta+1}} \Big\rceil, \quad h_n = \frac{1}{2M}, \quad t_j = \frac{2j-1}{2M},$$

and

$$\phi_j(t) = L h_n^{\beta} f\Big(\frac{t - t_j}{h_n}\Big), \quad j = 1, ...M, \quad t \in [0, 1],$$

where f is defined the same as in (3.5). Meanwhile, we consider the set of all binary sequences of length M: $\Omega = \Big\{\omega = (\omega_1, ..., \omega_M), \omega_i \in \{0, 1\}\Big\} = \{0, 1\}^M$. By Varshamov-Gilbert bound, there exists a subset $\Omega_0 = \{\omega^0, ..., \omega^N\}$ of Ω such that $\omega^0 = (0, ..., 0) \in \Omega_0$, and $d(\omega^j, \omega^k) \geq \frac{M}{8}$, $\forall\, 0 \leq j < k \leq N$, and $N \geq 2^{\frac{M}{8}}$, where $d(\cdot, \cdot)$ denotes the Hamming distance of two binary sequences. Then we define a collection of functions based on Ω_0: $\mathcal{E} = \Big\{f_\omega(t) = \sum_{j=1}^{M} \omega_j \phi_j(t) : \omega \in \Omega_0\Big\}$. From the result of Varshamov-Gilbert bound, we know that $S := \mathrm{Card}(\mathcal{E}) = \mathrm{Card}(\Omega_0) \geq 2^{\frac{M}{8}} + 1$. It is also easy to check that for all $f_\omega,\ f_{\omega'} \in \mathcal{E}$,

$$\begin{aligned} \int_0^1 (f_\omega(t) - f_{\omega'}(t))^2 dt &= \sum_{j=1}^{M} (\omega_j - \omega_j')^2 \int_{\Delta_j} \phi_j^2(t) dt \\ &= L^2 h_n^{2\beta+1} \|f\|_2^2 \sum_{j=1}^{M} (\omega_j - \omega_j')^2 \\ &\geq L^2 h_n^{2\beta} \|f\|_2^2 / 16, \end{aligned} \tag{3.10}$$

where $\Delta_j = [(j-1)/M, j/M]$.

In what follows, we combine two fundamental results in coding theory: one is Varshamov-Gilbert bound ([25, 26]) in its general form of a q-ary code, the other is the volume estimate of Hamming balls. Let $A_q(n,d)$ denote the largest size of a q-ary code of block length n with minimal Hamming distance d.

Proposition 3.3. The maximal size of a $q-ary$ code of block length n with minimal Hamming distance $d=pn$, satisfies

$$A_q(n,d+1) \geq q^{n(1-h_q(p))}, \tag{3.11}$$

where $p \in [0, 1-1/q]$, $h_q(p) = p\log_q(q-1) - p\log_q p - (1-p)\log_q(1-p)$ is the $q-ary$ entropy function.

We now have all the elements needed in hand to construct our hypotheses set. Denote $\Omega_1 = \{\omega^1, ..., \omega^N\}$, which is a subset of Ω_0 without ω^0. We then consider a subset $\mathcal{E}_1$ of $\mathcal{E}$ which is given by $\mathcal{E}_1 := \Big\{ f_\omega(t) = \sum_{j=1}^{M} \omega_j \phi_j(t) : \omega \in \Omega_1 \Big\}$. Clearly, $S_1 := \mathrm{Card}(\mathcal{E}_1) \geq 2^{M/8}$. Then we define a new collection of matrix valued functions as

$$\mathcal{C} = \big\{ \tilde{A} = (a_{ij}) \in \mathbb{C}^{\frac{m}{2}\times\frac{r}{2}} : a_{ij} \in \{\delta f_\omega : \omega \in \Omega_1\big\},\ \delta \in \mathbb{C}, \forall 1 \leq i \leq m/2, 1 \leq j \leq r/2\}.$$

Obviously, the collection $\mathcal{C}$ is a S_1-ary code of block length $mr/4$. Thus, we can apply the result of Proposition 3.3. It is easy to check that for $p = 1/4$, and $q \geq 4$

$$1 - h_q(p) = 1 - p\log_q \frac{q-1}{p} + (1-p)\log_q(1-p) \geq \frac{1}{4}. \tag{3.12}$$

In our case, $q = S_1 \geq 2^{M/8}$ and $n = mr/4$. If we take $p = 1/4$, we know that

$$A_{S_1}(mr/4, mr/16) \geq A_{S_1}(mr/4, mr/16+1) \geq S_1^{mr/16}. \tag{3.13}$$

In other words, (3.13) guarantees that there exists a subset $\mathcal{H}^0 \subset \mathcal{C}$ with $\mathrm{Card}(\mathcal{H}^0) \geq$

$2^{Mmr/128}$ such that for any $A_1, A_2 \in \mathcal{H}^0$, the Hamming distance between A_1 and A_2 is at least $mr/16$. Now we define the building blocks of our hypotheses set

$$\mathcal{H} := \mathcal{H}^0 \cup \Big\{ O_{\frac{m}{2}\times\frac{r}{2}} \Big\},$$

where $O_{\frac{m}{2}\times\frac{r}{2}}$ is the $\frac{m}{2} \times \frac{r}{2}$ zero matrix. Obviously, $\mathcal{H}$ has size $\mathrm{Card}(\mathcal{H}) \geq 2^{Mmr/64} + 1$, and for any $A_1(t), A_2(t) \in \mathcal{H}$, the minimum Hamming distance is still greater than $mr/16$. We consider the set of matrix valued functions

$$\mathcal{B}(\mathcal{H}) = \left\{ \begin{bmatrix} \tilde{A} & \tilde{A} & \dots & \tilde{A} & O \end{bmatrix} : \tilde{A} \in \mathcal{H} \right\},$$

where O denotes the $m/2 \times (m/2 - r\lfloor m/r \rfloor/2)$ zero matrix. Finally, our hypotheses set of matrix valued functions $\mathcal{H}_m$ is defined as

$$\mathcal{H}_m = \left\{ \begin{bmatrix} \tilde{O} & \widehat{A} \\ \widehat{A}^* & \tilde{O} \end{bmatrix} \in \mathbb{C}^{m\times m} : \widehat{A} \in \mathcal{B}(\mathcal{H}) \right\}.$$

By the definition of $\mathcal{H}_m$ and similar to the arguments in proof of Theorem 3.1, it is easy to check that $\mathcal{H}_m \subset \mathcal{A}(r, a)$, and also

$$\mathrm{Card}(\mathcal{H}_m) \geq 2^{Mmr/64} + 1. \tag{3.14}$$

Now we consider any two different hypotheses $A_j(t), A_k(t) \in \mathcal{H}_m$.

$$\int_0^1 \|A_j(t) - A_k(t)\|_2^2 dt \geq \gamma^2 \frac{mr}{16} 2\Big\lfloor \frac{m}{r} \Big\rfloor \int_0^1 (f_\omega(t) - f_{\omega'}(t))^2 dt, \tag{3.15}$$

where $\omega \neq \omega'$. Based on (3.10), we have

$$\frac{1}{m^2} \int_0^1 \|A_j(t) - A_k(t)\|_2^2 dt \geq \frac{\gamma^2 L^2 h_n^{2\beta} \|f\|_2^2}{256} \geq c_* \Big(\frac{mr}{n}\Big)^{\frac{2\beta}{2\beta+1}}. \tag{3.16}$$

where c_* is a constant depending on $\|f\|_2$, L, c_1 and γ.

On the other hand, we repeat the same analysis on the Kullback-Leibler divergence $K(\mathbb{P}_0, \mathbb{P}_A)$ as in the proof of Theorem 3.1. One can get

$$K(\mathbb{P}_0, \mathbb{P}_A) \leq \frac{n}{8a^2}\mathbb{E}\langle A(\tau), X\rangle^2 \leq \frac{n}{8a^2}\gamma^2 \sum_{j=1}^{M}\int_0^1 \phi_j^2(\tau)d\tau \leq \frac{\gamma^2 c_1^{2\beta+1} L^2 Mmr\|f\|_2^2}{8a^2}, \tag{3.17}$$

where $A(\tau) \in \mathcal{H}_m$. Combine (3.14) and (3.17) we know that

$$\frac{1}{\mathrm{Card}(\mathcal{H}_m) - 1}\sum_{A(t)\in\mathcal{H}_m} K(\mathbb{P}_0, \mathbb{P}_A) \leq \alpha \log(\mathrm{Card}(\mathcal{H}_m) - 1) \tag{3.18}$$

is satisfied for any $\alpha > 0$ if γ is chosen as a sufficiently small constant. In view of (3.16) and (3.18), the lower bound follows from Theorem 2.5 in [24]. □

Now we consider the minimax lower bound on integrated risk measured by L_∞-norm for general matrix valued functions without any rank information. Denote

$$\mathcal{A}(a) = \big\{A(t) \in \mathbb{H}_m,\ \forall t \in [0,1] :\ |a_{ij}(t)| \leq a,\ a_{ij} \in \Sigma(\beta, L)\big\}.$$

We denote by $\mathcal{P}(a)$ the class of distributions of random triplet (τ, X, Y) that satisfies model (1.1) with any $A \in \mathcal{A}(a)$.

Theorem 3.4. *Under model (1.1), let (τ_j, X_j, Y_j), $j = 1, ..., n$ be i.i.d. copies of the random triplet (τ, X, Y) with X uniformly distributed in $\mathcal{X}$, τ uniformly distributed in $[0,1]$, X and τ are independent, and $|Y| \leq a$, a.s. for some constant $a > 0$; let $A(t)$ be any matrix valued function in $\mathcal{A}(a)$. Then there exist an absolute constant $\eta \in (0,1)$ such that*

$$\inf_{\widehat{A}}\sup_{P^A_{\tau,X,Y}\in\mathcal{P}(a)} \mathbb{P}_{P^A_{\tau,X,Y}}\Big\{\sup_{t\in(0,1)}\frac{1}{m^2}\big\|\widehat{A}(t) - A(t)\big\|^2 > \bar{C}(\beta, L, a)\Big(\frac{m\vee\log n}{n}\Big)^{\frac{2\beta}{2\beta+1}}\Big\} \geq \eta. \tag{3.19}$$

where $\bar{C}(\beta, L, a)$ is a constant depending on β, L and a.

Recall that in the real valued case, the minimax lower bound measured by $\sup$ norm of Hölder class is $O((\frac{\log n}{n})^{2\beta/(2\beta+1)})$. In our result (3.19), if the dimension m degenerates to 1, we get the same result as in real valued case and it is optimal. While the dimension m is large enough such that $m \gg \log n$, the lower bound (3.19) shows that the estimator (2.4) achieves a near minimax optimal rate up to a logarithmic factor.

Proof. Without loss of generality, assume that m is an even number. For some constant $\gamma > 0$, denote $\mathcal{V} = \Big\{v \in \mathbb{C}^{\frac{m}{2}} : a_i \in \{0, \gamma\}, \forall\ 1 \le i \le m/2\Big\}$. Due to the Varshamov-Gilbert bound (see Lemma 2.9 in [24]), there exists a subset $\mathcal{V}^0 \subset \mathcal{V}$ with cardinality $\mathrm{Card}(\mathcal{V}^0) \ge 2^{m/16} + 1$ containing the zero vector $\mathbf{0} \in \mathbb{C}^{\frac{m}{2}}$, and such that for any two distinct elements v_1 and v_2 of $\mathcal{V}^0$,

$$\|v_1 - v_2\|_2^2 \ge \frac{m}{16}\gamma^2. \tag{3.20}$$

Consider the set of matrices

$$\mathcal{B}(\mathcal{V}) = \Big\{ \begin{bmatrix} v & v & \dots & v \end{bmatrix} \in \mathbb{C}^{\frac{m}{2}\times\frac{m}{2}} : v \in \mathcal{V}_0 \Big\}.$$

Clearly, $\mathcal{B}(\mathcal{V})$ is a collection of rank one matrices. Then we construct another matrix set $\mathcal{V}_m$,

$$\mathcal{V}_m = \left\{ \begin{bmatrix} \tilde{O} & V \\ V^* & \tilde{O} \end{bmatrix} \in \mathbb{C}^{m\times m} : V \in \mathcal{B}(\mathcal{V}) \right\}$$

where $\tilde{O}$ is the $m/2 \times m/2$ zero matrix. Apparently, $\mathcal{V}_m \subset \mathbb{H}_m$.

On the other hand, we define the grid on $[0, 1]$

$$M = \Big\lceil c_2 \Big(\frac{n}{m + \log n}\Big)^{\frac{1}{2\beta+1}} \Big\rceil, \quad h_n = \frac{1}{2M}, \quad t_j = \frac{2j-1}{2M},$$

and

$$\phi_j(t) = L h_n^\beta f\Big(\frac{t - t_j}{h_n}\Big), \quad j = 1, ...M, \quad t \in [0, 1]$$

where f is defined the same as in (3.5), and c_2 is some constant. Denote $\Phi := \Big\{\phi_j : j = 1, ...M\Big\}$. We consider the following set of hypotheses: $\mathcal{A}_{\mathcal{B}}^{\beta} := \{\widehat{A}(t) = V\phi_j(t) : V \in \mathcal{V}_m,\ \phi_j \in \Phi\}$. One can immediately get that the size of $\mathcal{A}_{\mathcal{B}}^{\beta}$ satisfies

$$\mathrm{Card}(\mathcal{A}_{\mathcal{B}}^{\beta}) \geq (2^{m/16} + 1)M. \tag{3.21}$$

By construction, the following claims are obvious: any element $\widehat{A}(t)$ of $\mathcal{A}_{\mathcal{B}}^{\beta}$ has rank at most 2; the entries of $\widehat{A}(t) \in \mathcal{A}_{\mathcal{B}}^{\beta}$ are uniformly bounded for some sufficiently small γ, and $\widehat{A}_{ij}(t) \in \Sigma(\beta, L)$. Thus $\mathcal{A}_{\mathcal{B}}^{\beta} \subset \mathcal{A}(a)$.

Now we consider the distance between two distinct elements $A(t)$ and $A'(t)$ of $\mathcal{A}_{\mathcal{B}}^{\beta}$. An immediate observation is that

$$\sup_{t\in[0,1]} \|A(t) - A'(t)\|^2 \geq \frac{1}{4} \sup_{t\in[0,1]} \|A(t) - A'(t)\|_2^2,$$

due to the fact that $\forall t \in (0,1)$, $\mathrm{rank}(A(t) - A'(t)) \leq 4$. Then we turn to get lower bound on $\sup_{t\in(0,1)} \|A(t) - A'(t)\|_2^2$. Recall that by construction of $\mathcal{A}_{\mathcal{B}}^{\beta}$, we have for any $A \neq A'$, $A(t) = A_1\phi_j(t),\ A'(t) = A_2\phi_k(t)$, where $A_1, A_2 \in \mathcal{V}_m$. There are three cases need to be considered: 1). $A_1 \neq A_2$ and $j = k$; 2). $A_1 = A_2 \neq 0$ and $j \neq k$; 3). $A_1 \neq A_2$ and $j \neq k$.

For case 1,

$$\sup_{t\in[0,1]} \|A(t) - A'(t)\|_2^2 = \|A_1 - A_2\|_2^2\|\phi_j\|_\infty^2 \geq \frac{m^2}{16}\gamma^2L^2h_n^{2\beta}\|f\|_\infty^2 \geq c_1^* m^2\Big(\frac{m + \log n}{n}\Big)^{\frac{2\beta}{2\beta+1}},$$

where c_1^* is a constant depending on $\|f\|_\infty^2$, β, L and γ.

For case 2,

$$\sup_{t\in[0,1]} \|A(t) - A'(t)\|_2^2 = \|A_1\|_2^2\|\phi_j - \phi_k\|_\infty^2 \geq \frac{m^2}{16}\gamma^2L^2h_n^{2\beta}\|f\|_\infty^2 \geq c_2^* m^2\Big(\frac{m + \log n}{n}\Big)^{\frac{2\beta}{2\beta+1}},$$

where c_2^* is a constant depending on $\|f\|_\infty^2$, β, L and γ.

For case 3,

$$\sup_{t\in[0,1]}\|A(t)-A'(t)\|_2^2 \geq (\|A_1\|_2^2\|\phi_j\|_\infty^2 \vee \|A_2\|_2^2\|\phi_k\|_\infty^2) \geq \frac{m^2}{16}\gamma^2L^2h_n^{2\beta}\|f\|_\infty^2 \geq c_3^* m^2\Big(\frac{m+\log n}{n}\Big)^{\frac{2\beta}{2\beta+1}},$$

where c_3^* is a constant depending on $\|f\|_\infty^2$, β, L and γ.

Therefore, by the analysis above we conclude that for any two distinct elements $A(t)$ and $A'(t)$ of $\mathcal{A}_{\mathcal{B}}^{\beta}$,

$$\sup_{t\in[0,1]}\|A(t)-A'(t)\|^2 \geq \frac{1}{4}\sup_{t\in[0,1]}\|A(t)-A'(t)\|_2^2 \geq c_* m^2\Big(\frac{m+\log n}{n}\Big)^{\frac{2\beta}{2\beta+1}}, \tag{3.22}$$

where c_* is a constant depending on $\|f\|_\infty^2$, L, γ and β.

Meanwhile, we repeat the same analysis on the Kullback-Leibler divergence $K(\mathbb{P}_0,\mathbb{P}_A)$ as in the proof of Theorem 3.1. One can get that for any $A\in\mathcal{A}_{\mathcal{B}}^{\beta}$, the Kullback-Leibler divergence $K(\mathbb{P}_0,\mathbb{P}_A)$ between $\mathbb{P}_0$ and $\mathbb{P}_A$ satisfies

$$K(\mathbb{P}_0,\mathbb{P}_A) \leq \frac{n}{8a^2}\mathbb{E}|\langle A(\tau),X\rangle|^2 \leq \frac{n}{8a^2}\gamma^2\int_0^1\phi_j^2(\tau)d\tau \leq \frac{\gamma^2c_2^{2\beta+1}L^2(m+\log n)\|f\|_2^2}{8a^2}. \tag{3.23}$$

Combine (3.21) and (3.23) we know that

$$\frac{1}{\mathrm{Card}(\mathcal{A}_{\mathcal{B}}^{\beta})-1}\sum_{A\in\mathcal{A}_{\mathcal{B}}^{\beta}}K(\mathbb{P}_0,\mathbb{P}_A) \leq \alpha\log(\mathrm{Card}(\mathcal{A}_{\mathcal{B}}^{\beta})-1) \tag{3.24}$$

is satisfied for any $\alpha>0$ if γ is chosen as a sufficiently small constant. In view of (3.22) and (3.24), the lower bound follows from Theorem 2.5 in [24]. □

4 Model selection

Despite the fact that estimators (1.3) and (2.1) achieve near optimal minimax rates in theory with properly chosen bandwidth h and order of degree ℓ, such parameters depend on

quantities like β and L which are unknown to us in advance. In this section, we propose an adaptive estimation procedure to choose h and ℓ adaptively. Two popular methods to address such problems are proposed in the past few decades. One is Lepskii's method, and the other is aggregation method. In the 1990s, many data-driven procedures for selecting the "best" estimator emerged. Among them, a series of papers stood out and shaped a method what is now called "Lepskii's method". This method has been described in its general form and in great detail in [19]. Later, [27] proposed a bandwidth selection procedure based on pointwise adaptation of a kernel estimator that achieves optimal minimax rate of point estimation over Hölder class, and [28] proposed a new bandwidth selector that achieves optimal rates of convergence over Besov classes with spatially imhomogeneous smoothness. The basic idea of Lepskii's method is to choose a bandwidth from a geometric grid to get an estimator not "very different" from those indexed by smaller bandwidths on the grid. Although Lepskii's method is shown to give optimal rates in pointwise estimation over Hölder class in [27], it has a major defect when applied to our problem: the procedure already requires a huge amount of computational cost when real valued functions are replaced by matrix valued functions. Indeed, with Lepskii's method, in order to get a good bandwidth, one needs to compare all smaller bandwidth with the target one, which leads to dramatically growing computational cost. Still, we have an extra parameter ℓ that needs to fit with h. As a result, we turn to aggregation method to choose a bandwidth from the geometric grid introduced by Lepskii's method, which is more computationally efficient for our problem. The idea of aggregation method can be briefly summarized as follows: one splits the data set into two parts; the first is used to build all candidate estimators and the second is used to aggregate the estimates to build a new one (aggregation) or select one (model selection) which is as good as the best candidate among all constructed. The model selection procedure we use was initially introduced by [20] in classical nonparametric estimation with bounded response. [21] generalized this method to the case where the noise can be unbounded but with a finite p-th moment for some $p > 2$. One can find a more

detailed review on such penalization methods in [29].

Firstly, we introduce the geometric grid created by [27] where to conduct our model selection procedure. Assume that the bandwidth being considered falls into the range $[h_{min}, h_{max}]$. Recall that the "ideal" bandwidth $\widehat{h}_n$ which is given as

$$\widehat{h}_n = C_1\Big(\frac{\ell^3\big(\ell!\Phi R(T)a\big)^2 mr\log n}{L^2 n}\Big)^{\frac{1}{2\beta+1}}, \tag{4.1}$$

$h_{\max}, h_{\min}$ can be chosen as

$$h_{\max} = C_1\Big(\frac{\ell^{*3}\big(\ell^*!\Phi R(T)a\big)^2 mr\log n}{L_*^2 n}\Big)^{\frac{1}{2\beta^*+1}}, \quad h_{\min} = C_1\Big(\frac{\ell_*^3\big(\ell_*!\Phi R(T)a\big)^2 mr\log n}{L^{*2} n}\Big)^{\frac{1}{2\beta_*+1}},$$

where $[\beta_*, \beta^*]$ and $[L_*, L^*]$ are the possible ranges of β, L respectively. Obviously, β is the most important parameter among all. Note that when those ranges are not so clear, a natural upper bound of $h_{\max}$ is 1, and a typical choice of $h_{\min}$ can be set to $n^{-1/2}$. Denote

$$d(h) = \sqrt{1 \vee 2\log\Big(\frac{h_{max}}{h}\Big)}, \quad d_n = \sqrt{2\log\Big(\frac{h_{\max}}{h_{\min}}\Big)}, \quad \alpha(h) = \frac{1}{\sqrt{d(h)}}.$$

Apparently, $d_n = O(\sqrt{\log n})$. Define grid $\mathcal{H}$ inductively by

$$\mathcal{H} = \Big\{h_k \in [h_{min}, h_{max}] : h_0 = h_{max}, h_{k+1} = \frac{h_k}{1+\alpha(h_k)}, k = 0, 1, 2, ...\Big\}. \tag{4.2}$$

Note that the grid $\mathcal{H}$ is a decreasing sequence and the sequence becomes denser as k grows.

We consider possible choices of ℓ_k for each h_k. A trivial candidate set is $\ell_k \in \mathcal{L} := \{\lfloor\beta_*\rfloor, \lfloor\beta_*\rfloor + 1, ..., \lfloor\beta^*\rfloor\} \subset \mathbb{N}^*$. If the size of this set is large, one can shrink it through the correspondence (4.1) between h and β, $\ell_k \le \big\lfloor \frac{\frac{\log n^{-1} + \log mr\log 2m}{\log h_k} - 1}{2} \big\rfloor$. If $n \ge m^d$ for some $d > 1$, $\big\lfloor \frac{\frac{(1-\frac{1}{d})\log n^{-1}}{\log h_k} - 1}{2} \big\rfloor \le \ell_k \le \big\lfloor \frac{\frac{\log n^{-1}}{\log h_k} - 1}{2} \big\rfloor$, which indicates the more the data, the narrower the range. We denote the candidate set for ℓ as $\mathcal{L}$. Then the set

$$\tilde{\mathcal{H}} = \mathcal{H} \times \mathcal{L} := \{(h, \ell) : h \in \mathcal{H},\ \ell \in \mathcal{L}\}$$

indexed a countable set of candidate estimators. Once (h_k, ℓ_k) is fixed, one can take $\varepsilon_k = D(\ell_k + 1)R(T)\Phi a\sqrt{\frac{\log 2m}{nmh_k}}$.

Now we introduce our model selection procedure based on $\tilde{\mathcal{H}}$. We split the data (τ_j, X_j, Y_j), $j = 1, ..., 2n$, into two parts with equal size. The first part of the observations $\{(\tau_j, X_j, Y_j) : j \in \hbar_n\}$ contains n data points, which are randomly drawn without replacement from the original data set. We construct a sequence of estimators $\widehat{A}^k$, $k = 1, 2, ...$ based on the training data set $\hbar_n$ through (2.1) for each pair in $\tilde{\mathcal{H}}$. Our main goal is to select an estimator $\widehat{A}$ among $\{\widehat{A}^k\}$, which is as good as the one that has the smallest mean square error. We introduce an quantity π_k associated with each estimator $\widehat{A}^k$ which serves as a penalty term. We use the remaining part of the data set $\{(\tau_j, X_j, Y_j) : j \in \ell_n\}$ to perform the selection procedure:

$$k^* = \arg\min_k \frac{1}{n}\sum_{j\in\ell_n}(Y_j - \langle\widehat{A}^k(\tau_j), X_j\rangle)^2 + \frac{\pi_k}{n}. \tag{4.3}$$

Denote $\widehat{A}^* = \widehat{A}^{k^*}$ as the adaptive estimator. In practice, we suggest one to rank all estimators $\widehat{A}^k$ according to the following rule: 1. pairs with bigger h always have smaller index; 2. if two pairs have the same h, the one with smaller ℓ has smaller index. Our selection procedure can be summarized in Algorithm 1:

Algorithm 1: Model Selection Procedure

1. Construct the geometric grid $\mathcal{H}$ defined as in (4.2) and compute the candidate set $\tilde{\mathcal{H}}$;
2. Equally split the data set (τ_j, X_j, Y_j), $j = 1, ..., N$ into two parts ($\hbar_n$ and ℓ_n) by randomly drawing without replacement;
3. For each pair in $\tilde{\mathcal{H}}$, construct an estimator $\widehat{A}^k$ defined as in (2.1) using data set $\hbar_n$;
4. Using the second data set ℓ_n to perform the selection rule defined as in (4.3).

The selection procedure described in Algorithm 1 have several advantages: firstly, it chooses a global bandwidth instead of a local one; secondly, since our selection procedure is only based on computations of entries of $\widehat{A}^k$, no matrix computation is involved in the last step, which saves a lot of computational cost and can be easily applied to high dimensional problems; finally, step 3 and 4 can be easily parallelized. The following theorem shows that the integrated risk of $\widehat{A}^*$ measured by L_2-norm can be bounded by the smallest one among all candidates plus an extra term of order n^{-1} which is negligible.

Theorem 4.1. *Under model (1.1), let (τ_j, X_j, Y_j), $j = 1, ..., 2n$ be i.i.d. copies of the random triplet (τ, X, Y) with X uniformly distributed in $\mathcal{X}$, τ uniformly distributed in $[0, 1]$, X and τ are independent, and $|Y| \leq a$, a.s. for some constant $a > 0$; let A be a matrix valued function satisfying A1, A2, A3, and A4; let $\{\widehat{A}^k\}$ be a sequence of estimators constructed from $\tilde{\mathcal{H}}$; let $\widehat{A}^*$ be the adaptive estimator selected through Algorithm 1. Then with probability at least $1 - \frac{1}{n^{mr}}$*

$$\frac{1}{m^2}\int_0^1 \left\|\widehat{A}^*(t) - A(t)\right\|_2^2 dt \leq 3\min_k \left\{\frac{1}{m^2}\int_0^1 \left\|\widehat{A}^k(t) - A(t)\right\|_2^2 dt + \frac{\pi_k}{n}\right\} + \frac{C(a)}{n}, \quad (4.4)$$

where $C(a)$ is a constant depending on a.

Recall that $\mathrm{Card}(\mathcal{H}) = O(\log n)$, we can take $\pi_k = kmr$. Then $\pi_k \leq c_1 mr \log n$ uniformly for all k with some numerical constant c_1. According to Lepskii's method that at least one candidate in $\mathcal{H}$ gives the optimal bandwidth associated with the unknown smoothness parameter β, together with the result of Theorem 2.1, the following corollary follows from Theorem 4.1.

Proof. For any $\widehat{A}^k$, denote the difference in empirical loss between $\widehat{A}^k$ and A by

$$r_n(\widehat{A}^k, A) := \frac{1}{n}\sum_{j=1}^n (Y_j - \langle \widehat{A}^k(\tau_j), X_j\rangle)^2 - \frac{1}{n}\sum_{j=1}^n (Y_j - \langle A(\tau_j), X_j\rangle)^2 = -\frac{1}{n}\sum_{j=1}^n U_j,$$

where $U_j = (Y_j - \langle A(\tau_j), X_j\rangle)^2 - (Y_j - \langle \widehat{A}^k(\tau_j), X_j\rangle)^2$. It is easy to check that

$$U_j = 2(Y_j - \langle A(\tau_j), X_j\rangle)\langle \widehat{A}^k(\tau_j) - A(\tau_j), X_j\rangle - \langle \widehat{A}^k(\tau_j) - A(\tau_j), X_j\rangle^2. \tag{4.5}$$

We denote $r(\widehat{A}^k, A) := \mathbb{E}\langle \widehat{A}^k(\tau) - A(\tau), X\rangle^2$. The following concentration inequality developed by [30] to prove Bernstein's inequality is key to our proof.

Lemma 2. Let U_j, $j = 1, ..., n$ be independent bounded random variables satisfying $|U_j - \mathbb{E}U_j| \leq M$ with $h = M/3$. Set $\bar{U} = n^{-1}\sum_{j=1}^n U_j$. Then for all $t > 0$

$$\mathbb{P}\Big\{\bar{U} - \mathbb{E}\bar{U} \geq \frac{t}{n\varepsilon} + \frac{n\varepsilon \mathrm{var}(\bar{U})}{2(1-c)}\Big\} \leq e^{-t},$$

with $0 < \varepsilon h \leq c < 1$.

Firstly, we bound the variance of U_j. Under the assumption that $|Y|$ and $|\langle A(\tau), X\rangle|$ are bounded by a constant a, one can easily check that $h = 8a^2/3$. Given $\mathbb{E}(Y_j|\tau_j, X_j) = \langle A(\tau_j), X_j\rangle$, we know that the covariance between the two terms on the right hand side of (4.5) is zero. Conditionally on (τ, X), the second order moment of the first term satisfies

$$4\mathbb{E}\sigma^2_{Y|\tau,X}\langle \widehat{A}^k(\tau_j) - A(\tau_j), X_j\rangle^2 \leq 4a^2 r(\widehat{A}^k, A).$$

To see why, one can consider the random variable $\tilde{Y}$ with the distribution $\mathbb{P}\{\tilde{Y} = a\} = \mathbb{P}\{\tilde{Y} = -a\} = 1/2$. The variance of Y is always bounded by the variance of $\tilde{Y}$ which is a^2 under the assumption that $|Y_j|$ and $|\langle \widehat{A}^k(\tau_j), X_j\rangle|$ are bounded by a constant $a > 0$. Similarly, we can get that the variance of the second term conditioned on (τ, X) is also bounded by $4a^2\mathbb{E}\langle \widehat{A}^k(\tau_j) - A(\tau_j), X_j\rangle^2$. As a result, $n\mathrm{var}(\bar{U}) \leq 8a^2 r(\widehat{A}^k, A)$. By the result of Lemma 2, we have for any $\widehat{A}^k$ with probability at least $1 - e^{-t}$

$$r(\widehat{A}^k, A) - r_n(\widehat{A}^k, A) < \frac{t}{n\varepsilon} + \frac{4a^2\varepsilon r(\widehat{A}^k, A)}{1-c}.$$

Set $t = \varepsilon\pi_k + \log 1/\delta$, we get with probability at least $1 - \delta/e^{\varepsilon\pi_k}$

$$(1-\alpha)r(\widehat{A}^k, A) < r_n(\widehat{A}^k, A) + \frac{\pi_k}{n} + \frac{4a^2}{(1-c)\alpha}\Big(\frac{\log 1/\delta}{n}\Big).$$

where $\alpha = 4a^2\varepsilon/(1-c) < 1$. Denote

$$\tilde{k}^* = \arg\min_k \Big\{ r(\widehat{A}^k, A) + \frac{\pi_k}{n} \Big\}.$$

By the definition of $\widehat{A}^*$, we have with probability at least $1 - \delta/e^{\varepsilon\widehat{\pi}^*}$

$$(1-\alpha)r(\widehat{A}^*, A) < r_n(\widehat{A}^{\tilde{k}^*}, A) + \frac{\pi_{\tilde{k}^*}}{n} + \frac{4a^2}{(1-c)\alpha}\Big(\frac{\log 1/\delta}{n}\Big). \tag{4.6}$$

where $\widehat{\pi}^*$ is the penalty terms associated with $\widehat{A}^*$.

Now we apply the result of Lemma 2 one more time and set $t = \log 1/\delta$, we get with probability at least $1 - \delta$

$$r_n(\widehat{A}^{\tilde{k}^*}, A) \le (1+\alpha)r(\widehat{A}^{\tilde{k}^*}, A) + \frac{4a^2}{(1-c)\alpha}\frac{\log 1/\delta}{n}. \tag{4.7}$$

Apply the union bound of (4.6) and (4.7), we get with probability at least $1 - \delta(1 + e^{-\varepsilon\widehat{\pi}^*})$

$$r(\widehat{A}^*, A) \le \frac{(1+\alpha)}{(1-\alpha)}\Big(r(\widehat{A}^{\tilde{k}^*}, A) + \frac{\pi_{\tilde{k}^*}}{n}\Big) + \frac{4a^2}{(1-c)\alpha(1-\alpha)}\frac{\log 1/\delta}{n}.$$

By taking $\varepsilon = 3/32a^2$ and $c = \varepsilon h$,

$$r(\widehat{A}^*, A) \le 3\Big(r(\widehat{A}^{\tilde{k}^*}, A) + \frac{\pi_{\tilde{k}^*}}{n}\Big) + \frac{64a^2}{3}\frac{\log 1/\delta}{n}.$$

By taking $\delta = 1/n^{mr}$ and adjusting the constant, we have with probability at least $1 - 1/n^{mr}$

$$\frac{1}{m^2}\int_0^1 \big\|\widehat{A}^*(t) - A(t)\big\|_2^2 dt \le 3\min_k \Big\{ \frac{1}{m^2}\int_0^1 \big\|\widehat{A}^k(t) - A(t)\big\|_2^2 dt + \frac{\pi_k}{n} \Big\} + C(a)\frac{mr\log n}{n}$$

where $C(a)$ is a constant depending on a. $\square$

Corollary 4.1. Assume that the conditions of Theorem 4.1 hold with $\pi_k = kmr$, and $n > mr \log n$. Then with probability at least $1 - \frac{1}{n^{mr-1}}$

$$\frac{1}{m^2} \int_0^1 \left\| \widehat{A}^*(t) - A(t) \right\|_2^2 dt \leq C(a, \Phi, \ell, L) \Big(\frac{mr \log n}{n} \Big)^{\frac{2\beta}{2\beta+1}} \tag{4.8}$$

where $C(a, \Phi, \ell, L)$ is a constant depending on a, Φ, ℓ, and L.

CHAPTER 3

SIMULATION RESULTS OF NONPARAMETRIC ESTIMATION OF LOW RANK MATRIX VALUED FUNCTION

1 An ADMM Algorithm

The alternating direction method of multipliers (ADMM) is a powerful algorithm to solve convex optimization problems, see a comprehensive introduction in [31]. The application of ADMM to matrix recovery problems can be found in [32, 33] and the references therein.

Recall that the major optimization problem to solve is

$$\widehat{S}^h = \arg \frac{1}{nh} \sum_{j=1}^{n} K\Big(\frac{\tau_j - t_0}{h}\Big)\Big(Y_j - \Big\langle \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau_j - t_0}{h}\Big), X_j \Big\rangle\Big)^2 + \varepsilon \|S\|_1. \tag{1.1}$$

where $\mathbb{D} \subset \mathbb{H}_{(\ell+1)m}$ is a closed subset of block diagonal matrices with $S_j \in \mathbb{H}_m$ on its diagonal, and $\{p_i\}$ is a sequence of orthogonal polynomials with nonnegative weight function K. Such a problem belongs to the standard form of optimization problems considered in ADMM applications, see [31]. To be more specific, the optimization problem involves the sum of two functions of the solution, i.e. a loss function and a penalization term. Instead of solving the original optimization problem, our ADMM algorithm presented below introduces a new variable such that the original problem to solve is transformed into the following one:

$$\widehat{S}^h = \arg \min_{S=\bar{S}\ S,\bar{S}\in\mathbb{D}} \frac{1}{nh} \sum_{j=1}^{n} K\Big(\frac{\tau_j - t_0}{h}\Big)\Big(Y_j - \Big\langle \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau_j - t_0}{h}\Big), X_j \Big\rangle\Big)^2 + \varepsilon \|\bar{S}\|_1. \tag{1.2}$$

Then the corresponding augmented Lagrangian multipliers of function of (1.2) is defined

as

$$\widehat{S}^h = \arg \min_{S=\bar{S}\ S,\bar{S}\in\mathbb{D}} \frac{1}{nh}\sum_{j=1}^{n} K\Big(\frac{\tau_j - t_0}{h}\Big)\Big(Y_j - \Big\langle \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau_j - t_0}{h}\Big), X_j\Big\rangle\Big)^2 + \varepsilon\|\bar{S}\|_1 + \frac{\rho}{2}\|S - \bar{S}\|_2^2 + \langle Z, S - \bar{S}\rangle. \tag{1.3}$$

where $\rho > 0$ is a constant and $Z \in \mathbb{D}$. Our ADMM algorithm is presented in Algorithm 2. It updates S and $\bar{S}$ alternatively and the multiplier Z is updated by the difference between the iterates of S and $\bar{S}$. Note that in order to update $\bar{S}^{(k)}$, it is equivalent to solve the following optimization problem:

$$\bar{S}^{(k+1)} = \arg\min_{\bar{S}\in\mathbb{D}} \varepsilon\|\bar{S}\|_1 + \frac{\rho}{2}\|S^{(k+1)} - \bar{S} + \frac{Z^{(k)}}{\rho}\|_2^2. \tag{1.4}$$

It was proved by [9] that the solution to this problem has a simple form which can be obtained by soft thresholding of singular values of $S^{(k+1)} + \frac{Z^{(k)}}{\rho}$ as

$$\bar{S}^{(k+1)} = \sum_j (\sigma_j(\tilde{S}) - \frac{\varepsilon}{\rho})_+ u_j(\tilde{S}) v_j(\tilde{S})^T, \tag{1.5}$$

where $\tilde{S} = S^{(k+1)} + \frac{Z^{(k)}}{\rho}$, $x_+ = \max\{x, 0\}$, and $\sigma_j(\tilde{S}),\ u_j,\ v_j$ are the singular values, left and right singular vectors of $\tilde{S}$ respectively.

2 Numerical results

In this section, we present the numerical simulation results of our estimators (1.3) and (2.1), and simulation results of our model selection procedure in Algorithm 1. The underlying matrix valued function we create is in Hölder class $\Sigma(\beta, L)$ with $\beta = 3/2$, $L = 24$ and rank constraint $\mathrm{rank}_A(t) \leq 3$. The orthogonal polynomial we choose is Chebyshev polynomials of the second kind.

Algorithm 2: ADMM Algorithm

Set up the values of $max\_Iteration$ and tolerance $\varepsilon_{tol} > 0$; Initialize $S^{(0)}, \bar{S}^{(0)} \in \mathbb{D}$ and $Z^{(0)} = \mathbf{0}$; **while** $k < max\_Iteration$ **do**

$$S^{(k+1)} = \arg\min_{S\in\mathbb{D}} \frac{1}{nh}\sum_{j=1}^{n} K\Big(\frac{\tau_j - t_0}{h}\Big)\Big(Y_j - \Big\langle \sum_{i=0}^{\ell} S_i p_i\Big(\frac{\tau_j - t_0}{h}\Big), X_j\Big\rangle\Big)^2 + \frac{\rho}{2}\|S - \bar{S}^{(k)}\|_2^2 + \langle Z^{(k)}, S - \bar{S}^{(k)}\rangle;$$

$$\bar{S}^{(k+1)} = \arg\min_{\bar{S}\in\mathbb{D}} \varepsilon\|\bar{S}\|_1 + \frac{\rho}{2}\|S^{(k+1)} - \bar{S}\|_2^2 + \langle Z^{(k)}, S^{(k+1)} - \bar{S}\rangle;$$

$$Z^{(k+1)} = Z^{(k)} + \rho(S^{(k+1)} - \bar{S}^{(k+1)});$$

if $\|\bar{S}^{(k+1)} - \bar{S}^{(k)}\|_2^2 \le \varepsilon_{tol}$ *or* $\|Z^{(k+1)} - Z^{(k)}\|_2^2 \le \rho^2\varepsilon_{tol}$ **then**

  Reaching the tolerance;

end

Return $\bar{S}^{(k+1)}$. $k = k + 1$;

end

Return $\bar{S}^{(k+1)}$.

In Fig. 3.1 and Fig. 3.2 we plot the pointwise error at $t_0 = 0.5$ and integrated risk against the iteration number of Algorithm 2. As we can see, our ADMM algorithm converges really fast when ρ is small. We should also mention that according to our experiments, smaller ρ value gives faster convergence speed typically. But it doesn't mean that smaller ρ is always better. There is an optimal value of ρ that gives the best accuracy. One needs to tune this parameter in order to get the best accuracy and fairly good convergence speed.

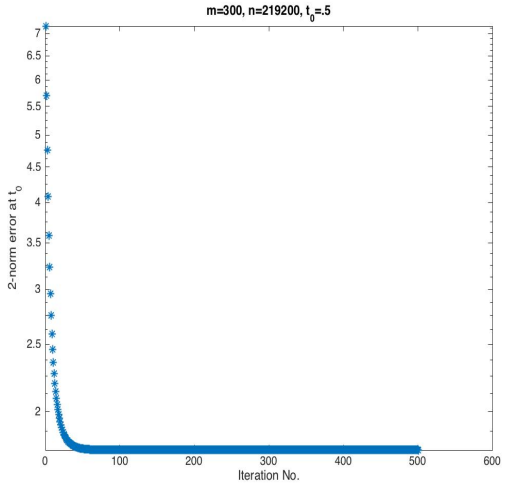


Figure 3.1: Pointwise risk convergence of ADMM Algorithm

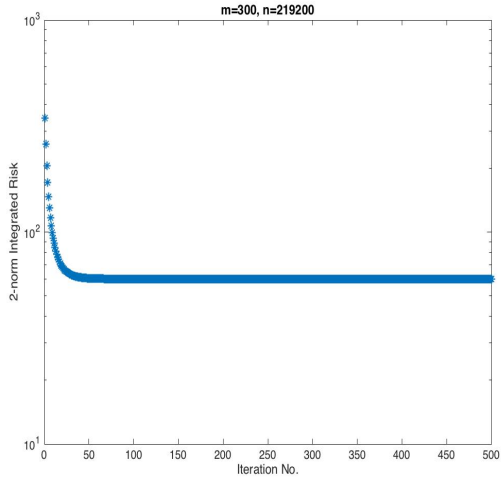


Figure 3.2: Integrated risk convergence of ADMM Algorithm

2.1 Pointwise estimation simulation

By plug in the optimal bandwidth in Theorem 1.1, we run our Algorithm 2 to solve the point estimator at $t_0 = 0.5$ with $m = 150$. Fig. 3.3 - Fig. 3.10 show different levels of recovery for the true underlying data matrix. As we can see, the recovery quality increases evidently as sample size n grows.

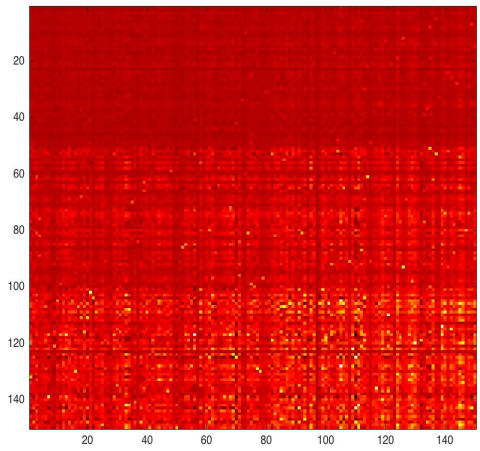


Figure 3.3: $n = 1600$

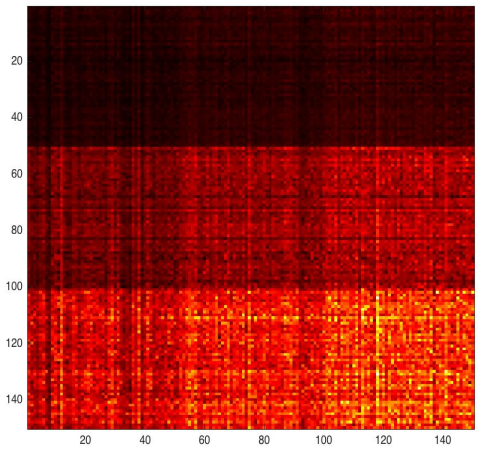


Figure 3.4: $n = 6400$

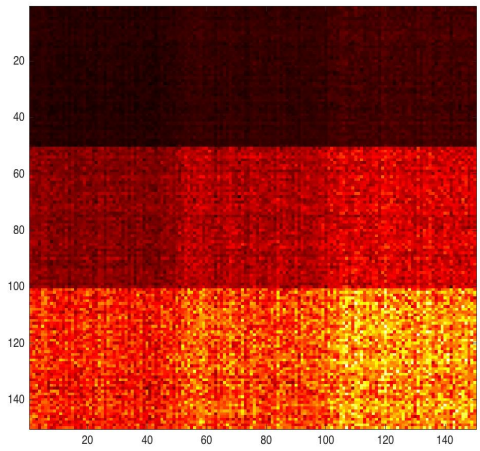


Figure 3.5: $n = 25600$

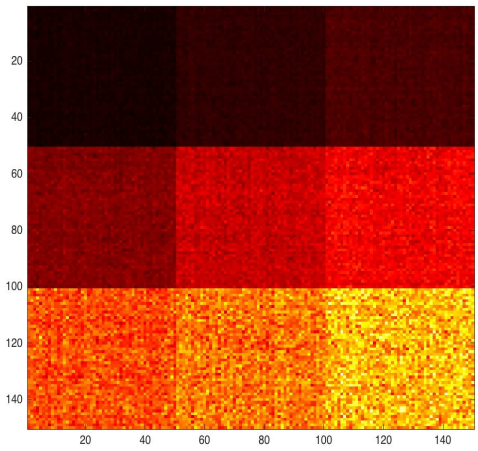


Figure 3.6: $n = 102400$

In table 3.1 and table 3.2 we display the comparison of pointwise risk measured by $\frac{1}{m^2}\|\widehat{A}(t_0) - A(t_0)\|_2^2$ between our theoretical bounds proved in (1.4), (3.1) and our simula-

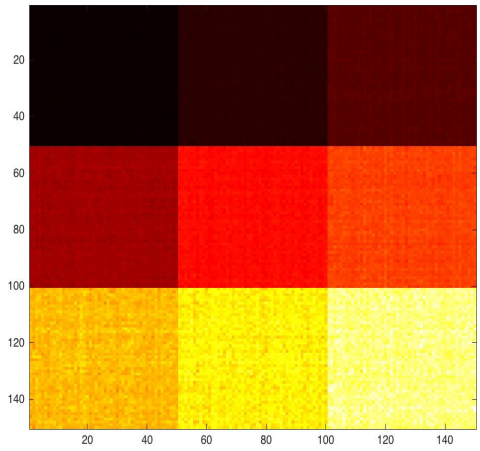


Figure 3.7: $n = 409600$

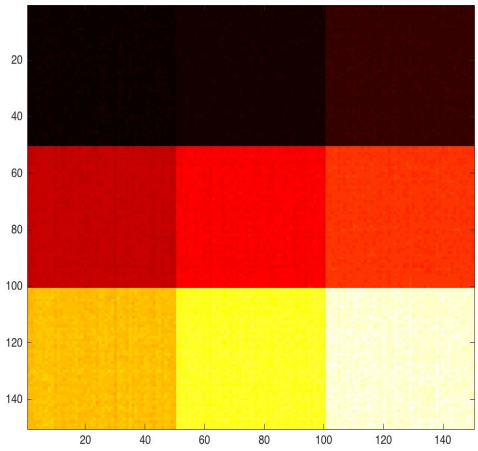


Figure 3.8: $n = 1638400$

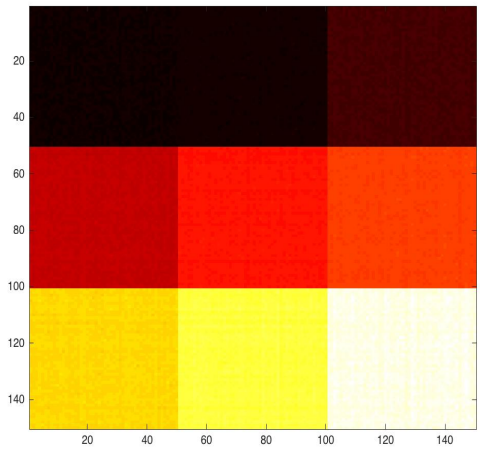


Figure 3.9: $n = 3276800$

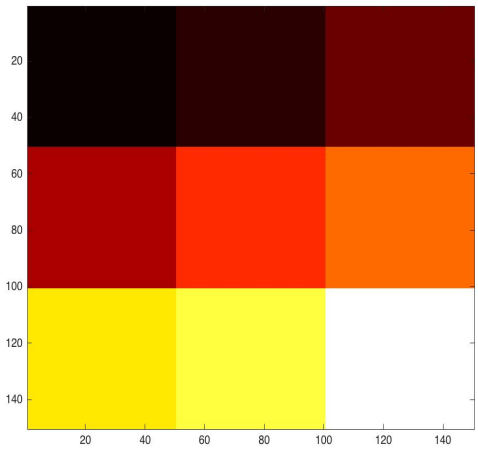


Figure 3.10: True data

tion results. The data is plotted in Fig. 3.11. As we can see, the simulation results match well with the minimax lower bound (3.1).

Sample size	800	1600	3200	6400	12800	25600
Theoretical upper bound	19.1780	11.4033	6.7805	4.0317	2.3973	1.4254
Minimax lower bound	5.1962	3.0897	1.8371	1.0924	0.6495	0.3862
Experimental error rate	7.2122	4.4569	1.9499	0.8600	0.5302	0.4329

Table 3.1: Pointwise error rate comparison with different sample size n

Sample size	51200	102400	204800	409600	819200	1638400
Theoretical upper bound	0.8476	0.5040	0.2997	0.1782	0.1059	0.0630
Minimax lower bound	0.2296	0.1365	0.0812	0.0483	0.0287	0.0171
Experimental error rate	0.2518	0.1156	0.0584	0.0466	0.0354	0.0194

Table 3.2: Pointwise error rate comparison with different sample size n

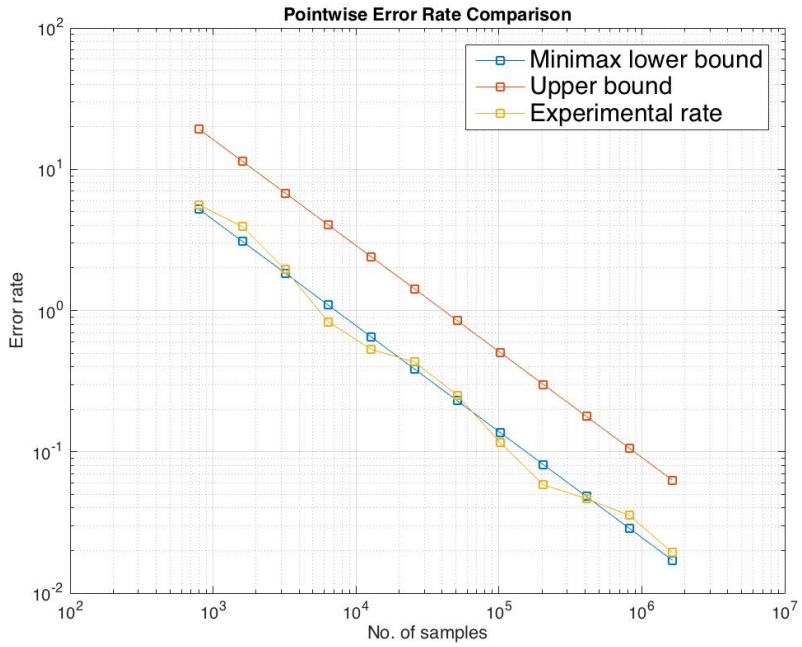


Figure 3.11: The point risk comparison at $t_0 = 0.5$

2.2 Integrated risk estimation simulation

In table 3.3 and table 3.4 we display the comparison of integrated risk measured by the L_2-norm between the theoretical bounds proved in (2.2), (3.9) and our simulation results. Since $\beta = 3/2$ and $\ell = 1$, we use piecewise linear polynomials to approximate the underlying function. The data is plotted in Fig. 3.11. As we can see, the simulation results match well with the minimax lower bound (3.9).

Sample size	800	1600	3200	6400	12800	25600
Theoretical upper bound	1683.1	1000.8	595.1	353.8	210.4	125.1
Minimax lower bound	456.0163	271.1489	161.2261	95.8656	57.0020	33.8936
Experimental error rate	457.7443	293.3489	170.4948	106.8291	57.8282	37.2912

Table 3.3: Integrated error rate comparison with different sample size n

Sample size	51200	102400	204800	409600	819200	1638400
Theoretical upper bound	74.4	44.2	26.3	15.6	9.3	5.5
Minimax lower bound	20.1533	11.9832	7.1253	4.2367	2.5192	1.4979
Experimental error rate	19.1367	11.4798	8.0132	4.0110	2.9849	1.5030

Table 3.4: Integrated error rate comparison with different sample size n

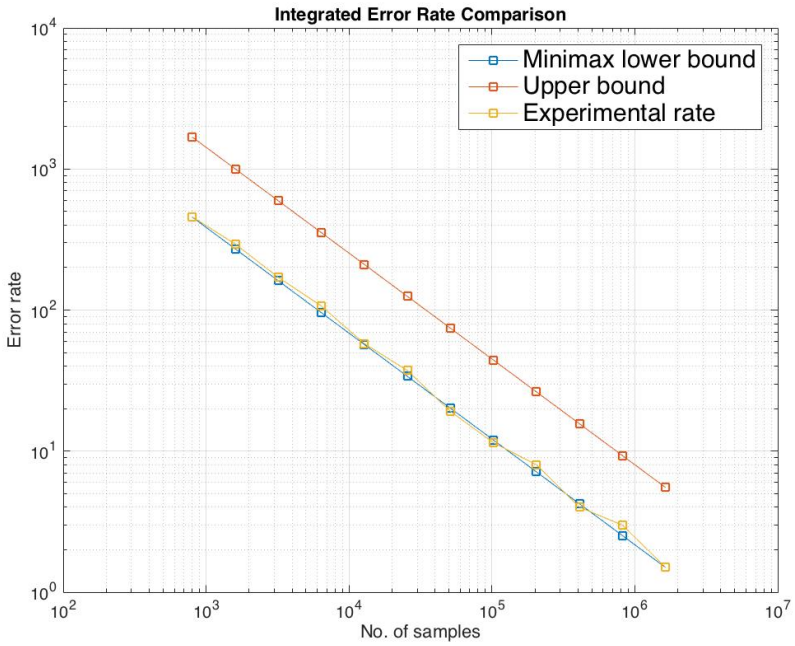


Figure 3.12: The integrated risk comparison with different sample size n

2.3 Simulation of model selection

Recall that in section 4, we emphasized that choosing a good bandwidth h is crucial to get better estimation. We developed Algorithm 1 to choose the optimal bandwidth. We implement Algorithm 1 in this section, and perform simulation with $m = 90$ and $n = 3200000$. We choose $h_{\max} = 1.0$ and $h_{\min} = 1/\sqrt{n}$ to construct the geometric grid $\mathcal{H}$ as in (4.2). We

display the simulation results in table 3.5 and 3.6. To be more specific, we computed each global estimator as in (2.1) with each bandwidth on the $\mathcal{H}$. The corresponding integrated risks measured by L_2-norm are displayed in second row and our model selection criterion computed as in (4.3) are displayed in the third row. The smaller value of the third row, the better. The data are plotted in Fig. 3.13. As we can see, our selector selects $\widehat{h} = 0.0853$ with the smallest criterion value of 0.3490. The corresponding integrated risk is also the smallest among all candidates on the grid.

Bandwidth on grid $\mathcal{H}$	1.0000	0.5000	0.2602	0.1461
Integrated risk	68.1239	45.0275	1.0207	0.0657
Model selection criterion	5.8238	4.7442	1.0100	0.3862

Table 3.5: Model Selection

Bandwidth on grid $\mathcal{H}$	0.0853	0.0510	0.0311	0.0192	0.0121
Integrated risk	0.0333	0.04371	0.0538	0.0663	0.0807
Model selection criterion	0.3490	0.4821	0.6741	0.9771	1.3199

Table 3.6: Model Selection

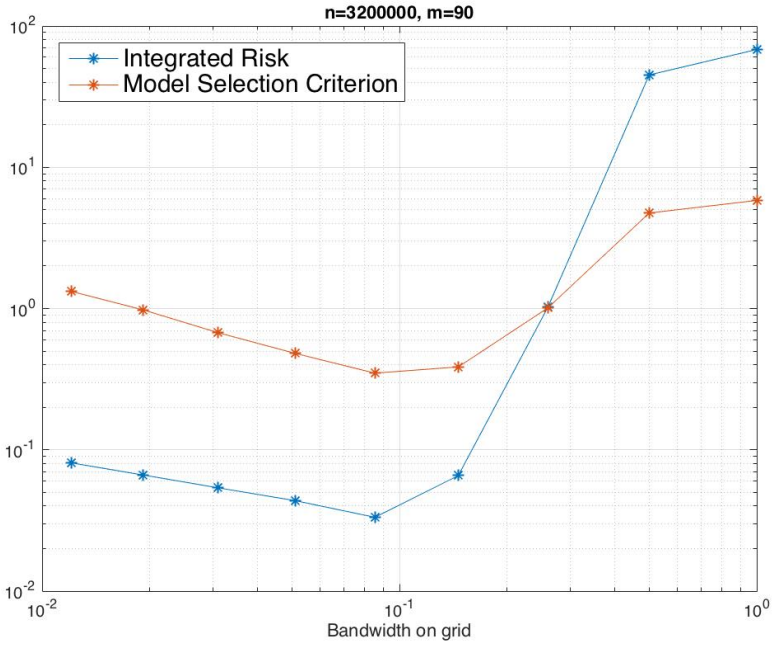


Figure 3.13: Model Selection on Grid $\mathcal{H}$

CHAPTER 4

THE ℓ_∞ PERTURBATION OF HOSVD AND LOW RANK TENSOR DENOISING

1 Introduction

A tensor is a mutliarray 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 [34], [35], [36], [37], [38]); latent variable modeling (see [39], [40], [41]); genomic signal processing ([42], [43] and [44]) 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 while it can be implemented fast for matrices, see [45].

The higher order singular value decomposition (HOSVD) is one machinery to deal with tensors which generalizes the matrix SVD to higher order tensors, see [46],[47], [48], [49] and [50]. The conceptual simplicity and computational efficiency make HOSVD popular and successful on several applications including face recognition (see [38]), genomic signal processing (see [43]) and more examples in a survey paper [51]. 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 HOSVD is appealing, there are several fundamental theoretical mysteries yet to be uncovered.

One important problem is to study the perturbation of HOSVD when stochastic noise is observed. The difficulty comes from both methodological and theoretical aspects. The computation of HOSVD is essentially reduced to matrix SVD which can be achieved efficiently. This naive estimator is actually statistically suboptimal and further power iterations

can lead to a minimax optimal estimator, see [52], [53], [54], [55] and references therein. Another intriguing phenomenon is on the signal-to-noise ratio (SNR) exhibiting distinct computational and statistical phase transitions, which do not exist for matrices. In particular, there is a gap on SNR between statistical optimality and computational optimality for HOSVD, see [53]. For introductory simplicity *, we consider the third-order tensors where 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

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

with $Z(i, j, k) \sim \mathcal{N}(0, \sigma^2)$ being i.i.d. for $i, j, k \in [d]$ and $[d] := \{1, \ldots, d\}$. The signal strength $\underline{\Lambda}(\mathbf{A})$ is defined as the smallest nonzero singular values of matricizations of $\mathbf{A}$, 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 [53] and [55]) that if the signal strength $\underline{\Lambda}(\mathbf{A}) \geq D_1\sigma d^{3/4}$ for a large enough constant $D_1 > 0$, the following bound holds

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

where $\widehat{\mathbf{U}}, \widehat{\mathbf{V}}, \widehat{\mathbf{W}}$ 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 [56]) can improve the estimate (denoted by $\widetilde{\mathbf{U}}, \widetilde{\mathbf{V}}, \widetilde{\mathbf{W}}$) to

$$r^{-1/2}\max\left\{\left\|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\right\|_{\ell_2}, \left\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^\top - \mathbf{V}\mathbf{V}^\top\right\|_{\ell_2}, \left\|\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^\top - \mathbf{W}\mathbf{W}^\top\right\|_{\ell_2}\right\} = O_p\left(\frac{\sigma d^{1/2}}{\underline{\Lambda}(\mathbf{A})}\right), \quad (1.1)$$

which is minimax optimal (see [53]). Moreover, it is demonstrated in [53] via an assump-

*Results of this paper cover the general case where $\mathbf{A}$ is $d_1 \times d_2 \times d_3$ with multilinear ranks (r_1, r_2, r_3), and can be easily generalized to higher order tensors.

tion on hypergraphical planted clique detection that if $\underline{\Lambda}(\mathbf{A}) = o\big(\sigma d^{3/4}\big)$, then all polynomial time algorithms produce trivial estimates of $\mathbf{U}, \mathbf{V}, \mathbf{W}$.

This work is focused on the estimation of linear forms of tensor singular vectors. More specifically, consider singular vectors $\mathbf{U} = \big(\mathbf{u}_1, \ldots, \mathbf{u}_r\big) \in \mathbb{R}^{d\times r}$ and our goal is to estimate $\langle \mathbf{u}_j, \mathbf{x}\rangle$ for fixed $\mathbf{x} \in \mathbb{R}^d$ and $j = 1, \ldots, r$. By choosing $\mathbf{x}$ over the canonical basis vectors in $\mathbb{R}^d$, we end up with an estimation of $\mathbf{u}_j$ whose componentwise perturbation bound can be attained. Unlike the ℓ_2-norm perturbation bound, the ℓ_∞ bound can characterize the entrywise sign consistency and entrywise significance (i.e. entrywise magnitude) of singular vectors. The componentwise signs of singular vectors have been utilized in numerous applications, such as community detection (see [57], [58], [59] and [60]). The entrywise significance is useful in submatrix localizations, see [61], [62] and references therein. We show in Section 4 that ℓ_∞ bounds require a weaker condition than ℓ_2 bounds to guarantee exact clustering in high dimensions. Furthermore, it enables us to construct a low rank estimator of $\mathbf{A}$ with a sharp bound on $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{\ell_\infty}$. To the best of our knowledge, ours is the first result concerning the low rank tensor denoising with sharp ℓ_∞ bound.

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

$$\mathbf{A} = \sum_{k=1}^{r} \lambda_k \big(\mathbf{u}_k \otimes \mathbf{v}_k \otimes \mathbf{w}_k\big), \quad \lambda_1 \geq \ldots \geq \lambda_r > 0 \tag{1.2}$$

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 $d \times r$ orthonormal matrices. 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 matricization (see Section 2) and we preset $\lambda_0 = +\infty$ and $\lambda_{r+1} = 0$. We show that if $\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big) \geq D_1\big(\sigma\lambda_1 d^{1/2} + \sigma^2 d^{3/2}\big)$, the following bound holds for any $\mathbf{x} \in \mathbb{R}^d$,

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

where $b_k \in [-1/2, 0]$ is an absolute constant which does not depend on $\mathbf{x}$.

If $r = 1$ (rank one spiked tensor PCA model, see [52]) such that $\underline{\Lambda}(\mathbf{A}) = \bar{g}_1\big(\mathcal{M}_1(\mathbf{A})\big) = \lambda_1$, we get

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

By taking $\mathbf{x}$ over the canonical basis vectors in $\mathbb{R}^d$, the above fact implies that

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

under the eigengap condition $\lambda_1 \geq D_1\sigma d^{3/4}$ which is a standard requirement in tensor PCA [†]. Moreover, a low rank estimator (denoted by $\widehat{\mathbf{A}}$) is constructed under the same conditions such that

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

implying that the ℓ_∞ bound is determined by the coherence $\max\big\{\|\mathbf{u}_1\|_{\ell_\infty}, \|\mathbf{v}_1\|_{\ell_\infty}, \|\mathbf{w}_1\|_{\ell_\infty}\big\}$.

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 ([64]) 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 [65], [66] and references therein, which provide sharp perturbation bounds in ℓ_2-norm. There are recent bounds (see [67] and [68]) in ℓ_∞-norm developed under additional constraint (incoherent singular spaces) and structural noise (sparse noise). To obtain a sharp ℓ_∞-norm bound, we borrow the instruments invented by [69] and extensively applied in [63]. Our framework is built upon a second order method of estimating the singular

[†]We shall point out that a similar result on matrix SVD has appeared in [63] which is suboptimal for tensors. Indeed, the result in [63] is established under the eigengap condition $\lambda_1 \geq D_1\sigma d$.

subspaces, which improves the eigengap requirement than the first order method. Similar techniques have been proposed for solving tensor completion ([70]) and tensor PCA ([55]). The success of this seemingly natural treatment hinges upon delicate dealing with the correlations among higher order terms. We benefit from these ℓ_∞-norm spectral bound by proposing a low rank estimator for tensor denoising such that entrywise perturbation is guaranteed through the tensor incoherence.

We organize this chapter 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 on singular vector perturbation and the 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. The proofs are provided in Section 5.

2 Preliminaries on Tensor and HOSVD

2.1 Notations

We first review some notations which 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 $\mathcal{M}_1 : \mathbb{R}^{d_1 \times d_2 \times d_3} \mapsto \mathbb{R}^{d_1 \times (d_2 d_3)}$ 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 $\mathbf{A}$. 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 through 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 $\mathbf{r}(\mathbf{A}) := \big(r_1(\mathbf{A}), r_2(\mathbf{A}), r_3(\mathbf{A})\big)$.

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

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

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

$$\mathbf{A} = \mathbf{C} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W} \tag{2.1}$$

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:

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

where $\mathbf{Z} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is a noise tensor with i.i.d. entries satisfying $Z(i,j,k) \sim \mathcal{N}(0, \sigma^2)$. By observing $\mathbf{Y}$, the goal is to estimate $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$. An immediate solution is to compute HOSVD of $\mathbf{Y}$. To this end, let $\widehat{\mathbf{U}} \in \mathbb{R}^{d_1 \times r_1}$, $\widehat{\mathbf{V}} \in \mathbb{R}^{d_2 \times r_2}$, $\widehat{\mathbf{W}} \in \mathbb{R}^{d_3 \times r_3}$ 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 of $\widehat{\mathbf{U}}, \widehat{\mathbf{V}}$ and $\widehat{\mathbf{W}}$ is the so-called eigengap.

Observe that the computing of $\widehat{\mathbf{U}}$ is essentially via matrix SVD on $\mathcal{M}_1(\mathbf{A})$. It suffices to consider eigengaps for 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 singular values $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ and $\{\mathbf{g}_1, \ldots, \mathbf{g}_r\}$ are the corresponding left singular vectors and $\{\mathbf{h}_1, \ldots, \mathbf{h}_r\}$ are its corresponding right singular vectors. Introduce further $\lambda_0 = +\infty$ 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 $\mathbf{U}, \widehat{\mathbf{U}} \in \mathbb{R}^{d_1 \times r_1}$ are the top-r_1 left singular vectors of $\mathcal{M}_1(\mathbf{A})$ and $\mathcal{M}_1(\mathbf{Y})$ respectively. By Wedin's $\sin\Theta$ theorem ([64]),

$$\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \mathbf{U}\mathbf{U}^\top\| = O\bigg(\frac{\|\mathcal{M}_1(\mathbf{Z})\|}{\bar{g}_{r_1}\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big)}\bigg), \tag{2.2}$$

which is generally suboptimal 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 [65] and [66]. In this paper, we focus on the perturbation bound in ℓ_∞-norm. To this end, write $\mathbf{U} = \big(\mathbf{u}_1, \ldots, \mathbf{u}_r\big)$ and $\widehat{\mathbf{U}} = \big(\widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_r\big)$. We are interested in the perturbation of linear forms $\langle \widehat{\mathbf{u}}_k, \mathbf{x}\rangle$ for $\mathbf{x} \in \mathbb{R}^{d_1}$. Similar results can be obtained for singular vectors $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{W}}$.

3 Main Results

3.1 Second Order Spectral Analysis

The ℓ_∞-norm spectral perturbation for balanced matrices has been developed in [63]. Recall that $\mathbf{u}_k$ denotes the k-th left singular vector of $\mathcal{M}_1(\mathbf{A})$ and $\widehat{\mathbf{u}}_k$ 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 determined by the larger dimension $(d_1 \vee d_2 d_3)$, see Section 5. It turns out that the machinery in [63] is suboptimal meaning that the eigengap requirement becomes $\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big) \geq D_1\sigma\big(d_1 \vee d_2 d_3\big)^{1/2}$, which shall be unnecessarily strong in view of the recent results in [66], [53] and [55].

In this paper, we conduct a second order spectral analysis for $\widehat{\mathbf{U}}$. Basically, 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 spectral perturbation on $\mathcal{M}_1(\mathbf{Y})\mathcal{M}_1^\top(\mathbf{Y})$ instead of on $\mathcal{M}_1(\mathbf{Y})$. Clearly,

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

where $\mathbf{\Gamma} = \mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{Z}) + \mathcal{M}_1(\mathbf{Z})\mathcal{M}_1^\top(\mathbf{A}) + \mathcal{M}_1(\mathbf{Z})\mathcal{M}_1^\top(\mathbf{Z})$. Note that $\mathbf{U}$ are the leading eigenvectors of $\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})$ and $\widehat{\mathbf{U}}$ are the top-r_1 eigenvectors of $\mathcal{M}_1(\mathbf{Y})\mathcal{M}_1^\top(\mathbf{Y})$. Moreover, the following fact 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 our 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 5)

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

Note that subtracting by an identity matrix does not change 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 $D_1\sigma\sqrt{d_1}\big\|\mathcal{M}_1(\mathbf{A})\big\|$ with high probability, which creates a constraint on the condition number of $\mathcal{M}_1(\mathbf{A})$. Moreover, in order to characterize a sharp perturbation bound of linear forms $\langle\widehat{\mathbf{u}}_k, \mathbf{x}\rangle$, we need to pay more attention to dealing with correlations among the higher order terms than the first order method in [63].

3.2 Perturbation of Linear Forms of Singular Vectors

In this section, we present our main theorem characterizing the perturbation of linear forms $\langle\widehat{\mathbf{u}}_k, \mathbf{x}\rangle$ for any $\mathbf{x} \in \mathbb{R}^{d_1}$, where $\widehat{\mathbf{u}}_k$ is the k-th left singular vector of $\mathcal{M}_1(\mathbf{Y})$. Our results have similar implications as the previous work [63], meaning that the bias $\mathbb{E}\widehat{\mathbf{u}}_k - \mathbf{u}_k$ is well aligned with $\mathbf{u}_k$. Therefore, by 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, 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.

Theorem 3.1. *Let*[‡] $\mathbf{M} := \mathcal{M}_1(\mathbf{A})$ *and* $\delta(d_1, d_2, d_3) = \sigma d_1^{1/2}\|\mathbf{M}\| + \sigma^2(d_1 d_2 d_3)^{1/2}$ *and suppose* $d_2 d_3 e^{-d_1/2} \le 1$. *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*

[‡]Observe that if we set $d_3 = 1$ and consider the case with $d_1 \ll d_2$, then Theorem 3.1 elaborates the one-sided spectral perturbation in ℓ_∞-norm for unbalanced (or fat) matrices.

any $\mathbf{x}$*, the following bound holds with probability at least* $1-e^{-t}$*,*

$$\begin{aligned}&\big|\langle\widehat{\mathbf{u}}_k,\mathbf{x}\rangle-(1+b_k)^{1/2}\langle\mathbf{u}_k,\mathbf{x}\rangle\big|\\&\quad\leq D_2\bigg(t^{1/2}\frac{\sigma\|\mathbf{M}\|+\sigma^2(d_2d_3)^{1/2}}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}+\frac{\sigma^2 d_1}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}\Big(\frac{\delta(d_1,d_2,d_3)}{\bar{g}_k(\mathbf{M}\mathbf{M}^\top)}\Big)\bigg)\|\mathbf{x}\|_{\ell_2}\end{aligned}\tag{3.1}$$

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

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

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

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

If $\kappa\big(\mathcal{M}_1(\mathbf{A})\big)\leq\big(\frac{d_2d_3}{d_1}\big)^{1/4}$, the above bound becomes $\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\big)\geq D_1\sigma(d_1d_2d_3)^{1/4}$ which is a standard requirement in tensor SVD or PCA, see [53], [54] and [52]. 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 $\widehat{\mathbf{u}}_k$.

Corollary 3.1. Under the conditions in Theorem 3.1, there exists a universal constant $D_1>0$ such that the following bound holds with probability at least $1-\frac{1}{d_1}$,

$$\big\|\widehat{\mathbf{u}}_k-(1+b_k)^{1/2}\mathbf{u}_k\big\|_{\ell_\infty}\leq D_1\bigg(\Big(\frac{\log d_1}{d_1}\Big)^{1/2}+\Big(\frac{d_1}{d_2d_3}\Big)^{1/2}\bigg).$$

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

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

which has an analogous form to the perturbation bound in [63] implying a famous delo-

calization phenomenon in random matrix theory, see [71] and [72] and references therein. The bias b_k is usually unknown and we borrow the idea in [63] to estimate b_k based on two independent samples, which happens in the application of tensor decomposition for gene expression, usually multiple independent copies are available, see [73].

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. Let $\widehat{\mathbf{u}}_k^{(1)}$ and $\widehat{\mathbf{u}}_k^{(2)}$ 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 $\widehat{\mathbf{u}}_k^{(1)}$ and $\widehat{\mathbf{u}}_k^{(2)}$ are chosen such that $\langle \widehat{\mathbf{u}}_k^{(1)}, \widehat{\mathbf{u}}_k^{(2)} \rangle \geq 0$. Define the estimator of b_k by

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

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

Theorem 3.2. *Under the assumptions in Theorem 3.1, 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 $t \geq 0$,*

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

and

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

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

Remark 1. If $d/2 \leq \min_k d_k \leq \max_k d_k \leq 2d$, we get

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

Moreover, if $\mathrm{rank}(\mathbf{A}) = (1,1,1)$, we can write $\|\widetilde{\mathbf{u}}_1 - \mathbf{u}_1\|_{\ell_\infty} = O_p\Big(\frac{\sigma \log^{1/2} d}{\underline{\Lambda}(\mathbf{A})} + \frac{\sigma^2 d \log^{1/2} d}{\underline{\Lambda}^2(\mathbf{A})}\Big)$ where $\underline{\Lambda}(\mathbf{A}) = \lambda_{\min}\big(\mathcal{M}_1(\mathbf{A})\big)$. Note that $\|\widetilde{\mathbf{u}}_1 - \mathbf{u}_1\|_{\ell_2} = O_p\Big(\frac{\sigma d^{1/2}}{\underline{\Lambda}(\mathbf{A})} + \frac{\sigma^2 d^{3/2}}{\underline{\Lambda}^2(\mathbf{A})}\Big)$, see [53]. Therefore, our one-sided SVD perturbation bound in ℓ_∞-norm for a matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ is optimal if it is ultra-fat such that $d_1^2 \leq d_2$.

3.3 Low Rank Tensor Denoising ℓ_∞ Bound

In this section, we consider low rank estimate of $\mathbf{A}$ through 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 3.2. 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 entrywise 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 on $\tilde{\mathbf{u}}_k, \tilde{\mathbf{v}}_k, \tilde{\mathbf{w}}_k$. We shall need the following conditions on the eigengaps: for a large enough constant $D_1 > 0$,

$$\bar{g}_k\big(\mathcal{M}_1(\mathbf{A})\mathcal{M}_1^\top(\mathbf{A})\big) \geq D_1\Big(\sigma d_1^{1/2}\overline{\Lambda}(\mathbf{A}) + \sigma^2(d_1 d_2 d_3)^{1/2}\Big), \quad 1 \leq k \leq r_1, \tag{3.2}$$

$$\bar{g}_k\big(\mathcal{M}_2(\mathbf{A})\mathcal{M}_2^\top(\mathbf{A})\big) \geq D_1\Big(\sigma d_2^{1/2}\overline{\Lambda}(\mathbf{A}) + \sigma^2(d_1 d_2 d_3)^{1/2}\Big), \quad 1 \leq k \leq r_2, \tag{3.3}$$

$$\bar{g}_k\big(\mathcal{M}_3(\mathbf{A})\mathcal{M}_3^\top(\mathbf{A})\big) \geq D_1\Big(\sigma d_3^{1/2}\overline{\Lambda}(\mathbf{A}) + \sigma^2(d_1 d_2 d_3)^{1/2}\Big), \quad 1 \leq k \leq r_3, \tag{3.4}$$

where

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

Similarly, define

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

and the overall eigengap

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

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

Theorem 3.3. *Suppose conditions (3.2) (3.3) (3.4) hold and assume that for all* $i \in [d_1], j \in [d_2], k \in [d_3]$,

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

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

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

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

Remark 2. To highlight the contribution of Theorem 3.3, 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 $\bar{g}_{\min}(\mathbf{A}) \gtrsim \sigma d^{3/4}$, then

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

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 [53]. Theorem 3.3 is more interesting when $\mathbf{A}$ is incoherent such that $\mu_{\mathbf{U}}, \mu_{\mathbf{V}}, \mu_{\mathbf{W}} = O(1)$. We conclude that

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

4 Applications

In this section, we review two applications of ℓ_∞-norm. It is interesting to observe that it is unnecessary to estimate the bias b_k in these applications.

4.1 High Dimensional Clustering

Many statistical and machine learning tasks are associated with clustering high dimensional data, see [74], [75], [76], [77], [78] 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

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

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 $p \gg n$.

Given the data matrix

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

the goal is to conduct bi-clustering. Let $n_k := \mathrm{Card}\big(\{1 \le i \le n : \ell_i = k\}\big)$ for $k = 0, 1$ such that $n_0 + n_1 = n$. 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 produce the cluster membership. Moreover, the leading singular value of $\mathbb{E}\mathbf{Y}$ is $n^{1/2}\|\boldsymbol{\beta}\|_{\ell_2}$. Let $\widehat{\mathbf{u}}$ denotes the leading left singular vector of $\mathbf{Y}$. By Corollary 3.1, 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

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

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

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

implying that if $\ell_i = \ell_j$, then $\operatorname{sign}\big(\widehat{u}(i)\big) = \operatorname{sign}\big(\widehat{u}(j)\big)$ for all $1 \le i, j \le n$. The above analysis also implies that it is unnecessary to estimate b_k in this application, since scaling will not change the entrywise signs. Therefore, in order to guarantee exact clustering, the ℓ_∞ bound requires

$$\|\boldsymbol{\beta}\|_{\ell_2} \ge D_1\Big(n^{1/6} \vee p^{1/8} \vee \big(p\log(n)/n\big)^{1/4}\Big),$$

while the ℓ_2 bound in [66] requires

$$\|\boldsymbol{\beta}\|_{\ell_2} \ge D_1\big(n^{1/2} \vee p^{1/4} \vee (p/n)^{1/4}\big)$$

for exact clustering.

Remark 3. 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 $\mathrm{rank}(\mathbf{M}) \leq k$, 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 clustering membership. By Theorem 3.1, if $\Delta_j \geq C_1\sqrt{k}\|\boldsymbol{\beta}_1\|_{\ell_2} + C_2(kp/n)^{1/2}$ where $\Delta_j = \min\{\big(\|\boldsymbol{\beta}_j\|_{\ell_2}^2 - \|\boldsymbol{\beta}_{j+1}\|_{\ell_2}^2\big), \big(\|\boldsymbol{\beta}_{j-1}\|_{\ell_2}^2 - \|\boldsymbol{\beta}_j\|_{\ell_2}^2\big)\}$, then

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

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

4.2 Subtensor Localization

In gene expression association analysis (see [73], [79], [80] and [81]) and planted clique detection (see [82], [83] and [84]), the goal is equivalent to localizing a sub-tensor whose entries are statistically more significant than the others. One simple model characterizing this type of tensor data is as

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

with $C_k = \cup_{j=1}^{b_k} C_k^{(j)} \subset [d_k]$ where $\big\{C_k^{(1)}, \ldots, C_k^{(b_k)}\big\}$ are disjoint subsets of $[d_k]$ for $k = 1, 2, 3$, i.e., there are b_k dense blocks in the k-th direction. Then, in total, there are $b_1b_2b_3$ dense blocks in $\mathbb{E}\mathbf{Y}$. The vector $\mathbf{1}_{C_k} \in \mathbb{R}^{p_k}$ 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 of $\mathbf{Y}$, the goal is to localize the unknown subsets $\{C_1^{(j)}\}_{j=1}^{b_1}, \{C_2^{(j)}\}_{j=1}^{b_2}$ and $\{C_3^{(j)}\}_{j=1}^{(b_3)}$. 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

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

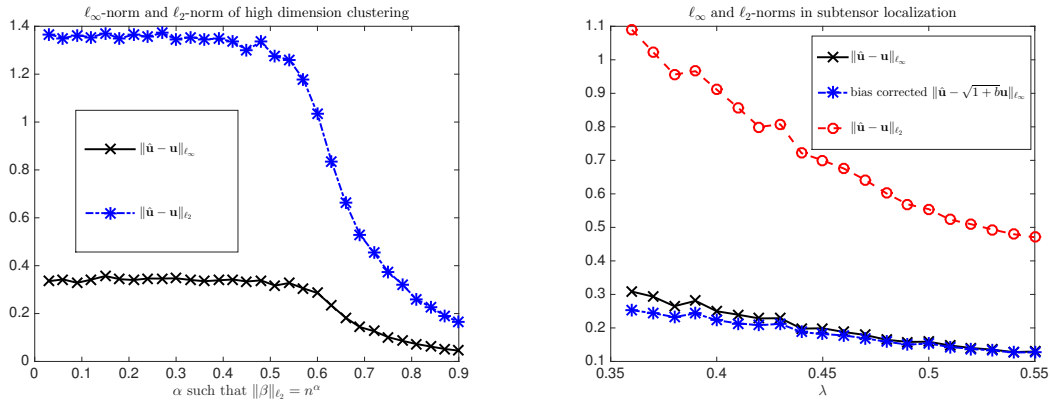
where $|C|$ denotes the cardinality of C. By Theorem 3.1, if $\lambda \geq D_1 \frac{(d_1d_2d_3)^{1/4}}{|C_1|^{1/2}|C_2|^{1/2}|C_3|^{1/2}}$ for a large enough constant $D_1 > 0$, then with probability at least $1 - \frac{1}{d_{\max}}$ where $d_{\max} := (d_1 \vee d_2 \vee d_3)$ and we assume $d_{\max} \leq D_1(d_1d_2d_3)^{1/2}$,

$$\begin{aligned}
&\|\widehat{\mathbf{u}} - (1+b_1)^{1/2}\mathbf{u}\|_{\ell_\infty} \\
&\quad \leq \frac{D_1}{\lambda|C_1|^{1/2}|C_2|^{1/2}|C_3|^{1/2}} + \frac{D_1(d_2d_3)^{1/2}}{\lambda^2|C_1||C_2||C_3|} + \frac{D_1d_1}{\lambda^2|C_1||C_2||C_3|}\left(\frac{(d_1d_2d_3)^{1/2}}{\lambda^2|C_1||C_2||C_3|}\right). \\
&\quad \leq D_1\left(\frac{1}{d_1^{1/2}} + \left(\frac{d_1}{d_2d_3}\right)^{1/2}\right). \qquad (4.1)
\end{aligned}$$

If we let $\widehat{C}_1$ denote the locations of entries of $\widehat{\mathbf{u}}$ whose magnitudes are among the $|C_1|$ largest, it is straightforward to see that $\widehat{C}_1 = C_1$ on the above event if $D_2|C_1|d_1 \leq d_2d_3$ for a large enough constant $D_2 > 0$. Note that it is also unnecessary to estimate b_1 if we are only interested in the top-$|C_1|$ largest entries of $|\widehat{\mathbf{u}}|$.

4.3 Numerical Experiments

We present simulation results of experiments on the above applications. In high dimensional clustering, we randomly sample a vector $\boldsymbol{\beta} \in \mathbb{R}^p$ with $p = 1000$. Fixed $\boldsymbol{\beta}$, $n/2 = 50$ random vectors are sampled from distribution $\mathcal{N}(\boldsymbol{\beta}, \mathbf{I}_p)$ and $n/2 = 50$ random vectors are sampled from distribution $\mathcal{N}(-\boldsymbol{\beta}, \mathbf{I}_p)$. Then, we compare between the leading left singular



(a) High dimension clustering: a significant gap between ℓ_∞-norm and ℓ_2-norm.

(b) Subtensor localization: the bias correction indeed improves the ℓ_∞ norm.

Figure 4.1: Comparison on ℓ_∞-norm, ℓ_2-norm and bias corrected ℓ_∞ norm in high dimension clustering and subtensor localization.

vector of $\mathbf{Y}$ and leading left singular vector of $\mathbb{E}\mathbf{Y}$, i.e., $\|\widehat{\mathbf{u}} - \mathbf{u}\|_\infty$ and $\|\widehat{\mathbf{u}} - \mathbf{u}\|_{\ell_2}$, without bias correction. For each $\|\boldsymbol{\beta}\|_{\ell_2} \in [n^{0.03}, n^{0.9}]$, the loss is reported by averaging 30 independent simulations. The results are displayed in Figure 4.1a where we can observe a significant gap between ℓ_2-norm and ℓ_∞-norm. It explains why ℓ_∞-norm is more powerful for exact clustering than ℓ_2-norm in this application.

In subtensor localization, we fix $d_1 = d_2 = d_3 = 100$ and $C_1 = C_2 = C_3 = [20]$, i.e., the subtensor is the bottom-left-front corner of $\mathbb{E}\mathbf{Y}$. The λ is varied from 0.36 to 0.55. For each λ, we report the average ℓ_∞-norm, ℓ_2-norm and bias corrected ℓ_∞-norm, all from 30 independent simulations. It is interesting to observe that actually the bias correction indeed can improve the ℓ_∞-norm when λ is small. The results are displayed in Figure 4.1b.

5 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 $m_1 \lesssim m_2$. 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 $\{\mathbf{h}_1, \ldots, \mathbf{h}_{r_1}\} \subset \operatorname{span}\big\{\mathbf{v}_j \otimes \mathbf{w}_k^\top : j \in [r_2], k \in [r_3]\big\}$ 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

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

Even though Theorem 3.1 and Theorem 3.2 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 $s \leq r_1$. Denote $\Delta_k := \{j : \lambda_j = \mu_k\}$ for $1 \leq k \leq s$ and $\nu_k := \operatorname{Card}(\Delta_k)$ 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 $\mathbf{Y}\mathbf{Y}^\top$,

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

We develop a sharp concentration bound for bilinear forms $\langle \widehat{\mathbf{P}}_k^{uu}\mathbf{x}, \mathbf{y}\rangle$ 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 $\widehat{\mathbf{\Gamma}} := \mathbf{\Gamma} - m_2\sigma^2\mathbf{I}_{m_1}$ and the spectral analysis shall be realized on $\mathbf{A}\mathbf{A}^\top + \widehat{\mathbf{\Gamma}}$.

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 3 is provided in the Appendix.

Lemma 3. For any deterministic matrix $\mathbf{B} \in \mathbb{R}^{m_3\times m_2}$, the following bounds hold

$$\begin{aligned} \mathbb{E}\|\mathbf{B}\mathbf{Z}^\top\| &\lesssim \sigma\|\mathbf{B}\|\Big(m_1^{1/2} + m_3^{1/2} + (m_1m_3)^{1/4}\Big) \qquad (5.1)\\ \big\|\mathbb{E}\mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1}\big\| &\lesssim \sigma^2(m_1m_2)^{1/2}. \end{aligned}$$

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

$$\begin{aligned} \|\mathbf{B}\mathbf{Z}^\top\| &\lesssim \sigma\|\mathbf{B}\|\Big(m_1^{1/2} + m_3^{1/2} + (m_1m_3)^{1/4} + t^{1/2} + (m_1t)^{1/4}\Big) \qquad (5.2)\\ \big\|\mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1}\big\| &\lesssim \sigma^2 m_2^{1/2}\big(m_1^{1/2} + t^{1/2}\big). \end{aligned}$$

5.1 Proof of Theorem 3.1

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 3.1 is decomposed of two separate components. Theorem 5.1 provides the concentration bound for $\left|\langle \mathbf{P}_k \mathbf{x}, \mathbf{y} \rangle - \mathbb{E}\langle \mathbf{P}_k \mathbf{x}, \mathbf{y} \rangle\right|$ by Gaussian isoperimetric inequality and the proof is postponed to the Appendix. In Theorem 5.2, we characterize the bias $\mathbb{E}\widehat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu}$.

Theorem 5.1. *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* $1 - e^{-t}$*,*

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

The following spectral representation formula is needed whose proof can be found in [69].

Lemma 4. The following bound holds

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

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

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

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

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

Theorem 5.2. *Let $\delta(m_1, m_2) := \mu_1\sigma m_1^{1/2} + \sigma^2(m_1m_2)^{1/2}$ 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 $m_2e^{-m_1/2} \leq 1$. Then there exists an absolute constant $D_2 > 0$ such that*

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

Proof of Theorem 3.1. Combining Theorem 5.1 and Theorem 5.2, we conclude that for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m_1}$ with probability at least $1 - e^{-t}$ for all $\log 8 \leq t \leq m_1$,

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

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

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

where

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

Moreover, a simple fact is $b_k \leq \mathbb{E}\|\widehat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu}\| \lesssim \frac{\delta(m_1,m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}$ by Wedin's sinΘ theorem ([64]). If $\bar{g}_k(\mathbf{A}\mathbf{A}^\top) \geq D\delta(m_1, m_2)$ for a large enough constant $D > 0$, we can ensure $b_k \in [-1/2, 0]$. Then, with probability at least $1 - e^{-t}$,

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

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

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

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 $\mathcal{E}_1$, $\langle \widehat{\mathbf{u}}_k, \mathbf{u}_k \rangle^2 \geq \frac{1}{4}$. Then, on event $\mathcal{E}_1$,

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

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

Proof of Theorem 5.2. Recall the representation formula of $\widehat{\mathbf{P}}_k^{uu}$ in Lemma 4 that

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

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

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

such that we can write $\mathbb{E}\widehat{\mathbf{P}}_k^{uu} = \mathbf{P}_k^{uu} + \mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) + \Big(\mathbb{E}\mathbf{S}_k(\widehat{\mathbf{\Gamma}}) - \mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) \Big)$. We derive an upper bound on $\big\| \mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbb{E}\mathbf{S}_k(\widehat{\mathbf{\Gamma}}) \big\|$ and the proof can be found in the Appendix. Lemma 5 implies that our analysis can be proceeded by replacing $\widehat{\mathbf{\Gamma}}$ with $\widetilde{\mathbf{\Gamma}}$.

Lemma 5. There exists a universal constant $D_1 > 0$ such that if $m_2 e^{-m_1/2} \le 1$, then

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

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

$$\begin{aligned}
&\mathbb{E}\widehat{\mathbf{P}}_k^{uu} - \mathbf{P}_k^{uu} - \mathbf{P}_k^{uu}\mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu} \\
&\quad = \mathbb{E}\mathbf{S}_k(\widehat{\mathbf{\Gamma}}) - \mathbb{E}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) \\
&\quad + \mathbb{E}\Big(\mathbf{P}_k^{uu}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})(\mathbf{P}_k^{uu})^\perp + (\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu} + (\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})(\mathbf{P}_k^{uu})^\perp\Big)\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\| \le \delta_t\big) \\
&\quad + \mathbb{E}\Big(\mathbf{P}_k^{uu}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})(\mathbf{P}_k^{uu})^\perp + (\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu} + (\mathbf{P}_k^{uu})^\perp\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})(\mathbf{P}_k^{uu})^\perp\Big)\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\| > \delta_t\big).
\end{aligned} \tag{5.3}$$

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 [63], 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,

$$
\begin{aligned}
&(\mathbf{P}_k^{uu})^{\perp}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu}\\
&=-\sum_{r\geq 2}(-1)^r\frac{1}{2\pi i}\oint_{\gamma_k}\bigg(\frac{1}{(\mu_k^2-\eta)^2}\sum_{s=2}^{r}\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-1}\big(\mathbf{P}_k^{uu}\widetilde{\mathbf{\Gamma}}\big)\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\\
&\quad+\frac{1}{\mu_k^2-\eta}\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^r\mathbf{P}_k^{uu}\bigg)d\eta. \qquad (5.4)
\end{aligned}
$$

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

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

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

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

Observe that

$$
\big|\big\langle\widetilde{\mathbf{\Gamma}}\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y},\mathbf{u}_j\big\rangle\big|\leq\|\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\|^{r-s}\|\widetilde{\mathbf{\Gamma}}\|^{r-s+1}\|\mathbf{y}\|_{\ell_2}
$$

$$\leq \Big(\frac{2}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{(r-s)} \|\widetilde{\mathbf{\Gamma}}\|^{r-s+1}\|\mathbf{y}\|_{\ell_2}.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\Big\langle \mathbf{x}, \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-1}\big(\mathbf{P}_k^{uu}\widetilde{\mathbf{\Gamma}}\big)\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y}\Big\rangle \mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| \leq \delta_t\Big) \\
&= \sum_{j\in\Delta_k} \mathbb{E}\big\langle \widetilde{\mathbf{\Gamma}}\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y}, \mathbf{u}_j\big\rangle\big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-1}\mathbf{u}_j, \mathbf{x}\big\rangle \mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| \leq \delta_t\Big) \\
&\leq \sum_{j\in\Delta_k} \mathbb{E}^{1/2}\Big|\big\langle \widetilde{\mathbf{\Gamma}}\big(\mathbf{R}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{r-s}\mathbf{P}_k^{uu}\mathbf{y}, \mathbf{u}_j\big\rangle \mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| \leq \delta_t\Big)\Big|^2 \\
&\qquad\qquad \times \mathbb{E}^{1/2}\Big|\big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-1}\mathbf{u}_j, \mathbf{x}\big\rangle \mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| \leq \delta_t\Big)\Big|^2 \\
&\leq \Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{r-s}\delta_t\|\mathbf{y}\|_{\ell_2} \sum_{j\in\Delta_k} \mathbb{E}^{1/2}\Big|\big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\mathbf{u}_j, \mathbf{x}\big\rangle \mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\| \leq \delta_t\Big)\Big|^2.
\end{aligned} \tag{5.5}$$

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

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

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

$$\begin{aligned}
&\Big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\mathbf{u}_j, \mathbf{x}\Big\rangle \\
&\qquad = \Big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{Z}\mathbf{A}^\top\mathbf{u}_j, \mathbf{x}\Big\rangle
\end{aligned} \tag{5.6}$$

$$+\Big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{A}\mathbf{Z}^\top\mathbf{u}_j, \mathbf{x}\Big\rangle \tag{5.7}$$

$$+\Big\langle \big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\Big(\mathbf{Z}\sum_{k'\neq k}\mathbf{P}^{hh}_{k'}\mathbf{Z}^\top\mathbf{u}_j - \sigma^2(m_2-\nu_k)\mathbf{u}_j\Big), \mathbf{x}\Big\rangle. \tag{5.8}$$

The upper bounds of (5.6), (5.7), and (5.8) 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 $\{\mathbf{e}_1, \ldots, \mathbf{e}_{m_1}\}$ denote the canonical basis vectors in $\mathbb{R}^{m_1}$ and $\{\mathbf{z}_1^\top, \ldots, \mathbf{z}_{m_1}^\top\}$ denote the rows of $\mathbf{Z}$. Therefore,

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

It is clear that $\langle \mathbf{z}_i, \mathbf{h}_j\rangle, i = 1, \ldots, m_1$ are i.i.d. and $\langle \mathbf{z}_i, \mathbf{h}_j\rangle \sim \mathcal{N}(0, \sigma^2)$. Recall that $\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta) = \sum_{k'\neq k}\frac{\mathbf{P}^{uu}_{k'}}{\mu_{k'}^2-\eta}$, 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

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

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

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

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

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

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,

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

As a result,

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

$$\leq \frac{\sigma\mu_k}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\Big)^{s-2}\|\mathbf{x}\|_{\ell_2}.$$

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

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

To this end, we write

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

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

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

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

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

Bound of $\mathbb{E}^{1/2}\Big|\Big\langle\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\widetilde{\mathbf{\Gamma}}\big)^{s-2}\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\big(\mathbf{Z}\sum_{k'\neq k}\mathbf{P}^{hh}_{k'}\mathbf{Z}^\top\big)\mathbf{u}_j,\mathbf{x}\Big\rangle\Big|^2\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\|\leq\delta_t\big)$. Note that we used the fact $\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^\top}(\eta)\mathbf{u}_j=\mathbf{0}$ in (5.8). Again, let $\{\mathbf{z}_1,\ldots,\mathbf{z}_{m_2}\}\subset\mathbb{R}^{m_1}$ denote the

corresponding columns of $\mathbf{Z}$. We write

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

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,

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

where we used the fact $\mathbb{E}^{1/2}\big\|(\sum_{k'\neq k}\mathbf{P}_{k'}^{hh})\mathbf{Z}^\top\big\|^2\lesssim\sigma m_2^{1/2}$ from Lemma 3.

Finalize the proof of Theorem. Combining the above bounds into (5.7), (5.6) and (5.8), we conclude that

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

Continue from (5.5) and we end up with

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

Plug the bounds into (5.4),

$$\begin{aligned}
&\big|\mathbb{E}\big\langle(\mathbf{P}_k^{uu})^{\perp}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu}\mathbf{y},\mathbf{x}\big\rangle\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\le\delta_t\|\big)\big|\\
&\quad\lesssim\sum_{r\ge2}\frac{\pi\bar{g}_k(\mathbf{A}\mathbf{A}^{\top})}{2\pi}\Big(\frac{2}{\bar{g}_k(\mathbf{A}\mathbf{A}^{\top})}\Big)^2(r-1)\nu_k\delta_t\frac{\sigma^2m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^{\top})}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^{\top})}\Big)^{r-2}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}\\
&\quad\le D_1\nu_k\frac{\sigma^2m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^{\top})}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}\sum_{r\ge2}(r-1)\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^{\top})}\Big)^{r-1}
\end{aligned}$$

where we used the fact $\oint_{\gamma_k}\big(\widetilde{\mathbf{R}}_{\mathbf{A}\mathbf{A}^{\top}}(\eta)\widetilde{\mathbf{\Gamma}}\big)^r\mathbf{P}_k^{uu}d\eta=\mathbf{0}$. By the inequality $\sum_{r\ge1}rq^r=\frac{q}{(1-q)^2},\forall q<1$ and the fact $D_1\delta_t\le\bar{g}_k(\mathbf{A}\mathbf{A}^{\top})$ for some large constant $D_1>0$ and $t\le m_1$, we conclude with

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

implying that

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

The same bound holds for

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

following the same arguments. As a result,

$$\begin{aligned}
&\Big\|\mathbb{E}\Big((\mathbf{P}_k^{uu})^{\perp}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})\mathbf{P}_k^{uu}+\mathbf{P}_k^{uu}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})(\mathbf{P}_k^{uu})^{\perp}+(\mathbf{P}_k^{uu})^{\perp}\mathbf{S}_k(\widetilde{\mathbf{\Gamma}})(\mathbf{P}_k^{uu})^{\perp}\Big)\mathbf{1}\big(\|\widetilde{\mathbf{\Gamma}}\|\le\delta_t\big)\Big\|\\
&\qquad\lesssim\nu_k\frac{\sigma^2m_2^{1/2}+\sigma\mu_1}{\bar{g}_k(\mathbf{A}\mathbf{A}^{\top})}\Big(\frac{2\delta_t}{\bar{g}_k(\mathbf{A}\mathbf{A}^{\top})}\Big).
\end{aligned}\tag{5.9}$$

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

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

which is clearly dominated by (5.9). Substitute the above bounds into (5.3) and we get

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

□

5.2 Proof of Theorem 3.2

The proof of Theorem 3.2 is identical to the proof of Corollary 1.5 in [63] and will be skipped here.

5.3 Proof of Theorem 3.3

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

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

Some preliminary facts shall be concluded from Theorem 3.1. By Theorem 3.2, there exists an event $\mathcal{E}_2$ with $\mathbb{P}\big(\mathcal{E}_2\big) \geq 1 - \frac{1}{d^2}$ on which

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

and

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

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

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

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 $\mathcal{E}_2$,

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

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 $\mathcal{E}_2$,

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

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 $\mathcal{E}_2$,

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

and

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

$$\lesssim \overline{\Lambda}(\mathbf{A})\bigg(\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2 dr}{\bar{g}^2_{\min}(\mathbf{A})}\log^{1/2} d\bigg)^3.$$

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

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

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 (3.2) (3.3) (3.4) 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 $\mathcal{E}_2$,

$$\begin{aligned}
&\Big|\big\langle \mathbf{A}\cdot\big(\mathbf{P}_{\widetilde{\mathbf{U}}},\mathbf{P}_{\widetilde{\mathbf{V}}},\mathbf{P}_{\widetilde{\mathbf{W}}}\big)-\mathbf{A},\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\big\rangle\Big| \\
&\qquad\lesssim \sigma r^3\bigg(\frac{\sigma\widetilde{\kappa}(\mathbf{A})}{\bar{g}_{\min}(\mathbf{A})}+\frac{\widetilde{\kappa}^2(\mathbf{A})}{d}\bigg)(\mu_{\mathbf{U}}\mu_{\mathbf{V}}+\mu_{\mathbf{U}}\mu_{\mathbf{W}}+\mu_{\mathbf{V}}\mu_{\mathbf{W}})\log^{3/2} d
\end{aligned}\tag{5.10}$$

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 $1-\frac{1}{d^2}$,

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

where we used Lemma 3 for the upper bound of $\big\|\mathcal{M}_1(\mathbf{Z})(\mathbf{V}\otimes\mathbf{W})\big\|$. Moreover, since $\mu_{\mathbf{U}}\ge 1$,

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

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

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

Similar bounds can be attained for

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

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

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

and

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

Observe by conditions (3.2) (3.3) (3.4) that

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

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

$$\left|\left\langle \mathbf{Z}\cdot\left(\mathbf{P}_{\widetilde{\mathbf{U}}},\mathbf{P}_{\widetilde{\mathbf{V}}},\mathbf{P}_{\widetilde{\mathbf{W}}}\right),\mathbf{e}_i\otimes\mathbf{e}_j\otimes\mathbf{e}_k\right\rangle\right| \lesssim \frac{\sigma r^2}{d^{1/2}}\left(\frac{\sigma\overline{\Lambda}(\mathbf{A})r+\sigma^2 dr}{\bar{g}^2_{\min}(\mathbf{A})}\right)\mu_{\mathbf{U}}\mu_{\mathbf{V}}\mu_{\mathbf{W}}\log^{3/2} d. \quad (5.11)$$

By combining (5.10) and (5.11), we get on event $\mathcal{E}_2 \cap \mathcal{E}_3$,

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

$$\lesssim \sigma r^3 \left(\frac{\sigma \widetilde{\kappa}(\mathbf{A})}{\bar{g}_{\min}(\mathbf{A})} + \frac{\widetilde{\kappa}^2(\mathbf{A})}{d} \right) (\mu_{\mathbf{U}}\mu_{\mathbf{V}} + \mu_{\mathbf{U}}\mu_{\mathbf{W}} + \mu_{\mathbf{V}}\mu_{\mathbf{W}}) \log^{3/2} d,$$

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

CHAPTER 5

APPENDICES

Appendices

1 Proof of Lemma 1

The proof of Lemma 1 follows from a similar approach introduced by [22].

Proof. For any $S \in \mathbb{H}_m$ of $\operatorname{rank} r$, $S = \sum_{j=1}^{r} \lambda_i (e_j \otimes e_j)$, where λ_j are non-zero eigenvalues of S (repeated with their multiplicities) and $e_j \in \mathbb{C}^m$ are the corresponding orthonormal eigenvectors. Denote $\operatorname{sign}(S) := \sum_{j=1}^{r} \operatorname{sign}(\lambda_i)(e_j \otimes e_j)$. Let $\mathcal{P}_L$, $\mathcal{P}_L^{\perp}$ be the following orthogonal projectors in the space $(\mathbb{H}_m, \langle \cdot, \cdot \rangle)$:

$$\mathcal{P}_L(A) := A - P_{L^{\perp}} A P_{L^{\perp}},\ \mathcal{P}_L^{\perp}(A) := P_{L^{\perp}} A P_{L^{\perp}},\ \forall A \in \mathbb{H}_m$$

where P_L denotes the orthogonal projector on the linear span of $\{e_1, ..., e_r\}$, and $P_{L^{\perp}}$ is its orthogonal complement. Clearly, this formulation provides a decomposition of a matrix A into a "low rank part" $\mathcal{P}_L(A)$ and a "high rank part" $\mathcal{P}_L^{\perp}(A)$ if $\operatorname{rank}(S) = r$ is small. Given $b > 0$, define the following cone in the space $\mathbb{H}_m$:

$$\mathcal{K}(\mathbb{D}; L; b) := \{A \in \mathbb{D} : \|\mathcal{P}_L^{\perp} A\|_1 \leq b \|\mathcal{P}_L(A)\|_1\}$$

which consists of matrices with a "dominant" low rank part if S is low rank.

Firstly, we can rewrite (1.1) as

$$\widehat{S}^h = \arg\min_{S \in \mathbb{D}} \frac{1}{n} \sum_{j=1}^{n} \left(\tilde{Y}_j - \left\langle S, \tilde{X}_j \right\rangle \right)^2 + \varepsilon \|S\|_1. \tag{1.1}$$

where $\tilde{X}_j = \operatorname{Diag}\Big[\sqrt{\frac{1}{h} K\Big(\frac{\tau_j - t_0}{h}\Big)} p_0\Big(\frac{\tau_j - t_0}{h}\Big) X_j, \sqrt{\frac{1}{h} K\Big(\frac{\tau_j - t_0}{h}\Big)} p_1\Big(\frac{\tau_j - t_0}{h}\Big) X_j, ..., \sqrt{\frac{1}{h} K\Big(\frac{\tau_j - t_0}{h}\Big)} p_\ell\Big(\frac{\tau_j - t_0}{h}\Big) X$ and $\tilde{Y}_j = \sqrt{\frac{1}{h} K\Big(\frac{\tau_j - t_0}{h}\Big)} Y_j$.

Denote the loss function as

$$\mathcal{L}\big(\tilde{Y}; \langle S(\tau), \tilde{X} \rangle\big) := \left(\tilde{Y}_j - \left\langle S, \tilde{X}_j \right\rangle \right)^2,$$

and the risk

$$P\mathcal{L}\big(\tilde{Y};\langle S(\tau),\tilde{X}\rangle\big) := \mathbb{E}\mathcal{L}\big(\tilde{Y};\langle S(\tau),\tilde{X}\rangle\big) = \sigma^2 + \mathbb{E}\frac{1}{h}K\Big(\frac{\tau-t_0}{h}\Big)\big(Y-\langle S(\tau),X\rangle\big)^2$$

Since $\widehat{S}^h$ is a solution of the convex optimization problem (1.1), there exists a $\widehat{V} \in \partial\big\|\widehat{S}^h\big\|_1$, such that for $\forall S \in \mathbb{D}$ (see [85] Chap. 2)

$$\frac{2}{n}\sum_{j=1}^{n}\Big(\Big\langle \widehat{S}^h, \tilde{X}_j\Big\rangle - \tilde{Y}_j\Big)\langle \widehat{S}^h - S, \tilde{X}_j\rangle + \varepsilon\langle \widehat{V}, \widehat{S}^h - S\rangle \le 0.$$

This implies that, for all $S \in \mathbb{D}$,

$$\begin{aligned}
&\mathbb{E}\mathcal{L}'(\tilde{Y};\langle \widehat{S}, \tilde{X}\rangle)\langle \widehat{S}^h - S, \tilde{X}\rangle + \varepsilon\langle \widehat{V}, \widehat{S}^h - S\rangle \\
&\le \mathbb{E}\mathcal{L}'(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle)\langle \widehat{S}^h - S, \tilde{X}\rangle - \frac{2}{n}\sum_{j=1}^{n}(\langle \widehat{S}^h, \tilde{X}_j\rangle - \tilde{Y}_j)\langle \widehat{S}^h - S, \tilde{X}_j\rangle.
\end{aligned} \tag{1.2}$$

where $\mathcal{L}'$ denotes the partial derivative of $\mathcal{L}(y;u)$ with respect to u. One can easily check that for $\forall S \in \mathbb{D}$,

$$\mathbb{E}\mathcal{L}'(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle)\langle \widehat{S}^h - S, \tilde{X}\rangle \ge \mathbb{E}(\mathcal{L}(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle) - \mathcal{L}(\tilde{Y};\langle S, \tilde{X}\rangle)) + \|\widehat{S}^h - S\|^2_{L_2(\tilde{\Pi})}. \tag{1.3}$$

where $\tilde{\Pi}$ denotes the distribution of $\tilde{X}$. If $\mathbb{E}\mathcal{L}(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle) \le \mathbb{E}\mathcal{L}(\tilde{Y};\langle S, \tilde{X}\rangle)$ for $\forall S \in \mathbb{D}$, then the oracle inequality in Lemma 1 holds trivially. So we assume that $\mathbb{E}\mathcal{L}(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle) > \mathbb{E}\mathcal{L}(\tilde{Y};\langle S, \tilde{X}\rangle)$ for some $S \in \mathbb{D}$. Thus, inequalities (1.2) and (1.3) imply that

$$\begin{aligned}
&\mathbb{E}\mathcal{L}(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle) + \|\widehat{S}^h - S\|^2_{L_2(\tilde{\Pi})} + \varepsilon\langle \widehat{V}, \widehat{S}^h - S\rangle \\
&\le \mathbb{E}\mathcal{L}(\tilde{Y};\langle S, \tilde{X}\rangle) + \mathbb{E}\mathcal{L}'(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle)\langle \widehat{S}^h - S, \tilde{X}\rangle - \frac{2}{n}\sum_{j=1}^{n}(\langle \widehat{S}^h, \tilde{X}_j\rangle - \tilde{Y}_j)\langle \widehat{S}^h - S, \tilde{X}_j\rangle.
\end{aligned} \tag{1.4}$$

According to the well known representation of subdifferential of nuclear norm, see [86]

Sec. A.4, for any $V \in \partial\|S\|_1$, we have

$$V := \mathrm{sign}(S) + \mathcal{P}_L^{\perp}(W),\ W \in \mathbb{H}_m\ , \|W\| \leq 1.$$

By the duality between nuclear norm and operator norm

$$\langle \mathcal{P}_L^{\perp}(W), \widehat{S}^h - S\rangle = \langle \mathcal{P}_L^{\perp}(W), \widehat{S}^h\rangle = \langle W, \mathcal{P}_L^{\perp}(\widehat{S}^h)\rangle = \|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1.$$

Therefore, by the monotonicity of subdifferentials of convex function $\|\cdot\|_1$, for any $V := \mathrm{sign}(S) + \mathcal{P}_L^{\perp}(W) \in \partial\|S\|_1$, we have

$$\langle V, \widehat{S}^h - S\rangle = \langle \mathrm{sign}(S), \widehat{S}^h - S\rangle + \|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1 \leq \langle \widehat{V}, \widehat{S}^h - S\rangle, \tag{1.5}$$

we can use (1.5) to change the bound in (1.4) to get

$$\begin{aligned}
&\mathbb{E}\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X}\rangle) + \|S - \widehat{S}^h\|^2_{L_2(\tilde{\Pi})} + \varepsilon\|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1 \\
&\leq \mathbb{E}\mathcal{L}(\tilde{Y}; \langle S, \tilde{X}\rangle) + \varepsilon\langle \mathrm{sign}(S), S - \widehat{S}^h\rangle + \mathbb{E}\mathcal{L}'(\tilde{Y}; \langle \widehat{S}^h, \tilde{X}\rangle)\langle \widehat{S}^h - S, \tilde{X}\rangle \\
&- \frac{2}{n}\sum_{j=1}^{n}(\langle \widehat{S}^h, \tilde{X}_j\rangle - \tilde{Y}_j)\langle \widehat{S}^h - S, \tilde{X}_j\rangle.
\end{aligned} \tag{1.6}$$

For the simplicity of representation, we use the following notation to denote the empirical process:

$$\begin{aligned}
&(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle \widehat{S}^h, \tilde{X}\rangle))\langle \widehat{S}^h - S, \tilde{X}\rangle := \\
&\mathbb{E}\mathcal{L}'(\tilde{Y}; \langle \widehat{S}^h, \tilde{X}\rangle)\langle \widehat{S}^h - S, \tilde{X}\rangle - \frac{2}{n}\sum_{j=1}^{n}(\langle \widehat{S}^h, \tilde{X}_j\rangle - \tilde{Y}_j)\langle \widehat{S}^h - S, \tilde{X}_j\rangle.
\end{aligned} \tag{1.7}$$

The following part of the proof is to derive an upper bound on the empirical process (1.7). Before we start with the derivation, let us present several vital ingredients that will be used

in the following literature. For a given $S \in \mathbb{D}$ and for $\delta_1, \delta_2, \delta_3, \delta_4 \geq 0$, denote

$$\mathcal{A}(\delta_1, \delta_2) := \{A \in \mathbb{D} : A - S \in \mathcal{K}(\mathbb{D}; L; b), \|A - S\|_{L_2(\tilde{\Pi})} \leq \delta_1, \|\mathcal{P}_L^{\perp} A\|_1 \leq \delta_2\},$$

$$\tilde{\mathcal{A}}(\delta_1, \delta_2, \delta_3) := \{A \in \mathbb{D} : \|A - S\|_{L_2(\tilde{\Pi})} \leq \delta_1, \|\mathcal{P}_L^{\perp} A\|_1 \leq \delta_2, \|\mathcal{P}_L(A - S)\|_1 \leq \delta_3\},$$

$$\check{\mathcal{A}}(\delta_1, \delta_4) := \{A \in \mathbb{D} : \|A - S\|_{L_2(\tilde{\Pi})} \leq \delta_1, \|A - S\|_1 \leq \delta_4\},$$

and

$$\alpha_n(\delta_1, \delta_2) := \sup\{|(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle A, \tilde{X}\rangle))\langle A - S, \tilde{X}\rangle| : A \in \mathcal{A}(\delta_1, \delta_2)\},$$

$$\tilde{\alpha}_n(\delta_1, \delta_2, \delta_3) := \sup\{|(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle A, \tilde{X}\rangle))\langle A - S, \tilde{X}\rangle| : A \in \tilde{\mathcal{A}}(\delta_1, \delta_2, \delta_3)\},$$

$$\check{\alpha}_n(\delta_1, \delta_4) := \sup\{|(P - P_n)(\mathcal{L}'(\tilde{Y}; \langle A, \tilde{X}\rangle))\langle A - S, \tilde{X}\rangle| : A \in \check{\mathcal{A}}(\delta_1, \delta_4)\}.$$

Given the definitions above, Lemma 6 below shows upper bounds on the three quantities $\alpha_n(\delta_1, \delta_2)$, $\tilde{\alpha}_n(\delta_1, \delta_2, \delta_3)$, $\check{\alpha}_n(\delta_1, \delta_4)$. The proof of Lemma 6 can be found in section 2. Denote

$$\Xi := n^{-1} \sum_{j=1}^{n} \varepsilon_j \tilde{X}_j \tag{1.8}$$

where ε_j are i.i.d. Rademacher random variables.

Lemma 6. Suppose $0 < \delta_k^- < \delta_k^+$, $k = 1, 2, 3, 4$. Let $\eta > 0$ and

$$\bar{\eta} := \eta + \sum_{k=1}^{2} \log([\log_2(\frac{\delta_k^+}{\delta_k^-})] + 2) + \log 3,$$

$$\tilde{\eta} := \eta + \sum_{k=1}^{3} \log([\log_2(\frac{\delta_k^+}{\delta_k^-})] + 2) + \log 3,$$

$$\check{\eta} := \eta + \sum_{k=1, k=4} \log([\log_2(\frac{\delta_k^+}{\delta_k^-})] + 2) + \log 3.$$

Then with probability at least $1-e^{-\eta}$, for all $\delta_k \in [\delta_k^-, \delta_k^+]$, k=1,2,3

$$\alpha_n(\delta_1,\delta_2) \le \frac{C_1(\ell+1)R(T)\Phi a}{\sqrt{h}}\Big\{\mathbb{E}\|\Xi\|(\sqrt{\mathrm{rank}(S)}m\delta_1+\delta_2)+\frac{2(\ell+1)R(T)\Phi a\bar{\eta}}{n\sqrt{h}}+\delta_1\sqrt{\frac{\bar{\eta}}{n}}\Big\} \tag{1.9}$$

$$\tilde{\alpha}_n(\delta_1,\delta_2,\delta_3) \le \frac{C_2(\ell+1)R(T)\Phi a}{\sqrt{h}}\Big\{\mathbb{E}\|\Xi\|(\delta_2+\delta_3)+\frac{2(\ell+1)R(T)\Phi a\tilde{\eta}}{n\sqrt{h}}+\delta_1\sqrt{\frac{\tilde{\eta}}{n}}\Big\} \tag{1.10}$$

$$\check{\alpha}_n(\delta_1,\delta_4) \le \frac{C_3(\ell+1)R(T)\Phi a}{\sqrt{h}}\Big\{\mathbb{E}\|\Xi\|\delta_4+\frac{2(\ell+1)R(T)\Phi a\check{\eta}}{n\sqrt{h}}+\delta_1\sqrt{\frac{\check{\eta}}{n}}\Big\} \tag{1.11}$$

where C_1, C_2, and C_3 are numerical constants.

Since both $\widehat{S}^h$ and S are in $\mathbb{D}$, by the definition of $\tilde{\alpha}$ and $\check{\alpha}$, we have

$$(P-P_n)(\mathcal{L}'(\tilde{Y};\langle\widehat{S}^h,\tilde{X}\rangle))\langle\widehat{S}^h-S,\tilde{X}\rangle \le \tilde{\alpha}(\|\widehat{S}^h-S\|_{L_2(\tilde{\Pi})};\|\mathcal{P}_L^{\perp}\widehat{S}^h\|_1;\|\mathcal{P}_L(\widehat{S}^h-S)\|_1), \tag{1.12}$$

and

$$(P-P_n)(\mathcal{L}'(\tilde{Y};\langle\widehat{S}^h,\tilde{X}\rangle))\langle\widehat{S}^h-S,\tilde{X}\rangle \le \check{\alpha}(\|\widehat{S}^h-S\|_{L_2(\tilde{\Pi})};\|\widehat{S}^h-S\|_1), \tag{1.13}$$

If $\widehat{S}^h - S \in \mathcal{K}(\mathbb{D};L;b)$, by the definition of α, we have

$$(P-P_n)(\mathcal{L}'(\tilde{Y};\langle\widehat{S}^h,\tilde{X}\rangle))\langle\widehat{S}^h-S,\tilde{X}\rangle \le \alpha(\|\widehat{S}^h-S\|_{L_2(\tilde{\Pi})};\|\mathcal{P}_L^{\perp}\widehat{S}^h\|_1), \tag{1.14}$$

Assume for a while that

$$\|\widehat{S}^h-S\|_{L_2(\tilde{\Pi})} \in [\delta_1^-,\delta_1^+],\ \|\mathcal{P}_L^{\perp}\widehat{S}^h\|_1 \in [\delta_2^-,\delta_2^+],\ \|\mathcal{P}_L^{\perp}(\widehat{S}^h-S)\|_1 \in [\delta_3^-,\delta_3^+]. \tag{1.15}$$

By the definition of subdifferential, for any $\widehat{V} \in \partial\|\widehat{S}^h\|_1$,

$$\langle \widehat{V}, S - \widehat{S}^h \rangle \leq \|S\|_1 - \|\widehat{S}^h\|_1.$$

Then we apply (1.13) in bound (1.4) and use the upper bound on $\check{\alpha}_n(\delta_1, \delta_4)$ of Lemma 6, and get with probability at least $1 - e^{-\eta}$,

$$\begin{aligned}
&P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)) + \|\widehat{S}^h - S\|^2_{L_2(\tilde{\Pi})} \\
&\leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \varepsilon(\|S\|_1 - \|\widehat{S}^h\|_1) + \check{\alpha}_n(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}, \|\widehat{S}^h - S\|_1) \\
&\leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \varepsilon(\|S\|_1 - \|\widehat{S}^h\|_1) \\
&+ \frac{C_3(\ell+1)R(T)\Phi a}{\sqrt{h}} \Big\{ \mathbb{E}\|\Xi\| \|\widehat{S}^h - S\|_1 + \frac{2(\ell+1)R(T)\Phi a \check{\eta}}{n\sqrt{h}} + \|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})} \sqrt{\frac{\check{\eta}}{n}} \Big\}.
\end{aligned} \tag{1.16}$$

Assuming that

$$\varepsilon > \frac{C(\ell+1)R(T)\Phi a}{\sqrt{h}} \mathbb{E}\|\Xi\|, \tag{1.17}$$

where $C = C_1 \vee 4C_2 \vee C_3$. From (1.16)

$$P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)) \leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + 2\varepsilon\|S\|_1 + \frac{C_3(\ell+1)^2 R(T)^2 \Phi^2 a^2 \tilde{\eta}}{nh}. \tag{1.18}$$

We now apply the upper bound on $\tilde{\alpha}_n(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}, \|\mathcal{P}_L^{\perp}\widehat{S}^h)\|_1, \|\mathcal{P}_L(\widehat{S}^h - S)\|_1)$ to (1.6) and get with probability at least $1 - e^{-\eta}$,

$$\begin{aligned}
&P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X} \rangle)) + \|\widehat{S}^h - S\|^2_{L_2(\tilde{\Pi})} + \varepsilon\|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1 \\
&\leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \varepsilon\|\mathcal{P}_L(\widehat{S}^h - S)\|_1 + \tilde{\alpha}_n(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})}, \|\mathcal{P}_L^{\perp}\widehat{S}^h)\|_1, \|\mathcal{P}_L(\widehat{S}^h - S)\|_1) \\
&\leq P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X} \rangle)) + \varepsilon\|\mathcal{P}_L(\widehat{S}^h - S)\|_1 \\
&+ \frac{C_2(\ell+1)R(T)\Phi a}{\sqrt{h}} \Big\{ \mathbb{E}\|\Xi\| (\|\mathcal{P}_L^{\perp}\widehat{S}^h)\|_1 + \|\mathcal{P}_L(\widehat{S}^h - S)\|_1) \Big\} + \frac{C_2(\ell+1)^2 R(T)^2 \Phi^2 a^2 \tilde{\eta}}{nh},
\end{aligned} \tag{1.19}$$

where the first inequality is due to the fact that

$$|\langle \mathrm{sign}(S), S-\widehat{S}^h\rangle| = |\langle \mathrm{sign}(S), \mathcal{P}_L(S-\widehat{S}^h)\rangle| \leq \|\mathrm{sign}(S)\|\|\mathcal{P}_L(S-\widehat{S}^h)\|_1 \leq \|\mathcal{P}_L(S-\widehat{S}^h)\|_1.$$

With assumption (1.17) holds, we get from (1.19)

$$\begin{aligned} &P\mathcal{L}(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle) + \varepsilon\|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1 \\ &\leq P\mathcal{L}(\tilde{Y};\langle S, \tilde{X}\rangle) + \frac{5\varepsilon}{4}\|\mathcal{P}_L(\widehat{S}^h - S)\|_1 + \frac{\varepsilon}{4}\|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1 + \frac{C_2(\ell+1)^2R(T)^2\Phi^2a^2\tilde{\eta}}{nh}. \end{aligned} \tag{1.20}$$

If the following is satisfied:

$$\frac{C_2(\ell+1)^2R(T)^2\Phi^2a^2\tilde{\eta}}{nh} \geq \frac{5\varepsilon}{4}\|\mathcal{P}_L(\widehat{S}^h - S)\|_1 + \frac{\varepsilon}{4}\|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1, \tag{1.21}$$

we can just conclude that

$$P(\mathcal{L}(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle)) \leq P(\mathcal{L}(\tilde{Y};\langle S, \tilde{X}\rangle)) + \frac{C_2(\ell+1)^2R(T)^2\Phi^2a^2\tilde{\eta}}{nh}, \tag{1.22}$$

which is sufficient to meet the bound of Lemma 1. Otherwise, by the assumption that $P(\mathcal{L}(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle)) > P(\mathcal{L}(\tilde{Y};\langle S, \tilde{X}\rangle))$, one can easily check that

$$\|\mathcal{P}_L^{\perp}(\widehat{S}^h - S)\|_1 \leq 5\|\mathcal{P}_L(\widehat{S}^h - S)\|_1,$$

which implies that $\widehat{S}^h - S \in \mathcal{K}(\mathbb{D}; L; 5)$. This fact allows us to use the bound on $\alpha_n(\delta_1, \delta_2)$ of Lemma 6. We get from (1.6)

$$\begin{aligned} &P(\mathcal{L}(\tilde{Y};\langle \widehat{S}^h, \tilde{X}\rangle)) + \|\widehat{S}^h - S\|^2_{L_2(\tilde{\Pi})} + \varepsilon\|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1 \\ &\leq P(\mathcal{L}(\tilde{Y};\langle S, \tilde{X}\rangle)) + \varepsilon\langle \mathrm{sign}(S), S - \widehat{S}^h\rangle \\ &+ \frac{C_1(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|(\sqrt{\mathrm{rank}(S)}m\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})} + \|\mathcal{P}_L^{\perp}(\widehat{S}^h)\|_1) + \frac{C_1(\ell+1)^2R(T)^2\Phi^2a^2\bar{\eta}}{nh}. \end{aligned} \tag{1.23}$$

By applying the inequality

$$\left|\langle \mathrm{sign}(S), \widehat{S}^h - S\rangle\right| \le m\sqrt{\mathrm{rank}(S)}\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})},$$

and the assumption (1.17), we have with probability at least $1 - e^{-\eta}$,

$$P(\mathcal{L}(\tilde{Y}; \langle \widehat{S}^h, \tilde{X}\rangle)) \le P(\mathcal{L}(\tilde{Y}; \langle S, \tilde{X}\rangle)) + \varepsilon^2 m^2 \mathrm{rank}(S) + \frac{C_1(\ell+1)^2 R(T)^2 \Phi^2 a^2 \bar{\eta}}{nh}. \tag{1.24}$$

To sum up, the bound of Lemma 1 follows from (1.18), (1.22) and (1.24) provided that condition (1.17) and condition (1.15) hold.

We still need to specify δ_k^-, δ_k^+, $k = 1, 2, 3, 4$ to establish the bound of the theorem. By the definition of $\widehat{S}^h$, we have

$$P_n(\mathcal{L}(\tilde{Y}; \langle X, \widehat{S}^h\rangle)) + \varepsilon\|\widehat{S}^h\|_1 \le P_n(\mathcal{L}(\tilde{Y}; \langle X, 0\rangle)) \le Q,$$

implying that $\|\widehat{S}^h\|_1 \le \frac{Q}{\varepsilon}$. Next, $\|\mathcal{P}_L^\perp \widehat{S}^h\|_1 \le \|\widehat{S}^h\|_1 \le \frac{Q}{\varepsilon}$ and $\|\mathcal{P}_L(\widehat{S}^h - S)\|_1 \le 2\|\widehat{S}^h - S\|_1 \le \frac{2Q}{\varepsilon} + 2\|S\|_1$. Finally, we have $\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})} \le 2a$. Thus, we can take $\delta_1^+ := 2a$, $\delta_2^+ := \frac{Q}{\varepsilon}$, $\delta_3^+ := \frac{2Q}{\varepsilon} + 2\|S\|_1$, $\delta_4^+ := \frac{Q}{\varepsilon} + \|S\|_1$. With these choices, $\delta_k^+,\ k = 1, 2, 3, 4$ are upper bounds on the corresponding norms in condition (1.15). We choose $\delta_1^- := \frac{a}{\sqrt{n}}$, $\delta_2^- := \frac{a^2}{n\varepsilon} \wedge \frac{\delta_2^+}{2}$, $\delta_3^- := \frac{a^2}{n\varepsilon} \wedge \frac{\delta_3^+}{2}$, $\delta_4^- := \frac{a^2}{n\varepsilon} \wedge \frac{\delta_4^+}{2}$. Let $\eta^* := \eta + 3\log(B\log_2(\|S\|_1 \vee n \vee \varepsilon \vee a^{-1} \vee Q))$. It is easy to verify that $\bar{\eta} \vee \tilde{\eta} \vee \tilde{\eta} \le \eta^*$. for a proper choice of numerical constant B in the definition of η^*. When condition (1.15) does not hold, which means at least one of the numbers $\delta_k^-,\ k = 1, 2, 3, 4$ we chose is not a lower bound on the corresponding norm, we can still use the bounds

$$\begin{aligned}
&(P - P_n)(\mathcal{L}'(Y; \langle \widehat{S}^h, \tilde{X}\rangle))\langle \widehat{S}^h - S, \tilde{X}\rangle \\
&\le \tilde{\alpha}(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})} \vee \delta_1^-; \|\mathcal{P}_L^\perp \widehat{S}^h\|_1 \vee \delta_2^-; \|\mathcal{P}_L(\widehat{S}^h - S)\|_1 \vee \delta_3^-),
\end{aligned} \tag{1.25}$$

and

$$(P - P_n)(\mathcal{L}'(Y; \langle \widehat{S}^h, \tilde{X} \rangle))\langle \widehat{S}^h - S, \tilde{X} \rangle \leq \breve{\alpha}(\|\widehat{A}(t)^{\varepsilon} - S\|_{L_2(\tilde{\Pi})} \vee \delta_1^-; \|\widehat{S}^h - S\|_1 \vee \delta_4^-), \tag{1.26}$$

instead of (1.12), (1.13). In the case when $\widehat{S}^h - S \in \mathcal{K}(\mathbb{D}; L; 5)$, we can use the bound

$$(P - P_n)(\mathcal{L}'(Y; \langle \widehat{S}^h, \tilde{X} \rangle))\langle \widehat{S}^h - S, \tilde{X} \rangle \leq \alpha(\|\widehat{S}^h - S\|_{L_2(\tilde{\Pi})} \vee \delta_1^-; \|\mathcal{P}_L^{\perp} \widehat{S}^h\|_1 \vee \delta_2^-), \tag{1.27}$$

instead of bound (1.14). Then one can repeat the arguments above with only minor modifications. By the adjusting the constants, the result of Lemma 1 holds.

The last thing we need to specify is the size of ε which controls the nuclear norm penalty. Recall that from condition (1.17), the essence is to control $\mathbb{E}\|\Xi\|$. Here we use a simple but powerful noncommutative matrix Bernstein inequalities. The original approach was introduced by [87]. Later, the result was improved by [23] based on the classical result of [88]. We give the following lemma which is a direct consequence of the result proved by [23], and we omit the proof here.

Lemma 7. Under the model (1.1), Ξ is defined as in (1.8) with τ_j are i.i.d. uniformly distributed in [0,1], and ε_j are i.i.d. Rademacher random variables, and X_j are i.i.d uniformly distributed in $\mathcal{X}$. Then for any $\eta > 0$, with probability at least $1 - e^{-\eta}$

$$\|\Xi\| \leq 4\Big(\sqrt{\frac{(\eta + \log 2m)}{nm}} \bigvee \frac{(\eta + \log 2m)\Phi}{n\sqrt{h}}\Big),$$

and by integrating its exponential tail bounds

$$\mathbb{E}\|\Xi\| \leq C\Big(\sqrt{\frac{\log 2m}{nm}} \bigvee \frac{(\log 2m)\Phi}{n\sqrt{h}}\Big)$$

where C is a numerical constant.

Together with (1.17), we know for some numerical constant $D > 0$,

$$\varepsilon \geq D\frac{\Phi a(\ell+1)R(T)}{\sqrt{h}}\Big(\sqrt{\frac{\log 2m}{nm}} \bigvee \frac{(\log 2m)\Phi}{n\sqrt{h}}\Big).$$

which completes the proof of Lemma 1. □

2 Proof of Lemma 6

Proof. We only prove the first bound in detail, and proofs of the rest two bounds are similar with only minor modifications. By Talagrand's concentration inequality [89], and its Bousquet's form [90], with probability at least $1 - e^{-\eta}$,

$$\alpha_n(\delta_1, \delta_2) \leq 2\mathbb{E}\alpha_n(\delta_1, \delta_2) + \frac{24(\ell+1)^2R(T)^2\Phi^2a^2\eta}{nh} + \frac{12(\ell+1)R(T)\Phi a\delta_1\sqrt{\eta}}{\sqrt{nh}}. \quad (2.1)$$

By standard Rademacher symmetrization inequalities, see [86], Sec. 2.1, we can get

$$\mathbb{E}\alpha_n(\delta_1, \delta_2) \leq 4\mathbb{E}\sup\Big\{\Big|\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j(\langle A, \tilde{X}_j\rangle - \tilde{Y}_j)\langle A - S, \tilde{X}_j\rangle\Big| : A \in \mathcal{A}(\delta_1, \delta_2)\Big\}, \quad (2.2)$$

where $\{\varepsilon_j\}$ are i.i.d. Rademacher random variables independent of $\{(\tau_j, X_j, \tilde{Y}_j)\}$. Then we consider the function $f(u) = (u - y + v)u$, where $|y| \leq \frac{2\Phi a}{\sqrt{h}}$ and $|v|,\ |u| \leq \frac{2(\ell+1)R(T)\Phi a}{\sqrt{h}}$. Clearly, this function has a Lipschitz constant $\frac{6(\ell+1)R(T)\Phi a}{\sqrt{h}}$. Thus by comparison inequality, see [86], Sec. 2.2, we can get

$$\begin{aligned}
&\mathbb{E}\sup\Big\{\Big|n^{-1}\sum_{j=1}^{n}\varepsilon_j(\langle A, \tilde{X}_j\rangle - \tilde{Y}_j)\langle A - S, \tilde{X}_j\rangle\Big| : A \in \mathcal{A}(\delta_1, \delta_2)\Big\} \\
&\leq \frac{6(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\sup\Big\{n^{-1}\Big|\sum_{j=1}^{n}\varepsilon_i\langle A - S, \tilde{X}_j\rangle\Big| : A \in \mathcal{A}(\delta_1, \delta_2)\Big\}.
\end{aligned} \quad (2.3)$$

As a consequence of (2.2) and (2.3), we have

$$\mathbb{E}\alpha_n(\delta_1,\delta_2) \le \frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\sup\Big\{n^{-1}\Big|\sum_{j=1}^{n}\varepsilon_i\langle A-S,\tilde{X}_j\rangle\Big| : A\in\mathcal{A}(\delta_1,\delta_2)\Big\}. \quad (2.4)$$

The next step is to get an upper bound on $\big|n^{-1}\sum_{j=1}^{n}\varepsilon_i\langle A-S,\tilde{X}_j\rangle\big|$. Recall that $\Xi := n^{-1}\sum_{j=1}^{n}\varepsilon_j\tilde{X}_j$, then we have $n^{-1}\sum_{j=1}^{n}\varepsilon_i\langle A-S,\tilde{X}_j\rangle = \langle A-S,\Xi\rangle$. One can check that

$$\begin{aligned}
|\langle A-S,\Xi\rangle| &\le |\langle\mathcal{P}_L(A-S),\mathcal{P}_L\Xi\rangle| + |\langle\mathcal{P}_L^{\perp}(A-S),\Xi\rangle| \\
&\le \|\mathcal{P}_L\Xi\|_2\|\mathcal{P}_L(A-S)\|_2 + \|\Xi\|\|\mathcal{P}_L^{\perp}A\|_1 \\
&\le m\sqrt{2\mathrm{rank}(S)}\|\Xi\|\|A-S\|_{L_2(\tilde{\Pi})} + \|\Xi\|\|\mathcal{P}_L^{\perp}A\|_1.
\end{aligned}$$

The second line of this inequality is due to Hölder's inequality and the third line is due to the facts that $(A-S)\in\mathcal{K}(\mathbb{D};L;5)$, $\mathrm{rank}(\mathcal{P}_L(\Xi))\le 2\mathrm{rank}(S)$, $\|\mathcal{P}_L\Xi\|_2 \le 2\sqrt{\mathrm{rank}(\mathcal{P}_L(\Xi))}\|\Xi\|$, and $\big\|A-S\big\|^2_{L_2(\tilde{\Pi})} = \frac{1}{m^2}\big\|A-S\big\|_2$. Therefore,

$$\begin{aligned}
&\frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\sup\Big\{\Big|n^{-1}\sum_{j=1}^{n}\varepsilon_i\langle A-S,\tilde{X}_j\rangle\Big| : A\in\mathcal{A}(\delta_1,\delta_2)\Big\} \\
&\le \frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|(2\sqrt{2\mathrm{rank}(S)}m\delta_1+\delta_2).
\end{aligned} \quad (2.5)$$

It follows from (2.1), (2.4) and (2.5) that with probability at least $1-e^{-\eta}$,

$$\begin{aligned}
\alpha_n(\delta_1,\delta_2) \le &\Big(\frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|(\sqrt{\mathrm{rank}(S)}m\delta_1+\delta_2)\Big) \\
&+\frac{24(\ell+1)^2R(T)^2\Phi^2a^2\eta}{nh} + \frac{12(\ell+1)R(T)\Phi a\delta_1\sqrt{\eta}}{\sqrt{nh}}.
\end{aligned}$$

Now similar to the approach in [22], we make this bound uniform in $\delta_k\in[\delta_k^-,\delta_k^+]$. Let $\delta_k^{j_k} = \delta_k^+2^{-j_k}$, $j_k = 0,...,[\log_2(\delta_k^+/\delta_k^-)]+1$, $k=1,2$. By the union bound, with probability

at least $1-e^{-\eta}/3$, for all $j_k = 0, ..., [\log_2(\delta_k^+/\delta_k^-)]+1,\ k=1,2$, we have

$$\alpha_n(\delta_1,\delta_2) \le \Big(\frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|(\sqrt{\mathrm{rank}(S)}m\delta_1^{j_1}+\delta_2^{j_2})\Big) + \frac{24(\ell+1)^2R(T)^2\Phi^2a^2\eta}{nh} + \frac{12(\ell+1)R(T)\Phi a\delta_1^{j_1}\sqrt{\eta}}{\sqrt{nh}}.$$

which implies that for all $\delta_k \in [\delta_k^-, \delta_k^+],\ k=1,2$,

$$\alpha_n(\delta_1,\delta_2) \le \Big(\frac{12(\ell+1)R(T)\Phi a}{\sqrt{h}}\mathbb{E}\|\Xi\|(\sqrt{\mathrm{rank}(S)}m\delta_1+\delta_2)\Big) + \frac{24(\ell+1)^2R(T)^2\Phi^2a^2\bar{\eta}}{nh} + \frac{12(\ell+1)R(T)\Phi a\delta_1\sqrt{\bar{\eta}}}{\sqrt{nh}}.$$

The proofs of the second and the third bounds are similar to this one, we omit the repeated arguments. □

3 Proof of Lemma 3

Let $\mathbf{z}_i \in \mathbb{R}^{m_1}, i=1,\dots,m_2$ 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 $\tilde{\mathbf{z}}_j \in \mathbb{R}^{m_1}, j=1,\dots,m_1$ 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

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

The inequalities (5.7) and (5.2) are on the concentration of sample covariance operator, where a sharp bound has been derived in [91] and will be skipped here.

4 Proof of Theorem 5.1

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

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

Lemma 8. 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 $1 - e^{-t}$,

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

Proof. Recall that

$$\widehat{\mathbf{\Gamma}} = \mathbf{A}\mathbf{Z}^\top + \mathbf{Z}\mathbf{A}^\top + \mathbf{Z}\mathbf{Z}^\top - m_2\sigma^2\mathbf{I}_{m_1}.$$

Then, we write $\big\langle \mathbf{x}, \mathbf{L}_k(\widehat{\mathbf{\Gamma}})\mathbf{y}\big\rangle$ as

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

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,

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

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

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

Similarly, for all $t \geq 0$,

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

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 $1 \leq i \leq m_2$. 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 $t \geq 0$,

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

By [92, 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 $t \lesssim m_1 \leq m_2$, we conclude that there exists an absolute constant $D_1 > 0$ such that

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

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

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

which concludes the proof. □

It remains to derive the upper bound of $\big| \langle \mathbf{x}, \mathbf{S}_k(\widehat{\mathbf{\Gamma}})\mathbf{y} \rangle - \mathbb{E}\langle \mathbf{x}, \mathbf{S}_k(\widehat{\mathbf{\Gamma}})\mathbf{y} \rangle \big|$.The following lemma is due to [69].

Lemma 9. Let $\delta(m_1, m_2) := \sigma\mu_1 m_1^{1/2} + \sigma^2 (m_1 m_2)^{1/2}$ and suppose that $\delta(m_1, m_2) \leq \frac{1-\gamma}{2(1+\gamma)} \bar{g}_k(\mathbf{A}\mathbf{A}^\top)$ for some $\gamma \in (0, 1)$. There exists a constant $D_\gamma > 0$ such that, for all symmetric $\widehat{\mathbf{\Gamma}}_1, \widehat{\mathbf{\Gamma}}_2 \in \mathbb{R}^{m_1 \times m_1}$ satisfying the condition $\max\big\{ \|\widehat{\mathbf{\Gamma}}_1\|, \|\widehat{\mathbf{\Gamma}}_2\| \big\} \leq (1+\gamma)\delta(m_1, m_2)$,

$$\|\mathbf{S}_k(\widehat{\mathbf{\Gamma}}_1) - \mathbf{S}_k(\widehat{\mathbf{\Gamma}}_2)\| \leq D_\gamma \frac{\delta(m_1, m_2)}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)} \|\widehat{\mathbf{\Gamma}}_1 - \widehat{\mathbf{\Gamma}}_2\|.$$

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

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

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

Lemma 10. For any $\delta_1 \leq \frac{1-\gamma}{2(1+\gamma)} \bar{g}_k(\mathbf{A}\mathbf{A}^\top)$ 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 of Lemma 10. Since $\varphi(\frac{\|\widehat{\mathbf{\Gamma}}(\mathbf{Z})\|}{\delta_1})\varphi(\frac{\|\mathbf{Z}\|}{\delta_2}) \neq 0$ only if $\|\widehat{\mathbf{\Gamma}}(\mathbf{Z})\| \le (1+\gamma)\delta_1$ and $\|\mathbf{Z}\| \le (1+\gamma)\delta_2$, Lemma 4 implies that

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

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

By the Lipschitzity of function φ, Lemma 9 and definition of $\widehat{\mathbf{\Gamma}}(\mathbf{Z})$, it is easy to check

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

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

$$\begin{aligned}
&\big|F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_1) - F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_2)\big| \\
&\quad = \Big|\Big\langle \mathbf{x}, \mathbf{S}_k\big(\widehat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\mathbf{y}\Big\rangle \varphi\Big(\frac{\|\widehat{\mathbf{\Gamma}}(\mathbf{Z}_1)\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}_1\|}{\delta_2}\Big)\Big| \\
&\quad = \Big|\Big\langle \mathbf{x}, \mathbf{S}_k\big(\widehat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\mathbf{y}\Big\rangle \varphi\Big(\frac{\|\widehat{\mathbf{\Gamma}}(\mathbf{Z}_1)\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}_1\|}{\delta_2}\Big) - \Big\langle \mathbf{x}, \mathbf{S}_k\big(\widehat{\mathbf{\Gamma}}(\mathbf{Z}_1)\big)\mathbf{y}\Big\rangle \varphi\Big(\frac{\|\widehat{\mathbf{\Gamma}}(\mathbf{Z}_2)\|}{\delta_1}\Big)\varphi\Big(\frac{\|\mathbf{Z}_1\|}{\delta_2}\Big)\Big|
\end{aligned}$$

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

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

Case 4. If $\|\widehat{\mathbf{\Gamma}}(\mathbf{Z}_1)\| \leq (1+\gamma)\delta_1, \quad \|\widehat{\mathbf{\Gamma}}(\mathbf{Z}_2)\| \geq (1+\gamma)\delta_1, \quad \|\mathbf{Z}_1\| \geq (1+\gamma)\delta_2, \quad \|\mathbf{Z}_2\| \geq (1+\gamma)\delta_2$. 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 $\max\big\{\|\widehat{\mathbf{\Gamma}}(\mathbf{Z}_1)\|, \|\widehat{\mathbf{\Gamma}}(\mathbf{Z}_2)\|\big\} \leq (1+\gamma)\delta_1, \quad \|\mathbf{Z}_1\| \leq (1+\gamma)\delta_2, \quad \|\mathbf{Z}_2\| \geq (1+\gamma)\delta_2$. Again, we have $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}_2) = 0$. Then,

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

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 10 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 11. Let $\delta(m_1, m_2) := \sigma\mu_1 m_1^{1/2} + \sigma^2(m_1 m_2)^{1/2}$ and suppose that $\mathbb{E}\|\widehat{\mathbf{\Gamma}}\| \leq \frac{1-\gamma}{2}\bar{g}_k(\mathbf{A}\mathbf{A}^\top)$ 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

$\log 8 \le t \le m_1$, the following inequality holds with probability at least $1 - e^{-t}$,

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

Proof of Lemma 11. Choose $\delta_1 = \delta_1(m_1, m_2)$ and $\delta_2 = \delta_2(m_1, m_2)$ as follows where $\log 8 \le t \le m_1$ is to be determined:

$$\begin{aligned}
\delta_1(m_1, m_2) :&= \delta_1(m_1, m_2, t) := \mathbb{E}\|\widetilde{\mathbf{\Gamma}}\| + D_1 t^{1/2}(\sigma\mu_1 + \sigma^2 m_2^{1/2}) \\
\delta_2(m_1, m_2) :&= \delta_2(m_1, m_2, t) := \mathbb{E}\|\mathbf{Z}\| + D_2 \sigma t^{1/2}
\end{aligned}$$

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

Case 1. If $D_1 t^{1/2}(\mu_1\sigma + \sigma^2 m_2^{1/2}) \le \frac{\gamma}{4}\bar{g}_k(\mathbf{A}\mathbf{A}^\top)$. Then, $\delta_1 \le (1 - \frac{\gamma}{2})\frac{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}{2} = \frac{1-2\gamma'}{1+2\gamma'}\frac{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}{2}$ for some $\gamma' \in (0, 1/2)$. By Lemma 10, $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\cdot)$ satisfies the Lipschitz condition. By definition of $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z})$, we have $F_{\delta_1,\delta_2,\mathbf{x},\mathbf{y}}(\mathbf{Z}) = \langle \mathbf{x}, \mathbf{S}_k(\widehat{\mathbf{\Gamma}})\mathbf{y}\rangle$ on the event $\{\|\widehat{\mathbf{\Gamma}}\| \le \delta_1, \|\mathbf{Z}\| \le \delta_2\}$. By Lemma 3 and $t \ge \log 8$,

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

and similarly,

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

It follows from Gaussian isoperimetric inequality (see [63, Lemma 2.6]) and Lemma 10 that with some constant $D_\gamma > 0$, for all $t \geq \log 8$ with probability at least $1 - e^{-t}$,

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

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

$$\left|\langle \mathbf{x}, \mathbf{S}_k(\widehat{\mathbf{\Gamma}})\mathbf{y}\rangle - M\right| \leq D_\gamma \frac{\sigma\mu_1 t^{1/2} + \sigma^2 m_2^{1/2} t^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\left(\frac{\delta(m_1,m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\right)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}. \tag{4.1}$$

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

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

and $\delta_1 \leq D_\gamma t^{1/2}(\sigma\mu_1 + \sigma^2 m_2^{1/2})$. By Lemma 3 and Lemma 4, with probability at least $1 - e^{-t}$,

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

which immediately yields that

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

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

$$\begin{aligned}
|\langle \mathbf{x}, \mathbf{S}_k(\widehat{\mathbf{\Gamma}})\mathbf{y}\rangle - M| \leq & D_\gamma t \frac{(\sigma\mu_1 + \sigma^2 m_2^{1/2})^2}{\bar{g}_k^2(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2} \\
\leq & D_\gamma \frac{\sigma\mu_1 t^{1/2} + \sigma^2 m_2^{1/2} t^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\left(\frac{\delta(m_1,m_2)}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\right)\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}.
\end{aligned} \tag{4.2}$$

Therefore, bounds (4.1) and (4.2) hold in both cases. The rest of the proof is quite standard by integrating the exponential tails and will be skipped here, see [63]. □

Proof of Theorem 5.1. By Lemma 8 and Lemma 11, 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 $1 - 2e^{-t}$,

$$\left|\left\langle \mathbf{x}, \widehat{\mathbf{P}}_k \mathbf{y}\right\rangle\right| \le Dt^{1/2}\frac{\sigma\mu_1 + \sigma^2 m_2^{1/2}}{\bar{g}_k(\mathbf{A}\mathbf{A}^\top)}\|\mathbf{x}\|_{\ell_2}\|\mathbf{y}\|_{\ell_2}$$

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

5 Proof of Lemma 5

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

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

where the constants $D_1, D_2 > 0$ are chosen such that $\max\left\{\mathbb{P}\big(\|\widetilde{\mathbf{\Gamma}}\| \ge \delta_t\big), \mathbb{P}\big(\|\widehat{\mathbf{\Gamma}}\| \ge \delta_t\big)\right\} \le e^{-t}$. By Lemma 9,

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

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 [91],

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

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

$$\begin{aligned}\mathbb{E}\Big\|\mathbf{S}_k(\widetilde{\mathbf{\Gamma}}) - \mathbf{S}_k(\widehat{\mathbf{\Gamma}})\Big\|\mathbf{1}\Big(\|\widetilde{\mathbf{\Gamma}}\|\le(1+\gamma)\delta_t\Big)\mathbf{1}\Big(\|\widehat{\mathbf{\Gamma}}\|\le(1+\gamma)\delta_t\Big) \\ \lesssim_\gamma \frac{\delta_t}{\bar g_k(\mathbf{A}\mathbf{A}^\top)}\Big(\frac{m_1\sigma^2}{\bar g_k(\mathbf{A}\mathbf{A}^\top)}\Big).\end{aligned} \tag{5.1}$$

Choose $t = m_1$, by Lemma 4 and Lemma 3,

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

which is clearly dominated by (5.1) for $t = m_1$ and $m_2e^{-m_1/2} \le 1$. The other terms are bounded in a similar fashion. To sum up, we obtain

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

REFERENCES

[1] F. Zhou, "Nonparametric estimation of low rank matrix valued function," *ArXiv preprint arXiv:1802.06292*, 2018.

[2] D. Xia and F. Zhou, "The ℓ_∞ perturbation of hosvd and low rank tensor denoising," *ArXiv preprint arXiv:1707.01207*, 2017.

[3] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, p. 717, 2009.

[4] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[5] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[6] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, 2010.

[7] V. Koltchinskii, "Von Neumann entropy penalization and low-rank matrix estimation," *The Annals of Statistics*, pp. 2936–2973, 2011.

[8] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[9] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, vol. 39, no. 5, pp. 2302–2329, 2011.

[10] A. Rohde and A. B. Tsybakov, "Estimation of high-dimensional low-rank matrices," *The Annals of Statistics*, vol. 39, no. 2, pp. 887–930, 2011.

[11] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *Journal of Machine Learning Research*, vol. 13, no. May, pp. 1665–1697, 2012.

[12] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *The Annals of Statistics*, vol. 43, no. 1, pp. 177–214, 2015.

[13] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.

[14] Y. Koren, "Collaborative filtering with temporal dynamics," *Communications of the ACM*, vol. 53, no. 4, pp. 89–97, 2010.

[15] L. Zhang, G. Wahba, and M. Yuan, "Distance shrinkage and euclidean embedding via regularized kernel estimation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 4, pp. 849–867, 2016.

[16] A. Singer and M. Cucuringu, "Uniqueness of low-rank matrix completion by rigidity theory," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 4, pp. 1621–1641, 2010.

[17] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[18] J. Fan and I. Gijbels, *Local polynomial modelling and its applications: Monographs on statistics and applied probability 66*. CRC Press, 1996, vol. 66.

[19] O. V. Lepskii, "On a problem of adaptive estimation in gaussian white noise," *Theory of Probability & Its Applications*, vol. 35, no. 3, pp. 454–466, 1991.

[20] A. R. Barron, "Complexity regularization with application to artificial neural networks," *Nonparametric Functional Estimation and Related Topics*, vol. 335, pp. 561–576, 1991.

[21] M. Wegkamp, "Model selection in nonparametric regression," *The Annals of Statistics*, vol. 31, no. 1, pp. 252–273, 2003.

[22] V. Koltchinskii, "Sharp oracle inequalities in low rank estimation," in *Empirical Inference*, Springer, 2013, pp. 217–230.

[23] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[24] A. B. Tsybakov, *Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats*, 2009.

[25] E. N. Gilbert, "A comparison of signalling alphabets," *Bell Labs Technical Journal*, vol. 31, no. 3, pp. 504–522, 1952.

[26] R. R. Varshamov, "Estimate of the number of signals in error correcting codes," in *Dokl. Akad. Nauk SSSR*, vol. 117, 1957, pp. 739–741.

[27] O. V. Lepski and V. G. Spokoiny, "Optimal pointwise adaptive methods in nonparametric estimation," *The Annals of Statistics*, pp. 2512–2546, 1997.

[28] O. V. Lepski, E. Mammen, and V. G. Spokoiny, "Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors," *The Annals of Statistics*, pp. 929–947, 1997.

[29] V. Koltchinskii, "Local rademacher complexities and oracle inequalities in risk minimization," *The Annals of Statistics*, vol. 34, no. 6, pp. 2593–2656, 2006.

[30] C. C. Craig, "On the Tchebychef inequality of Bernstein," *The Annals of Mathematical Statistics*, vol. 4, no. 2, pp. 94–102, 1933.

[31] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[32] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *ArXiv preprint arXiv:1009.5055*, 2010.

[33] C. Chen, B. He, and X. Yuan, "Matrix completion via an alternating direction method," *IMA Journal of Numerical Analysis*, vol. 32, no. 1, pp. 227–245, 2012.

[34] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 208–220, 2013.

[35] C.-F. Westin, S. E. Maier, H. Mamata, A. Nabavi, F. A. Jolesz, and R. Kikinis, "Processing and visualization for diffusion tensor mri," *Medical image analysis*, vol. 6, no. 2, pp. 93–108, 2002.

[36] T Hildebrand and P Rüegsegger, "A new method for the model-independent assessment of thickness in three-dimensional images," *Journal of microscopy*, vol. 185, no. 1, pp. 67–75, 1997.

[37] N. Li and B. Li, "Tensor completion for on-board compression of hyperspectral images," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, IEEE, 2010, pp. 517–520.

[38] M Vasilescu and D. Terzopoulos, "Multilinear analysis of image ensembles: Tensorfaces," *Computer Vision?ECCV 2002*, pp. 447–460, 2002.

[39] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2773–2832, 2014.

[40] A. Cichocki, D. Mandic, L. De Lathauwer, G. Zhou, Q. Zhao, C. Caiafa, and H. A. Phan, "Tensor decompositions for signal processing applications: From two-way to multiway component analysis," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 145–163, 2015.

[41] A. T. Chaganty and P. Liang, "Spectral experts for estimating mixtures of linear regressions," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1040–1048.

[42] L. Omberg, G. H. Golub, and O. 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*, vol. 104, no. 47, pp. 18 371–18 376, 2007.

[43] C. Muralidhara, A. M. Gross, R. R. Gutell, and O. Alter, "Tensor decomposition reveals concurrent evolutionary convergences and divergences and correlations with structural motifs in ribosomal rna," *PloS one*, vol. 6, no. 4, e18768, 2011.

[44] S. P. Ponnapalli, M. A. Saunders, C. F. Van Loan, and O. Alter, "A higher-order generalized singular value decomposition for comparison of global mrna expression from multiple organisms," *PloS one*, vol. 6, no. 12, e28072, 2011.

[45] C. J. Hillar and L.-H. Lim, "Most tensor problems are NP-hard," *Journal of the ACM (JACM)*, vol. 60, no. 6, p. 45, 2013.

[46] Q. Zheng and R. Tomioka, "Interpolating convex and non-convex tensor decompositions via the subspace norm," in *Advances in Neural Information Processing Systems*, 2015, pp. 3106–3113.

[47] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[48] G. Bergqvist and E. G. Larsson, "The higher-order singular value decomposition: Theory and an application [lecture notes]," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 151–154, 2010.

[49] J. Chen and Y. Saad, "On the tensor svd and the optimal low rank orthogonal approximation of tensors," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 4, pp. 1709–1734, 2009.

[50] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[51] E. Acar and B. Yener, "Unsupervised multiway data analysis: A literature survey," *IEEE transactions on knowledge and data engineering*, vol. 21, no. 1, pp. 6–20, 2009.

[52] E. Richard and A. Montanari, "A statistical model for tensor PCA," in *Advances in Neural Information Processing Systems*, 2014, pp. 2897–2905.

[53] A. Zhang and D. Xia, "Tensor svd: Statistical and computational limits," *IEEE Transactions on Information Theory, to appear*, 2018+.

[54] S. B. Hopkins, J. Shi, and D. Steurer, "Tensor principal component analysis via sum-of-square proofs.," in *COLT*, 2015, pp. 956–1006.

[55] T. Liu, M. Yuan, and H. Zhao, "Characterizing spatiotemporal transcriptome of human brain via low rank tensor decomposition," *ArXiv preprint arXiv:1702.07449*, 2017.

[56] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the best rank-1 and rank-(r1, r2,..., rn) approximation of higher-order tensors," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1324–1342, 2000.

[57] L. Florescu and W. Perkins, "Spectral thresholds in the bipartite stochastic block model," *ArXiv preprint arXiv:1506.06737*, 2015.

[58] M. E. Newman, "Detecting community structure in networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 38, no. 2, pp. 321–330, 2004.

[59] P. Mitra, "Entrywise bounds for eigenvectors of random graphs," *Electronic journal of combinatorics*, vol. 16, no. 1, R131, 2009.

[60] J. Jin, "Fast community detection by score," *The Annals of Statistics*, vol. 43, no. 1, pp. 57–89, 2015.

[61] T. T. Cai, T. Liang, and A. Rakhlin, "Computational and statistical boundaries for submatrix localization in a large noisy matrix," *ArXiv preprint arXiv:1502.01988*, 2015.

[62] Z. Ma and Y. Wu, "Computational barriers in minimax submatrix detection," *The Annals of Statistics*, vol. 43, no. 3, pp. 1089–1116, 2015.

[63] V. Koltchinskii and D. Xia, "Perturbation of linear forms of singular vectors under gaussian noise," in *High Dimensional Probability VII*, Springer, 2016, pp. 397–423.

[64] P. Wedin, "Perturbation bounds in connection with singular value decomposition," *BIT Numerical Mathematics*, vol. 12, no. 1, pp. 99–111, 1972.

[65] R. Wang, "Singular vector perturbation under gaussian noise," *SIAM Journal on Matrix Analysis and Applications*, vol. 36, no. 1, pp. 158–177, 2015.

[66] T. T. Cai and A. Zhang, "Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics," *ArXiv preprint arXiv:1605.00353*, 2016.

[67] J. Fan, W. Wang, and Y. Zhong, "An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation," *ArXiv preprint arXiv:1603.03516*, 2016.

[68] J. Cape, M. Tang, and C. E. Priebe, "The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics," *ArXiv preprint arXiv:1705.10735*, 2017.

[69] V. Koltchinskii and K. Lounici, "Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Institut Henri Poincaré, vol. 52, 2016, pp. 1976–2013.

[70] D. Xia and M. Yuan, "On polynomial time methods for exact low rank tensor completion," *ArXiv preprint arXiv:1702.06980*, 2017.

[71] M. Rudelson, R. Vershynin, *et al.*, "Delocalization of eigenvectors of random matrices with independent entries," *Duke Mathematical Journal*, vol. 164, no. 13, pp. 2507–2538, 2015.

[72] V. Vu and K. Wang, "Random weighted projections, random quadratic forms and random eigenvectors," *Random Structures & Algorithms*, vol. 47, no. 4, pp. 792–821, 2015.

[73] V. Hore, A. Viñuela, A. Buil, J. Knight, M. I. McCarthy, K. Small, and J. Marchini, "Tensor decomposition for multiple-tissue gene expression experiments," *Nature Genetics*, vol. 48, no. 9, pp. 1094–1100, 2016.

[74] A. McCallum, K. Nigam, and L. 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*, ACM, 2000, pp. 169–178.

[75] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," *Acm Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 90–105, 2004.

[76] J. Fan and Y. Fan, "High dimensional classification using features annealed independence rules," *Annals of statistics*, vol. 36, no. 6, p. 2605, 2008.

[77] T. Hastie, R. Tibshirani, and J. Friedman, "Unsupervised learning," in *The elements of statistical learning*, Springer, 2009, pp. 485–585.

[78] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989.

[79] Q. Xiong, N. Ancona, E. R. Hauser, S. Mukherjee, and T. S. Furey, "Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets," *Genome research*, vol. 22, no. 2, pp. 386–397, 2012.

[80] M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh, "Minimax localization of structural information in large noisy matrices," in *Advances in Neural Information Processing Systems*, 2011, pp. 909–917.

[81] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, "Discovering local structure in gene expression data: The order-preserving submatrix problem," *Journal of computational biology*, vol. 10, no. 3-4, pp. 373–384, 2003.

[82] S. C. Brubaker and S. S. Vempala, "Random tensors and planted cliques," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, Springer, 2009, pp. 406–419.

[83] A. Anandkumar, R. Ge, D. Hsu, and S. Kakade, "A tensor spectral approach to learning mixed membership community models," in *Conference on Learning Theory*, 2013, pp. 867–881.

[84] L. Gauvin, A. Panisson, and C. Cattuto, "Detecting the community structure and activity patterns of temporal networks: A non-negative tensor factorization approach," *PloS one*, vol. 9, no. 1, e86028, 2014.

[85] J.-P. Aubin and I. Ekeland, *Applied nonlinear analysis*. Courier Corporation, 2006.

[86] V. Koltchinskii, "Oracle inequalities in empirical risk minimization and sparse recovery problems," 2011.

[87] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels," *IEEE Transactions on Information Theory*, vol. 48, no. 3, pp. 569–579, 2002.

[88] E. H. Lieb, "Convex trace functions and the wigner-yanase-dyson conjecture," *Advances in Mathematics*, vol. 11, no. 3, pp. 267–288, 1973.

[89] M. Talagrand, "New concentration inequalities in product spaces," *Inventiones mathematicae*, vol. 126, no. 3, pp. 505–563, 1996.

[90] O. Bousquet, "A bennett concentration inequality and its application to suprema of empirical processes," *Comptes Rendus Mathematique*, vol. 334, no. 6, pp. 495–500, 2002.

[91] V. Koltchinskii and K. Lounici, "Concentration inequalities and moment bounds for sample covariance operators," *Bernoulli*, vol. 23, no. 1, pp. 110–133, 2017.

[92] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *ArXiv preprint arXiv:1011.3027*, 2010.