HIGH-DIMENSIONAL INFERENCE FOR LOW-DIMENSIONAL STRUCTURES: DOUBLE SPARSE VECTORS AND LOW-RANK TENSORS

by

Yuchen Zhou

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2021

Date of final oral examination: 04/19/2021

The dissertation is approved by the following members of the Final Oral Committee:
Anru Zhang, Assistant Professor, Statistics
Yazhen Wang, Professor, Statistics
Garvesh Raskutti, Associate Professor, Statistics
Nicolas Garcia Trillos, Assistant Professor, Statistics
Kangwook Lee, Assistant Professor, Electrical and Computer Engineering
Ramya Korlakai Vinayak, Assistant Professor, Electrical and Computer Engineering

Acknowledgments

First, I would like to express my deepest and greatest gratitude to my advisors, Professor Anru Zhang and Professor Yazhen Wang. Anru guided me to the charming area – high-dimensional statistics. During my Ph.D., we worked on many interesting projects together and I benefited a lot from extensive discussions with him. He is very smart and experienced and can always point out possible directions for me when my projects got stuck. He set an extraordinary example of how to do research. Yazhen is an amazing statistician and also a fantastic teacher helping me build a good foundation of mathematical statistics. He has great statistical insight and complicated statistical theory becomes much easier to understand after his explanation. He also provided me with a lot of useful advice on both my research and life. The guidance I received from these two advisors has deeply shaped my thoughts on statistics, and I couldn't finish this dissertation without their help.

I would also like to thank Professor Grace Wahba, a legend in statistics and a perfect role model who is inspiring to me. I really appreciate her encouragement and helpful suggestions during my PhD. I am extremely grateful to my undergraduate advisor, Professor Liwei Wang. I spent two wonderful years in Liwei's group and finished my first research project under his supervision, which gave me the confidence to pursue a doctoral degree.

I would like to thank my final oral committee members, Professors Garvesh Raskutti, Nicolas Garcia Trillos, Kangwook Lee, Ramya Korlakai Vinayak, for their time to read my thesis and their valuable questions and helpful suggestions. Many

thanks to Professors Po-Ling Loh and Varun Jog for being part of my preliminary committee.

I also feel very fortunate to collaborate with Professor Tony Cai, Chen Dan, Jun Gao, Professor Kristoffer Arnsfelt Hansen, He Jiang, Wenlong Mou, Hong Wang, Professor Pradeep Ravikumar, Professor Pixu Shi, Professor Dong Xia, and Hongyang Zhang. I really learned a lot from them.

I would like to thank the staff members in IFDS and the Department of Statistics, Janine Gumley, Dan Barnish, Nancy Brinkerhoff, and John Schuppel. They are very nice and helped me a lot in these five years. I also want to thank my friends, Fan Chen, Rui Chen, Yue Gao, Rungang Han, Tzu-Hsiang Hung, Yutian Liu, Yuetian Luo, Yifan Mei, Tun Lee Ng, Yuling Yan, Yilin Zhang, Zihao Zheng, and Hao Zhou, for their help in both my life and research.

Finally, I would like to particularly thank my parents, Bin Zhou and Guifeng Yu, and my girlfriend, Lili Zheng, for their love and support.

Contents

List of Figures

Abstract

High-dimensional statistics has attracted considerable attention in recent years. To achieve reliable estimation and uncertainty quantification, some low-dimensional structures, including sparsity and low-rankness, are usually assumed. In this thesis, we introduce some recent advances in high-dimensional statistics with these two structures.

In Chapter 2, we study the sparse group Lasso for high-dimensional double sparse linear regression, where the parameter of interest is simultaneously element-wise and group-wise sparse. This problem is an important instance of the simultaneously structured model – an actively studied topic in statistics and machine learning. In the noiseless case, we establish matching upper and lower bounds on the sample complexity for the exact recovery of sparse vectors and for stable estimation of approximately sparse vectors, respectively. In the noisy case, upper and matching minimax lower bounds for the estimation error are obtained. We also consider the debiased sparse group Lasso and investigate its asymptotic property for the purpose of statistical inference. Numerical studies are provided to support the theoretical results.

In Chapter 3, we consider the statistical inference for several low-rank tensor models. Specifically, in the Tucker low-rank tensor PCA or regression model, provided with any estimates achieving some attainable error rate, we develop the data-driven confidence regions for the singular subspace of the parameter tensor based on the asymptotic distribution of an updated estimate by two-iteration alternating minimization. The asymptotic distributions are established under some essential conditions on the signal-to-noise ratio (in the PCA model) or sample size (in the regression model). If the parameter tensor is further orthogonally decomposable, we develop the methods and theory for inference on each individual singular vector. For the rank-one tensor PCA model, we establish the asymptotic distribution for general linear forms of principal components and confidence interval for each entry of the parameter tensor. Numerical simulations are presented to corroborate our theoretical discoveries.

Finally, Chapter 4 studies a general framework for high-order tensor SVD. We propose a new computationally efficient algorithm, tensor-train orthogonal iteration (TTOI), that aims to estimate the low tensor-train rank structure from the noisy high-order tensor observation. We develop the general upper bound on estimation error for TTOI with the support of several new representation lemmas on tensor matricizations. By developing a matching information-theoretic lower bound, we also prove that TTOI achieves the minimax optimality under the spiked tensor model. The merits of the proposed TTOI are illustrated through applications to estimation and dimension reduction of high-order Markov processes, numerical studies, and a real data example on New York City taxi travel records.

Chapter 1

Introduction

The past few decades have witnessed the astonishing development of high-dimensional statistics, which enables us to cope with large-scale data arising from contemporary applications in many areas, including biology, engineering, finance, etc. In literature, some hidden low-dimensional structures are usually assumed to break the curse of dimensionality. *Sparsity* and *low rankness* are two such low-dimensional assumptions and play important roles in high-dimensional settings. These two assumptions make models not only easier to estimate but also more interpretable. While extensive research has been devoted to studying high-dimensional statistical inference under these two structures, there remain many important problems that are less understood. In this thesis, we tackle some of these problems, including estimation and uncertainty quantification for **double sparse vectors** and **low-rank tensors**.

1.1 High-Dimensional Double Sparse Regression

In literature, the Lasso (Tibshirani, 1996) and group Lasso (Yuan and Lin, 2006) are widely used to estimate entry-wise sparse and group-wise sparse vectors under high-dimensional regression models, respectively, and their theoretical properties have been thoroughly investigated. However, in some applications including gene expression data analysis, cancer biology, climate prediction amongst others, the coefficient vector β^* may be double sparse – that is to say, it satisfies both element-wise and group-wise sparsity. To estimate β^* more accurately, Friedman et al. (2010); Simon et al. (2013) proposed the sparse group Lasso that linearly combines the ℓ_1 and group ℓ_1 penalties. Although the sparse group Lasso enjoys huge success in practice, there is still a lack of theoretical understandings, e.g., if it can achieve a smaller statistical error rate than the Lasso and group Lasso and if we can make inference based on the sparse group Lasso estimator.

In Chapter 2 (based on Cai et al., 2019b), we provide theoretical results for the sparse group Lasso under the double sparse linear regression model. Specifically, in the noiseless case, by proving matching upper and lower bounds, we show that the $\ell_1 + \ell_{1,2}$ minimization achieves optimal sample complexity to exactly recover the double sparse vector and to stably estimate approximately double sparse vector β^*, respectively. In the noisy case, we confirm that the error rate of the sparse group Lasso is optimal by establishing matching upper and minimax lower bounds. The proofs of upper bounds are based on a novel construction of an approximate dual certificate. Furthermore, inspired by Javanmard and Montanari (2014), we propose the debiased sparse group Lasso and derive its asymptotic distribution, which could be used to construct a confidence interval for β^*. Interestingly, different from other simultaneously structured models studied in the literature, our results show that **the multi-objective optimization with norms associated with entry-wise and group-wise sparsity (the sparse group Lasso or $\ell_1 + \ell_{1,2}$ minimization) indeed help us achieve better statistical performance in double sparse linear regression** than exploiting just one structure.

1.2 Tensor Data Analysis

Tensors have attracted a flurry of interest in machine learning, computational mathematics, and statistics. Different from the matrix setting, tensors have more sophisticated structures and are more challenging to handle: even the best low-rank tensor approximation and tensor operator/nuclear norms are computationally intractable. The challenge is further compounded by the explosion of dimensionality of tensors. These facts call for the development of novel theoretical analysis as well as new methods.

1.2.1 Inference for Low-rank Tensors

The estimation of low-rank tensors and their associated subspaces has been extensively studied in the literature. However, the statistical inference or uncertainty quantification of low-rank tensors, i.e., deriving asymptotic distributions and constructing the confidence intervals/regions of tensors/subspaces, have been much less investigated. In Chapter 3 (based on Xia et al., 2020), we consider this problem under the Tucker low-rank tensor PCA and regression models. Our main contributions are summarized as follows:

1. For the target low Tucker-rank tensor $\mathcal{T}$ with subspaces U_j and any estimates $\hat{U}_j^{(0)}$ achieving some attainable error rate, we propose a two-step alternating minimization algorithm with output $\hat{U}_j$ and derive data-driven confidence regions for U_j based on the asymptotic distributions of $\|\sin\Theta(U_j, \hat{U}_j)\|$ under some essential conditions on the signal-to-noise ratio (under the PCA model) and sample size (under the regression model).

2. Specifically, under the PCA model, if $\mathcal{T} = \sum_{i=1}^{r} \lambda_i \cdot u_i \otimes v_i \otimes w_i$ is orthogonal decomposable, we make inference for single principle components u_i, v_i, w_i.

3. Furthermore, under the rank-1 PCA model (i.e., $r = 1$), we prove the asymptotic distributions of linear forms of principle components u_1, v_1, w_1 and construct confidence intervals for each entry of $\mathcal{T}$.

Surprisingly, different from the matrix/vector cases considered in the literature, **making inference for low-rank tensors does not rely on any debiasing procedure.** In the literature of low-rank tensor estimation, it is widely investigated that achieving an accurate and computationally feasible estimation usually requires much stronger conditions than the one needed in the information-theoretic limit (also known as statistical-computational gap). Such essential conditions allow us to make inference without debiasing.

1.2.2 High-order Tensor SVD

In modern applications, it is increasingly more common to encounter high-order tensors, i.e., tensors with large values of order number. Compared to low-order tensors, high-order tensors contain much more parameters, which leads to enormous challenges in storage and processing.

To address this issue, (Oseledets, 2009; Oseledets and Tyrtyshnikov, 2010; Oseledets, 2011) introduced an elegant sequential low-rank structure, the tensor-train (TT) decomposition. For an order-d dimensional-p tensor, the TT-decomposition only involves $O((d-2)pr^2 + 2pr)$ parameters, which is much less than the one for the Tucker decomposition ($O(r^d + dpr)$) and the total number of the tensor entries (p^d) if d is large and thus significantly reduces the storage burden. While the low-rank tensor-train approximation under the deterministic setting is considered in the literature (Oseledets, 2011; Oseledets and Tyrtyshnikov, 2010; Bigoni et al., 2016), estimating the true low TT-rank structure from a noisy observation is more crucial in some cases (e.g., the transition probability estimation in high-order Markov chains/decision processes) and is much less studied.

In Chapter 4 (based on Zhou et al., 2020), we consider the high-order tensor SVD model. To accurately estimate the true tensor, we propose a new algorithm, Tensor-Train Orthogonal Iteration (TTOI), which consists of the initialization via TT-SVD (Oseledets, 2011) and new iterative back/forward updates. We establish lemmas that help us better understand the TT-structure and provide upper bounds for the estimation error. In addition, under the probabilistic spiked tensor model, we also prove matching minimax lower bound indicating that the TTOI can achieve the sharp error rate. As a by-product, we show that the proposed TTOI can be used to further improve the approximation result obtained by TT-SVD. As an example, we study the application of TTOI on estimating transition probabilities of high-order Markov processes and performing state aggregation. Finally, synthetic and real data analysis is provided to validate the performance of TTOI.

Chapter 2

High-Dimensional Double Sparse Regression [*]

2.1 Introduction

Consider the *high-dimensional double sparse regression* with simultaneously group-wise and element-wise sparsity structures

$$y = X\beta^* + \varepsilon, \quad \text{or equivalently} \quad y_i = X_i^\top \beta^* + \varepsilon_i, \quad i = 1, \dots, n. \tag{2.1}$$

Here, the covariates $X \in \mathbb{R}^{n \times p}$ and parameter β^* are divided into d known groups, where the jth group contains b_j variables,

$$X = [X_{(1)} \ \cdots \ X_{(d)}], \quad \beta^* = \left((\beta^*_{(1)})^\top, \cdots (\beta^*_{(d)})^\top \right)^\top, \quad X_{(j)} \in \mathbb{R}^{n \times b_j}, \beta^*_{(j)} \in \mathbb{R}^{b_j}; \tag{2.2}$$

β^* is a (s, s_g)-*sparse vector* in the sense that

$$\|\beta^*\|_{0,2} := \sum_{j=1}^{d} 1_{\{\beta^*_{(j)} \neq 0\}} \leqslant s_g \quad \text{and} \quad \|\beta^*\|_0 := \sum_{i=1}^{p} 1_{\{\beta^*_i \neq 0\}} \leqslant s. \tag{2.3}$$

[*] This work is based on Cai et al. (2019b) (`https://arxiv.org/abs/1909.09851`).

The focus of this chapter is on the estimation of and inference for β^* based on (y, X). This problem has great importance in a variety of applications. For example in genome-wide association studies (GWAS) Silver et al. (2013), the genes can be grouped into pathways and it is believed that only a small portion of the pathways contain causal single nucleotide polymorphisms (SNPs), and the number of causal SNPs is much less than the one of non-causal SNPs in a causal pathway. The sparse group Lasso has been applied to identify causal genes or SNPs associated with a certain trait Silver et al. (2013). Other examples include cancer diagnosis and therapy Vidyasagar (2014); Allahyar and De Ridder (2015), classification Rao et al. (2015), and climate prediction Chatterjee et al. (2012) among many others. The problem can also be viewed as a prototype of various problems in statistics and machine learning, such as the sparse multiple response regression Wang et al. (2013) and multiple task learning Lounici et al. (2009); Lozano and Swirszcz (2012); Zhou et al. (2017).

The sparse group Lasso Friedman et al. (2010); Simon et al. (2013); Li et al. (2015) provides a classic and straightforward estimator for β^*:

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 + \lambda_g\|\beta\|_{1,2}. \tag{2.4}$$

Here, $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$ and $\|\beta\|_{1,2} = \sum_j \|\beta_{(j)}\|_2$ are ℓ_1 and $\ell_{1,2}$ convex regularizers to account for element-wise and group-wise sparsity structures, respectively. $\lambda, \lambda_g \geqslant 0$ are tuning parameters. In the noiseless setting that $\varepsilon = 0$, one can apply the constrained $\ell_1 + \ell_{1,2}$ minimization instead to estimate β^*:

$$\begin{aligned} \hat{\beta} \quad = \quad & \arg\min \quad \lambda\|\beta\|_1 + \lambda_g\|\beta\|_{1,2} \\ & \text{subject to} \quad y = X\beta. \end{aligned} \tag{2.5}$$

In fact, when λ, λ_g tend to zero while λ/λ_g is fixed as a constant, the sparse group Lasso (2.4) tends to the $\ell_1 + \ell_{1,2}$ minimization (2.5).

When β^* is only element-wise sparse, the regular Lasso Tibshirani (1996)

$$\hat{\beta}^{L} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{2.6}$$

can be applied and its theoretical properties have been well studied. See, for example, Bickel et al. (2009); Verzelen (2012). When β^* is only group-wise sparse, the group Lasso

$$\hat{\beta}^{GL} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda_g\|\beta\|_{1,2} \tag{2.7}$$

and its variations have been widely investigated Yuan and Lin (2006); Lounici et al. (2011); Bunea et al. (2013). However, to estimate the simultaneously element-wise and group-wise sparse vector β^*, despite many empirical successes of sparse group Lasso in practice, the theoretical properties, including optimal rate of convergence and sample complexity, are still unclear so far to the best of our knowledge.

2.1.1 Simultaneously Structured Models

More broadly speaking, the simultaneously structured models, i.e., the parameter of interest has multiple structures at the same time, have attracted enormous attention in many fields including statistics, applied mathematics, and machine learning. In addition to the high-dimensional double sparse regression, other simultaneously structured models include sparse principal component analysis Johnstone and Lu (2009); Ma (2013), tensor singular value decomposition Zhang and Xia (2018); Wang and Li (2018), simultaneously sparse and low-rank matrix/tensor recovery Oymak et al. (2015); Hao et al. (2018), sparse matrix/tensor SVD Zhang and Han (2018), and sparse phase retrieval Jaganathan et al. (2013); Shechtman et al. (2014); Cai et al. (2016a). As shown in Oymak et al. (2013, 2015), by minimizing multi-objective regularizers with norms associated with these structures (such as ℓ_1 norm for element-wise sparsity, nuclear norm for low-rankness, and total variation norm for piecewise constant structures), one usually cannot do better than applying an algorithm that only exploits one structure. They particularly illustrated that simultaneously sparse and low-rank structured matrix cannot be well estimated by penalizing ℓ_1 and nuclear norm regularizers. Instead, non-convex methods were proposed and shown to achieve better performance.

However based on their results, it remains an open question whether the convex regularization, such as sparse group Lasso or $\ell_1 + \ell_{1,2}$ minimization, can achieve good performance in estimation of parameter with two types of sparsity structures, such as the aforementioned high-dimensional double sparse regression. Specifically, as illustrated in Section 2.2.2, a direct application of Oymak et al. (2015) does not provide a sample complexity lower bound for exact recovery that matches our upper bound.

2.1.2 Optimality and Related Literature

This chapter fills the void of statistical limits of sparse group Lasso and provides an affirmative answer to the aforementioned question: by exploiting both element-wise and group-wise sparsity structures, the $\ell_1 + \ell_{1,2}$ regularization does provide better performance in high-dimensional double sparse regression. Particularly in the noiseless case, it is shown that (s, s_g)-sparse vectors can be exactly recovered and approximately (s, s_g)-sparse vectors can be stably estimated with high probability whenever the sample size satisfies $n \gtrsim s_g \log(d/s_g) + s \log(es_g b)$, where $b = \max_{1 \leqslant i \leqslant d} b_i$. On the other hand, we prove that exact recovery cannot be achieved by $\ell_1 + \ell_{1,2}$ regularization and stable estimation of approximately (s, s_g)-sparse vectors is impossible in general unless $n \gtrsim s_g \log(d/s_g) + s \log(es_g b/s)$. We then consider the noisy case and develop the matching upper and lower bounds on the convergence rate for the estimation error. Simulation studies are carried out and the results support our theoretical findings. In addition, statistical inference for the individual coordinates of β^* is studied. A confidence interval is constructed based on the debiased sparse group Lasso estimator and its asymptotic property. The results show that by exploring the simultaneously element-wise and group-wise sparsity structures, the debiased sparse group Lasso requires less sample size than the debiased Lasso and debiased group Lasso in the literature Zhang and Zhang (2014); Javanmard and Montanari (2014); Mitra and Zhang (2016); Cai and Guo

(2017).

The theoretical analysis of sparse group Lasso and $\ell_1+\ell_{1,2}$ minimization is highly non-trivial. First, the regularizer $\lambda\|\cdot\|_1+\lambda_g\|\cdot\|_{1,2}$ is not decomposable with respect to the support of β^* so that the classic techniques of decomposable regularizers Negahban et al. (2012) and null space property Stojnic et al. (2008) may not be suitable here. Despite a substantial body of literature on high-dimensional element-wise sparse vector estimation based on restricted isometry property (RIP) Candes et al. (2006); Candes and Tao (2007); Cai and Zhang (2013a,b, 2014) and restricted eigenvalue Bickel et al. (2009), these techniques cannot provide nearly optimal results for sparse group Lasso here as it is technically difficult to partition general vectors into simultaneously element-wise and group-wise ones that preserves some ordering structures. Departing from the previous literature, our theoretical analysis relies on a novel construction of approximate dual certificate. See Section 2.2.3 for further details. Although our results mostly focus on the performance of sparse group Lasso and $\ell_1+\ell_{1,2}$ estimators, the techniques of approximate dual certificate on multi-norm structures here can also be of independent interest.

The statistical properties of sparse group Lasso and related estimators have been studied previously. For example, Chatterjee et al. (2012) developed consistency results for estimators with a general tree-structured norm regularizers, of which the sparse group Lasso is a special case. Poignard (2018) analyzed the asymptotic behaviors of the adaptive sparse group Lasso estimator. Rao et al. (2015, 2013) studied the multi-task learning and classification problems based on a variant of sparse group Lasso estimator. Li et al. (2015) studied multivariate linear regression via sparse group Lasso. Ahsen and Vidyasagar (2017) provided a theoretical framework for developing error bounds of the group Lasso, sparse group Lasso, and group Lasso with tree structured overlapping groups. Specifically, their results imply that the group-wise sparse signal can be exactly recovered with high probability by solving (2.5) if the sample size satisfies $n \gtrsim s_g\left(b+\log d\right)$. Different from previous results, this chapter focused on both the required sample size and convergence rate of estimation error of sparse group Lasso. To the best of our knowledge, this is the first result that provides optimal theoretical guarantees for both the sample complexity and estimation error of sparse group Lasso.

2.1.3 Organization of the Chapter

The rest of the article is organized as follows. After a brief introduction to notation and preliminaries in Section 2.2.1, the main theoretical results on constrained $\ell_1 + \ell_{1,2}$ minimization in the noiseless setting is presented in Section 2.2.2 and the key proof ideas are explained in Section 2.2.3. Results for sparse group Lasso in the noisy setting are discussed in Section 2.3. In particular, the optimal rate of estimation error and statistical inference are studied in Sections 2.3.1 and 2.3.2, respectively. In Section 2.4.1, we introduce a practical scheme to select tuning parameters. In Section 2.4.2, we provide simulation results in both noiseless and noisy cases to justify our theoretical findings. The proofs of technical results are given in Section 5.1. All technical lemmas and their proofs can be found in Appendix 5.1.9.

2.2 $\ell_1 + \ell_{1,2}$ Minimization in Noiseless Case

2.2.1 Notation and Preliminaries

The following notation will be used throughout the chapter. We denote $a \wedge b = \min\{a, b\}$, $a \vee b = \max\{a, b\}$. Let $\text{sgn}(\cdot)$ be the sign function, i.e., $\text{sgn}(x) = 1, 0$, or -1, if $x > 0, x = 0$, or $x < 0$, respectively. $H_\alpha(\cdot)$ is the soft-thresholding function such that $H_\alpha(x) = \text{sgn}(x) \cdot \{(|x| - \alpha) \vee 0\}$ for any $x \in \mathbb{R}$. We say $a \lesssim b$ and $a \gtrsim b$ if $a \leqslant Cb$ and $b \leqslant Ca$ for some uniform constant $C > 0$, respectively. $a \asymp b$ means $a \lesssim b$ and $a \gtrsim b$ both hold. Let the uppercase $C, C_1, C_0, \ldots$ and lowercase $c, c_1, c_0, \ldots$ denote large and small positive constants respectively, whose actual values vary from time to time. Throughout the chapter, we focus on the parameter index set $\{1, \ldots, p\}$ partitioned into d groups. Denote $(1), \ldots, (d) \subseteq \{1, \ldots, p\}$ as the index sets belonging to each group. Additionally, for any group index subset $G \subseteq \{1, \ldots, d\}$, define $(G) = \cup_{j \in G}(j)$, $(G^c) = \cup_{j \notin G}(j)$. For any vector γ and index subset T, $\gamma_T \in \mathbb{R}^{|T|}$ represents the sub-vector of γ with index set T. In particular, $\gamma_{(G)}$ represents the sub-vector of γ in the union of Groups $j \in G$. Define the ℓ_q norm of any vector γ as $\|\gamma\|_q = (\sum_i |\gamma_i|^q)^{1/q}$. For any vector $\beta \in \mathbb{R}^p$ with group structures, we also define the ℓ_{q_1,q_2} norm for any $0 \leqslant q_1, q_2 \leqslant \infty$ as

$$\|\gamma\|_{q_1,q_2} = \left(\sum_{j=1}^{d} \|\gamma_{(j)}\|_{q_2}^{q_1}\right)^{1/q_1} = \left\{\sum_{j=1}^{d} \left(\sum_{i\in(j)} |\gamma_i|^{q_2}\right)^{q_1/q_2}\right\}^{1/q_1}.$$

In particular, $\|\gamma\|_{0,2} = \sum_{j=1}^{d} 1_{\{\gamma_{(j)}\neq 0\}}$ is the number of non-zero groups of γ, $\|\gamma\|_{\infty,2} = \max_j \|\gamma_{(j)}\|_2$ is the maximum ℓ_2 norm among all groups of γ, and $\|\gamma\|_{1,2} = \sum_{j=1}^{d} \|\gamma_{(j)}\|_2$ is the group-wise ℓ_1 penalty. With a slight abuse of notation, we simply denote $\|\gamma_T\|_{q_1,q_2} = \|u\|_{q_1,q_2}$ if $u \in \mathbb{R}^p$, u restricted on subset T is γ_T and u restricted on T^c is 0.

The focus of this chapter is on simultaneously element-wise and group-wise sparse vectors defined as follows.

Definition 2.2.1 (Simultaneous element-wise and group-wise sparsity)**.** *Assume $\beta^* \in \mathbb{R}^p$ is associated with group partition $(1), \ldots, (d)$. For positive integers s, s_g satisfying $s_g \leqslant d$ and $s_g \leqslant s \leqslant \max_{\Omega\subseteq\{1,\ldots,d\},|\Omega|=s_g} \sum_{i\in\Omega} b_i$, we say β^* is (s, s_g)-sparse if*

$$\|\beta^*\|_{0,2} = \sum_{j=1}^{d} 1_{\{\beta^*_{(j)}\neq 0\}} \leqslant s_g, \quad \|\beta^*\|_0 = \sum_{i} 1_{\{\beta^*_i\neq 0\}} \leqslant s.$$

2.2.2 Noiseless Case and Sample Complexity

To analyze the performance of sparse group Lasso and $\ell_1 + \ell_{1,2}$ minimization, we first introduce the following assumption on the design matrix X.

Assumption 2.2.1 (Sub-Gaussian assumption)**.** *Suppose all rows of X are i.i.d. centered sub-Gaussian distributed. Specifically, $\mathbb{E}X_{i\cdot} = 0, \mathrm{Var}(X_{i\cdot}^\top) = \Sigma$, and for any $\alpha \in \mathbb{R}^p$, we have $\mathbb{E}\exp\left(\alpha^\top\Sigma^{-1/2}X_{i\cdot}^\top\right) \leqslant \exp\left(\kappa^2\|\alpha\|_2^2/2\right)$ for constant $\kappa > 0$. We also assume there* *exist two constants* $C_{\max} \geqslant c_{\min} > 0$ *such that* $c_{\min} \leqslant \sigma_{\min}(\Sigma) \leqslant \sigma_{\max}(\Sigma) \leqslant C_{\max}$*, where* $\sigma_{\max}(\Sigma)$ *and* $\sigma_{\min}(\Sigma)$ *are the largest and smallest eigenvalues of* Σ*, respectively.*

Clear, a random matrix X with i.i.d. standard normal entries satisfies this assumption – this design is referred to as the Gaussian ensemble and has been considered as a benchmark setting in compressed sensing and high-dimensional regression literature Candes and Plan (2011); Javanmard and Montanari (2018).

The following theorem shows that the $\ell_1 + \ell_{1,2}$ minimization achieves the exact recovery with high probability when β^* is simultaneously element-wise and group-wise sparse, X is weakly dependent, and Assumption 2.2.1 holds. The theorem also provides a more general upper bound on estimation error if β^* is approximately element-wise and group-wise sparse.

Theorem 2.2.1 (ℓ_1+$\ell_{1,2}$ minimization in noiseless case)**.** *Suppose one observes* $y = X\beta^*$*, where* X *has the group structure* (2.2) *and satisfies Assumption 2.2.1,* β^* *is* (s, s_g)*-sparse, and* $b = \max_{1 \leqslant i \leqslant d} b_i$*. Let* T *be the support of* β^**. Suppose there exist uniform constants* $C, c > 0$ *such that*

$$n \geqslant C\left(s_g \log(d/s_g) + s\log(es_g b)\right), \tag{2.8}$$

$$\max_{i \in T^c} \left\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\right\|_2 \leqslant c/\sqrt{s}, \tag{2.9}$$

then the constrained $\ell_1 + \ell_{1,2}$ *minimization* (2.5) *with* $\lambda_g = \sqrt{s/s_g}\lambda$ *achieves the exact recovery with probability at least* $1 - C\exp(-cn/s)$*.*

Moreover, if $\beta^* \in \mathbb{R}^p$ *is a general vector and* $\hat{\beta}$ *is the solution to the constrained* $\ell_1 + \ell_{1,2}$ *minimization* (2.5) *with* $\lambda_g = \sqrt{s/s_g}\lambda$*, then*

$$\|\hat{\beta} - \beta^*\|_2 \lesssim \min_{\substack{S: \|\beta^*_S\|_0 \leqslant s, \|\beta^*_S\|_{0,2} \leqslant s_g, \\ \max_{i \in S^c} \|\Sigma_{i,S}\Sigma^{-1}_{S,S}\|_2 \leqslant c/\sqrt{s}}} \left(\frac{1}{\sqrt{s}}\|\beta^*_{S^c}\|_1 + \frac{1}{\sqrt{s_g}}\|\beta^*_{S^c}\|_{1,2}\right). \tag{2.10}$$

with probability at least $1 - C\exp(-cn/s)$*.*

Remark 2.2.1 (Interpretation and comparison)**.** *In Theorem 2.2.1, the required sample size for achieving exact recovery contains two terms:* $s_g \log(d/s_g)$ *and* $s\log(es_g b)$*. In-*

tuitively speaking, $s_g \log(d/s_g)$ *corresponds to the complexity of identifying* s_g *non-zero groups and* $s \log(es_g b)$ *corresponds to the complexity of estimating* s *non-zero elements of* β *in* s_g *known groups.*

When β^* *is only element-wise or group-wise sparse, one can apply respectively the classic* ℓ_1 *or* $\ell_{1,2}$ *minimization to recover* β^**,*

$$\hat{\beta}^{\ell_1} = \arg\min_{\beta} \|\beta\|_1 \quad \textit{subject to} \quad y = X\beta, \tag{2.11}$$

$$\hat{\beta}^{\ell_{1,2}} = \arg\min_{\beta} \|\beta\|_{1,2} \quad \textit{subject to} \quad y = X\beta. \tag{2.12}$$

The ℓ_1 *minimization and* $\ell_{1,2}$ *minimization here are respectively the special form of the regular Lasso and group Lasso (if* $\lambda, \lambda_g = 0_+$ *in* (2.6) *and* (2.7)*), respectively. Especially if the group size* $b_1 \asymp \cdots \asymp b_d \asymp b$*, to ensure exact recovery in the noiseless setting with high probability,* (2.11) *requires* $n \gtrsim C s \log(ebd/s)$ *Foucart and Rauhut (2013) and group Lasso requires* $n \gtrsim s_g(b + \log(ed/s_g))$*. The* $\ell_1 + \ell_{1,2}$ *minimization* (2.5) *has provable advantages over both regular and group Lasso when* $b \gg \log(d) \gg \log(es_g b)$ *and* $s_g b / \log(es_g b) \gg s \gg s_g$*. In particular, when* $s_g = s$*, the double sparse regression reduces to the vanilla sparse linear regression, and the upper bound* (2.10) *matches the classic upper bound for* ℓ_1 *minimization Candes and Plan (2011).*

In addition, Condition (2.9) *is an important technical condition we used in our theoretical analysis.*

Next, we consider the sample complexity lower bound. Suppose $b_1 = b_2 = \cdots = b_d$ and $d \geqslant 2s_g$. Recall that one observes $y = X\beta^*$ without noise and aims to estimate the (s, s_g)-sparse vector β^* based on y and X. As indicated by classic results in compressed sensing Candes and Tao (2005), with sufficient computing power, the ℓ_0 minimization below achieves exact recovery of β^*

$$\hat{\beta}^{\ell_0} = \arg\min \|\beta\|_0 \quad \text{subject to} \quad X\beta = y \tag{2.13}$$

as along as X is non-degenerate and $n \geqslant 2s$. This bound is actually sharp: when

$n < 2s$, for any set $T \subseteq \{1, \dots, db\}$ with cardinality $2s$, one can find a vector γ such that $\text{supp}(\gamma) \subseteq T$ and $X\gamma = 0$. By choosing an appropriate T, we can split the support γ to obtain two (s, s_g)-sparse vectors β_1, β_2 satisfying $\beta_1 + \beta_2 = \gamma$. Then, $X\beta_1 = X(-\beta_2)$ but there is no way to distinguish β_1 and β_2 merely based on X and $y = X\beta_1 = X(-\beta_2)$.

However, the ℓ_0 minimization (2.13) is computational infeasible in practice while a larger sample size is required for applying more practical methods. The following theorem shows that by performing the convex ℓ_1 regularization, $\ell_{1,2}$ regularization, or any weighted combination of them, one requires at least $\Omega(s_g \log(d/s_g) + s \log(es_g b/s))$ observations to ensure exact recovery of (s, s_g)-sparse vectors.

Theorem 2.2.2 (Sample complexity lower bound for exact recovery)**.** *Suppose* $b_1 = \cdots = b_d = b$, $d, b \geqslant 3$. *Suppose* X *is an* n*-by-*(db) *matrix. If every* $(2s, 2s_g)$*-sparse vector* $\beta \in \mathbb{R}^{db}$ *is a minimizer of the following programming for some* $(\lambda, \lambda_g) \in \{(\lambda, \lambda_g) : \lambda, \lambda_g \geqslant 0, \lambda + \lambda_g > 0\}$*:*

$$\min_z \lambda \|z\|_1 + \lambda_g \|z\|_{1,2} \quad \textit{subject to} \quad Xz = y = X\beta.$$

In other words, if the $\ell_1 + \ell_{1,2}$ *minimization exactly recover all* $(2s, 2s_g)$*-sparse vector* β*, then we must have* $n \gtrsim s_g \log(d/s_g) + s \log(es_g b/s)$*.*

The following sample complexity lower bound shows that for arbitrary methods, to ensure stable estimation of all approximately sparse vectors, one requires at least $\Omega(s_g \log(d/s_g) + s \log(es_g b/s))$ observations.

Theorem 2.2.3 (Sample complexity lower bound for stable estimation)**.** *Suppose* $b_1 = \cdots = b_d = b$, $b, d \geqslant 3$. *Assume there exists a matrix* $X \in \mathbb{R}^{n \times (bd)}$*, a map* $\Delta : \mathbb{R}^n \to \mathbb{R}^{bd}$ *(*Δ *may depend on* X*), and a constant* $C > 0$ *satisfying*

$$\|\beta - \Delta(X\beta)\|_2 \leqslant C \left(\frac{\|\beta\|_1}{\sqrt{s}} + \frac{\|\beta\|_{1,2}}{\sqrt{s_g}} \right) \tag{2.14}$$

for all $\beta \in \mathbb{R}^p$ *and some* s, s_g *satisfying* $d \geqslant s_g, s_g b \geqslant s \geqslant s_g$*. There exists constants* C_0

and c_0 *that depend only on* C *such that whenever* $s_g \geqslant C_0$*, we must have*

$$n \geqslant c_0(s_g \log(d/s_g) + s \log(es_g b/s)).$$

Remark 2.2.2 (Optimality and comparison with previous results)**.** *Theorems 2.2.2 and 2.2.3 show that the sample complexity upper bound in Theorem 2.2.1 is rate-optimal under a weak condition:* $\log(es_g b) \asymp \log(es_g b) - \log(s)$ *or* $\log(d) \geqslant 2s\log(s)/s_g$*. Oymak, et al. Oymak et al. (2015) provided a general analysis for convex regularization of simultaneously structured parameter estimation. Specifically for the high-dimensional double sparse regression, a direct application of their Theorem 3.2 and Corollary 3.1 implies that if* $\ell_1 + \ell_{1,2}$ *minimization can exactly recover* (s, s_g)*-sparse vector* β^* *with a constant probability, one must have* $n \gtrsim s$*. We can see that Theorem 2.2.2 provides a sharper lower bound on sample complexity.*

In addition, by setting $s_g = s$*, the lower bound in Theorems 2.2.2 and 2.2.3 reduces to* $n \gtrsim s\log(p/s)$*, which matches the optimal sample complexity lower bound for exact recovery of s-sparse vectors (Foucart and Rauhut, 2013, Theorem10.11, Proposition 10.7). By setting* $s = s_g b$*, we obtain a sample complexity lower bound* $n \gtrsim s_g(b + \log(d/s_g))$ *for (approximate)* s_g*-group-wise sparse vector recovery and stable estimation. To the best of our knowledge, this is the first sample complexity lower bound for group Lasso.*

2.2.3 Proof Sketches

We briefly discuss the proof sketches of the main technical results in this section. The detailed proofs are postponed to Section 5.1.

The proof of Theorem 2.2.1 is based on a novel dual certificate scheme. The dual certificate Bertsekas and Nedic (2003) has been used in the theoretical analysis for various convex optimization methods in high-dimensional problems, such as matrix completion Candès and Recht (2009); Gross (2011), compressed sensing Candes and Plan (2011), robust PCA Candès et al. (2011), tensor completion Yuan and Zhang (2016), etc. The high-dimensional double sparse linear regression exhibits different aspects from these previous works due to the simultaneous sparsity structure. In particular, we can show that if the u_{et} defined below is in the row space of X, it can be used as an exact dual certificate for recovery of (s, s_g)-sparse vector β^*:

$$u_{et} = v_{et} + w_{et} \in \mathbb{R}^p, \quad \begin{cases} (v_{et})_{(j)} = \sqrt{s/s_g}\beta^*_{(j)}/\|\beta^*_{(j)}\|_2, & j \in G; \\ \|(v_{et})_{(j)}\|_2 < \sqrt{s/s_g}, & j \in G^c; \end{cases} \quad \begin{cases} (w_{et})_T = \operatorname{sgn}(\beta^*_T) \\ \|(w_{et})_{T^c}\|_\infty < 1. \end{cases} \tag{2.15}$$

Here, T and G are the element-wise and group-wise supports of β^*:

$$T = \{i : \beta_i \neq 0\} \subseteq \{1, \dots, p\}, \quad G = \{j : \beta_{(j)} \neq 0\} \subseteq \{1, \dots, d\}.$$

Roughly speaking, u_{et} is the sub-gradient of objective function (2.5) evaluated at $\beta = \beta^*$. If u_{et} is in the row space of X, the sub-gradient will be perpendicular to the feasible set of (2.5), which implies that β^* is the unique minimizer of $\ell_1 + \ell_{1,2}$ minimization (2.5).

For more general vector β^* that does not necessarily have a sparse support T or G, we consider the following (s, s_g)-sparse approximation:

$$\begin{aligned} \beta^{ap} = & \arg\min_S \frac{1}{\sqrt{s}}\|\beta^*_{S^c}\|_1 + \frac{1}{\sqrt{s_g}}\|\beta^*_{S^c}\|_{1,2} \\ & \text{subject to} \quad \|\beta^*_S\|_0 \leqslant s. \quad \|\beta^*_S\|_{0,2} \leqslant s_g, \quad \max_{i \in S^c} \|\Sigma_{i,S}\Sigma^{-1}_{S,S}\|_2 \leqslant c/\sqrt{s}. \end{aligned} \tag{2.16}$$

Let $T = \{i : \beta^{ap}_i \neq 0\}$ and $G = \{j : (\beta^{ap})_{(j)} \neq 0\}$ be the element-wise and group-wise supports of β^{ap}. Define

$$\widetilde{u}_0 = \widetilde{v}_0 + \widetilde{w}_0 \in \mathbb{R}^p, \quad \begin{cases} (\widetilde{v}_0)_{(j)} = \sqrt{s/s_g}\beta^*_{T,(j)}/\|\beta^*_{T,(j)}\|_2, & j \in G; \\ \|(\widetilde{v}_0)_{(j)}\|_2 < \sqrt{s/s_g}, & j \in G^c; \end{cases} \quad \begin{cases} (\widetilde{w}_0)_T = \operatorname{sgn}(\beta^*_T) \\ \|(\widetilde{w}_0)_{T^c}\|_\infty < 1. \end{cases} \tag{2.17}$$

Here $\beta^*_{T,(j)} \in \mathbb{R}^{b_j}$ is the subvector β^* restricted on the j-th group with all entries in T^c set to zero. Similarly to the exactly sparse case, if $\widetilde{u}_0$ is in the row space of X and the true β^* is approximately (s, s_g)-sparse, the minimizer of (2.5) will be close to β^*.

However, it is often difficult to find an exact dual certificate that lies in the row space of X and satisfies stringent conditions in (2.15) or (2.17). We instead propose to analyze via the *approximate dual certificate* defined as (2.18) in the following lemma.

Lemma 2.2.1 (Approximate dual certificate for sparse group Lasso)**.** *Suppose* T, G *are element-wise and group-wise support defined in* (2.16)*.* $\widetilde{u}_0$ *is defined in* (2.17)*. Assume* X *satisfies* $\sigma_{\min}\left(X_T^\top X_T/n\right) \geqslant c_{\min}/2$*. If there exists* $u \in \mathbb{R}^p$ *in the row span of* X *satisfying*

$$\begin{aligned} &\|u_T - (\widetilde{u}_0)_T\|_2 \cdot \max_{i \in T^c} \left\|X_T^\top X_i/n\right\|_2 \leqslant c_{\min}/8, \\ &\|H_{1/2}(u_{(G^c)})\|_{\infty,2} \leqslant \sqrt{s_0}/2, \quad \|u_{(G)\setminus T}\|_\infty \leqslant 1/2, \end{aligned} \tag{2.18}$$

Then the conclusion of Theorem 2.2.1 (2.10) *holds with probability at least* $1-2e^{-cn}$*. Here,* $H_{1/2}(\cdot)$ *is the soft-thresholding operator defined at the beginning of Section 2.2.*

If we additionally assume β^* *is* (s, s_g)*-sparse, then* β^* *is the unique solution to the sparse group* $\ell_1 + \ell_{1,2}$ *minimization* (2.5) *with probability at least* $1 - 2e^{-cn}$*.*

Lemma 2.2.1 shows that the conclusion of Theorem 2.2.1 holds if there exists an approximate dual certificate u satisfying the condition (2.18). The following lemma shows that, under the assumptions in Theorem 2.2.1, one can find such an approximate dual certificate with high probability.

Lemma 2.2.2. *Suppose* X *has group structure* (2.2) *and satisfies Assumption 2.2.1. Recall* $\sigma_{\min}(X_T^\top X_T/n)$ *is the least eigenvalue of* $X_T^\top X_T/n$*. Then* $\sigma_{\min}\left(X_T^\top X_T/n\right) \geqslant 1/2$ *and* (2.18) *holds with probability at least* $1 - Ce^{-cn/s}$*, where* T *is defined in* (2.16)*.*

Another key technical tool to the proof of Theorem 2.2.1 is the following Lemma, which shows that X satisfies the restricted isometry property for all simultaneously element-wise and group-wise sparse vectors with high probability when there are enough samples.

Lemma 2.2.3. *If* $n \geqslant C(s_g \log(d/s_g) + s \log(es_g b))$*,*

$$\frac{c_{\min}}{2}\|\gamma\|_2^2 \leqslant \frac{1}{n}\|X\gamma\|_2^2 \leqslant (C_{\max}+\frac{c_{\min}}{2})\|\gamma\|_2^2, \quad \forall \gamma \in \{\gamma \in \mathbb{R}^p : \|\gamma\|_0 \leqslant 2s, \|\gamma\|_{0,2} \leqslant 2s_g\} \tag{2.19}$$

with probability at least $1 - 2e^{-cn}$.

Next we briefly discuss the proof of Theorem 2.2.2. Consider the quotient space $\mathbb{R}^{db}/\ker(X) = \{[\gamma] := x + \ker(X), \gamma \in \mathbb{R}^{db}\}$ and define an associated norm as $\|[\gamma]\| = \inf_{\nu \in \ker(X)}\{\lambda\|\gamma - \nu\|_1 + \lambda_g\|\gamma - \nu\|_{1,2}\}$. We show that there exist N different (s, s_g)-sparse vectors $\beta^{(1)}, \ldots, \beta^{(N)}$ such that $\log(N) \asymp s\log(es_g b/s) + s_g\log(d/s_g)$ and $\|[\beta^{(i)}]\| = 1, \|[\beta^{(i)}] - [\beta^{(j)}]\| \geqslant 2/9$ for all $1 \leqslant i \neq j \leqslant N$. By a property of the packing number and the fact that $\dim(\mathbb{R}^{db}/\ker(X)) \leqslant n$, we must have $N \leqslant 10^n$. Thus $n \gtrsim \log(N) \asymp s\log(es_g b/s) + s_g\log(d/s_g)$.

We prove Theorem 2.2.3 by contradiction. Assume that

$$n < c_0\left(s\log(es_g b/s) + s_g\log(d/s_g)\right) \tag{2.20}$$

for a sufficiently small constant c_0. Let $\|\cdot\| = \|\cdot\|_1 + \sqrt{s/s_g}\|\cdot\|_{1,2}$ and $B = \{x \in \mathbb{R}^{db} : \|x\| \leqslant 1\}$ be the unit ball associated with $\|\cdot\|$. Define

$$d^n(B, \mathbb{R}^p) = \inf_{\substack{L^n \text{ is a subspace of } \mathbb{R}^p \\ \text{with } \dim(\mathbb{R}^p/L^n) \leqslant n}} \left\{\sup_{\beta \in B \cap L^n} \|\beta\|_2\right\},$$

We have $d^n(B, \mathbb{R}^p) \leqslant \frac{C}{\sqrt{s}}$ by the assumption of this theorem. We can also show that there exists a uniform constant $c > 0$ such that

$$d^n(B, \mathbb{R}^p) \geqslant c\min\left\{\frac{1}{\sqrt{s_0}}, \left[\left(\frac{s_g}{s}\log\left(\frac{c\frac{s}{s_g}d\log(es_g b/s)}{n}\right) + \log(es_g b/s)\right)/n\right]^{1/2}\right\}.$$

The previous two inequalities and (2.20) together imply that

$$n \geqslant c\left(s_g\log\left(\frac{c\frac{s}{s_g}d\log(es_g b/s)}{n}\right) + s\log(es_g b/s)\right) \geqslant c_0\left(s\log(es_g b/s) + s_g\log(d/s_g)\right) > n.$$

This contradiction shows that $n \geqslant c_0\left(s\log(es_g b/s) + s_g\log(d/s_g)\right)$.

2.3 Sparse Group Lasso in Noisy Case

We now turn to the noisy case.

2.3.1 Optimal Rate of Estimation Error of Sparse Group Lasso

When observations are noisy, we have the following theoretical guarantee for the sparse group Lasso.

Theorem 2.3.1 (Upper bound of estimation error)**.** *Suppose* $y = X\beta^* + \varepsilon$*,* X *satisfies Assumption 2.2.1,* $n \geqslant C\,(s_g \log(d/s_g) + s\log(es_g b))$ *for some uniform constant* $C > 0$*,* $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$*, and* $b = \max_{1 \leqslant i \leqslant d} b_i$*. Then the sparse group Lasso estimator* (2.4) *with*

$$\lambda = C\sigma\sqrt{(s\log(es_g b) + s_g \log(ed/s_g))n/s} \quad \textit{and} \quad \lambda_g = \sqrt{s/s_g}\lambda$$

satisfies

$$\begin{aligned}&\|\hat{\beta} - \beta^*\|_2 \\ &\lesssim \min_{\substack{S: \|\beta^*_S\|_0 \leqslant s, \|\beta^*_S\|_{0,2} \leqslant s_g, \\ \max_{i \in S^c} \|\Sigma_{i,S}\Sigma^{-1}_{S,S}\|_2 \leqslant c/\sqrt{s}}} \left\{ \sqrt{\frac{\sigma^2(s_g \log(d/s_g) + s\log(es_g b))}{n}} + \frac{\|\beta^*_{S^c}\|_1}{\sqrt{s}} + \frac{\|\beta^*_{S^c}\|_{1,2}}{\sqrt{s_g}} \right\}\end{aligned}$$

with probability at least $1 - C\exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s}\right)$.

Especially, if β^* *is exactly* (s, s_g)*-sparse and* $\max_{i \in T^c} \|\Sigma_{i,T}\Sigma^{-1}_{T,T}\|_2 \leqslant c/\sqrt{s}$ *holds, then*

$$\|\hat{\beta} - \beta^*\|_2^2 \lesssim \frac{\sigma^2(s_g \log(d/s_g) + s\log(es_g b))}{n} \tag{2.21}$$

with probability at least $1 - C\exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s}\right)$.

In addition, we focus on the following class of simultaneously element-wise and group-wise sparse vectors,

$$\mathcal{F}_{s,s_g} = \{\beta : \|\beta\|_0 \leqslant s, \|\beta\|_{0,2} \leqslant s_g\}.$$

The following minimax lower bound of estimation error holds.

Theorem 2.3.2 (Lower bound of estimation error)**.** *Suppose* X *satisfies Assumption 2.2.1,* $b_1 = \cdots = b_d = b$, *and* $d, b \geqslant 3$. *Then we have*

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathcal{F}_{s,s_g}} \mathbb{E}\|\hat{\beta} - \beta\|_2^2 \gtrsim \frac{\sigma^2(s_g \log(ed/s_g) + s \log(es_g b/s))}{n}.$$

Remark 2.3.1. *Theorems 2.3.1 and 2.3.2 together show that the sparse group Lasso yields the minimax optimal rate of convergence as long as the following condition holds:* $\log(es_g b) \asymp \log(es_g b) - \log(s)$ *or* $\log(d) \gtrsim s \log(s)/s_g$.

Remark 2.3.2. *We briefly discuss the main proof ideas of Theorem 2.3.2 here. First, we randomly generate a series of subsets* $\Omega^{(i)} \subseteq \{1, \ldots, p\}$ *as feasible supports of* (s, s_g)*-sparse vectors. Then, we prove by a probabilistic argument that there exist* $N \gtrsim (s_g \log(d/s_g) + s \log(es_g b/s))$ *subsets* $\{\Omega^{(i)}\}_{i=1}^N$ *such that* $|\Omega^{(i)} \cap \Omega^{(j)}| < 8s_g \lfloor s/s_g \rfloor /9$ *for any* $i < j$. *Next, we construct a series of candidate* (s, s_g)*-sparse vectors* $\beta^{(i)}$ *such that* $\beta_k^{(i)} = \tau 1_{\{k \in \Omega^{(i)}\}}$. *Intuitively speaking,* $\{\beta^{(i)}\}_{i=1}^N$ *are non-distinguishable based only on observations* (y, X) *by such a construction. Theorem 2.3.2 then follows by choosing an appropriate* τ *and the generalized Fano's lemma.*

2.3.2 Statistical Inference via Debiased Sparse Group Lasso

We further consider the statistical inference for β^* under the double sparse linear regression model. First, let $\hat{\beta}$ be the sparse group Lasso estimator given by (2.4). Inspired by the recent advances in inference for high-dimensional linear regression Zhang and Zhang (2014); Van de Geer et al. (2014); Javanmard and Montanari (2014); Cai and Guo (2017), we propose the following *debiased sparse group Lasso estimator,*

$$\hat{\beta}^u = \hat{\beta} + \frac{1}{n}\hat{M}X^\top \left(Y - X\hat{\beta}\right). \tag{2.22}$$

Here, $\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^n X_k X_k^\top$ is the sample covariance matrix and $\hat{M} = [\hat{m}_1 \ \cdots \ \hat{m}_p]^\top$ is an approximation of the inverse covariance matrix Σ^{-1}, where $\hat{m}_i$ is the solution to the following convex optimization,

$$\begin{aligned} \text{minimize} \quad & m^\top \hat{\Sigma} m \\ \text{subject to} \quad & \|H_\alpha(\hat{\Sigma} m - e_i)\|_{\infty,2} \leqslant \gamma. \end{aligned} \tag{2.23}$$

Here, H_α is the soft-thresholding operator with thresholding level α defined at the beginning of Section 2.2 and e_i is the i-th vector in the canonical basis of $\mathbb{R}^p$. The following theorem establishes an asymptotic result for debiased sparse group Lasso.

Theorem 2.3.3 (Asymptotic distribution of debiased sparse group Lasso)**.** *Suppose* $\beta^* \in \mathbb{R}^p$ *is* (s, s_g)*-sparse,* $X \in \mathbb{R}^{n\times p}$ *satisfies Assumption 2.2.1, and* $\max_{i\in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2 \leqslant c/\sqrt{s}$*. Set* $\lambda = C\sigma\sqrt{\frac{(s\log(es_g b)+s_g\log(d/s_g))n}{s}}$ *and* $\lambda_g = \sqrt{\frac{s}{s_g}}\lambda$ *in* (2.4)*,* $\alpha = \frac{\lambda}{n\sigma}, \gamma = \sqrt{\frac{s}{s_g}}\frac{\lambda}{n\sigma}$ *in* (2.23)*. Then with probability at least* $1-C\exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}\right)$*, the debiased sparse group Lasso estimator* $\hat{\beta}^u$ *can be decomposed as* $\sqrt{n}\left(\hat{\beta}^u - \beta^*\right) = \Delta + w$*, where*

$$\|\Delta\|_\infty \leqslant \frac{C\left(s\log(es_g b) + s_g\log(ed/s_g)\right)}{\sqrt{n}}\sigma, \quad w|X \sim N\left(0, \sigma^2\hat{M}\hat{\Sigma}\hat{M}^\top\right). \tag{2.24}$$

In particular, if $\sqrt{n} \gg s\log(es_g b) + s_g\log(ed/s_g)$*, for any* $1 \leqslant i \leqslant p$*,*

$$\frac{\sqrt{n}\left(\hat{\beta}_i^u - \beta_i^*\right)}{\sqrt{\hat{m}_i^\top\hat{\Sigma}\hat{m}_i}} \to N\left(0, \sigma^2\right). \tag{2.25}$$

Remark 2.3.3. (2.25) *provides a method to construct confidence intervals for* β^**. Specifically if* $\hat{\sigma}$ *is a consistent estimator of* σ*, such as the scaled sparse group Lasso to be discussed in Section 2.5,*

$$\left[\hat{\beta}_i^u - \Phi^{-1}(1-\alpha/2)\hat{\sigma}\sqrt{\frac{\hat{m}_i^\top\hat{\Sigma}\hat{m}_i}{n}}, \quad \hat{\beta}_i^u + \Phi^{-1}(1-\alpha/2)\hat{\sigma}\sqrt{\frac{\hat{m}_i^\top\hat{\Sigma}\hat{m}_i}{n}}\right]$$

would be an asymptotic $(1-\alpha)$*-confidence interval for* β_i^**. We can see that the debiased* *sparse group Lasso estimator has the provably advantage on sample complexity* ($n \gg (s\log(es_g b) + s_g \log(ed/s_g))^2$*) over the ones via debiased Lasso* ($n \gg s \log p$*, see Zhang and Zhang (2014); Javanmard and Montanari (2014); Cai and Guo (2017)) or debiased group Lasso* ($n \gg (s_g b + s_g \log p)^2$*, see Mitra and Zhang (2016)) for constructing asymptotic confidence intervals of* β^**.*

2.4 Simulation Studies

In this section, we investigate the numerical performance of the sparse group Lasso and $\ell_1 + \ell_{1,2}$ minimization for double sparse regression. The results support our theoretical findings in Sections 2.2 and 2.3. We first discuss the practical choice for the tuning parameters used in the proposed algorithms.

2.4.1 Practical Selection of Tuning Parameters

By introducing τ as a surrogate for $(\lambda_g/\lambda)^2$, we can rewrite the $\ell_1 + \ell_{1,2}$ minimization and the sparse group Lasso as

$$\hat{\beta} = \arg\min \|\beta\|_1 + \sqrt{\tau}\|\beta\|_{1,2} \quad \text{subject to} \quad y = X\beta, \tag{2.26}$$

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 + \lambda\sqrt{\tau}\|\beta\|_{1,2}. \tag{2.27}$$

As suggested by Theorems 2.2.1 and 2.3.1, the theoretical choice of the tuning parameters (λ, τ) relies on σ, s, and s_g in sparse group Lasso and $\ell_1 + \ell_{1,2}$ minimization for double sparse regression. These values, however, are usually unknown in practice. In addition, those theoretical values of tuning parameters may not achieve the best finite-sample numerical performance. We thus introduce in this section a data-driven approach to tuning parameter selection using K-fold cross-validation.

We first discuss how to select τ in the $\ell_1 + \ell_{1,2}$ minimization (2.26). Recall n is the sample size, p is the total number of covariates, d is the number of groups, $b_1, \dots, b_d$ are the number of covariates in each group, and $b = \max_j b_j$. Since the theoretical value $\tau = s/s_g$ and s/s_g must satisfy $1 \leqslant s/s_g \leqslant b$, for a given integer $L \geqslant 1$, we introduce a grid

$$S_0 = \{b^{(l-1)/(L-1)} : 1 \leqslant l \leqslant L\} \tag{2.28}$$

as a set of candidate values for τ. Here, the grid size L can be set to a typical value of 10, or a larger value if more computing power is available. We split the data $\{X_i, y_i\}_{i=1}^n$ into K groups. For $1 \leqslant k \leqslant K$, let $J_k \subset \{1, \ldots, n\}$ be the index set of the kth group and $J_k^c = \{1, \ldots, n\} \backslash J_k$. For each $\tau \in S_0$, we solve

$$\hat{\beta}^{(k)}(\tau) = \arg\min \|\beta\|_1 + \sqrt{\tau}\|\beta\|_{1,2} \quad \text{subject to} \quad y_{J_k^c} = X_{[J_k^c,:]}\beta$$

and calculate the prediction error

$$\hat{R}(\tau) = \sum_{k=1}^{K} \sum_{j \in J_k} \left(y_j - X_{[j,:]}\hat{\beta}^{(k)}(\tau)\right)^2.$$

Let τ_* be the minimizer of the prediction error: $\tau_* = \arg\min_{\tau \in S_0} \hat{R}(\tau)$. Then, the final estimator $\hat{\beta}$ is calculated using (2.26) with τ_*.

Then we consider the sparse group Lasso (2.27), which includes two tuning parameters (τ, λ). We still define S_0 in (2.28) as a grid of candidate values of τ. Following the idea in (Simon et al., 2013, Section 3.3), for each $\tau \in S_0$, we begin with a large value of $\lambda_{\max}(\tau)$ so that $\hat{\beta}$, the outcome of sparse group Lasso (2.27) with tuning parameters $(\tau, \lambda_{\max}(\tau))$, is zero (this can be achieved by the SGL package[†]). Let $\lambda_{\min}(\tau)$ be a small fraction of $\lambda_{\max}(\tau)$ (e.g., $\lambda_{\min} = 0.1\lambda_{\max}$ as suggested in (Simon et al., 2013, Section 5)). Then we define $\Lambda(\tau) = \left\{\{\lambda_{\min}(\tau)\}^{(L-l)/(L-1)} \cdot \{\lambda_{\max}(\tau)\}^{(l-1)/(L-1)} : l = 1, \ldots, L\right\}$. Next, we split the data $\{X_i, y_i\}_{i=1}^n$ into K groups. For $1 \leqslant k \leqslant K$, let $J_k \subset \{1, \ldots, n\}$ be the index set of the kth group and $J_k^c = \{1, \ldots, n\} \backslash J_k$. For each $\tau \in S_0$, $\lambda \in \Lambda(\tau)$,

[†] https://cran.r-project.org/web/packages/SGL/index.html

and $k \in \{1, \dots, K\}$, we solve

$$\hat{\beta}^{(k)}(\tau, \lambda) = \arg\min_{\beta} \left\| y_{J_k^c} - X_{[J_k^c,:]}\beta \right\|_2^2 + \lambda\|\beta\|_1 + \lambda\sqrt{\tau}\|\beta\|_{1,2}$$

and calculate the prediction error

$$\hat{R}(\tau, \lambda) = \sum_{k=1}^{K} \sum_{j \in J_k} \left(y_j - X_{[j,:]}\hat{\beta}^{(k)}(\tau, \lambda) \right)^2.$$

Let (τ_*, λ_*) be the minimizer of the prediction error: $(\tau_*, \lambda_*) = \arg\min_{\tau \in S_0, \lambda \in \Lambda(\tau)} \hat{R}(\tau, \lambda)$. The final estimator $\hat{\beta}$ is calculated using (2.27) with (τ_*, λ_*).

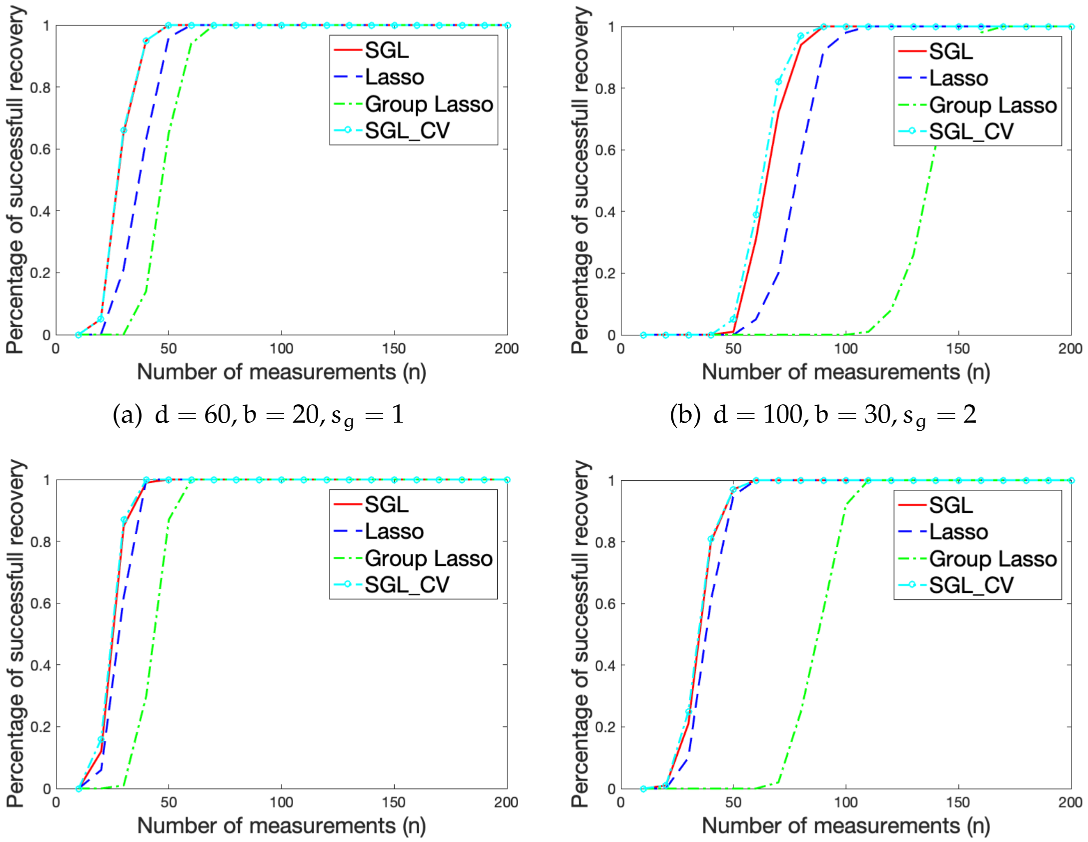
In our simulation studies next, we will examine the performance of this cross-validation scheme with $K = L = 10$, $\lambda_{\text{min}} = 0.1\lambda_{\text{max}}$.

2.4.2 Numerical Results

We begin by considering the sample complexity for the exact recovery in the noiseless case. Suppose all group sizes are equal ($b_1 = \cdots = b_d = b$) and the number of observations n varies from 5 to 200. We consider four simulation designs with (1) $d = 60, b = 20, s_g = 1$; (2) $d = 100, b = 30, s_g = 2$; (3) $d = b = 20, s_g = 1$; and (4) $d = b = 40, s_g = 1$. For each setting, we randomly draw $X \in \mathbb{R}^{n \times db}$ with i.i.d. standard normal entries, construct the fixed vector $\beta^* \in \mathbb{R}^{db}$ satisfying

$$\beta^*_{(j)} = \begin{cases} (1, 2, 3, 4, 5, 0, \dots, 0) \in \mathbb{R}^b & j = 1, \dots, s_g; \\ 0 & j = s_g + 1, \dots, d, \end{cases}$$

and generate $y = X\beta^* = \sum_{j=1}^{s_g} X_{(j)}\beta^*_{(j)}$. We implement the $\ell_1 + \ell_{1,2}$ minimization (2.5) with $\lambda_g = \sqrt{s/s_g}\lambda$ (`SGL`), ℓ_1 minimization (2.11) (`Lasso`), and $\ell_{1,2}$ minimization (2.12) (`Group Lasso`), and $\ell_1 + \ell_{1,2}$ minimization (2.5) with the tuning parameter λ_g/λ selected using cross validation discussed in Section 2.4.1 (`SGL_CV`). An exact recovery of β^* is considered to be successful if $\|\hat{\beta} - \beta^*\|_2 \leqslant 10^{-4}$. The successful recovery rate based on 100 replicates is shown in Figure 2.1. It can be seen that `SGL` and SGL_CV have comparable performance and both methods have significantly better performance than Lasso and Group Lasso. This is in line with our theoretical results.



(a) $d = 60, b = 20, s_g = 1$

(b) $d = 100, b = 30, s_g = 2$

(c) $d = 20, b = 20, s_g = 1$

(d) $d = 40, b = 40, s_g = 1$

Figure 2.1: Exact recovery rate in the noiseless case

Then we consider the noisy case and focus on average estimation errors of different methods. We generate

$$y = X\beta^* + \varepsilon = \sum_{j=1}^{s_g} X_{(j)}\beta^*_{(j)} + \varepsilon,$$

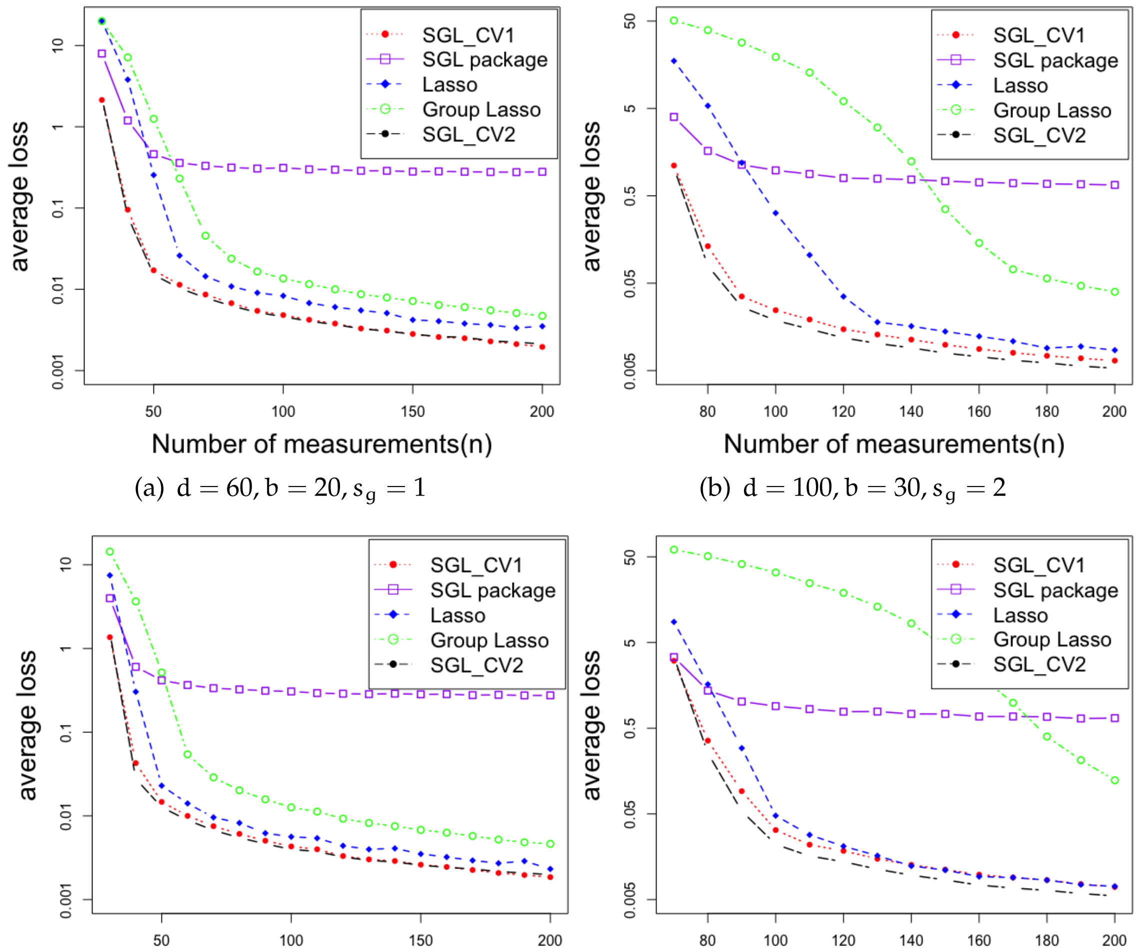
where X, β^* are drawn in the same way as the previous setting and $\varepsilon \overset{\text{iid}}{\sim} N(0, 0.1^2)$. We consider four designs: i. $d = 60, b = 20, s_g = 1$; ii. $d = 100, b = 30, s_g = 2$; iii. $d = b = 20, s_g = 1$; and iv. $d = b = 40, s_g = 2$. For each case, the number of observations n is chosen from an equally spaced sequence from 5 to 200 and the simulation is replicated for 500 times. We compare the average estimation error of (a) `SGL_CV1`: sparse group Lasso with theoretical value $\lambda_g = \sqrt{s/s_g}\lambda$ and λ selected via cross validation; (b) `SGL_package`: sparse group Lasso via SGL package[‡] in R with the option of automatic tuning parameter selection; (c) `Lasso`: regular Lasso with tuning parameter selected via cross validation; (d) `group Lasso`: group Lasso with tuning parameter selected via cross validation; (e) `SGL_CV2`: sparse group Lasso with both λ and λ_g selected using the proposed cross validation scheme. We can see the proposed method `SGL_CV2` achieves smaller estimation error than all other methods, including `SGL_CV1`, the focus of our theory. These experimental results demonstrate our theory and the applicability of the proposed cross-validation scheme.

2.5 Discussions

In this chapter, we study the high-dimensional double sparse regression and investigate the theoretical properties of the sparse group Lasso and $\ell_1 + \ell_{1,2}$ minimization. Particularly, we develop the matching upper and lower bounds on the sample complexity for $\ell_1 + \ell_{1,2}$ minimization in the noiseless case. We also prove that the sparse group Lasso achieves minimax optimal rate of convergence in a range of settings in the noisy case. Our results give an affirmative answer to the open question for high-dimensional statistical inference for simultaneously structured model: by introducing both ℓ_1 and $\ell_{1,2}$ penalties, one can achieve better performance on estimation and statistical inference for simultaneously element-wise and group-wise sparse vectors.

In addition to β^*, the estimation and inference for noise level σ is another

[‡] `https://cran.r-project.org/web/packages/SGL/index.html`



(a) $d = 60, b = 20, s_g = 1$

(b) $d = 100, b = 30, s_g = 2$

(c) $d = 20, b = 20, s_g = 1$

(d) $d = 40, b = 40, s_g = 2$

Figure 2.2: Average estimation error in the noisy case

importance task in high-dimensional double sparse regression. Motivated by the recent development of scaled Lasso Sun and Zhang (2012), one may consider the following scaled sparse group Lasso estimator:

$$\{\hat{\beta}^s, \hat{\sigma}\} = \arg\min_{\beta\in\mathbb{R}^p, \sigma>0} \left\{ \frac{\|y - X\beta\|_2^2}{\sigma} + n\sigma + \widetilde{\lambda}\|\beta\|_1 + \widetilde{\lambda}_g\|\beta\|_2 \right\},$$

where $\widetilde{\lambda}$ and $\widetilde{\lambda}_g$ are tuning parameters that do not rely on σ. The consistency of $\hat{\sigma}$ can be established based on similar ideas of scaled Lasso in the literature Sun and Zhang (2012); Javanmard and Montanari (2014) and the approximate dual certificate in this work.

Moreover, our technical results can be useful in a variety of other problems with simultaneous sparsity structures. For example, Tibshirani et al. (2005); Rinaldo (2009) considered the estimation of piece-wise constant sparse signals, i.e., both the signal vector and the difference between successive entries of the signal vector are sparse. Jalali and Fazel (2013); Jalali et al. (2019) discussed the estimation of structured parameters where both the number of non-zero elements and the number of distinct values of the parameter vectors are small. Sprechmann et al. (2010) considered the estimation of matrices with simultaneous sparsity structures within each block and among different blocks. It is interesting to further study the statistical limits, including the sample complexity and minimax optimal rate of convergence for these problems. In particular, based on the specific sparsity structures of each problem, we can introduce corresponding multi-objective regularizers and the convex regularization methods. The corresponding approximate dual certificates can be proposed, constructed, and analyzed to provide strong theoretical guarantees.

Chapter 3

Inference for Low-rank Tensors [*]

3.1 Introduction

An mth order tensor is a multiway array along m directions. Recent years have witnessed a fast growing demand for the collection, processing, and analysis of data in the form of tensors. These tensor data commonly arise, to name a few, when features are collected from different domains, or when multiple data copies are provided by various agents or sources. For instances, the worldwide food trading flows (De Domenico et al., 2015; Jing et al., 2020) produce a fourth order tensor (countries × countries × food × years); the online click-through data (Han et al., 2020; Sun et al., 2017) in e-commerce form a third order tensor (users × categories × periods); Berkeley human mortality data (Wilmoth and Shkolnikov, 2006; Zhang and Han, 2019) yield a third order tensor (ages × years × countries). In addition, the applications of tensor also include collaborative filtering (Karatzoglou et al., 2010; Shah and Yu, 2019), recommender system design (Bi et al., 2018), computational imaging (Zhang et al., 2020b), and neuroimaging (Zhou et al., 2013). Researchers have made tremendous efforts to innovate effective methods for the analysis of tensor data.

[*] This work is based on Xia et al. (2020) (`https://arxiv.org/abs/2012.14844`).

Low-rank models have rendered fundamental toolkits to analyze tensor data. A tensor $\mathcal{T} \in \mathbb{R}^{p_1 \times \cdots \times p_m}$ has low Tucker rank (or multilinear rank) if all fibers[†] of $\mathcal{T}$ along different ways lie in rank-reduced subspaces of high-dimension, say $\{U_j\}_{j=1}^m$, respectively (Tucker, 1966). The core assumption of low-rank tensor models is that the observed data is driven by an *unknown* low-rank tensor $\mathcal{T}$, while the Tucker low-rank conditions can significantly reduce the model complexity. Consequently, the analysis of tensor data often boils down to the estimation and inference of the low-rank tensor $\mathcal{T}$ or its principal components based on the given datasets.

In the literature, a rich list of methods have been developed for the *estimation* of low-rank tensor $\mathcal{T}$ and the associated subspace U_j, such as alternating minimization (Anandkumar et al., 2014a), convex regularization (Tomioka and Suzuki, 2013; Yuan and Zhang, 2016), power iterations (Anandkumar et al., 2014a), orthogonal iteration (De Lathauwer et al., 2000b; Zhang and Xia, 2018), vanilla gradient descent with spectral initialization (Cai et al., 2019a), projected gradient descent (Chen et al., 2019a), simultaneous gradient descent (Han et al., 2020), etc. However, in many practical scenarios, to enable more reliable decision making and prediction, it is important to quantify the estimation error in addition to point estimations. This task, referred to as *uncertainty quantification* or *statistical inference*, usually involves the construction of confidence intervals/regions for the unknown parameters through the development of the (approximate) distributions of the estimators. The statistical inference or uncertainty quantification for low-rank tensor models remains largely unexplored. In this chapter, we aim to make an attempt to this fundamental and challenging problem. Our focus is on two basic yet important settings: low-rank tensor PCA and tensor regression, which we briefly summarize as follows.

Tensor principal component analysis (PCA) is among the most basic problem of *unsupervised inference* for low-rank tensors. We consider the tensor PCA model (Anandkumar et al., 2014a; Richard and Montanari, 2014; Liu et al., 2017; Zhang

[†] Here, the tensor fibers are the counterpart of matrix columns and rows for tensors. See Kolda and Bader (2009) for a review.

and Xia, 2018; Chen, 2019; Perry et al., 2020), which assumes

$$\mathcal{A} = \mathcal{T} + \mathcal{Z}, \tag{3.1}$$

where the signal $\mathcal{T}$ admits a low-rank decomposition (3.6) and the noise $\mathcal{Z}$ contains i.i.d. entries with mean zero and variance σ^2. A central goal of tensor PCA is on the estimation and inference of $\mathcal{T}$ and/or $\{U_j\}_j$, i.e. the low-rank structure from $\mathcal{A}$. Tensor PCA has been proven effective for learning hidden components in Gaussian mixture models (Anandkumar et al. (2014a)), where $\{U_j\}_j$ represent the hidden components. By constructing confidence regions of $\{U_j\}_j$, we are able to make uncertainty quantifications for the hidden components of Gaussian mixture models. In addition, confidence regions of $\{U_j\}_j$ can be useful for the inference of spatial and temporal patterns of gene regulation during brain development (Liu et al. (2017)).

Low-rank tensor regression can be seen as one of the most basic setting of *supervised inference* for low-rank tensors. Specifically, suppose we observe a set of random pairs $\{\mathcal{X}_i, Y_i\}_{i=1}^n$ associated as

$$Y_i = \langle \mathcal{T}, \mathcal{X}_i \rangle + \xi_i. \tag{3.2}$$

Here, the main point of interest is $\mathcal{T}$, a low-rank tensor that characterizes the association between response Y and covariate $\mathcal{X}$, and ξ_i is the noise term. When the tensor order is $m = 2$, this problem is reduced to the widely studied *trace matrix regression model* in the literature (Candes and Plan, 2010; Koltchinskii et al., 2011; Tomioka and Suzuki, 2013; Cai et al., 2013; Chen et al., 2019a; Koltchinskii and Xia, 2015; Rauhut et al., 2017; Raskutti et al., 2019; Fan et al., 2019). This model can also be used as the prototype of many problems in high-dimensional statistics and machine learning, including phase retrieval (Candes et al., 2013b) and blind deconvolution (Li et al., 2019b). When $m \geqslant 3$, this problem has been studied under the scenario of high-order interaction pursuit (Hao et al., 2020) and large-scale linear system from partial differential equations (Lynch et al., 1964). In applications of tensor regression to neuroimaging analysis, the principal components of $\mathcal{T}$ are useful in the understanding of the association between disease outcomes and brain image patterns (Zhou et al. (2013)). In addition, the principal components determine the cluster memberships of neuroimaging data (Sun and Li, 2019). Confidence regions of $\{U_j\}_j$ in the aforementioned applications allow us to make significance test for the detected regions of interest, and to make uncertainty quantifications for clustering outcomes, respectively.

In addition to tensor PCA and regression, there is a broad range of low-rank tensor models, such as tensor completion (Yuan and Zhang, 2016; Montanari and Sun, 2018; Xia and Yuan, 2019; Zhang, 2019), generalized tensor estimation (Han et al., 2020), and tensor high-order clustering (Wu et al., 2016; Feizi et al., 2017; Chi et al., 2018; Sun and Li, 2019; Wang and Zeng, 2019; Luo and Zhang, 2020c). A common goal of these problems is to accurately estimate and make inference on some type of low-rank structures.

3.1.1 Summary of the Main Results

In this chapter, we aim to develop the methods and theory for statistical inference under the low-rank tensor PCA and regression models. First, suppose the target tensor $\mathcal{T}$ is Tucker low-rank with singular subspace U_j as point of interest. Given any estimator $\hat{U}_j^{(0)}$ that achieves some reasonable estimation error, we introduce a straightforward two-iteration alternating minimization scheme (Algorithms 1 and 2 in Section 3.3.1) and obtain $\hat{U}_j$. Surprisingly, we are able to derive an asymptotic distribution of $\|\sin\Theta(\hat{U}_j, U_j)\|_F^2$ (definition of sin-theta distance is postponed to Section 3.2) even though $\hat{U}_j$ is from non-convex iterations. Under the tensor PCA model with some essential conditions on SNR, we prove that

$$\frac{\|\sin\Theta(\hat{U}_j, U_j)\|_F^2 - p_j\sigma^2\|\Lambda_j^{-1}\|_F^2}{\sqrt{2p_j}\sigma^2\|\Lambda_j^{-2}\|_F} \xrightarrow{d.} N(0,1) \quad \text{as} \quad p_j \to \infty. \tag{3.3}$$

Here, Λ_j is the diagonal matrix containing all non-zero singular values of the jth matricization of $\mathcal{G}$ (see definition of matricization in Section 3.2). Under the tensor regression model with some essential conditions on sample size and SNR, we prove that

$$\frac{\|\sin\Theta(\hat{U}_j, U_j)\|_F^2 - p_j n^{-1}\sigma^2\|\Lambda_j^{-1}\|_F^2}{\sqrt{2p_j}n^{-1}\sigma^2\|\Lambda_j^{-2}\|_F} \xrightarrow{d.} N(0,1) \quad \text{as} \quad p_j \to \infty. \tag{3.4}$$

Then, we consider a special class of *orthogonally decomposable tensors* $\mathcal{T}$ in the sense that $\mathcal{T} = \sum_{j=1}^r \lambda_j \cdot u_j \otimes v_j \otimes w_j \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ for orthonormal vectors $\{u_j\}_j$, $\{v_j\}_j$, and $\{w_j\}_j$. The orthogonally decomposable tensor has been widely studied as a benchmark setting for tensor decomposition in the literature (Kolda, 2001; Chen and Saad, 2009; Robeva, 2016; Belkin et al., 2018; Auddy and Yuan, 2020). In addition, the (near-)orthogonally decomposable tensors have been used in various applications of statistics and machine learning, such as latent variable model (Anandkumar et al., 2014a), hidden Markov models (Anandkumar et al., 2012), etc. Under the tensor PCA model, we prove that

$$\frac{\langle \hat{u}_j, u_j\rangle^2 - (1 - p_j\sigma^2\lambda_j^{-2})}{\sqrt{2p_j}\sigma^2\lambda_j^{-2}} \xrightarrow{d.} N(0,1) \quad \text{as} \quad p_1 \to \infty \tag{3.5}$$

for $j = 1, \cdots, r$ when some essential SNR condition holds. Here, $\{\hat{u}_j, \hat{v}_j, \hat{w}_j\}_j$ are the estimates of $\{u_j, v_j, w_j\}_j$ (up to some permutation of index j) based on a two-step power iteration (Algorithm 3). Similar results can also be obtained for $\langle \hat{v}_j, v_j\rangle^2$ and $\langle \hat{w}_j, w_j\rangle^2$.

Next, we propose the estimates of $\Lambda_j, \lambda_j, \sigma^2$ that are involved in the asymptotic distributions of $\|\sin\Theta(\hat{U}_j, U_j)\|_F^2$ in (3.3)(3.4) and $\langle \hat{u}_j, u_j\rangle^2$ in (3.5). We prove that the asymptotic normality in (3.3)(3.4)(3.5) still hold after plugging in these estimates. These results immediately yield the data-driven confidence regions for U_j (Tucker low-rank settings) or $\{u_j\}_j$ (orthogonally decomposable settings).

If $\mathcal{A}$ is a rank-1 tensor, the low-rank tensor PCA model reduces to the widely studied *rank-1 tensor PCA* (see a literature survey in Section 3.1.2). Under this model, we establish the asymptotic normality of any linear functionals for the power iteration estimators $\hat{u}, \hat{v}, \hat{w}$: for all unit vectors $q_i \in \mathbb{R}^{p_i}$, under regularity conditions, we have

$$\left(\frac{\langle q_1, \hat{u}-u\rangle + \frac{p_1\langle q_1,u\rangle}{2(\lambda/\sigma)^2}}{\sqrt{\frac{p_1\langle q_1,u\rangle^2}{2(\lambda/\sigma)^4}+\frac{1-\langle q_1,u\rangle^2}{(\lambda/\sigma)^2}}}, \frac{\langle q_2, \hat{v}-v\rangle + \frac{p_2\langle q_2,v\rangle}{2(\lambda/\sigma)^2}}{\sqrt{\frac{p_2\langle q_2,v\rangle^2}{2(\lambda/\sigma)^4}+\frac{1-\langle q_2,v\rangle^2}{(\lambda/\sigma)^2}}}, \frac{\langle q_3, \hat{w}-w\rangle + \frac{p_3\langle q_3,w\rangle}{2(\lambda/\sigma)^2}}{\sqrt{\frac{p_3\langle q_3,w\rangle^2}{2(\lambda/\sigma)^4}+\frac{1-\langle q_3,w\rangle^2}{(\lambda/\sigma)^2}}}\right)^\top \xrightarrow{d} N(0, I_3)$$

as $p_1, p_2, p_3 \to \infty$. We further derive the entrywise asymptotic distribution for each entry of the estimator $\hat{\mathcal{T}}$, and propose a thresholding procedure to construct the asymptotic $1-\alpha$ entrywise confidence interval for $\mathcal{T}$, which is the first of such work to our best knowledge.

Our theoretical results reveal a key message: under the tensor PCA and regression model, *the inference of principal components can be efficiently done when a computationally feasible optimal estimate is achievable.* In recent literature, it is widely observed in many low-rank tensor models (See 3.1.2 for a review of literature) that in order to achieve an accurate estimation in polynomial time, one often requires a more stringent condition than what is needed in the statistical (or information-theoretic) limit. Such a statistical and computational gap becomes a “blessing" to the statistical inference of low-rank tensor models, as debiasing can become unnecessary if those strong but essential conditions for computational feasibility are met!

3.1.2 Related Prior Work

This chapter is related to a broad range of literature in high-dimensional statistics and matrix/tensor analysis. First, a variety of methods have been proposed for tensor PCA in the literature. A non-exhaustive list include high-order orthogonal iteration (De Lathauwer et al., 2000b); sequential-HOSVD (Vannieuwenhoven et al., 2012), inference for low-rank matrix completion (Foucart et al., 2017; Chen et al., 2019b), (truncated) power iteration (Anandkumar et al., 2014b; Sun et al., 2017; Liu et al., 2017), STAT-SVD (Zhang and Han, 2019). In addition, the computational hardness was widely considered for tensor PCA. Particularly in the worse case scenario, the best low-rank approximation of tensors can be NP hard (De Silva and

Lim, 2008; Hillar and Lim, 2013). The average-case computational complexity for tensor PCA model has also been widely studied under various computational models, including the Sum-of-Squares (Hopkins et al., 2015), optimization landscape (Arous et al., 2019), average-case reduction (Zhang and Xia, 2018; Luo and Zhang, 2020c; Brennan and Bresler, 2020; Luo and Zhang, 2020a), and statistical query (Dudeja and Hsu, 2020). It has now been widely justified that the SNR condition $\lambda_{\min}/\sigma \geqslant Cp^{3/4}$ is essential to ensure tensor PCA is solvable in polynomial time.

Regression of low-rank tensor has attracted enormous attention recently. Various methods, such as the (regularized) alternating minimization (Zhou et al., 2013; Sun and Li, 2017; Li et al., 2018), convex regularization (Tomioka and Suzuki, 2013; Raskutti et al., 2019), projected gradient descent (Chen et al., 2019a; Rauhut et al., 2017), importance sketching (Zhang et al., 2020a) were studied. Recently, Han et al. (2020) proved that a gradient descent algorithm can recover a low-rank third order tensor $\mathcal{T}$ with statistically optimal convergence rate when the sample size n is much greater than the tensor dimension $p^{3/2}$. It was widely conjectured that $n \geqslant Cp^{3/2}$ is essential for the problem being solvable in polynomial time (see Barak and Moitra (2016) for the evidence).

While the statistical inference for low-rank tensor models remain largely unexplored, there have been several recent results demystifying the statistical inference for low-rank *matrix* models. For matrix PCA, Xia (2019b) introduced an explicit representation formula for $\hat{U}_j\hat{U}_j^\top$. A more precise characterization of the distribution of $\|\sin\Theta(\hat{U}_j, U_j)\|_F^2$ was established in Bao et al. (2018) by random matrix theory. On the other hand, the estimators of tensor PCA are often calculated from iterative optimization algorithms (e.g., power iterations or gradient descent) in existing literature, while the estimator of matrix PCA is based on non-iterative schemes. Due to the complex statistical dependence involved in iterative optimization algorithms, it is significantly more challenging to analyze the asymptotic distribution of the estimator in tensor PCA than the one in matrix PCA. We also note that, when studying the asymptotic distributions of individual eigenvectors, an eigengap condition is often crucial for matrix PCA but not required for tensor PCA.

The inference and uncertainty quantification were also considered for low-rank matrix regression. For example, Carpentier et al. (2019) introduced a debiased estimator based on the nuclear norm penalized low-rank estimator. Cai et al. (2016b) introduced another debiasing technique and characterize the entrywise distribution of the debiased estimator under the restricted isometry property. Xia (2019a) studied a debiased estimator for matrix regression under the isotropic Gaussian design and established the distribution of $\|\sin\Theta(\hat{U}_j, U_j)\|_F^2$ under nearly optimal sample size conditions. All these approaches rely on suitable debiasing of certain initial estimates. In addition to low-rank estimation, an appropriate debias was found crucial for high-dimensional sparse regression (Zhang and Huang, 2008), and various debiasing schemes were introduced (Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014). Interestingly, as will be shown in Section 3.3, our estimating and inference procedure for low-rank tensor regression does not involve debiasing.

Statistical inference for low-rank models are particularly challenging for tensor problems. In a concurrent work, Huang et al. (2020) studied the statistical inference for tensor spiked model. Recently, Cai et al. (2020) studied the entrywise statistical inference for noisy low-rank tensor completion based on an incoherence condition on U_js, i.e., all the rows of U_j have comparable magnitudes. In comparison, our results do not require further conditions on U_js or debiasing.

3.1.3 Organizations

The rest of the chapter is organized as follows. After an introduction on notation and preliminaries in Section 3.2, we discuss the inference for principal components under the Tucker low-rank models in Section 3.3. Specifically, a general two-iteration alternating minimization procedure, inference for tensor PCA, inference for tensor regression, and a proof sketch are given in Sections 3.3.1, 3.3.2, 3.3.3, and 3.3.4, respectively. In Section 3.4, we focus on the inference for individual singular vectors of orthogonally decomposable tensors. The asymptotic distribution and entrywise confidence interval are discussed for rank-1 tensor PCA model in Section 3.5. Section

5.2 includes some algorithms for tensor PCA and regression in the literature and all proofs of the main technical results.

3.2 Notation and Preliminaries

We use calligraphic letters $\mathcal{T}, \mathcal{G}$ to denote tensors, upper-case letters U, W to denote matrices, and lower-case letters u, w to denote vectors or scalars. For a random variable X and $\alpha > 0$, the Orlicz ψ_α-norm of X is defined as

$$\|X\|_{\psi_\alpha} = \inf\{K > 0 : \mathbb{E}\{\exp(|X|/K)^\alpha\} \leqslant 2\}.$$

Specifically, a random variable with finite ψ_2-norm or ψ_1-norm is called the sub-Gaussian or sub-exponential random variable, respectively. Let e_j denote the jth canonical basis vector whose dimension varies at different places. Let $\mathrm{rank}(\mathcal{T})$ be the Tucker rank of $\mathcal{T}$ and write $(a_1, \dots, a_m) \leqslant (b_1, \dots, b_m)$ if $a_j \leqslant b_j$ for all $j \in [m]$. We use $\|\cdot\|_F$ for Frobenius norm, $\|\cdot\|$ for matrix spectral norm and $\|\cdot\|_2$ for vector ℓ_2-norm. Denote $\mathbb{S}^{p-1} = \{v \in \mathbb{R}^p : \|v\|_2 \leqslant 1\}$ as the set of p-dimensional unit vectors. Define $\mathbb{O}_{p,r} = \{U \in \mathbb{R}^{p\times r} : U^\top U = I_r\}$ as the set of all p-by-r matrices with orthonormal columns. In particular, $\mathbb{O}_r$ is the set of all $r \times r$ orthogonal matrices.

We denote $\times_j$ the jth multi-linear product between a tensor and matrix. For instance, if $\mathcal{G} \in \mathbb{R}^{r_1\times r_2\times r_3}$ and $V_1 \in \mathbb{R}^{p_1\times r_1}$, then

$$\mathcal{G} \times_1 V_1 = \Big(\sum_{j_1=1}^{r_1} \mathcal{G}(j_1, i_2, i_3)V(i_1, j_1)\Big)_{i_1\in[p_1], i_2\in[r_2], i_3\in[r_3]}.$$

We write $(U_1, \cdots, U_m) \cdot \mathcal{G}$ in short for $\mathcal{G} \times_1 U_1 \times_2 \cdots \times_m U_m$. Let $\mathcal{M}_j$ be the jth tensor matricization that rearranges each mode-j fiber of $\mathcal{T} \in \mathbb{R}^{p_1\times\cdots\times p_d}$ to a column of $\mathcal{M}_j(\mathcal{T}) \in \mathbb{R}^{p_j\times(p_1\cdots p_d/p_j)}$.

We say $\mathcal{T}$ has Tucker rank $(r_1, \cdots, r_m)$ if it admits a Tucker decomposition

$$\mathcal{T} = (U_1, \cdots, U_m) \cdot \mathcal{G}, \tag{3.6}$$

where $\mathcal{G} \in \mathbb{R}^{r_1 \times \cdots \times r_m}$ and $U_i \in \mathbb{O}_{p_i, r_i}$ for $i \in [m]$. The Tucker decomposition (3.6) can be roughly seen as a generalization of matrix singular value decomposition (SVD) to higher-order tensors, where U_j can be viewed as principal components of the jth matricization of $\mathcal{T}$, and $\mathcal{G}$ contains the singular values. In the case that $r_1 = \cdots = r_m = r$ and $\mathcal{G}$ is diagonalizable, we say $\mathcal{T}$ is *orthogonally decomposable*. If $\mathcal{T}$ satisfies Tucker decomposition (3.6), one has

$$\mathcal{M}_j(\mathcal{T}) = U_j \mathcal{M}_j(\mathcal{G}) \big(U_1 \otimes \cdots \otimes U_{j-1} \otimes U_{j+1} \otimes \cdots \otimes U_m \big)^\top \in \mathbb{R}^{p_j \times (p_1 \cdots p_m / p_j)}.$$

Here $\otimes$ stands for Kronecker product so that $U \otimes W \in \mathbb{R}^{(p_1 p_2) \times (r_1 r_2)}$ if $U \in \mathbb{R}^{p_1 \times r_1}$ and $W \in \mathbb{R}^{p_2 \times r_2}$. The readers are referred to Kolda and Bader (2009) for a comprehensive survey on tensor algebra.

Let $\sigma_r(\cdot)$ be the rth largest singular value of a matrix. If $\mathcal{T}$ has Tucker ranks $(r_1, \cdots, r_m)$, the signal strength of $\mathcal{T}$ is defined by

$$\lambda_{\min} := \lambda_{\min}(\mathcal{T}) = \min \left\{ \sigma_{r_1}\big(\mathcal{M}_1(\mathcal{T})\big), \sigma_{r_2}\big(\mathcal{M}_2(\mathcal{T})\big), \cdots, \sigma_{r_m}\big(\mathcal{M}_m(\mathcal{T})\big) \right\},$$

i.e., the smallest positive singular value of all matricizations. Similarly, define $\lambda_{\max} := \lambda_{\max}(\mathcal{T}) = \max_j \sigma_1\big(\mathcal{M}_j(\mathcal{T})\big)$. The condition number of $\mathcal{T}$ is defined by $\kappa(\mathcal{T}) := \lambda_{\max}(\mathcal{T}) \lambda_{\min}^{-1}(\mathcal{T})$. We let Λ_j be the $r_j \times r_j$ diagonal matrix containing the singular values of $\mathcal{M}_j(\mathcal{G})$ (or equivalently the singular values of $\mathcal{M}_j(\mathcal{T})$). Note that Λ_js are not necessarily equal for different j, although $\|\Lambda_1\|_F = \cdots = \|\Lambda_m\|_F = \|\mathcal{T}\|_F$.

We define the principle angles between $U, \widehat{U} \in \mathbb{O}_{p,r}$ as an r-by-r diagonal matrix: $\Theta(U, \widehat{U}) = \operatorname{diag}(\arccos(\sigma_1), \ldots, \arccos(\sigma_r))$, where $\sigma_1 \geqslant \cdots \geqslant \sigma_r \geqslant 0$ are the singular values of $U^\top \widehat{U}$. Then the $\sin\Theta$ distances between $\hat{U}$ and U are defined as

$$\|\sin\Theta(U, \widehat{U})\| = \|\operatorname{diag}\left(\sin(\arccos(\sigma_1)), \ldots, \sin(\arccos(\sigma_r))\right)\| = \sqrt{1 - \sigma_r^2},$$

$$\|\sin\Theta(U, \widehat{U})\|_F = \left(\sum_{i=1}^r \sin^2(\arccos(\sigma_i)) \right)^{1/2} = \left(r - \sum_{i=1}^r \sigma_i^2 \right)^{1/2}.$$

3.3 Inference for Principal Components of Tucker Low-rank Tensor

For notational simplicity, we focus on the inference for third-order tensors, i.e., $m = 3$, while the results for general mth order tensor essentially follows and will be briefly discussed in Section 3.7.

3.3.1 Estimating Procedure

An accurate estimation is often the starting point for statistical inference and uncertainty quantification. In this section, we briefly discuss the estimation procedure for both tensor regression and PCA models. First, we summarize both models as follows:

$$Y_i = \langle \mathcal{X}_i, \mathcal{T} \rangle + \xi_i, \quad i = 1, \ldots, n.$$

Here, $\mathcal{X}_i$ can be the covariate in tensor regression; $n = p_1p_2p_3$, $Y_i = \mathcal{A}(j_1, j_2, j_3)$, and $\mathcal{X}_i = (e_{j_1}, e_{j_2}, e_{j_3}) \cdot 1$ with $i = (j_1 - 1)p_2p_3 + (j_2 - 1)p_3 + j_3, j_1 \in [p_1], j_2 \in [p_2], j_3 \in [p_3]$ in tensor PCA. Let $l_n(\mathcal{T}) = \sum_{i=1}^{n}(Y_i - \langle \mathcal{X}_i, \mathcal{T} \rangle)^2$ be the loss function in both settings. Then a straightforward solution to both problems is via the following Tucker rank constrained least squares estimator:

$$\begin{aligned} &\min_{\operatorname{rank}(\mathcal{T}) \leqslant (r_1, r_2, r_3)} \ell_n(\mathcal{T}) := \frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \langle \mathcal{X}_i, \mathcal{T} \rangle\big)^2, \\ \text{or equivalently} \quad &(\hat{\mathcal{G}}, \hat{U}_1, \hat{U}_2, \hat{U}_3) := \operatorname*{arg\,min}_{\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3},\ U_j \in \mathbb{O}_{p_j, r_j}} \ell_n\big((U_1, U_2, U_3) \cdot \mathcal{G}\big). \end{aligned} \tag{3.7}$$

Since the objective function (3.7) is highly non-convex, an efficient algorithm with provable guarantees is crucial for both tensor PCA and regression. As discussed earlier, various computationally feasible procedures have been proposed in the literature. For tensor regression, Han et al. (2020) recently introduced a simultaneous gradient descent algorithm and proved their proposed procedure achieves the minimax optimal estimation error; for tensor PCA, a simpler and more direct approach, higher-order orthogonal iteration (HOOI), was introduced by De Lathauwer et al. (2000b). The implementation details of both algorithms are provided in Section 5.2.1 in the supplementary materials.

Moreover, the primary interest of this chapter is on the statistical inference for $\mathcal{T}$ or U_j, far beyond deriving estimators achieving optimal estimation error. In general, even estimators achieving minimax optimal estimation error rate may not enjoy a proper asymptotic distribution. For example, the true parameter $\mathcal{T}$ or U_j plus a small enough perturbation can achieve optimal estimation error but does not satisfy any tractable distribution.

To this end, we introduce a *two-iteration alternating minimization* algorithm for both Tucker low-rank tensor PCA and tensor regression in Algorithms 1 and 2, respectively. Our theory in later this section reveals a surprising fact: if any estimator $\tilde{\mathcal{T}} = (\hat{U}_1^{(0)}, \hat{U}_2^{(0)}, \hat{U}_3^{(0)}) \cdot \hat{\mathcal{G}}^{(0)}$ achieving some attainable estimation error is provided as the input, the two-iteration alternating minimization in Algorithms 1 and 2 will provide an estimator enjoying asymptotic normality and being ready to use for confidence region construction.

Algorithm 1 Power Iteration for Tensor PCA

```
Input: $\ell_n(\cdot)$: Objective function (3.7); Initializations $(\hat{U}_1^{(0)}, \hat{U}_2^{(0)}, \hat{U}_3^{(0)})$;
for $t = 0, 1$ do
    $\hat{U}_1^{(t+1)}$ = leading $r_1$ left singular vectors of $\mathcal{M}_1(\mathcal{A} \times_2 \hat{U}_2^{(t)\top} \times_3 \hat{U}_3^{(t)\top})$;
    $\hat{U}_2^{(t+1)}$ = leading $r_2$ left singular vectors of $\mathcal{M}_2(\mathcal{A} \times_1 \hat{U}_1^{(t)\top} \times_3 \hat{U}_3^{(t)\top})$;
    $\hat{U}_3^{(t+1)}$ = leading $r_3$ left singular vectors of $\mathcal{M}_3(\mathcal{A} \times_1 \hat{U}_1^{(t)\top} \times_2 \hat{U}_2^{(t)\top})$;
end for
Output: Test statistic $\hat{U}_1 := \hat{U}_1^{(2)}, \hat{U}_2 := \hat{U}_2^{(2)}, \hat{U}_3 := \hat{U}_3^{(2)}$, and $\hat{\mathcal{G}} = (\hat{U}_1^{(2)\top}, \hat{U}_2^{(2)\top}, \hat{U}_3^{(2)\top}) \cdot \mathcal{A}$.
```

Remark 3.3.1 (Interpretation of Alternating Minimization Update in Tensor PCA)**.** *A key observation by (De Lathauwer et al., 2000b, Theorems 4.1, 4.2) shows minimizing* $\min_{\mathrm{rank}(\mathcal{T})\leqslant(r_1,r_2,r_3)} \|\mathcal{T} - \mathcal{A}\|_F^2$ *is equivalent to maximizing* $\max_{U_j \in \mathbb{O}_{p_j,r_j}} \|(U_1^\top, U_2^\top, U_3^\top) \cdot$

Algorithm 2 Alternating Minimization for Tensor Regression

```
Input: $\ell_n(\cdot)$: Objective function (3.7); Initializations $(\hat{U}_1^{(0)}, \hat{U}_2^{(0)}, \hat{U}_3^{(0)})$, and $\hat{\mathcal{G}}^{(0)}$ is the solution of $\arg\min_{\mathcal{G}} \ell_n\big((\hat{U}_1^{(0)}, \hat{U}_2^{(0)}, \hat{U}_3^{(0)}) \cdot \mathcal{G}\big)$ for tensor regression model;
for $t = 0, 1$ do
    Solve $\nabla_{U_1} \ell_n\big((\hat{U}_1^{(t+0.5)}, \hat{U}_2^{(t)}, \hat{U}_3^{(t)}) \cdot \hat{\mathcal{G}}^{(t)}\big) = 0$ to obtain $\hat{U}_1^{(t+0.5)}$;
    Update by $\hat{U}_1^{(t+1)} = \mathrm{SVD}_{r_1}\big(\hat{U}_1^{(t+0.5)}\big)$;
    Solve $\nabla_{U_2} \ell_n\big((\hat{U}_1^{(t)}, \hat{U}_2^{(t+0.5)}, \hat{U}_3^{(t)}) \cdot \hat{\mathcal{G}}^{(t)}\big) = 0$ to obtain $\hat{U}_2^{(t+0.5)}$;
    Update by $\hat{U}_2^{(t+1)} = \mathrm{SVD}_{r_2}\big(\hat{U}_2^{(t+0.5)}\big)$;
    Solve $\nabla_{U_3} \ell_n\big((\hat{U}_1^{(t)}, \hat{U}_2^{(t)}, \hat{U}_3^{(t+0.5)}) \cdot \hat{\mathcal{G}}^{(t)}\big) = 0$ to obtain $\hat{U}_3^{(t+0.5)}$;
    Update by $\hat{U}_3^{(t+1)} = \mathrm{SVD}_{r_3}\big(\hat{U}_3^{(t+0.5)}\big)$;
    Solve $\nabla_{\mathcal{G}} \ell_n\big((\hat{U}_1^{(t+1)}, \hat{U}_2^{(t+1)}, \hat{U}_3^{(t+1)}) \cdot \hat{\mathcal{G}}^{(t+1)}\big) = 0$ to obtain $\hat{\mathcal{G}}^{(t+1)}$;
end for
Output: Test statistic $\hat{U}_1 := \hat{U}_1^{(2)}, \hat{U}_2 := \hat{U}_2^{(2)}, \hat{U}_3 := \hat{U}_3^{(2)}$, and $\hat{\mathcal{G}} := \hat{\mathcal{G}}^{(2)}$.
```

$\mathcal{A}\|_F^2$. *Therefore, the optimization in tensor PCA is equivalent to*

$$\begin{aligned}(\hat{U}_1, \hat{U}_2, \hat{U}_3) :=& \mathop{\arg\min}_{U_j \in \mathbb{O}_{p_j, r_j}} \ell_n((U_1, U_2, U_3) \cdot \mathcal{G}) := \mathop{\arg\max}_{U_j \in \mathbb{O}_{p_j, r_j}} \|(U_1^\top, U_2^\top, U_3^\top) \cdot \mathcal{A}\|_F^2 \\ =& \mathop{\arg\max}_{U_j \in \mathbb{O}_{p_j, r_j}} \|U_j \mathcal{M}_j(\mathcal{A} \times_{j+1} U_{j+1} \times_{j+2} U_{j+2})\|_F^2 .\end{aligned}$$

Here, for convenience of notation, $U_4 = U_1, U_5 = U_2, r_4 = r_1, r_5 = r_2$. *Note that, given fixed* $\hat{U}_{j+1}^{(t)}$ *and* $\hat{U}_{j+2}^{(t)}$, *Eckart-Young-Mirsky Theorem (Eckart and Young, 1936) implies the optimal solution to* $\max_{U_j \in \mathbb{O}_{p_j, r_j}} \|(U_j^\top, \hat{U}_{j+1}^{(t)\top}, \hat{U}_{j+2}^{(t)\top}) \cdot \mathcal{A}\|_F^2$ *is attainable via singular value decomposition:*

$$\hat{U}_j^{(t+1)} = \textit{leading } r_j \textit{ left singular vectors of } \mathcal{M}_j\left(\mathcal{A} \times_{j+1} \hat{U}_{j+1}^{(t)\top} \times_{j+2} \hat{U}_{j+2}^{(t)\top}\right).$$

This explains the alternating minimization update steps for tensor PCA in Algorithm 1.

Hereinafter, we denote $\hat{U}_j$ the output of Algorithms 1 and 2, $p = \max\{p_1, p_2, p_3\}$ and $r_{\max} = \max\{r_1, r_2, r_3\}$. Next, we establish the asymptotic distribution and develop the inference procedure for $\|\sin\Theta(\hat{U}_j, U_j)\|_F^2$ in tensor PCA and tensor regression models when $\mathcal{T}$ admits the Tucker decomposition (3.6).

3.3.2 Inference for Tucker Low-rank Tensor PCA

We assume the following condition on initialization $(\hat{U}_1^{(0)}, \hat{U}_2^{(0)}, \hat{U}_3^{(0)})$ of Algorithm 1 holds.

Assumption 3.3.1. *Under tensor PCA model (3.1) with* $\mathcal{Z}_{i_1,i_2,i_3} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$*, there is an event* $\mathcal{E}_0$ *with* $\mathbb{P}(\mathcal{E}_0) \geqslant 1 - C_1 e^{-c_1 p}$ *for some absolute constants* $c_1, C_1 > 0$ *so that, under* $\mathcal{E}_0$*, the initialization* $(\hat{U}_1^{(0)}, \hat{U}_2^{(0)}, \hat{U}_3^{(0)})$ *satisfy* $\max_{j=1,2,3} \|\sin\Theta(\hat{U}_j^{(0)}, U_j)\| \leqslant C_2\sqrt{p}\sigma/\lambda_{\min}$ *for some absolute constant* $C_2 > 0$*.*

The claimed error rates in Assumption 3.3.1 are attainable by the algorithm HOOI under the SNR condition $\lambda_{\min}/\sigma \geqslant Cp^{3/4}$ (Zhang and Xia, 2018, Theorem 1). Such the SNR condition is essential to ensure a consistent estimator is achievable in polynomial time as illustrated by the literature reviewed in Section 3.1.2. Note that (Zhang and Xia, 2018, Theorem 1) presented an expectation error bound $\mathbb{E}\|\sin\Theta(\hat{U}_j^{(0)}, U_j)\|$, while its proof indeed involved a desired probabilistic bound as claimed by Assumption 3.3.1. If a given initialization estimation error upper bound is in a metric other than the $\sin\Theta$ distance described in Assumption 3.3.1, we may apply Lemma 5.2.4 in the supplementary materials to “translate" the upper bound in another metric to the desired $\sin\Theta$ distance.

Suppose $\hat{U}_j$ is the output of Algorithm 1. Built on Assumption 3.3.1, we characterize the distribution of $\|\sin\Theta(\hat{U}_j, U_j)\|_F^2$ by the following theorem.

Theorem 3.3.1 (Asymptotic normality of principal components in tensor PCA). *Suppose Assumption 3.3.1 holds for tensor PCA model (3.1),* $\mathcal{Z}(i_1, i_2, i_3) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, $p_j \asymp p$ *for* $j = 1, 2, 3$*, and* $\kappa(\mathcal{T}) \leqslant \kappa_0$*. Let* $\hat{U}_j$*s be the output of Algorithm 1 for tensor PCA model. There exist absolute constants* $c_1, C_0, C_1, C_2, C_3 > 0$ *such that if* $\lambda_{\min}/\sigma \geqslant$

$C_0(p^{3/4} + \kappa_0^2 p^{1/2})$, *then*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\|\sin\Theta(\hat{U}_j, U_j)\|_F^2 - p_j\sigma^2\|\Lambda_j^{-1}\|_F^2}{\sqrt{2p_j}\sigma^2\|\Lambda_j^{-2}\|_F} \leqslant x \right) - \Phi(x) \right|$$
$$\leqslant C_1 e^{-c_1 p} + C_2\left(\frac{\kappa_0^6 (p r_{\max})^{3/2}}{(\lambda_{\min}/\sigma)^2} + \frac{\kappa_0^2 (p\log p)^{1/2}}{\lambda_{\min}/\sigma} \right) + C_3 \frac{r_{\max}^{3/2}}{\sqrt{p}},$$

where $\Lambda_j = \mathrm{diag}(\lambda_1^{(j)}, \ldots, \lambda_{r_j}^{(j)})$ *is the diagonal matrix containing the singular values of* $\mathcal{M}_j(\mathcal{G})$*, and* $\Phi(x)$ *is the cumulative distribution function of* $N(0,1)$.

If the condition number $\kappa_0 = O(1)$, $(p r_{\max})^{3/4}(\lambda_{\min}/\sigma)^{-1} \to 0$ and $r_{\max}^3/p \to 0$ as $p \to \infty$, Theorem 3.3.1 yields

$$\frac{\|\sin\Theta(\hat{U}_j, U_j)\|_F^2 - p_j\sigma^2\|\Lambda_j^{-1}\|_F^2}{\sqrt{2p_j}\sigma^2\|\Lambda_j^{-2}\|_F} \xrightarrow{d.} N(0,1) \quad \text{as} \quad p \to \infty.$$

By the proof of Theorem 3.3.1, we can further establish the following joint distribution of all U_js:

$$\begin{pmatrix} \frac{\|\sin\Theta(\hat{U}_1, U_1)\|_F^2 - p_1\sigma^2\|\Lambda_1^{-1}\|_F^2}{\sqrt{2p_1}\sigma^2\|\Lambda_1^{-2}\|_F} \\ \frac{\|\sin\Theta(\hat{U}_2, U_2)\|_F^2 - p_2\sigma^2\|\Lambda_2^{-1}\|_F^2}{\sqrt{2p_2}\sigma^2\|\Lambda_2^{-2}\|_F} \\ \frac{\|\sin\Theta(\hat{U}_3, U_3)\|_F^2 - p_3\sigma^2\|\Lambda_3^{-1}\|_F^2}{\sqrt{2p_3}\sigma^2\|\Lambda_3^{-2}\|_F} \end{pmatrix} \xrightarrow{d.} N(0, I_3) \quad \text{as} \quad p \to \infty.$$

Remark 3.3.2. *We briefly compare Theorem 3.3.1 with the existing results in the literature. The asymptotic normality of* $\|\sin\Theta(\hat{U}_j, U_j)\|_F^2$ *in Theorem 3.3.1 requires SNR condition* $\lambda_{\min} \gg (r_{\max} p)^{3/4}$*, which is slightly stronger than the optimal SNR condition* $\lambda_{\min} \geqslant C_0 p^{3/4}$ *for achieving the consistent estimation in (Zhang and Xia, 2018, Theorem 1) (if* $r \geqslant 1$*), matches the condition in (Zheng and Tomioka, 2015, Theorem 1) (if* $r = 1$*), and weaker than the condition in (Richard and Montanari, 2014, Theorem 4) (if* $r = 1$*). Second, note that Theorem 3.3.1 implies* $\mathbb{E}\|\sin\Theta(\hat{U}_j, U_j)\|_F^2 = \big(1 + o(1)\big)p_j\sigma^2\|\Lambda_j^{-1}\|_F^2$*. To the best of our knowledge, this is the first result with a precise constant characterization of the estimation error in tensor PCA.*

While Theorem 3.3.1 characterizes the asymptotic distribution of $\|\sin\Theta(\hat{U}_j, U_j)\|_F^2$ for tensor PCA model, the result is not immediately applicable to uncertainty quantification of $\hat{U}_j$ since $\|\Lambda_j^{-1}\|_F^2$, $\|\Lambda_j^{-2}\|_F$, and σ^2 are often unknown in practice. We thus propose an estimate for Λ_j, σ:

$$\begin{aligned}&\hat{\Lambda}_j \text{ is the diagonal matrix containing the top } r_j \text{ singular values of } \mathcal{M}_j\left(\mathcal{A} \times_{j+1} \hat{U}_{j+1}^\top \times_{j+2} \hat{U}_{j+2}^\top\right),\\ &\hat{\sigma} = \left\|\mathcal{A} - \mathcal{A} \times_1 \hat{U}_1\hat{U}_1^\top \times_2 \hat{U}_2\hat{U}_2^\top \times_3 \hat{U}_3\hat{U}_3^\top\right\|_F / \sqrt{p_1p_2p_3}.\end{aligned} \tag{3.8}$$

We can prove a deviation bound for $\hat{\sigma}$ and the normal approximation for $\|\sin\Theta(\hat{U}_j, U_j)\|_F^2$ with the proposed plug-in estimators.

Lemma 3.3.1. *Under conditions of Theorem 3.3.1, there exist two constants* $C_1, C_2 > 0$ *such that*

$$\mathbb{P}\left\{|\hat{\sigma}^2/\sigma^2 - 1| \leqslant C_2(\kappa_0\sqrt{r_{\max}}p^{-1} + p^{-3/4}\sqrt{\log(p)})\right\} \geqslant 1 - C_1p^{-3}.$$

Theorem 3.3.2 (Inference for Tucker Low-rank Tensor PCA). *Suppose the conditions in Theorem 3.3.1 hold. Let* $\hat{\Lambda}_1 \in \mathbb{R}^{r_1\times r_1}$ *and* $\hat{\sigma}$ *be defined as* (3.8). *There exist absolute constants* $c_1, C_0, C_1, C_2, C_3 > 0$ *such that if* $\lambda_{\min}/\sigma \geqslant C_0(p^{3/4} + \kappa_0^2 p^{1/2})$, *then for* $j = 1, 2, 3$,

$$\begin{aligned}&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\|\sin\Theta(\hat{U}_j, U_j)\|_F^2 - p_j\hat{\sigma}^2\|\hat{\Lambda}_j^{-1}\|_F^2}{\sqrt{2p_j}\hat{\sigma}^2\|\hat{\Lambda}_j^{-2}\|_F} \leqslant x\right) - \Phi(x)\right|\\ &\leqslant C_1e^{-c_1p} + C_2\left(\frac{r_{\max}^{3/2}\kappa_0^6p^{3/2}}{(\lambda_{\min}/\sigma)^2} + \frac{\kappa_0^3\sqrt{pr_{\max}(r_{\max}^2 + \log p)}}{\lambda_{\min}/\sigma} + \frac{\sqrt{\log(p)}}{p^{1/4}} + \frac{\kappa_0\sqrt{r_{\max}}}{\sqrt{p}}\right) + C_3\frac{r_{\max}^{3/2}}{\sqrt{p}}.\end{aligned}$$

When the condition number $\kappa_0 = O(1)$, $(pr_{\max})^{3/4}(\lambda_{\min}/\sigma)^{-1} \to 0$, and $r_{\max}^3/p \to 0$ as $p \to \infty$, Theorem 3.3.2 implies

$$\frac{\|\sin\Theta(\hat{U}_j, U_j)\|_F^2 - p_j\hat{\sigma}^2\|\hat{\Lambda}_j^{-1}\|_F^2}{\sqrt{2p_j}\hat{\sigma}^2\|\hat{\Lambda}_j^{-2}\|_F} \xrightarrow{d.} N(0,1) \quad \text{as} \quad p \to \infty. \tag{3.9}$$

Equation (3.9) is readily applicable to statistical inference for U_j. After getting $\hat{U}_j$ by Algorithm 1, we propose a $(1-\alpha)$-level confidence region for U_j as

$$\mathrm{CR}_\alpha(\hat{U}_j) := \Big\{ V \in \mathbb{O}_{p_j,r_j} : \|\sin\Theta(\hat{U}_j, V)\|_F^2 \leqslant p_j\hat{\sigma}^2\|\hat{\Lambda}_j^{-1}\|_F^2 + z_\alpha\sqrt{2p_j}\hat{\sigma}^2\|\hat{\Lambda}_j^{-2}\|_F \Big\}, \tag{3.10}$$

where $z_\alpha = \Phi^{-1}(1-\alpha)$ is the $(1-\alpha)$ quantile of the standard normal distribution. The following corollary is an immediate result of Theorem 3.3.2, which confirms that the confidence region $\mathrm{CR}_\alpha(\hat{U}_1)$ is indeed asymptotically accurate.

Corollary 3.3.1 (Confidence region for tensor PCA)**.** *Suppose the conditions of Theorem 3.3.2 hold and the confidence region* $\mathrm{CR}_\alpha(\hat{U}_j)$ *is defined in (3.10). If* $\kappa_0^6(r_{\max}^{3/2}p^{3/2} + r_{\max}p\log p)(\lambda_{\min}/\sigma)^{-2} \to 0$ *and* $r_{\max}^3/p \to 0$ *as* $p \to \infty$*, then*

$$\lim_{p\to\infty} \mathbb{P}\big(U_j \in \mathrm{CR}_\alpha(\hat{U}_j)\big) = 1-\alpha.$$

3.3.3 Inference for Tucker Low-rank Tensor Regression

This section is devoted to the asymptotic distribution and inference in low-rank tensor regression. We first introduce the following assumption on the initialization for Algorithm 2.

Assumption 3.3.2. *Under tensor regression model (3.2) with* $\mathcal{X}(i_1,i_2,i_3) \overset{\text{i.i.d.}}{\sim} N(0,1)$, $\mathrm{Var}(\xi_i) = \sigma^2$ *and* $\|\xi_i\|_{\psi_2} \leqslant C\sigma$ *for some constant* $C > 0$*, there is an event* $\mathcal{E}_0$ *with* $\mathbb{P}(\mathcal{E}_0) \geqslant 1 - C_1e^{-c_1p}$ *for some absolute constants* $c_1, C_1 > 0$ *so that, under* $\mathcal{E}_0$*, the initialization* $\tilde{\mathcal{T}} = (\hat{U}_1^{(0)}, \hat{U}_2^{(0)}, \hat{U}_3^{(0)})\cdot\hat{\mathcal{G}}^{(0)}$ *satisfy* $\|\tilde{\mathcal{T}} - \mathcal{T}\|_F^2 \leqslant C_2pr_{\max}\sigma^2/n$ *or* $\max_j \|\sin\Theta(\hat{U}_j^{(0)}, U_j)\| \leqslant C_2\sqrt{p/n}\sigma/\lambda_{\min}$ *for some absolute constant* $C_2 > 0$.

The claimed bound of $\|\tilde{\mathcal{T}} - \mathcal{T}\|_F^2$ in Assumption 3.3.2 is attainable, for instance, by the gradient descent algorithm developed in Han et al. (2020) and the importance sketching algorithm developed in Zhang et al. (2020a) under the SNR condition

$n(\lambda_{\min}/\sigma)^2 \geqslant Cp^{3/2}$ and the sample size condition $n \geqslant Cp^{3/2}r_{\max}$. The theoretical guarantees for this claim can be found in (Han et al., 2020, Theorem 4.2) and (Zhang et al., 2020a, Theorem 4).

Based on Assumption 3.3.2, we establish the following asymptotic results for tensor regression.

Theorem 3.3.3. *Suppose Assumption 3.3.2 holds for tensor regression model (3.2),* $\mathcal{X}(i_1, i_2, i_3) \overset{\text{i.i.d.}}{\sim} N(0,1)$*,* $\mathrm{Var}(\xi_i) = \sigma^2$*, and* $\|\xi_i\|_{\psi_2} \leqslant C\sigma$ *for some constant* $C > 0$*,* $p_j \asymp p$ *for* $j = 1, 2, 3$*, and* $\kappa(\mathcal{T}) \leqslant \kappa_0$*. Let* $\hat{U}_j$*s be the output of two-iteration alternating minimization (Algorithm 2). There exist absolute constants* $c_1, C_0, C_1, C_2, C_3, C_4 > 0$ *such that if* $n(\lambda_{\min}/\sigma)^2 \geqslant C_0(p^{3/2} \vee \kappa_0^4 p r_{\max}^2)$ *and* $n \geqslant C_2(p^{3/2} \vee \kappa_0^2 p r_{\max}^3)$*, then*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\|\sin\Theta(\hat{U}_j, U_j)\|_F^2 - p_j n^{-1}\sigma^2\|\Lambda_j^{-1}\|_F^2}{\sqrt{2p_j} n^{-1}\sigma^2 \|\Lambda_j^{-2}\|_F} \leqslant x \right) - \Phi(x) \right|$$
$$\leqslant C_3\left(\frac{\kappa_0^4 r_{\max}^{5/2} p^{3/2}}{n} + \kappa_0^3 \left(\frac{r_{\max}^5 p \log^2 n}{n} \right)^{1/2} + \frac{p^{3/2}}{n} \left(\frac{\kappa_0^5 r_{\max}^2}{\lambda_{\min}/\sigma} + \frac{\kappa_0^5 r_{\max}^{3/2}}{(\lambda_{\min}/\sigma)^2} \right) + \kappa_0^4 \left(\frac{p r_{\max}^3 + r_{\max} p \log p}{n(\lambda_{\min}/\sigma)^2} \right)^{1/2} \right)$$
$$+ C_1 e^{-c_1 p} + C_4 \frac{r_{\max}^{3/2}}{\sqrt{p}},$$

where Λ_j *is the* $r_j \times r_j$ *diagonal matrix containing the singular values of* $\mathcal{M}_j(\mathcal{T})$*.*

If the condition number $\kappa_0 = O(1)$, $(r_{\max}^{5/2}p^{3/2} + r_{\max}^5 p \log^2 n)/n \to 0$, $r_{\max}^{3/2}p^{3/2}/(n(\lambda_{\min}/\sigma)^2) \to 0$ and $r_{\max}^3/p \to 0$ as $p \to \infty$, Theorem 3.3.3 implies

$$\frac{\|\sin\Theta(\hat{U}_j, U_j)\|_F^2 - p_j n^{-1}\sigma^2\|\Lambda_j^{-1}\|_F^2}{\sqrt{2p_j} n^{-1}\sigma^2 \|\Lambda_j^{-2}\|_F} \xrightarrow{\text{d.}} N(0,1) \quad \text{as} \quad p \to \infty.$$

To make inference for tensor regression, we develop the following asymptotic normal distribution for $\|\sin\Theta(\hat{U}_j, U_j)\|_F$ with the plug-in estimates of Λ_j.

Theorem 3.3.4 (Tensor regression)**.** *Suppose the conditions in Theorem 3.3.3 hold. Let* $\hat{\Lambda}_j = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{r_j})$ *be a diagonal matrix containing the singular values of* $\mathcal{M}_1(\hat{\mathcal{G}})$*, where*

$\hat{\mathcal{G}}$ is the output of Algorithm 2. There exist absolute constants $c_1, C_0, C_1, C_2, C_3, C_4 > 0$ such that if $n(\lambda_{\min}/\sigma)^2 \geqslant C_0(p^{3/2} \vee \kappa_0^6 p r_{\max}^2)$ and $n \geqslant C_2(p^{3/2} \vee \kappa_0^8 p r_{\max}^3)$, then for $j = 1, 2, 3$,

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\|\sin\Theta(\hat{U}_j, U_j)\|_F^2 - p_j n^{-1}\sigma^2 \|\hat{\Lambda}_j^{-1}\|_F^2}{\sqrt{2p_j}\, n^{-1}\sigma^2 \|\hat{\Lambda}_j^{-2}\|_F} \leqslant x \right) - \Phi(x) \right|$$
$$\leqslant C_3\bigg(\frac{\kappa_0^4 r_{\max}^{5/2} p^{3/2}}{n} + \kappa_0^3\Big(\frac{r_{\max}^5 p \log^2 n}{n}\Big)^{1/2} + \frac{p^{3/2}}{n}\Big(\frac{\kappa_0^5 r_{\max}^2}{\lambda_{\min}/\sigma} + \frac{\kappa_0^5 r_{\max}^{3/2}}{(\lambda_{\min}/\sigma)^2}\Big) + \kappa_0^4\Big(\frac{p r_{\max}^3 + r_{\max} p \log p}{n(\lambda_{\min}/\sigma)^2}\Big)^{1/2}\bigg)$$
$$+ C_1 e^{-c_1 p} + C_4 \frac{r_{\max}^{3/2}}{\sqrt{p}}.$$

We propose the following $(1-\alpha)$-level confidence region for U_j:

$$\widetilde{\mathrm{CR}}_\alpha(\hat{U}_j) := \left\{ V \in \mathbb{O}_{p_j, r_j} : \|\sin\Theta(\hat{U}_j, V)\|_F^2 \leqslant \frac{p_j \sigma^2 \|\hat{\Lambda}_j^{-1}\|_F^2}{n} + z_\alpha \frac{\sqrt{2p_j}\,\sigma^2 \|\hat{\Lambda}_j^{-2}\|_F}{n} \right\}. \tag{3.11}$$

The following corollary establishes the coverage probability of the proposed confidence region.

Corollary 3.3.2 (Confidence region for tensor regression)**.** *Suppose the conditions of Theorem 3.3.4 hold and the confidence region $\widetilde{\mathrm{CR}}_\alpha(\hat{U}_j)$ is defined by (3.11). If $(\kappa_0^5 r_{\max}^{5/2} p^{3/2} + \kappa_0^6 r_{\max}^5 p \log^2 n)/n \to 0$, $\kappa_0^5 r_{\max}^{3/2} p^{3/2}/(n(\lambda_{\min}/\sigma)^2) \to 0$ and $r_{\max}^3/p \to 0$ as $p \to \infty$, then*

$$\lim_{p\to\infty} \mathbb{P}\big(U_j \in \widetilde{\mathrm{CR}}_\alpha(\hat{U}_j)\big) = 1 - \alpha.$$

Remark 3.3.3 (Selection of σ)**.** *When σ is unknown, we can estimate it by a sample splitting scheme as follows. First, we retain a part of sample $\{(\mathcal{X}_k, Y_k)\}_{k=1}^{\lceil p^{3/2}\rceil}$ and use the other samples to compute the estimator $\tilde{\mathcal{T}}$. Define*

$$\hat{\sigma}^2 := \sum\nolimits_{k=1}^{\lceil p^{3/2}\rceil} \big(Y_k - \langle \tilde{\mathcal{T}}, \mathcal{X}_k\rangle\big)^2 / \lceil p^{3/2}\rceil.$$

Under Assumption 3.3.2 and conditions of Theorem 3.3.3, we can show with probability at least $1 - p^{-3}$*,* $\left|\hat{\sigma}^2/\sigma^2 - 1\right| = O\big(p^{-3/4}\sqrt{\log p} + r_{\max} p n^{-1}\big)$*. By plugging in* $\hat{\sigma}$ *to* (3.11)*, we obtain a data-driven* $(1-\alpha)$ *asymptotic confidence region for* U_j*.*

3.3.4 Proof Sketch

In this section, we briefly explain the proof strategy for tensor PCA model, i.e., Theorem 3.3.1. The proof for tensor regression model is more complicated but shares similar spirits. Without loss of generality, we assume $\sigma = 1$. First,

$$2\|\sin\Theta(\hat{U}_1, U_1)\|_F^2 = \|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 = 2r_1 - 2\big\langle \hat{U}_1\hat{U}_1^\top, U_1U_1^\top\big\rangle = -2\big\langle U_1U_1^\top, \hat{U}_1\hat{U}_1^\top - U_1U_1^\top\big\rangle.$$

It thus suffices to investigate the distribution of $\big\langle U_1U_1^\top, \hat{U}_1\hat{U}_1^\top - U_1U_1^\top\big\rangle$. By Algorithm 1, $\hat{U}_1$ are the top-r_1 left singular vectors of $\mathcal{M}_1\big(\mathcal{A} \times_2 \hat{U}_2^{(1)\top} \times_3 \hat{U}_3^{(1)\top}\big)$. As a result, $\hat{U}_1\hat{U}_1^\top$ is the spectral projector and can be decomposed as

$$\mathcal{M}_1(\mathcal{A})\Big(\hat{U}_2^{(1)}\hat{U}_2^{(1)\top} \otimes \hat{U}_3^{(1)}\hat{U}_3^{(1)\top}\Big)\mathcal{M}_1^\top(\mathcal{A}) =: \mathcal{M}_1(\mathcal{T})\mathcal{M}_1^\top(\mathcal{T}) + D_1^{(1)}.$$

The high-level ideas of the proof include the following steps.

Step 1: We apply the spectral representation formula (Xia (2019b); also see the statement in Lemma 5.2.2 from the supplementary materials) and expand

$$\hat{U}_1\hat{U}_1^\top = U_1U_1^\top + S_1(D_1^{(1)}) + S_2(D_1^{(1)}) + S_3(D_1^{(1)}) + \sum_{k \geqslant 4} S_k(D_1^{(1)}),$$

where $S_k(\cdot)$ denotes the kth order perturbation term:

$$S_k(D_1^{(1)}) = \sum_{s_1+\cdots+s_{k+1}=k} (-1)^{1+\tau(\mathbf{s})} \cdot B_1^{-s_1}D_1^{(1)}B_1^{-s_2}D_1^{(1)}B_1^{-s_3}\cdots B_1^{-s_k}D_1^{(1)}B_1^{-s_{k+1}},$$

where $B_1^{-k} = U_1\Lambda_1^{-2k}U_1^\top$ for each positive integer k, $B_1^0 := I_{p_1} - U_1U_1^\top$, $s_1, \cdots, s_{k+1}$ are non-negative integers, and $\tau(\mathbf{s}) = \sum_{j=1}^{k+1}\mathbb{I}(s_j > 0)$.

Step 2: Since $\big\langle U_1U_1^\top, S_1(D_1^{(1)})\big\rangle = 0$ and $\|S_k(D_1^{(1)})\| \leqslant (C_1\kappa_0^2\sqrt{p}/\lambda_{\min})^k$ with high probability, we can write

$$\left\langle \hat{U}_1\hat{U}_1^\top - U_1U_1^\top, U_1U_1^\top \right\rangle = \left\langle S_2(D_1^{(1)}), U_1U_1^\top \right\rangle + \left\langle S_3(D_1^{(1)}), U_1U_1^\top \right\rangle + O\Big(\frac{r_{\mathsf{max}}\kappa_0^8 p^2}{\lambda_{\mathsf{min}}^4}\Big).$$

In other words, the higher order terms ($k \geqslant 4$) can be bounded with high probability, which becomes small order terms.

Step 3: We show, with high probability, the third order term can be bounded by

$$\left|\left\langle S_3(D_1^{(1)}), U_1U_1^\top \right\rangle\right| = O\Big(\frac{\kappa_0^3 p\sqrt{r_{\mathsf{max}}\log p}}{\lambda_{\mathsf{min}}^3} + \frac{\kappa_0^3 p^2 r_{\mathsf{max}}^{3/2}}{\lambda_{\mathsf{min}}^4}\Big)$$

and becomes small order term. Now, it suffices to only investigate the second order term carefully.

Step 4: We decompose the second order term $\left\langle S_2(D_1^{(1)}), U_1U_1^\top \right\rangle$ into a leading term and remainder terms. Similarly to *Step 2* and *Step 3*, we show that the remainder terms are, with high probability, bounded by $O(\kappa_0^3 p\sqrt{r_{\mathsf{max}}\log p}\lambda_{\mathsf{min}}^{-3} + \kappa_0^3 p^2 r_{\mathsf{max}}^{3/2}\lambda_{\mathsf{min}}^{-4})$.

Step 5: We prove that the leading term of $\left\langle S_2(D_1^{(1)}), U_1U_1^\top \right\rangle$ can be written as a sum of independent random variables, which yields a normal approximation by Berry-Essen Theorem. Finally, combining all these steps, we get the normal approximation for $\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2$.

Among these steps, Steps 4 and 5 are the most technically involved. Throughout the proof, we apply the spectral representation formula at multiple stages to prove sharp upper bounds for higher-order terms, and establish central limit theorem for the second-order term.

The following lemmas are used in our proof and could be of independent interest. First, Lemma 3.3.2 is used to establish the concentration inequalities for the sum of random variables that have heavier tails than Gaussian.

Lemma 3.3.2 (Orlicz ψ_α-norm for product of random variables)**.** *Suppose* $X_1, \ldots, X_n$ *are* n *random variables (not necessarily independent) satisfying* $\|X_i\|_{\psi_{\alpha_i}} \leqslant K_i$*. Define*

$\bar{\alpha} = \left(\sum_{i=1}^n \alpha_i^{-1}\right)^{-1}$. *Then*

$$\left\| \prod_{i=1}^n X_i \right\|_{\psi_{\bar{\alpha}}} \leqslant \prod_{i=1}^n K_i.$$

Next, Lemma 3.3.3 provides a tight probabilistic upper bound for sum of third moments of Gaussian random matrices.

Lemma 3.3.3. *Suppose* $Z_1, \dots, Z_n \in \mathbb{R}^{p \times r}$ *are independent random matrices satisfying* $Z_i(j,k) \overset{\text{i.i.d.}}{\sim} N(0,1)$. *Then there exist two universal constants* $C, C_1 > 0$ *such that for fixed* $M_1, \dots, M_n \in \mathbb{R}^{p \times r}$,

$$\mathbb{P}\left(\left| \sum_{i=1}^n \|Z_i\|_F^2 \langle Z_i, M_i \rangle \right| \geqslant Cpr \left(\sum_{i=1}^n \|M_i\|_F^2 \right)^{1/2} \sqrt{\log(p)} \right) \leqslant p^{-C_1}.$$

3.4 PCA for Orthogonally Decomposable Tensors

In this section, we specifically focus on the tensor PCA model (3.1) with orthogonally decomposable signal tensor $\mathcal{T}$:

$$\mathcal{T} = \sum_{i=1}^r \lambda_i \cdot u_i \otimes v_i \otimes w_i, \tag{3.12}$$

where $U = (u_1, \cdots, u_r) \in \mathbb{O}_{p_1,r}$, $V = (v_1, \cdots, v_r) \in \mathbb{O}_{p_2,r}$, and $W = (w_1, \cdots, w_r) \in \mathbb{O}_{p_3,r}$ all have orthonormal columns; the singular values satisfy $\lambda_{\min} = \min\{\lambda_1, \dots, \lambda_r\} > 0$. Here, for any $u \in \mathbb{R}^{p_1}, v \in \mathbb{R}^{p_2}, w \in \mathbb{R}^{p_3}$, $u \otimes v \otimes w$ is a $p_1 \times p_2 \times p_3$ tensor whose (i,j,k)th entry is $u(i)v(j)w(k)$.

Our goal is to make inference on the principal components based on a noisy observation $\mathcal{A} = \mathcal{T} + \mathcal{Z}$. Different from the inference for Tucker low-rank tensor discussed in Section 3.3, where an accurate estimation is hopeful only for the joint column space of U_j due to the non-identifiability of Tucker decomposition, we can make inference for each individual vector $\{u_j, v_j, w_j\}$ if $\mathcal{T}$ is orthogonally decomposable as (3.12). Given some estimates $\{\hat{u}_j^{(0)}, \hat{v}_j^{(0)}, \hat{w}_j^{(0)}\}_{j=1}^r$, we propose to pass them to a post-processing step by two-iteration procedure in Algorithm 3 to obtain the test statistics $\{\hat{u}_j, \hat{v}_j, \hat{w}_j\}_{j=1}^r$.

Algorithm 3 Power Iterations for Orthogonally decomposable $\mathcal{T}$

```
Input: $\mathcal{A}$, initialization $\{\hat{u}_j^{(0)}, \hat{v}_j^{(0)}, \hat{w}_j^{(0)}\}_{j=1}^r$;
for $t = 0, 1$ do
    for $j = 1, 2, \cdots, r$ do
        Compute $\hat{u}_j^{(t+0.5)} = \mathcal{A} \times_2 \hat{v}_j^{(t)\top} \times_3 \hat{w}_j^{(t)\top}$; Update $\hat{u}_j^{(t+1)} = \hat{u}_j^{(t+0.5)} \|\hat{u}_j^{(t+0.5)}\|_2^{-1}$;
        Compute $\hat{v}_j^{(t+0.5)} = \mathcal{A} \times_1 \hat{u}_j^{(t)\top} \times_3 \hat{w}_j^{(t)\top}$; Update $\hat{v}_j^{(t+1)} = \hat{v}_j^{(t+0.5)} \|\hat{v}_j^{(t+0.5)}\|_2^{-1}$;
        Compute $\hat{w}_j^{(t+0.5)} = \mathcal{A} \times_1 \hat{u}_j^{(t)\top} \times_2 \hat{v}_j^{(t)\top}$; Update $\hat{w}_j^{(t+1)} = \hat{w}_j^{(t+0.5)} \|\hat{w}_j^{(t+0.5)}\|_2^{-1}$;
    end for
end for
Output: $\hat{u}_j = \hat{u}_j^{(2)}, \hat{v}_j = \hat{v}_j^{(2)}$ and $\hat{w}_j = \hat{w}_j^{(2)}$ for all $j = 1, \cdots, r$.
```

Since our primary interest is about the statistical inference for $\{u_j, v_j, w_j\}$, we assume that the initializations of Algorithm 3 satisfies the following Assumption 3.4.1. Such an assumption is achievable by the power iteration method with k-means initialization introduced in Anandkumar et al. (2014a) along with the theoretical guarantees developed in Liu et al. (2017) when $\lambda/\sigma \geqslant Cp^{3/4}$.

Assumption 3.4.1. *Under the tensor PCA model (3.1) with $\mathcal{T}$ being orthogonally decomposable as (3.12), there is an event $\mathcal{E}_0$ with $\mathbb{P}(\mathcal{E}_0) \geqslant 1 - C_1 e^{-c_1 p}$ for some absolute constants $c_1, C_1 > 0$ such that, under $\mathcal{E}_0$, the initializations $\{\hat{u}_j^{(0)}, \hat{v}_j^{(0)}, \hat{w}_j^{(0)}\}_j$ satisfy $\max\left\{\|\hat{u}_{\pi(j)}^{(0)} - u_j\|_2, \|\hat{v}_{\pi(j)}^{(0)} - v_j\|_2, \|\hat{w}_{\pi(j)}^{(0)} - w_j\|_2\right\} \leqslant C_2\sigma\sqrt{p}/\lambda_j$ for some permutation $\pi : [r] \to [r]$, all $1 \leqslant j \leqslant r$, and some absolute constant $C_2 > 0$.*

We establish the asymptotic normality for the outcome of Algorithm 3 as follows.

Theorem 3.4.1 (PCA for orthogonally decomposable tensors)**.** *Suppose Assumption 3.4.1 holds for tensor PCA model (3.1) with an orthogonally decomposable $\mathcal{T}$ as (3.12),*

$\mathcal{Z}(i_1, i_2, i_3) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, $p_j \asymp p$ *for* $j = 1, 2, 3$, *and* $\kappa(\mathcal{T}) \leqslant \kappa_0$. *Let* $\{\hat{u}_j, \hat{v}_j, \hat{w}_j\}_{j=1}^{r}$ *be the output of Algorithm 3. There exist absolute constants* $c_1, C_0, C_1, C_2, C_3 > 0$ *such that if* $\lambda_{\min}/\sigma \geqslant C_0(p^{3/4} + \kappa_0^2 p^{1/2})$*, then*

$$\begin{aligned}
&\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\langle \hat{u}_{\pi(j)}, u_j \rangle^2 - (1 - p_j \sigma^2 \lambda_j^{-2})}{\sqrt{2p_j}\, \sigma^2 \lambda_j^{-2}} \leqslant x \right) - \Phi(x) \right| \\
&\leqslant C_1 e^{-c_1 p} + C_2 \left(\frac{\kappa_0^6 \sigma^2 (pr)^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^2 \sigma (p \log p)^{1/2}}{\lambda_{\min}} \right) + C_3 \frac{r^{3/2}}{\sqrt{p}}
\end{aligned} \tag{3.13}$$

for all $j = 1, \cdots, r$. *Here,* $\pi(\cdot)$ *is the permutation introduced in Assumption 3.4.1. Moreover, let* $\hat{\lambda}_j = \|\mathcal{A} \times_2 \hat{v}_j^\top \times_3 \hat{w}_j^\top\|_2$. *Then, (3.13) also holds if* λ_j *is replaced by* $\hat{\lambda}_j$ *and* $\kappa_0^2 \sigma (p \log p)^{1/2} \lambda_{\min}^{-1}$ *is replaced by* $\kappa_0^3 \sigma \sqrt{pr(r^2 + \log p)} \lambda_{\min}^{-1}$. *Similar results also hold for* $\langle \hat{v}_{\pi(j)}, v_j \rangle^2$ *and* $\langle \hat{w}_{\pi(j)}, w_j \rangle^2$.

By Theorem 3.4.1, if $\lambda_{\min}/\sigma \gg \kappa_0^3 (pr)^{3/4} + \kappa_0^2 (p \log p)^{1/2}$ and $r \ll p^{1/3}$, then for each $j = 1, \cdots, r$,

$$\frac{\langle \hat{u}_{\pi(j)}, u_j \rangle^2 - (1 - p_j \sigma^2 \lambda_j^{-2})}{\sqrt{2p_j}\, \sigma^2 \lambda_j^{-2}} \xrightarrow{\text{d.}} N(0, 1) \quad \text{as} \quad p \to \infty.$$

Similarly to Section 3.3.2, we plug in data-driven estimates of λ_j and σ^2 and construct a $(1 - \alpha)$ confidence region for u_j as

$$\mathrm{CR}_\alpha(\hat{u}_{\pi(j)}) := \left\{ v \in \mathbb{R}^{p_j} : \|v\|_2 = 1 \text{ and } \langle \hat{u}_{\pi(j)}, v \rangle^2 \geqslant (1 - p_j \hat{\sigma}^2 \hat{\lambda}_{\pi(j)}^{-2}) - z_\alpha \sqrt{2p_j}\, \hat{\sigma}^2 \hat{\lambda}_{\pi(j)}^{-2} \right\}. \tag{3.14}$$

The confidence region for v_j, w_j can be constructed similarly.

3.5 Entry-wise Inference for Rank-1 Tensors

In this section, we consider the statistical inference for tensor PCA model with a rank-1 signal tensor:

$$\mathcal{A} = \mathcal{T} + \mathcal{Z}, \quad \mathcal{T} = \lambda \cdot u \otimes v \otimes w. \tag{3.15}$$

Here, $u \in \mathbb{S}^{p_1-1}, v \in \mathbb{S}^{p_2-1}, w \in \mathbb{S}^{p_3-1}$, the singular value $\lambda > 0$, and $\mathcal{Z} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. We specifically aim to study the inference for any linear form of u, v, w, i.e., $\langle q_1, u\rangle, \langle q_2, v\rangle$, and $\langle q_3, w\rangle$, with arbitrary deterministic unit vectors $\{q_1, q_2, q_3\}$. We also aim to study the inference for each entry $\mathcal{T}_{ijk}, i \in [p_1], j \in [p_2], k \in [p_3]$. To this end, we first apply the rank-1 power iteration in Algorithm 4 (Zhang and Golub, 2001; Richard and Montanari, 2014). Algorithm 4 can be roughly seen as a rank-1 special case of Algorithm 3 for the Tucker low-rank tensor PCA and Algorithm 9 for the orthogonally decomposable tensor PCA.

Algorithm 4 Power iterations for rank-1 tensor $\mathcal{T}$

```
    Input: $\mathcal{A}$
    Initialize $\hat{u}^{(0)} = \mathrm{SVD}_1(\mathcal{M}_1(\mathcal{A}))$, $\hat{v}^{(0)} = \mathrm{SVD}_1(\mathcal{M}_2(\mathcal{A}))$, $\hat{w}^{(0)} = \mathrm{SVD}_1(\mathcal{M}_3(\mathcal{A}))$, $t = 1$;
    while $t < t_{\max}$ do
        Compute $\hat{u}^{(t+0.5)} = \mathcal{A} \times_2 \hat{v}^{(t)\top} \times_3 \hat{w}^{(t)\top}$; Update $\hat{u}^{(t+1)} = \hat{u}^{(t+0.5)} \|\hat{u}^{(t+0.5)}\|_2^{-1}$;
        Compute $\hat{v}^{(t+0.5)} = \mathcal{A} \times_1 \hat{u}^{(t)\top} \times_3 \hat{w}^{(t)\top}$; Update $\hat{v}^{(t+1)} = \hat{v}^{(t+0.5)} \|\hat{v}^{(t+0.5)}\|_2^{-1}$;
        Compute $\hat{w}^{(t+0.5)} = \mathcal{A} \times_1 \hat{u}^{(t)\top} \times_2 \hat{v}^{(t)\top}$; Update $\hat{w}^{(t+1)} = \hat{w}^{(t+0.5)} \|\hat{w}^{(t+0.5)}\|_2^{-1}$;
        $t = t + 1$;
    end while
    Output: $\hat{u} = \hat{u}^{(t_{\max})}, \hat{v} = \hat{v}^{(t_{\max})}$ $\hat{w} = \hat{w}^{(t_{\max})}$, $\hat{\lambda}$ and $\hat{\mathcal{T}}$.
```

Next, we establish the asymptotic normality for the output of Algorithm 4, $\hat{u}, \hat{v}, \hat{w}$, under the essential SNR condition that ensures tensor PCA is solvable in polynomial time. Without loss of generality, we assume that the signs of $\hat{u}, \hat{v}, \hat{w}$ satisfy $\langle \hat{u}, u\rangle \geqslant 0, \langle \hat{v}, v\rangle \geqslant 0$ and $\langle \hat{w}, w\rangle \geqslant 0$ (otherwise one can flip the sign of $\hat{u}, \hat{v}, \hat{w}$ without changing the problem essentially). With a slight abuse of notation, let u_i, v_j, and w_k be the ith entry of u, the jth entry of v, and the kth entry of w, respectively.

Theorem 3.5.1. *Consider the tensor PCA model* (3.1) *with Gaussian noise* $\mathcal{Z}(i_1, i_2, i_3) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ *and rank*$(\mathcal{T}) = 1$, $p_j \asymp p$ *for* $j = 1, 2, 3$. *Let* $\hat{\lambda}, \hat{u}, \hat{v}, \hat{w}, \hat{\mathcal{T}}$ *be the outputs of Algorithm 4 with iteration number* $t_{\max} \geqslant C_1 \log(p)$ *for constant* $C_1 > 0$. *Suppose*

$\lambda/\sigma \gg p^{3/4}$. *For any deterministic array* $\{q_1^{(k)}, q_2^{(k)}, q_3^{(k)}\}_{k=1}^{\infty}$ *satisfying* $q_i^{(k)} \in \mathbb{S}^{k-1}$, *denote*

$$T_{q_1^{(p_1)},q_2^{(p_2)},q_3^{(p_3)}} = \left(\frac{\langle q_1^{(p_1)}, \hat{u} - u\rangle + \frac{p_1\langle q_1^{(p_1)}, u\rangle}{2(\lambda/\sigma)^2}}{\sqrt{\frac{p_1\langle q_1^{(p_1)},u\rangle^2}{2(\lambda/\sigma)^4} + \frac{1-\langle q_1^{(p_1)},u\rangle^2}{(\lambda/\sigma)^2}}}, \frac{\langle q_2^{(p_2)}, \hat{v} - v\rangle + \frac{p_2\langle q_2^{(p_2)}, v\rangle}{2(\lambda/\sigma)^2}}{\sqrt{\frac{p_2\langle q_2^{(p_2)},v\rangle^2}{2(\lambda/\sigma)^4} + \frac{1-\langle q_2^{(p_2)},v\rangle^2}{(\lambda/\sigma)^2}}}, \frac{\langle q_3^{(p_3)}, \hat{w} - w\rangle + \frac{p_3\langle q_3^{(p_3)}, w\rangle}{2(\lambda/\sigma)^2}}{\sqrt{\frac{p_3\langle q_3^{(p_3)},w\rangle^2}{2(\lambda/\sigma)^4} + \frac{1-\langle q_3^{(p_3)},w\rangle^2}{(\lambda/\sigma)^2}}} \right)^{\top}.$$

Then

$$T_{q_1^{(p_1)},q_2^{(p_2)},q_3^{(p_3)}} \xrightarrow{d.} N(0, I_3) \quad \text{as} \quad p \to \infty. \tag{3.16}$$

Specifically, if $|u_i|, |v_j|, |w_k| \ll \min\{\lambda/(\sigma p), 1\}$ *for some* $i \in [p_1], j \in [p_2], k \in [p_3]$, *then*

$$\left(\tfrac{\lambda}{\sigma}(\hat{u}_i - u_i),\ \tfrac{\lambda}{\sigma}(\hat{v}_j - v_j),\ \tfrac{\lambda}{\sigma}(\hat{w}_k - w_k)\right)^{\top} \xrightarrow{d.} N(0, I_3) \quad \text{as} \quad p \to \infty. \tag{3.17}$$

If, furthermore, $\sigma/\lambda \ll |u_i|, |v_j|, |w_k| \ll \min\{\lambda/(\sigma p), 1/\sqrt{\log(p)}\}$, *then*

$$\frac{\hat{\mathcal{T}}_{ijk} - \mathcal{T}_{ijk}}{\sigma\sqrt{\hat{u}_i^2\hat{v}_j^2 + \hat{v}_j^2\hat{w}_k^2 + \hat{w}_k^2\hat{u}_i^2}} \xrightarrow{d.} N(0, 1) \quad \text{as} \quad p \to \infty. \tag{3.18}$$

Theorem 3.5.1 establishes the asymptotic distribution for any linear functional $q_1^{\top}\hat{u}$, $q_2^{\top}\hat{v}$, $q_3^{\top}\hat{w}$. Theorem 3.5.1 also implies that $[\hat{\mathcal{T}}_{ijk} - z_{\alpha/2}\sigma\sqrt{\hat{u}_i^2\hat{v}_j^2 + \hat{v}_j^2\hat{w}_k^2 + \hat{w}_k^2\hat{u}_i^2}, \hat{\mathcal{T}}_{ijk} + z_{\alpha/2}\sigma\sqrt{\hat{u}_i^2\hat{v}_j^2 + \hat{v}_j^2\hat{w}_k^2 + \hat{w}_k^2\hat{u}_i^2}]$ is an asymptotic $(1-\alpha)$ confidence interval for T_{ijk} under some boundedness conditions of $|u_i|, |v_j|, |w_k|$. Here, the upper bound $|u_i|, |v_j|, |w_k| \ll \min\{\lambda/(\sigma p), 1/\sqrt{\log(p)}\}$ is significantly weaker than the incoherence condition commonly used in the matrix/tensor estimation/inference literature. On the other hand, the lower bound condition, $|u_i|, |v_j|, |w_k| \gg \sigma/\lambda$, is essential to ensure the asymptotic normality of $\hat{\mathcal{T}}$. To see this, consider a special case that $u_i = v_j = w_k = 0$, then (3.17) implies

$$\frac{\lambda^2\hat{\mathcal{T}}_{ijk}}{\sigma^3} \xrightarrow{d.} G_1G_2G_3 \ \text{as}\ p \to \infty, \quad (G_1, G_2, G_3)^{\top} \sim N(0, I_3).$$

In other words, $\hat{\mathcal{T}}_{ijk}$ satisfies a third moment Gaussian, not a Gaussian distribution.

To cover the broader scenarios that the lower bound conditions are absent, we consider the following lower-thresholding procedure. Let $s(t) = \max\{t, \log(p)\sigma^2\hat{\lambda}^{-2}\}$[‡] for $t \geqslant 0$ and define the confidence interval for $\mathcal{T}_{ijk}$ as

$$\begin{aligned}\widetilde{\mathrm{CI}}_\alpha(\hat{\mathcal{T}}_{ijk}) :=& \Big[\hat{\mathcal{T}}_{ijk} - z_{\alpha/2}\sigma\sqrt{s(\hat{u}_i^2)s(\hat{v}_j^2) + s(\hat{v}_j^2)s(\hat{w}_k^2) + s(\hat{w}_k^2)s(\hat{u}_i^2)},\\ &\hat{\mathcal{T}}_{ijk} + z_{\alpha/2}\sigma\sqrt{s(\hat{u}_i^2)s(\hat{v}_j^2) + s(\hat{v}_j^2)s(\hat{w}_k^2) + s(\hat{w}_k^2)s(\hat{u}_i^2)}\Big]. \end{aligned} \tag{3.19}$$

We can prove $\widetilde{\mathrm{CI}}_\alpha(\hat{\mathcal{T}}_{ijk})$ is a valid $(1-\alpha)$-level asymptotic confidence interval.

Theorem 3.5.2. *Suppose the conditions in Theorem 3.5.1 hold. If $\lambda/\sigma \gg p^{3/4}$ and $|u_i|, |v_j|, |w_k| \ll \min\{\lambda/(\sigma p), 1/\sqrt{\log(p)}\}$ for $i \in [p_1], j \in [p_2], k \in [p_3]$, then*

$$\liminf_{p\to\infty} \mathbb{P}\left(\mathcal{T}_{ijk} \in \widetilde{\mathrm{CI}}_\alpha(\hat{\mathcal{T}}_{ijk})\right) \geqslant 1-\alpha. \tag{3.20}$$

Remark 3.5.1 (Proof sketch of Theorem 3.5.1)**.** *The proof scheme for Theorem 3.5.1 is essentially different from many recent literature on the entrywise inference (Chen et al., 2019b; Xia and Yuan, 2020; Cai et al., 2020) and we provide a proof sketch here. Without loss generality, we assume $\sigma = 1$ and $\langle u, \hat{u}\rangle, \langle v, \hat{v}\rangle, \langle w, \hat{w}\rangle \geqslant 0$. First, we can decompose $\langle \hat{u}, q_1\rangle$ into two terms:*

$$\langle q_1, \hat{u}\rangle = \langle \hat{u}, uu^\top q_1\rangle + \langle \hat{u}, (I - uu^\top)q_1\rangle = (q_1^\top u)\hat{u}^\top u + (U_\perp^\top q_1)^\top U_\perp^\top \hat{u}. \tag{3.21}$$

Similar decompositions hold for $\langle \hat{v}, q_2\rangle$ and $\langle \hat{w}, q_3\rangle$. For any $O_i \in \mathbb{O}_{p_i-1}$, we construct three rotation matrices as

$$\tilde{O}_1 = uu^\top + U_\perp O_1 U_\perp^\top \in \mathbb{O}_{p_1}, \quad \tilde{O}_2 = vv^\top + V_\perp O_2 V_\perp^\top \in \mathbb{O}_{p_2}, \quad \tilde{O}_3 = ww^\top + W_\perp O_3 W_\perp^\top \in \mathbb{O}_{p_3},$$

where $U_\perp \in \mathbb{O}_{p_1,p_1-1}, V_\perp \in \mathbb{O}_{p_2,p_2-1}, W_\perp \in \mathbb{O}_{p_3,p_3-1}$ are the orthogonal complement of u, v, w, respectively. A key observation is that $\tilde{\mathcal{A}} = \mathcal{A} \times_1 \tilde{O}_1^\top \times_2 \tilde{O}_2^\top \times_3 \tilde{O}_3^\top$ and $\mathcal{A}$

[‡] Here, $\log(p)$ can be replaced by any value that grows to infinity as p grows.

share the same distribution. Suppose $\tilde{u}, \tilde{v}, \tilde{w}$ *are the outputs of Algorithm 4. Then we have* $\tilde{u} = \tilde{O}_1^\top \hat{u}, \tilde{v} = \tilde{O}_2^\top \hat{v}, \tilde{w} = \tilde{O}_3^\top \hat{w}$ *and can further prove that given* $\langle u, \hat{u}\rangle, \langle v, \hat{v}\rangle$ *and* $\langle w, \hat{w}\rangle$, $\left(\frac{\hat{u}^\top U_\perp}{\|U_\perp^\top \hat{u}\|_2}, \frac{\hat{v}^\top V_\perp}{\|V_\perp^\top \hat{v}\|_2}, \frac{\hat{w}^\top W_\perp}{\|W_\perp^\top \hat{w}\|_2}\right)$ *and* $\left(\frac{\hat{u}^\top U_\perp}{\|U_\perp^\top \hat{u}\|_2} O_1, \frac{\hat{v}^\top V_\perp}{\|V_\perp^\top \hat{v}\|_2} O_2, \frac{\hat{w}^\top W_\perp}{\|W_\perp^\top \hat{w}\|_2} O_3\right)$ *have the same distribution. By the uniqueness of the Haar measure (Neumann, 1936; Weil, 1940) and Theorem 3.3.1, we can further prove for any fixed vectors* $f_1 \in \mathbb{S}^{p_1-2}, f_2 \in \mathbb{S}^{p_2-2}, f_3 \in \mathbb{S}^{p_3-2}$, *we have*

$$\Big(\lambda \hat{u}^\top U_\perp f_1,\ \lambda \hat{v}^\top V_\perp f_2,\ \lambda \hat{w}^\top W_\perp f_3,$$

$$\frac{\hat{u}^\top u - (1 - p_1\lambda^{-2}/2)}{\sqrt{p_1/2}\lambda^{-2}},\ \frac{\hat{v}^\top v - (1 - p_2\lambda^{-2}/2)}{\sqrt{p_2/2}\lambda^{-2}},\ \frac{\hat{w}^\top w - (1 - p_3\lambda^{-2}/2)}{\sqrt{p_3/2}\lambda^{-2}}\Big)^\top \xrightarrow{d} N(0, I_6).$$

This inequality and (3.21) *result in* (3.16)(3.17)(3.18).

Remark 3.5.2. *The entrywise inference for Tucker low-rank or orthogonal decomposable tensor PCA can be significantly more challenging due to the dependence among different factors. We leave it as future research.*

3.6 Numerical Simulations

We now conduct numerical studies to support our theoretical findings in previous sections. Each experiment is repeated for 2000 times, from which we obtain 2000 realizations of the respective statistics. Then we draw histograms or boxplots, and compare with the corresponding baselines. In each histogram, the red line is the density of the standard normal distribution.

We begin with the inference for principal components of Tucker low-rank tensors. Specifically, we randomly draw $\check{U}_j \in \mathbb{R}^{p_j \times r_j}$ with i.i.d. standard normal entries and normalize to $U_j = \text{QR}(\check{U}_j)$. We then draw core tensor $\check{\mathcal{G}} \in \mathbb{R}^{r\times r\times r}$ with i.i.d. standard normal entries and rescale to $\mathcal{G} = \check{\mathcal{G}} \cdot p^\gamma/(\lambda_{\min}(\check{\mathcal{G}}))$. Consequently, U_j is uniform randomly selected from $\mathbb{O}_{p_j, r_j}$ and $\lambda_{\min}(\mathcal{G}) = \lambda = p^\gamma$. For $p_1 = p_2 = p_3 = 200$, $r = 3$, and $\sigma = 1$, each value of $\gamma \in \{0.80, 0.85, 0.90, 0.95\}$, we observe $\mathcal{A}$ under tensor PCA model (3.1) and apply Algorithm 1 to obtain a realization of

$$\mathsf{T} = \frac{\|\sin\Theta(\hat{U}_1, U_1)\|_F^2 - p\|\Lambda_1^{-1}\|_F^2}{\sqrt{2p_1}\|\Lambda_1^{-2}\|_F}.$$

We repeat this procedure for 2000 times, from which we obtain 2000 realizations of the respective statistics and plot the density histograms in Figure 3.1. We can see T achieves good normal approximation in these settings.

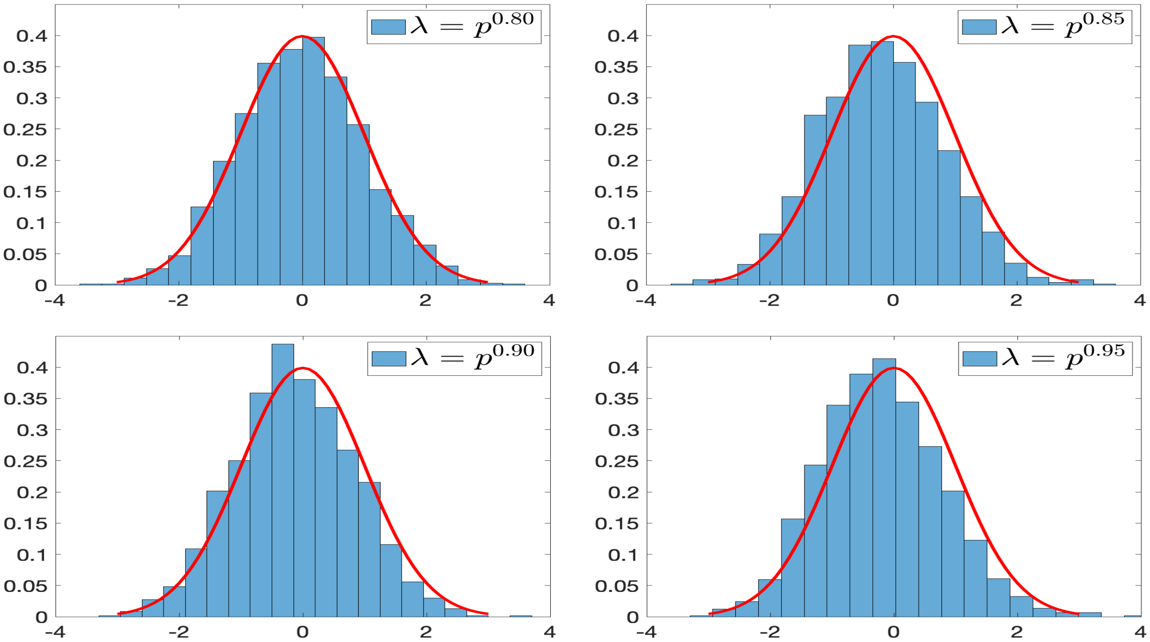


Figure 3.1: Normal approximation of $\frac{\|\sin\Theta(\hat{U}_1,U_1)\|_F^2 - p\|\Lambda_1^{-1}\|_F^2}{\sqrt{2p}\|\Lambda_1^{-2}\|_F}$ for order-3 Tucker low-rank tensor PCA model (3.1). Here, $p_1 = p_2 = p_3 = p = 200$, $r = 3$, $\sigma = 1$.

We then consider the asymptotic normality in orthogonally decomposable tensors under the tensor PCA model. Similarly, we fix $p = 200$, $r = 3$, and construct the orthogonally decomposable tensor as $\mathcal{T} = \sum_{i=1}^{r}(r+1-i)\lambda \cdot (u_i \otimes v_i \otimes w_i)$, where $[u_1, \ldots, u_r], [v_1, \ldots, v_r], [w_1, \ldots, w_r]$ are drawn uniform randomly from $\mathbb{O}_{p,r}$ similarly to the previous setting and $\lambda = p^\gamma$ with $\gamma = 0.80, 0.85, 0.90, 0.95$. For each γ, we obtain 2000 replicates of $\mathsf{T} = \frac{\langle \hat{u}_3, u_3\rangle^2 - (1-p\lambda^{-2})}{\sqrt{2p}\lambda^{-2}}$, draw the density histogram, and plot the results in Figure 3.2. We can see the normal approximation of T becomes more accurate as the signal strength λ grows.

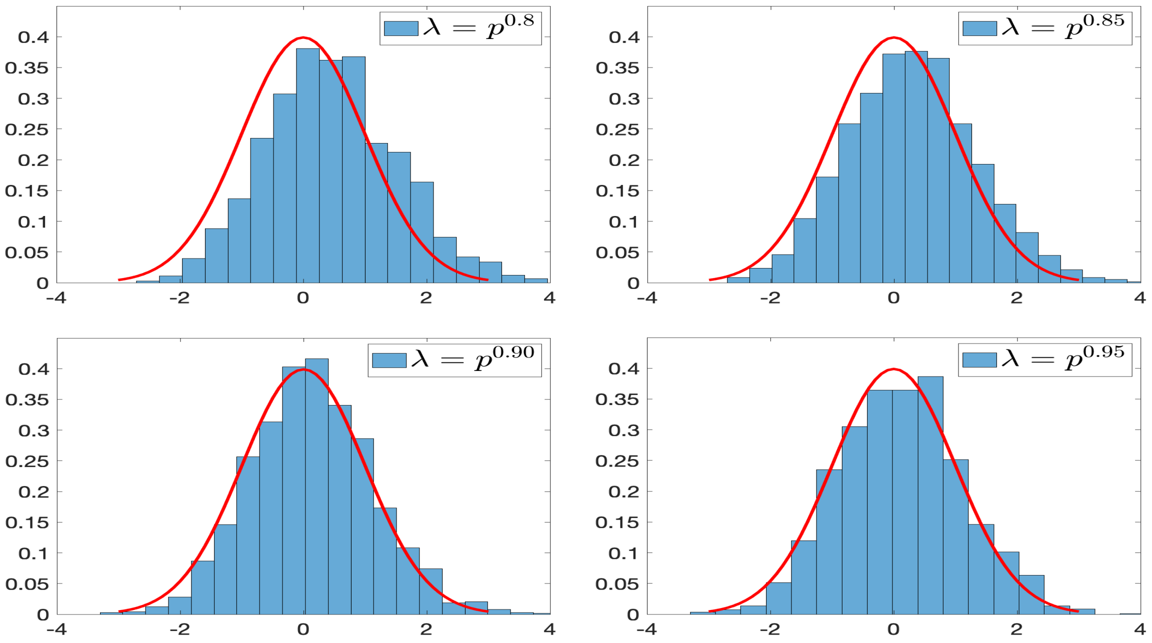


Figure 3.2: Normal approximation of $\frac{\langle \hat{u}_3, u_3 \rangle^2 - (1 - p\lambda^{-2})}{\sqrt{2}p\lambda^{-2}}$ for tensor PCA model (3.1) when $\mathcal{T}$ is a third-order orthogonally decomposable tensor and $\sigma = 1$. Here, $p_1 = p_2 = p_3 = p = 200, r = 3, \lambda_{\text{min}} = \lambda$.

Though the focus of this chapter is on third-order tensors, we will explain later in Section 3.7 that the results can be generalized to higher-order ones. Next, we conduct simulation study on tensor PCA model for fourth-order orthogonally decomposable tensors when $p = 100$ and $r = 1$. With a few modifications on the proof, we can show $\left(\langle \hat{u}_1, u_1 \rangle^2 - (1 - p\lambda^{-2})\right)(\sqrt{2}p\lambda^{-2})^{-1}$ is asymptotically normal under the required SNR assumption for efficient computation: $\text{SNR} \geqslant Cp$. The simulation results in Figure 3.3 show that equipped with a warm initialization, the two-iteration alternating minimization yields an estimator achieving good normal approximation even if $\text{SNR} \approx p^{0.9}$, which is strictly weaker than the required SNR assumption for efficient computation. See more discussions in Section 3.7.

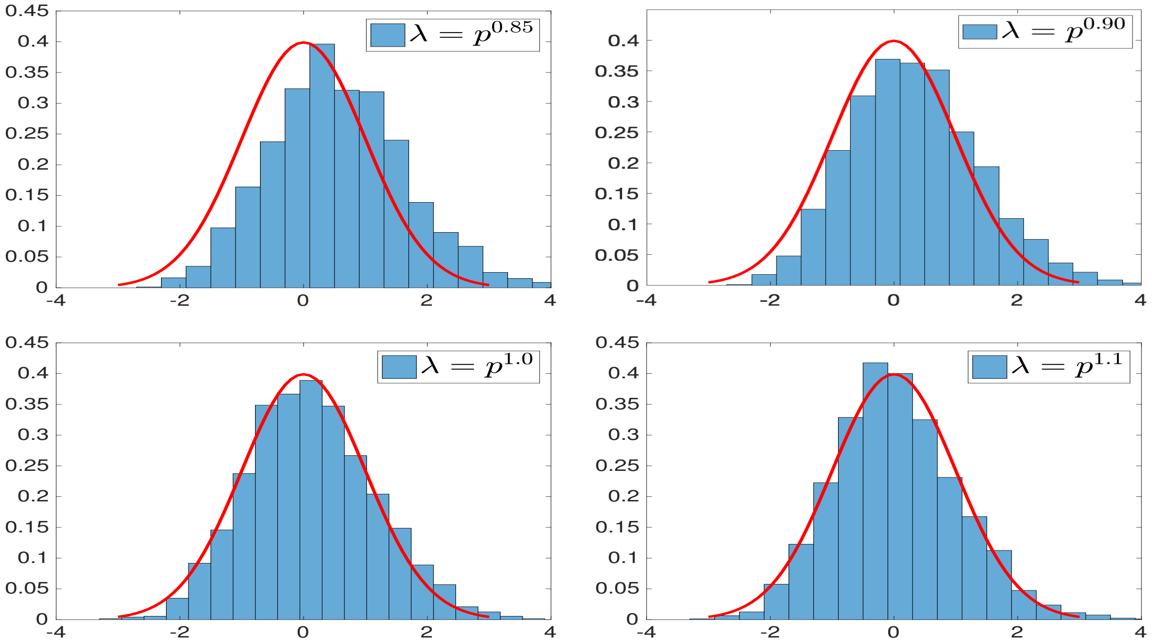


Figure 3.3: Normal approximation of $\frac{\langle \hat{u}_1, u_1\rangle^2 - (1 - p\lambda^{-2})}{\sqrt{2}p\lambda^{-2}}$ for tensor PCA model (3.1) when $\mathcal{T} = \lambda \cdot (u_1 \otimes v_1 \otimes w_1 \otimes q_1)$ is a fourth-order tensor and $\sigma = 1$. Here, $p_1 = p_2 = p_3 = p_4 = p = 100, r = 1$ and $\lambda_{\min} = \lambda$.

Then, we consider the entrywise inference under the rank-1 tensor PCA model. We construct $\mathcal{T} = \lambda \cdot u \otimes v \otimes w \in \mathbb{R}^{p \times p \times p}$, where $u = v = w = (1/\sqrt{p}, \ldots, 1/\sqrt{p})^\top$ and $\lambda = p^\gamma$ with $\gamma \in \{0.80, 0.85, 0.90, 0.95\}$. For each value of γ, we draw a random observation $\mathcal{A}$ under the tensor PCA model (3.1) and apply Algorithm 4 with $t_{\max} = 10$. We present the histogram in Figure 3.4 based on 2000 replicate values of $\frac{\hat{\mathcal{T}}_{1,1,1} - \mathcal{T}_{1,1,1}}{\sqrt{\hat{u}_1^2 \hat{v}_1^2 + \hat{v}_1^2 \hat{w}_1^2 + \hat{w}_1^2 \hat{u}_1^2}}$. The simulation results validate the asymptotic normality of $\frac{\hat{\mathcal{T}}_{ijk} - \mathcal{T}_{ijk}}{\sqrt{\hat{u}_i^2 \hat{v}_j^2 + \hat{v}_j^2 \hat{w}_k^2 + \hat{w}_k^2 \hat{u}_i^2}}$ when u, v, w have balanced entry values, which are in line with the theory in Theorem 3.5.1.

Finally, we consider the accuracy of the asymptotic entrywise confidence interval proposed in (3.19) under the tensor PCA model. Let $\mathcal{T} = \lambda \cdot u \otimes v \otimes w$ be a rank-1 tensor, where u, v, w are uniform randomly drawn from $\mathbb{S}^{p-1}$ for $p \in \{100, 200\}$ and $\lambda = p^\gamma$ for $\gamma \in \{0.80, 0.85, 0.90, 0.95\}$. For each combination of (p, γ), we report the

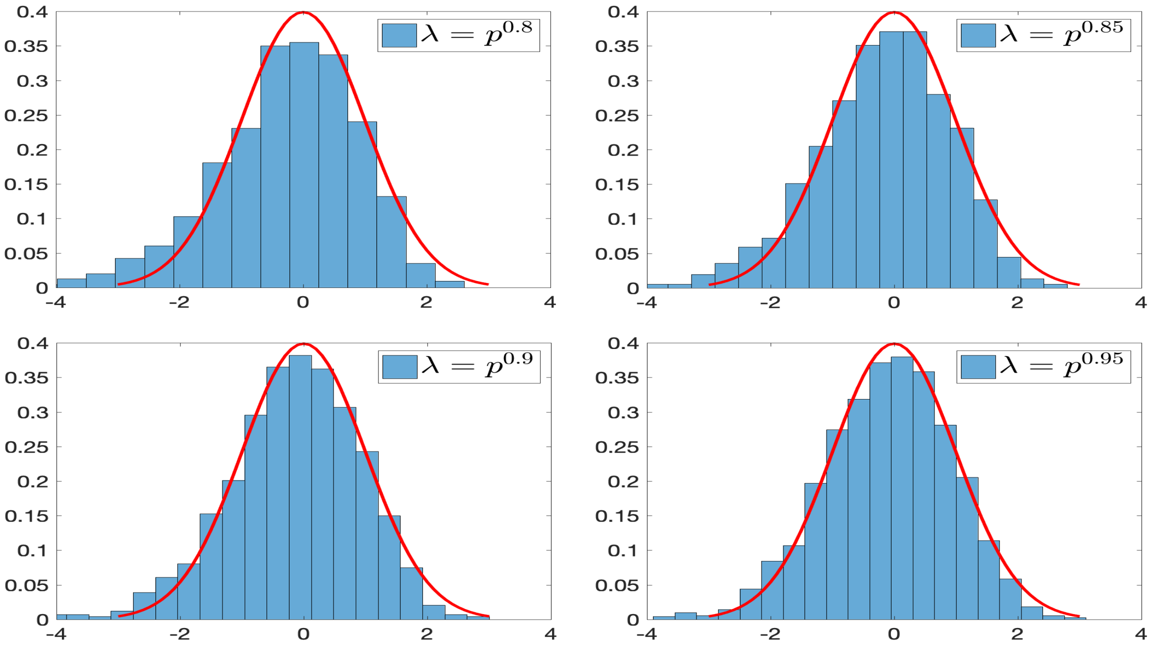


Figure 3.4: Normal approximation of $\frac{\hat{\mathcal{T}}_{1,1,1}-\mathcal{T}_{1,1,1}}{\sqrt{\hat{u}_1^2\hat{v}_1^2+\hat{v}_1^2\hat{w}_1^2+\hat{w}_1^2\hat{u}_1^2}}$ for tensor PCA model (3.1) when $\mathcal{T}$ is a rank-1 tensor and $\sigma = 1$. The parameters are $p_1 = p_2 = p_3 = p = 200$ with signal strength λ.

empirical coverage rates for the 0.95-confidence interval $\widehat{\mathrm{CR}}_{ijk}$ by boxplots in Figure 3.5. The results show the empirical coverage rates are close to 0.95 in all settings and larger values of (γ, p) lead to more accurate coverage.

3.7 Discussion

In this chapter, we investigate the inference for low-rank tensors under two basic and fundamentally important tensor models: tensor PCA and regression. Based on an initial estimator achieving a reasonable estimation error, we propose to update by a two-iteration alternating minimization algorithm then establish the asymptotic distribution for the singular subspace outcomes. Distributions of general linear forms of the singular vectors are also established for rank-one tensor PCA model,

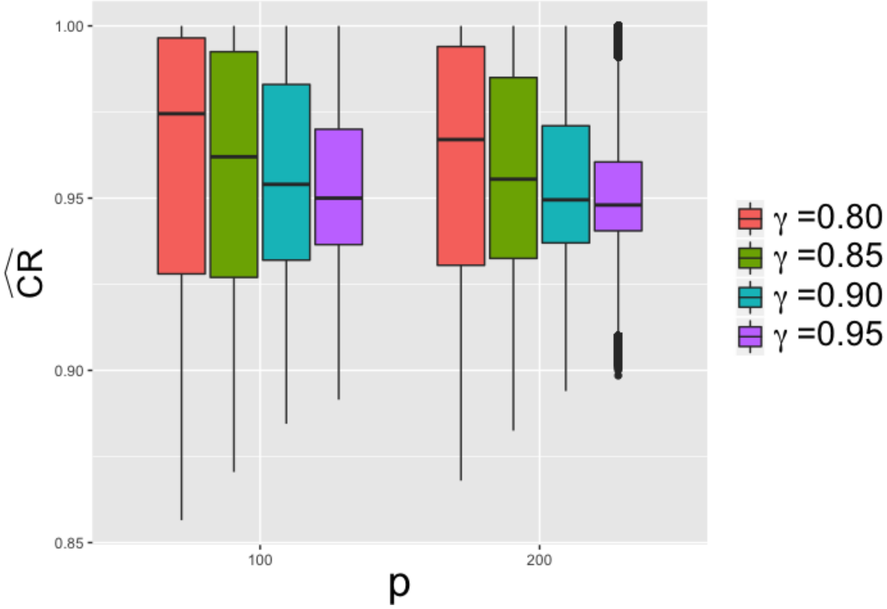


Figure 3.5: Boxplots for empirical coverage of entrywise confidence interval $\widehat{\mathrm{CR}}_{ijk}$

which further enables the entrywise inference on the parameter tensor.

Although our main focus is on third-order tensors, the results in this chapter can be extended to higher-order tensors. For example, suppose $m \geqslant 4$ and $\mathcal{T} = \sum_{j=1}^r \lambda_j \cdot u_j^{(1)} \otimes \cdots \otimes u_j^{(m)}$ is orthogonally decomposable. Given $\mathcal{A}$ from the tensor PCA model (3.1) and Assumption 3.4.1 holds, we can refine by two power iterations similarly to Algorithm 3, then obtain $\{\hat{u}_j^{(1)}, \hat{u}_j^{(2)}, \cdots, \hat{u}_j^{(m)}\}_{j=1}^r$. Similarly to Theorem 3.4.1, we can prove

$$\frac{\langle u_j^{(k)}, \hat{u}_j^{(k)} \rangle^2 - (1 - p_k \lambda_j^{-2})}{\sqrt{2p_k}\lambda_j^{-2}} \xrightarrow{d.} N(0,1), \quad k = 1, \ldots, m,$$

if $\lambda_{\mathsf{min}}/\sigma \gg p^{3/4}$ and other regularity conditions holds. If $m \geqslant 4$, the SNR condition $\lambda_{\mathsf{min}} \gg p^{3/4}$ is weaker than the condition that ensure a computationally feasible estimator exists, i.e., $\lambda_{\mathsf{min}}/\sigma \gg p^{m/4}$ (Zhang and Xia, 2018). In other words, if an sufficiently good initial estimate is already available, a weaker SNR condition $\lambda_{\mathsf{min}} \gg p^{3/4}$ is sufficient to guarantee the asymptotic normality of our final estimates.

This phenomenon is further justified by the simulation results in Figure 3.3.

Chapter 4

High-order Tensor SVD [*]

4.1 Introduction

Tensors, or high-order arrays, have attracted increasing attention in modern machine learning, computational mathematics, statistics, and data science. Some specific examples include recommender systems (Nasiri et al., 2014; Bi et al., 2018), neuroimaging analysis (Zhou et al., 2013; Wozniak et al., 2007), latent variable learning (Anandkumar et al., 2014a), multidimensional convolution (Oseledets and Tyrtyshnikov, 2009b), signal processing (Cichocki et al., 2015), neural network (Zhong et al., 2017; Mondelli and Montanari, 2019), computational imaging (Li and Li, 2010; Zhang et al., 2020b), contingency table (Dunson and Xing, 2009; Bhattacharya and Dunson, 2012). In addition to low-order tensors (e.g., tensor with a relatively small value of order number), the high-order tensors also commonly arise in applications in statistics and machine learning. For example, in convolutional neural networks, parameters in fully connected layers can be represented as high-order tensors (Novikov et al., 2015; Calvi et al., 2019). In an order-d Markov process, where the future states depend on jointly the current and $(d-1)$ previous states, the transition probabilities form an order-$(d+1)$ tensor. For an order-d Markov decision process, the transition probabilities can be represented by an order-$(2d+1)$

[*] This work is based on Zhou et al. (2020) (`https://arxiv.org/abs/2010.02482`).

tensor, with additional d directions representing past d actions. High-order tensors are also used to represent the joint probability in Markov random fields (Novikov et al., 2014).

Compared to the low-order tensors, high-order tensors encompass much more parameters and sophisticated structure, while leading to inhibitive cost in storage, processing, and analysis: an order-d dimension-p tensor contains p^d parameters. To address this issue, some low-dimensional parametrization is usually considered to capture the most informative subspaces in the tensor. In particular, the tensor-train (TT) decomposition (Oseledets, 2009; Oseledets and Tyrtyshnikov, 2009a; Oseledets, 2011; Fannes et al., 1992; Orús, 2019) introduced a classic low-dimensional parameterization to model the subspaces and latent cores in high-order tensor structures. TT decomposition has been used in a wide range of applications in physics and quantum computation (Bravyi et al., 2019; Fannes et al., 1992; Orús, 2019; Schollwöck, 2011; Rakhuba and Oseledets, 2016), signal processing (Cichocki et al., 2015), and supervised learning (Stoudenmire and Schwab, 2016) among many others. For example, the TT decomposition framework is utilized in quantum information science for modeling complex quantum states and handling the quantum mean value problem (Bravyi et al., 2019; Fannes et al., 1992; Orús, 2019; Rakhuba and Oseledets, 2016). The TT-decomposition of a tensor $\mathcal{X} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$ is defined as below:

$$\begin{aligned}\mathcal{X}_{i_1,\cdots,i_d} =& G_{1,[i_1,:]}\mathcal{G}_{2,[:,i_2,:]}\cdots\mathcal{G}_{d-1,[:,i_{d-1},:]}G^\top_{d,[i_d,:]}\\ =&\sum_{\alpha_1=1}^{r_1}\cdots\sum_{\alpha_{d-1}=1}^{r_{d-1}} G_{1,[i_1,\alpha_1]}\mathcal{G}_{2,[\alpha_1,i_2,\alpha_2]}\cdots\mathcal{G}_{d-1,[\alpha_{d-2},i_{d-1},\alpha_{d-1}]}G_{d,[i_d,\alpha_{d-1}]}.\end{aligned} \tag{4.1}$$

Here, the smallest values of $r_1, \ldots, r_{d-1}$ that enable the decomposition (4.1) are called the *TT-rank* of $\mathcal{X}$. Oseledets (2011) shows that the TT-rank $r_k = \mathrm{rank}([\mathcal{X}]_k)$, i.e., the rank of the kth sequential unfolding of $\mathcal{X}$ (see formal definition of sequential unfolding in Section 4.2.1). $G_1 \in \mathbb{R}^{p_1 \times r_1}$, $\mathcal{G}_k \in \mathbb{R}^{r_{k-1} \times p_k \times r_k}$, $G_d \in \mathbb{R}^{p_d \times r_{d-1}}$ are the *TT-cores* that multiply sequentially like a “train": $\mathcal{X}_{i_1,\cdots,i_d}$ equals the product of i_1th vector in G_1, i_2th matrix in $\mathcal{G}_2$, ..., i_{d-1}th matrix in $\mathcal{G}_{d-1}$, and i_dth vector in G_d. For convenience of presentation, we simplify (4.1) to

$$\mathcal{X} = [\![G_1, \mathcal{G}_2, \ldots, \mathcal{G}_{d-1}, G_d]\!]$$

and denote $r_0 = r_d = 1$ throughout the chapter. In particular, the TT rank and TT decomposition reduce to the regular matrix rank and decomposition when $d = 2$. If all dimensions p and ranks r are the same, the TT-parametrization involves $O(2pr + (d-2)pr^2)$ values, which can be significantly smaller than the ones for Tucker-decomposition $O(r^d + dpr)$ and the regular parameterization $O(p^d)$.

In most of the existing literature, the TT-decomposition was considered under the deterministic settings, and the central goal was often to *approximate* the nonrandom high-order tensors by low-dimensional structures (Oseledets, 2011; Oseledets and Tyrtyshnikov, 2010; Bigoni et al., 2016). However, in modern applications in data science such as Markov processes, Markov decision processes, and Markov random fields, the (transition) probability tensor computed based on data is often a random realization of the underlying true tensor. In these cases, the *estimation* of the underlying low-dimensional parameters hidden in the noisy observations can be more important: an accurate estimation of the transition tensor renders reliable prediction for future states in high-order Markov chains and better decision-making in high-order Markov decision processes; an accurate estimation of probability tensor sheds light to the underlying relationship among different variables in a random system (Novikov et al., 2014). To achieve such a goal, it is crucial to develop dimension reduction methods that can incorporate TT-decomposition into probabilistic models. Since singular value decomposition (SVD) is one of the most important dimension reduction methods involving probabilistic models for matrices, and there is no counterpart of it for high-order tensors, we aim to fill this void by developing a statistical framework and a computationally feasible method for high-order tensor SVD in this chapter.

4.1.1 Problem Formulation

This chapter focuses on the following *high-order tensor SVD model*. Suppose we observe an order-d tensor $\mathcal{Y}$ that contains a hidden tensor-train (TT) low-rank structure:

$$\mathcal{Y} = \mathcal{X} + \mathcal{Z}, \quad \mathcal{Y}, \mathcal{X}, \mathcal{Z} \in \mathbb{R}^{\otimes_{k=1}^{d} p_k}. \tag{4.2}$$

Here, $\mathcal{X}$ is TT-decomposable as (4.1) and $\mathcal{Z}$ is a noise tensor. Our goal is to estimate $\mathcal{X}$ and the TT cores of $\mathcal{X}$ based on $\mathcal{Y}$. To this end, a straightforward idea is to minimize the approximation error as follows,

$$\widehat{\mathcal{X}} = \underset{\mathcal{A} \text{ is decomposable as (4.1)}}{\arg\min} \|\mathcal{Y} - \mathcal{A}\|_F^2. \tag{4.3}$$

However, the approximation error minimization (4.3) is highly non-convex and finding the global optimal solution, even if the rank $r_1 = \cdots = r_{d-1} = 1$, is NP-hard in general (Hillar and Lim, 2013). Instead, a variety of computationally feasible methods have been proposed to approximate the best tensor-train low-rank decomposition in the literature. TT-SVD, a sequential singular value thresholding scheme, was introduced by Oseledets (2011) to be discussed in detail later. Oseledets (2011) also proposed TT-rounding via sequential QR decompositions, which reduces the TT-rank while ensures approximation accuracy. Dolgov and Savostyanov (2014) introduced the alternating minimal energy algorithm to reconstruct a TT-low-rank tensor approximately based on only a small proportion of revealed entries of the target tensor. (Song et al., 2017, Section L.2) proposed a sketching-based algorithm for fast low TT rank approximation of arbitrary tensors. Bigoni et al. (2016) studied the tensor-train decomposition for functional tensors. Li et al. (2019a) proposed the FastTT algorithm for fast sparse tensor decomposition based on parallel vector rounding and TT-rounding. Lubich et al. (2013) studied dynamical approximation with TT format for time-dependent tensors. Grasedyck et al. (2015) proposed the alternating least squares for tensor completion in the TT format. Bengua et al. (2017) studied the completion of low TT rank tensor and the applications to color image and video recovery. Steinlechner (2016) studied the Riemannian optimization methods for TT decomposition and completion. Also see Novikov et al. (2020) for a TT decomposition library in TensorFlow. To our best knowledge, the estimation performance of most procedures here remains unclear. Departing from these existing work, in this chapter, we make a first attempt to minimize the estimation error of $\mathcal{X}$ in addition to achieving the minimal approximation error under possibly random settings.

4.1.2 Our Contributions

Under Model (4.2), we make the following contributions to high-order tensor SVD in this chapter.

First, we propose a new algorithm, *Tensor-Train Orthogonal Iteration* (TTOI), that provides a computationally efficient estimation of the low-rank TT structure from the noisy observation. The proposed algorithm includes two major steps. First, we obtain initial estimates $\widehat{G}_1^{(0)}, \widehat{\mathcal{G}}_2^{(0)}, \dots, \widehat{\mathcal{G}}_{d-1}^{(0)}, \widehat{G}_d$ by performing forward sequential SVD based on matricizations and projections. This step was known as TT-SVD in the literature (Oseledets, 2011). Next, we utilize the initialization and perform the newly developed *backward updates* and *forward updates* alternatively and iteratively. The TTOI procedure will be discussed in detail in Section 4.2.

To see why the TTOI iterations yield better estimation than the classic TT-SVD method, recall that TT-SVD first performs singular value thresholding on $[\mathcal{Y}]_1$, i.e., the unfolding of $\mathcal{Y}$, without any additional updates (see detailed procedure of TT-SVD and formal definition of $[\mathcal{Y}]_1$ in Section 4.2.1), which can be inaccurate since $[\mathcal{Y}]_1$, a p_1-by-$\prod_{k=2}^d p_k$ matrix, has a great number of columns. In contrast, TTOI iteration utilizes the intermediate outcome of the previous iteration to substantially reduce the dimension of $[\mathcal{Y}]_1$ while performing singular value thresholding. In Figure 4.1, we provide a simple simulation example to show that even one TTOI iteration can significantly improve the estimation of the left singular subspace of G_1 (left panel) and the overall tensor $\mathcal{X}$ (right panel). Therefore, a one-step TTOI, i.e., the initialization with one TTOI iteration, can be used in practice when the computational cost is a concern.

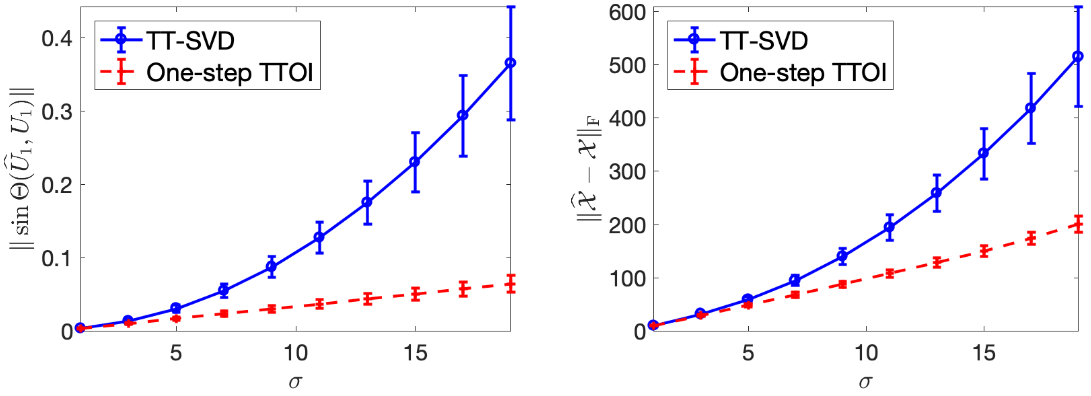


Figure 4.1: Average estimation error (dots) and standard deviation (bars) of $\|\sin\Theta(\widehat{U}_1, U_1)\|$ and $\|\widehat{\mathcal{X}} - \mathcal{X}\|_F$ by TT-SVD and one-step TTOI. Both algorithms are performed based on the observation $\mathcal{Y}$ generated from (4.2), where $\mathcal{Z} \overset{iid}{\sim} N(0, \sigma^2)$, $\mathcal{X}$ is a randomly generated order-5 tensor based on (4.1) with $p = 20, r = 1$, $G_1, \mathcal{G}_2, \dots, \mathcal{G}_{d-1}, G_d \overset{iid}{\sim} N(0, 1)$.

We develop theoretical guarantees for TTOI. In particular, we introduce a series of representation lemmas for tensor matricizations with TT format. Based on them, we develop a deterministic upper bound of estimation error for both forward and backward updates in TTOI iterations. Under the benchmark setting of spiked tensor model, we develop matching upper/lower bounds and prove that the proposed TTOI algorithm achieves the minimax optimal rate of estimation error. To the best of our knowledge, this is the first statistical optimality result for high-order tensors with TT format. We also prove for any high-order tensor, TTOI iteration has monotone decreasing approximation error with respect to the iteration index.

Moreover, to break the curse of dimensionality in high-order Markov processes, we study the state aggregatable high-order Markov processes and establish a key connection to TT decomposable tensors. We propose a TTOI estimator for the transition probability tensor in high-order state-aggregatable Markov processes and establish the theoretical guarantee. We conduct simulation experiments to demonstrate the performance of TTOI and validate our theoretical findings. We also apply our method to analyze a New York taxi dataset. By modeling taxi trips as trajectories realized from a citywide Markov chain, we found that the Manhattan traffic zone exhibits high-order Markovian dependence and the proposed TTOI reveals latent traffic patterns and meaningful partition of Manhattan traffic zones. Finally, we discuss several applications that our proposed algorithm is applicable to, including transition probability tensor estimation in high-order Markov decision processes and joint probability tensor estimation in Markov random fields.

4.1.3 Related Literature

In addition to the aforementioned literature on TT decomposition, our work is also related to a substantial body of work on matrix/tensor decomposition and SVD, spiked tensor model, etc. These literature are from a range of communities including applied mathematics, information theory, machine learning, scientific computing, signal processing, and statistics. Here we try to review existing literature in these communities without claiming this literature survey is exhaustive.

First, the matrix singular value thresholding was commonly used and extensively studied in various problems in data science, including matrix denoising (Candes et al., 2013a; Donoho and Gavish, 2014; Cai and Zhang, 2018), matrix completion (Cai et al., 2010; Zhang et al., 2011; Klopp, 2015; Chatterjee, 2015), principal component analysis (PCA) (Nadler, 2008), Markov chain state aggregation (Zhang and Wang, 2020). Such the task was also widely considered for tensors of order-3 or higher. In particular, to perform SVD and decomposition for tensors with Tucker low-rank structures, De Lathauwer et al. (2000a,b) introduced the higher-order SVD (HOSVD) and higher-order orthogonal iteration (HOOI). Zhang and Xia (2018) established the statistical and computational limits of tensor SVD, compared the theoretical properties of HOSVD and HOOI, and proved that HOOI achieves both statistical and computational optimality. Vannieuwenhoven et al. (2012) introduced the sequentially truncated higher-order singular value decomposition (ST-HOSVD).

Zhang and Han (2019) introduced a thresholding & projection based algorithm for sparse tensor SVD. A non-exhaustive list of methods for SVD and decomposition for tensors with CP low-rank structures include alternating least squares (Kolda and Bader, 2009; Sharan and Valiant, 2017), eigendecomposition-based approach (Leurgans et al., 1993), enhanced line search (Rajih et al., 2008), power iteration with SVD-based initialization (Anandkumar et al., 2014a), simultaneous diagonalization and higher-order SVD (Colombo and Vlassis, 2016).

In addition, the spiked tensor model and tensor principal component analysis (tensor PCA) are widely discussed in the literature. Richard and Montanari (2014); Hopkins et al. (2015); Anandkumar et al. (2016); Perry et al. (2020); Luo and Zhang (2020c); Arous et al. (2019) considered the statistical and computational limits of rank-1 spiked tensor model. Lesieur et al. (2017) studied the statistical and computational phase transitions and theoretical properties of the approximate message passing algorithm (AMP) under a Bayesian spiked tensor model. Allen (2012a,b) developed the regularization-based methods for tensor PCA. Lu et al. (2016); Zhou and Feng (2017); Liu et al. (2018); Lu et al. (2019) studied the robust tensor PCA to handle the possible outliers from the tensor observation.

Different from Tucker and CP decompositions, which have been a pinpoint in the enormous existing literature on tensors, we focus on the TT-structure associated with high-order tensors for the following reasons: (1) Tucker and CP decompositions do not involve the sequential structure of different modes, i.e., the Tucker and CP decompositions still hold if the d modes are arbitrarily permuted. While in applications such as high-order Markov process, high-order Markov decision process, and fully connected layers of deep neural networks, the order of different modes can be crucial; (2) the number of entries involved in the low-Tucker-rank parameterization grows exponentially with respect to the order d (r^d); (3) methods that explore CP low-rank structure can be numerically unstable for high-order tensors in computation as pointed out by Oseledets and Tyrtyshnikov (2010). In comparison, the TT-structure incorporates the order of different modes sequentially and involves much fewer parameters for high-order tensors, which renders it more suitable in many scenarios.

In Section 4.5, we will further discuss the application of TTOI on high-order Markov processes and state aggregation. This problem is related to a body of literature on dimension reduction and state aggregation for Markov processes that we will discuss in Section 4.5.

4.1.4 Organization

The rest of the article is organized as follows. In Section 4.2, after a brief introduction of the notation and preliminaries, we introduce the procedure of the tensor-train orthogonal iteration. The theoretical results, including three representation lemmas, a general estimation error bound, and the minimax optimal upper and lower bounds under the spiked tensor model, are provided in Sections 4.3 and 4.4. The application to high-order Markov chains is discussed in Section 4.5. The simulation and real data analysis are provided in Sections 4.6.1 and 4.6.2, respectively. Discussions and further applications to Markov random fields and high-order Markov decision processes are briefly discussed in Section 4.7. All technical proofs are provided in Section 5.3.

4.2 Procedure of Tensor-Train Orthogonal Iteration

4.2.1 Notation and Preliminaries

We first introduce the notation and preliminaries to be used throughout the chapter. We use the lowercase letters, e.g., x, y, z, to denote scalars or vectors. We use $C, c, C_0, c_0, \dots$ to denote generic constants, whose actual values may change from line to line. A random variable z is σ-sub-Gaussian if $\mathbb{E}e^{t(z-\mathbb{E}z)} \leqslant e^{\sigma^2 t^2/2}$ for any $t \in \mathbb{R}$. We say $a \lesssim b$ or $a = O(b)$ if $a \leqslant Cb$ for some uniform constant $C > 0$. We write $a = \widetilde{O}(b)$ if $a = O(b \log^{C'}(b))$ for constant $C' > 0$. The capital letters, e.g., X, Y, Z, are used to denote matrices. Specifically, $\mathbb{O}_{p,r} := \{U \in \mathbb{R}^{p \times r} : U^\top U = I_r\}$ is the set of all p-by-r matrices with orthogonal columns.

For $U \in \mathbb{O}_{p,r}$, let $U_{\perp} \in \mathbb{O}_{p,p-r}$ be the orthonormal complement of U, and let $P_U = UU^{\top}$ denote the projection matrix onto the column space of U. For any matrix $A \in \mathbb{R}^{p_1 \times p_2}$, let $A = \sum_{i=1}^{p_1 \wedge p_2} s_i u_i v_i^{\top}$ be the singular value decomposition, where $s_1(A) \geqslant \cdots \geqslant s_{p_1 \wedge p_2}(A) \geqslant 0$ are the singular values of A in non-increasing order. Define $s_{\min}(A) = s_{p_1 \wedge p_2}(A)$, $\mathrm{SVD}_r^L(A) = [u_1 \dots u_r] \in \mathbb{O}_{p_1,r}$, and $\mathrm{SVD}_r^R(A) = [v_1 \dots v_r] \in \mathbb{O}_{p_2,r}$ be the smallest non-trivial singular value, leading r left singular vectors, and leading r right singular vectors of A, respectively. We also write $\mathrm{SVD}^L(A) = \mathrm{SVD}^L_{p_1 \wedge p_2}(A)$ and $\mathrm{SVD}^R(A) = \mathrm{SVD}^L_{p_1 \wedge p_2}(A)$ as the collection of all left and right singular vectors of A, respectively. Define the Frobenius and spectral norms of A as $\|A\|_F = \sqrt{\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} A_{ij}^2} = \sqrt{\sum_{i=1}^{p_1 \wedge p_2} s_i^2(A)}$ and $\|A\| = s_1(A) = \max_{x \in \mathbb{R}^{p_2}} \|Ax\|_2 / \|x\|_2$. For any two matrices $U \in \mathbb{R}^{m_1 \times n_1}$ and $V \in \mathbb{R}^{m_2 \times n_2}$, let

$$U \otimes V = \begin{bmatrix} U_{11} \cdot V & \dots & U_{1n_1} \cdot V \\ \vdots & & \vdots \\ U_{m_1 1} \cdot V & \dots & U_{m_1 n_1} \cdot V \end{bmatrix} \in \mathbb{R}^{(m_1 m_2) \times (n_1 n_2)}$$

be their Kronecker product. To quantify the distance among subspaces, we define the principle angles between $U, \widehat{U} \in \mathbb{O}_{p,r}$ as an r-by-r diagonal matrix: $\Theta(U, \widehat{U}) = \mathrm{diag}(\arccos(s_1), \dots, \arccos(s_r))$, where $s_1 \geqslant \cdots \geqslant s_r \geqslant 0$ are the singular values of $U^{\top}\widehat{U}$. Define the $\sin\Theta$ norm as

$$\|\sin\Theta(U, \widehat{U})\| = \|\,\mathrm{diag}\left(\sin(\arccos(s_1)), \dots, \sin(\arccos(s_r))\right)\| = \sqrt{1 - s_r^2}.$$

The boldface calligraphic letters, e.g., $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Z}}$, are used to denote tensors. For an order-d tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{\otimes_{i=1}^d p_i}$ and $1 \leqslant k \leqslant d-1$, we define $[\boldsymbol{\mathcal{X}}]_k \in \mathbb{R}^{(p_1 \times \cdots \times p_k) \times (p_{k+1} \cdots p_d)}$ as the *sequential unfolding* of $\boldsymbol{\mathcal{X}}$ with rows enumerating all indices in Modes $1, \dots, k$ and columns enumerating all indices in Modes $(k+1), \cdots, d$, respectively. That is, for any $1 \leqslant k \leqslant d$ and $1 \leqslant i_k \leqslant p_k$,

$$([\boldsymbol{\mathcal{X}}]_k)_{(i_k-1)p_1 \cdots p_{k-1} + (i_{k-1}-1)p_1 \cdots p_{k-2} + \cdots + i_1, (i_d-1)p_{k+1} \cdots p_{d-1} + (i_{d-1}-1)p_{k+1} \cdots p_{d-2} + \cdots + i_{k+1}} = \boldsymbol{\mathcal{X}}_{i_1 \dots i_d}.$$

We also define the tensor Frobenius norm of $\mathcal{X}$ as $\|\mathcal{X}\|_F^2 = \sum_{i_1=1}^{p_1} \cdots \sum_{i_d=1}^{p_d} \mathcal{X}_{i_1,\dots,i_d}^2$. For any matrix $A \in \mathbb{R}^{p_1 \times p_2}$ and any tensor $\mathcal{B} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$, let $\text{vec}(A)$ and $\text{vec}(\mathcal{B})$ be the vectorization of A and $\mathcal{B}$, respectively. Formally,

$$(\text{vec}(\mathcal{B}))_{(i_d-1)p_1\cdots p_{d-1}+(i_{d-1}-1)p_1\cdots p_{d-2}+\cdots+i_1} = \mathcal{B}_{i_1,\dots,i_d}, \quad 1 \leqslant i_k \leqslant p_k, \quad k = 1, \dots, d.$$

4.2.2 Procedure of Tensor-Train Orthogonal Iteration

We are now in position to introduce the procedure of Tensor-Train Orthogonal Iteration (TTOI). The pseudocode of the overall procedure is given in Algorithm 1. TTOI includes three main parts: we first run *initialization*, then perform *backward update* and *forward update* alternatively and iteratively.

Algorithm 1 Tensor-Train Orthogonal Iteration (TTOI)

```
Input: $\mathcal{Y}$, $\{p_k\}_{k=1}^d$, $\{r_k\}_{k=1}^{d-1}$, increment tolerance $\varepsilon > 0$, maximum number of iterations $t_{\max}$
Obtain Initialization $\widetilde{R}_1^{(0)}, \dots, \widetilde{R}_{d-1}^{(0)}, \widehat{\mathcal{X}}^{(0)}$ by Algorithm 1(a)
for $t = 1, \dots, t_{\max}$ do
    if t is odd then
        Apply Algorithm 1(b) with input $\widetilde{R}_1^{(t-1)}, \dots, \widetilde{R}_{d-1}^{(t-1)}$ to obtain $\widehat{V}_1^{(t)}, \dots, \widehat{V}_d^{(t)}, \widehat{\mathcal{X}}^{(t)}$
    else
        Apply Algorithm 1(c) with input $\widehat{V}_1^{(t-1)}, \dots, \widehat{V}_d^{(t-1)}$ to obtain $\widetilde{R}_1^{(t)}, \dots, \widetilde{R}_{d-1}^{(t)}, \widehat{\mathcal{X}}^{(t)}$
    end if
    If $\|\widehat{\mathcal{X}}^{(t)}\|_F^2 - \|\widehat{\mathcal{X}}^{(t-1)}\|_F^2 \leqslant \varepsilon$ then break from the for loop
end for
Output: $\widehat{\mathcal{X}} = \widehat{\mathcal{X}}^{(t)}$
```

- **Part 1: Initialization.** First, we obtain an initial estimate of TT-cores $G_1, \mathcal{G}_2, \dots, \mathcal{G}_{d-1}, G_d$. This step is the tensor-train-singular value decomposition (TT-SVD) originally introduced by Oseledets (2011).

(i) Let $R_1^{(0)}$ be the unfolding of $\mathcal{Y}$ along Mode 1. We compute the top-r_1 SVD of $R_1^{(0)}$. Let $\widehat{U}_1^{(0)} \in \mathbb{O}_{p_1,r_1}$ be the first r_1 left singular vectors of $R_1^{(0)}$ and calculate $\widetilde{R}_1^{(0)} = (\widehat{U}_1^{(0)})^\top R_1^{(0)} \in \mathbb{R}^{r_1 \times (p_2 \dots p_d)}$. Then, $\widehat{U}_1^{(0)}$ is an initial estimate of the subspace that G_1 lies in and $\widetilde{R}_1^{(0)}$ can be seen as the projection residual.

(ii) Next, we realign the entries of $\widetilde{R}_1^{(0)} \in \mathbb{R}^{r_1 \times (p_2 \dots p_d)}$ to $R_2^{(0)} \mathbb{R}^{(r_1 p_2) \times (p_3 \dots p_d)}$, where the rows and columns of $R_2^{(0)}$ correspond to indices of Modes-1, 2 and Modes-3, $\dots$, d, respectively. Then, we evaluate the top-r_2 SVD of $R_2^{(0)}$. Let $\widehat{U}_2^{(0)}$ be the first r_2 left singular vectors of $R_2^{(0)}$ and evaluate $\widetilde{R}_2^{(0)} = (\widehat{U}_2^{(0)})^\top R_2^{(0)} \in \mathbb{R}^{r_2 \times p_3 \dots p_d}$. Again, $\widehat{U}_2^{(0)}$ is an estimate of the singular subspace that $\mathcal{G}_2$ lies on and $\widetilde{R}_2^{(0)}$ is the projection residual for the next calculation.

(iii) We apply Step (ii) on $\widetilde{R}_2^{(0)}$ to obtain $\widehat{U}_3^{(0)} \in \mathbb{O}_{r_2 p_3, r_3}$ and $\widetilde{R}_3^{(0)} \in \mathbb{R}^{r_3 \times (p_4 \cdots p_d)}$; $\dots$; apply Step (ii) on $\widetilde{R}_{d-2}^{(0)}$ to obtain $\widehat{U}_{d-1}^{(0)} \in \mathbb{O}_{r_{d-2} p_{d-1}, r_{d-1}}$ and $\widetilde{R}_{d-1}^{(0)} \in \mathbb{R}^{r_{d-1} \times p_d}$. Then we reshape matrix $\widehat{U}_k^{(0)} \in \mathbb{R}^{(p_k r_{k-1}) \times r_k}$ to tensor $\widehat{\mathcal{U}}_k^{(0)} \in \mathbb{R}^{r_{k-1} \times p_k \times r_k}$ for $k = 2, \dots, d-1$. Now, $\left(\widehat{U}_1^{(0)}, \widehat{\mathcal{U}}_2^{(0)}, \dots, \widehat{\mathcal{U}}_{d-1}^{(0)}, \widetilde{R}_{d-1}^{(0)\top}\right)$ yield the initial estimates of TT-cores of $\mathcal{X}$ and we expect that

$$\mathcal{X} \approx \mathcal{X}^{(0)} = [\![\widehat{U}_1^{(0)}, \widehat{\mathcal{U}}_2^{(0)}, \cdots, \widehat{\mathcal{U}}_{d-1}^{(0)} \widetilde{R}_{d-1}^{(0)}]\!].$$

The initialization step is summarized to Algorithm 1(a) and illustrated in Figure 4.2. In summary, we perform SVD on some "residual" $R_k^{(0)}$ sequentially for $k = 1, \dots, d-1$. As will be shown in Lemma 4.3.3, $R_k^{(0)}$ satisfies

$$R_k^{(0)} = (I_{p_k} \otimes \widehat{U}_{k-1}^{(0)\top}) \cdots (I_{p_2 \cdots p_k} \otimes \widehat{U}_1^{(0)\top})[\mathcal{Y}]_k,$$

where $[\mathcal{Y}]_k \in \mathbb{R}^{(p_1 \cdots p_k) \times (p_{k+1} \cdots p_d)}$ is the kth sequential unfolding of $\mathcal{Y}$ (see definition in Section 4.2.1). This quantity plays a key role in the backward update next.

- **Part 2: Backward update.** For iterations $t = 1, 3, 5, \dots$, we perform backward

Algorithm 1(a) Initialization (TT-SVD (Oseledets, 2011))

```
Input: $\mathcal{Y}, \{r_k\}_{k=1}^{d-1}, \{p_k\}_{k=1}^{d}$
Calculate $R_1^{(0)} = [\mathcal{Y}]_1$
for $k = 1, \ldots, d-1$ do
    $\widehat{U}_k^{(0)} = \mathrm{SVD}_{r_k}^{L}(R_k^{(0)})$
    If $k = 1$ then $U_{\mathrm{prod}}^{(0)} = \widehat{U}_k^{(0)}$ else $U_{\mathrm{prod}}^{(0)} = (I_{p_k} \otimes U_{\mathrm{prod}}^{(0)})\widehat{U}_k^{(0)}$
    $\widetilde{R}_k^{(0)} = \widehat{U}_k^{(0)\top} R_k^{(0)}$
    If $k < d-1$ then $R_{k+1}^{(0)} = \mathrm{reshape}(\widetilde{R}_k^{(0)}, r_k p_{k+1}, p_{k+2} \cdots p_d)$
end for
$[\widehat{X}^{(0)}]_{d-1} = U_{\mathrm{prod}}^{(0)} \widetilde{R}_{d-1}^{(0)}$
Reshape $[\widehat{X}^{(0)}]_{d-1} \in \mathbb{R}^{(p_1 \cdots p_{d-1}) \times p_d}$ to $\widehat{\mathcal{X}}^{(0)} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$
Output: $\widetilde{R}_1^{(0)}, \ldots, \widetilde{R}_{d-1}^{(0)}, \widehat{\mathcal{X}}^{(0)}$
```

Algorithm 1(b) TT-Backward Update

```
Input: $\mathcal{Y}, \{r_k\}_{k=1}^{d-1}, \{p_k\}_{k=1}^{d}, \widetilde{R}_1^{(t-1)}, \ldots, \widetilde{R}_{d-1}^{(t-1)}$ for odd iteration number t
for $k = 1, \ldots, d-1$ do
    if $k = 1$ then
        $\widehat{V}_{d-k+1}^{(t)} = \mathrm{SVD}_{r_{d-k}}^{R}\left(\widetilde{R}_{d-k}^{(t-1)}\right), \quad V_{\mathrm{prod}}^{(t)} = \widehat{V}_{d-k+1}^{(t)}$
    else
        $\widehat{V}_{d-k+1}^{(t)} = \mathrm{SVD}_{r_{d-k}}^{R}\left(\widetilde{R}_{d-k}^{(t-1)}(V_{\mathrm{prod}}^{(t)} \otimes I_{p_{d-k+1}})\right), \quad V_{\mathrm{prod}}^{(t)} = (V_{\mathrm{prod}}^{(t)} \otimes I_{p_{d-k+1}})\widehat{V}_{d-k+1}^{(t)}$
    end if
end for
$\widehat{V}_1^{(t)} = [\mathcal{Y}]_1 V_{\mathrm{prod}}^{(t)}, \quad [\widehat{X}^{(t)}]_1 = \widehat{V}_1^{(t)} V_{\mathrm{prod}}^{(t)\top}, \quad$ reshape $[\widehat{X}^{(t)}]_1 \in \mathbb{R}^{p_1 \times (p_2 \cdots p_d)}$ to $\widehat{\mathcal{X}}^{(t)} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$
Output: $\widehat{V}_1^{(t)}, \ldots, \widehat{V}_d^{(t)}, \widehat{\mathcal{X}}^{(t)}$
```

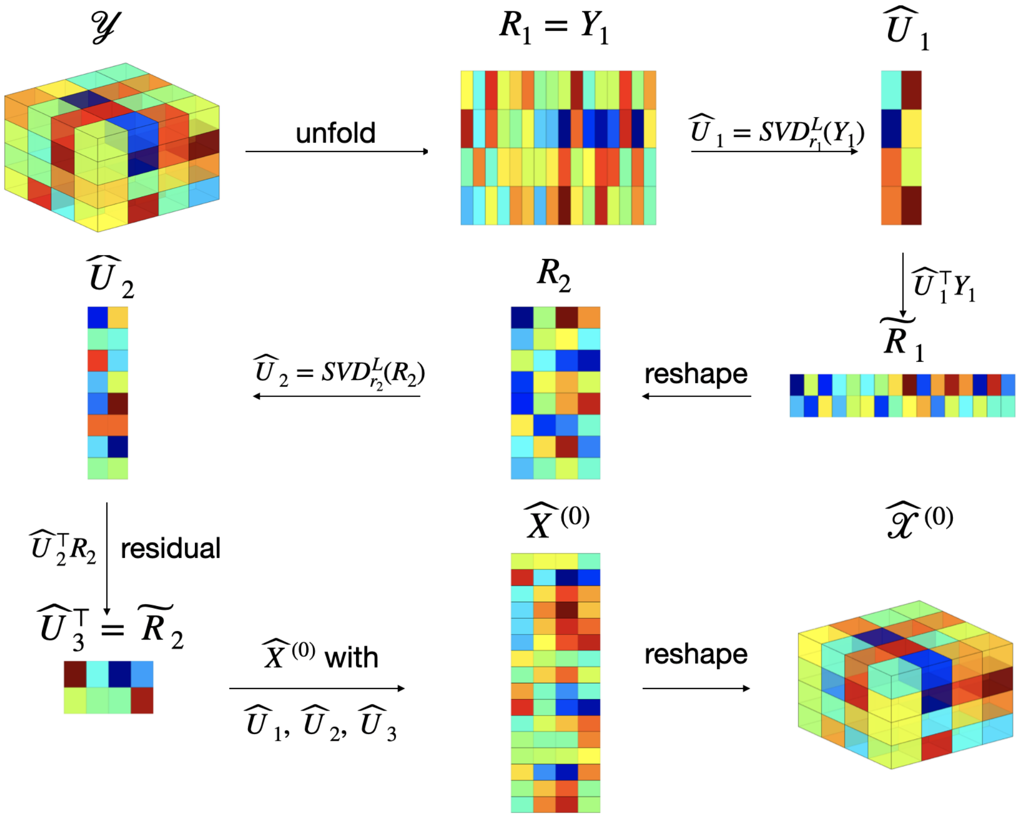


Figure 4.2: A Pictorial Illustration of Initialization (Algorithm 1(a), $d = 3$)

update, i.e., to sequentially obtain $\widehat{V}_d^{(t)}, \dots, \widehat{V}_2^{(t)}$ based on the intermediate results from the $(t-1)$st iteration (0th iteration is the initialization). The pseudocode of backward update is provided in Algorithm 1(b). The calculation in Algorithm 1(c) is equivalent to

$$\widehat{V}_d^{(t)} = \mathrm{SVD}^R\left(\widetilde{R}_{d-1}^{(t-1)}\right),$$

$$\widehat{V}_k^{(t)} = \mathrm{SVD}^R\left(\widetilde{R}_{k-1}^{(t-1)}(\widehat{V}_d^{(t)} \otimes I_{p_k \cdots p_{d-1}}) \cdots (\widehat{V}_{k+1}^{(t)} \otimes I_{p_k})\right), \quad k = d-1, \dots, 2,$$

and

$$\widehat{V}_1^{(t)} = [\mathcal{Y}]_1 (\widehat{V}_d^{(t)} \otimes I_{p_2 \cdots p_{d-1}}) \cdots (\widehat{V}_3^{(t)} \otimes I_{p_2}) \widehat{V}_2^{(t)} \in \mathbb{R}^{p_1 \times r_1}.$$

Algorithm 1(c) TT-Forward Update

```
Input: $\mathcal{Y}, \{r_k\}_{k=1}^{d-1}, \{p_k\}_{k=1}^{d}, \widehat{V}_1^{(t-1)}, \ldots, \widehat{V}_d^{(t-1)}$ for even iteration number $t$
$R_1^{(t)} = [\mathcal{Y}]_1$
for $k = 1, \ldots, d-1$ do
    if $k = 1$ then
        $\widehat{U}_1^{(t)} = \mathrm{SVD}_{r_1}^{L}\left(\widehat{V}_1^{(t)}\right), \quad U_{\mathrm{prod}}^{(t)} = \widehat{U}_1^{(t)}$
    else
        $\widehat{U}_k^{(t)} = \mathrm{SVD}_{r_k}^{R}\left(R_k^{(t)}(\widehat{V}_d^{(t-1)} \otimes I_{p_{k+1} \ldots p_{d-1}}) \cdots (\widehat{V}_{k+2}^{(t-1)} \otimes I_{p_{k+1}})\widehat{V}_{k+1}^{(t-1)}\right)$
        $U_{\mathrm{prod}}^{(t)} = (I_{p_k} \otimes U_{\mathrm{prod}}^{(t)})\widehat{U}_k^{(t)}$
    end if
    $\widetilde{R}_k^{(t)} = \widehat{U}_k^{(t)\top} R_k^{(t)}$
    If $k < d-1$ then $R_{k+1}^{(t)} = \mathrm{reshape}\left(\widetilde{R}_k^{(t)}, r_k p_{k+1}, p_{k+2} \cdots p_d\right)$
end for
$[\widehat{X}^{(t)}]_{d-1} = U_{\mathrm{prod}}^{(t)} \widetilde{R}_{d-1}^{(0)}$,   reshape $[\widehat{X}^{(t)}]_{d-1} \in \mathbb{R}^{(p_1 \cdots p_{d-1}) \times p_d}$ to $\widehat{\mathcal{X}}^{(t)} \in \mathbb{R}^{p_1 \times \cdots \times p_d}$
Output: $\widetilde{R}_1^{(t)}, \ldots, \widetilde{R}_{d-1}^{(t)}, \widehat{\mathcal{X}}^{(t)}$
```

Here,

$$\widetilde{R}_k^{(t-1)} = (\widehat{U}_k^{(t-1)})^\top (I_{p_k} \otimes \widehat{U}_{k-1}^{(t-1)\top}) \cdots (I_{p_2 \cdots p_k} \otimes \widehat{U}_1^{(t-1)\top})[\mathcal{Y}]_k$$

are the projection residual term in the intermediate outcome of the $(t-1)$st iteration. Then, we reshape $\widehat{V}_k^{(t)\top} \in \mathbb{R}^{r_{k-1} \times (p_k r_k)}$ to $\mathcal{V}_k^{(t)} \in \mathbb{R}^{r_{k-1} \times p_k \times r_k}$. The backward updated estimate is

$$\widehat{\mathcal{X}}^{(t)} = [\![\widehat{V}_1^{(t)}, \widehat{\mathcal{V}}_2^{(t)}, \ldots, \widehat{\mathcal{V}}_{d-1}^{(t)}, \widehat{V}_d^{(t)}]\!].$$

Remark 4.2.1 (Interpretation of backward update)**.** *The backward updates utilize and extract the right singular vectors of the intermediate products of the* $(t-1)$*st iteration,*

$$\widetilde{R}_k^{(t-1)} = (\widehat{U}_k^{(t-1)})^\top (I_{p_k} \otimes \widehat{U}_{k-1}^{(t-1)\top}) \cdots (I_{p_2 \cdots p_k} \otimes \widehat{U}_1^{(t-1)\top})[\mathcal{Y}]_k,$$

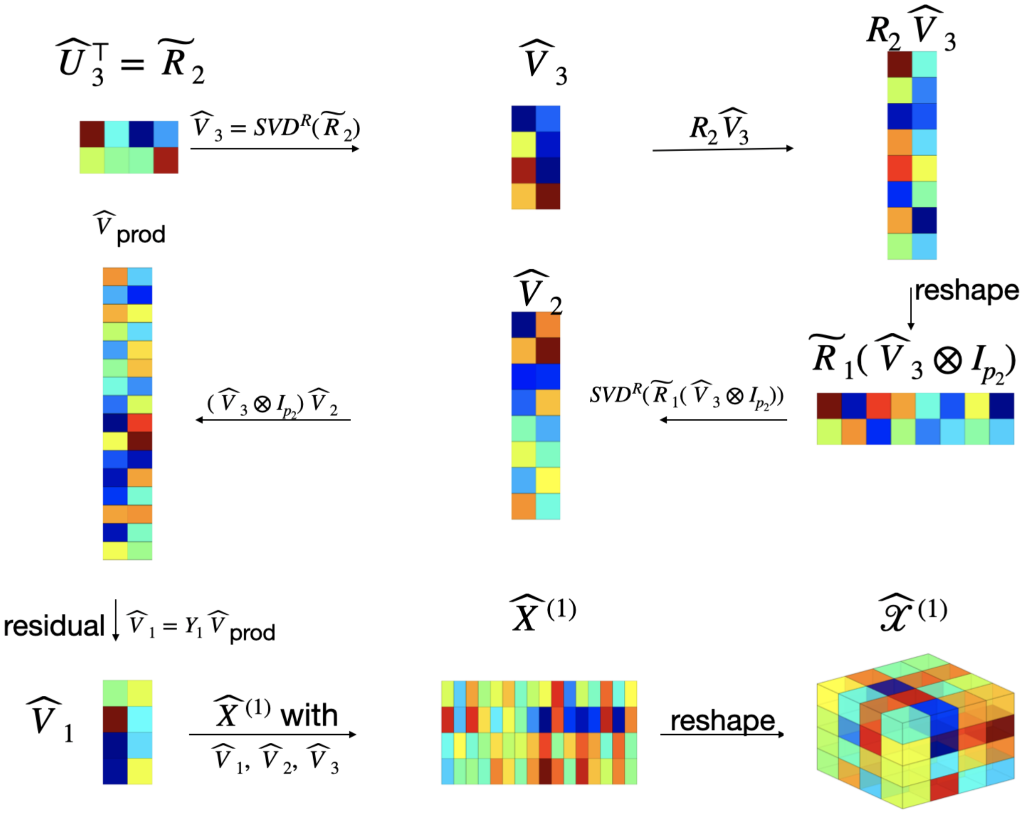

$\widehat{U}_3^\top = \widetilde{R}_2$
$\widehat{V}_3 = SVD^R(\widetilde{R}_2)$
$\widehat{V}_3$
$R_2\widehat{V}_3$
$R_2\widehat{V}_3$
reshape
$\widetilde{R}_1(\widehat{V}_3 \otimes I_{p_2})$
$\widehat{V}$ prod
$\widehat{V}_2$
$(\widehat{V}_3 \otimes I_{p_2})\widehat{V}_2$
$SVD^R(\widetilde{R}_1(\widehat{V}_3 \otimes I_{p_2}))$
residual
$\widehat{V}_1 = Y_1\widehat{V}$ prod
$\widehat{X}^{(1)}$
$\widehat{\mathcal{X}}^{(1)}$
$\widehat{V}_1$
$\widehat{X}^{(1)}$ with
$\widehat{V}_1, \widehat{V}_2, \widehat{V}_3$
reshape

Figure 4.3: A pictorial illustration of TT-Backward update (Algorithm 1(b), $d = 3$)

as opposed to the entire data $[\mathcal{Y}]_k$*. Such a dimension reduction scheme is the key to the backward update: it can simultaneously reduce the dimension of the matrix of interest,* $[\mathcal{Y}]_k$*, and the noise therein, while preserving the signal strength. Different from the initialization in Step 1, the backward update utilizes the information from both the forward and backward singular subspaces of the tensor-train structure of* $\mathcal{X}$*. See Section 4.3 for more illustration.*

- **Part 3: Forward Update.** For iteration $t = 2, 4, 6, \dots$, we perform forward update, i.e., to sequentially obtain $\widehat{U}_1^{(t)}, \dots, \widehat{U}_d^{(t)}$ based on the intermediate results from the $(t-1)$st iteration. Essentially, the forward update can be seen as a reversion of the backward update by flipping all modes of tensor $\mathcal{Y}$. The pseudocode of this procedure is collected in Algorithm 1(c). Recall

$[\mathcal{Y}]_1(\widehat{V}_d^{(t-1)} \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_3^{(t-1)} \otimes I_{p_2})\widehat{V}_2^{(t-1)}$ is the intermediate product from the $(t-1)$st update. We sequentially compute

$$\widehat{U}_1^{(t)} = \mathrm{SVD}^L \left([\mathcal{Y}]_1(\widehat{V}_d^{(t-1)} \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_3^{(t-1)} \otimes I_{p_2})\widehat{V}_2^{(t-1)} \right);$$

$$\begin{aligned}\widehat{U}_k^{(t)} = \mathrm{SVD}^L \Big(&(I_{p_k} \otimes \widehat{U}_{k-1}^{(t)\top}) \cdots (I_{p_2 \cdots p_k} \otimes \widehat{U}_1^{(t)\top})[\mathcal{Y}]_k \\ &\cdot (\widehat{V}_d^{(t-1)} \otimes I_{p_{k+1} \dots p_{d-1}}) \cdots (\widehat{V}_{k+2}^{(t-1)} \otimes I_{p_{k+1}})\widehat{V}_{k+1}^{(t-1)} \Big)\end{aligned}$$

for $k = 2, \dots, d-1$, and

$$\widehat{U}_d^{(t)} = [(\widehat{U}_{d-1}^{(t)})^\top (I_{p_{d-1}} \otimes (\widehat{U}_{d-2}^{(t)})^\top) \cdots (I_{p_{d-1} \dots p_2} \otimes (\widehat{U}_1^{(t)})^\top)[\mathcal{Y}]_{d-1}]^\top \in \mathbb{R}^{p_d \times r_{d-1}}.$$

Reshape $\widehat{U}_k^{(t)} \in \mathbb{R}^{(p_k r_{k-1}) \times r_k}$ to $\widehat{\mathcal{U}}_k^{(t)} \in \mathbb{R}^{r_{k-1} \times p_k \times r_k}$ for $k = 2, \dots, d-1$. Then, compute

$$\widehat{\mathcal{X}}^{(t)} = \llbracket \widehat{G}_1^{(t)}, \widehat{\mathcal{G}}_2^{(t)}, \dots, \widehat{\mathcal{G}}_{d-1}^{(t)}, \widehat{G}_d^{(t)} \rrbracket.$$

We will explain the algebraic schemes in the TTOI procedure through several representation lemmas in Section 4.3.1. We will also show in Theorem 4.2 that the objective function $\|\mathcal{Y} - \widehat{\mathcal{X}}^{(t)}\|_F^2$ is monotone decreasing with respect to the iteration index t. In the large-scale scenarios that performing iterations is beyond the capacity of computing, we can reduce the number of iterations, and even to $t_{\max} = 1$, i.e., the one-step iteration, which have often yielded sufficiently accurate estimation as we will illustrate in both theory and simulation studies. Such the phenomenon has been recently discovered for HOOI in the Tucker low-rank tensor decomposition (Luo et al., 2020).

Remark 4.2.2 (Computational and storage costs of TTOI)**.** *We consider the computational and storage costs of TTOI on the* p*-dimensional, rank-*r*, order-*d*, and dense tensor. Since computing the first* r *singular vectors of an* $m \times n$ *matrix via block power method requires* $\widetilde{O}(mnr)$ *operations, initialization costs* $\widetilde{O}(p^d r)$ *operations, each iteration of TTOI,* *including forward and backward updates, costs* $O(p^d r)$. *Therefore, the total number of operations of TTOI with* T *iterations is* $\widetilde{O}(p^d r) + O(Tp^d r)$, *which is not significantly more than the number of elements of the target tensor. Moreover, TTOI requires* $O(p^d)$ *storage cost, which is not significantly more than the storage cost of the original tensor.*

4.3 Theoretical Analysis

This section is devoted to the theoretical analysis of the proposed procedure. For convenience, we introduce the following two abbreviations for matrix sequential products: for $M_i \in \mathbb{R}^{(p_i r_{i-1}) \times r_i}, 1 \leqslant i \leqslant d-1$ and $B_j \in \mathbb{R}^{(r_j p_j) \times r_{j-1}}, 2 \leqslant j \leqslant d$, we denote

$$M^{(L)}_{\text{prod},k} = (I_{p_2 \cdots p_k} \otimes M_1) \cdots (I_{p_k} \otimes M_{k-1}) M_k \in \mathbb{R}^{(p_1 \cdots p_k) \times r_k}, \quad \forall 1 \leqslant k \leqslant d-1,$$

$$B^{(R)}_{\text{prod},k} = (B_d \otimes I_{p_k \cdots p_{d-1}}) \cdots (B_{k+1} \otimes I_{p_k}) B_k \in \mathbb{R}^{(p_k \cdots p_d) \times r_{k-1}}, \quad \forall 2 \leqslant k \leqslant d.$$

Equivalently, $M^{(L)}_{\text{prod},k}$ and $B^{(R)}_{\text{prod},k}$ can be defined sequentially as

$$M^{(L)}_{\text{prod},1} = M_1, \quad M^{(L)}_{\text{prod},k+1} = (I_{p_{k+1}} \otimes M^{(L)}_{\text{prod},k}) M_{k+1}, \quad 1 \leqslant k \leqslant d-2,$$

$$B^{(R)}_{\text{prod},d} = B_d, \quad B^{(R)}_{\text{prod},k} = (B^{(R)}_{\text{prod},k+1} \otimes I_{p_k}) B_k, \quad 2 \leqslant k \leqslant d-1.$$

4.3.1 Representation Lemmas for high-order tensors

Since the computation of high-order tensors with tensor-train structures involves extensive tensor algebra, we introduce the following three lemmas on the matrix representation of high-order tensors. These lemmas play a fundamental role in the later theoretical analysis.

Lemma 4.3.1 (Representation for sequential matricization of TT-decomposable tensor). *Suppose* $\mathcal{X} = [\![G_1, \mathcal{G}_2, \ldots, \mathcal{G}_{d-1}, G_d]\!]$. *Then the sequential matricization of* $\mathcal{X}$

can be written as

$$
\begin{aligned}
[\mathcal{X}]_k =& (I_{p_2\cdots p_k} \otimes G_1)(I_{p_3\cdots p_k} \otimes [\mathcal{G}_2]_2)\cdots(I_{p_k} \otimes [\mathcal{G}_{k-1}]_2)\,[\mathcal{G}_k]_2\,[\mathcal{G}_{k+1}]_1 \\
&\cdot \left([\mathcal{G}_{k+2}]_1 \otimes I_{p_{k+1}}\right)\cdots\left([\mathcal{G}_{d-1}]_1 \otimes I_{p_{k+1}\cdots p_{d-2}}\right)\left(G_d^\top \otimes I_{p_{k+1}\cdots p_{d-1}}\right).
\end{aligned}
\tag{4.4}
$$

Lemma 4.3.2 (Representation of tensor reshaping). *For any tensor* $\mathcal{T} \in \mathbb{R}^{\otimes_{k=1}^d p_k}$ *and* $1 \leqslant i < j \leqslant d-1$, *we have*

$$
[\mathcal{T}]_j = (I_{p_{i+1}\cdots p_j} \otimes [\mathcal{T}]_i)A^{(p_{i+1}\cdots p_j, p_{j+1}\cdots p_d)}, \quad [\mathcal{T}]_i = A^{(p_{i+1}\cdots p_j, p_1\cdots p_i)\top}([\mathcal{T}]_j \otimes I_{p_{i+1}\cdots p_j}).
$$

Here, we define $e_k^{(ij)}$ *as the* k*th canonical basis of* $\mathbb{R}^{ij}$ *and*

$$
A^{(i,j)} = \begin{bmatrix} e_1^{(ij)} & e_{i+1}^{(ij)} & \cdots & e_{i(j-1)+1}^{(ij)} \\ e_2^{(ij)} & e_{i+2}^{(ij)} & \cdots & e_{i(j-1)+2}^{(ij)} \\ \vdots & \vdots & \ddots & \vdots \\ e_i^{(ij)} & e_{2i}^{(ij)} & \cdots & e_{ij}^{(ij)} \end{bmatrix} \in \mathbb{R}^{(i^2 j)\times j}. \tag{4.5}
$$

Lemmas 4.3.1 and 4.3.2 can be proved by checking each entry of the corresponding matricizations. In addition, the following lemma provides a representation of sequential reshaping tensor, in particular for $R_k^{(t)}$ and $\tilde{R}_k^{(t)}$, the key intermediate outcomes in TTOI procedure.

Lemma 4.3.3 (Representation of sequential reshaping tensor). *Suppose* $\mathcal{T} \in \mathbb{R}^{\otimes_{k=1}^d p_k}$, $M_i \in \mathbb{R}^{(r_{i-1}p_i)\times r_i}$ *for* $1 \leqslant i \leqslant d-1$, $B_i \in \mathbb{R}^{(p_i r_i)\times r_{i-1}}$ *for* $2 \leqslant i \leqslant d$, *where* $r_0 = r_d = 1$. *Consider the following sequential multiplication:*

Forward sequential multiplication: *Let* $S_1 = [\mathcal{T}]_1$. *For* $k = 1, \ldots, d-1$, *calculate*

$$
\begin{aligned}
&\widetilde{S}_k = M_k^\top S_k \in \mathbb{R}^{r_k \times (p_{k+1}\cdots p_d)}, \\
&S_{k+1} = \textit{Reshape}(\widetilde{S}_k, r_k p_{k+1}, p_{k+2}\cdots p_d) \quad \textit{if } k < d-1.
\end{aligned}
$$

Then for any $1 \leqslant k \leqslant d-1$,

$$S_k = (I_{p_k} \otimes M^{(L)\top}_{\mathit{prod},k-1})[\mathcal{T}]_k, \quad \widetilde{S}_k = M^{(L)\top}_{\mathit{prod},k}[\mathcal{T}]_k. \tag{4.6}$$

Here, $I_{p_k} \otimes M^{(L)\top}_{\mathit{prod},k-1} = I_{p_1}$ *if* $k = 1$.

Backward sequential multiplication: *Let* $W_{d-1} = [\mathcal{T}]_{d-1}$. *For* $k = d-1, \ldots, 1$, *calculate*

$$\begin{aligned} &\widetilde{W}_k = W_k B_{k+1} \in \mathbb{R}^{(p_1 \cdots p_k) \times r_k}, \\ &W_{k-1} = \mathit{Reshape}(\widetilde{W}_k, p_1 \cdots p_{k-1}, p_k r_k) \quad \mathit{if}\ k > 1. \end{aligned}$$

Then for any $1 \leqslant k \leqslant d-1$,

$$W_k = [\mathcal{T}]_k (B^{(R)}_{\mathit{prod},k+2} \otimes I_{p_{k+1}}), \quad \widetilde{W}_k = [\mathcal{T}]_k B^{(R)}_{\mathit{prod},k+1}.$$

Here, $B^{(R)}_{\mathit{prod},k+2} \otimes I_{p_{k+1}} = I_{p_d}$ *if* $k = d-1$.

In particular, $R_k^{(0)}, \widetilde{R}_k^{(0)}$ *in Algorithm 1(a) and* $R_k^{(t)}, \widetilde{R}_k^{(t)} (t \in \{2,4,6,\ldots\})$ *in Algorithm 1(c) satisfy*

$$R_k^{(t)} = \left(I_{p_k} \otimes (\widehat{U}^{(t)})^{(L)\top}_{\mathit{prod},k-1}\right) [\mathcal{Y}]_k, \quad \widetilde{R}_k^{(t)} = (\widehat{U}^{(t)})^{(L)\top}_{\mathit{prod},k} [\mathcal{Y}]_k, \quad \forall 1 \leqslant k \leqslant d-1. \tag{4.7}$$

The proof of Lemma 4.3.3 is provided in Section 5.3.8.

4.3.2 Deterministic Upper Bounds for Estimation Error of TTOI

Now we are in position to analyze the performance of TTOI. The following Theorem 4.1 introduces an upper bound on estimation error of $\widehat{\mathcal{X}}^{(2t+1)}$ (backward update) and $\widehat{\mathcal{X}}^{(2t+2)}$ (forward update).

Theorem 4.1. *Suppose we observe* $\mathcal{Y} = \mathcal{X} + \mathcal{Z}$, *where* $\mathcal{X}$ *admits a TT decomposition as* (4.1).

(A deterministic estimation error bound for backward updates) *Let* $\widetilde{U}_1^{(2t)} = U_1 \in \mathbb{R}^{p_1 \times r_1}$ *be the left singular space of* $[\mathcal{X}]_1$*. For* $2 \leqslant k \leqslant d-1$*, define* $\widetilde{U}_k^{(2t)} \in \mathbb{R}^{p_k r_{k-1} \times r_k}$ *as the left singular subspace of* $\left(I_{p_k} \otimes (\widehat{U}^{(2t)})^{(L)\top}_{\mathit{prod},k-1}\right)[\mathcal{X}]_k$*. If for some constant* $c_0 \in (0,1)$*,*

$$\left\|\sin\Theta\left(\widehat{U}_k^{(2t)}, \widetilde{U}_k^{(2t)}\right)\right\| \leqslant c_0, \quad \forall 1 \leqslant k \leqslant d-1, \tag{4.8}$$

then there exists a constant $C > 0$ *such that the outcome of Algorithm 1(b) satisfies*

$$\left\|\widehat{\mathcal{X}}^{(2t+1)} - \mathcal{X}\right\|_F^2 \leqslant C\left(\sum_{k=1}^{d-1} A_k^{(2t+1)} + B^{(2t+1)}\right), \tag{4.9}$$

where

$$A_k^{(2t+1)} = \left\|(\widehat{U}^{(2t)})^{(L)\top}_{\mathit{prod},k}[\mathcal{Z}]_k\left((\widehat{V}^{(2t+1)})^{(R)}_{\mathit{prod},k+2} \otimes I_{p_{k+1}}\right)\right\|_F^2,$$

$$B^{(2t+1)} = \left\|[\mathcal{Z}]_1(\widehat{V}^{(2t+1)})^{(R)}_{\mathit{prod},2}\right\|_F^2.$$

Here, $(\widehat{V}^{(2t+1)})^{(R)}_{\mathit{prod},k+2} \otimes I_{p_{k+1}} = I_{p_d}$ *if* $k = d-1$.

(A deterministic estimation error bound for forward updates) *For* $2 \leqslant k \leqslant d-1$*, let* $\widetilde{V}_k^{(2t+1)} \in \mathbb{R}^{(p_k r_k) \times r_{k-1}}$ *be the right singular space of* $[\mathcal{X}]_{k-1}\left((\widehat{V}^{(2t+1)})^{(R)}_{\mathit{prod},k+1} \otimes I_{p_k}\right)$ *and let* $\widetilde{V}_d^{(2t+1)} = V_d \in \mathbb{R}^{p_d \times r_{d-1}}$ *be the right singular space of* $[\mathcal{X}]_{d-1}$*. If for some constant* $c_0 \in (0,1)$*,*

$$\left\|\sin\Theta\left(\widehat{V}_k^{(2t+1)}, \widetilde{V}_k^{(2t+1)}\right)\right\| \leqslant c_0, \quad \forall 2 \leqslant k \leqslant d,$$

then there exists a constant $C > 0$ *such that the outcome of Algorithm 1(c) satisfies*

$$\left\|\widehat{\mathcal{X}}^{(2t+2)} - \mathcal{X}\right\|_F^2 \leqslant C\left(\sum_{k=1}^{d-1} A_k^{(2t+2)} + B^{(2t+2)}\right), \tag{4.10}$$

where

$$A_k^{(2t+2)} = \left\|\left(I_{p_k} \otimes (\widehat{U}^{(2t+2)})^{(L)\top}_{\mathit{prod},k-1}\right)[\mathcal{Z}]_k(\widehat{V}^{(2t+1)})^{(R)}_{\mathit{prod},k+1}\right\|_F^2,$$

$$B^{(2t+2)} = \left\| (\widehat{U}^{(2t+2)})^{(L)\top}_{\textit{prod},d-1} [\mathcal{Z}]_{d-1} \right\|_F^2 .$$

Here, $I_{p_k} \otimes (\widehat{U}^{(2t+2)})^{(L)\top}_{\textit{prod},k-1} = I_{p_1}$ *if* $k = 1$.

The proof of Theorem 4.1 is provided in Section 5.3.1. Theorem 4.1 shows the estimation error $\|\widehat{\mathcal{X}}^{(t+1)} - \mathcal{X}\|_F^2$ can be bounded by the projected noise $\mathcal{Z}$, i.e., $A_k^{(t+1)}$ and $B^{(t+1)}$, if the estimates in initialization $(t = 0)$ or the previous iteration $(t \geqslant 1)$, $\{\widehat{U}_k^{(t)}\}_{k=1}^{d-1}$ or $\{\widehat{V}_k^{(t)}\}_{k=2}^{d}$, are within constant distance to the true underlying subspaces. The developed upper bound can be significantly smaller than $C\,\|\mathcal{Z}\|_F^2$, the classic upper bound induced from the approximation error (e.g., Theorem 2.2 in Oseledets (2011)), especially in the high-dimensional setting $(p \gg r)$.

Remark 4.3.1 (Interpretation of error bounds in Theorem 4.1)**.** *Here, we provide some explanation for* $A_k^{(2t+1)}$ *and* $B^{(2t+1)}$ *in the error bound* (4.9)*. By algebraic calculation, the TT-core estimation via backward update can be written as*

$$\widehat{V}_{k+1}^{(2t+1)} = \mathrm{SVD}^R\Big\{ (\widehat{U}^{(2t)})^{(L)\top}_{\textit{prod},k} ([\mathcal{X}]_k + [\mathcal{Z}]_k) \Big((\widehat{V}^{(2t+1)})^{(R)}_{\textit{prod},k+2} \otimes I_{p_{k+1}} \Big) \Big\}, \quad \forall 1 \leqslant k \leqslant d-1$$

and

$$\widehat{V}_1^{(2t+1)} = ([\mathcal{X}]_1 + [\mathcal{Z}]_1)(\widehat{V}^{(2t+1)})^{(R)}_{\textit{prod},2}.$$

From the definition of $A_k^{(2t+1)}$*, we have see* $A_k^{(2t+1)}$ *quantifies the error of the singular subspace estimate* $\widehat{V}_{k+1}^{(2t+1)}$ *and* $B^{(2t+1)}$ *quantifies the error of the projected residual* $\widehat{V}_1^{(2t+1)}$*. By symmetry, similar interpretation also applies to* $A_k^{(2t+2)}$ *and* $B^{(2t+2)}$ *for the error bound of forward update* (4.10)*.*

Remark 4.3.2 (Proof Sketch of Theorem 4.1)**.** *While the complete proof of Theorem 4.1 is provided in Section 5.3.1, we provide a brief proof sketch here.*

Without loss of generality, we focus on (4.9) *for* $t = 0$ *while other cases follows similarly. For convenience, we simply let* $\widehat{U}_i, \widehat{V}_i$ *denote* $\widehat{U}_i^{(0)}, \widehat{V}_i^{(1)}$*, respectively. First, by Lemma 4.3.1, we can transform* $[\widehat{X}^{(1)}]_1$*, the outcome of backward update, to*

$$[\widehat{X}^{(1)}]_1 = [\mathcal{Y}]_1 P_{(\widehat{V}_d \otimes I_{p_2 \cdots p_{d-1}}) \cdots (\widehat{V}_3 \otimes I_{p_2}) \widehat{V}_2}.$$

Then we can further bound the estimation error of $\widehat{\mathcal{X}}^{(1)}$ *as*

$$\begin{aligned}\|\widehat{\mathcal{X}}^{(1)} - \mathcal{X}\|_F^2 \leqslant & C\left\|[\mathcal{Z}]_1(\widehat{V}_d \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_3 \otimes I_{p_2})\widehat{V}_2\right\|_F^2 \\ & + C\sum_{k=2}^{d}\left\|[\mathcal{X}]_1(\widehat{V}_d \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_2 \cdots p_k})(\widehat{V}_{k\perp} \otimes I_{p_2 \cdots p_{k-1}})\right\|_F^2 .\end{aligned}$$

Next, based on Lemma 4.3.2 and (4.8)*, we can prove*

$$\begin{aligned}&\left\|[\mathcal{X}]_1(\widehat{V}_d \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_2 \cdots p_k})(\widehat{V}_{k\perp} \otimes I_{p_2 \cdots p_{k-1}})\right\|_F \\ = & \left\|[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_F \\ \leqslant & C\left\|\widehat{U}_{k-1}^\top(I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_F .\end{aligned}$$

Finally, we apply the perturbation projection error bound (Lemma 5.3.3) to prove that

$$\begin{aligned}& C\left\|\widehat{U}_{k-1}^\top(I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_F \\ \leqslant & C\left\|\widehat{U}_{k-1}^\top(I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{Z}]_{k-1}(\widehat{V}_d \otimes I_{p_k \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\right\|_F .\end{aligned}$$

Theorem (4.4.1) *is proved by combing all inequalities above.*

Next, we establish a decomposition formula for the approximation error, i.e., the objective function in (4.3) $\left\|\mathcal{Y} - \mathcal{X}^{(t)}\right\|_F^2$, and show that the approximation error is monotone decreasing through TTOI iterations.

Theorem 4.2 (Approximation error decays through iterations)**.** *We implement TTOI on* $\mathcal{Y}$*. Let* $\widehat{\mathcal{X}}^{(t)}$ *be the outcome after the* t*th iteration. For any* $k \geqslant 1$*, we have*

$$\textit{(Approximation error decay)} \quad \|\mathcal{Y}\|_F^2 - \|\widehat{\mathcal{X}}^{(t+1)}\|_F^2 \leqslant \|\mathcal{Y}\|_F^2 - \|\widehat{\mathcal{X}}^{(t)}\|_F^2, \tag{4.11}$$

$$\textit{(Approximation error decomposition)} \quad \|\mathcal{Y} - \widehat{\mathcal{X}}^{(k+1)}\|_F^2 = \|\mathcal{Y}\|_F^2 - \|\widehat{\mathcal{X}}^{(k+1)}\|_F^2. \tag{4.12}$$

See Section 5.3.2 for the proof of Theorem 4.2.

4.4 TTOI for Tensor-Train Spiked Tensor Model

In this section, we further focus on a probabilistic setting, *spiked tensor model*, where the noise tensor $\mathcal{Z}$ has independent, mean zero, and σ-sub-Gaussian entries (see definition in Section 4.2.1). The spiked tensor model has been widely studied as a benchmark setting for tensor PCA/SVD and dimension reduction in recent literature in machine learning, information theory, statistics, and data science (Richard and Montanari, 2014; Lesieur et al., 2017; Zhang and Xia, 2018; Wein et al., 2019; Perry et al., 2020). The central goal therein is to discover the underlying low-rank tensor $\mathcal{X}$. Most of the existing works focused on tensors with Tucker or CP decomposition.

Under the spiked tensor model, we can verify that the initialization step of TTOI gives sufficiently good initial estimations with high probability that matches the required condition in Theorem 4.1.

Theorem 4.3 (Probabilistic bound for initial estimates and projected noise)**.** *Suppose* $\mathcal{X}$ *is TT-decomposable as* (4.1) *and* $\mathcal{Z}$ *have independent zero mean and* σ*-sub-Gaussian random variables. Denote* $p = \min\{p_1, \cdots, p_d\}$*. If there exists a constant* C_{gap} *such that* $\lambda_k = s_{r_k}([\mathcal{X}]_k) \geqslant C_{gap}\left((\sum_{i=1}^{d} p_i r_{i-1} r_i)^{1/2} + (p_{k+1}\cdots p_d)^{1/2}\right)\sigma$ *for* $1 \leqslant k \leqslant d-1$, *then there exist some constants* $C, c > 0$*, with probability at least* $1 - C\exp(-cp)$*,*

$$\max_{k=1,\dots,d-1}\left\|\sin\Theta\left(\widehat{U}_k^{(0)}, \widetilde{U}_k^{(0)}\right)\right\| \leqslant \frac{1}{2}, \tag{4.13}$$

$$\max_{\substack{k=1,\dots,d-1\\ t=2,4,6,\dots}}\left\|\sin\Theta\left(\widehat{U}_k^{(t)}, \widetilde{U}_k^{(t)}\right)\right\|, \max_{\substack{k=2,\dots,d\\ t=1,3,5,\dots}}\left\|\sin\Theta\left(\widehat{V}_k^{(t)}, \widetilde{V}_k^{(t)}\right)\right\| \leqslant \frac{1}{2}, \tag{4.14}$$

and for all $t \geqslant 1$,

$$\max\{A_k^{(t)}, B^{(t)}\} \leqslant C\sigma^2 \sum_{i=1}^{d} p_i r_i r_{r-1}. \tag{4.15}$$

Here, $\widetilde{U}_k^{(t)}$, $\widetilde{V}_k^{(t)}$, $A_k^{(t)}$ *and* $B^{(t)}$ *are defined in Theorem 4.1.*

The proof of Theorem 4.3 is provided in Section 5.3.3. Based on Theorems 4.1 and 4.3, we can further prove:

Corollary 4.4.1 (Upper bound for estimation error)**.** *Suppose* $\mathcal{X}$ *can be decomposed as* (4.1), $\mathcal{Z}_{i_1,\dots,i_d}$ *are independent zero mean and* σ*-sub-Gaussian random variables,* $p = \min\{p_1, \cdots, p_d\}$*. Suppose there exists a constant* C_{gap} *such that* $\lambda_k = s_{r_k}([\mathcal{X}]_k) \geqslant C_{gap}\left((\sum_{i=1}^{d} p_i r_{i-1} r_i)^{1/2} + (p_{k+1}\cdots p_d)^{1/2}\right)\sigma$ *for* $1 \leqslant k \leqslant d-1$*. Then with probability at least* $1 - Ce^{-cp}$*, for all* $t \geqslant 1$*,*

$$\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\|_F^2 \leqslant C\sigma^2 \sum_{i=1}^{d} p_i r_i r_{i-1}. \tag{4.16}$$

The proof of Corollary 4.4.1 is provided in Section 5.3.4.

Remark 4.4.1 (Interpretation of Corollary 4.4.1)**.** *Note that the TT-cores* $G_1, \mathcal{G}_i, G_d$ *respectively have* $p_1 r_1, p_i r_i r_{i-1}, p_d r_{d-1}$ *free parameters, the upper bound* (4.16) *can be seen as the noise level* σ^2 *times the degrees of freedom of the low TT rank tensors.*

Next, we develop a minimax lower bound for the low TT rank structure estimation. Consider the following general class of tensors with dimension $\mathbf{p} = (p_1, \dots, p_d)$ and TT rank $\mathbf{r} = (r_1, \dots, r_{d-1})$.

$$\mathcal{F}_{\mathbf{p},\mathbf{r}}(\boldsymbol{\lambda}) = \left\{\mathcal{X} \in \mathbb{R}^{p_1\times\cdots\times p_d}, \mathcal{X} \text{ can be decomposed as (4.1)}, s_{r_k}([\mathcal{X}]_k) \geqslant \lambda_k, 1 \leqslant k \leqslant d-1\right\}. \tag{4.17}$$

Here, the constraint on the least singular value of $[\mathcal{X}]_k$ corresponds to the condition required for upper bound in Theorem 4.3.

Theorem 4.4 (Lower bound)**.** *Consider the order-*d *TT spiked tensor model* (4.2)*, where* $\mathcal{Z} \overset{iid}{\sim} N(0, \sigma^2)$*. Assume* $p = \min\{p_1, \dots, p_d\} \geqslant C_0$ *for some large constant* C_0*,* $r_1 \leqslant p_1/2, r_i \leqslant p_i r_{i-1}/2, r_{i-1} \leqslant p_i r_i/2$ *for* $2 \leqslant i \leqslant d-1$*,* $r_{d-1} \leqslant p_d$*, and* $\lambda_i > 0$*. Also* *assume* $r_1 r_2 \leqslant p_1$ *if* $d = 3$*. Then there exists a universal constant* $c > 0$ *such that*

$$\inf_{\widehat{\mathcal{X}}} \sup_{\mathcal{X} \in \mathcal{F}_{\mathbf{p},\mathbf{r}}(\lambda)} \mathbb{E} \left\| \widehat{\mathcal{X}} - \mathcal{X} \right\|_F^2 \geqslant c\sigma^2 \sum_{i=1}^{d} p_i r_i r_{i-1}. \tag{4.18}$$

See Section 5.3.5 for the proof of Theorem 4.4.

Remark 4.4.2. *Corollary 4.4.1 and Theorem 4.4 together show TTOI achieves the minimax optimal rate of estimation error in the low TT-rank class* $\mathcal{F}_{\mathbf{p},\mathbf{r}}(\lambda)$*.*

4.5 TTOI for Dimension Reduction and State Aggregation in High-order Markov Chain

Since the introduction at the beginning of the 20th century, the Markov process has been ubiquitous in a variety of disciplines. In the literature, the first order Markov process, i.e., the future observation at $(t+1)$ is conditionally independent of those at times $1, \ldots, (t-1)$ given the immediate past observation at time t, has been commonly used and extensively studied. Moreover, the high-order Markov process often appear in many scenarios, where the future observation is affected by a longer history. For example, in the taxi travel trajectory, the future stop of a taxi not only depends on the current location but also the past path that reveals the direction this taxi is heading to (Benson et al., 2017). The high-order Markov processes have also been applied to inter-personal relationship (Raftery, 1985), financial econometrics (Tsay, 2005), traffic flow (Zhao and Sun, 2016), among many other applications.

We specifically consider an ergodic, time-invariant, and $(d-1)$st order Markov process on a finite state space $\{1, \ldots, p\}$. That is, the future state X_{t+d} depends on the current state X_{t+d-1} and the previous $(d-2)$ states $(X_{t+d-2}, \ldots, X_{t+1})$ jointly:

$$\mathbb{P}\left(X_{t+d} | X_1, \ldots, X_{t+d-1}\right) = \mathbb{P}\left(X_{t+d} | X_{t+1}, \ldots, X_{t+d-1}\right) = \mathcal{P}_{[X_{t+1}, \ldots, X_{t+d}]}. \tag{4.19}$$

Our goal is to achieve a reliable estimation of the transition tensor $\mathcal{P}$ and to predict the future state X_{t+d} based on an observable trajectory. Since the total number of free parameters in a $(d-1)$st order Markov transition tensor $\mathcal{P}$ is $O(p^d)$ without further assumptions, it may be prohibitively difficult to infer $\mathcal{P}$ in both statistics and computation even if p and d are only of moderate scale. Instead, a sufficient dimension reduction for high-order Markov processes is in demand.

To enable the statistical inference and dimension reduction for high-order Markov processes, a powerful tool, mixed transition distribution model (MTD), was introduced (Raftery, 1985). The MTD model assumes that the distribution of future state is a linear combination of the distributions associated with the $(d-1)$ immediate past states. The readers are also referred to Berchtold and Raftery (2002) for a survey on mixed transition distribution model. The linear assumption, however, does not take into account the potential interactions of past states that commonly appear in practice. For example in the New York taxi trip data, the interaction among past locations of a taxi indicates its potential future direction.

On the other hand, there is a recent surge of development in dimension reduction and state aggregation for first order Markov chains. For example, Ganguly et al. (2014) considered the Markov chain aggregation and the application to biology; Du et al. (2019) considered the rank-reduced Markov model and mode clustering; Zhang and Wang (2020) considered Markov rank, aggregagability, and lumpability of Markov processes and proposed the dimension reduction and state aggregation methods through spectral decomposition with theoretical guarantees; Sanders et al. (2020) proposed clustering block model and proposed efficient algorithm to solve it; Zhu et al. (2019) introduced a convex and non-convex methods to estimate the rank-reduced low-rank Markov transition matrix.

Inspired by these work, we propose and study the *state aggregation model* for the discrete-time high-order Markov processes as follows.

Definition 4.5.1 ($(d-1)$st order state aggregatable Markov process)**.** *Suppose there exist maps* $G_1 : [p] \to \mathbb{R}^{r_1}$, $G_k : [p] \times \mathbb{R}^{r_{k-1}} \to \mathbb{R}^{r_k}$, $G_d : [p] \times \mathbb{R}^{r_{d-1}} \to \mathbb{R}$ *such that* $G_2, \ldots, G_d$ *are linear:* $G_k(X, \lambda_1 u + \lambda_2 v) = \lambda_1 G_k(X, u) + \lambda_2 G_k(X, v)$ *for any vectors*

u, v, scalars $\lambda_1, \lambda_2 \in \mathbb{R}$. We say a Markov process $\{X_1, X_2, \dots\}$ is $(d-1)$st order state aggregatable if for all $t \geqslant 0$, the transition can be sequentially generated as follows,

$$\begin{aligned}
&\tilde{P}_1(X_{t+1}) = G_1(X_{t+1}) \in \mathbb{R}^{r_1},\\
&\tilde{P}_k(X_{t+1}, \dots, X_{t+k}) = G_k(X_{t+k}, \tilde{P}_{k-1}(X_{t+1}, \dots, X_{t+k-1})) \in \mathbb{R}^{r_k}, \quad k = 2, \dots, d-1,\\
&\mathbb{P}\left(X_{t+d} | X_1, \dots, X_{t+d-1}\right) = \mathbb{P}\left(X_{t+d} | X_{t+1}, \dots, X_{t+d-1}\right) = G_d(X_{t+d}, \tilde{P}_{d-1}(X_{t+1}, \dots, X_{t+d-1})).
\end{aligned}$$

In a $(d-1)$st order state aggregatable Markov process, the future state X_{t+d} relies on a sequential aggregation of the previous $d-1$ states $X_{t+1}, \dots, X_{t+d-1}$ as follows: we first project X_{t+1} to a r_1-dimensional vector $\tilde{P}_1(X_{t+1})$ via G_1, then project $\tilde{P}_1(X_{t+1})$ jointly with X_{t+2} to a r_2-dimensional vector $\tilde{P}_1(X_{t+1}, X_{t+2})$ via G_2. We repeat such the projection sequentially for $X_{t+3}, \dots, X_{t+d}$ and yield the transition probability $\mathbb{P}\left(X_{t+d} | X_{t+1}, \dots, X_{t+d-1}\right)$. Also, see Figure 4.4 for a pictorial illustration.

Based on the definition of the state aggregatable Markov chain, we can prove the corresponding probability transition tensor $\mathcal{P}$ will have low TT rank.

Proposition 4.5.1. *The transition tensor $\mathcal{P}$ of the rank reduced high-order Markov model in Definition 4.5.1 has TT-rank no more than $(r_1, \dots, r_{d-1})$. In other words, $\mathcal{P}$ satisfies* $\text{rank}([\mathcal{P}]_k) \leqslant r_k$.

The proof of Proposition 4.5.1 is provided in Section 5.3.6.

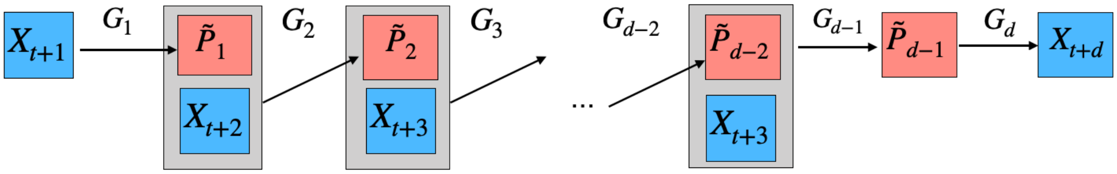


Figure 4.4: A pictorial illustration of a $(d-1)$st order state aggregatable Markov chain

Next, we focus on a *synchronous* or *generative setting*, which can be seen as a high-order generalization of the classic observation model for the analysis of Markov (decision/reward) processes (see Kearns and Singh (1999) for an introduction), for the high-order Markov process. To be specific, for each sample index $k = 1, \dots, n$ and previous states $(i_1, \dots, i_{d-1}) \in [p]^{d-1}$, suppose we observe the next state $X(i_1, \dots, i_{d-1}; k)$ drawn from the Markov transition tensor $\mathcal{P}$. It is natural to estimate $\mathcal{P}$ via the empirical transition tensor:

$$\widehat{\mathcal{P}}^{\mathrm{emp}}_{i_1,\dots,i_d} = \sum_{k=1}^{n} 1_{\{X(i_1,\dots,i_{d-1};k)=i_d\}} \Big/ n, \quad i_1, \dots, i_d \in \{1, \dots, p\}^d.$$

Then, $\widehat{\mathcal{P}}^{\mathrm{emp}}$ is an unbiased estimator of $\mathcal{P}$. However, if the entries of $\mathcal{P}$ are approximately balanced, the mean squared error of $\widehat{\mathcal{P}}^{\mathrm{emp}}$ satisfies

$$\begin{aligned} &\mathbb{E}\left\|\widehat{\mathcal{P}}^{\mathrm{emp}} - \mathcal{P}\right\|_F^2 = \sum_{i_1,\dots,i_d} \mathrm{Var}\left(\widehat{\mathcal{P}}^{\mathrm{emp}}_{i_1,\dots,i_d}\right) \\ = &\sum_{i_1,\dots,i_{d-1}} \sum_{i_d} \frac{\mathbb{P}\left(i_d|i_1, \dots, i_{d-1}\right)\left(1 - \mathbb{P}\left(i_d|i_1, \dots, i_{d-1}\right)\right)}{n} \asymp \frac{p^{d-1}}{n}, \end{aligned} \tag{4.20}$$

To obtain a more accurate estimator, we propose to first perform TTOI on $\widehat{\mathcal{P}}^{\mathrm{emp}}$ to obtain $\widehat{\mathcal{P}}^{(1)}$, then project each row of $[\widehat{\mathcal{P}}^{(1)}]_{d-1}$, or equivalently, each mode-d fiber of $\widehat{\mathcal{P}}^{(1)}$, onto the simplex $S^{p-1} = \{x \in \mathbb{R}^p : \sum_{i=1}^p x_i = 1, x_i \geqslant 0 \text{ for all } 1 \leqslant i \leqslant p\}$ via probability simplex projection (see an implementation in Duchi et al. (2008)) and obtain $\widehat{\mathcal{P}}$.

We establish an upper bound on estimation error for the TTOI estimator $\widehat{\mathcal{P}}$.

Proposition 4.5.2. *Consider the synchronous or generative model for a $(d-1)$st order state aggregatable Markov process described above. Suppose the initialization condition* (4.8) *in Theorem 4.1 holds. Then with probability at least $1 - Ce^{-cp}$, the output of one-step TTOI followed by the probability simplex projection satisfies*

$$\left\|\widehat{\mathcal{P}} - \mathcal{P}\right\|_F^2 \leqslant C\left(\max_{1\leqslant i\leqslant d-1} r_i\right) \sum_{i=1}^{d} p_i r_i r_{i-1} \Big/ n.$$

The proof of Proposition 4.5.2 is provided in Section 5.3.7. Compared to the estimation error rate of $\widehat{\mathcal{P}}^{\text{emp}}$ in (4.20), Proposition 4.5.2 shows TTOI achieves significantly reduced estimation error by exploiting the low TT rank structure of the high-order Markov process.

Remark 4.5.1. *If the observations form one transition trajectory* $\{X_0, \ldots, X_N\}$*, we can work on the following empirical transition tensor*

$$\widehat{\mathcal{P}}^{\text{emp}}_{i_1,\ldots,i_d} = \begin{cases} \frac{\sum_{t=0}^{N-d+1} 1_{\{X_t=i_1,\ldots,X_{t+d-1}=i_d\}}}{\sum_{t=0}^{N-d+1} 1_{\{X_t=i_1,\ldots,X_{t+d-2}=i_{d-1}\}}}, & \sum_{t=1}^{N-d+1} 1_{\{X_t=i_1,\ldots,X_{t+d-2}=i_{d-1}\}} > 0; \\ 1/p, & \sum_{t=1}^{N-d+1} 1_{\{X_t=i_1,\ldots,X_{t+d-2}=i_{d-1}\}} = 0. \end{cases} \tag{4.21}$$

Then $\widehat{\mathcal{P}}^{\text{emp}}$ *can be a nearly unbiased and strongly consistent estimator for* $\mathcal{P}$*. When the Markov process is* $(d-1)$*st order state aggregatable, we can apply TTOI to obtain a better estimate. As will be explored by numerical studies in Section 4.6.1, the TTOI estimator achieves favorable performance on the estimation of* $\mathcal{P}$*.*

4.6 Numerical Studies

In this section, we investigate the numerical performance of TTOI.

4.6.1 Simulation

In each simulation setting, we present the numerical results in both average estimation error (denoted by dots) and standard deviation (denoted by bars) based on 100 repetitions.

We first consider the tensor-train spiked tensor model (4.2) discussed in Section 4.4. Specifically, we randomly generate $G_1, \mathcal{G}_2, \ldots, \mathcal{G}_{d-1}, G_d$ with i.i.d. standard normal entries, and generate $\mathcal{Z}$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ or $\text{Unif}(-b, b)$ entries. Let $p_1 = \cdots = p_d = p, r_1 = \cdots = r_{d-1} = r$, and consider four settings: (1) $p = 100, d = 3, r = 1$; (2) $p = 50, d = 4, r = 1$; (3) $p = 20, d = 5, r = 1$; (4) $p = 20, d = 5, r = 2$. For varying values of $\sigma \in [1, 19]$ and $b \in [3, 30]$, we evaluate the estimation error $\left\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\right\|_F$ of the TT-SVD and TTOI estimators with 1 or 2 iterations, i.e., $t_{\max} =$

$0,1,2$. From the results summarized in Figure 4.5 (normal noise) and Figure 4.6 (uniform noise), we can see TTOI, even with one iteration, performs significantly better than TT-SVD, and the advantage becomes more significant as the noise level σ, b grows. This suggests that the proposed TTOI is effective for high-order tensor SVD compared to the classic TT-SVD, especially when the observations are corrupted by substantial noise.

Next, we demonstrate the performance of TTOI on transition tensor estimation for the high-order state-aggregatable Markov chains studied in Section 4.5. We consider the $(d-1)$st order Markov chain on p states. To generate the transition tensor $\mathcal{P}$, we first draw $\tilde{G}_1 \in \mathbb{R}^{p\times r}, \tilde{\mathcal{G}}_2 \in \mathbb{R}^{r\times p\times r}, \dots, \tilde{G}_d \in \mathbb{R}^{r\times p}$ with i.i.d. standard normal entries, then normalize the rows of $\tilde{G}_1, \tilde{\mathcal{G}}_2, \dots, \tilde{G}_d$ in absolute values as

$$G_{1,[i,j]} = \frac{|\tilde{G}_{1,[i,j]}|}{\sum_{j'} |\tilde{G}_{1,[i,j']}|}, \quad \mathcal{G}_{k,[i_1,i_2,j]} = \frac{|\tilde{\mathcal{G}}_{k,[i_1,i_2,j]}|}{\sum_{j'} |\tilde{\mathcal{G}}_{k,[i_1,i_2,j']}|}, \quad G_{d,[i,j]} = \frac{|\tilde{G}_{d,[i,j]}|}{\sum_{j'} |\tilde{G}_{d,[i,j']}|}.$$

By this means, $\mathcal{P} = [\![G_1, \mathcal{G}_2, \dots, \mathcal{G}_{d-1}, G_d]\!]$ satisfies $\mathcal{P}_{i_1,\dots,i_d} \geqslant 0$, $\sum_{i_d=1}^{p} \mathcal{P}_{i_1,\dots,i_d} = 1$ for any $(i_1, \dots, i_{d-1})$, so $\mathcal{P}$ forms a Markov transition tensor. To generate the trajectory $\{X_1, \dots, X_N\}$, we generate the initial $d-1$ states $X_1, \dots, X_{d-1}$ i.i.d. uniformly from $[p]$, then generate $X_d, \dots, X_N$ sequentially according to (4.19). To estimate $\mathcal{P}$, we construct the empirical probability tensor $\widehat{\mathcal{P}}^{\text{emp}}$ by (4.21), then apply TT-SVD and TTOI with input $\widehat{\mathcal{P}}^{\text{emp}}$ as detailed in Section 4.5 to obtain $\widehat{\mathcal{P}}$. We consider two numerical settings: (1) $p = 100, d = 3, r = 1$; (2) $p = 50, d = 4, r = 1$. We evaluate the estimation error $\|\widehat{\mathcal{P}}^{(i)} - \mathcal{P}\|_F$ for each setting and summarize the results to Figure 4.7. Again, TTOI exhibits clear advantage over the existing methods in all simulation settings.

4.6.2 Real Data Experiments

We apply the proposed method to investigate the Manhattan taxi data[†]. This dataset contains the New York City taxi trip records from 14,144 drivers in 2013. We treat

[†] 2013 Trip Data, available at `https://chriswhong.com/open-data/foil_nyc_taxi/`

each travel record as a transition among different locations at New York City, then the overall dataset can be organized as a collection of fragmented sample trajectories of a Markov chain on New York City traffic. Some recent analysis on such data can be seen at, e.g., Liu et al. (2012); Benson et al. (2017); Zhang and Wang (2020).

Due to the high-dimensional spatiotemporal nature of the dataset, a sufficient dimension reduction or state aggregation is often a crucial first step to study a metropolitan-wide traffic pattern. To this end, we apply the high-order Markov model as described in Section 4.5. Specifically, we discretize the Manhattan region into a grid of $p = 119$ states that forms a state space. Then, we collect all travel records in Manhattan of each driver from the dataset, sort them by time, and form into Markovian transition trajectories. In particular, each travel record is treated as a transition from the pickup to the drop-off location. If the drop-off location i of the previous trip is different from the pickup location j of the next trip by the same driver, we also form a transition from states i to j. Based on the trajectories, we can construct a high-order Markov chain with an order d empirical transition probability tensor $\widehat{\mathcal{P}}^{\mathrm{emp}} \in \mathbb{R}^{\otimes_{k=1}^{d} p}$ as described in Section 4.5. Assuming the true probability tensor is state aggregatable (Definition 4.5.1), we apply one-step TTOI proposed in Section 4.5 and obtain $\widehat{\mathcal{P}}$. It is noteworthy if $d = 2$, the described procedure of $\widehat{\mathcal{P}}$ is equivalent to the classic matrix spectral decomposition in the literature. Figure 4.8 plots the singular values of the sequential unfolding matrices of $\widehat{\mathcal{P}}^{\mathrm{emp}}$ for $d = 3$, which clearly demonstrates the low-TT-rankness of the probability transition tensor $\mathcal{P}$. In the following experiments, we focus on the order-2 Markov model and analyze all consecutive two transitions: $i \to j \to k$, corresponding to the $d = 3$ case.

Inspired by the classic methods of matrix spectral decomposition, we aggregate all location states in Manhattan into a few clusters via both $\widehat{\mathcal{P}}$ and $\widehat{\mathcal{P}}^{\mathrm{emp}}$. Specifically, we calculate $\widehat{G}_d^{\top}$, i.e., the last TT-core of $\widehat{\mathcal{P}}$, and $[\widehat{\mathcal{P}}^{\mathrm{emp}}]_{d-1}$, i.e., the matricization of $\widehat{\mathcal{P}}^{\mathrm{emp}}$ whose columns correspond to the last mode. Then we perform k-means on all columns of $\widehat{G}_d^{\top}$ and $[\widehat{\mathcal{P}}^{\mathrm{emp}}]_{d-1}$, record the cluster index, associate the index to each location state, and plot the results in Figure 4.9 (Panels (a)(b) are for TTOI and Panels (c)(d) are for empirical estimate). From Figure 4.9 (a)(b), we can clearly identify four regions: (i) lower Manhattan (orange), (ii) midtown (dark blue), (iii) upper west side (green), and (iv) upper east side (brown or black). In contrast, direct clustering on $\mathcal{P}^{\text{emp}}$ yields less interpretable results as the majority points go to one cluster. It is also worth noting even the location information is not provided to this experiment, the resulting clusters in Figures 4.9 (a)(b) show good spatial proximity between locations. This illustrates the effectiveness of TTOI in dimension reduction and state-aggregation for high-order Markov processes.

Next, we illustrate the high-order nature of the city-wide taxi trip through the following experiment. For each initial state $i \in [p]$, we apply k-means to cluster the column span of $\widehat{\mathcal{P}}_{[i,:,:]}$, where $\widehat{\mathcal{P}}$ is the outcome of TTOI. We present the results in Figure 4.10, where the red triangles denote the given first state i and $r = k = 7$. If the city-wide taxi trips do not have significant high-order effects, $\widehat{\mathcal{P}}$ should be reducible to a first order Markov process and $\widehat{\mathcal{P}}_{[i,:,:]}$ should have similar values for different i. However, as we can see from Figure 4.10 that the clustering results highly depends on the first state i, the high-order effects exist in the city-wide taxi trip Markov process. In addition, the states in different directions of i are often clustered to different regions, which shows that the taxi drivers may tend to move to the same direction in consecutive trips, which yields the high-order effects in the driving trajectories.

4.7 Discussions and Additional Applications

In this chapter, we propose a general framework for high-order SVD. We introduce a novel procedure, tensor-train orthogonal iteration (TTOI), that efficiently estimates the low tensor train rank structure from the high-order tensor observation. TTOI has significant advantages over the classic ones in the literature. We establish a general deterministic error bound for TTOI with the support of several new representation lemmas for tensor matricizations. Under the commonly studied spiked tensor model, we establish an upper bound for TTOI and a matching information-theoretic lower bound. We also illustrate the merits of TTOI through simulation studies and a real data example in New York City taxi trips.

In addition to the high-order Markov processes, the proposed TTOI can also be applied to the *Markov random field (MRF)* estimation. We give a brief description of MRF below. Consider an undirected graph $G = (V, E)$, where $V = \{1, \dots, d\}$ is a set of vertices and $E \subseteq V \times V$ is a collection of edges. Each vertex $i \in V$ is associated with a random variable X_i, taking values in $\{s_1, \dots, s_p\}$. In an MRF model, the distribution of $\{X_1, \dots, X_d\}$ can be factorized as

$$\mathbb{P}(X_1, \dots, X_d) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(X_C),$$

where $\mathcal{C}$ is a collection of subgraphs of G and $X_C = (X_v, v \in C)$ denotes the random vector corresponding to vertices in C. The joint probability function $\mathbb{P}(\cdot)$ can be written as a tensor $\mathcal{P} \in \mathbb{R}^{\otimes_{k=1}^{d} p}$, where $\mathcal{P}_{i_1,\dots,i_d} = \mathbb{P}(X_1 = s_{i_1}, \dots, X_d = s_{i_d})$. The MRFs have a wide range of applications, including image analysis (Li, 2009; Zhang et al., 2001), genomic study (Wei and Li, 2007), and natural language processing (Chaplot et al., 2015). The readers are referred to, e.g., Wainwright and Jordan (2008) for an introduction to MRFs.

A central problem of MRF is how to estimate the population density $\mathcal{P}$ based on a limited number of samples $\{X_1^{(i)}, \dots, X_d^{(i)}\}_{i=1}^{n}$. It is straightforward to estimate $\mathcal{P}$ via the empirical probability tensor $\widehat{\mathcal{P}}^{\text{emp}}$:

$$\widehat{\mathcal{P}}^{\text{emp}}_{i_1,\dots,i_d} = \sum_{i=1}^{n} \prod_{k=1}^{d} 1(X_k^{(i)} = s_{i_k}) \Big/ n.$$

We can show that $\widehat{\mathcal{P}}^{\text{emp}}$ is unbiased for $\mathcal{P}$. Recently, Novikov et al. (2014) pointed out that $\mathcal{P}$ is often approximately low tensor-train rank in practice. To further exploit such the structure, we can conduct TTOI on $\widehat{\mathcal{P}}^{\text{emp}}$. Under regularity conditions, it can be shown that the entries of $\mathcal{Z}$ are bounded and weakly independent, then Corollary 4.4.1 suggests the following estimation error rate of the TTOI estimator:

$\|\widehat{\mathcal{P}} - \mathcal{P}\|_F^2 \leqslant C \sum_{i=1}^{d} r_i r_{i-1}/(np^{2d-1})$, which can be significantly smaller than the estimation error of original empirical estimator $\widehat{\mathcal{P}}^{\text{emp}}$.

Moreover, the proposed framework can be also applied to *high-order Markov decision process (high-order MDP)*. MDP has been commonly used as a baseline in control theory and reinforcement learning (Singh et al., 1995; Sutton et al., 1998; Puterman, 2014; Duan et al., 2020). Despite the wide applications of MDPs, most of the existing work focus on the first-order Markov processes. However, the high-order effects often appear, i.e., the transition probability at the current time depends not only on current, but also the past $(d-1)$ states and actions. See Figure 4.11 for an example. Since the number of free parameters in such MDPs can be huge, a sufficient dimension reduction for the state and action space can be a crucial first step. Similarly to the example of high-order Markov process in Section 4.5, the TTOI can be applied to achieve better dimension reduction and state aggregation for the high-order Markov decision processes.

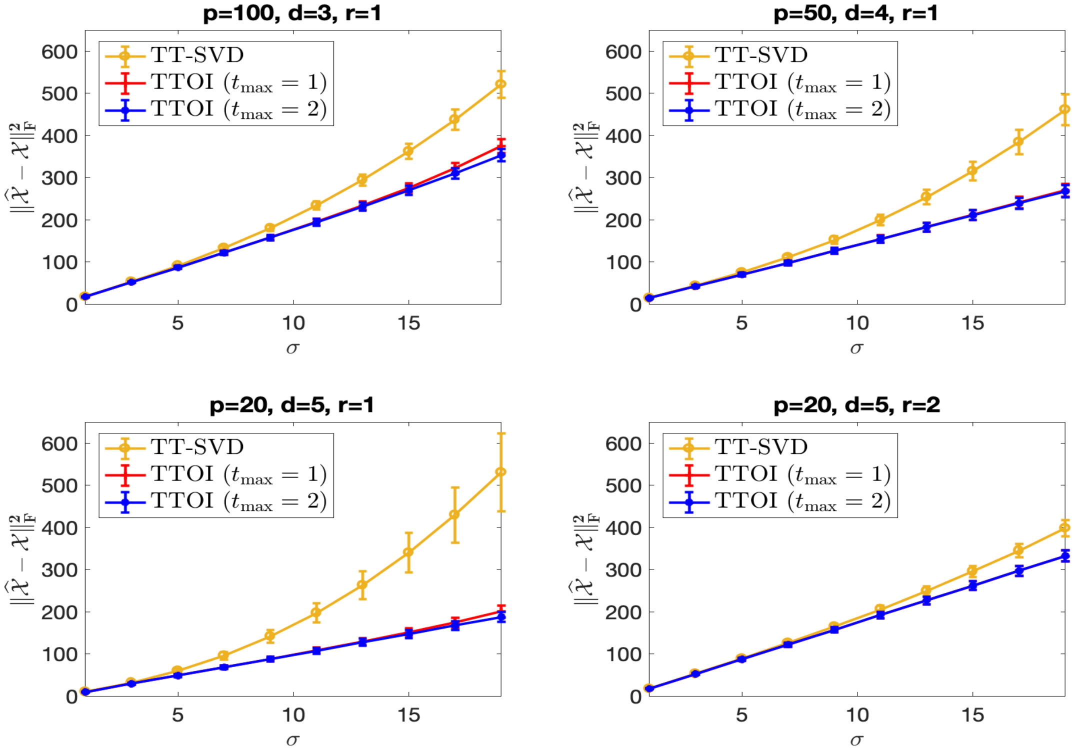


Figure 4.5: Estimation error of TT-SVD and TTOI for high-order spiked tensor model. Here, $\mathcal{Z} \overset{iid}{\sim} N(0, \sigma^2)$.

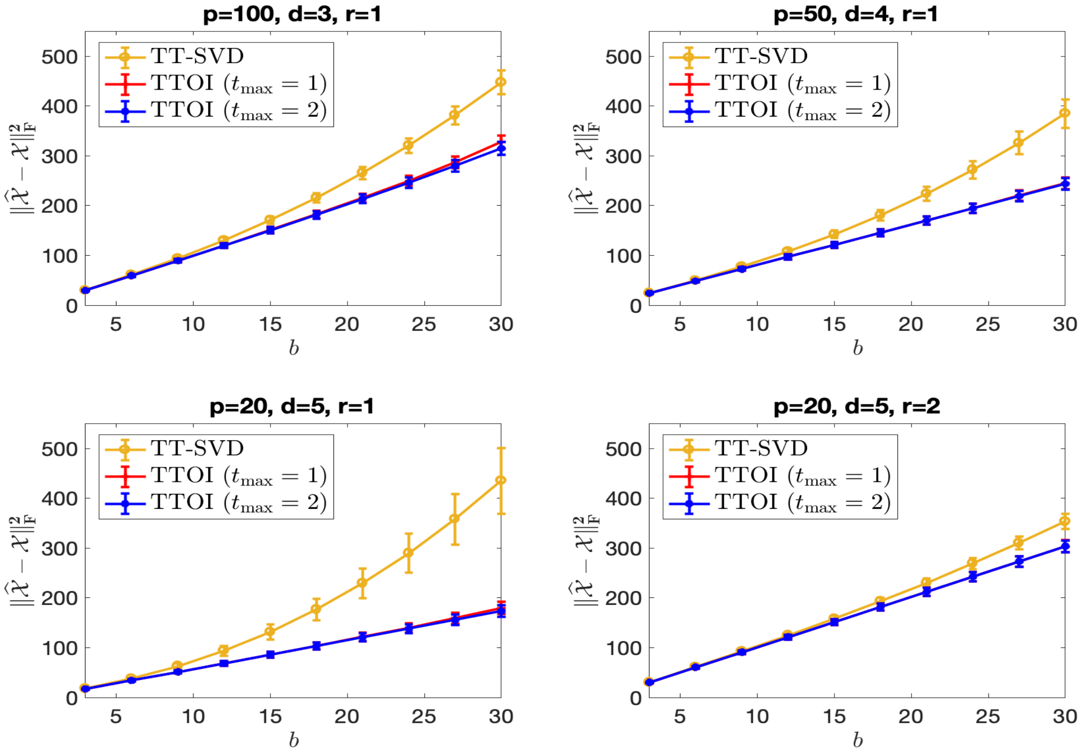


Figure 4.6: Estimation error of TT-SVD and TTOI for high-order spiked tensor model. Here, $\mathcal{Z} \overset{iid}{\sim} \mathrm{Unif}(-b, b)$.

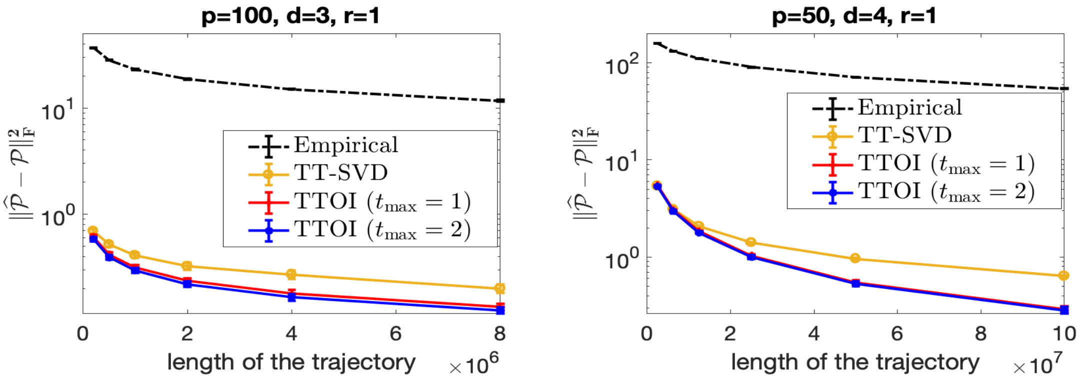


Figure 4.7: Estimation error of the transition tensor versus length of the observable trajectory in high order state-aggregatable Markov chain estimation.

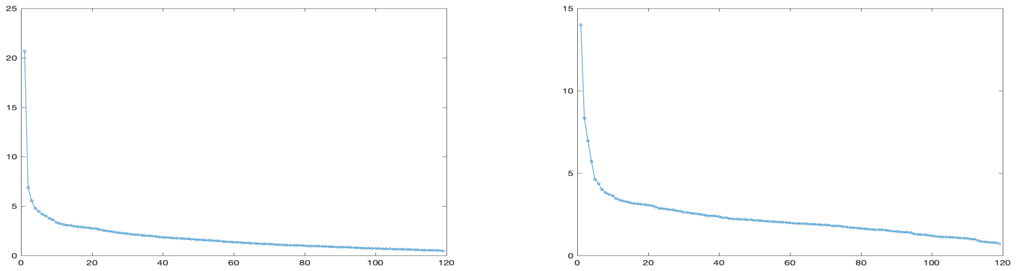
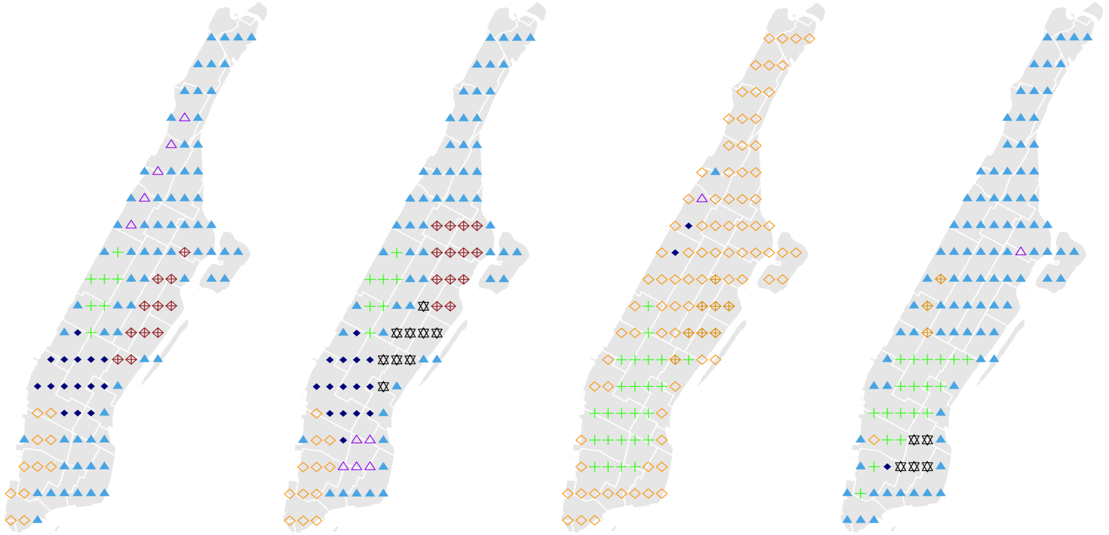


Figure 4.8: Singular values of sequential unfolding matrices $[\widehat{\mathcal{P}}^{\text{emp}}]_1$ (left panel) and $[\widehat{\mathcal{P}}^{\text{emp}}]_2$ (right panel)



(a) $\widehat{G}_d$, $r = 6, k = 6$ (b) $\widehat{G}_d$, $r = 7, k = 7$ (c) $[\widehat{\mathcal{P}}^{\text{emp}}]_{d-1}$, $k = 6$ (d) $[\widehat{\mathcal{P}}^{\text{emp}}]_{d-1}$, $k = 7$

Figure 4.9: State aggregation based on TTOI and empirical estimate

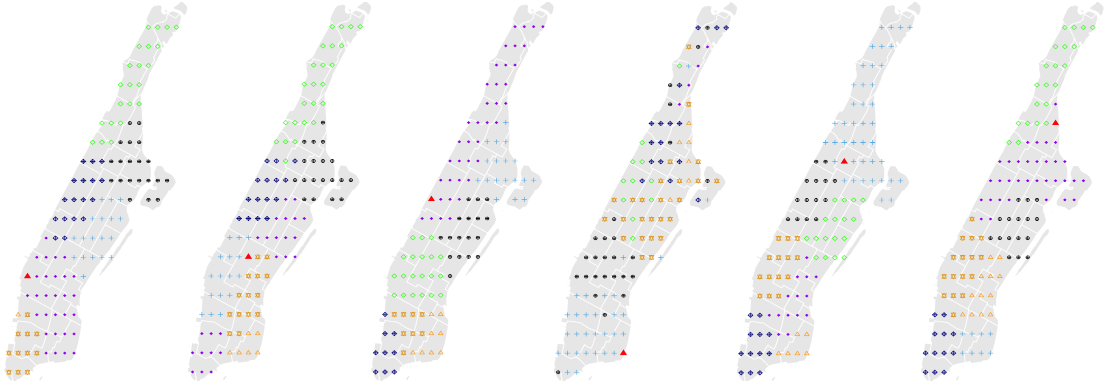


Figure 4.10: Based on second order Markov model, state aggregation results are different with different initial state (the red triangle denotes the initial state i in each subfigure)

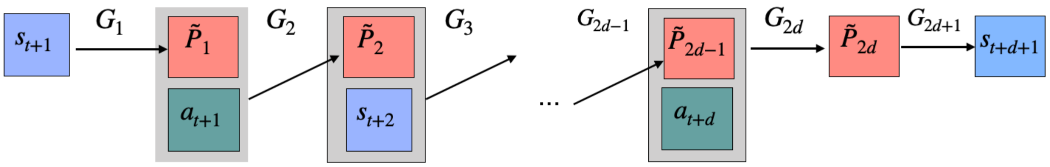


Figure 4.11: Illustration of a high-order state aggregatable Markov decision process

Chapter 5

Appendices

5.1 Appendix to Chapter 2

We collect the proofs of technical results in this section.

5.1.1 Proof of Lemma 2.2.1

Let T satisfy (2.16). For convenience, we denote $s_0 = s/s_g$ and decompose u as

$$u = v + w, \quad v_i = \begin{cases} u_i - \sqrt{s_0}\beta_i^*/\|\beta_{T,(j)}^*\|_2, & i \in T, i \in (j); \\ u_i, & i \in (G)\backslash T; \\ u_i - H_{1/2}(u_i), & i \in (G^c). \end{cases}$$
$$w_{(j)} = \begin{cases} \sqrt{s_0}\beta_{T,(j)}^*/\|\beta_{T,(j)}^*\|_2, & j \in G; \\ H_{1/2}(u_{(j)}), & j \notin G. \end{cases} \tag{5.1}$$

Note that $|H_{1/2}(x) - x| \leqslant 1/2$ for any $x \in \mathbb{R}$. Based on the property of (2.18), $\|u_{(G)\backslash T}\|_\infty \leqslant 1/2$, then

$$\max_{i \in T^c} |v_i| \leqslant 1/2, \quad \|v_T - \mathrm{sgn}(\beta_T^*)\|_2 = \|u_T - (\widetilde{u}_0)_T\|_2 \leqslant \frac{c_{\min}}{8\max_{i\in T^c}\|X_T^\top X_i/n\|_2}; \tag{5.2}$$

$$w_{(j)} = \sqrt{s_0}\beta^*_{T,(j)}/\|\beta^*_{T,(j)}\|_2, \text{ if } j \in G; \quad \|w_{(j)}\|_2 \leqslant \sqrt{s_0}/2, \text{ if } j \notin G. \tag{5.3}$$

Suppose $\hat{\beta}$ is the minimizer to (2.5), $h = \hat{\beta} - \beta^*$, then based on the sub-differential of $\|\beta\|_1$ and $\|\beta\|_{1,2}$, we have

$$\begin{aligned}
\mathcal{P}(\hat{\beta}) =& \|\hat{\beta}\|_1 + \sqrt{s_0}\|\hat{\beta}\|_{1,2} = \|\beta^* + h\|_1 + \sqrt{s_0}\|\beta^* + h\|_{1,2} \\
\geqslant& \|\beta^*_T\|_1 + \operatorname{sgn}(\beta^*_T)^\top h_T + \|h_{T^c}\|_1 + \sqrt{s_0}\left(\|\beta^*_T\|_{1,2} + \sum_{j\in G}\frac{\beta^{*\top}_{T,(j)}h_{(j)}}{\|\beta^*_{T,(j)}\|_2} + \sum_{j\notin G}\|h_{(j)}\|_2\right) \\
&- \|\beta^*_{T^c}\|_1 - \sqrt{s_0}\|\beta^*_{T^c}\|_{1,2} \\
\geqslant& \mathcal{P}(\beta^*) + \|h_{T^c}\|_1 + \sqrt{s_0}\|h_{(G^c)}\|_{1,2} + \operatorname{sgn}(\beta^*_T)^\top h_T + \sum_{j\in G}\frac{\sqrt{s_0}\beta^{*\top}_{T,(j)}h_{(j)}}{\|\beta^*_{T,(j)}\|_2} \\
&- 2\|\beta^*_{T^c}\|_1 - 2\sqrt{s_0}\|\beta^*_{T^c}\|_{1,2}.
\end{aligned} \tag{5.4}$$

The last inequality comes from $\|\beta^*\|_1 = \|\beta^*_T\|_1 + \|\beta^*_{T^c}\|_1$ and $\|\beta^*\|_{1,2} \leqslant \|\beta^*_T\|_{1,2} + \|\beta^*_{T^c}\|_{1,2}$.

In particular, given $Xh = 0$ and that u lies in the row span of X, we have $v^\top h + w^\top h = u^\top h = 0$. Therefore,

$$\begin{aligned}
&\operatorname{sgn}(\beta^*_T)^\top h_T + \sum_{j\in G}\frac{\sqrt{s_0}\beta^{*\top}_{T,(j)}h_{(j)}}{\|\beta^*_{T,(j)}\|_2} = \operatorname{sgn}(\beta^*_T)^\top h_T - v^\top h + \sum_{j\in G}\frac{\sqrt{s_0}\beta^{*\top}_{T,(j)}h_{(j)}}{\|\beta^*_{T,(j)}\|_2} - w^\top h \\
=& -(v_T - \operatorname{sgn}(\beta^*_T))^\top h_T - v_{T^c}^\top h_{T^c} - \sum_{j\in G}\left(w_{(j)} - \sqrt{s_0}\beta^*_{T,(j)}/\|\beta^*_{T,(j)}\|_2\right)^\top h_{(j)} - (w_{(G^c)})^\top h_{(G^c)} \\
\geqslant& -\|v_T - \operatorname{sgn}(\beta^*_T)\|_2\|h_T\|_2 - \|v_{T^c}\|_\infty \cdot \|h_{T^c}\|_1 \\
&- \max_{j\in G}\left\|w_{(j)} - \sqrt{s_0}\beta^*_{T,(j)}/\|\beta^*_{T,(j)}\|_2\right\|_2 \cdot \|h_{(G)}\|_{1,2} - \|w_{(G^c)}\|_{\infty,2}\|h_{(G^c)}\|_{1,2} \\
\overset{(5.2)(5.3)}{\geqslant}& -\|v_T - \operatorname{sgn}(\beta^*_T)\|_2 \cdot \|h_T\|_2 - \|h_{T^c}\|_1/2 - \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2.
\end{aligned} \tag{5.5}$$

Next note that $h = h_T + h_{T^c}$, we must have $X_T h_T = -X_{T^c} h_{T^c}$, then

$$\begin{aligned}\|h_T\|_2 =& \|(X_T^\top X_T/n)^{-1} X_T^\top X_T h_T/n\|_2 \leqslant \sigma_{\min}^{-1}(X_T^\top X_T/n)\|X_T X_{T^c} h_{T^c}/n\|_2 \\ \leqslant& \frac{2}{c_{\min}} \cdot \max_{i \in T^c} \|X_T^\top X_i/n\|_2 \cdot \|h_{T^c}\|_1.\end{aligned} \tag{5.6}$$

Combining (5.2), (5.5), and (5.6), one obtains

$$\operatorname{sgn}(\beta_T^*)^\top h_T + \sum_{j \in G} \frac{\sqrt{s_0}\beta_{T,(j)}^{*\top} h_{(j)}}{\|\beta_{T,(j)}^*\|_2} \geqslant -3/4 \cdot \|h_{T^c}\|_1 - \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2.$$

Plug this inequality to (5.4), we finally have

$$\mathcal{P}(\hat{\beta}) \geqslant \mathcal{P}(\beta^*) + \|h_{T^c}\|_1/4 + \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2 - 2\|\beta_{T^c}^*\|_1 - 2\sqrt{s_0}\|\beta_{T^c}^*\|_{1,2}.$$

Since $\hat{\beta}$ is the minimizer to (2.5), we must have $\mathcal{P}(\hat{\beta}) \leqslant \mathcal{P}(\beta^*)$, then

$$\|h_{T^c}\|_1/4 + \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2 \leqslant 2\|\beta_{T^c}^*\|_1 + 2\sqrt{s_0}\|\beta_{T^c}^*\|_{1,2}. \tag{5.7}$$

If β^* is (s, s_g)-sparse, immediately we have $h_{T^c} = 0$. Then $0 = X_T^\top X h = (X_T^\top X_T)h_T$. By $\sigma_{\min}(X_T^\top X_T/n) \geqslant c_{\min}/2$, we know $X_T^\top X_T/n$ is non-singular, then $h_T = 0$.

Now, we consider the general case. Without loss of generality, suppose $G = \{1, \ldots, g\}$, where $g \leqslant s_g$. Denote T_1 as the indices of the s largest entries of $h_{(G)\setminus T}$, T_2 as the indices of the s largest entries of $h_{(G)\setminus[T \cup T_1]}$, and so on. For $s_g + 1 \leqslant i \leqslant d$, denote $S_{i,1}$ as the indices of the $\lfloor s/s_g \rfloor$ largest entries of $h_{(i)}$, $S_{i,2}$ as the indices of the $\lfloor s/s_g \rfloor$ largest entries of $h_{(i)\setminus S_{i,1}}$, and so on. Let $\widetilde{S}_1, \ldots, \widetilde{S}_{\sum_{i=g+1}^d \lceil b_i/\lfloor s/s_g \rfloor \rceil}$ be an arrangement of $S_{i,j}(1 \leqslant j \leqslant \lceil b_i/\lfloor s/s_g \rfloor \rceil, g+1 \leqslant i \leqslant d)$ such that $\|h_{\widetilde{S}_1}\|_2^2 \geqslant \cdots \geqslant \|h_{\widetilde{S}_{\sum_{i=g+1}^d \lceil b_i/\lfloor s/s_g \rfloor \rceil}}\|_2^2$. Let $R_1 = \cup_{i=1}^{s_g} \widetilde{S}_i$, $R_2 = \cup_{i=s_g+1}^{2s_g} \widetilde{S}_i$, and so on. Then $(T_1, T_2, \ldots, R_1, R_2, \ldots)$ is a partition of T^c, and $|T_i|, |R_j| \leqslant s, |g(T_i)|, |g(R_j)| \leqslant s_g$, where $g(S) = \{i_1, \ldots, i_k\}$ if $S \subseteq \cup_{j=1}^k (i_j)$ and $S \cap (i_j)$ are not empty for all $1 \leqslant j \leqslant k$. Let

$\widetilde{T} = T \cup T_1 \cup R_1$. If (2.19) holds, then

$$\frac{c_{\min}}{2}\|h_{\widetilde{T}}\|_2^2 \leqslant \frac{1}{n}\|X_{\widetilde{T}}h_{\widetilde{T}}\|_2^2 = \frac{1}{n}\langle X_{\widetilde{T}}h_{\widetilde{T}}, Xh\rangle - \frac{1}{n}\langle X_{\widetilde{T}}h_{\widetilde{T}}, X_{\widetilde{T}^c}h_{\widetilde{T}^c}\rangle. \tag{5.8}$$

Since $Xh = 0$, we have

$$\langle X_{\widetilde{T}}h_{\widetilde{T}}, Xh\rangle = 0. \tag{5.9}$$

Now, we consider $\left|\langle X_{\widetilde{T}}h_{\widetilde{T}}, X_{\widetilde{T}^c}h_{\widetilde{T}^c}\rangle\right|$. By triangle inequality,

$$\left|\langle X_{\widetilde{T}}h_{\widetilde{T}}, X_{\widetilde{T}^c}h_{\widetilde{T}^c}\rangle\right| \leqslant \left|\langle X_T h_T, X_{\widetilde{T}^c}h_{\widetilde{T}^c}\rangle\right| + \left|\langle X_{T_1}h_{T_1}, X_{\widetilde{T}^c}h_{\widetilde{T}^c}\rangle\right| + \left|\langle X_{R_1}h_{R_1}, X_{\widetilde{T}^c}h_{\widetilde{T}^c}\rangle\right|.$$

The triangle inequality shows that

$$\left|\langle X_T h_T, X_{\widetilde{T}^c}h_{\widetilde{T}^c}\rangle\right| \leqslant \sum_{i\geqslant 2}|\langle X_T h_T, X_{T_i}h_{T_i}\rangle| + \sum_{j\geqslant 2}\left|\langle X_T h_T, X_{R_j}h_{R_j}\rangle\right|.$$

Combine the parallelogram identity and (2.19) together, we have

$$|\langle X_T h_T, X_{T_i}h_{T_i}\rangle| \leqslant C_{\max} n\|h_T\|_2\|h_{T_i}\|_2, \quad \left|\langle X_T h_T, X_{R_j}h_{R_j}\rangle\right| \leqslant C_{\max} n\|h_T\|_2\|h_{R_j}\|_2.$$

Thus,

$$\left|\langle X_T h_T, X_{\widetilde{T}^c}h_{\widetilde{T}^c}\rangle\right| \leqslant C_{\max} n\|h_T\|_2\Big(\sum_{i\geqslant 2}\|h_{T_i}\|_2 + \sum_{j\geqslant 2}\|h_{R_j}\|_2\Big). \tag{5.10}$$

By (3.10) in Candes and Tao (2007), we have

$$\sum_{i\geqslant 2}\|h_{T_i}\|_2 \leqslant s^{-1/2}\|h_{(G)\setminus T}\|_1, \tag{5.11}$$

and

$$\begin{aligned}\sum_{j\geqslant 2}\|h_{R_j}\|_2 =& \sum_{j\geqslant 2}\left(\sum_{i=(j-1)s_g+1}^{js_g}\|h_{\widetilde{S}_i}\|_2^2\right)^{1/2} \leqslant \sum_{j\geqslant 2}\sqrt{s_g}\|h_{\widetilde{S}_{(j-1)s_g}}\|_2 \leqslant \sum_{j\geqslant 2}\sqrt{s_g}\sum_{i=(j-2)s_g+1}^{(j-1)s_g}\|h_{\widetilde{S}_i}\|_2/s_g\\ =& s_g^{-1/2}\sum_k\|h_{\widetilde{S}_k}\|_2 = s_g^{-1/2}\sum_{i=g+1}^{d}\sum_j\|h_{S_{i,j}}\|_2.\end{aligned}$$

For all $g+1 \leqslant i \leqslant d$, apply (3.10) in Candes and Tao (2007) again,

$$\sum_{j \geqslant 2} \|h_{S_{i,j}}\|_2 \leqslant (\lfloor s/s_g \rfloor)^{-1/2} \|h_{(i)}\|_1 \leqslant \sqrt{2}(s/s_g)^{-1/2} \|h_{(i)}\|_1.$$

Moreover, by the definition of $S_{i,1}$,

$$\sum_{i=g+1}^{d} \|h_{S_{i,1}}\|_2 \leqslant \sum_{i=g+1}^{d} \|h_{(i)}\|_2 = \|h_{(G^c)}\|_{1,2}.$$

Therefore,

$$\begin{aligned} \sum_{j \geqslant 2} \|h_{R_j}\|_2 \leqslant & s_g^{-1/2} \left(\sum_{i=g+1}^{d} \sqrt{2}(s/s_g)^{-1/2} \|h_{(i)}\|_1 \right) + s_g^{-1/2} \|h_{(G^c)}\|_{1,2} \\ = & \sqrt{2} s^{-1/2} \|h_{(G^c)}\|_1 + s_g^{-1/2} \|h_{(G^c)}\|_{1,2}. \end{aligned} \tag{5.12}$$

Combine (5.10), (5.11) and (5.12) together, if (2.19) holds, we have

$$\begin{aligned} \left| \langle X_T h_T, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle \right| \leqslant & C_{\max} n \|h_T\|_2 (s^{-1/2} \|h_{(G) \setminus T}\|_1 + \sqrt{2} s^{-1/2} \|h_{(G^c)}\|_1 + s_g^{-1/2} \|h_{(G^c)}\|_{1,2}) \\ \leqslant & C_{\max} n \|h_T\|_2 (\sqrt{2} s^{-1/2} \|h_{T^c}\|_1 + s_g^{-1/2} \|h_{(G^c)}\|_{1,2}). \end{aligned}$$

Similarly, if (2.19) holds, then $\left| \langle X_{T_1} h_{T_1}, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle \right| \leqslant C_{\max} n \|h_{T_1}\|_2 (\sqrt{2} s^{-1/2} \|h_{T^c}\|_1 + s_g^{-1/2} \|h_{(G^c)}\|_{1,2})$ and $\left| \langle X_{R_1} h_{R_1}, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle \right| \leqslant C_{\max} n \|h_{R_1}\|_2 (\sqrt{2} s^{-1/2} \|h_{T^c}\|_1 + s_g^{-1/2} \|h_{(G^c)}\|_{1,2})$. Thus, with probability at least $1 - 2e^{-cn}$,

$$\begin{aligned} \left| \langle X_{\widetilde{T}} h_{\widetilde{T}}, X_{\widetilde{T}^c} h_{\widetilde{T}^c} \rangle \right| \leqslant & C_{\max} n \left(\|h_T\|_2 + \|h_{T_1}\|_2 + \|h_{R_1}\|_2 \right) \left(\sqrt{2} s^{-1/2} \|h_{T^c}\|_1 + s_g^{-1/2} \|h_{(G^c)}\|_{1,2} \right) \\ \leqslant & \sqrt{3} C_{\max} n \|h_{\widetilde{T}}\|_2 (\sqrt{2} s^{-1/2} \|h_{T^c}\|_1 + s_g^{-1/2} \|h_{(G^c)}\|_{1,2}). \end{aligned} \tag{5.13}$$

The last inequality holds due to Cauchy-Schwarz inequality. Combine (5.8), (5.9), (5.13) and Lemma 2.2.3 together, we know that with probability at least $1 - 2e^{-cn}$,

$$\frac{c_{\min}}{2} \|h_{\widetilde{T}}\|_2^2 \leqslant \sqrt{3} C_{\max} \|h_{\widetilde{T}}\|_2 (\sqrt{2} s^{-1/2} \|h_{T^c}\|_1 + s_g^{-1/2} \|h_{(G^c)}\|_{1,2}),$$

i.e., with probability at least $1-2e^{-cn}$,

$$\|h_{\widetilde{T}}\|_2 \leqslant 2\sqrt{3}\frac{C_{\max}}{c_{\min}}(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}).$$

Finally, by (5.7), (5.11), (5.12) and the previous inequality, with probability at least $1-2e^{-cn}$,

$$\begin{aligned}
\|h\|_2 \leqslant & \|h_{\widetilde{T}}\|_2 + \sum_{i\geqslant 2}\|h_{T_i}\|_2 + \sum_{j\geqslant 2}\|h_{R_j}\|_2 \\
\leqslant & 2\sqrt{3}\frac{C_{\max}}{c_{\min}}(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}) + \sqrt{2}s^{-1/2}\|h_{T^c}\|_2 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2} \\
\leqslant & C\left(\frac{1}{\sqrt{s}}\|\beta^*_{T^c}\|_1 + \frac{1}{\sqrt{s_g}}\|\beta^*_{T^c}\|_{1,2}\right).
\end{aligned}$$

In summary, we have finished the proof of this lemma. □

5.1.2 Proof of Lemma 2.2.2

Let T satisfy (2.16). Given $\|\beta^*_T\|_{0,2} \leqslant s_g$, without loss of generally we assume that

$$\beta^*_{T,(s_g+1)}, \cdots, \beta^*_{T,(d)} = 0.$$

We also denote $T_{(j)}$ as the support of $\beta^*_{T,(j)}$. First by Lemma 5.1.3 Part 3 with

$$v \in \mathbb{R}^p, v_k = \begin{cases} 1, & k = i; \\ 0, & k \neq i; \end{cases} \qquad U \in \mathbb{R}^{p\times|T|} = \mathbb{R}^{(\sum_{i=1}^d b_i)\times|T|}, U_{[T,:]} = I; U_{[T^c,:]} = 0,$$

and notice that $x\log(eu/x) \geqslant \log(eu)$ for all $1 \leqslant x \leqslant u$, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{i\in T^c}\left\|X_T^\top X_i/n\right\|_2 \geqslant 1/2\right) \leqslant \sum_{i\in T^c}\mathbb{P}\left(\left\|X_T^\top X_i/n\right\|_2 \geqslant 1/2\right) \\
\leqslant & \sum_{i\in T^c}\mathbb{P}\left(\left\|X_T^\top X_i/n - \mathbb{E}X_T^\top X_i/n\right\|_2 + \|\mathbb{E}X_T^\top X_i/n\|_2 \geqslant 1/2\right)
\end{aligned}$$

$$\begin{aligned}
&\leqslant \sum_{i\in T^c} \mathbb{P}\left(\left\|X_T^\top X_i/n - \mathbb{E}X_T^\top X_i/n\right\|_2 \geqslant 1/2 - \|\Sigma_{T,T}\|\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2\right)\\
&\leqslant \sum_{i\in T^c} \mathbb{P}\left(\left\|X_T^\top X_i/n - \mathbb{E}X_T^\top X_i/n\right\|_2 \geqslant 1/4\right)\\
&\leqslant db \cdot C\exp\left(Cs - n\right) \leqslant C\exp\left(\log(d) + \log(b) + Cs - n\right)\\
&\leqslant C\exp\left(s_g\log(ed/s_g) + s\log(es_gb/s) + Cs - n\right) \leqslant C\exp(-cn) \qquad (5.14)
\end{aligned}$$

provided that $n \geqslant C\left(s\log(es_gb/s) + s_g\log(d/s_g)\right)$ for some large constant $C > 0$. Note that the fourth inequality comes from the facts that $\|\Sigma_{T,T}\| \leqslant \|\Sigma\| \leqslant C_{\max}$ and $\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2 \leqslant c/\sqrt{s} \leqslant 1/(4C_{\max})$. By Lemma 5.1.4 Part 1, we also know

$$\begin{aligned}
\mathbb{P}\left(\sigma_{\min}(X_T^\top X_T/n) \leqslant c_{\min}/2\right) \leqslant& \mathbb{P}\left(\|X_T^\top X_T/n - \Sigma_{T,T}\| \geqslant c_{\min}/2\right)\\
\leqslant& \mathbb{P}\left(\|X_T^\top X_T\Sigma_{T,T}^{-1}/n - I_{|T|}\|\|\Sigma_{T,T}\| \geqslant c_{\min}/2\right)\\
\leqslant& \mathbb{P}\left(\|X_T^\top X_T\Sigma_{T,T}^{-1}/n - I_{|T|}\| \geqslant c_{\min}/(2C_{\max})\right)\\
\leqslant& C\exp\left(Cs - cn\right) \leqslant C\exp(-cn).
\end{aligned}$$

Next, we apply the well-regarded golfing scheme Gross (2011); Candes and Plan (2011) to find an approximate dual certificate u that satisfies (2.18). Let

$$u_0 \in \mathbb{R}^p, \quad (u_0)_{(j)} = \begin{cases} \sqrt{s/s_g}\beta_{T,(j)}^*/\|\beta_{T,(j)}^*\|_2 + \mathrm{sgn}(\beta_{T,(j)}^*), & j \in G;\\ 0, & j \in G^c. \end{cases} \qquad (5.15)$$

Immediately we have $(u_0)_T = (\widetilde{u}_0)_T$. We divide n rows of X into non-overlapping batches, say $X_{[I_1,:]}, X_{[I_2,:]}, \ldots$, with $|I_l| = n_l$. Here, $n_1, n_2, \ldots$ will be specified a little while later. Consider the following sequences

$$\begin{aligned}
&\alpha_0 = u_0,\\
&\gamma_l = X_{[I_l,:]}^\top X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l \cdot (\alpha_{l-1})_T, \quad \alpha_l = \alpha_{l-1} - \gamma_l, \quad l = 1, 2, \ldots, l_{\max}.
\end{aligned} \qquad (5.16)$$

Finally the approximate dual certificate is defined as

$$u = \sum_{l=1}^{l_{\max}} \gamma_l = \sum_{l=1}^{l_{\max}} X_{[I_l,:]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot (\alpha_{l-1})_T. \tag{5.17}$$

From the inductive definition we can see

$$(\alpha_l)_T = (I - X_{[I_l,T]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l)(\alpha_{l-1})_T, \quad (\gamma_l)_{T^c} = X_{[I_l,T^c]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l \cdot (\alpha_{l-1})_T, \quad l = 1, 2, \ldots.$$

Next, we apply the random matrix results (Lemmas 5.1.4 and 5.1.3) and obtain the following tail probabilities.

- if $n_l \geqslant Cst_l$ for large constant $C > 0$ and $t_l \geqslant C$, by Part 1 of Lemma 5.1.4,

$$\begin{aligned} &\mathbb{P}\left(\|X_{[I_l,T]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1}/n_l - I_{|T|}\| \geqslant C\sqrt{st_l/n_l} \right) \\ \leqslant &C \exp\left(Cs - n_l \min\left\{ \frac{st_l}{n_l}, \left(\frac{st_l}{n_l}\right)^{1/2} \right\} \right) \leqslant C \exp\left(-cst_l\right); \end{aligned} \tag{5.18}$$

- Suppose $q_{l-1} = (\alpha_{l-1})_T \in \mathbb{R}^{|T|}$ is independent of $X_{[I_l,:]}$. If $n_l \geqslant \frac{C(s_0 \log(es_g b/s) + \log d)}{\min\{s_0 \delta_l^2, \sqrt{s_0}\delta_l\}}$ for $\delta_l \geqslant C \max_{i \in T^c} \|\Sigma_{i,T} \Sigma_{T,T}^{-1}\|_2 \geqslant C(\max_{i \in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2) \|\Sigma_{T,T}^{-1} q_{l-1}\|_2 / \|q_{l-1}\|_2$, by Lemma 5.1.4 Part 2,

$$
\begin{aligned}
&\mathbb{P}\left(\max_{j\in G^c}\left\|H_{\|q_{l-1}\|_2\delta_l}\left(X^\top_{[I_l,(j)]}X_{[I_l,T]}\Sigma^{-1}_{T,T}/n_l\cdot q_{l-1}\right)\right\|_2 \geqslant \sqrt{s_0}\|q_{l-1}\|_2\delta_l\right)\\
\leqslant&\sum_{j\in G^c}\mathbb{P}\left(\left\|H_{\|q_{l-1}\|_2\delta_l}\left(X^\top_{[I_l,(j)]}X_{[I_l,T]}(\Sigma^{-1}_{T,T}q_{l-1})/n_l\right)\right\|_2 \geqslant \sqrt{s_0}\|q_{l-1}\|_2\delta_l\right)\\
\leqslant& d\cdot\binom{b}{\lceil s_0\rceil}\exp\left(Cs_0 - cn_l\min\left\{\frac{s_0\|q_{l-1}\|_2^2\delta_l^2}{\kappa^4\|\Sigma^{-1}_{T,T}q_{l-1}\|_2^2},\frac{\sqrt{s_0}\|q_{l-1}\|_2\delta_l}{\kappa^2\|\Sigma^{-1}_{T,T}q_{l-1}\|_2}\right\}\right)\\
&+d\cdot\binom{b}{\lfloor s_0\rfloor}\exp\left(Cs_0 - cn_l\min\left\{\frac{s_0\|q_{l-1}\|_2^2\delta_l^2}{\kappa^4\|\Sigma^{-1}_{T,T}q_{l-1}\|_2^2},\frac{\sqrt{s_0}\|q_{l-1}\|_2\delta_l}{\kappa^2\|\Sigma^{-1}_{T,T}q_{l-1}\|_2}\right\}\right)\\
\leqslant& 2d\cdot\left(\frac{eb}{\lfloor s_0\rfloor}\right)^{\lceil s_0\rceil}\exp\left(Cs_0 - cn_l\min\{s_0\delta_l^2,\sqrt{s_0}\delta_l\}\right)\\
\leqslant& C\exp\left(\log(d) + Cs_0\log(2es_gb/s) + Cs_0 - cn_l\min\{s_0\delta_l^2,\sqrt{s_0}\delta_l\}\right)\\
\leqslant& C\exp(-cn_l\min\{s_0\delta_l^2,\sqrt{s_0}\delta_l\});
\end{aligned}
\tag{5.19}
$$

The third inequality comes from $\|\Sigma^{-1}_{T,T}\| \leqslant \frac{1}{c_{\min}}$.

- Suppose $q_{l-1} = (\alpha_{l-1})_T \in \mathbb{R}^{|T|}$ is fixed. If $n_l\min\{\theta_l^2,\theta_l\} \geqslant C\log(es_gb), \theta_l \geqslant$

$2\max_{i\in T^c}\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2$, by Lemma 5.1.3 Part 2,

$$\begin{aligned}
&\mathbb{P}\left(\|X_{[I_l,(G)\setminus T]}^\top X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l\cdot q_{l-1}\|_\infty \geqslant \theta_l\|q_{l-1}\|_2\right)\\
\leqslant&\sum_{i\in(G)\setminus T}\mathbb{P}\left(|X_{[I_l,i]}^\top X_{[I_l,T]}/n_l\cdot(\Sigma_{T,T}^{-1}q_{l-1})|\geqslant\theta_l\|q_{l-1}\|_2\right)\\
\leqslant&\sum_{i\in(G)\setminus T}\mathbb{P}\left(|X_{[I_l,i]}^\top X_{[I_l,T]}/n_l\cdot(\Sigma_{T,T}^{-1}q_{l-1})-\Sigma_{i,T}\Sigma_{T,T}^{-1}q_{l-1}|\geqslant\theta_l\|q_{l-1}\|_2-|\Sigma_{i,T}\Sigma_{T,T}^{-1}q_{l-1}|\right)\\
\leqslant&\sum_{i\in(G)\setminus T}\mathbb{P}\left(|X_{[I_l,i]}^\top X_{[I_l,T]}/n_l\cdot(\Sigma_{T,T}^{-1}q_{l-1})-\Sigma_{i,T}\Sigma_{T,T}^{-1}q_{l-1}|\geqslant\theta_l\|q_{l-1}\|_2-\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2\|q_{l-1}\|_2\right)\\
\leqslant&\sum_{i\in(G)\setminus T}\mathbb{P}\left(|X_{[I_l,i]}^\top X_{[I_l,T]}/n_l\cdot(\Sigma_{T,T}^{-1}q_{l-1})-\Sigma_{i,T}\Sigma_{T,T}^{-1}q_{l-1}|\geqslant\frac{1}{2}\theta_l\|q_{l-1}\|_2\right)\\
\leqslant&\sum_{i\in(G)\setminus T}\mathbb{P}\left(|X_{[I_l,i]}^\top X_{[I_l,T]}/n_l\cdot(\Sigma_{T,T}^{-1}q_{l-1})-\Sigma_{i,T}\Sigma_{T,T}^{-1}q_{l-1}|\geqslant\frac{c_{\min}}{2}\theta_l\|\Sigma_{T,T}^{-1}q_{l-1}\|_2\right)\\
\leqslant& s_g b\cdot C\exp\left(-cn_l\min\{\theta_l^2,\theta_l\}\right)=C\exp(\log(s_g b)-cn_l\min\{\theta_l^2,\theta_l\})\\
\leqslant& C\exp\left(-cn_l\min\{\theta_l^2,\theta_l\}\right).
\end{aligned}\tag{5.20}$$

Then we specify $\{n_l,t_l,\delta_l,\theta_l\}_{l\geqslant 1}$ as follows,

- $n_1=n_2\geqslant C(s\log(es_g b)+s_g\log(d/s_g))$, $t_1=t_2=cn_1/(s\log(es))\geqslant C$, $\delta_1=\delta_2=1/(16\sqrt{s})$, $\theta_1=\theta_2=1/(16\sqrt{s})$;
- $n_3=\cdots=n_{l_{\max}}\asymp\frac{n_1}{l_{\max}-2}\geqslant C(s\log(es_g b)+s_g\log(d/s_g))/\log(es)$, $t_3=\cdots=t_{l_{\max}}=cn_3/s\geqslant C$, $\delta_3=\cdots=\delta_{l_{\max}}=\log(es)/(16\sqrt{s})\geqslant\max\{(\log(es)/s)^{1/2}/16,\log(es)\sqrt{s_0}/(16s)\}$, $\theta_3=\cdots=\theta_{l_{\max}}=(\log(es)/s)^{1/2}/16$, with $l_{\max}=\lceil C\log(es)\rceil+2$.

We can see the following events happen

$$\begin{aligned}
&\|X_{[I_l,T]}^\top X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l-I_{|T|}\|\leqslant C\sqrt{st_l/n_l}\leqslant\sqrt{1/\log(es)},\quad l=1,2;\\
&\|X_{[I_l,T]}^\top X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l-I_{|T|}\|\leqslant C\sqrt{st_l/n_l}\leqslant 1/2,\quad l=3,\ldots,l_{\max};
\end{aligned}\tag{5.21}$$

$$\begin{aligned}&\max_{j\in G^c}\left\|H_{\|q_{l-1}\|_2/(16\sqrt{s})}\left(X_{[I_l,(j)]}^{\top}X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l\cdot q_{l-1}\right)\right\|_2\leqslant\sqrt{s_0}\|q_{l-1}\|_2/(16\sqrt{s}),\quad l=1,2;\\&\max_{j\in G^c}\left\|H_{\|q_{l-1}\|_2\cdot\log(es)/(16\sqrt{s})}\left(X_{[I_l,(j)]}^{\top}X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l\cdot q_{l-1}\right)\right\|_2\\\leqslant&\sqrt{s_0}\|q_{l-1}\|_2\log(es)/(16\sqrt{s}),\quad l=3,\dots,l_{\max};\end{aligned}\tag{5.22}$$

$$\begin{aligned}&\left\|X_{[I_l,(G)\setminus T]}^{\top}X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l\cdot q_{l-1}\right\|_\infty\leqslant\|q_{l-1}\|_2/(16\sqrt{s}),\quad l=1,2\\&\left\|X_{[I_l,(G)\setminus T]}^{\top}X_{[I_l,T]}\Sigma_{T,T}^{-1}/n_l\cdot q_{l-1}\right\|_\infty\leqslant\|q_{l-1}\|_2\cdot(\log(es)/s)^{1/2}/16,\quad l=3,\dots,l_{\max}.\end{aligned}\tag{5.23}$$

with probability at least $1-C\log(es)\exp(-c\frac{n}{\log(es)})-C\log(es)\exp\left(-c\frac{n}{s_g}\right)-C\log(es)\exp\left(-c\frac{n}{s}\right)$. By triangle inequality, u_0 satisfies

$$\|u_0\|_2\leqslant\sqrt{s/s_g}\left(\sum_{j\in G}\left\|\frac{\beta_{T,(j)}^*}{\|\beta_{T,(j)}^*\|_2}\right\|_2^2\right)^{1/2}+\|\operatorname{sgn}(\beta_T^*)\|_2\leqslant2\sqrt{s}.\tag{5.24}$$

When $\max_{i\in T^c}\|X_T^{\top}X_i/n\|_2\leqslant\frac{1}{2}$ and (5.21)-(5.24) hold, we have

$$\begin{aligned}&\|q_0\|_2\leqslant2\sqrt{s},\\&\|q_1\|_2=\left\|(I_{|T|}-X_{I_1,T}^{\top}X_{I_1,T}\Sigma_{T,T}^{-1}/n_1)q_0\right\|\leqslant\|I_{|T|}-X_{I_1,T}^{\top}X_{I_1,T}\Sigma_{T,T}^{-1}/n_1\|\cdot\|q_0\|_2\\&\qquad\leqslant2\sqrt{s/\log(es)};\\&\text{similarly, }\|q_2\|_2\leqslant\|q_1\|_2/\sqrt{\log(es)}\leqslant2\sqrt{s}/(\log(es));\\&\qquad\|q_l\|_2\leqslant\|q_{l-1}\|_2/2\leqslant\cdots\leqslant\|q_2\|/2^{l-2}\leqslant2^{3-l}\sqrt{s}/(\log(es)),\quad l\geqslant3.\end{aligned}\tag{5.25}$$

For large constant $C>0,\|q_{l_{\max}}\|_2\leqslant2^{3-C\log(es)}\sqrt{s}/\log(es)\leqslant c_{\min}/8$. Notice that

$$u_T=(\sum_{l=1}^{l_{\max}}\gamma_l)_T=(\sum_{l=1}^{l_{\max}}(\alpha_{l-1}-\alpha_l))_T=(\alpha_0-\alpha_{l_{\max}})_T=(\widetilde{u}_0)_T-(q_{l_{\max}})_T,$$

we know that

$$\|u_T - (\widetilde{u}_0)_T\|_2 \cdot \max_{i \in T^c} \|X_T^\top X_i / n\|_2 = \|q_{l_{\max}}\|_2 \cdot \max_{i \in T^c} \|X_T^\top X_i / n\|_2 \leqslant \frac{c_{\min}}{8} \cdot \frac{1}{2} < \frac{c_{\min}}{8}.$$

In addition,

$$\begin{aligned}
\|u_{(G)\setminus T}\|_\infty \leqslant & \sum_{l=1}^{l_{\max}} \left\| X_{[I_l,(G)\setminus T]}^\top X_{[I_l,T]} \Sigma_{T,T}^{-1} / n_l \cdot (\alpha_{l-1})_T \right\|_\infty \\
\leqslant & \|q_0\|_2 / (16\sqrt{s}) + \|q_1\|_2 / (16\sqrt{s}) + \sum_{l=3}^{l_{\max}} \|q_{l-1}\|_2 \cdot (\log(es)/s)^{1/2} / 16 \\
\leqslant & 1/8 + 1/8 + \sum_{l=3}^{\infty} 2^{4-l} / 16 \leqslant 1/2.
\end{aligned}$$

Since

$$\begin{aligned}
& \|q_0\|_2 / (16\sqrt{s}) + \|q_1\|_2 / (16\sqrt{s}) + \sum_{l=3}^{l_{\max}} \|q_{l-1}\|_2 \cdot \log(es) / (16\sqrt{s}) \\
\leqslant & \frac{1}{8} + \frac{1}{8} + \sum_{l=3}^{l_{\max}} 2^{4-l} \sqrt{s} / (\log(es)) \cdot \log(es) / (16\sqrt{s}) \leqslant \frac{1}{2},
\end{aligned}$$

$$\begin{aligned}
\left\| H_{1/2}(u_{(G^c)}) \right\|_{\infty,2} \leqslant & \|H_{\|q_0\|_2/(16\sqrt{s}) + \|q_1\|_2/(16\sqrt{s}) + \sum_{l=3}^{l_{\max}} \|q_{l-1}\|_2 \cdot \log(es)/(16\sqrt{s})}(u_{(G^c)})\|_{\infty,2} \\
\leqslant & \sum_{l=1}^{2} \left\| H_{\|q_{l-1}\|_2/(16\sqrt{s})} \left(X_{[I_l,(G^c)]}^\top X_{[I_l,T^c]} q_{l-1} \right) \right\|_{\infty,2} \\
& + \sum_{l=3}^{l_{\max}} \left\| H_{\|q_{l-1}\|_2 \cdot \log(es)/(16\sqrt{s})} \left(X_{[I_l,(G^c)]}^\top X_{[I_l,T^c]} q_{l-1} \right) \right\|_{\infty,2} \\
\leqslant & \sum_{l=1}^{2} \sqrt{s_0} \|q_{l-1}\|_2 / (16\sqrt{s}) + \sum_{l=3}^{l_{\max}} \sqrt{s_0} \|q_{l-1}\|_2 \cdot \log(es) / (16\sqrt{s}) \\
\leqslant & \sqrt{s_0}/2.
\end{aligned}$$

Thus, the construction of u satisfies all required condition in Lemma 2.2.1 with probability at least $1 - C\exp\left(-c\frac{n}{s}\right)$. This has finished the proof of this lemma. □

5.1.3 Proof of Lemma 2.2.3

Let $g(S)$ be the group support of set S, that is, $g(S) = \{i_1, \ldots, i_k\}$ if $S \subset \cup_{j=1}^{k}(i_j)$ and $S \cap (i_j)$ are not empty for all $1 \leqslant j \leqslant k$. Lemma 5.1.4 Part 1 and the union bound show that

$$
\begin{aligned}
& \mathbb{P}\left(\exists \gamma \in \mathbb{R}^p, \|\gamma\|_0 \leqslant 2s, \|\gamma\|_{0,2} \leqslant 2s_g, \frac{1}{n}\|X\gamma\|_2^2 \notin \left[\frac{c_{\min}}{2}\|\gamma\|_2^2, (C_{\min} + \frac{c_{\min}}{2})\|\gamma\|_2^2\right]\right) \\
=& \mathbb{P}\left(\exists x \in \mathbb{R}^{2s\wedge p}, S \subseteq 1, \cdots, p, |S| = 2s \wedge p, |g(S)| \leqslant 2s_g, \frac{1}{n}\|X_S x\|_2^2 \notin \left[\frac{c_{\min}}{2}\|\gamma\|_2^2, (C_{\min} + \frac{c_{\min}}{2})\|\gamma\|_2^2\right]\right) \\
\leqslant & \sum_{S \subseteq \{1,\ldots,p\}, |S| = 2s\wedge p, |g(S)| \leqslant 2s_g} \mathbb{P}\left(\forall x \in \mathbb{R}^{2s\wedge p}, \frac{1}{n}\|X_S x\|_2^2 \notin \left[\frac{c_{\min}}{2}\|\gamma\|_2^2, (C_{\min} + \frac{c_{\min}}{2})\|\gamma\|_2^2\right]\right) \\
\leqslant & \sum_{S \subseteq \{1,\ldots,p\}, |S| = 2s\wedge p, |g(S)| \leqslant 2s_g} \mathbb{P}\left(\|\frac{1}{n}X_S^\top X_S - \Sigma_{S,S}\| \geqslant \frac{c_{\min}}{2}\right) \\
\leqslant & \sum_{S \subseteq \{1,\ldots,p\}, |S| = 2s\wedge p, |g(S)| \leqslant 2s_g} \mathbb{P}\left(\|\frac{1}{n}X_S^\top X_S \Sigma_{S,S}^{-1} - I_{|S|}\| \geqslant \frac{c_{\min}}{2C_{\max}}\right) \\
\leqslant & \left[\binom{d}{2s_g} \vee 1\right] \binom{2s_g b}{2s} \cdot 2\exp\left(Cs - cn\right) \\
\leqslant & \left(\frac{ed}{2s_g}\right)^{2s_g} \left(\frac{e \cdot 2s_g b}{2s}\right)^{2s} \cdot 2\exp\left(Cs - cn\right) \\
\leqslant & 2\exp\left(2s\log(es_g b/s) + 2s_g\log(ed/s_g) + Cs - cn\right) \\
\leqslant & 2e^{-cn}.
\end{aligned}
$$

□

5.1.4 Proof of Theorem 2.2.2

If $d \geqslant 3s_g$ and $b \geqslant 3s/s_g$, by (5.52), we can find $\Omega^{(1)}, \ldots, \Omega^{(N)} \subset \{1, \ldots, db\}$ such that $|\Omega^{(i)}| = s_g\lfloor s/s_g \rfloor$, $|\Omega_{(k)}^{(i)}| = \lfloor s/s_g \rfloor 1_{\{\Omega_{(k)}^{(i)} \text{ is not empty}\}}$ for all $1 \leqslant i \leqslant N$, $1 \leqslant k \leqslant d$, and

$$\left|\Omega^{(i)} \cap \Omega^{(j)}\right| \leqslant 8s_g\lfloor s/s_g \rfloor/9, \quad 1 \leqslant i \neq j \leqslant N, \tag{5.26}$$

$$\left|\left\{k \middle| \left|\Omega_{(k)}^{(i)} \cap \Omega_{(k)}^{(j)}\right| \geqslant 2\lfloor s/s_g \rfloor/3\right\}\right| \leqslant 2s_g/3, \quad 1 \leqslant i \neq j \leqslant N, \tag{5.27}$$

where $N = \left\lfloor \left(\frac{d}{2\sqrt{2}s_g}\right)^{s_g/3} \left(\frac{b}{2\sqrt{2}\lfloor s/s_g \rfloor}\right)^{s/9} \right\rfloor$. For any $1 \leqslant j \leqslant db, 1 \leqslant i \leqslant N$, define

$$\beta_j^{(i)} = \begin{cases} \frac{1}{\lambda s_g \lfloor s/s_g \rfloor + \lambda_g s_g \sqrt{\lfloor s/s_g \rfloor}}, & j \in \Omega^{(i)} \\ 0 & j \notin \Omega^{(i)}, \end{cases}$$

then $\|\beta^{(i)}\|_0 \leqslant s, \|\beta^{(i)}\|_{0,2} \leqslant s_g$. We consider the quotient space

$$\mathbb{R}^{db}/\ker(X) = \left\{[x] := x + \ker(X), x \in \mathbb{R}^{db}\right\}.$$

Then the dimension of $\mathbb{R}^{db}/\ker(X)$ is $\operatorname{rank}(X) \leqslant n$. Define the norm $\|[x]\| = \inf_{v \in \ker(X)}\{\lambda\|x - v\|_1 + \lambda_g\|x - v\|_{1,2}\}$. For any vector $x \in \mathbb{R}^{db}$ satisfying $\|x\|_0 \leqslant 2s, \|x\|_{0,2} \leqslant 2s_g$, note that $x - v$ with $v \in \ker(X)$ satisfies $X(x - v) = Xx$, by our assumption, we have $\|[x]\| = \lambda\|x\|_1 + \lambda_g\|x\|_{1,2}$. Thus $\|[\beta^{(1)}]\| = \cdots = \|[\beta^{(N)}]\| = 1$. Moreover, by (5.26) and (5.27),

$$\begin{aligned} \|\beta^{(i)} - \beta^{(j)}\|_1 =& \frac{1}{\lambda s_g \lfloor s/s_g \rfloor + \lambda_g s_g \sqrt{\lfloor s/s_g \rfloor}} \left(|\Omega^{(i)}| + |\Omega^{(j)}| - 2|\Omega^{(i)} \cap \Omega^{(j)}|\right) \\ \geqslant& \frac{2s_g \lfloor s/s_g \rfloor}{9(\lambda s_g \lfloor s/s_g \rfloor + \lambda_g s_g \sqrt{\lfloor s/s_g \rfloor})}, \end{aligned}$$

and

$$\begin{aligned} \|\beta^{(i)} - \beta^{(j)}\|_{1,2} =& \sum_{k=1}^{d} \|\beta^{(i)}_{(k)} - \beta^{(j)}_{(k)}\|_2 \\ \geqslant& \sum_{k \in S_{i,j}} \|\beta^{(i)}_{(k)} - \beta^{(j)}_{(k)}\|_2 \\ \geqslant& \frac{1}{\lambda s_g \lfloor s/s_g \rfloor + \lambda_g s_g \sqrt{\lfloor s/s_g \rfloor}} \sqrt{\frac{2\lfloor s/s_g \rfloor}{3}} \cdot |S_{i,j}| \\ \geqslant& \frac{1}{\lambda s_g \lfloor s/s_g \rfloor + \lambda_g s_g \sqrt{\lfloor s/s_g \rfloor}} \sqrt{\frac{2\lfloor s/s_g \rfloor}{3}} \cdot \frac{s_g}{3}, \end{aligned}$$

where $S_{i,j} = \left\{ k | \Omega^{(i)}_{(k)}, \Omega^{(j)}_{(k)} \text{ are not empty sets}, \left| \Omega^{(i)}_{(k)} \cap \Omega^{(j)}_{(k)} \right| < 2\lfloor s/s_g \rfloor / 3 \right\}$. Since $\beta^{(i)} - \beta^{(j)}$ is $(2s, 2s_g)$-sparse,

$$\left\| [\beta^{(i)}] - [\beta^{(j)}] \right\| = \left\| [\beta^{(i)} - \beta^{(j)}] \right\| = \lambda \left\| \beta^{(i)} - \beta^{(j)} \right\|_1 + \lambda_g \left\| \beta^{(i)} - \beta^{(j)} \right\|_{1,2} \geqslant 2/9.$$

By (Foucart and Rauhut, 2013, Proposition C.3), we have $N \leqslant 10^{\text{rank}(X)} \leqslant 10^n$. Therefore we have

$$\left\lfloor \left(\frac{d}{2\sqrt{2} s_g} \right)^{s_g/3} \left(\frac{b}{2\sqrt{2} \lfloor s/s_g \rfloor} \right)^{s/9} \right\rfloor \leqslant 10^n,$$

which means that $n \geqslant c(s_g \log(d/s_g) + s \log(e s_g b/s))$.

If $d < 3s_g$ or $b < 3s/s_g$, let $s'_g = [s_g/3] \vee 1 \geqslant s_g/5, s' = [s/15] \vee s'_g$, then $d \geqslant 3s'_g$ and $b \geqslant 3s'/s'_g$. Since all $(2s, 2s_g)$-sparse vectors can be exactly recovered by the $\ell_1 + \ell_{1,2}$ minimization and $s' \leqslant s, s'_g \leqslant s_g$, we know that the $\ell_1 + \ell_{1,2}$ minimization exactly recover all $(2s', 2s'_g)$-sparse vectors. Therefore, we have

$$\begin{aligned} n \geqslant & c(s'_g \log(d/s'_g) + s' \log(e s'_g b/s')) \geqslant c\left(\frac{s_g}{5} \cdot \log\left(\frac{d}{s_g} \right) + \frac{s}{15} \cdot \log\left(\frac{eb(s_g/5)}{s/15} \vee eb \right) \right) \\ \geqslant & c'(s_g \log(d/s_g) + s \log(e s_g b/s)). \end{aligned} \tag{5.28}$$

□

5.1.5 Proof of Theorem 2.2.3

We would like prove Theorem 2.2.3 by contradiction. Let

$$c = \min\left\{ \frac{1}{8}, c', \sqrt{\frac{c'}{256}} \right\}, \quad c_0 = \min\left\{ \frac{c}{2e}, \frac{c^2}{2C^2}, 16c^2 \right\}, \quad C_0 = \max\left\{ \frac{C^2}{c^2}, \frac{1}{32c^2} \right\},$$

where c' is a uniform constant such that $n \geqslant c'(s\log(es_g b/s) + s_g\log(d/s_g))$ if the conditions in Theorem 2.2.2 are satisfied. Assume for contradiction that

$$n < c_0(s\log(es_g b/s) + s_g\log(d/s_g)). \tag{5.29}$$

Let $s_0 = s/s_g$, define the norm $\|\cdot\| = \|\cdot\|_1 + \sqrt{s_0}\|\cdot\|_{1,2}$. Let $B = \{x \in \mathbb{R}^p | \|x\| \leqslant 1\}$,

$$d^n(B,\mathbb{R}^p) = \inf_{\substack{L^n \text{ is a subspace of } \mathbb{R}^p \\ \text{with } \dim(\mathbb{R}^p/L^n)\leqslant n}} \left\{ \sup_{\beta\in B\cap L^n} \|\beta\|_2 \right\}.$$

By (Foucart and Rauhut, 2013, Theorem 10.4), we have

$$d^n(B,\mathbb{R}^p) \leqslant \sup_{\beta\in B} \|\beta - \Delta(X\beta)\|_2 \leqslant \frac{C}{\sqrt{s}} \sup_{\beta\in B} (\|\beta\|_1 + \sqrt{s_0}\|\beta\|_{1,2}) = \frac{C}{\sqrt{s}}. \tag{5.30}$$

If

$$d^n(B,\mathbb{R}^p) \geqslant c\min\left\{ \frac{1}{\sqrt{s_0}}, \left[\left(\frac{s_g}{s} \log\left(\frac{c\frac{s}{s_g} d\log(es_g b/s)}{n} \right) + \log(es_g b/s) \right) /n \right]^{1/2} \right\}, \tag{5.31}$$

since

$$\frac{C}{\sqrt{s}} \leqslant \frac{c\sqrt{C_0}}{\sqrt{s}} \leqslant \frac{c\sqrt{s_g}}{\sqrt{s}} = \frac{c}{\sqrt{s_0}},$$

(5.30) and (5.31) together imply that

$$n \geqslant \frac{c^2}{C^2}\left(s_g\log\left(\frac{c\frac{s}{s_g} d\log(es_g b/s)}{n} \right) + s\log(es_g b/s) \right). \tag{5.32}$$

By (5.29),

$$\begin{aligned} \frac{c\frac{s}{s_g} d\log(es_g b/s)}{n} >& \frac{c\frac{s}{s_g} d\log(es_g b/s)}{c_0(s\log(es_g b/s) + s_g\log(d/s_g))} \\ \geqslant& 2e\frac{\frac{s}{s_g} d\log(es_g b/s)}{s\log(es_g b/s) + s_g\log(d/s_g)} \end{aligned}$$

$$\begin{aligned}&\geqslant \min\left\{e\frac{\frac{s}{s_g}d\log(es_gb/s)}{s\log(es_gb/s)}, \frac{e\frac{s}{s_g}d\log(es_gb/s)}{s_g\log(ed/s_g)}\right\}\\ &\geqslant \min\left\{\frac{ed}{s_g}, \frac{\frac{ed}{s_g}}{\log(\frac{ed}{s_g})}\right\} \geqslant \left(\frac{ed}{s_g}\right)^{1/2}. \qquad (5.33)\end{aligned}$$

In the last inequality, we used $x^{1/2} \geqslant \log(x)/2$ for all $x \geqslant 1$.
Combine (5.32) and (5.33) together, we have

$$n \geqslant \frac{c^2}{2C^2}\left(s\log(es_gb/s) + s_g\log(d/s_g)\right) \geqslant c_0\left(s\log(es_gb/s) + s_g\log(d/s_g)\right) > n,$$

contradiction!

Thus, we only need to prove (5.31) based on (5.29). We still use the proof of contradiction. If

$$d^n(B,\mathbb{R}^p) < c\min\left\{\frac{1}{\sqrt{s_0}}, \left[\left(\frac{s_g}{s}\log\left(\frac{c\frac{s}{s_g}d\log(es_gb/s)}{n}\right) + \log(es_gb/s)\right)/n\right]^{1/2}\right\} := \mu,$$

then there exists a subspace L^n of $\mathbb{R}^p$ with $\dim(\mathbb{R}^p/L^n) \leqslant n$ such that for all $v \in L^n\backslash\{0\}$,

$$\|v\|_2 < \mu\left(\|v\|_1 + \sqrt{s_0}\|v\|_{1,2}\right).$$

Let $B \in \mathbb{R}^{n\times p}$ satisfying $\ker(B) = L^n$. Let $s' = \lfloor\frac{1}{32\mu^2}\rfloor, s'_g = \lfloor s'/s_0\rfloor$, by (5.29) and (5.33),

$$\frac{1}{8}s_0^{-1/2} \geqslant cs_0^{-1/2} \geqslant \mu \geqslant c\min\left\{\sqrt{\frac{C_0}{s}}, \left(\frac{\frac{s_g}{2s}\log(d/s_g) + \log(es_gb/s)}{c_0(s_g\log(d/s_g) + s\log(es_gb/s))}\right)^{1/2}\right\} \geqslant \frac{1}{4\sqrt{2}}s^{-1/2},$$

which means that

$$1 \leqslant s' \leqslant s, \quad 1 \leqslant s'_g \leqslant s_g.$$

Moreover, we have $\frac{1}{64\mu^2} < s' \leqslant \frac{1}{32\mu^2}$. For any $(2s', 2s'_g)$-sparse β with support set T and group support set G, and $\nu \in \ker(A)$, by Cauchy-Schwarz inequality,

$$\begin{aligned}\|\nu_T\|_1 + \sqrt{s_0}\|\nu_{(G)}\|_{1,2} \leqslant & \sqrt{2s'}\|\nu_T\|_2 + \sqrt{s_0}\sqrt{2s'_g}\|\nu_T\|_2 \leqslant 2\sqrt{2s'}\|\nu_T\|_2 \\ < & 2\sqrt{2}\frac{1}{4\sqrt{2}\mu}\mu\left(\|\nu\|_1 + \sqrt{s_0}\|\nu\|_{1,2}\right) = \frac{1}{2}\left(\|\nu\|_1 + \sqrt{s_0}\|\nu\|_{1,2}\right),\end{aligned}$$

i.e.,

$$\|\nu_T\|_1 + \sqrt{s_0}\|\nu_{(G)}\|_{1,2} < \|\nu_{T^c}\|_1 + \sqrt{s_0}\|\nu_{(G^c)}\|_{1,2}.$$

Based on Cauchy-Schwarz inequality and the sub-differential of $\|\beta\|_1$ and $\|\beta\|_{1,2}$, we have

$$\begin{aligned}&\|\beta+\nu\|_1 + \sqrt{s_0}\|\beta+\nu\|_{1,2} \\ \geqslant & \|\beta\|_1 + \operatorname{sgn}(\beta)^\top \nu_T + \|\nu_{T^c}\|_1 + \sqrt{s_0}\left(\|\beta\|_{1,2} + \sum_{j\in G}\frac{\beta_{(j)}^\top \nu_{(j)}}{\|\beta_{(j)}\|_2} + \sum_{j\in G^c}\|\nu_j\|_2\right) \\ \geqslant & \|\beta\|_1 - \|\nu_T\|_1 + \|\nu_{T^c}\|_1 + \sqrt{s_0}\left(\|\beta\|_{1,2} - \|\nu_{(G)}\|_2 + \|\nu_{(G^c)}\|_2\right) \\ > & \|\beta\|_1 + \sqrt{s_0}\|\beta\|_{1,2}.\end{aligned}$$

By Theorem 2.2.2,

$$n \geqslant c'(s'\log(es'_g b/s') + s'_g\log(d/s'_g)) \geqslant c's'\left\{\log\left(\frac{es_g b}{2s}\right) + \frac{1}{2s_0}\log(s_0 d/s')\right\} \geqslant c's'\log\left(\frac{es_g b}{2s}\right).$$

Thus

$$\begin{aligned}n \geqslant & c's'\left(\log\left(\frac{es_g b}{2s}\right) + \frac{s_g}{s}\log\left(\frac{c'\frac{s}{s_g}d\log(es_g b/s)}{n}\right)\right) \\ > & \frac{c'}{64\mu^2}\left(\frac{1}{4}\log(es_g b/s) + \frac{s_g}{s}\log\left(\frac{c\frac{s}{s_g}d\log(es_g b/s)}{n}\right)\right) \\ \geqslant & n\end{aligned}$$

provided that $c = \min\left\{\frac{1}{8}, c', \sqrt{\frac{c'}{256}}\right\}$, contradiction! This means that (5.31) holds if (5.29) is true.

Therefore, we have finished the proof of Theorem 2.2.3. □

5.1.6 Proof of Theorem 2.3.1

Let $\lambda = C\sigma\sqrt{\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}n}, \lambda_g = \sqrt{s/s_g}\lambda$. By (5.58) in Lemma 5.1.2 and (5.73), one has

$$\begin{aligned}
&\mathbb{P}\left(\left\|H_{\frac{1}{10}\lambda}(X^\top\varepsilon)\right\|_{\infty,2} \geqslant \frac{1}{10}\lambda_g\right)\\
\leqslant&\mathbb{P}\left(\exists 1 \leqslant j \leqslant d, \left\|H_{\frac{1}{10}\lambda}(X_{(j)}^\top\varepsilon)\right\|_2 \geqslant \frac{1}{10}\lambda_g, \|\varepsilon\|_2 \geqslant 5\sqrt{n\sigma^2}\right) + \mathbb{P}\left(\|\varepsilon\|_2 \geqslant 5\sqrt{n\sigma^2}\right)\\
\leqslant&\mathbb{P}\left(\exists 1 \leqslant j \leqslant d, \left\|H_{\frac{1}{10}\lambda}(X_{(j)}^\top\varepsilon)\right\|_2 \geqslant \frac{1}{10}\lambda_g \middle| \|\varepsilon\|_2 \geqslant 5\sqrt{n\sigma^2}\right) + \mathbb{P}\left(\|\varepsilon\|_2 \geqslant 5\sqrt{n\sigma^2}\right)\\
\leqslant&d\exp\left(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right) + e^{-n}\\
=&\exp\left(\log(s_g) + \log(d/s_g) - C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right) + e^{-n}\\
\leqslant&\exp\left(-C\frac{s\log(es_g b) + s_g\log(d/s_g)}{s_g}\right) + e^{-n}.
\end{aligned} \tag{5.34}$$

By the definition of $\hat{\beta}$ and KKT condition, we have

$$X^\top(y - X\hat{\beta}) + \lambda z_1 + \lambda_g z_2 = 0,$$

where

$$\begin{cases} (z_1)_i = \operatorname{sgn}(\hat{\beta}_i), & \hat{\beta}_i \neq 0; \\ |(z_1)_i| \leqslant 1, & \hat{\beta}_i = 0; \end{cases} \qquad \begin{cases} (z_2)_{(j)} = \frac{\hat{\beta}_{(j)}}{\|\hat{\beta}_{(j)}\|_2}, & \hat{\beta}_{(j)} \neq 0; \\ \|(z_2)_{(j)}\|_2 \leqslant 1, & \hat{\beta}_{(j)} = 0. \end{cases}$$

Therefore,

$$\|H_\lambda(X^\top(X\hat{\beta} - y))\|_{\infty,2} \leqslant \lambda_g.$$

(5.34), Lemma 5.1.5 Part 1 and the previous inequality together imply that

$$\mathbb{P}\left(\left\|H_{(1+\frac{1}{10})\lambda}(X^\top Xh)\right\|_{\infty,2} \leqslant (1+\frac{1}{10})\lambda_g\right) \geqslant 1-\exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}\right)-e^{-n}, \tag{5.35}$$

where $h=\hat{\beta}-\beta^*$. By the definition of $\hat{\beta}$, we have

$$\|y-X\hat{\beta}\|_2^2+\lambda\|\hat{\beta}\|_1+\lambda_g\|\hat{\beta}\|_{1,2} \leqslant \|y-X\beta^*\|_2^2+\lambda\|\beta^*\|_1+\lambda_g\|\beta^*\|_{1,2}.$$

(5.4) and the previous inequality show that

$$\begin{aligned} &\|Xh\|_2^2+\lambda\|h_{T^c}\|_1+\lambda_g\|h_{(G^c)}\|_{1,2} \\ \leqslant& 2\langle Xh,\varepsilon\rangle-\lambda\cdot\mathrm{sgn}(\beta_T^*)^\top h_T-\lambda_g\sum_{j\in G}\frac{\beta_{T,(j)}^{*\top}h_{(j)}}{\|\beta_{T,(j)}^*\|_2}+2\lambda\|\beta_{T^c}^*\|_1+2\lambda_g\|\beta_{T^c}^*\|_{1,2}. \end{aligned} \tag{5.36}$$

First, we consider $\langle Xh,\varepsilon\rangle$. Denote $P=X_T(X_T^\top X_T)^{-1}X_T^\top$, since $Xh=X_Th_T+X_{T^c}h_{T^c}$ and $(I_n-P)X_T=0$,

$$\begin{aligned} &|\langle Xh,\varepsilon\rangle| \leqslant |\langle PXh,\varepsilon\rangle|+|\langle (I_n-P)Xh,\varepsilon\rangle| \\ =&\left|\langle X_T^\top Xh,(X_T^\top X_T)^{-1}X_T^\top\varepsilon\rangle\right|+|\langle (I_n-P)X_{T^c}h_{T^c},\varepsilon\rangle| \\ =&\left|\langle X_T^\top Xh,(X_T^\top X_T)^{-1}X_T^\top\varepsilon\rangle\right|+|\langle X_{T^c}h_{T^c},(I_n-P)\varepsilon\rangle|. \end{aligned} \tag{5.37}$$

Therefore, to give an upper bound of $|\langle Xh,\varepsilon\rangle|$, we only need to bound $\left|\langle X_T^\top Xh,(X_T^\top X_T)^{-1}X_T^\top\varepsilon\rangle\right|$ and $|\langle X_{T^c}h_{T^c},(I_n-P)\varepsilon\rangle|$, respectively. By Part 1 of Lemma 5.1.4 and also notice that $c_{\min}\leqslant\sigma_{\min}(\Sigma)\leqslant\sigma_{\max}(\Sigma)\leqslant C_{\max}$,

$$\begin{aligned} \mathbb{P}\left(\left\|\left(\frac{1}{n}X_T^\top X_T\right)^{-1}\right\|\geqslant\frac{2}{c_{\min}}\right) \leqslant&\mathbb{P}\left(\|\frac{1}{n}X_T^\top X_T-\Sigma_{T,T}\|\geqslant\frac{c_{\min}}{2}\right) \\ \leqslant&\mathbb{P}\left(\|\frac{1}{n}X_T^\top X_T\Sigma_{T,T}^{-1}-I_s\|\geqslant\frac{c_{\min}}{2C_{\max}}\right) \\ \leqslant&2\exp\left(Cs-cn\right)\leqslant 2\exp\left(-cn\right). \end{aligned} \tag{5.38}$$

(5.38), Lemma 5.1.6 and Cauchy-Schwarz inequality together imply that with probability at least $1 - \exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}\right) - 2\exp(-cn)$,

$$\begin{aligned}\|(X_T^\top X_T)^{-1}X_T^\top \varepsilon\|_1 \leqslant & \frac{2}{c_{\min}}\frac{\sqrt{s}}{n}\|X_T^\top \varepsilon\|_2 \leqslant \frac{2}{c_{\min}}\frac{s}{n}\|X_T^\top \varepsilon\|_\infty \\ \leqslant & C\frac{s}{n}\sqrt{n\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}\sigma^2} \leqslant C\frac{s}{n}\lambda,\end{aligned}$$

$$\|(X_T^\top X_T)^{-1}X_T^\top \varepsilon\|_{1,2} \leqslant \sqrt{s_g}\|(X_T^\top X_T)^{-1}X_T^\top \varepsilon\|_2 \leqslant \frac{2}{c_{\min}}\frac{\sqrt{s_g}}{n}\|X_T^\top \varepsilon\|_2 \leqslant C\frac{\sqrt{s\cdot s_g}}{n}\lambda.$$

Combine Lemma 5.1.5 Part 2, (5.35) and the previous two inequalities together, with probability at least $1 - 2\exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}\right) - 3e^{-cn}$,

$$\begin{aligned}\left|\langle X_T^\top Xh, (X_T^\top X_T)^{-1}X_T^\top \varepsilon\rangle\right| \leqslant & \frac{11}{10}\lambda\|(X_T^\top X_T)^{-1}X_T^\top \varepsilon\|_1 + \frac{11}{10}\lambda_g\|(X_T^\top X_T)^{-1}X_T^\top \varepsilon\|_{1,2} \\ \leqslant & C\frac{s}{n}\lambda^2.\end{aligned} \tag{5.39}$$

Similarly to the proof of (5.34), also notice that $\|(I_n - P)\varepsilon\|_2 \leqslant \|\varepsilon\|_2$ and $X_{(G^c)}$ is independent of $I_n - P$, we have

$$\begin{aligned}&\mathbb{P}\left(\left\|H_{\frac{1}{10}\lambda}\left(X_{(G^c)}^\top(I_n - P)\varepsilon\right)\right\|_{\infty,2} \geqslant \frac{1}{10}\lambda_g\right) \\ \leqslant & \mathbb{P}\left(\exists j \in G^c, \left\|H_{\frac{1}{10}\lambda}\left(X_{(j)}^\top(I_n - P)\varepsilon\right)\right\|_2 \geqslant \frac{1}{10}\lambda_g \middle| \|(I_n - P)\varepsilon\|_2 \geqslant 5\sqrt{n\sigma^2}\right) \\ & + \mathbb{P}\left(\|\varepsilon\|_2 \geqslant 5\sqrt{n\sigma^2}\right) \\ \leqslant & \exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}\right) + e^{-n}.\end{aligned}$$

By Lemma 5.1.5 Part 2 and (5.34), with probability at least $1-\exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}\right)-e^{-n}$,

$$\left|\langle X_{(G^c)}h_{(G^c)}, (I_n - P)\varepsilon\rangle\right| = \left|\langle h_{(G^c)}, X_{(G^c)}^\top(I_n - P)\varepsilon\rangle\right| \leqslant \frac{1}{10}\lambda\|h_{(G^c)}\|_1 + \frac{1}{10}\lambda_g\|h_{(G^c)}\|_{1,2}.$$

Notice that $X_{T^c\backslash(G^c)}$ and $I_n - P$ are independent and $|T^c\backslash(G^c)| \leqslant |G| \leqslant s_g b$, by Lemma 5.1.6, with probability at least $1 - \exp(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}) - e^{-n}$,

$$\begin{aligned}\left|\langle X_{T^c\backslash(G^c)}h_{T^c\backslash(G^c)}, (I_n - P)\varepsilon\rangle\right| &\leqslant \|h_{T^c\backslash(G^c)}\|_1\|X^\top_{T^c\backslash(G^c)}(I_n - P)\varepsilon\|_\infty \\ &\leqslant C\sqrt{n\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}}\sigma^2\|h_{T^c\backslash(G^c)}\|_1 \\ &\leqslant \frac{1}{10}\lambda\|h_{T^c\backslash(G^c)}\|_1.\end{aligned}$$

Combine the previous two inequalities together, we have

$$\begin{aligned}|\langle X_{T^c}h_{T^c}, (I_n - P)\varepsilon\rangle| &\leqslant \left|\langle X_{(G^c)}h_{(G^c)}, (I_n - P)\varepsilon\rangle\right| + \left|\langle X_{T^c\backslash(G^c)}h_{T^c\backslash(G^c)}, (I_n - P)\varepsilon\rangle\right| \\ &\leqslant \frac{1}{10}\lambda\|h_{T^c}\|_1 + \frac{1}{10}\lambda_g\|h_{(G^c)}\|_{1,2}\end{aligned} \tag{5.40}$$

with probability $1-C\exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}\right)-Ce^{-cn}$. Combine (5.37), (5.39) and (5.40) together, we know that with probability at least $1-C\exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}\right)-Ce^{-cn}$,

$$|\langle Xh, \varepsilon\rangle| \leqslant C\frac{s}{n}\lambda^2 + \frac{1}{10}\lambda\|h_{T^c}\|_1 + \frac{1}{10}\lambda_g\|h_{(G^c)}\|_{1,2}. \tag{5.41}$$

Moreover, by the proof of Theorem 2.2.1, with probability at least $1-C\exp(-cn/s)$, there exists an approximate dual certificate $u \in \mathbb{R}^p$ in the row span of X satisfying (2.18), and $\|v_T - \mathrm{sgn}(\beta^*_T)\|_2 \leqslant \frac{1}{8}$, where v is defined in (5.1). Similarly to (5.5), we have

$$\begin{aligned}&\mathrm{sgn}(\beta^*_T)^\top h_T + \sum_{j\in G}\frac{\sqrt{s_0}\beta^{*\top}_{T,(j)}h_{(j)}}{\|\beta^*_{T,(j)}\|_2} \\ &\geqslant -\|v_T - \mathrm{sgn}(\beta^*_T)\|_2\cdot\|h_T\|_2 - \|h_{T^c}\|_1/2 - \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2 + \langle h, u\rangle \\ &\geqslant -\frac{c_{\min}}{8}\cdot\|h_T\|_2 - \|h_{T^c}\|_1/2 - \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2 + \langle h, u\rangle.\end{aligned}$$

By Lemma 5.1.7, with probability at least $1 - Ce^{-cn/s}$, $u = X^\top w$ with $\|w\|_2 \leqslant$

$C\sqrt{s/n}$. Therefore, with probability at least $1 - Ce^{-cn/s}$,

$$|\langle h, u\rangle| = |\langle Xh, w\rangle| \leqslant \|Xh\|_2\|w\|_2 \leqslant C\sqrt{s/n}\|Xh\|_2.$$

The two previous inequalities together imply that

$$\operatorname{sgn}(\beta_T^*)^\top h_T + \sum_{j\in G} \frac{\sqrt{s_0}\beta_{T,(j)}^{*\top} h_{(j)}}{\|\beta_{T,(j)}^*\|_2} \geqslant -\frac{c_{\min}}{8}\cdot\|h_T\|_2 - \|h_{T^c}\|_1/2 - \sqrt{s_0}\|h_{(G^c)}\|_{1,2}/2 - C\sqrt{s/n}\|Xh\|_2 \tag{5.42}$$

with probability at least $1 - Ce^{-cn/s}$.

Combine (5.36), (5.41) and (5.42) together, with probability at least $1 - Ce^{-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}} - Ce^{-cn/s}$,

$$\begin{aligned} &\|Xh\|_2^2 + \frac{3}{10}\lambda\|h_{T^c}\|_1 + \frac{3}{10}\lambda_g\|h_{(G^c)}\|_{1,2} \\ \leqslant &C\frac{s}{n}\lambda^2 + \frac{c_{\min}}{8}\lambda\|h_T\|_2 + C\sqrt{s/n}\lambda\|Xh\|_2 + 2\lambda\|\beta_{T^c}^*\|_1 + 2\lambda_g\|\beta_{T^c}^*\|_{1,2}. \end{aligned} \tag{5.43}$$

By (5.14), (5.35) and (5.38), with probability at least $1 - \exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}\right) - Ce^{-cn}$,

$$\begin{aligned} \|h_T\|_2 \leqslant &\|(X_T^\top X_T)^{-1}\|\|X_T^\top X_T h_T\|_2 \\ \leqslant &\frac{2}{c_{\min} n}\|X_T^\top Xh - X_T^\top X_{T^c} h_{T^c}\|_2 \\ \leqslant &\frac{2}{c_{\min} n}\left(\|X_T^\top Xh\|_2 + \|X_T^\top X_{T^c} h_{T^c}\|_2\right) \\ \leqslant &\frac{2}{c_{\min} n}\left(\|H_{\frac{11}{10}\lambda}(X_T^\top Xh)\|_2 + \frac{11}{10}\sqrt{s}\lambda + n\sum_{i\in T^c}\|X_T^\top X_i/n\|_2|h_i|\right) \\ \leqslant &\frac{2}{c_{\min} n}\left(\sqrt{s_g}\|H_{\frac{11}{10}\lambda}(X_T^\top Xh)\|_{\infty,2} + \frac{11}{10}\sqrt{s}\lambda + n\max_{i\in T^c}\|X_T^\top X_i/n\|_2\|h_{T^c}\|_1\right) \\ \leqslant &\frac{2}{c_{\min} n}\left(\sqrt{s_g}\frac{11}{10}\lambda_g + \frac{11}{10}\sqrt{s}\lambda + \frac{n}{2}\|h_{T^c}\|_1\right) \\ \leqslant &\frac{5}{c_{\min}}\frac{\sqrt{s}}{n}\lambda + \frac{1}{c_{\min}}\|h_{T^c}\|_1. \end{aligned} \tag{5.44}$$

The fourth inequality comes from $\|x\|_2 \leqslant \|H_\alpha(x)\|_2 + \sqrt{s}\alpha$ for $x \in \mathbb{R}^s$; the fifth inequality holds since $\|X_T^\top Xh\|_{0,2} \leqslant s_g$.
(5.43) and (5.44) together imply that

$$\|Xh\|_2^2 + \frac{7}{40}\lambda\|h_{T^c}\|_1 + \frac{3}{10}\lambda_g\|h_{(G^c)}\|_{1,2} \leqslant C\frac{s}{n}\lambda^2 + C\sqrt{s/n}\lambda\|Xh\|_2 + 2\lambda\|\beta^*_{T^c}\|_1 + 2\lambda_g\|\beta^*_{T^c}\|_{1,2}$$

with probability at least $1 - C\exp(-C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s}) - Ce^{-cn/s}$. Also notice that

$$C\sqrt{s/n}\lambda\|Xh\|_2 \leqslant \|Xh\|_2^2 + C\frac{s}{n}\lambda^2,$$

with probability at least $1 - C\exp(-C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s}) - Ce^{-cn/s}$,

$$\|h_{T^c}\|_1 + \sqrt{s_0}\|h_{(G^c)}\|_{1,2} \leqslant C\left(\frac{s}{n}\lambda + \|\beta^*_{T^c}\|_1 + \sqrt{s_0}\|\beta^*_{T^c}\|_{1,2}\right). \tag{5.45}$$

From the proof of Lemma 2.2.1, we know that (5.8) and (5.13) hold with probability at least $1 - 2e^{-cn}$. By Lemma 5.1.5 Part 2 and (5.35), with probability at least $1 - \exp\left(-C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s_g}\right) - e^{-n}$,

$$\begin{aligned} \left|\langle X_{\tilde{T}}h_{\tilde{T}}, Xh\rangle\right| = &\left|\langle h_{\tilde{T}}, X_{\tilde{T}}^\top Xh\rangle\right| \leqslant \frac{11}{10}\left(\lambda\|h_{\tilde{T}}\|_1 + \lambda_g\|h_{\tilde{T}}\|_{1,2}\right) \\ \leqslant &\frac{11}{10}\left(\lambda\cdot\sqrt{3s}\|h_{\tilde{T}}\|_2 + \lambda_g\sqrt{2s_g}\|h_{\tilde{T}}\|_2\right) \leqslant 4\lambda\sqrt{s}\|h_{\tilde{T}}\|_2. \end{aligned} \tag{5.46}$$

The second inequality is due to $\|h_{\tilde{T}}\|_0 \leqslant 3s$, $\|h_{\tilde{T}}\|_{0,2} \leqslant 2s_g$ and Cauchy-Schwarz inequality.

Combine (5.8), (5.13), (5.45) and (5.46) together, with probability at least $1 - C\exp(-C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s}) - Ce^{-cn/s}$, we have

$$\begin{aligned} \frac{c_{\min}}{2}\|h_{\tilde{T}}\|_2^2 \leqslant &\frac{1}{n}4\lambda\sqrt{s}\|h_{\tilde{T}}\|_2 + \sqrt{3}C_{\max}\|h_{\tilde{T}}\|_2(\sqrt{2}s^{-1/2}\|h_{T^c}\|_1 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}) \\ \leqslant &\frac{1}{n}4\lambda\sqrt{s}\|h_{\tilde{T}}\|_2 + \sqrt{3}C_{\max}\|h_{\tilde{T}}\|_2\cdot\frac{C}{\sqrt{s}}\left(\frac{s}{n}\lambda + \|\beta^*_{T^c}\|_1 + \sqrt{s_0}\|\beta^*_{T^c}\|_{1,2}\right) \end{aligned}$$

$$\leqslant C\left(\frac{\sqrt{s}}{n}\lambda + \frac{1}{\sqrt{s}}\|\beta^*_{T^c}\|_1 + \frac{1}{\sqrt{s_g}}\|\beta^*_{T^c}\|_{1,2}\right)\|h_{\tilde{T}}\|_2.$$

Therefore, with probability at least $1 - C\exp(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}) - Ce^{-cn/s}$,

$$\|h_{\tilde{T}}\|_2 \leqslant C\left(\frac{\sqrt{s}}{n}\lambda + \frac{1}{\sqrt{s}}\|\beta^*_{T^c}\|_1 + \frac{1}{\sqrt{s_g}}\|\beta^*_{T^c}\|_{1,2}\right). \tag{5.47}$$

By (5.11), (5.12), (5.45) and the previous inequality, also notice that $e^{-cn/s} \leqslant e^{-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}}$, with probability at least $1 - C\exp(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s})$,

$$\begin{aligned}\|h\|_2 \leqslant &\|h_{\tilde{T}}\|_2 + \sum_{i\geqslant 2}\|h_{T_i}\|_2 + \sum_{j\geqslant 2}\|h_{R_j}\|_2 \leqslant \|h_{\tilde{T}}\|_2 + \sqrt{2}s^{-1/2}\|h_{T^c}\|_2 + s_g^{-1/2}\|h_{(G^c)}\|_{1,2}\\ \leqslant &C\left(\frac{\sqrt{s}}{n}\lambda + \frac{1}{\sqrt{s}}\|\beta^*_{T^c}\|_1 + \frac{1}{\sqrt{s_g}}\|\beta^*_{T^c}\|_{1,2}\right),\end{aligned} \tag{5.48}$$

i.e., with probability at least $1 - C\exp(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s})$,

$$\|h\|_2 \leqslant C\left(\sqrt{\frac{\sigma^2(s_g\log(d/s_g) + s\log(es_g b))}{n}} + \frac{1}{\sqrt{s}}\|\beta^*_{T^c}\|_1 + \frac{1}{\sqrt{s_g}}\|\beta^*_{T^c}\|_{1,2}\right).$$

Moreover, if β^* is (s, s_g)-sparse, then $\|\beta^*_{T^c}\|_1 = \|\beta^*_{T^c}\|_{1,2} = 0$. Therefore, with probability at least $1 - C\exp(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s})$,

$$\|h\|_2^2 \leqslant \frac{C\sigma^2(s_g\log(d/s_g) + s\log(es_g b))}{n}.$$

□

5.1.7 Proof of Theorem 2.3.2

First, we consider the case that $d \geqslant 3s_g$ and $b \geqslant 3s/s_g$. Let $\omega^{(1)}, \dots, \omega^{(N)}$ be uniformly randomly vectors from

$$A = \{\omega \in \{0,1\}^{db} | \sum_j 1_{\{\omega_{(j)} \neq 0\}} = s_g, \|\omega_{(j)}\|_0 = \lfloor s/s_g \rfloor \text{ if } \omega_{(j)} \neq 0\}.$$

Denote $\Omega^{(i)} = \{j | \omega_j^{(i)} \neq 0\}$, $\Omega^{(i)}_{(k)} = \{j | j \in (k), \omega_j^{(i)} \neq 0\}$ and $\beta^{(i)} = \tau\omega^{(i)}$, for all $1 \leqslant i \leqslant N, 1 \leqslant k \leqslant d$, where τ is a parameter that will be specified later. Obviously, $\|\beta^{(i)}\|_0 = s_g \lfloor s/s_g \rfloor \leqslant s$, therefore $\|\beta^{(i)} - \beta^{(j)}\|_2^2 \leqslant 2s_g\lfloor s/s_g \rfloor \tau^2 \leqslant 2s\tau^2$.

Moreover, if $|\Omega^{(i)} \cap \Omega^{(j)}| \geqslant 8s_g\lfloor s/s_g \rfloor/9$, then we must have

$$\left| \left\{ k | \omega^{(i)}_{(k)}, \omega^{(j)}_{(k)} \neq 0, \left| \Omega^{(i)}_{(k)} \cap \Omega^{(j)}_{(k)} \right| \geqslant 2\lfloor s/s_g \rfloor/3 \right\} \right| \geqslant 2s_g/3,$$

otherwise $|\Omega^{(i)} \cap \Omega^{(j)}| \leqslant \frac{2s_g}{3}\lfloor s/s_g \rfloor + \frac{s_g}{3}2\lfloor s/s_g \rfloor/3 \leqslant 8s_g\lfloor s/s_g \rfloor/9$, which is a contradiction.

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\|\beta^{(i)} - \beta^{(j)}\|_2^2 \leqslant 2s_g\lfloor s/s_g \rfloor\tau^2/9\right) \\
=&\mathbb{P}\left(|\Omega^{(i)} \cap \Omega^{(j)}| \geqslant 8s_g\lfloor s/s_g \rfloor/9\right) \\
\leqslant&\mathbb{P}\left(\left| \left\{ k | \omega^{(i)}_{(k)}, \omega^{(j)}_{(k)} \neq 0, \left| \Omega^{(i)}_{(k)} \cap \Omega^{(j)}_{(k)} \right| \geqslant 2\lfloor s/s_g \rfloor/3 \right\} \right| \geqslant 2s_g/3\right) \\
\leqslant&\frac{\sum_{l=\lceil 2s_g/3 \rceil}^{s_g} \binom{s_g}{l} \left[\sum_{t=\lceil 2\lfloor s/s_g \rfloor/3 \rceil}^{\lfloor s/s_g \rfloor} \binom{\lfloor s/s_g \rfloor}{t}\binom{b-\lfloor s/s_g \rfloor}{\lfloor s/s_g \rfloor - t}\right]^l \binom{b}{\lfloor s/s_g \rfloor}^{s_g - l}\binom{d-l}{s_g-l}}{\binom{d}{s_g}\binom{b}{\lfloor s/s_g \rfloor}^{s_g}} \qquad (5.49) \\
=&\sum_{l=\lceil 2s_g/3 \rceil}^{s_g} \binom{s_g}{l} \frac{\binom{d-l}{s_g-l}}{\binom{d}{s_g}} \cdot \left[\sum_{t=\lceil 2\lfloor s/s_g \rfloor/3 \rceil}^{\lfloor s/s_g \rfloor} \binom{\lfloor s/s_g \rfloor}{t} \frac{\binom{b-\lfloor s/s_g \rfloor}{\lfloor s/s_g \rfloor - t}}{\binom{b}{\lfloor s/s_g \rfloor}}\right]^l.
\end{aligned}$$

Note that

$$\frac{\binom{d-l}{s_g-l}}{\binom{d}{s_g}} = \frac{\frac{(d-l)\cdots(d-s_g+1)}{(s_g-l)!}}{\frac{d(d-1)\cdots(d-s_g+1)}{s_g!}} = \frac{s_g(s_g-1)\cdots(s_g-l+1)}{d(d-1)\cdots(d-l+1)} \leqslant \left(\frac{s_g}{d}\right)^l,$$

The inequality holds since $\frac{s_g-i}{d-i} \leqslant \frac{s_g}{d}$ for all $1 \leqslant i \leqslant s_g$.
Similarly, for $1 \leqslant t \leqslant \lfloor s/s_g \rfloor$,

$$\frac{\binom{b-\lfloor s/s_g\rfloor}{\lfloor s/s_g\rfloor-t}}{\binom{b}{\lfloor s/s_g\rfloor}} \leqslant \frac{\binom{b-t}{\lfloor s/s_g\rfloor-t}}{\binom{b}{\lfloor s/s_g\rfloor}} \leqslant \left(\frac{\lfloor s/s_g\rfloor}{b}\right)^t.$$

Combine (5.49) and the previous two inequalities together, we have

$$\begin{aligned}
&\mathbb{P}\left(\|\beta^{(i)}-\beta^{(j)}\|_2^2 \leqslant 2s_g\lfloor s/s_g\rfloor\tau^2/9\right)\\
\leqslant &\sum_{l=\lceil 2s_g/3\rceil}^{s_g}\binom{s_g}{l}\left(\frac{s_g}{d}\right)^l\cdot\left[\sum_{t=\lceil 2\lfloor s/s_g\rfloor/3\rceil}^{\lfloor s/s_g\rfloor}\binom{\lfloor s/s_g\rfloor}{t}\left(\frac{\lfloor s/s_g\rfloor}{b}\right)^t\right]^l\\
\leqslant &\sum_{l=\lceil 2s_g/3\rceil}^{s_g}\binom{s_g}{l}\left(\frac{s_g}{d}\right)^l\cdot\left[\sum_{t=\lceil 2\lfloor s/s_g\rfloor/3\rceil}^{\lfloor s/s_g\rfloor}\binom{\lfloor s/s_g\rfloor}{t}\left(\frac{\lfloor s/s_g\rfloor}{b}\right)^{2\lfloor s/s_g\rfloor/3}\right]^l\\
\leqslant &\sum_{l=\lceil 2s_g/3\rceil}^{s_g}\binom{s_g}{l}\left(\frac{s_g}{d}\right)^l\cdot\left[2^{\lfloor s/s_g\rfloor}\left(\frac{\lfloor s/s_g\rfloor}{b}\right)^{2\lfloor s/s_g\rfloor/3}\right]^l\\
\leqslant &\sum_{l=\lceil 2s_g/3\rceil}^{s_g}\binom{s_g}{l}\left(\frac{s_g}{d}\right)^{2s_g/3}\cdot\left[\left(\frac{2\sqrt{2}\lfloor s/s_g\rfloor}{b}\right)^{2\lfloor s/s_g\rfloor/3}\right]^{2s_g/3}\\
\leqslant &\left(\frac{2\sqrt{2}s_g}{d}\right)^{2s_g/3}\cdot\left(\frac{2\sqrt{2}\lfloor s/s_g\rfloor}{b}\right)^{2s/9}.
\end{aligned} \tag{5.50}$$

Set $N = \left\lfloor \left(\frac{d}{2\sqrt{2}s_g}\right)^{s_g/3}\left(\frac{b}{2\sqrt{2}\lfloor s/s_g\rfloor}\right)^{s/9}\right\rfloor$, then

$$\mathbb{P}\left(\forall 1 \leqslant i \neq j \leqslant N, \|\beta^{(i)}-\beta^{(j)}\|_2^2 > 2s_g\lfloor s/s_g\rfloor\tau^2/9\right)$$

$$\begin{aligned}&\geqslant 1 - \frac{N(N-1)}{2}\left(\frac{2\sqrt{2}s_g}{d}\right)^{2s_g/3} \cdot \left(\frac{2\sqrt{2}\lfloor s/s_g\rfloor}{b}\right)^{2s/9}\\ &> 0.\end{aligned}$$

i.e., the probability that $\beta^{(1)}, \ldots, \beta^{(N)}; \Omega^{(1)}, \cdots, \Omega^{(N)}$ satisfy

$$\frac{s}{9}\tau^2 < 2s_g\lfloor s/s_g\rfloor\tau^2/9 < \min_{i\neq j}\|\beta^{(i)} - \beta^{(j)}\|_2^2 \leqslant 2s\tau^2, \tag{5.51}$$

$$|\Omega^{(i)} \cap \Omega^{(j)}| < 8s_g\lfloor s/s_g\rfloor/9, \quad \forall 1 \leqslant i < j \leqslant N \tag{5.52}$$

is positive. For convenience, we fix $\beta^{(1)}, \ldots, \beta^{(N)}$ to be the vectors satisfying (5.51).

Denote $y^{(i)} = X\beta^{(i)} + \varepsilon$ for all $1 \leqslant i \leqslant N$. We consider the Kullback-Leibler divergence between different distribution pairs:

$$D_{KL}\left((y^{(i)}, X), (y^{(j)}, X)\right) = \mathbb{E}_{(y^{(j)}, X)}\left[\log\left(\frac{p(y^{(i)}, X)}{p(y^{(j)}, X)}\right)\right],$$

where $p(y^{(i)}, X)$ is the probability density of $(y^{(i)}, X)$. Conditioning on X, we have

$$\mathbb{E}_{(y^{(j)}, X)}\left[\log\left(\frac{p(y^{(i)}, X)}{p(y^{(j)}, X)}\right)|X\right] = \frac{\|X(\beta^{(i)} - \beta^{(j)})\|_2^2}{2\sigma^2}.$$

Thus for $1 \leqslant i \neq j \leqslant N$,

$$\begin{aligned}D_{KL}\left((y^{(i)}, X), (y^{(j)}, X)\right) =& \mathbb{E}_X\frac{\|X(\beta^{(i)} - \beta^{(j)})\|_2^2}{2\sigma^2} = \frac{n(\beta^{(i)} - \beta^{(j)})^\top\Sigma(\beta^{(i)} - \beta^{(j)})}{2\sigma^2}\\ \leqslant& \frac{3n\|\beta^{(i)} - \beta^{(j)}\|_2^2}{4\sigma^2} \leqslant \frac{3ns\tau^2}{2\sigma^2}.\end{aligned} \tag{5.53}$$

In the first inequality, we used $\sigma_{\max}(\Sigma) \leqslant \frac{3}{2}$.

By generalized Fano's Lemma,

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathbb{F}_{s,s_g}} \mathbb{E}\|\hat{\beta} - \beta\|_2 \geqslant \frac{\sqrt{s\tau^2/9}}{2}\left(1 - \frac{\frac{3ns\tau^2}{2\sigma^2} + \log 2}{\log N}\right).$$

Since $\log N \asymp s_g \log(\frac{d}{s_g}) + s \log\left(\frac{es_g b}{s}\right)$, by setting $\tau = c\sqrt{\frac{\sigma^2\left(s_g \log(\frac{d}{s_g}) + s\log\left(\frac{es_g b}{s}\right)\right)}{ns}}$, we have

$$\inf_{\hat{\beta}} \sup_{\beta \in \mathbb{F}_{s,s_g}} \mathbb{E}\|\hat{\beta} - \beta\|_2^2 \geqslant \left(\inf_{\hat{\beta}} \sup_{\beta \in \mathbb{F}_{s,s_g}} \mathbb{E}\|\hat{\beta} - \beta\|_2\right)^2 \geqslant c\frac{\sigma^2\left(s_g \log(d/s_g) + s\log(es_g b/s)\right)}{n}.$$

If $d < 3s_g$ or $b < 3s/s_g$, let $s_g' = [s_g/3] \vee 1 \geqslant s_g/5, s' = [s/15] \vee s_g'$, then $d \geqslant 3s_g'$ and $b \geqslant 3s'/s_g'$. Similarly to (5.28), we have

$$\begin{aligned}\inf_{\hat{\beta}} \sup_{\beta \in \mathbb{F}_{s,s_g}} \mathbb{E}\|\hat{\beta} - \beta\|_2^2 \geqslant & \inf_{\hat{\beta}} \sup_{\beta \in \mathbb{F}_{s',s_g'}} \mathbb{E}\|\hat{\beta} - \beta\|_2^2 \geqslant c\frac{\sigma^2\left(s_g' \log(d/s_g') + s' \log(es_g' b/s')\right)}{n} \\ \geqslant & c'\frac{\sigma^2\left(s_g \log(d/s_g) + s\log(es_g b/s)\right)}{n}.\end{aligned}$$

□

5.1.8 Proof of Theorem 2.3.3

The proof of Theorem 2.3.3 relies on the following key lemma, which shows that Σ^{-1} is in the feasible set of the optimization problem (2.23) with high probability by choosing appropriate α and γ.

Lemma 5.1.1. *By setting* $\alpha = C\sqrt{\frac{s\log(es_g b) + s_g \log(d/s_g)}{sn}}, \gamma = \sqrt{\frac{s}{s_g}}\alpha$ *in* (2.23), *we have*

$$\mathbb{P}\left(\max_{1 \leqslant i \leqslant p} \|H_\alpha(e_i - \frac{1}{n}X^\top X\Sigma^{-1}e_i)\|_{\infty,2} \leqslant \gamma\right) \geqslant 1 - 4\exp\left(-C\frac{s\log(es_g b) + s_g \log(d/s_g)}{s_g}\right).$$

Note that $Y = X\beta^* + \varepsilon$, we have

$$\sqrt{n}(\hat{\beta}^u - \beta^*) = \sqrt{n}\left(\hat{\beta} - \beta^* + \frac{1}{n}\hat{M}X^\top\left(Y - X\hat{\beta}\right)\right) = \sqrt{n}\left(I - \frac{1}{n}\hat{M}X^\top X\right)(\hat{\beta} - \beta^*) + \frac{1}{\sqrt{n}}\hat{M}X^\top\varepsilon.$$

Since $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, we know that

$$\frac{1}{\sqrt{n}}\hat{M}X^\top\varepsilon|X \sim N\left(0, \hat{M}\hat{\Sigma}\hat{M}^\top\right).$$

Denote $h = \hat{\beta} - \beta^*$. Since β^* is (s, s_g)-sparse, by (5.45), (5.48) and Cauchy-Schwarz inequality, with probability at least $1 - C\exp(-C\frac{s\log(es_gb) + s_g\log(d/s_g)}{s})$,

$$\|h\|_1 \leqslant \|h_T\|_1 + \|h_{T^c}\|_1 \leqslant \sqrt{s}\|h_T\|_2 + \|h_{T^c}\|_1 \leqslant \sqrt{s}\|h\|_2 + \|h_{T^c}\|_1 \leqslant C\frac{s}{n}\lambda.$$

$$\|h\|_{1,2} \leqslant \|h_{(G)}\|_{1,2} + \|h_{(G^c)}\|_{1,2} \leqslant \sqrt{s_g}\|h_{(G)}\|_2 + \|h_{(G^c)}\|_{1,2} \leqslant \sqrt{s_g}\|h\|_2 + \|h_{(G^c)}\|_{1,2} \leqslant C\frac{\sqrt{s \cdot s_g}}{n}\lambda.$$

In addition, Lemma 5.1.1 shows that Σ^{-1} is in the feasible set of (2.23) with probability at least $1 - C\exp(-C\frac{s\log(es_gb) + s_g\log(d/s_g)}{s})$. By the definition of $\hat{M}$,

$$\max_i \|H_\alpha(e_i - \hat{\Sigma}\hat{M}^\top e_i)\|_{\infty,2} = \max_i \|H_\alpha(e_i - \hat{\Sigma}\hat{m}_i)\|_{\infty,2} \leqslant \gamma. \tag{5.54}$$

Combining these facts, by Lemma 5.1.5 Part 2, we must have

$$\begin{aligned}\left\|(I - \frac{1}{n}\hat{M}XX^\top)(\hat{\beta} - \beta^*)\right\|_\infty =& \max_i \left|\langle e_i - \hat{\Sigma}\hat{M}^\top e_i, h\rangle\right| \\ \leqslant& \alpha\|h\|_1 + \gamma\|h\|_{1,2} \\ \leqslant& C\frac{s}{n}\alpha\lambda + C\frac{\sqrt{s \cdot s_g}}{n}\gamma\lambda \\ =& \frac{C(s\log(es_gb) + s_g\log(d/s_g))}{n}\sigma\end{aligned}$$

with probability at least $1 - C\exp(-C\frac{s\log(es_gb) + s_g\log(d/s_g)}{s})$. This has finished the proof of (2.24).

Next, we consider $\hat{m}_i^\top \hat{\Sigma} \hat{m}_i$. By (5.54) and Lemma 5.1.5 Part 2, we have

$$1 - \langle e_i, \hat{\Sigma}\hat{m}_i \rangle = \langle e_i, e_i - \hat{\Sigma}\hat{m}_i \rangle \leqslant \alpha \|e_i\|_1 + \gamma \|e_i\|_{1,2} = \alpha + \gamma.$$

Therefore, for any $c \geqslant 0$,

$$\hat{m}_i^\top \hat{\Sigma} \hat{m}_i \geqslant \hat{m}_i^\top \hat{\Sigma} \hat{m}_i + c(1-\alpha-\gamma) - c\langle e_i, \hat{\Sigma}\hat{m}_i \rangle \geqslant \min_m \left\{ m^\top \hat{\Sigma} m + c(1-\alpha-\gamma) - c\langle e_i, \hat{\Sigma} m \rangle \right\}.$$

Since $m = ce_i/2$ achieves the minimum of the right hand side, we have

$$\hat{m}_i^\top \hat{\Sigma} \hat{m}_i \geqslant c(1-\alpha-\gamma) - \frac{c^2}{4} \hat{\Sigma}_{i,i}.$$

If $\hat{\Sigma}_{ii} > 0$ for all $1 \leqslant i \leqslant p$, by setting $c = 2(1-\alpha-\gamma)/\hat{\Sigma}_{i,i}$, we have

$$\hat{m}_i^\top \hat{\Sigma} \hat{m}_i \geqslant \frac{(1-\alpha-\gamma)^2}{\hat{\Sigma}_{i,i}}, \quad \forall 1 \leqslant i \leqslant p. \tag{5.55}$$

Moreover, by Lemma 5.1.3 Part 2 with $u = v = e_i$, we have

$$\mathbb{P}\left(\left| \hat{\Sigma}_{i,i} - \Sigma_{i,i} \right| \geqslant \frac{c_{\min}}{2} \right) \leqslant 2 \exp(-cn).$$

By the union bound,

$$\begin{aligned} &\mathbb{P}\left(\exists 1 \leqslant i \leqslant p, \left| \hat{\Sigma}_{i,i} - \Sigma_{i,i} \right| \geqslant \frac{c_{\min}}{2} \right) \\ \leqslant &\sum_{i=1}^{p} \mathbb{P}\left(\left| \hat{\Sigma}_{i,i} - \Sigma_{i,i} \right| \geqslant \frac{c_{\min}}{2} \right) \\ \leqslant &db \cdot 2 \exp(-cn) \\ \leqslant &2 \exp(-cn). \end{aligned}$$

Therefore, with probability at least $1 - 2\exp(-cn)$,

$$\frac{c_{\min}}{2} \leqslant \hat{\Sigma}_{i,i} \leqslant C_{\max} + \frac{c_{\min}}{2}, \quad \forall 1 \leqslant i \leqslant p.$$

(5.55) and the previous inequality together imply that with probability at least $1 - 2\exp(-cn)$,

$$\hat{m}_i^\top \hat{\Sigma} \hat{m}_i \geqslant \frac{1}{2C_{\max}}, \quad \forall 1 \leqslant i \leqslant p.$$

(2.24) and the previous inequality together imply (2.25). □

5.1.9 Technical Lemmas

We collect all additional technical lemmas and their proofs in this section.

Lemma 5.1.2 (Bernstein-type Inequality for Soft-thresholded Sub-Gaussian Vectors)**.**

Suppose the rows of $X \in \mathbb{R}^{n \times p}$ *are independent sub-Gaussian vectors satisfying Assumption 2.2.1.* $w \in \mathbb{R}^n$ *is a fixed vector,* Ω *is a subset of* $\{1, \dots, p\}$ *with* $|\Omega| = r$*. Then*

$$\mathbb{P}\left(\left\| \sum_{k=1}^n w_k X_{k,\Omega} \right\|_2 \geqslant \sqrt{C_{\max}} \kappa \|w\|_2 \cdot \left(\sqrt{r} + \sqrt{2t} \right) \right) \leqslant \exp(-t). \tag{5.56}$$

For any fixed vector $w \in \mathbb{R}^n$ *and fixed index subset* $\Omega \subseteq \{1, \dots, p\}$ *with* $|\Omega| = r$*,*

$$\begin{aligned}\mathbb{P}\left(\left\| H_{(\delta \|w\|_2)}(w^\top X_\Omega) \right\|_2 \geqslant t\|w\|_2 \right) \leqslant & \binom{r}{\lfloor (t/\delta)^2 \rfloor \wedge r} \cdot \exp\left(-(t/(\kappa\sqrt{C_{\max}}) - (t/\delta) \wedge \sqrt{r})_+^2 / 2 \right) \\ & + \binom{r}{\lceil (t/\delta)^2 \rceil} \cdot \exp\left(-(t/(\kappa\sqrt{C_{\max}}) - \sqrt{\lceil (t/\delta)^2 \rceil})_+^2 / 2 \right).\end{aligned} \tag{5.57}$$

In particular, for any $b \geqslant r$*, if* $\bar{\lambda} = C\|w\|_2 \sqrt{\frac{s \log(es_g b) + s_g \log(d/s_g)}{s}}$*,* $\bar{\lambda}_g = \sqrt{s/s_g}\bar{\lambda}$*, we have*

$$\mathbb{P}\left(\left\| H_{\bar{\lambda}}(w^\top X_\Omega) \right\|_2 \geqslant \bar{\lambda}_g \right) \leqslant \exp\left(-C \frac{s \log(es_g b) + s_g \log(d/s_g)}{s_g} \right). \tag{5.58}$$

Proof of Lemma 5.1.2. We only need to focus on the case where $\|w\|_2 = 1$. Let $W_\Omega = X_\Omega \Sigma_{\Omega,\Omega}^{-1/2}$, immediately we know that $W_{1,\Omega}, \dots, W_{n,\Omega}$ are isotropic sub-

Gaussian distributed. Then for any fixed w, $w^\top W_\Omega$ is also an isotropic sub-Gaussian vector such that for any $\alpha \in \mathbb{R}^r$,

$$\begin{aligned}\mathbb{E}\exp\left(w^\top W_\Omega \alpha\right) =& \mathbb{E}\exp\left(w^\top X_\Omega \Sigma_{\Omega,\Omega}^{-1/2}\alpha\right) = \mathbb{E}\exp\left(w^\top X\Sigma^{-1/2}(\Sigma^{1/2})_{\cdot,\Omega}\Sigma_{\Omega,\Omega}^{-1/2}\alpha\right)\\ &\leqslant \exp\left(\kappa^2\|(\Sigma^{1/2})_{\cdot,\Omega}\Sigma_{\Omega,\Omega}^{-1/2}\alpha\|_2^2/2\right) = \exp\left(\kappa^2\|\alpha\|_2^2/2\right).\end{aligned}$$

The last equation holds since $(\Sigma^{1/2})_{\Omega,\cdot}(\Sigma^{1/2})_{\cdot,\Omega} = (\Sigma^{1/2}\Sigma^{1/2})_{\Omega,\Omega} = \Sigma_{\Omega,\Omega}$.
By the tail inequality of sub-Gaussian quadratic form ((Hsu et al., 2012, Theorem 2.1)),

$$\mathbb{P}\left(\left\|w^\top W_\Omega\right\|_2^2 \geqslant \kappa^2\left(r + 2\sqrt{rt} + 2t\right)\right) \leqslant \exp(-t).$$

By taking square-root of the previous inequality, we have

$$\mathbb{P}\left(\left\|w^\top W_\Omega\right\|_2 \geqslant \kappa\|w\|_2 \cdot \left(\sqrt{r} + \sqrt{2t}\right)\right) \leqslant \exp(-t).$$

Also note that

$$\left\|w^\top X_\Omega\right\|_2 = \left\|w^\top W_\Omega \Sigma_{\Omega,\Omega}^{1/2}\right\|_2 \leqslant \left\|\Sigma_{\Omega,\Omega}^{1/2}\right\| \left\|w^\top W_\Omega\right\|_2 \leqslant \|\Sigma\|^{1/2}\left\|w^\top W_\Omega\right\|_2 \leqslant \sqrt{C_{\max}}\left\|w^\top W_\Omega\right\|_2,$$

we obtain (5.56).

For the second part of proof, note that

$$\begin{aligned}&\mathbb{P}\left(\left\|H_\delta(w^\top X_\Omega)\right\| \geqslant t\right)\\ \leqslant& \mathbb{P}\left(\exists \Lambda \subseteq \Omega, \text{ such that all entries of } |w^\top X_\Lambda| \geqslant \delta \text{ and } \|w^\top X_\Lambda\|_2 \geqslant t\right)\\ \leqslant& \mathbb{P}\left(\exists \Lambda \subseteq \Omega, \sqrt{|\Lambda|}\delta \leqslant t, \|w^\top X_\Lambda\|_2 \geqslant t\right)\\ &+ \mathbb{P}\left(\exists \Lambda \subseteq \Omega, \sqrt{|\Lambda|}\delta > t, \text{ all entries of } |w^\top X_\Lambda| \geqslant \delta\right)\\ \leqslant& \sum_{\substack{\Lambda\subseteq\Omega\\ |\Lambda| = \lfloor (t/\delta)^2\rfloor \wedge r}} \mathbb{P}\left(\|w^\top X_\Lambda\|_2 \geqslant t\right) + \sum_{\substack{\Lambda\subseteq\Omega\\ |\Lambda| = \lceil (t/\delta)^2\rceil}} \mathbb{P}\left(\|w^\top X_\Lambda\|_2 \geqslant t\right).\end{aligned}$$

By the first part of this lemma,

$$\mathbb{P}\left(\|w^\top X_\Lambda\|_2 \geqslant \sqrt{C_{\max}}\kappa\|w\|_2 t\right) \leqslant \exp\left(-\left(t-\sqrt{|\Lambda|}\right)_+^2/2\right).$$

Plug in this to the previous inequality, one has

$$\begin{aligned}\mathbb{P}\left(\left\|H_\delta(w^\top X_\Omega)\right\| \geqslant t\right) \leqslant & \binom{r}{\lfloor (t/\delta)^2\rfloor \wedge r} \cdot \exp\left(-(t/(\kappa\sqrt{C_{\max}}) - (t/\delta)\wedge\sqrt{r})_+^2/2\right) \\ & + \binom{r}{\lceil (t/\delta)^2\rceil} \cdot \exp\left(-(t/(\kappa\sqrt{C_{\max}}) - \sqrt{\lceil (t/\delta)^2\rceil})_+^2/2\right).\end{aligned}$$

Specifically, if $\delta = C\sqrt{\frac{s\log(es_g b)+s_g\log(d/s_g)}{s}}$, $t=\sqrt{s/s_g}\delta$,

$$\begin{aligned}t/(\kappa\sqrt{C_{\max}}) - \sqrt{\lceil (t/\delta)^2\rceil} \geqslant & C\sqrt{\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}} - \sqrt{2\frac{s}{s_g}} \\ \geqslant & C\sqrt{\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}}.\end{aligned}$$

Therefore, (5.57) shows that

$$\begin{aligned}\mathbb{P}\left(\left\|H_{\bar\lambda}(w^\top X_\Omega)\right\|_2 \geqslant \bar\lambda_g\right) \leqslant & r^{\lfloor (t/\delta)^2\rfloor}\exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}\right) \\ & + r^{\lceil (t/\delta)^2\rceil}\exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}\right) \\ \leqslant & 2r^{2s/s_g}\exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}\right) \\ \leqslant & \exp\left(\log 2 + \frac{2s\log(eb)}{s_g} - C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}\right) \\ \leqslant & \exp\left(-C\frac{s\log(es_g b)+s_g\log(d/s_g)}{s_g}\right).\end{aligned}$$

□

Lemma 5.1.3 (sub-Gaussian quadratic form concentrations). *Suppose* $Z \in \mathbb{R}^p$ *is a sub-Gaussian vector satisfying Assumption 2.2.1.*

1. *For any fixed* $u, v \in \mathbb{R}^p, u, v \neq 0$, $u^\top ZZ^\top v$ *is sub-exponential such that for every* $t > 0$,

$$\mathbb{P}\left(\left|u^\top ZZ^\top v - \mathbb{E}u^\top ZZ^\top v\right| \geqslant t\|u\|_2\|v\|_2\right) \leqslant C\exp(-ct/\kappa^2). \tag{5.59}$$

2. *In addition, suppose* $X = [X_1^\top, \ldots, X_n^\top]^\top \in \mathbb{R}^{n\times p}$ *is a random matrix with independent random sub-Gaussian rows satisfying Assumption 2.2.1,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} u^\top X_k X_k^\top v - u^\top \Sigma v\right| \geqslant t\|u\|_2\|v\|_2\right) \leqslant 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right). \tag{5.60}$$

3. *More generally, for any fixed matrix* $U \in \mathbb{R}^{p\times r}$, *the following concentration inequality in spectral norm holds,*

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{k=1}^{n} U^\top X_k X_k^\top v - U^\top \Sigma v\right\|_2 \geqslant t\|U\|\|v\|_2\right) \leqslant 2\exp\left(Cr - cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right). \tag{5.61}$$

Proof of Lemma 5.1.3. Since we can rescale u and v without essentially changing the problem, without loss of generality we assume $\|u\|_2 = \|v\|_2 = 1$. Let $A = uv^\top$, then $u^\top ZZ^\top v = Z^\top uv^\top Z = Z^\top AZ$. By Assumption 2.2.1, $\mathbb{E}Z = 0$ and $\|\langle Z, e_i\rangle\|_{\psi_2} \leqslant C\kappa$. By Hanson-Wright inequality ((Rudelson and Vershynin, 2013, Theorem 1.1)),

$$\begin{aligned}
\mathbb{P}\left(\left|u^\top ZZ^\top v - \mathbb{E}u^\top ZZ^\top v\right| \geqslant t\right) =& \mathbb{P}\left(|Z^\top AZ - \mathbb{E}Z^\top AZ| \geqslant t\right) \\
\leqslant& 2\exp\left[-c\min\left(\frac{t^2}{\kappa^4\|A\|_{HS}^2}, \frac{t}{\kappa^2\|A\|}\right)\right] \\
\leqslant& 2\exp\left[-c\min\left(\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right)\right],
\end{aligned}$$

where

$$\|A\|_{HS} = \left(\sum_{i,j} |a_{i,j}|^2\right)^{1/2} = \left(\sum_{i,j} |u_i v_j|^2\right)^{1/2} = \|u\|_2\|v\|_2 = 1,$$

$$\|A\| = \max_{\|x\|_2\leqslant 1} \|Ax\|_2 = \max_{\|x\|_2\leqslant 1} \|uv^\top x\|_2 = \|u\|_2 \max_{\|x\|_2\leqslant 1} |v^\top x| = \|u\|_2\|v\|_2 = 1.$$

Therefore, for every $t \geqslant \kappa^2$,

$$\mathbb{P}\left(\left|u^\top ZZ^\top v - \mathbb{E}u^\top ZZ^\top v\right| \geqslant t\right) \leqslant 2\exp\left(-ct/\kappa^2\right).$$

Thus, there exists a constant $c < \log 2$, for every $t \geqslant 0$,

$$\mathbb{P}\left(\left|u^\top ZZ^\top v - \mathbb{E}u^\top ZZ^\top v\right| \geqslant t\right) \leqslant 2\exp\left(-ct/\kappa^2\right).$$

Notice that $\mathbb{E}u^\top X_k X_k^\top v = u^\top \Sigma v$ for all $1 \leqslant k \leqslant n$, by Bernstein-type concentration inequality (c.f., (Vershynin, 2010, Proposition 5.16)),

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} u^\top X_k X_k^\top v - u^\top \Sigma v\right| \geqslant t\right) \leqslant 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right).$$

This has finished the proof of (5.60).

Finally, we consider (5.61), which can be done by an ε-net argument and the result in (5.60). For any $w \in \mathbb{R}^r$, $\|w\|_2 = 1$, set $u = Uw$ in (5.60), we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^{n} w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \geqslant \frac{t}{2}\|Uw\|_2\|v\|_2\right) \leqslant 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right).$$

By (Vershynin, 2010, Lemma 5.3), we can find a $\frac{1}{2}$-net $\mathcal{N}_{\frac{1}{2}}$ of $S^{r-1} = \{x | x \in \mathbb{R}^r, \|x\|_2 = 1\}$ with $|\mathcal{N}_{\frac{1}{2}}| \leqslant 5^r$. By the union bound,

$$\mathbb{P}\left(\forall w \in \mathcal{N}_{\frac{1}{2}}, \left|\frac{1}{n}\sum_{k=1}^{n} w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \geqslant \frac{t}{2}\|Uw\|_2\|v\|_2\right)$$

$$\leqslant 5^r \cdot 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right). \tag{5.62}$$

For any $g \in \mathbb{R}^r, g \neq 0$, set $x = \frac{g}{\|g\|_2} \in \arg\max_{w\in\mathbb{R}^r,\|w\|_2=1} |w^\top g|$, we can find $y \in \mathcal{N}_{\frac{1}{2}}$ such that $\|x-y\|_2 \leqslant \frac{1}{2}$. By triangle inequality,

$$\|g\|_2 - |y^\top g| = |x^\top g| - |y^\top g| \leqslant |x^\top g - y^\top g| \leqslant \|x-y\|_2\|g\|_2 \leqslant \frac{1}{2}\|g\|_2.$$

Therefore,

$$\sup_{w\in\mathbb{R}^r,\|w\|_2=1}\left|\frac{1}{n}\sum_{k=1}^n w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \leqslant 2\sup_{w\in\mathcal{N}_{\frac{1}{2}}}\left|\frac{1}{n}\sum_{k=1}^n w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right|.$$

The (5.62) and the previous inequality together, also notice that $\|U\| = \sup_{w\in\mathbb{R}^r,\|w\|_2=1}\|Uw\|_2$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{w\in\mathbb{R}^r,\|w\|_2=1}\left|\frac{1}{n}\sum_{k=1}^n w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \geqslant t\|U\|\|v\|_2\right)\\
\leqslant&\mathbb{P}\left(\sup_{w\in\mathcal{N}_{\frac{1}{2}}}\left|\frac{1}{n}\sum_{k=1}^n w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \geqslant \frac{t}{2}\|U\|\|v\|_2\right)\\
\leqslant&\mathbb{P}\left(\forall w \in \mathcal{N}_{\frac{1}{2}}, \left|\frac{1}{n}\sum_{k=1}^n w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right| \geqslant \frac{t}{2}\|Uw\|_2\|v\|_2\right)\\
\leqslant&5^r \cdot 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right).
\end{aligned} \tag{5.63}$$

Finally, note that

$$\left\|\frac{1}{n}\sum_{k=1}^n U^\top X_k X_k^\top v - U^\top \Sigma v\right\|_2 = \sup_{w\in\mathbb{R}^r,\|w\|_2=1}\left|\frac{1}{n}\sum_{k=1}^n w^\top U^\top X_k X_k^\top v - w^\top U^\top \Sigma v\right|,$$

we have proved (5.61). □

We collect the random matrix properties of X in the following lemma. These properties will be extensively used in the main content of the paper.

Lemma 5.1.4. *Suppose* $X = [X_1^\top, \dots, X_n^\top]^\top \in \mathbb{R}^{n\times p}$ *is a random matrix with independent random sub-Gaussian rows satisfying Assumption 2.2.1.*

1. *Suppose* $T \subseteq \{1, \dots, p\}$ *is with cardinality* s. *Then,*

$$\mathbb{P}\left(\left\|\frac{1}{n}X_T^\top X_T \Sigma_{T,T}^{-1} - I_s\right\| \geqslant t\right) \leqslant 2\exp\left(Cs - cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right); \qquad (5.64)$$

2. *For any fixed vector* $\alpha \in \mathbb{R}^s$, $\delta > 0$, *and fixed index subset* $\Omega \subseteq T^c$ *satisfying* $|\Omega| = r$, $t \geqslant \delta \geqslant C(\max_{i\in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2)\|\alpha\|_2$,

$$\begin{aligned}&\mathbb{P}\left(\left\|H_\delta(\alpha^\top X_T^\top X_\Omega/n)\right\|_2 \geqslant t\right)\\ \leqslant&\binom{r}{\lfloor (t/\delta)^2\rfloor \wedge r}\exp\left(C\lfloor (t/\delta)^2\rfloor \wedge r - cn\min\left\{\frac{t^2}{\kappa^4\|\alpha\|_2^2}, \frac{t}{\kappa^2\|\alpha\|_2}\right\}\right) \qquad (5.65)\\ &+\binom{r}{\lceil (t/\delta)^2\rceil}_+\exp\left(C\lceil (t/\delta)^2\rceil - cn\min\left\{\frac{t^2}{\kappa^4\|\alpha\|_2^2}, \frac{t}{\kappa^2\|\alpha\|_2}\right\}\right).;\end{aligned}$$

Here, $H_\lambda(\cdot)$ *is the soft-thresholding estimator at level* λ.

Proof of Lemma 5.1.4.

1. The first statement is via ε-net. Denote $W_T = X_T\Sigma_{T,T}^{-1/2}$, then the rows of W_T are independent isotropic sub-Gaussian distributed. For any fixed vector $x \in S^{s-1} = \{x : x \in \mathbb{R}^s, \|x\|_2 = 1\}$, by (Vershynin, 2010, Lemma 5.5), $Z_i = \langle (W_T)_{i\cdot}^\top, x\rangle$ are independent sub-Gaussian random variables with $\mathbb{E}Z_i^2 = 1$ and $\|Z_i\|_{\psi_2} \leqslant C\kappa$. Therefore, by Remark 5.18 and Lemma 5.14 in Vershynin (2010),$\|Z_i^2 - 1\|_{\psi_1} \leqslant 2\|Z_i^2\|_{\psi_1} \leqslant 4\|Z_i\|_{\psi_2}^2 \leqslant C\kappa^2$. Bernstein-type inequality shows that

$$\mathbb{P}\left(\left|\frac{1}{n}\|W_T x\|_2^2 - 1\right| \geqslant \frac{t}{2}\right) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (Z_i^2 - 1)\right| \geqslant \frac{t}{2}\right) \leqslant 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right).$$

By (Vershynin, 2010, Lemma 5.2), we can find a $\frac{1}{4}$-net $\mathcal{N}_{\frac{1}{4}}$ of $S^{s-1} = \{x : x \in \mathbb{R}^s, \|x\|_2 = 1\}$ with $|\mathcal{N}_{\frac{1}{4}}| \leqslant 9^s$. The union bound tells us

$$\mathbb{P}\left(\max_{x\in\mathcal{N}_{\frac{1}{4}}}\left|\frac{1}{n}\|W_T x\|_2^2 - 1\right| \geqslant \frac{t}{2}\right) \leqslant 9^s \cdot 2\exp\left(-cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right). \quad (5.66)$$

By (Vershynin, 2010, Lemma 5.4),

$$\|\frac{1}{n}W_T^\top W_T - I_s\| \leqslant 2\max_{x\in\mathcal{N}_{\frac{1}{4}}}\left|\langle(\frac{1}{n}W_T^\top W_T - I_s)x, x\rangle\right| = 2\max_{x\in\mathcal{N}_{\frac{1}{4}}}\left|\frac{1}{n}\|W_T x\|_2^2 - 1\right|. \quad (5.67)$$

Since $c_{\min} \leqslant \sigma_{\min}(\Sigma) \leqslant \sigma_{\max}(\Sigma) \leqslant C_{\max}$, we have $\|\Sigma_{T,T}^{1/2}\| \leqslant \sqrt{C_{\max}}$ and $\|\Sigma_{T,T}^{-1/2}\| \leqslant 1/\sqrt{c_{\min}}$. Therefore,

$$\begin{aligned}\|\frac{1}{n}X_T^\top X_T\Sigma_{T,T}^{-1} - I_s\| =&\|\Sigma_{T,T}^{1/2}(\frac{1}{n}W_T^\top W_T - I_s)\Sigma_{T,T}^{-1/2}\| \\ \leqslant&\|\Sigma_{T,T}^{1/2}\|\|\frac{1}{n}W_T^\top W_T - I_s\|\|\Sigma_{T,T}^{-1/2}\| \\ \leqslant&\sqrt{\frac{C_{\max}}{c_{\min}}}\|\frac{1}{n}W_T^\top W_T - I_s\|.\end{aligned} \quad (5.68)$$

Combine (5.66), (5.67) and (5.68) together, we have arrived at the conclusion.

2. Now we consider the proof for (5.65). Note that $\|H_\delta(\alpha^\top X_T^\top X_\Omega)\|_2 \geqslant t$ implies that there exists $\Lambda \subset \Omega$ such that all entry of $|\alpha^\top X_T^\top X_\Lambda|$ are greater than δ,

and $\left\| |\alpha^\top X_T^\top X_\Lambda| - \delta \right\|_2 \geqslant t$. Thus,

$$\begin{aligned}
&\mathbb{P}\left(\left\|H_\delta(\alpha^\top X_T^\top X_\Omega/n)\right\|_2 \geqslant t\right)\\
\leqslant&\mathbb{P}\left(\exists\Lambda \subseteq \Omega, \text{ such that all entries of } \left|\alpha^\top X_T^\top X_\Lambda/n\right| \geqslant \delta, \text{and } \|\alpha^\top X_T^\top X_\Lambda/n\|_2 \geqslant t\right)\\
\leqslant&\mathbb{P}\left(\exists\Lambda \subseteq \Omega, \sqrt{|\Lambda|}\delta \leqslant t, \|\alpha^\top X_T^\top X_\Lambda/n\|_2 \geqslant t\right)\\
&+\mathbb{P}\left(\exists\Lambda \subseteq \Omega, \sqrt{|\Lambda|}\delta > t, \text{ all entries of } \left|\alpha^\top X_T^\top X_\Lambda/n\right| \geqslant \delta\right)\\
\leqslant&\sum_{\substack{\Lambda\subseteq\Omega\\ |\Lambda|=\lfloor (t/\delta)^2\rfloor\wedge r}} \mathbb{P}\left(\|\alpha^\top X_T^\top X_\Lambda/n\|_2 \geqslant t\right) + \sum_{\substack{\Lambda\subseteq\Omega\\ |\Lambda|=\lceil (t/\delta)^2\rceil}} \mathbb{P}\left(\text{all entries of } \left|\alpha^\top X_T^\top X_\Lambda/n\right| \geqslant \delta\right)\\
\leqslant&\sum_{\substack{\Lambda\subseteq\Omega\\ |\Lambda|=\lfloor (t/\delta)^2\rfloor\wedge r}} \mathbb{P}\left(\|\alpha^\top X_T^\top X_\Lambda/n\|_2 \geqslant t\right) + \sum_{\substack{\Lambda\subseteq\Omega\\ |\Lambda|=\lceil (t/\delta)^2\rceil}} \mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geqslant t\right).
\end{aligned} \tag{5.69}$$

Since $t \geqslant \delta \geqslant C \max_{i\in T^c} \|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2\|\alpha\|_2$, we know that no matter $|\Lambda| = \lfloor (t/\delta)^2 \rfloor \wedge r$ or $\lceil (t/\delta)^2\rceil$,

$$2C_{\max}\sqrt{\lceil (t/\delta)^2\rceil}(\max_{i\in T^c}\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2)\|\alpha\|_2 \leqslant 2C_{\max}\sqrt{2}(t/\delta)(\max_{i\in T^c}\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2)\|\alpha\|_2 \leqslant t.$$

By Part 3 of Lemma 5.1.3, for any $\Lambda \subseteq \Omega$, $t \geqslant 2C_{\max}\sqrt{|\Lambda|}\max_{i\in T^c}\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2\|\alpha\|_2$, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geqslant t\right)\\
\leqslant&\mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geqslant t - \left\|\mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2\right)\\
\leqslant&\mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geqslant t - \|\Sigma_{\Lambda,T}\alpha\|_2\right)\\
=&\mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geqslant t - \left(\sum_{i\in\Lambda}(\Sigma_{i,T}\alpha)^2\right)^{1/2}\right)\\
\leqslant&\mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geqslant t - \sqrt{|\Lambda|}\max_{i\in T^c}|\Sigma_{i,T}\alpha|\right)\\
\leqslant&\mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geqslant t - \sqrt{|\Lambda|}\max_{i\in T^c}\|\Sigma_{i,T}\Sigma_{T,T}^{-1}\|_2\|\Sigma_{T,T}\|\|\alpha\|_2\right)
\end{aligned}$$

$$\leqslant \mathbb{P}\left(\left\|\alpha^\top X_T^\top X_\Lambda/n - \mathbb{E}\alpha^\top X_T^\top X_\Lambda/n\right\|_2 \geqslant t/2\right)$$
$$\leqslant 2\exp\left(C|\Lambda| - cn\min\left\{\frac{t^2}{\kappa^4\|\alpha\|_2^2}, \frac{t}{\kappa^2\|\alpha\|_2}\right\}\right).$$

Combine (5.69) and the previous inequality, one obtains

$$\begin{aligned}
&\mathbb{P}\left(\left\|H_\delta(\alpha^\top X_T^\top X_\Omega/n)\right\|_2 \geqslant t\right)\\
\leqslant&\binom{r}{\lfloor (t/\delta)^2\rfloor \wedge r}\exp\left(C\lfloor (t/\delta)^2\rfloor \wedge r - cn\min\left\{\frac{t^2}{\kappa^4\|\alpha\|_2^2}, \frac{t}{\kappa^2\|\alpha\|_2}\right\}\right)\\
&+\binom{r}{\lceil (t/\delta)^2\rceil}_+\exp\left(C\lceil (t/\delta)^2\rceil - cn\min\left\{\frac{t^2}{\kappa^4\|\alpha\|_2^2}, \frac{t}{\kappa^2\|\alpha\|_2}\right\}\right).
\end{aligned}$$

□

Lemma 5.1.5 (Properties of Soft-thresholding). *1. Suppose* $a, b > 0$, $x, y \in \mathbb{R}$, $H_\cdot(\cdot)$ *is the soft-thresholding operator satisfying* $H_a(x) = \operatorname{sgn}(x)\cdot(|x|-a)_+$. *Then the following triangular inequality holds,*

$$|H_{a+b}(x+y)| \leqslant |H_a(x)| + |H_b(y)|. \tag{5.70}$$

2. Suppose $a, b > 0$, $x, y \in \mathbb{R}^p$, *if* $\|H_a(x)\|_{\infty,2} \leqslant b$, *then*

$$|\langle x, y\rangle| \leqslant a\|y\|_1 + b\|y\|_{1,2}. \tag{5.71}$$

Proof of Lemma 5.1.5.

1.

$$\begin{aligned}
|H_{a+b}(x+y)| =&(|x+y| - a - b)_+ \leqslant (|x| - a + |y| - b)_+ \leqslant (|x|-a)_+ + (|y|-b)_+\\
=&|H_a(x)| + |H_b(y)|.
\end{aligned}$$

2.

$$
\begin{aligned}
|\langle x, y\rangle| \leqslant & |\langle H_a(x), y\rangle| + |\langle x - H_a(x), y\rangle| = |\sum_{j=1}^{d} \langle [H_a(x)]_{(j)}, y_{(j)}\rangle| + |\langle x - H_a(x), y\rangle| \\
\leqslant & \sum_{j=1}^{d} \|[H_a(x)]_{(j)}\|_2 \|y_{(j)}\|_2 + \|x - H_a(x)\|_\infty \|y\|_1 \leqslant \|H_a(x)\|_{\infty,2}\|y\|_{1,2} + \|x - a\|y\|_1 \\
\leqslant & b\|y\|_{1,2} + a\|y\|_1.
\end{aligned}
$$

□

Lemma 5.1.6. *Suppose* $X = [X_1^\top, \ldots, X_n^\top]^\top \in \mathbb{R}^{n \times p}$ *is a random matrix with independent random sub-Gaussian rows satisfying Assumption 2.2.1,* $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$*. Suppose* $T \subseteq \{1, \ldots, p\}$ *is with cardinality* s*,* $P \in \mathbb{R}^{n \times n}$ *is a projection matrix and independent of* X_T*. Then, for any* $t \geqslant \log(es)$*,*

$$
\mathbb{P}\left(\|X_T^\top P \varepsilon\|_\infty \geqslant C\kappa\sqrt{nt}\sigma^2\right) \leqslant e^{-n} + e^{-Ct}.
$$

Proof of Lemma 5.1.6. For fixed vector $w \in \mathbb{R}^n$, since Assumption 2 is satisfied, for $i \in T$, $X_{1i}, \ldots, X_{ni}$ are independent sub-Gaussian distributed such that

$$
\begin{aligned}
\mathbb{E}\exp\left(tX_{ji}\right) = & \mathbb{E}\exp\left(te_i^\top \Sigma^{1/2}\Sigma^{-1/2}X_{j\cdot}^\top\right) \leqslant \exp\left(\frac{\kappa^2\|\Sigma^{1/2}e_i\|_2^2 t^2}{2}\right) \leqslant \exp\left(\frac{\kappa^2\Sigma_{i,i}t^2}{2}\right) \\
\leqslant & \exp\left(\frac{C_{\max}\kappa^2 t^2}{2}\right).
\end{aligned}
$$

By Hoeffding-type inequality,

$$
\mathbb{P}\left(|X_{\cdot i}^\top w| \geqslant t\|w\|_2\right) \leqslant 2\exp\left(-c\frac{t^2}{\kappa^2}\right). \tag{5.72}
$$

Moreover, by (Laurent and Massart, 2000, Lemma 1), for any $x \geqslant 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} \varepsilon_i^2 \geqslant (n + 2\sqrt{nx} + 2x)\sigma^2\right) \leqslant e^{-x}.$$

Set $x = n$ in the last inequality, we have

$$\mathbb{P}\left(\|\varepsilon\|_2 \geqslant \sqrt{5n\sigma^2}\right) \leqslant e^{-n}. \tag{5.73}$$

Combine (5.72) and (5.73) together and notice that $\|P\varepsilon\|_2 \leqslant \|\varepsilon\|_2$, we have

$$\begin{aligned}
&\mathbb{P}\left(\|X_T^\top P\varepsilon\|_\infty \geqslant C\kappa\sqrt{nt\sigma^2}\right) \leqslant \sum_{i\in T} \mathbb{P}\left(|X_{\cdot i}^\top P\varepsilon| \geqslant C\kappa\sqrt{nt\sigma^2}\right)\\
\leqslant&\mathbb{P}\left(\|P\varepsilon\|_2 \geqslant \sqrt{5n\sigma^2}\right) + \sum_{i\in T} \mathbb{P}\left(|X_{\cdot i}^\top P\varepsilon| \geqslant C\kappa\sqrt{nt\sigma^2}, \|P\varepsilon\|_2 \leqslant \sqrt{5n\sigma^2}\right)\\
\leqslant&\mathbb{P}\left(\|\varepsilon\|_2 \geqslant \sqrt{5n\sigma^2}\right) + \sum_{i\in T} \mathbb{P}\left(|X_{\cdot i}^\top P\varepsilon| \geqslant C\kappa\sqrt{nt\sigma^2}\middle| \|P\varepsilon\|_2 \leqslant \sqrt{5n\sigma^2}\right)\\
\leqslant&e^{-n} + s \cdot 2\exp\left(-Ct\right) \leqslant e^{-n} + e^{-Ct}.
\end{aligned}$$

□

Lemma 5.1.7. *With probability at least* $1 - Ce^{-cn/s}$*, the approximate dual certificate defined in* (5.17) *can be written as* $u = X^\top w$*, where* $\|w\|_2 \leqslant C\sqrt{s/n}$*.*

Proof of Lemma 5.1.7. By (5.17), we have $u = X^\top w$, where $w = (w_1^\top, \dots, w_{l_{\max}}^\top)^\top$ and $w_l = \frac{1}{n_l} X_{I_l,T}\Sigma_{T,T}^{-1} q_{l-1}$. Thus $\|w\|_2^2 = \sum_{l=1}^{l_{\max}} \|w_l\|_2^2$. Also note that

$$\begin{aligned}
\frac{1}{n_l}\|X_{I_l,T}\Sigma_{T,T}^{-1}q_{l-1}\|_2^2 =&\langle \frac{1}{n_l} X_{I_l,T}^\top X_{I_l,T}\Sigma_{T,T}^{-1}q_{l-1}, \Sigma_{T,T}^{-1}q_{l-1}\rangle\\
=&\langle(\frac{1}{n_l} X_{I_l,T}^\top X_{I_l,T}\Sigma_{T,T}^{-1} - I_{|T|})q_{l-1}, \Sigma_{T,T}^{-1}q_{l-1}\rangle + \|\Sigma_{T,T}^{-1/2}q_{l-1}\|_2^2\\
=&\langle -q_l, \Sigma_{T,T}^{-1}q_{l-1}\rangle + \|\Sigma_{T,T}^{-1/2}q_{l-1}\|_2^2\\
\leqslant&\|q_l\|_2\|\Sigma_{T,T}^{-1}q_{l-1}\|_2 + \|\Sigma_{T,T}^{-1/2}q_{l-1}\|_2^2
\end{aligned}$$

$$\leqslant \frac{1}{c_{\min}} \|q_l\|_2 \|q_{l-1}\|_2 + \frac{1}{c_{\min}} \|q_{l-1}\|_2^2 \leqslant \frac{2}{c_{\min}} \|q_{l-1}\|_2^2.$$

By (5.25), with probability at least $1 - C\exp(-cn/s)$,

$$\begin{aligned}
\|w\|_2^2 \leqslant & \sum_{l=1}^{l_{\max}} \frac{C}{n_l} \|q_{l-1}\|_2^2 \\
\leqslant & \frac{C}{n}(2\sqrt{s})^2 + \frac{C}{n}\left(2\sqrt{s/\log(es)}\right)^2 + \frac{C\log(es)}{n} \sum_{l=3}^{l_{\max}} \left(2^{4-l}\sqrt{s}/\log(es)\right)^2 \\
\leqslant & C\frac{s}{n}.
\end{aligned}$$

□

Proof of Lemma 5.1.1. For any $1 \leqslant i \leqslant p, 1 \leqslant j \leqslant d, , \Lambda \subseteq (j), |\Lambda| = k$, by Lemma 5.1.3 with

$$v = \Sigma^{-1} e_i, \quad U \in \mathbb{R}^{p\times k}, U_{[\Lambda,:]} = I, U_{[\Lambda^c,:]} = 0,$$

we have

$$\mathbb{P}\left(\left\|(e_i)_\Lambda - \frac{1}{n} X_\Lambda^\top X \Sigma^{-1} e_i\right\|_2 \geqslant t\right) \leqslant 2\exp\left(Ck - cn\min\left\{\frac{t^2}{\kappa^4}, \frac{t}{\kappa^2}\right\}\right). \qquad (5.74)$$

By the same method in Lemma 5.1.2 Part 2,

$$\begin{aligned}
&\mathbb{P}\left(\left\|H_\alpha\left((e_i)_{(j)} - \frac{1}{n} X_{(j)}^\top X \Sigma^{-1} e_i\right)\right\|_2 \geqslant \gamma\right) \\
\leqslant & \mathbb{P}\Bigg(\exists \Lambda \subseteq (j), \text{ all entries of } |(e_i)_\Lambda - \frac{1}{n} X_\Lambda^\top X \Sigma^{-1} e_i| \geqslant \alpha \text{ and } \|(e_i)_\Lambda - \frac{1}{n} X_\Lambda^\top X \Sigma^{-1} e_i\|_2 \geqslant \gamma\Bigg) \\
\leqslant & \mathbb{P}\left(\exists \Lambda \subseteq (j), \sqrt{|\Lambda|}\alpha \leqslant \gamma, \|(e_i)_\Lambda - \frac{1}{n} X_\Lambda^\top X \Sigma^{-1} e_i\|_2 \geqslant \gamma\right) \\
& + \mathbb{P}\left(\exists \Lambda \subseteq (j), \sqrt{|\Lambda|}\alpha > \gamma, \text{ all entries of } |(e_i)_\Lambda - \frac{1}{n} X_\Lambda^\top X \Sigma^{-1} e_i| \geqslant \alpha\right) \\
\leqslant & \sum_{\substack{\Lambda \subseteq (j) \\ |\Lambda| = \lfloor s/s_g \rfloor}} \mathbb{P}\left(\|(e_i)_\Lambda - \frac{1}{n} X_\Lambda^\top X \Sigma^{-1} e_i\|_2 \geqslant \gamma\right)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{\substack{\Lambda\subseteq(j)\\|\Lambda|=\lceil s/s_g\rceil}} \mathbb{P}\left(\text{all entries of } |\,(e_i)_\Lambda-\frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i|\geqslant\alpha\right)\\
\leqslant&\sum_{\substack{\Lambda\subseteq(j)\\|\Lambda|=\lfloor s/s_g\rfloor}} \mathbb{P}\left(\|\,(e_i)_\Lambda-\frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i\|_2\geqslant\gamma\right)+\sum_{\substack{\Lambda\subseteq(j)\\|\Lambda|=\lceil s/s_g\rceil}} \mathbb{P}\left(\|\,(e_i)_\Lambda-\frac{1}{n}X_\Lambda^\top X\Sigma^{-1}e_i\|_2\geqslant\gamma\right).
\end{aligned}
$$

Combine (5.74) and the previous inequality together, we have

$$
\begin{aligned}
&\mathbb{P}\left(\left\|H_\alpha\left((e_i)_{(j)}-\frac{1}{n}X_{(j)}^\top X\Sigma^{-1}e_i\right)\right\|_2\geqslant\gamma\right)\\
\leqslant&\binom{b_j}{\lfloor s/s_g\rfloor}\cdot 2\exp\left(C\lfloor s/s_g\rfloor-cn\cdot C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s_gn}\right)\\
&+\binom{b_j}{\lceil s/s_g\rceil}\cdot 2\exp\left(C\lceil s/s_g\rceil-cn\cdot C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s_gn}\right)\\
\leqslant&4\left(\frac{2es_gb}{s}\right)^{2s/s_g}\exp\left(Cs/s_g-C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s_g}\right)\\
\leqslant&4\exp\left(\frac{2s}{s_g}\log\left(\frac{2es_gb}{s}\right)+Cs/s_g-C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s_g}\right).
\end{aligned}
\tag{5.75}
$$

By (5.75) and the union bound, we have

$$
\begin{aligned}
&\mathbb{P}\left(\max_{1\leqslant i\leqslant p}\|H_\alpha(e_i-\frac{1}{n}X^\top X\Sigma^{-1}e_i)\|_{\infty,2}\leqslant\gamma\right)\\
\leqslant&\sum_{i=1}^{p}\sum_{j=1}^{d}\mathbb{P}\left(\left\|H_\alpha\left((e_i)_{(j)}-\frac{1}{n}X_{(j)}^\top X\Sigma^{-1}e_i\right)\right\|_2\geqslant\gamma\right)\\
\leqslant&d^2b\cdot 4\exp\left(\frac{2s}{s_g}\log\left(\frac{2es_gb}{s}\right)+Cs/s_g-C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s_g}\right)\\
\leqslant&4\exp\left(2\log(s_g)+2\log(d/s_g)+\frac{3s}{s_g}\log\left(2eb\right)+Cs/s_g-C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s_g}\right)\\
\leqslant&4\exp\left(-C\frac{s\log(es_gb)+s_g\log(d/s_g)}{s_g}\right).
\end{aligned}
$$

□

5.2 Appendix to Chapter 3

In this section, we provide the optimal estimation procedures for Tucker low-rank tensor PCA and tensor regression, and proofs of technical results in Chapter 3. Without loss of generality, we assume that $r_j \asymp r_{\max} \asymp r$ for $j \in [3]$ throughout the proofs.

5.2.1 Optimal Estimation Procedure of Tucker Low-rank Tensor PCA and Tensor Regression

We collect the estimation procedures for Tucker low-rank tensor PCA and tensor regression in this section. Consider the Tucker low-rank tensor PCA: $\mathcal{A} = \mathcal{X} + \mathcal{Z}$, where $\mathcal{X} = (U_1, U_2, U_3)\mathcal{G}$. As proved by Zhang and Xia (2018), the following Algorithm 9 achieves the optimal rate in estimation error.

Algorithm 9 Higher Order Orthogonal Iteration (HOOI) (De Lathauwer et al., 2000b; Zhang and Golub, 2001; Richard and Montanari, 2014)

```
Input: $\mathcal{A}$, $r_1, r_2, r_3$, iteration $t_{\max}$;
Initialize $\hat{U}_1^{(0)} = \mathrm{SVD}_{r_1}(\mathcal{M}_1(\mathcal{A}))$, $\hat{U}_2^{(0)} = \mathrm{SVD}_{r_2}(\mathcal{M}_2(\mathcal{A}))$, $\hat{U}_3^{(0)} = \mathrm{SVD}_{r_3}(\mathcal{M}_3(\mathbf{A}))$, $t = 1$;
while $t \leqslant t_{\max}$ do
    $\hat{U}_1^{(t)}$ = leading $r_1$ left singular vectors of $\mathcal{M}_1(\mathcal{A}) \times_2 \hat{U}_2^{(t-1)\top} \times_3 \hat{U}_3^{(t-1)\top}$;
    $\hat{U}_2^{(t)}$ = leading $r_2$ left singular vectors of $\mathcal{M}_2(\mathcal{A}) \times_1 \hat{U}_1^{(t-1)\top} \times_3 \hat{U}_3^{(t-1)\top}$;
    $\hat{U}_3^{(t)}$ = leading $r_3$ left singular vectors of $\mathcal{M}_3(\mathcal{A}) \times_1 \hat{U}_1^{(t-1)\top} \times_2 \hat{U}_2^{(t-1)\top}$;
    $t = t + 1$;
end while
Output: $\hat{U}_1 = \hat{U}_1^{(t_{\max})}, \hat{U}_2 = \hat{U}_2^{(t_{\max})}, \hat{U}_3 = \hat{U}_3^{(t_{\max})}, \hat{\mathcal{G}}$.
```

Next, we introduce the simultaneous gradient descent in Algorithm 10 for Tucker low-rank regression. (Han et al., 2020, Theorem 4.2) proved that Algorithm 10 achieves the optimal rate of estimation error for $\mathcal{T}$.

Algorithm 10 Simultaneous Gradient Descent (Han et al., 2020)

```
    Input: $\ell_n(\cdot)$: the objective function (3.7) for tensor regression, $\{(\mathcal{X}_i, Y_i)\}_{i=1}^{n}$, $r_1, r_2, r_3$, tuning parameters $a, b > 0$, step size $\eta$;
$\tilde{U}_1, \tilde{U}_2, \tilde{U}_3, \tilde{\mathcal{G}} = \mathrm{HOOI}(\sum_{i=1}^{n} Y_i \mathcal{X}_i)$; Initialize $\hat{U}_j^{(0)} = b\tilde{U}_j$ for $j \in [3]$, $\hat{\mathcal{G}}^{(0)} = \tilde{\mathcal{G}}/b^3$;
for $t = 0, \dots, t_{\max} - 1$ do
    for $j = 1, 2, 3$ do
        $\hat{U}_j^{(t+1)} = U^{(t)} - \eta\big(\nabla_{U_j}\ell_n\big((\hat{U}_1^{(t)}, \hat{U}_2^{(t)}, \hat{U}_3^{(t)}) \cdot \hat{\mathcal{G}}^{(t)}\big) + a\hat{U}_j^{(t)}(\hat{U}_j^{(t)\top}\hat{U}_j^{(t)} - b^2 I_{r_j})\big)$;
    end for
    $\hat{\mathcal{G}}^{(t+1)} = \hat{\mathcal{G}}^{(t)} - \eta\nabla_{\mathcal{G}}\ell_n\big((\hat{U}_1^{(t)}, \hat{U}_2^{(t)}, \hat{U}_3^{(t)}) \cdot \hat{\mathcal{G}}^{(t)}\big)$;
    $t = t + 1$;
end for
$\hat{\mathcal{T}} = (\hat{U}_1^{(t_{\max})}, \hat{U}_2^{(t_{\max})}, \hat{U}_3^{(t_{\max})}) \cdot \hat{\mathcal{G}}^{(t_{\max})}$;
$\hat{U}_j = \mathrm{SVD}_{r_j}(\mathcal{M}_j(\hat{\mathcal{T}}))$ for $j \in [3]$;
$\hat{\mathcal{G}} = \hat{\mathcal{T}} \times_1 \hat{U}_1^\top \times_2 \hat{U}_2^\top \times_3 \hat{U}_3^\top$;
    Output: $\hat{U}_1, \hat{U}_2, \hat{U}_3, \hat{\mathcal{G}}$
```

5.2.2 Proof of Theorem 3.3.1

Note that $\mathcal{A}/\sigma = \mathcal{T}/\sigma + \mathcal{Z}/\sigma$. We can replace $\mathcal{A}, \mathcal{T}, \mathcal{Z}$ by $\mathcal{A}/\sigma, \mathcal{T}/\sigma, \mathcal{Z}/\sigma$ without essentially changing the problem. Thus, we assume that $\sigma = 1$ without loss of generality. To simplify the notations, we write $\mathcal{P}_U = UU^\top$ as the spectral projector for any orthonormal columns U, i.e., $U^\top U$ being an identity matrix. Then, write $\mathcal{P}_U^\perp = I - \mathcal{P}_U$. Denote $A_j = \mathcal{M}_j(\mathcal{A})$, $T_j = \mathcal{M}_j(\mathcal{T})$, $G_j = \mathcal{M}_j(\mathcal{G})$, and $Z_j = \mathcal{M}_j(\mathcal{Z})$ the corresponding matricizations for all $j = 1, 2, 3$.

Without loss of generality, we only consider $j = 1$ and prove the theorem for $\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2$. Notice that Theorem 3.3.1 automatically holds if $p \leqslant r^{1/3}$, we only need to consider the case $p \geqslant r^{1/3}$. By Algorithm 1, $\hat{U}_1 = \hat{U}_1^{(2)}$ contains the top-r_1 eigenvectors of $A_1(\mathcal{P}_{\hat{U}_2^{(1)}} \otimes \mathcal{P}_{\hat{U}_3^{(1)}})A_1^\top$. As a result, $\hat{U}_1^{(2)}\hat{U}_1^{(2)\top}$ is the spectral projector for the top-r_1 eigenvectors of

$$A_1(\mathcal{P}_{\hat{U}_2^{(1)}} \otimes \mathcal{P}_{\hat{U}_3^{(1)}})A_1^\top = T_1(\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3})T_1^\top + \mathfrak{J}_1 + \mathfrak{J}_2 + \mathfrak{J}_3 + \mathfrak{J}_4$$

where $\mathfrak{J}_1 = T_1(\mathcal{P}_{\hat{U}_2^{(1)}} \otimes \mathcal{P}_{\hat{U}_3^{(1)}})Z_1^\top$, $\mathfrak{J}_2 = Z_1(\mathcal{P}_{\hat{U}_2^{(1)}} \otimes \mathcal{P}_{\hat{U}_3^{(1)}})T_1^\top$, $\mathfrak{J}_3 = Z_1(\mathcal{P}_{\hat{U}_2^{(1)}} \otimes \mathcal{P}_{\hat{U}_3^{(1)}})Z_1^\top$, and $\mathfrak{J}_4 = T_1((\mathcal{P}_{\hat{U}_2^{(1)}} - \mathcal{P}_{U_2}) \otimes \mathcal{P}_{\hat{U}_3^{(1)}})T_1^\top + T_1(\mathcal{P}_{U_2} \otimes (\mathcal{P}_{\hat{U}_3^{(1)}} - \mathcal{P}_{U_3})T_1^\top$.

To this end, we write

$$\begin{aligned} A_1(\mathcal{P}_{\hat{U}_2^{(1)}} \otimes \mathcal{P}_{\hat{U}_3^{(1)}})A_1^\top =& U_1G_1G_1^\top U_1^\top + \mathfrak{J}_1 + \mathfrak{J}_2 + \mathfrak{J}_3 + \mathfrak{J}_4 \\ =:& U_1G_1G_1^\top U_1^\top + \mathfrak{E}_1. \end{aligned}$$

Lemma 5.2.1. *Under Assumption 3.3.1 and conditions of Theorem 3.3.1, there exist absolute constants* $c_1, C_1, C_2 > 0$ *so that with probability at least* $1 - C_1e^{-c_1p}$,

$$\|\mathfrak{J}_1\| = \|\mathfrak{J}_2\| \leqslant C_2\kappa_0\lambda_{\min}\sqrt{p}, \quad \|\mathfrak{J}_3\| \leqslant C_2p, \quad \|\mathfrak{J}_4\| \leqslant C_2\kappa_0^2p, \quad \|\mathfrak{E}_1\| \leqslant C_2\kappa_0\lambda_{\min}\sqrt{p}.$$

Moreover, by (Zhang and Xia, 2018, Theorem 1), the following bounds hold:

$$\max\{\|\hat{U}_k^{(0)}\hat{U}_k^{(0)\top} - U_kU_k^\top\|, \|\hat{U}_k^{(1)}\hat{U}_k^{(1)\top} - U_kU_k^\top\|, \|\hat{U}_k^{(2)}\hat{U}_k^{(2)\top} - U_kU_k^\top\|\} \leqslant C_2\sqrt{p}\lambda_{\min}^{-1} \tag{5.76}$$

for all $k \in [3]$. Denote $\mathcal{E}_0$ the event of Lemma 5.2.1 and (5.76) so that $\mathbb{P}(\mathcal{E}_0) \geqslant 1 - C_1e^{-c_1p}$. By definition, Λ_j^2 is a diagonal matrix containing the eigenvalues of $G_jG_j^\top$. Without loss of generality, we assume that $G_jG_j^\top = \Lambda_j^2$ is a diagonal matrix. Then immediately we have

$$\|\Lambda_j^{-1}G_j\| = 1, \forall j \in [3]. \tag{5.77}$$

Step 1: representation of spectral projector $\hat{U}_1\hat{U}_1^\top$. We write

$$\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 = 2r_1 - 2\langle\hat{U}_1\hat{U}_1^\top, U_1U_1^\top\rangle = -2\langle\hat{U}_1\hat{U}_1^\top - U_1U_1^\top, U_1U_1^\top\rangle.$$

Define, for a positive integer k, $\mathfrak{P}_j^{-k} = U_j\Lambda_j^{-2k}U_j^\top$. With a little abuse of notations, denote $\mathfrak{P}_j^0 := \mathfrak{P}_j^\perp := \mathcal{P}_{U_j}^\perp$. Note that, under the event $\mathcal{E}_0$ of Lemma 5.2.1

$$\|\mathfrak{E}_1\| \leqslant C_2\kappa_0\lambda_{\min}\sqrt{p} < \frac{\lambda_{\min}^2}{2}.$$

implying that the condition of (Xia, 2019b, Theorem 1) is satisfied.

Lemma 5.2.2. *(Xia, 2019b, Theorem 1) If $\|\mathfrak{E}_1\| \leqslant \frac{\lambda_{\min}^2}{2}$, the following equation holds*

$$\hat{U}_1\hat{U}_1^\top - U_1U_1^\top = \sum\nolimits_{k\geqslant 1} \mathcal{S}_{G_1,k}(\mathfrak{E}_1) \tag{5.78}$$

where for each positive integer k

$$\mathcal{S}_{G_1,k}(\mathfrak{E}_1) = \sum_{s_1+\cdots+s_{k+1}=k} (-1)^{1+\tau(\mathbf{s})} \cdot \mathfrak{P}_1^{-s_1}\mathfrak{E}_1\mathfrak{P}_1^{-s_2}\mathfrak{E}_1\mathfrak{P}_1^{-s_3}\cdots\mathfrak{P}_1^{-s_k}\mathfrak{E}_1\mathfrak{P}_1^{-s_{k+1}}$$

where $s_1, \cdots, s_{k+1}$ are non-negative integers and $\tau(\mathbf{s}) = \sum_{j=1}^{k+1}\mathbb{I}(s_j > 0)$.

By Lemma 5.2.1 and 5.2.2, eq.(5.78) holds under event $\mathcal{E}_0$ of Lemma 5.2.1. Since $\mathfrak{P}_j^0 U_jU_j^\top = U_jU_j^\top\mathfrak{P}_j^0 = 0$, we have

$$\langle \mathcal{S}_{G_1,1}(\mathfrak{E}_1), U_1U_1^\top \rangle = \langle \mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^{\perp} + \mathfrak{P}_1^{\perp}\mathfrak{E}_1\mathfrak{P}_1^{-1}, U_1U_1^\top \rangle = 0.$$

Similarly, $\langle \mathcal{S}_{G_1,2}(\mathfrak{E}_1), U_1U_1^\top \rangle = -\langle \mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^{\perp}\mathfrak{E}_1\mathfrak{P}_1^{-1}, U_1U_1^\top \rangle$ and

$\langle \mathcal{S}_{G_1,3}(\mathfrak{E}_1), U_1U_1^\top \rangle = -2\,\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^{\perp}\mathfrak{E}_1\mathfrak{P}_1^{\perp}\mathfrak{E}_1\mathfrak{P}_1^{-2}\right) + 2\,\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^{\perp}\mathfrak{E}_1\mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^{-1}\right).$

Note that

$$\begin{aligned}\|\mathcal{S}_{G_1,k}(\mathfrak{E}_1)\| \leqslant& \sum_{s_1+\cdots+s_{k+1}=k} \left\|(-1)^{1+\tau(\mathbf{s})} \cdot \mathfrak{P}_1^{-s_1}\mathfrak{E}_1\mathfrak{P}_1^{-s_2}\mathfrak{E}_1\mathfrak{P}_1^{-s_3}\cdots\mathfrak{P}_1^{-s_k}\mathfrak{E}_1\mathfrak{P}_1^{-s_{k+1}}\right\| \\ \leqslant& \binom{2k}{k}\frac{\|\mathfrak{E}_1\|^k}{\lambda_{\min}^{2k}} \leqslant \left(\frac{4\|\mathfrak{E}_1\|}{\lambda_{\min}^2}\right)^k,\end{aligned} \tag{5.79}$$

implying that

$$\left|\sum\nolimits_{k\geqslant 4}\langle \mathcal{S}_{G_1,k}(\mathfrak{E}_1), U_1U_1^\top \rangle\right| \leqslant r_1\sum\nolimits_{k\geqslant 4}\left(\frac{4\|\mathfrak{E}_1\|}{\lambda_{\min}^2}\right)^k \leqslant C_2r_1\frac{\kappa_0^4p^2}{\lambda_{\min}^4}$$

where the last inequality holds under event $\mathcal{E}_0$ by Lemma 5.2.1.

Therefore, under event $\mathcal{E}_0$, we write

$$\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 = -2\big\langle \mathcal{S}_{G_1,2}(\mathfrak{E}_1), U_1U_1^\top\big\rangle - 2\big\langle \mathcal{S}_{G_1,3}(\mathfrak{E}_1), U_1U_1^\top\big\rangle + O\Big(\frac{r_1\kappa_0^4p^2}{\lambda_{\min}^4}\Big).$$

Now, it suffices to investigate the first two terms on RHS of above equation.

Step 2: bounding $\big\langle \mathcal{S}_{G_1,3}(\mathfrak{E}_1), U_1U_1^\top\big\rangle$. Since $T_1^\top\mathfrak{P}_1^\perp = 0$ and $\mathfrak{P}_1^\perp T_1 = 0$,

$$\begin{aligned}
\big\langle \mathcal{S}_{G_1,3}(\mathfrak{E}_1), U_1U_1^\top\big\rangle =& -2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^\perp\mathfrak{E}_1\mathfrak{P}_1^\perp\mathfrak{E}_1\mathfrak{P}_1^{-2}\big) + 2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^\perp\mathfrak{E}_1\mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^{-1}\big)\\
=& -2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_3)\mathfrak{P}_1^\perp\mathfrak{J}_3\mathfrak{P}_1^\perp(\mathfrak{J}_2+\mathfrak{J}_3)\mathfrak{P}_1^{-2}\big)\\
&+ 2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_3)\mathfrak{P}_1^\perp(\mathfrak{J}_2+\mathfrak{J}_3)\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2+\mathfrak{J}_3+\mathfrak{J}_4)\mathfrak{P}_1^{-1}\big)\\
=& -2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_3)\mathfrak{P}_1^\perp\mathfrak{J}_3\mathfrak{P}_1^\perp(\mathfrak{J}_2+\mathfrak{J}_3)\mathfrak{P}_1^{-2}\big) \qquad (5.80)\\
&+ 2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2)\mathfrak{P}_1^{-1}\big) + 2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}(\mathfrak{J}_3+\mathfrak{J}_4)\mathfrak{P}_1^{-1}\big)\\
&+ 2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}\mathfrak{J}_3\mathfrak{P}_1^\perp(\mathfrak{J}_2+\mathfrak{J}_3)\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2+\mathfrak{J}_3+\mathfrak{J}_4)\mathfrak{P}_1^{-1}\big)\\
&+ 2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_3\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2+\mathfrak{J}_3+\mathfrak{J}_4)\mathfrak{P}_1^{-1}\big)
\end{aligned}$$

Define the term

$$\begin{aligned}
\mathfrak{M} =& \big\langle \mathcal{S}_{G_1,3}(\mathfrak{E}_1), U_1U_1^\top\big\rangle - 2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2)\mathfrak{P}_1^{-1}\big)\\
=& -2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_3)\mathfrak{P}_1^\perp\mathfrak{J}_3\mathfrak{P}_1^\perp(\mathfrak{J}_2+\mathfrak{J}_3)\mathfrak{P}_1^{-2}\big) + 2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}(\mathfrak{J}_3+\mathfrak{J}_4)\mathfrak{P}_1^{-1}\big)\\
&+ 2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}\mathfrak{J}_3\mathfrak{P}_1^\perp(\mathfrak{J}_2+\mathfrak{J}_3)\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2+\mathfrak{J}_3+\mathfrak{J}_4)\mathfrak{P}_1^{-1}\big)\\
&+ 2\,\mathrm{tr}\,\big(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_3\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2+\mathfrak{J}_3+\mathfrak{J}_4)\mathfrak{P}_1^{-1}\big).
\end{aligned}$$

By Lemma 5.2.1, under event $\mathcal{E}_0$,

$$\mathfrak{M} \leqslant C_2r_1\frac{\kappa_0^2p\cdot(\kappa_0\sqrt{p}\lambda_{\min})(\kappa_0\sqrt{p}\lambda_{\min})}{\lambda_{\min}^6} \leqslant C_2r_1\frac{\kappa_0^4p^2}{\lambda_{\min}^4}. \qquad (5.81)$$

Therefore, we conclude on event $\mathcal{E}_0$ that

$$\left|\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2\,\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^\perp\mathfrak{E}_1\mathfrak{P}_1^{-1}\right) + 4\,\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2)\mathfrak{P}_1^{-1}\right)\right| \leqslant C_2\frac{r_1\kappa_0^4p^2}{\lambda_{\min}^4}. \tag{5.82}$$

We begin with considering $\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2)\mathfrak{P}_1^{-1}\right)$. Clearly,

$$\begin{aligned} &\left|\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2)\mathfrak{P}_1^{-1}\right)\right| \\ \leqslant &\left|\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{-1}\right)\right| + \left|\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}\mathfrak{J}_2\mathfrak{P}_1^{-1}\right)\right|. \end{aligned} \tag{5.83}$$

It suffices to bound $\left|\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{-1}\right)\right|$ and $\left|\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^\perp\mathfrak{J}_2\mathfrak{P}_1^{-1}\mathfrak{J}_2\mathfrak{P}_1^{-1}\right)\right|$, respectively. By the proof of Lemma 5.2.1, on event $\mathcal{E}_0$, there exist two (random) matrices $R_2 \in \mathbb{O}_{r_2}$ and $R_3 \in \mathbb{O}_{r_3}$ such that $\|\hat{U}_2^{(1)} - U_2R_2\|, \|\hat{U}_3^{(1)} - U_3R_3\| \leqslant C_2\sqrt{p}/\lambda_{\min}$. Therefore, on event $\mathcal{E}_0$, (5.214),

$$\begin{aligned} &\left\|\mathfrak{J}_1 - T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^\top\right\| = \left\|T_1(\mathcal{P}_{\hat{U}_2^{(1)}}\otimes\mathcal{P}_{\hat{U}_3^{(1)}})Z_1^\top - T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^\top\right\| \\ \leqslant &\left\|\left[T_1(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right]\left[Z_1(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right]^\top - \left[T_1((U_2R_2)\otimes(U_3R_3))\right]\left[Z_1((U_2R_2)\otimes(U_3R_3))\right]^\top\right\| \\ \leqslant &\left\|Z_1\left[(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)}) - ((U_2R_2)\otimes(U_3R_3))\right]\right\|\left\|T_1(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right\| \\ &+ \|Z_1((U_2R_2)\otimes(U_3R_3))\|\left\|T_1\left[(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)}) - ((U_2R_2)\otimes(U_3R_3))\right]\right\| \\ \leqslant &\kappa_0\lambda_{\min}\left\|Z_1\left[(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)}) - ((U_2R_2)\otimes(U_3R_3))\right]\right\| \\ &+ C_0\sqrt{p}\left\|T_1\left[(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)}) - ((U_2R_2)\otimes(U_3R_3))\right]\right\| \\ \leqslant &\kappa_0\lambda_{\min}\left(\left\|Z_1\left((\hat{U}_2^{(1)} - U_2R_2)\otimes\hat{U}_3^{(1)}\right)\right\| + \left\|Z_1\left((U_2R_2)\otimes(\hat{U}_3^{(1)} - U_3R_3)\right)\right\|\right) \\ &+ C_0\sqrt{p}\Big(\left\|T_1\big((\hat{U}_2^{(1)} - U_2R_2)\otimes\hat{U}_3^{(1)}\big)\right\| + \left\|T_1\big((U_2R_2)\otimes(\hat{U}_3^{(1)} - U_3R_3)\big)\right\|\Big) \\ \overset{\text{proof of lemma 5.2.1, 5.2.3}}{\leqslant} &C_2\kappa_0\lambda_{\min}\cdot\sqrt{pr}\left(\left\|\hat{U}_2^{(1)} - U_2R_2\right\| + \left\|\hat{U}_3^{(1)} - U_3R_3\right\|\right) + C_3\kappa_0p^{3/2}/\lambda_{\min} \\ \leqslant &C_2\kappa_0p\sqrt{r}. \end{aligned} \tag{5.84}$$

By Lemma 5.2.1 and (5.84), on event $\mathcal{E}_0$,

$$\begin{aligned}
&\left|\operatorname{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{\perp}\mathfrak{J}_2\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{-1}\right)\right| \\
\leqslant &\left|\operatorname{tr}\left(\mathfrak{P}_1^{-1}T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top}\mathfrak{P}_1^{\perp}Z_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})T_1^{\top}\mathfrak{P}_1^{-1}T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top}\mathfrak{P}_1^{-1}\right)\right| \\
&+\left|\operatorname{tr}\left(\mathfrak{P}_1^{-1}(\mathfrak{J}_1-T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top})\mathfrak{P}_1^{\perp}Z_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})T_1^{\top}\mathfrak{P}_1^{-1}T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top}\mathfrak{P}_1^{-1}\right)\right| \\
&+\left|\operatorname{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{\perp}(\mathfrak{J}_1-T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top})^{\top}\mathfrak{P}_1^{-1}T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top}\mathfrak{P}_1^{-1}\right)\right| \\
&+\left|\operatorname{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{\perp}\mathfrak{J}_2\mathfrak{P}_1^{-1}(\mathfrak{J}_1-T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top})\mathfrak{P}_1^{-1}\right)\right| \\
\leqslant &\left|\operatorname{tr}\left(\mathfrak{P}_1^{-1}T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top}\mathfrak{P}_1^{\perp}Z_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})T_1^{\top}\mathfrak{P}_1^{-1}T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top}\mathfrak{P}_1^{-1}\right)\right| \\
&+C_2r_1\frac{\kappa_0 p\sqrt{r}\cdot\kappa_0\sqrt{p}\lambda_{\min}\cdot\kappa_0\sqrt{p}\lambda_{\min}}{\lambda_{\min}^6} \\
= &\left|\operatorname{tr}\left(U_1\Lambda_1^{-2}U_1^{\top}T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top}U_{1\perp}U_{1\perp}^{\top}Z_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})T_1^{\top}U_1\Lambda_1^{-2}U_1^{\top}T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^{\top}U_1\Lambda_1^{-2}U_1^{\top}\right)\right| \\
&+C_2\kappa_0^3r^{3/2}p^2\lambda_{\min}^{-4} \\
= &\left|\operatorname{tr}\left(U_1\Lambda_1^{-2}G_1W_2^{\top}W_2G_1^{\top}\Lambda_1^{-2}G_1W_1^{\top}\Lambda_1^{-2}U_1^{\top}\right)\right|+C_2\kappa_0^3r^{3/2}p^2\lambda_{\min}^{-4}
\end{aligned}\tag{5.85}$$

where we denote

$$W_1=U_1^{\top}Z_1(U_2\otimes U_3)\in\mathbb{R}^{r_1\times(r_2r_3)},\quad W_2=U_{1\perp}^{\top}Z_1(U_2\otimes U_3)\in\mathbb{R}^{(p_1-r_1)\times(r_2r_3)}.\tag{5.86}$$

Due to the property of Gaussian matrices, $[W_1\ W_2]=[U_1^{\top}\ U_{1\perp}^{\top}]Z_1(U_2\otimes U_3)\overset{\text{i.i.d.}}{\sim}N(0,1)$. Therefore, $W_1\overset{\text{i.i.d.}}{\sim}N(0,1)$, $W_{1\perp}\overset{\text{i.i.d.}}{\sim}N(0,1)$, and W_1, W_2 are independent. Conditioning on W_2, we have

$$\begin{aligned}
&\operatorname{tr}\left(U_1\Lambda_1^{-2}G_1W_2^{\top}W_2G_1^{\top}\Lambda_1^{-2}G_1W_1^{\top}\Lambda_1^{-2}U_1^{\top}\right)\Big|W_2=\operatorname{tr}\left(\Lambda_1^{-4}G_1W_2^{\top}W_2G_1^{\top}\Lambda_1^{-2}G_1W_1^{\top}\right)\Big|W_2 \\
\sim &N(0,\|\Lambda_1^{-4}G_1W_2^{\top}W_2G_1^{\top}\Lambda_1^{-2}G_1\|_F^2)\Big|W_2.
\end{aligned}$$

By the Gaussian concentration inequality, we get

$$\begin{aligned}
&\mathbb{P}\left(\left|\operatorname{tr}\left(U_1\Lambda_1^{-2}G_1W_2^{\top}W_2G_1^{\top}\Lambda_1^{-2}G_1W_1^{\top}\Lambda_1^{-2}U_1^{\top}\right)\right|\leqslant C_2\sqrt{\log(p)}\|\Lambda_1^{-4}G_1W_2^{\top}W_2G_1^{\top}\Lambda_1^{-2}G_1\|_F\Big|W_2\right) \\
&\geqslant 1-p^{-3}
\end{aligned}$$

for some absolute constant $C_2 > 0$. Denote the above event $\mathcal{E}_1$ so that $\mathbb{P}(\mathcal{E}_1) \geqslant 1-p^{-3}$.

In addition, by (5.77), on event $\mathcal{E}_0$,

$$\|\Lambda_1^{-4} G_1 W_2^\top W_2 G_1^\top \Lambda_1^{-2} G_1\|_{\mathrm{F}} \leqslant C_2\sqrt{r_1}\frac{\|W_2\|^2}{\lambda_{\min}^3} \leqslant C_2\sqrt{r_1}\frac{p}{\lambda_{\min}^3}.$$

By the previous two inequalities, on event $\mathcal{E}_0 \cap \mathcal{E}_1$,

$$\Bigg|\mathrm{tr}\left(U_1\Lambda_1^{-2}G_1 W_2^\top W_2 G_1^\top \Lambda_1^{-2} G_1 W_1^\top \Lambda_1^{-2} U_1^\top\right)\Bigg| W_2 \Bigg| \leqslant C_2\sqrt{r_1}\frac{p\sqrt{\log(p)}}{\lambda_{\min}^3}.$$

By combining eq. (5.85) and the above inequality, we conclude on event $\mathcal{E}_0 \cap \mathcal{E}_1$ that

$$\left|\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{\perp}\mathfrak{J}_2\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{-1}\right)\right| \leqslant C_2\left(r^{1/2}p\sqrt{\log(p)}\lambda_{\min}^{-3} + r^{3/2}\kappa_0^3 p^2\lambda_{\min}^{-4}\right).$$

Similarly, on event $\mathcal{E}_0 \cap \mathcal{E}_1$,

$$\left|\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{\perp}\mathfrak{J}_2\mathfrak{P}_1^{-1}\mathfrak{J}_2\mathfrak{P}_1^{-1}\right)\right| \leqslant C_2\left(r^{1/2}p\sqrt{\log(p)}\lambda_{\min}^{-3} + r^{3/2}\kappa_0^3 p^2\lambda_{\min}^{-4}\right).$$

Combining e.q. (5.83) and the above two inequalities, with probability at least $1 - C_1 p^{-3}$,

$$\left|\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{\perp}\mathfrak{J}_2\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_2)\mathfrak{P}_1^{-1}\right)\right| \leqslant C_2\left(r^{1/2}p\sqrt{\log(p)}\lambda_{\min}^{-3} + r^{3/2}\kappa_0^3 p^2\lambda_{\min}^{-4}\right). \tag{5.87}$$

Step 3: bounding smaller terms of $\langle \mathcal{S}_{G_1,2}(\mathfrak{E}_1), U_1U_1^\top\rangle$. Recall that $-\langle \mathcal{S}_{G_1,2}(\mathfrak{E}_1), U_1U_1^\top\rangle = \mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^{\perp}\mathfrak{E}_1\mathfrak{P}_1^{-1}\right)$. Since $T_1^\top\mathfrak{P}_1^{\perp} = 0$ and $\mathfrak{P}_1^{\perp}T_1 = 0$, we write

$$\begin{aligned}
&\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^{\perp}\mathfrak{E}_1\mathfrak{P}_1^{-1}\right) = \mathrm{tr}\left(\mathfrak{P}_1^{-1}(\mathfrak{J}_1+\mathfrak{J}_3)\mathfrak{P}_1^{\perp}(\mathfrak{J}_2+\mathfrak{J}_3)\mathfrak{P}_1^{-1}\right)\\
=&\,\mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{\perp}\mathfrak{J}_2\mathfrak{P}_1^{-1}\right) + \mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{\perp}\mathfrak{J}_3\mathfrak{P}_1^{-1}\right) + \mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_3\mathfrak{P}_1^{\perp}\mathfrak{J}_2\mathfrak{P}_1^{-1}\right) + \mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_3\mathfrak{P}_1^{\perp}\mathfrak{J}_3\mathfrak{P}_1^{-1}\right)\\
=&:\mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV}.
\end{aligned} \tag{5.88}$$

By Lemma 5.2.1, on event $\mathcal{E}_0$,

$$|\mathrm{IV}| \leqslant r_1 \|\mathfrak{P}_1^{-1}\|^2 \|\mathfrak{J}_3\|^2 \leqslant C_2 r_1 \frac{p^2}{\lambda_{\min}^4}. \tag{5.89}$$

Next, we show that $\mathrm{II} = \mathrm{III} \leqslant C_2 \left(r^{1/2} p \sqrt{\log(p)} \lambda_{\min}^{-3} + r^{3/2} \kappa_0 p^2 \lambda_{\min}^{-4} \right)$ with probability at least $1 - C_1 p^{-3}$. Similarly to (5.84), with probability at least $1 - C_1 e^{-c_1 p}$,

$$\begin{aligned}
&\left\|\mathfrak{J}_3 - Z_1(\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3}) Z_1^\top\right\| = \left\| Z_1(\mathcal{P}_{\hat{U}_2^{(1)}} \otimes \mathcal{P}_{\hat{U}_3^{(1)}}) Z_1^\top - Z_1(\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3}) Z_1^\top \right\| \\
\leqslant & \left\| \left[Z_1(\hat{U}_2^{(1)} \otimes \hat{U}_3^{(1)}) \right] \left[Z_1(\hat{U}_2^{(1)} \otimes \hat{U}_3^{(1)}) \right]^\top - [Z_1((U_2R_2) \otimes (U_3R_3))] \, [Z_1((U_2R_2) \otimes (U_3R_3))]^\top \right\| \\
\leqslant & \left\| Z_1 \left[(\hat{U}_2^{(1)} \otimes \hat{U}_3^{(1)}) - ((U_2R_2) \otimes (U_3R_3)) \right] \right\| \left\| Z_1(\hat{U}_2^{(1)} \otimes \hat{U}_3^{(1)}) \right\| \\
& + \|Z_1((U_2R_2) \otimes (U_3R_3))\| \left\| Z_1 \left[(\hat{U}_2^{(1)} \otimes \hat{U}_3^{(1)}) - ((U_2R_2) \otimes (U_3R_3)) \right] \right\| \\
= & \left\| Z_1 \left[(\hat{U}_2^{(1)} \otimes \hat{U}_3^{(1)}) - ((U_2R_2) \otimes (U_3R_3)) \right] \right\| (C_2\sqrt{p} + \|Z_1(U_2 \otimes U_3)\|) \\
\leqslant & C_2\sqrt{p} \left(\left\| Z_1 \left((\hat{U}_2^{(1)} - U_2R_2) \otimes \hat{U}_3^{(1)} \right) \right\| + \left\| Z_1 \left((U_2R_2) \otimes (\hat{U}_3^{(1)} - U_3R_3) \right) \right\| \right) \\
\overset{\text{proof of Lemma 5.2.1}}{\leqslant} & C_2\sqrt{p} \cdot \sqrt{pr} \left\| \hat{U}_2^{(1)} - U_2R_2 \right\| + C_2\sqrt{p} \cdot \sqrt{pr} \left\| \hat{U}_3^{(1)} - U_3R_3 \right\| \\
\leqslant & C_2 \frac{p^{3/2} r^{1/2}}{\lambda_{\min}}.
\end{aligned} \tag{5.90}$$

Combining (5.90) and (5.84) together, with probability at least $1 - C_1 e^{-c_1 p}$,

$$\begin{aligned}
& |\mathrm{II}| = |\mathrm{III}| \\
= & \left| \mathrm{tr} \left(U_1 \Lambda_1^{-2} U_1^\top \mathfrak{J}_1 U_{1\perp} U_{1\perp}^\top \mathfrak{J}_3 U_1 \Lambda_1^{-2} U_1^\top \right) \right| \\
\leqslant & \left| \mathrm{tr} \left(U_1 \Lambda_1^{-2} U_1^\top T_1 (\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3}) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3}) Z_1^\top U_1 \Lambda_1^{-2} U_1^\top \right) \right| \\
& + \left| \mathrm{tr} \left(U_1 \Lambda_1^{-2} U_1^\top (\mathfrak{J}_1 - T_1 (\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3}) Z_1^\top) U_{1\perp} U_{1\perp}^\top Z_1 (\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3}) Z_1^\top U_1 \Lambda_1^{-2} U_1^\top \right) \right| \\
& + \left| \mathrm{tr} \left(U_1 \Lambda_1^{-2} U_1^\top \mathfrak{J}_1 U_{1\perp} U_{1\perp}^\top (\mathfrak{J}_3 - Z_1 (\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3}) Z_1^\top) U_1 \Lambda_1^{-2} U_1^\top \right) \right| \\
\leqslant & \left| \mathrm{tr} \left(U_1 \Lambda_1^{-2} U_1^\top T_1 (\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3}) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3}) Z_1^\top U_1 \Lambda_1^{-2} U_1^\top \right) \right| \qquad (5.91) \\
& + r \frac{\|T_1 (\mathcal{P}_{\hat{U}_2} \otimes \mathcal{P}_{\hat{U}_3}) Z_1^\top\| \|\mathfrak{J}_3 - Z_1 (\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3}) Z_1^\top\|}{\lambda_{\min}^4}
\end{aligned}$$

$$
\begin{aligned}
&+ r\frac{\|\mathfrak{J}_1 - T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^\top\|\|Z_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^\top\|}{\lambda_{\min}^4}\\
&\leqslant \left|\mathrm{tr}\left(U_1\Lambda_1^{-2}U_1^\top T_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(\mathcal{P}_{U_2}\otimes\mathcal{P}_{U_3})Z_1^\top U_1\Lambda_1^{-2}U_1^\top\right)\right| + r^{3/2}\kappa_0\frac{p^2}{\lambda_{\min}^4}\\
&= \left|\mathrm{tr}\left(U_1\Lambda_1^{-2}G_1W_2^\top W_2W_1^\top\Lambda_1^{-2}U_1^\top\right)\right| + r^{3/2}\kappa_0\frac{p^2}{\lambda_{\min}^4}\\
&= \left|\mathrm{tr}\left(\Lambda_1^{-4}G_1W_2^\top W_2W_1^\top\right)\right| + r^{3/2}\kappa_0\frac{p^2}{\lambda_{\min}^4} \qquad (5.92)
\end{aligned}
$$

where W_1 and W_2 are defined in (5.86).
Observe that

$$
\mathrm{tr}\left(\Lambda_1^{-4}G_1W_2^\top W_2W_1^\top\right)\Big| W_2 = \langle\Lambda_1^{-4}G_1W_2^\top W_2, W_1\rangle|W_2 \sim N\left(0, \left\|\Lambda_1^{-4}G_1W_2^\top W_2\right\|_F^2\right)
$$

By the Gaussian concentration inequality, we have

$$
\mathbb{P}\left(\mathrm{tr}\left(\Lambda_1^{-4}G_1W_2^\top W_2W_1^\top\right) \geqslant C_2\sqrt{\log(p)}\left\|\Lambda_1^{-4}G_1W_2^\top W_2\right\|_F \Big| W_2\right) \leqslant p^{-3}. \qquad (5.93)
$$

Moreover, by (5.77), with probability at least $1 - C_1e^{-c_1p}$,

$$
\begin{aligned}
\left\|\Lambda_1^{-4}G_1W_2^\top W_2\right\|_F \leqslant & \sqrt{r_1}\left\|\Lambda_1^{-4}G_1W_2^\top W_2\right\| \leqslant C_2\sqrt{r_1}\frac{\|W_2\|^2}{\lambda_{\min}^3}\\
\leqslant & C_2\sqrt{r_1}\frac{\|Z_1(U_2\otimes U_3)\|^2}{\lambda_{\min}^3} \leqslant C_2\sqrt{r_1}\frac{p}{\lambda_{\min}^3}.
\end{aligned}
$$

By (5.93) and the above inequality, we get with probability at least $1 - C_1p^{-3}$,

$$
\mathrm{tr}\left(\Lambda_1^{-4}G_1W_2^\top W_2W_1^\top\right) \leqslant C_2\sqrt{r_1}\frac{p\sqrt{\log(p)}}{\lambda_{\min}^3}. \qquad (5.94)
$$

Recall that $\lambda_{\min} \gg p^{3/4}$, eq. (5.92) and (5.94) together imply that

$$
|\mathrm{II}| = |\mathrm{III}| \leqslant C_2\left(r^{1/2}p\sqrt{\log(p)}\lambda_{\min}^{-3} + r^{3/2}\kappa_0p^2\lambda_{\min}^{-4}\right) \qquad (5.95)
$$

with probability $1 - C_1 p^{-3}$.

Step 4: treating the leading term of $\langle \mathcal{S}_{G_1,2}(\mathfrak{E}_1), U_1U_1^\top\rangle$. Now, we consider the leading term $I = \operatorname{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{\perp}\mathfrak{J}_2\mathfrak{P}_1^{-1}\right)$. By definition and Algorithm 1, $\hat{U}_2^{(1)}\hat{U}_2^{(1)\top}$ is the spectral projector for the top-r_2 eigenvectors of

$$\begin{aligned} A_2(\mathcal{P}_{\hat{U}_1^{(0)}} \otimes \mathcal{P}_{\hat{U}_3^{(0)}})A_2^\top =& U_2G_2G_2^\top U_2^\top - U_2G_2(U_1^\top \mathcal{P}^{\perp}_{\hat{U}_1^{(0)}}U_1 \otimes U_3^\top \mathcal{P}_{\hat{U}_3^{(0)}}U_3)G_2^\top U_2^\top \\ &- U_2G_2(I_{r_1} \otimes U_3^\top \mathcal{P}^{\perp}_{\hat{U}_3^{(0)}}U_3)G_2^\top U_2^\top + T_2(\mathcal{P}_{\hat{U}_1^{(0)}} \otimes \mathcal{P}_{\hat{U}_3^{(0)}})Z_2^\top \\ &+ Z_2(\mathcal{P}_{\hat{U}_1^{(0)}} \otimes \mathcal{P}_{\hat{U}_3^{(0)}})T_2^\top + Z_2(\mathcal{P}_{\hat{U}_1^{(0)}} \otimes \mathcal{P}_{\hat{U}_3^{(0)}})Z_2^\top \\ =&: U_2G_2G_2^\top U_2^\top + \hat{\mathfrak{E}}_2. \end{aligned}$$

Similarly, we can define $\hat{\mathfrak{E}}_3$. Let $\hat{\Lambda}_2^2$ and $\hat{\Lambda}_3^2$ be the diagonal matrices containing the eigenvalues of $G_2(U_1\mathcal{P}_{\hat{U}_1^{(0)}}U_1^\top \otimes U_3\mathcal{P}_{\hat{U}_3^{(0)}}U_3^\top)G_2^\top$ and $G_3(U_1\mathcal{P}_{\hat{U}_1^{(0)}}U_1^\top \otimes U_2\mathcal{P}_{\hat{U}_2^{(0)}}U_2^\top)G_3^\top$ with decreasing order, respectively. Let $\hat{\lambda}_{\min}$ be the smallest eigenvalue among all eigenvalues of $\hat{\Lambda}_2^2$ and $\hat{\Lambda}_3^2$.

Recall that $\mathfrak{P}_j^k = U_j\Lambda_j^{-2k}U_j^\top$ for positive integer k, and $\mathfrak{P}_j^0 := \mathfrak{P}_j^\perp := \mathcal{P}_{U_j}^\perp$ for $j = 2, 3$. By Lemma 5.2.1 and Lemma 5.2.2, with probability at least $1 - C_1e^{-c_1p}$,

$$\|\hat{\mathfrak{E}}_j\| \leqslant C_2\kappa_0\sqrt{p}\lambda_{\min}$$

and

$$\hat{U}_j^{(1)}\hat{U}_j^{(1)\top} - U_jU_j^\top = \sum_{k \geqslant 1}\mathcal{S}_{G_j,k}(\hat{\mathfrak{E}}_j).$$

where for positive integer k,

$$\mathcal{S}_{G_j,k}(\hat{\mathfrak{E}}_j) = \sum_{s_1+\cdots+s_{k+1}=k}(-1)^{1+\tau(\mathbf{s})} \cdot \mathfrak{P}_j^{-s_1}\hat{\mathfrak{E}}_j\mathfrak{P}_j^{-s_2}\hat{\mathfrak{E}}_j\mathfrak{P}_j^{-s_3}\cdots\mathfrak{P}_j^{-s_k}\hat{\mathfrak{E}}_j\mathfrak{P}_j^{-s_{k+1}}.$$

For $k \geqslant 2$, similarly to (5.79), we have

$$\left\|\mathcal{S}_{G_j,k}(\hat{\mathfrak{E}}_j)\right\| \leqslant \left(\frac{4\|\hat{\mathfrak{E}}_j\|}{\lambda_{\min}^2}\right)^k.$$

Then with probability at least $1 - C_1e^{-c_1p}$,

$$\Big\|\sum_{k\geqslant 2}\mathcal{S}_{G_j,k}(\hat{\mathfrak{E}}_j)\Big\| \leqslant \sum_{k\geqslant 2}\left(\frac{4\|\hat{\mathfrak{E}}_j\|}{\lambda_{\min}^2}\right)^k \leqslant C_2\frac{\kappa_0^2p}{\lambda_{\min}^2}.$$

Note that

$$\mathcal{P}_{U_j}\mathcal{S}_{G_j,1}(\hat{\mathfrak{E}}_j) = \mathcal{P}_{U_j}\left(\mathfrak{P}_j^{-1}\hat{\mathfrak{E}}_j\mathfrak{P}_j^{\perp} + \mathfrak{P}_j^{\perp}\hat{\mathfrak{E}}_j\mathfrak{P}_j^{-1}\right) = U_j\Lambda_j^{-2}U_j^\top\hat{\mathfrak{E}}_j\mathcal{P}_{U_j}^{\perp}$$

Therefore, with probability at least $1 - C_1e^{-c_1p}$, for $j = 2, 3$,

$$\left\|U_j^\top\mathcal{P}_{\hat{U}_j^{(1)}} - U_j^\top - \Lambda_j^{-2}U_j^\top\hat{\mathfrak{E}}_j\mathcal{P}_{U_j}^{\perp}\right\| = \left\|\mathcal{P}_{U_j}\mathcal{P}_{\hat{U}_j^{(1)}} - \mathcal{P}_{U_j} - U_j\Lambda_j^{-2}U_j^\top\hat{\mathfrak{E}}_j\mathcal{P}_{U_j}^{\perp}\right\| \leqslant C_2\frac{\kappa_0^2p}{\lambda_{\min}^2}. \tag{5.96}$$

For $I = \mathrm{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{J}_1\mathfrak{P}_1^{\perp}\mathfrak{J}_2\mathfrak{P}_1^{-1}\right)$, with probability at least $1 - C_1e^{-c_1p}$,

$$\begin{aligned}
&\left|I - \mathrm{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\right| \\
=&\left|\mathrm{tr}\left(U_1\Lambda_1^{-2}U_1^\top\mathfrak{J}_1U_{1\perp}U_{1\perp}^\top\mathfrak{J}_1^\top U_1\Lambda_1^{-2}U_1^\top\right) - \mathrm{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\right| \\
=&\Big|\ \mathrm{tr}\left(\Lambda_1^{-4}G_1((U_2^\top\mathcal{P}_{\hat{U}_2^{(1)}})\otimes(U_3^\top\mathcal{P}_{\hat{U}_3^{(1)}}))Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1((\mathcal{P}_{\hat{U}_2^{(1)}}U_2)\otimes(\mathcal{P}_{\hat{U}_3^{(1)}}U_3))G_1^\top\right) \\
&\quad - \mathrm{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\Big|.
\end{aligned} \tag{5.97}$$

By (5.77), (5.96) and (5.212), with probability at least $1 - C_1e^{-c_1p}$ that

$$\begin{aligned}
&\Big|\ \mathrm{tr}\left(\Lambda_1^{-4}G_1((U_2^\top\mathcal{P}_{\hat{U}_2^{(1)}})\otimes(U_3^\top\mathcal{P}_{\hat{U}_3^{(1)}}))Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1((\mathcal{P}_{\hat{U}_2^{(1)}}U_2)\otimes(\mathcal{P}_{\hat{U}_3^{(1)}}U_3))G_1^\top\right) \\
&\quad - \mathrm{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\Big| \\
\leqslant& 2\left|\mathrm{tr}\left(\Lambda_1^{-4}G_1((\Lambda_2^{-2}U_2^\top\hat{\mathfrak{E}}_2\mathcal{P}_{U_2}^{\perp})\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\right| \\
&\quad + 2\left|\mathrm{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes(\Lambda_3^{-2}U_3^\top\hat{\mathfrak{E}}_3\mathcal{P}_{U_3}^{\perp}))Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\right|
\end{aligned}$$

$$
\begin{aligned}
&+ C_2 r_1 \max_{j=2,3} \left(\|\Lambda_j^{-2} U_j^\top \hat{\mathfrak{E}}_j \mathcal{P}_{U_j}^\perp \| \right)^2 \cdot (\sqrt{pr})^2 \cdot \lambda_{\min}^{-2} + C_2 r_1 \frac{\kappa_0^2 p}{\lambda_{\min}^2} \cdot \sqrt{pr} \cdot \sqrt{p} \cdot \lambda_{\min}^{-2} \\
&+ C_2 r_1 \frac{\kappa_0^2 p}{\lambda_{\min}^2} \cdot \max_{j=2,3} \|\Lambda_j^{-2} U_j^\top \hat{\mathfrak{E}}_j \mathcal{P}_{U_j}^\perp \| \cdot (\sqrt{pr})^2 \cdot \lambda_{\min}^{-2} + C_2 r_1 \left(\frac{\kappa_0^2 p}{\lambda_{\min}^2} \right)^2 \cdot (\sqrt{pr})^2 \cdot \lambda_{\min}^{-2} \\
\leqslant & 2 \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} U_2^\top \hat{\mathfrak{E}}_2 \mathcal{P}_{U_2}^\perp) \otimes U_3^\top) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (U_2 \otimes U_3) G_1^\top \right) \right| \\
&+ 2 \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 (U_2^\top \otimes (\Lambda_3^{-2} U_3^\top \hat{\mathfrak{E}}_3 \mathcal{P}_{U_3}^\perp)) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (U_2 \otimes U_3) G_1^\top \right) \right| + C_2 r^2 \kappa_0^2 p^2 \lambda_{\min}^{-4}.
\end{aligned}
\tag{5.98}
$$

By the definition of $\hat{\mathfrak{E}}_2$ and recall that $T_2^\top \mathcal{P}_{U_2}^\perp = 0$, eq. (5.77) and (5.212) and Lemma 5.2.1 imply that with probability at least $1 - C_1 e^{-c_p}$,

$$
\begin{aligned}
& \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} U_2^\top \hat{\mathfrak{E}}_2 \mathcal{P}_{U_2}^\perp) \otimes U_3^\top) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (U_2 \otimes U_3) G_1^\top \right) \right| \\
\leqslant & \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} U_2^\top T_2 (\mathcal{P}_{\hat{U}_1^{(0)}} \otimes \mathcal{P}_{\hat{U}_3^{(0)}}) Z_2^\top \mathcal{P}_{U_2}^\perp) \otimes U_3^\top) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (U_2 \otimes U_3) G_1^\top \right) \right| \\
&+ \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} U_2^\top Z_2 (\mathcal{P}_{\hat{U}_1^{(0)}} \otimes \mathcal{P}_{\hat{U}_3^{(0)}}) Z_2^\top \mathcal{P}_{U_2}^\perp) \otimes U_3^\top) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (U_2 \otimes U_3) G_1^\top \right) \right| \\
\leqslant & \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} U_2^\top T_2 (\mathcal{P}_{\hat{U}_1^{(0)}} \otimes \mathcal{P}_{\hat{U}_3^{(0)}}) Z_2^\top \mathcal{P}_{U_2}^\perp) \otimes U_3^\top) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (U_2 \otimes U_3) G_1^\top \right) \right| \\
&+ C_2 r_1 \lambda_{\min}^{-2} \cdot \sqrt{pr} p \lambda_{\min}^{-2} \cdot \sqrt{p} \\
\leqslant & \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} U_2^\top T_2 (\mathcal{P}_{U_1} \otimes \mathcal{P}_{U_3}) Z_2^\top \mathcal{P}_{U_2}^\perp) \otimes U_3^\top) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (U_2 \otimes U_3) G_1^\top \right) \right| \\
&+ \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} U_2^\top T_2 ((\mathcal{P}_{\hat{U}_1^{(0)}} - \mathcal{P}_{U_1}) \otimes \mathcal{P}_{U_3}) Z_2^\top \mathcal{P}_{U_2}^\perp) \otimes U_3^\top) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (U_2 \otimes U_3) G_1^\top \right) \right| \\
&+ \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} U_2^\top T_2 (\mathcal{P}_{\hat{U}_1^{(0)}} \otimes (\mathcal{P}_{\hat{U}_3^{(0)}} - \mathcal{P}_{U_3})) Z_2^\top \mathcal{P}_{U_2}^\perp) \otimes U_3^\top) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (U_2 \otimes U_3) G_1^\top \right) \right| \\
&+ C_2 r^{3/2} p^2 \lambda_{\min}^{-4} \\
\leqslant & \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} G_2 (U_1^\top \otimes U_3^\top) Z_2^\top \mathcal{P}_{U_2}^\perp) \otimes U_3^\top) Z_1^\top U_{1\perp} U_{1\perp}^\top Z_1 (U_2 \otimes U_3) G_1^\top \right) \right| \\
&+ C_2 r_1 \lambda_{\min}^{-2} \cdot \lambda_{\min}^{-1} \cdot \sqrt{pr} \frac{\sqrt{p}}{\lambda_{\min}} \cdot \sqrt{pr} \cdot \sqrt{p} + C_2 r^{3/2} p^2 \lambda_{\min}^{-4} \\
\leqslant & \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} G_2 W_3^\top \mathcal{P}_{U_2}^\perp) \otimes U_3^\top) W_4^\top W_4 (U_2 \otimes U_3) G_1^\top \right) \right| + C_2 r^2 p^2 \lambda_{\min}^{-4} \\
= & \left| \mathrm{tr} \left(\Lambda_1^{-4} G_1 ((\Lambda_2^{-2} G_2 W_5^\top) \otimes I_{r_3}) W_6^\top W_7 G_1^\top \right) \right| + C_2 r^2 p^2 \lambda_{\min}^{-4}.
\end{aligned}
\tag{5.99}
$$

where $W_3 = Z_2(U_1 \otimes U_3) \in \mathbb{R}^{p_2 \times (r_1 r_3)}$, $W_4 = U_{1\perp}^\top Z_1 \in \mathbb{R}^{(p_1 - r_1) \times (p_2 p_3)}$, $W_5 = U_{2\perp}^\top W_3 \in \mathbb{R}^{(p_2 - r_2) \times (r_1 r_3)}$, $W_6 = W_4(U_{2\perp} \otimes U_3) \in \mathbb{R}^{(p_1 - r_1) \times ((p_2 - r_2) r_3)}$, $W_7 = W_4(U_2 \otimes$

$U_3) \in \mathbb{R}^{(p_1-r_1)\times(r_2r_3)}$.

By definition, $W_3 \overset{\text{i.i.d.}}{\sim} N(0,1)$, $W_4 \overset{\text{i.i.d.}}{\sim} N(0,1)$, $W_5 \overset{\text{i.i.d.}}{\sim} N(0,1)$, and W_3 and W_4 are independent. Furthermore, since $W_4([U_2\ U_{2\perp}] \otimes U_3) \overset{\text{i.i.d.}}{\sim} N(0,1)$ and W_6, W_7 are two disjoint submatrices of $W_4([U_2\ U_{2\perp}] \otimes U_3)$. Therefore, $W_6 \overset{\text{iid}}{\sim} N(0,1)$, $W_7 \overset{\text{i.i.d.}}{\sim} N(0,1)$, W_5, W_6, and W_7 are jointly independent. Then,

$$\begin{aligned}
&\operatorname{tr}\left(\Lambda_1^{-4}G_1((\Lambda_2^{-2}G_2W_5^\top)\otimes I_{r_3})W_6^\top W_7 G_1^\top\right)\bigg| W_5, W_7 \\
=&\operatorname{tr}\left(W_7G_1^\top\Lambda_1^{-4}G_1((\Lambda_2^{-2}G_2W_5^\top)\otimes I_{r_3})W_6^\top\right)\bigg| W_5, W_7 \\
\sim& N(0, \|W_7G_1^\top\Lambda_1^{-4}G_1((\Lambda_2^{-2}G_2W_5^\top)\otimes I_{r_3})\|_F^2).
\end{aligned}$$

By the Gaussian concentration inequality, we have

$$\begin{aligned}
\mathbb{P}\Big(&\left|\operatorname{tr}\left(\Lambda_1^{-4}G_1((\Lambda_2^{-2}G_2W_5^\top)\otimes I_{r_3})W_6^\top W_7G_1^\top\right)\right| \\
&> C_2\sqrt{\log(p)}\|W_7G_1^\top\Lambda_1^{-4}G_1((\Lambda_2^{-2}G_2W_5^\top)\otimes I_{r_3})\|_F\bigg| W_5, W_7\Big) \leqslant p^{-3}.
\end{aligned}$$

In addition, with probability at least $1 - C_1e^{-c_1p}$, we have $\|W_5\|, \|W_7\| \leqslant C_2\sqrt{p_1}$ since $r = O(\sqrt{p})$. By (5.77), we obtain

$$\begin{aligned}
\|W_7G_1^\top\Lambda_1^{-4}G_1((\Lambda_2^{-2}G_2W_5^\top)\otimes I_{r_3})\|_F \leqslant&\sqrt{r_1}\|W_7G_1^\top\Lambda_1^{-4}G_1((\Lambda_2^{-2}G_2W_5^\top)\otimes I_{r_3})\| \\
\leqslant& C_2\sqrt{r_1}\sqrt{p}\lambda_{\min}^{-2}\cdot\lambda_{\min}^{-1}\sqrt{p} \leqslant C_2\sqrt{r_1}\frac{p}{\lambda_{\min}^3}.
\end{aligned}$$

Therefore, with probability at least $1 - C_1p^{-3}$,

$$\left|\operatorname{tr}\left(\Lambda_1^{-4}G_1((\Lambda_2^{-2}G_2W_5^\top)\otimes I_{r_3})W_6^\top W_7G_1^\top\right)\right| \leqslant C_2\sqrt{r_1}\frac{p\sqrt{\log(p)}}{\lambda_{\min}^3}.$$

Combining (5.99) and the above inequality, we get with probability at least $1-C_1p^{-3}$

that

$$\left|\operatorname{tr}\left(\Lambda_1^{-4}G_1((\Lambda_2^{-2}U_2^\top\hat{\mathfrak{E}}_2\mathcal{P}_{U_2}^\perp)\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\right| \leqslant C_2(\sqrt{r_1}p\sqrt{\log(p)}\lambda_{\min}^{-3}+r^2p^2\lambda_{\min}^{-4}).$$

Similarly, with probability at least $1-C_1p^{-3}$,

$$\left|\operatorname{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes(\Lambda_3^{-2}U_3^\top\hat{\mathfrak{E}}_3\mathcal{P}_{U_3}^\perp))Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\right| \leqslant C_2(\sqrt{r_1}p\sqrt{\log(p)}\lambda_{\min}^{-3}+r^2p^2\lambda_{\min}^{-4}).$$

By (5.97), (5.98) and the above two inequalities, we get with probability at least $1-C_1p^{-3}$ that

$$\left|I-\operatorname{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\right| \leqslant C_2(\sqrt{r_1}p\sqrt{\log(p)}\lambda_{\min}^{-3}+r^2\kappa_0^2p^2\lambda_{\min}^{-4}). \tag{5.100}$$

Combining (5.88), (5.89), (5.95) and the above inequality, we get with probability at least $1-C_1p^{-3}$ that

$$\begin{aligned}&\left|\operatorname{tr}\left(\mathfrak{P}_1^{-1}\mathfrak{E}_1\mathfrak{P}_1^\perp\mathfrak{E}_1\mathfrak{P}_1^{-1}\right)-\operatorname{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\right|\\ &\qquad\leqslant C_2(\sqrt{r_1}p\sqrt{\log(p)}\lambda_{\min}^{-3}+r^2\kappa_0^2p^2\lambda_{\min}^{-4}).\end{aligned}$$

By (5.82), (5.87) and the above inequality, with probability at least $1-C_1p^{-3}$,

$$\begin{aligned}&\left|\|\hat{U}_1\hat{U}_1^\top-U_1U_1^\top\|_F^2-2\operatorname{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\right|\\ &\qquad\leqslant C_2(\sqrt{r_1}p\sqrt{\log(p)}\lambda_{\min}^{-3}+r^2\kappa_0^4p^2\lambda_{\min}^{-4}).\end{aligned} \tag{5.101}$$

Final step: characterizing the distribution. By eq. (5.101), it suffices to prove the distribution of $\operatorname{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)$. We write

$$\begin{aligned}&\operatorname{tr}\left(\Lambda_1^{-4}G_1(U_2^\top\otimes U_3^\top)Z_1^\top U_{1\perp}U_{1\perp}^\top Z_1(U_2\otimes U_3)G_1^\top\right)\\ =&\|\Lambda_1^{-2}G_1(U_2^\top\otimes U_3^\top)Z_1^\top U_{1\perp}\|_F^2=\sum_{j=r_1+1}^{p_1}\|\Lambda_1^{-2}G_1(U_2^\top\otimes U_3^\top)Z_1^\top u_j\|_2^2,\end{aligned} \tag{5.102}$$

where $\{u_j\}_{j=r_1+1}^{p_1}$ are the columns of $U_{1\perp}$. For any $r_1 + 1 \leqslant j \leqslant p_1$, $Z_1^\top u_j \in N(0, I_{p_2 \times p_3})$, and

$$\mathbb{E}(Z_1^\top u_{j_1})(Z_1^\top u_{j_2})^\top = 0, \quad \forall r_1 + 1 \leqslant j_1 \neq j_2 \leqslant p_1.$$

Therefore, $\{Z_1^\top u_j\}_{j=r_1+1}^{p_1}$ are standard Gaussian random vectors. Recall that $G_1 G_1^\top = \Lambda_1^2$ and

$$\begin{aligned}\Lambda_1^{-2} G_1 (U_2^\top \otimes U_3^\top) Z_1^\top u_j \sim & N(0, \Lambda_1^{-2} G_1 (U_2^\top \otimes U_3^\top)[\Lambda_1^{-2} G_1 (U_2^\top \otimes U_3^\top)]^\top) \\ = & N(0, \Lambda_1^{-2} G_1 G_1^\top \Lambda_1) = N(0, \Lambda_1^{-2}).\end{aligned}$$

Therefore,

$$\|\Lambda_1^{-2} G_1 (U_2^\top \otimes U_3^\top) Z_1^\top U_{1\perp}\|_F^2 \overset{d.}{=} \sum_{i=1}^{p_1 - r_1} \|\Lambda_1^{-1} z_i\|_2^2,$$

where $z_i \overset{i.i.d.}{\sim} N(0, I_r)$. The RHS in above equation is a sum of independent random variables.

Clearly, $\mathbb{E}\|\Lambda_1^{-1} z_i\|_2^2 = \|\Lambda_1^{-1}\|_F^2$, $\operatorname{Var}\left(\|\Lambda_1^{-1} z_i\|_2^2\right) = 2\|\Lambda_1^{-2}\|_F^2$, and

$$\mathbb{E}\|\Lambda_1^{-1} z_i\|_2^6 \leqslant C_3 \sum_{j_1, j_2, j_3 = 1}^{r_1} \frac{1}{\lambda_{j_1}^{(1)2} \lambda_{j_2}^{(1)2} \lambda_{j_3}^{(1)2}} = C_3 \|\Lambda_1^{-1}\|_F^6$$

where we denote $\Lambda_1 = \operatorname{diag}\big(\lambda_1^{(1)}, \lambda_2^{(1)}, \cdots, \lambda_{r_1}^{(1)}\big)$. By Berry-Esseen theorem (Berry, 1941; Esseen, 1942), we get

$$\begin{aligned}&\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\|\Lambda_1^{-2} G_1 (U_2^\top \otimes U_3^\top) Z_1^\top U_{1\perp}\|_F^2 - 2(p_1 - r_1) \left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{8(p_1 - r_1)} \left\|\Lambda_1^{-2}\right\|_F} \leqslant x \right) - \Phi(x) \right| \\ \leqslant & C_3 \left(\frac{\|\Lambda_1^{-1}\|_F^4}{\|\Lambda_1^{-2}\|_F^2} \right)^{3/2} \cdot \frac{1}{\sqrt{p_1 - r_1}}.\end{aligned}$$

Note that $\sqrt{8(p_1 - r_1) \left\|\Lambda_1^{-2}\right\|_F} \geqslant \sqrt{2 p_1 r_1} \kappa_0^{-2} \lambda_{\min}^{-2}$. By eq. (5.101) and Lipschitz property of $\Phi(\cdot)$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2(p_1-r_1)\left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{8(p_1-r_1)}\left\|\Lambda_1^{-2}\right\|_F} \leqslant x\right)\\
\leqslant&\mathbb{P}\left(\frac{\|\Lambda_1^{-2}G_1(U_2^\top \otimes U_3^\top)Z_1^\top U_{1\perp}\|_F^2 - 2(p_1-r_1)\left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{8(p_1-r_1)}\left\|\Lambda_1^{-2}\right\|_F} \leqslant x + C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^2\sqrt{p\log p}}{\lambda_{\min}}\Big)\right)\\
&+ C_1p^{-3}\\
\leqslant&\Phi\left(x + C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^2\sqrt{p\log p}}{\lambda_{\min}}\Big)\right) + C_1p^{-3} + C_3\frac{r_1^{3/2}}{\sqrt{p_1-r_1}}\\
\leqslant&\Phi(x) + C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^2\sqrt{p\log p}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}}.
\end{aligned} \tag{5.103}$$

Similarly, we can show that

$$\begin{aligned}
\mathbb{P}\Big(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2(p_1-r_1)\left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{8(p_1-r_1)}\left\|\Lambda_1^{-2}\right\|_F} \leqslant x\Big)&\\
\geqslant \Phi(x) - C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^2\sqrt{p\log p}}{\lambda_{\min}}\Big) - C_3\frac{r^{3/2}}{\sqrt{p}}&.
\end{aligned}$$

Combining two inequalities above, we know that

$$\begin{aligned}
\sup_{x\in\mathbb{R}}\left|\mathbb{P}\Big(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2(p_1-r_1)\left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{8(p_1-r_1)}\left\|\Lambda_1^{-2}\right\|_F} \leqslant x\Big) - \Phi(x)\right|&\\
\leqslant C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^2\sqrt{p\log p}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}}&.
\end{aligned}$$

Moreover, by the previous inequality and the Lipschitz property of $\Phi(\cdot)$ and $|x|e^{-x^2/2} < 1$ for all $x \in \mathbb{R}$, for any $x \in \mathbb{R}$,

$$\left|\mathbb{P}\Big(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2p_1\left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{8p_1}\left\|\Lambda_1^{-2}\right\|_F} \leqslant x\Big) - \Phi(x)\right|$$

$$\begin{aligned}
&=\left|\mathbb{P}\Big(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2(p_1-r_1)\left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{8(p_1-r_1)}\left\|\Lambda_1^{-2}\right\|_F} \leqslant \sqrt{\frac{p_1}{p_1-r_1}}x + \frac{r_1\left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{2(p_1-r_1)}\left\|\Lambda_1^{-2}\right\|_F}\Big) - \Phi\left(x\right)\right| \\
&\leqslant C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^2\sqrt{p\log p}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}} + \left|\Phi\left(\sqrt{\frac{p_1}{p_1-r_1}}x + \frac{r_1\left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{2(p_1-r_1)}\left\|\Lambda_1^{-2}\right\|_F}\right) - \Phi(x)\right| \\
&\leqslant C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^2\sqrt{p\log p}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}} + \left|\Phi\left(\sqrt{\frac{p_1}{p_1-r_1}}x\right) - \Phi(x)\right| + C_3\frac{r_1\left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{2(p_1-r_1)}\left\|\Lambda_1^{-2}\right\|_F} \\
&\leqslant C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^2\sqrt{p\log p}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}} + \left(\sqrt{\frac{p_1}{p_1-r_1}} - 1\right)|x|e^{-x^2/2} \\
&\leqslant C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^2\sqrt{p\log p}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}}.
\end{aligned}$$

Therefore, we conclude the proof of Theorem 3.3.1.

5.2.3 Proof of Theorem 3.3.2

First, we show that

$$\begin{aligned}
&\mathbb{P}\left(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2p_1\sigma^2\left\|\hat{\Lambda}_1^{-1}\right\|_F^2}{\sqrt{8p_1}\sigma^2\left\|\hat{\Lambda}_1^{-2}\right\|_F} \leqslant x\right) \\
&\leqslant\Phi(x) + C_2\left(\frac{r^{3/2}\kappa_0^6p^{3/2}}{(\lambda_{\min}/\sigma)^2} + \frac{\kappa_0^3\sqrt{pr(r^2+\log(p))}}{\lambda_{\min}/\sigma}\right) + C_3\frac{r^{3/2}}{\sqrt{p}}. \qquad (5.104)
\end{aligned}$$

Without loss of generality, we assume $\sigma = 1$ and only prove the result for $\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2$. We denote $\tilde{U}_1 \in \mathbb{O}_{p_1,r_1}$ the top-r_1 left singular vectors of $\mathcal{M}_1(\mathcal{A}\times_2\hat{U}_2^\top\times_3\hat{U}_3^\top)$. By Theorem 3.3.1 and Zhang and Xia (2018), it is easy to show that $\|\tilde{U}_1\tilde{U}_1^\top - U_1U_1^\top\| \leqslant C_2\sqrt{p}\lambda_{\min}^{-1}$ with probability at least $1 - C_1p^{-3}$. By definition, we know that $\hat{\Lambda}_1^2 = \mathrm{diag}(\hat{\lambda}_1^2, \cdots, \hat{\lambda}_{r_1}^2)$ contains the eigenvalues of $\tilde{U}_1^\top A_1(\mathcal{P}_{\hat{U}_2}\otimes\mathcal{P}_{\hat{U}_3})A_1^\top\tilde{U}_1$. We denote $\Lambda_1 = \mathrm{diag}(\lambda_1, \cdots, \lambda_{r_1})$. Then,

$$\begin{aligned}
&\sup_{1\leqslant k\leqslant r_1}\left|\lambda_k^2 - \hat{\lambda}_k^2\right| \\
&\leqslant \inf_{R\in\mathbb{O}_{r_1}}\left\|\tilde{U}_1^\top A_1(\mathcal{P}_{\hat{U}_2}\otimes\mathcal{P}_{\hat{U}_3})A_1^\top\tilde{U}_1 - RG_1G_1^\top R\right\|
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \inf_{R\in\mathbb{O}_{r_1}} \left\|\tilde{U}_1^\top T_1(\mathcal{P}_{\hat{U}_2}\otimes\mathcal{P}_{\hat{U}_3})T_1^\top\tilde{U}_1 - R\Lambda_1^2 R\right\| \\
&\quad + 2\left\|\tilde{U}_1^\top T_1(\mathcal{P}_{\hat{U}_2}\otimes\mathcal{P}_{\hat{U}_3})Z_1^\top\tilde{U}_1\right\| + \left\|\tilde{U}_1^\top Z_1(\mathcal{P}_{\hat{U}_2}\otimes\mathcal{P}_{\hat{U}_3})Z_1^\top\tilde{U}_1\right\| \\
&\leqslant \left\|\tilde{U}_1^\top U_1 G_1((U_2^\top\mathcal{P}_{\hat{U}_2}U_2)\otimes(U_3^\top\mathcal{P}_{\hat{U}_3}U_3))G_1^\top U_1^\top\tilde{U}_1 - \tilde{U}_1^\top U_1 G_1 G_1^\top U_1^\top \tilde{U}_1\right\| \\
&\quad + \inf_{R\in\mathbb{O}_{r_1}}\left\|\tilde{U}_1^\top U_1\Lambda_1^2 U_1^\top\tilde{U}_1 - R\Lambda_1^2 R^\top\right\| + 2\kappa_0\lambda_{\min}\left\|\tilde{U}_1^\top Z_1(\hat{U}_2\otimes\hat{U}_3)\right\| + \left\|\tilde{U}_1^\top Z_1(\hat{U}_2\otimes\hat{U}_3)\right\|^2.
\end{aligned}
$$

By (5.214) and the Gaussian concentration inequality, with probability at least $1-C_1p^{-3}$,

$$
\left\|\tilde{U}_1^\top Z_1(\hat{U}_2\otimes\hat{U}_3)\right\|^2 \leqslant C_2 p,
$$

and

$$
\begin{aligned}
\kappa_0\lambda_{\min}\left\|\tilde{U}_1^\top Z_1(\hat{U}_2\otimes\hat{U}_3)\right\| &\leqslant \kappa_0\lambda_{\min}\left\|\tilde{U}_1^\top(\mathcal{P}_{U_1}+\mathcal{P}_{U_1}^\perp)Z_1(\hat{U}_2\otimes\hat{U}_3)\right\| \\
&\leqslant \kappa_0\lambda_{\min}\Big(\left\|U_1^\top Z_1(\hat{U}_2\otimes\hat{U}_3)\right\| + \left\|\tilde{U}_1^\top U_{1\perp}\right\|\left\|Z_1(\hat{U}_2\otimes\hat{U}_3)\right\|\Big) \\
&\leqslant \kappa_0\lambda_{\min}\left(\left\|U_1^\top Z_1(\hat{U}_2\otimes\hat{U}_3)\right\| + \left\|\tilde{U}_1^\top U_{1\perp}\right\|\left\|Z_1(\hat{U}_2\otimes\hat{U}_3)\right\|\right) \\
&\leqslant \kappa_0\lambda_{\min}\left(\left\|U_1^\top Z_1(U_2\otimes U_3)\right\| + C_2\sqrt{pr}\frac{\sqrt{p}}{\lambda_{\min}} + C_2\frac{\sqrt{p}}{\lambda_{\min}}\sqrt{p}\right) \\
&\leqslant C_2\kappa_0\lambda_{\min}\left(\sqrt{r^2+\log(p)} + \sqrt{pr}\frac{\sqrt{p}}{\lambda_{\min}} + \frac{\sqrt{p}}{\lambda_{\min}}\sqrt{p}\right) \\
&\leqslant C_2\kappa_0(\sqrt{r^2+\log(p)}\lambda_{\min} + p\sqrt{r}).
\end{aligned}
$$

Moreover, with probability at least $1-C_1e^{-c_1p}$,

$$
\begin{aligned}
&\left\|\tilde{U}_1^\top U_1 G_1((U_2^\top\mathcal{P}_{\hat{U}_2}U_2)\otimes(U_3^\top\mathcal{P}_{\hat{U}_3}U_3))G_1^\top U_1^\top\tilde{U}_1 - \tilde{U}_1^\top U_1 G_1 G_1^\top U_1^\top\tilde{U}_1\right\| \\
&\qquad\leqslant \left\|G_1((U_2^\top\mathcal{P}_{\hat{U}_2}U_2)\otimes(U_3^\top\mathcal{P}_{\hat{U}_3}U_3))G_1^\top - G_1G_1^\top\right\| \\
&\qquad\leqslant \left\|G_1((U_2^\top\mathcal{P}_{\hat{U}_2}^\perp U_2)\otimes(U_3^\top\mathcal{P}_{\hat{U}_3}U_3))G_1^\top\right\| + \left\|G_1(I_{r_2}\otimes(U_3^\top\mathcal{P}_{\hat{U}_3}^\perp U_3))G_1^\top\right\| \\
&\qquad\leqslant \kappa_0^2\lambda_{\min}^2\Big(\left\|U_2^\top\mathcal{P}_{\hat{U}_2}^\perp U_2\right\| + \left\|U_3^\top\mathcal{P}_{\hat{U}_3}^\top U_3\right\|\Big) \leqslant \kappa_0^2\lambda_{\min}^2\Big(\left\|U_2^\top\hat{U}_{2\perp}\right\|^2 + \left\|U_3^\top\hat{U}_{3\perp}\right\|^2\Big) \\
&\qquad\leqslant C_2\kappa_0^2 p.
\end{aligned}
$$

To deal with $\inf_{R\in\mathbb{O}_{r_1}}\left\|\tilde{U}_1^\top U_1\Lambda_1^2U_1^\top\tilde{U}_1 - R\Lambda_1^2R^\top\right\|$, we need the following lemma.

Lemma 5.2.3. *For* $U, \hat{U} \in \mathbb{O}_{p,r}$,

$$\inf_{R\in\mathbb{O}_r} \left\|\hat{U}^\top U - R\right\| \leqslant \|U_\perp^\top \hat{U}\|^2.$$

$$\inf_{R\in\mathbb{O}_r} \left\|\hat{U}^\top U - R\right\|_F \leqslant \min\left\{\|U_\perp^\top \hat{U}\|_F^2, \sqrt{r}\|U_\perp^\top \hat{U}\|^2\right\}.$$

By Lemma 5.2.3 and (5.213), with probability at least $1 - C_1 e^{-cp}$,

$$\begin{aligned}
&\inf_{R\in\mathbb{O}_{r_1}} \left\|\tilde{U}_1^\top U_1 \Lambda_1^2 U_1^\top \tilde{U}_1 - R\Lambda_1^2 R^\top\right\| \\
&\qquad\leqslant \inf_{R\in\mathbb{O}_{r_1}} \left\{\left\|(\tilde{U}_1^\top U_1 - R)\Lambda_1^2 U_1^\top \tilde{U}_1\right\| + \left\|R\Lambda_1^2(\tilde{U}_1^\top U_1 - R)^\top\right\|\right\} \\
&\qquad\leqslant 2\inf_{R\in\mathbb{O}_{r_1}} \left\|\tilde{U}_1^\top U_1 - R\right\|\|\Lambda_1^2\| \leqslant C_2(\sqrt{p}\lambda_{\min}^{-1})^2 \cdot \kappa_0^2\lambda_{\min}^2 \\
&\qquad\leqslant C_2\kappa_0^2 p.
\end{aligned}$$

Combining together the inequalities above, we get with probability at least $1-C_1p^{-3}$,

$$\sup_{1\leqslant k\leqslant r_1} \left|\lambda_k^2 - \hat{\lambda}_k^2\right| \leqslant C_2\left(\kappa_0^2\sqrt{r}p + \kappa_0 r\sqrt{\log(p)}\lambda_{\min}\right). \tag{5.105}$$

Therefore, with probability at least $1 - C_1p^{-3}$,

$$\left|\left\|\Lambda_1^{-1}\right\|_F^2 - \left\|\hat{\Lambda}_1^{-1}\right\|_F^2\right| \leqslant r_1 \sup_{1\leqslant k\leqslant r_1} \frac{\left|\lambda_k^2 - \hat{\lambda}_k^2\right|}{\lambda_k^2\hat{\lambda}_k^2} \leqslant C_2\left(\kappa_0^2 r^{3/2} p\lambda_{\min}^{-4} + \kappa_0 r\sqrt{r^2 + \log(p)}\lambda_{\min}^{-3}\right)$$

and as a result

$$\begin{aligned}
\frac{\left|\left\|\Lambda_1^{-2}\right\|_F - \left\|\hat{\Lambda}_1^{-2}\right\|_F\right|}{\left\|\Lambda_1^{-2}\right\|_F} &\leqslant \frac{\left\|\Lambda_1^{-2} - \hat{\Lambda}_1^{-2}\right\|_F}{\left\|\Lambda_1^{-2}\right\|_F} \leqslant \frac{\sup_{1\leqslant k\leqslant r_1} \frac{|\lambda_k^2 - \hat{\lambda}_k^2|}{\lambda_k^2\hat{\lambda}_k^2}}{\kappa_0^{-2}\lambda_{\min}^{-2}} \\
&\leqslant C_2\left(\kappa_0^4 r^{1/2} p\lambda_{\min}^{-2} + \kappa_0^3\sqrt{r^2 + \log(p)}\lambda_{\min}^{-1}\right).
\end{aligned}$$

Note that $\sqrt{8(p_1-r_1)\left\|\Lambda_1^{-2}\right\|_F} \geqslant \sqrt{2p_1r_1}\kappa_0^{-2}\lambda_{\min}^{-2}$. By eq. (5.103), we have

$$
\begin{aligned}
&\mathbb{P}\left(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2p_1\left\|\hat{\Lambda}_1^{-1}\right\|_F^2}{\sqrt{8p_1}\left\|\Lambda_1^{-2}\right\|_F} \leqslant x\right)\\
\leqslant&\mathbb{P}\left(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2p_1\left\|\Lambda_1^{-1}\right\|_F^2}{\sqrt{8p_1}\left\|\Lambda_1^{-2}\right\|_F} \leqslant x + C_2\left(\kappa_0^4 r\frac{p^{3/2}}{\lambda_{\min}^2} + \kappa_0^3\frac{\sqrt{pr(r^2+\log(p))}}{\lambda_{\min}}\right)\right) + C_1p^{-3}\\
\leqslant&\Phi(x) + C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^3\sqrt{pr(r^2+\log(p))}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}}.
\end{aligned}
$$

Furthermore, we have

$$
\begin{aligned}
&\mathbb{P}\left(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2p_1\left\|\hat{\Lambda}_1^{-1}\right\|_F^2}{\sqrt{8p_1}\left\|\hat{\Lambda}_1^{-2}\right\|_F} \leqslant x\right)\\
=&\mathbb{P}\left(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2p_1\left\|\hat{\Lambda}_1^{-1}\right\|_F^2}{\sqrt{8p_1}\left\|\Lambda_1^{-2}\right\|_F} \leqslant x\left(1 + \frac{\left\|\hat{\Lambda}_1^{-2}\right\|_F - \left\|\Lambda_1^{-2}\right\|_F}{\left\|\Lambda_1^{-2}\right\|_F}\right)\right)\\
\leqslant&\mathbb{P}\left(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2p_1\left\|\hat{\Lambda}_1^{-1}\right\|_F^2}{\sqrt{8p_1}\left\|\Lambda_1^{-2}\right\|_F} \leqslant x\Big(1 + C_2\Big(\frac{\kappa_0^4r^{1/2}p}{\lambda_{\min}^2} + \frac{\kappa_0^3\sqrt{r^2+\log(p)}}{\lambda_{\min}}\Big)\operatorname{sgn}(x)\Big)\right) + \frac{C_1}{p^3}\\
\leqslant&\Phi\left(x\Big(1 + C_2\Big(\frac{\kappa_0^4r^{1/2}p}{\lambda_{\min}^2} + \frac{\kappa_0^3\sqrt{r^2+\log(p)}}{\lambda_{\min}}\Big)\operatorname{sgn}(x)\Big)\right) + \frac{C_1}{p^3}\\
&+ C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^3\sqrt{pr(r^2+\log(p))}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}}\\
\leqslant&\Phi(x) + C_2\left(\Big(\frac{\kappa_0^4r^{1/2}p}{\lambda_{\min}^2} + \frac{\kappa_0^3\sqrt{r^2+\log(p)}}{\lambda_{\min}}\Big)|x|\cdot e^{-x^2/2}\right)\\
&+ C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^3\sqrt{pr(r^2+\log(p))}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}}\\
\leqslant&\Phi(x) + C_2\Big(\frac{\kappa_0^4r^{1/2}p}{\lambda_{\min}^2} + \frac{\kappa_0^3\sqrt{r^2+\log(p)}}{\lambda_{\min}}\Big) + C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^3\sqrt{pr(r^2+\log(p))}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}}\\
\leqslant&\Phi(x) + C_2\Big(\frac{r^{3/2}\kappa_0^6p^{3/2}}{\lambda_{\min}^2} + \frac{\kappa_0^3\sqrt{pr(r^2+\log(p))}}{\lambda_{\min}}\Big) + C_3\frac{r^{3/2}}{\sqrt{p}}, \qquad (5.106)
\end{aligned}
$$

which has proved (5.104).

Then, by Lemma 3.3.1 and the similar argument for proving (5.104), we further have

$$\begin{aligned}&\mathbb{P}\left(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2p_1\hat{\sigma}^2\|\hat{\Lambda}_1^{-1}\|_F^2}{\sqrt{8p_1}\hat{\sigma}^2\|\hat{\Lambda}_1^{-2}\|_F} \leqslant x\right)\\ \leqslant&\Phi(x) + C_2\left(\frac{r^{3/2}\kappa_0^6p^{3/2}}{(\lambda_{\min}/\sigma)^2} + \frac{\kappa_0^3\sqrt{pr(r^2+\log p)}}{\lambda_{\min}/\sigma} + \frac{\sqrt{\log(p)}}{p^{1/4}} + \frac{\kappa_0\sqrt{r}}{\sqrt{p}}\right) + C_3\frac{r^{3/2}}{\sqrt{p}}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\mathbb{P}\left(\frac{\|\hat{U}_1\hat{U}_1^\top - U_1U_1^\top\|_F^2 - 2p_1\hat{\sigma}^2\|\hat{\Lambda}_1^{-1}\|_F^2}{\sqrt{8p_1}\hat{\sigma}^2\|\hat{\Lambda}_1^{-2}\|_F} \leqslant x\right)\\ \geqslant&\Phi(x) - C_2\left(\frac{r^{3/2}\kappa_0^6p^{3/2}}{(\lambda_{\min}/\sigma)^2} + \frac{\kappa_0^3\sqrt{pr(r^2+\log p)}}{\lambda_{\min}/\sigma} + \frac{\sqrt{\log(p)}}{p^{1/4}} + \frac{\kappa_0\sqrt{r}}{\sqrt{p}}\right) - C_3\frac{r^{3/2}}{\sqrt{p}}.\end{aligned}$$

Therefore, we conclude the proof of Theorem 3.3.2.

5.2.4 Proof of Theorem 3.3.3

Note that

$$Y_i/\sigma = \langle(\mathcal{T}/\sigma), \mathcal{X}_i\rangle + (\xi_i/\sigma).$$

We can replace $Y_i, \mathcal{T}, \xi_i$ by Y_i/σ, $\mathcal{T}/\sigma$, and ξ_i/σ without changing this problem essentially. Therefore, we assume that $\sigma = 1$ without loss of generality. We only need to focus on the non-trivial case $p \geqslant r^{1/3}$. Since the proof is technical challenging and long, we divide the proof into several steps. Consider the SVD decomposition $\hat{U}_j^{(t)\top}U_j = L_j^{(t)}S_j^{(t)}D_j^{(t)\top}$ for $t = 0, 1$ and $j = 1, 2, 3$, where $L_j^{(t)}, D_j^{(t)} \in \mathbb{O}_{r_j}$, and $S_j^{(t)}$ is the diagonal matrix with all singular values of $\hat{U}_j^{(t)\top}U_j$ in decreasing order. Denote $R_j^{(t)} = L_j^{(t)}D_j^{(t)\top} \in \mathbb{O}_{r_j}$.

For $t = 0, 1$, denote

$$\Delta\mathcal{T}_1^{(t+0.5)} = \hat{\mathcal{G}}^{(t)} \times_1 \hat{U}_1^{(t+0.5)} \times_2 \hat{U}_2^{(t)} \times_3 \hat{U}_3^{(t)} - \mathcal{T} \tag{5.107}$$

and denote $\hat{G}_j^{(t)} = \mathcal{M}_j(\hat{\mathcal{G}}^{(t)})$.

Step 0: preliminary bounds on $\hat{\mathcal{G}}^{(t)}$ and $\hat{U}_j^{(t+0.5)}$. Before dealing with $\|\hat{U}_1^{(2)}\hat{U}_1^{(2)\top} - U_1U_1^\top\|_F^2$, we first prove some preliminary results on $\hat{\mathcal{G}}^{(t)}$ and $\hat{U}_j^{(t+0.5)}$ which shall be used later.

Step 0.1: the error of $\hat{\mathcal{G}}^{(t)}$. Without loss of generality, we only prove the bound for $t = 0$ and we write $\hat{\mathcal{G}} = \hat{\mathcal{G}}^{(0)}$ for brevity. Consider the SVD decomposition $\hat{U}_i^{(0)\top}U_i = L_iS_iD_i^\top$, where $L_i, D_i \in \mathbb{O}_{r_i}$, and S_i is the diagonal matrix with all singular values of $\hat{U}_i^{(0)\top}U_i$ in decreasing order. Note that we omitted the superscripts of L_i, D_i, S_i for brevity. Let $R_i = L_iD_i^\top \in \mathbb{O}_{r_i}$ and

$$\mathfrak{J}_1 = \frac{1}{n}R_1\left[U_1^\top\Big(\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)(U_2 \otimes U_3)\right](R_2^\top \otimes R_3^\top).$$

We aim to show that with probability at least $1 - C_1e^{-c_1p} - p^{-3}$,

$$\left\|\hat{G}_1 - R_1G_1(R_2^\top \otimes R_3^\top) - \mathfrak{J}_1\right\| \leqslant C\left(\kappa_0\frac{pr}{n} + \kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right) \tag{5.108}$$

where $\hat{G}_1 = \mathcal{M}_1(\hat{\mathcal{G}})$.

Since $\frac{\partial}{\partial\mathcal{G}}\ell_n\left(\hat{\mathcal{G}} \times_1 \hat{U}_1^{(0)} \times_2 \hat{U}_2^{(0)} \times_3 \hat{U}_3^{(0)}\right) = 0$, we have

$$2\sum_{i=1}^{n}\hat{U}_1^{(0)\top}\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\langle\hat{G}_1, \hat{U}_1^{(0)\top}\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\rangle - 2\sum_{i=1}^{n}Y_i\hat{U}_1^{(0)\top}\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)}) = 0. \tag{5.109}$$

Denote

$$\Delta\mathcal{G} = \hat{\mathcal{G}} - \mathcal{G} \times_1 R_1 \times_2 R_2 \times_3 R_3, \quad \Delta G_1 = \mathcal{M}_1(\Delta\mathcal{G}) = \hat{G}_1 - R_1G_1(R_2^\top \otimes R_3^\top)$$

and

$$\begin{aligned}\mathcal{A}_{\mathcal{T}}^{(0)} =&\mathcal{G} \times_1 (\hat{U}_1^{(0)} R_1) \times_2 (\hat{U}_2^{(0)} R_2) \times_3 (\hat{U}_2^{(0)} R_2) - \mathcal{T},\\ A_{T_1}^{(0)} =&\mathcal{M}_1(\mathcal{A}_{\mathcal{T}}^{(0)}) = (\hat{U}_1^{(0)} R_1) G_1 \left((\hat{U}_2^{(0)} R_2)^\top \otimes (\hat{U}_3^{(0)} R_3)^\top \right) - T_1.\end{aligned}$$

By (5.109), we have

$$\begin{aligned}&\Delta G_1 - \mathfrak{J}_1\\ =&\Big(\mathcal{M}_1(\Delta\mathcal{G}) - \frac{1}{n}\sum_{i=1}^{n} \hat{U}_1^{(0)\top} \mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) \langle \Delta\mathcal{G} \times_1 \hat{U}_1^{(0)} \times_2 \hat{U}_2^{(0)} \times_3 \hat{U}_3^{(0)}, \mathcal{X}_i \rangle \Big)\\ &-\Big(\frac{1}{n}\sum_{i=1}^{n} \hat{U}_1^{(0)\top} \mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) \langle \mathcal{A}_{\mathcal{T}}^{(0)}, \mathcal{X}_i \rangle - \hat{U}_1^{(0)\top} \mathcal{M}_1(\mathcal{A}_{\mathcal{T}}^{(0)})(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) \Big)\\ &+\Big(\frac{1}{n}\sum_{i=1}^{n} \xi_i \hat{U}_1^{(0)\top} \mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) - \mathfrak{J}_1 \Big) - \hat{U}_1^{(0)\top} A_{T_1}^{(0)} (\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}). \quad (5.110)\end{aligned}$$

Notice that $\text{rank}(\mathcal{M}_i(\Delta\mathcal{G})) \leqslant 2r_i$ and $\text{rank}(\mathcal{M}_i(\mathcal{A}_{\mathcal{T}}^{(0)})) \leqslant 2r_i$ for $i \in [3]$, by Lemma 5.2.9, with probability at least $1 - e^{-C_1 pr}$,

$$\begin{aligned}&\left\| \mathcal{M}_1(\Delta\mathcal{G}) - \frac{1}{n}\sum_{i=1}^{n} \hat{U}_1^{(0)\top} \mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) \langle \Delta\mathcal{G} \times_1 \hat{U}_1^{(0)} \times_2 \hat{U}_2^{(0)} \times_3 \hat{U}_3^{(0)}, \mathcal{X}_i \rangle \right\|\\ &\leqslant C_2 \sqrt{\frac{pr}{n}} \|\Delta\mathcal{G}\|_F, \quad (5.111)\end{aligned}$$

and

$$\begin{aligned}&\left\| \frac{1}{n}\sum_{i=1}^{n} \hat{U}_1^{(0)\top} \mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) \langle \mathcal{A}_{\mathcal{T}}^{(0)}, \mathcal{X}_i \rangle - \hat{U}_1^{(0)\top} \mathcal{M}_1(\mathcal{A}_{\mathcal{T}}^{(0)})(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) \right\|\\ &\leqslant C_2 \sqrt{\frac{pr}{n}} \left\| \mathcal{A}_{\mathcal{T}}^{(0)} \right\|_F. \quad (5.112)\end{aligned}$$

By the definition of R_i and Lemma 5.2.3, with probability at least $1 - C_1 e^{-c_1 p}$,

$$\left\|\hat{U}_i^{(0)\top} U_i - R_i\right\| \leqslant \left\|U_{i\perp}^{\top} \hat{U}_i^{(0)}\right\|^2 \leqslant C \frac{p}{n\lambda_{\min}^2}, \tag{5.113}$$

and

$$\begin{aligned}\left\|U_i - \hat{U}_i^{(0)} R_i\right\| \leqslant& \left\|\mathcal{P}_{\hat{U}_i^{(0)}}(U_i - \hat{U}_i^{(0)} R_i)\right\| + \left\|\mathcal{P}^{\perp}_{\hat{U}_i^{(0)}}(U_i - \hat{U}_i^{(0)} R_i)\right\| \\ =& \left\|\hat{U}_i^{(0)\top} U_i - R_i\right\| + \left\|\left(\hat{U}_i^{(0)}\right)_{\perp}^{\top} U_i\right\| \leqslant C \frac{\sqrt{p/n}}{\lambda_{\min}}.\end{aligned} \tag{5.114}$$

Thus with probability at least $1 - C_1 e^{-c_1 p}$,

$$\begin{aligned}\left\|\mathcal{A}_{\mathcal{T}}^{(0)}\right\|_F =& \left\|\mathcal{G} \times_1 (\hat{U}_1^{(0)} R_1) \times_2 (\hat{U}_2^{(0)} R_2) \times_3 (\hat{U}_3^{(0)} R_3) - \mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3\right\|_F \\ \leqslant& \left\|\mathcal{G} \times_1 (\hat{U}_1^{(0)} R_1 - U_1) \times_2 (\hat{U}_2^{(0)} R_2) \times_3 (\hat{U}_3^{(0)} R_3)\right\|_F + \left\|\mathcal{G} \times_1 U_1 \times_2 (\hat{U}_2^{(0)} R_2 - U_2) \times_3 (\hat{U}_3^{(0)} R_3)\right\|_F \\ &+ \left\|\mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 (\hat{U}_3^{(0)} R_3 - U_3)\right\|_F \\ \leqslant& \left\|\hat{U}_1^{(0)} R_1 - U_1\right\|_F \|G_1\| + \left\|\hat{U}_2^{(0)} R_2 - U_2\right\|_F \|G_2\| + \left\|\hat{U}_3^{(0)} R_3 - U_3\right\|_F \|G_3\| \\ \leqslant& C \frac{\sqrt{pr/n}}{\lambda_{\min}} \cdot \kappa_0 \lambda_{\min} = C\kappa_0 \sqrt{pr/n}.\end{aligned}$$

The previous inequality and (5.112), with probability at least $1 - e^{-C_1 pr} - C_1 e^{-c_1 p}$,

$$\left\|\frac{1}{n} \sum_{i=1}^{n} \hat{U}_1^{(0)\top} \mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) \langle \mathcal{A}_{\mathcal{T}}^{(0)}, \mathcal{X}_i \rangle - \hat{U}_1^{(0)\top} \mathcal{M}_1(\mathcal{A}_{\mathcal{T}}^{(0)})(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)})\right\| \leqslant C_2 \kappa_0 \frac{pr}{n}. \tag{5.115}$$

By Lemma 5.2.10 and (5.114), with probability at least $1 - e^{-C_1 pr} - C_1 e^{-c_1 p}$,

$$\begin{aligned}&\left\|\frac{1}{n} \sum_{i=1}^{n} \xi_i \hat{U}_1^{(0)\top} \mathcal{M}_1(\mathcal{X}_i) \left(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}\right) - \mathfrak{J}_1\right\| \\ \leqslant& \left\|\frac{1}{n} \sum_{i=1}^{n} \xi_i \left(\hat{U}_1^{(0)} - U_1 R_1^{\top}\right)^{\top} \mathcal{M}_1(\mathcal{X}_i) \left(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}\right)\right\|\end{aligned}$$

$$
\begin{aligned}
&+\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\hat{U}_1^{(0)\top}\mathcal{M}_1(\mathcal{X}_i)\left(\left(\hat{U}_2^{(0)}-U_2R_2^\top\right)\otimes\hat{U}_3^{(0)}\right)\right\|\\
&+\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\hat{U}_1^{(0)\top}\mathcal{M}_1(\mathcal{X}_i)\left(\hat{U}_2^{(0)}\otimes\left(\hat{U}_3^{(0)}-U_3R_3^\top\right)\right)\right\|\\
\leqslant & C\sqrt{\frac{pr}{n}}\left(\left\|\hat{U}_1^{(0)}-U_1R_1^\top\right\|+\left\|\hat{U}_2^{(0)}-U_2R_2^\top\right\|+\left\|\hat{U}_3^{(0)}-U_3R_3^\top\right\|\right)\leqslant C_2\frac{p\sqrt{r}}{n\lambda_{\min}}.
\end{aligned}
\tag{5.116}
$$

In addition, by (5.113), with probability at least $1-C_1e^{-c_1p}$,

$$
\begin{aligned}
&\left\|\hat{U}_1^{(0)\top}A_{T_1}^{(0)}(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\right\|\\
=&\left\|R_1G_1\left(R_2^\top\otimes R_3^\top\right)-\left(\hat{U}_1^{(0)\top}U_1\right)G_1\left(\left(\hat{U}_2^{(0)\top}U_2\right)^\top\otimes\left(\hat{U}_3^{(0)\top}U_3\right)^\top\right)\right\|\\
\leqslant&\left\|\left(R_1-\hat{U}_1^{(0)\top}U_1\right)G_1\left(R_2^\top\otimes R_3^\top\right)\right\|+\left\|\left(\hat{U}_1^{(0)\top}U_1\right)G_1\left(\left(R_2-\hat{U}_2^{(0)\top}U_2\right)^\top\otimes R_3^\top\right)\right\|\\
&+\left\|\left(\hat{U}_1^{(0)\top}U_1\right)G_1\left(\left(\hat{U}_2^{(0)\top}U_2\right)^\top\otimes\left(R_3-\hat{U}_3^{(0)\top}U_3\right)^\top\right)\right\|\\
\leqslant&\left(\left\|R_1-\hat{U}_1^{(0)\top}U_1\right\|+\left\|R_2-\hat{U}_2^{(0)\top}U_2\right\|+\left\|R_3-\hat{U}_3^{(0)\top}U_3\right\|\right)\kappa_0\lambda_{\min}\\
\leqslant&C_2\frac{p}{n\lambda_{\min}^2}\kappa_0\lambda_{\min}=C_2\kappa_0\frac{p}{n\lambda_{\min}}.
\end{aligned}
\tag{5.117}
$$

Putting (5.110), (5.111), (5.115), (5.116) and (5.117) and Lemma 5.2.9 together, we get with probability $1-p^{-3}-C_1e^{-c_1p}$ that

$$
\begin{aligned}
\|\Delta G_1\|\leqslant&\|\mathfrak{J}_1\|+C_2\left(\sqrt{\frac{pr}{n}}\|\Delta\mathcal{G}\|_F+\kappa_0\frac{pr}{n}+\frac{p\sqrt{r}}{n\lambda_{\min}}+\kappa_0\frac{p}{n\lambda_{\min}}\right)\\
\leqslant&C_2\sqrt{\frac{pr^2}{n}}\|\Delta G_1\|+C_2\left(\sqrt{\frac{r^2+\log(p)}{n}}+\kappa_0\frac{pr}{n}+\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right)\\
\leqslant&\frac{1}{2}\|\Delta G_1\|+C_2\left(\sqrt{\frac{r^2+\log(p)}{n}}+\kappa_0\frac{pr}{n}+\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right)
\end{aligned}
$$

and as a result

$$
\|\Delta G_1\|\leqslant C_2\left(\sqrt{\frac{r^2+\log(p)}{n}}+\kappa_0\frac{pr}{n}+\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right)
\tag{5.118}
$$

Therefore, with probability at least $1-p^{-3}-C_1e^{-c_1p}$,

$$\|\Delta G_1-\mathfrak{J}_1\|\leqslant C_2\left(\sqrt{\frac{pr^2}{n}}\|\Delta G_1\|+\kappa_0\frac{pr}{n}+\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right)\leqslant C_2\left(\kappa_0\frac{pr}{n}+\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right)$$

which proves (5.108).

Step 0.2: the error of $\hat{U}_j^{(t+0.5)}$. Without loss of generality, we only prove the bound for $t=0$ and $j=1$. Again, we denote $\hat{\mathcal{G}}=\hat{\mathcal{G}}^{(0)}$ and $\hat{G}_1=\mathcal{M}_1(\hat{\mathcal{G}})$ for brevity.

We aim to show that with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\hat{U}_1^{(0.5)}=U_1R_1^\top+\frac{1}{n}\mathcal{P}_{U_1}^\perp\Big(\sum_{j=1}^n\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)(U_2\otimes U_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top+\mathfrak{E},\tag{5.119}$$

where $\|\mathfrak{E}\|\leqslant C_2\left(\kappa_0\frac{pr}{n\lambda_{\min}}+\kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}\right)$.

Since $\frac{\partial}{\partial U_1}\ell_n(\hat{\mathcal{G}}\times_1\hat{U}_1^{(0.5)}\times_2\hat{U}_2^{(0)}\times_3\hat{U}_3^{(0)})=0$, we have

$$2\sum_{i=1}^n\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\hat{G}_1^\top\langle\hat{U}_1^{(0.5)},\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\hat{G}_1^\top\rangle-2\sum_{i=1}^n\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\hat{G}_1^\top Y_i=0.$$

Denote

$$\Delta\mathcal{T}_1=\hat{\mathcal{G}}\times_1\hat{U}_1^{(0.5)}\times_2\hat{U}_2^{(0)}\times_3\hat{U}_3^{(0)}-\mathcal{T}.$$

For brevity, denote $\hat{U}_1=\hat{U}_1^{(0.5)}$ for simplicity. Then, we write

$$\begin{aligned}&\hat{U}_1(\hat{G}_1\hat{G}_1^\top)-U_1R_1^\top(\hat{G}_1\hat{G}_1^\top)\\=&\left(\mathcal{M}_1(\Delta\mathcal{T}_1)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})-\frac{1}{n}\sum_{i=1}^n\langle\Delta\mathcal{T}_1,\mathcal{X}_i\rangle\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\right)\hat{G}_1^\top\\&+\Big(U_1G_1\big((\hat{U}_2^{(0)\top}U_2)^\top\otimes(\hat{U}_3^{(0)\top}U_3)^\top\big)\hat{G}_1^\top-U_1R_1^\top\big(\hat{G}_1\hat{G}_1^\top\big)\Big)\\&+\frac{1}{n}\sum_{i=1}^n\xi_i\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\hat{G}_1^\top,\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\hat{U}_1 - U_1 R_1^\top\\
=&\Big(\mathcal{M}_1(\Delta\mathcal{T}_1)(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) - \frac{1}{n}\sum_{i=1}^n \langle \Delta\mathcal{T}_1, \mathcal{X}_i\rangle \mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\Big)\hat{G}_1^\top\big(\hat{G}_1\hat{G}_1^\top\big)^{-1}\\
&+\Big(U_1G_1\big((\hat{U}_2^{(0)\top}U_2)^\top \otimes (\hat{U}_3^{(0)\top}U_3)^\top\big)\hat{G}_1^\top - U_1R_1^\top\big(\hat{G}_1\hat{G}_1^\top\big)\Big)\hat{G}_1^\top\big(\hat{G}_1\hat{G}_1^\top\big)^{-1}\\
&+\frac{1}{n}\sum_{i=1}^n \xi_i \mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes \hat{U}_3^{(0)})\hat{G}_1^\top\big(\hat{G}_1\hat{G}_1^\top\big)^{-1}\\
=:&\mathrm{I}+\mathrm{II}+\mathrm{III}.
\end{aligned} \tag{5.120}$$

By (5.108) and Lemma 5.2.9, with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}
&\left\|\Big(\mathcal{M}_1(\Delta\mathcal{T}_1)(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) - \frac{1}{n}\sum_{i=1}^n \langle \Delta\mathcal{T}_1, \mathcal{X}_i\rangle \mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\Big)\hat{G}_1^\top\big(\hat{G}_1\hat{G}_1^\top\big)^{-1}\right\|\\
\leqslant & C\sqrt{\frac{pr}{n}}\|\Delta\mathcal{T}_1\|_F\left\|\hat{G}_1^\top\big(\hat{G}_1\hat{G}_1^\top\big)^{-1}\right\|\\
\leqslant & C\sqrt{\frac{pr}{n}}\left\|\hat{\mathcal{G}}\times_1\hat{U}_1\times_2\hat{U}_2^{(0)}\times_3\hat{U}_3^{(0)}-\mathcal{T}\right\|_F\left(\lambda_{\min} - C_2\Big(\sqrt{\frac{r^2+\log(p)}{n}}+\kappa_0\frac{pr}{n}+\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\Big)\right)^{-1}\\
\leqslant & C\sqrt{\frac{pr}{n}}\lambda_{\min}^{-1}\left\|\hat{\mathcal{G}}\times_1\hat{U}_1\times_2\hat{U}_2^{(0)}\times_3\hat{U}_3^{(0)}-\mathcal{T}\right\|_F.
\end{aligned}$$

Moreover, with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}
&\big\|\hat{\mathcal{G}}\times_1\hat{U}_1\times_2\hat{U}_2^{(0)}\times_3\hat{U}_3^{(0)}-\mathcal{T}\big\|_F\\
\leqslant&\big\|\big(\hat{\mathcal{G}}-\mathcal{G}\times_1R_1\times_2R_2\times_3R_3\big)\times_1\hat{U}_1\times_2\hat{U}_2^{(0)}\times_3\hat{U}_3^{(0)}\big\|_F\\
&+\big\|\mathcal{G}\times_1\big(\hat{U}_1R_1\big)\times_2\big(\hat{U}_2^{(0)}R_2\big)\times_3\big(\hat{U}_3^{(0)}R_3\big)-\mathcal{G}\times_1U_1\times_2U_2\times_3U_3\big\|_F\\
\leqslant&\sqrt{r}\big\|\hat{\mathcal{G}}-\mathcal{G}\times_1R_1\times_2R_2\times_3R_3\big\|+\sqrt{r}\big\|U_1-\hat{U}_1R_1\big\|\|G_1\|+\sqrt{r}\big\|U_2-\hat{U}_2^{(0)}R_2\big\|\|G_2\|\\
&+\sqrt{r}\big\|U_3-\hat{U}_3^{(0)}R_3\big\|\|G_3\|\\
\leqslant&C_2\sqrt{r}\Big(\sqrt{\frac{r^2+\log(p)}{n}}+\kappa_0\frac{pr}{n}+\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\Big)+\sqrt{r}\kappa_0\lambda_{\min}\big\|U_1-\hat{U}_1R_1\big\|+C_2\kappa_0\sqrt{\frac{pr}{n}}
\end{aligned}$$

$$\leqslant C_2\left(\kappa_0\sqrt{\frac{pr}{n}} + \kappa_0\frac{pr^{3/2}}{n} + \kappa_0\frac{pr}{n\lambda_{\min}}\right) + \sqrt{r}\kappa_0\lambda_{\min}\big\|U_1 - \hat{U}_1R_1\big\|. \tag{5.121}$$

The two previous inequalities together imply that with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}
\|I\| =& \left\|\left(\mathcal{M}_1(\Delta\mathcal{T}_1)(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)}) - \frac{1}{n}\sum_{i=1}^{n}\langle\Delta\mathcal{T}_1, \mathcal{X}_i\rangle\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)} \otimes \hat{U}_3^{(0)})\right)\hat{G}_1^\top\left(\hat{G}_1\hat{G}_1^\top\right)^{-1}\right\| \\
\leqslant& C_2\left(\frac{\kappa_0 pr}{n\lambda_{\min}} + \kappa_0\frac{p^{3/2}r^2}{n^{3/2}\lambda_{\min}} + \kappa_0\frac{p^{3/2}r^{3/2}}{n^{3/2}\lambda_{\min}^2}\right) + C_2\kappa_0 r\sqrt{\frac{p}{n}}\big\|U_1 - \hat{U}_1R_1\big\|.
\end{aligned} \tag{5.122}$$

By (5.118), with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$, we have

$$\begin{aligned}
&\left\|\hat{G}_1^\top\left(\hat{G}_1\hat{G}_1^\top\right)^{-1} - (R_2 \otimes R_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\right\| \\
\leqslant& \left\|\left(\hat{G}_1 - R_1G_1(R_2^\top \otimes R_3^\top)\right)^\top\left(\hat{G}_1\hat{G}_1^\top\right)^{-1}\right\| \\
&+ \left\|(R_2 \otimes R_3)G_1^\top R_1^\top\left(R_1(G_1G_1^\top)^{-1}R_1^\top - \left(\hat{G}_1\hat{G}_1^\top\right)^{-1}\right)\right\| \\
\leqslant& C_2\left(\sqrt{\frac{r^2 + \log(p)}{n}} + \kappa_0\frac{pr}{n} + \kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right)\lambda_{\min}^{-2} \\
&+ \|G_1\|\big\|(G_1G_1^\top)^{-1}\big\|\left\|\left(\hat{G}_1\hat{G}_1^\top\right)^{-1}\right\|\big\|R_1G_1G_1^\top R_1^\top - \hat{G}_1\hat{G}_1^\top\big\| \\
\leqslant& C_2\left(\sqrt{\frac{r^2 + \log(p)}{n}} + \kappa_0\frac{pr}{n} + \kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right)\lambda_{\min}^{-2} \\
&+ C_2\kappa_0\lambda_{\min}^{-3}\left(\left\|\left(R_1G_1(R_2^\top \otimes R_3^\top) - \hat{G}_1\right)(R_2 \otimes R_3)G_1^\top R_1^\top\right\| + \left\|\hat{G}_1\left(R_1G_1(R_2^\top \otimes R_3^\top) - \hat{G}_1\right)^\top\right\|\right) \\
\leqslant& C_2\kappa_0^2\lambda_{\min}^{-2}\left(\sqrt{\frac{r^2 + \log(p)}{n}} + \kappa_0\frac{pr}{n} + \kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right),
\end{aligned} \tag{5.123}$$

and

$$\begin{aligned}
&\left\|U_1G_1\left((\hat{U}_2^{(0)\top}U_2)^\top \otimes (\hat{U}_3^{(0)\top}U_3)^\top\right) - U_1R_1^\top\hat{G}_1\right\| \\
\leqslant& \left\|G_1\left((\hat{U}_2^{(0)\top}U_2 - R_2)^\top \otimes (\hat{U}_3^{(0)\top}U_3)^\top\right)\right\| + \left\|G_1\left(R_2^\top \otimes (\hat{U}_3^{(0)\top}U_3 - R_3)^\top\right)\right\| \\
&+ \big\|R_1G_1\left(R_2^\top \otimes R_3^\top\right) - \hat{G}_1\big\|
\end{aligned}$$

$$
\begin{aligned}
&\leqslant C_2\kappa_0\lambda_{\min}\left(\frac{\sqrt{p/n}}{\lambda_{\min}}\right)^2 + C_2\left(\sqrt{\frac{r^2+\log(p)}{n}} + \kappa_0\frac{pr}{n} + \kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right) \\
&\leqslant C_2\left(\sqrt{\frac{r^2+\log(p)}{n}} + \kappa_0\frac{pr}{n} + \kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right).
\end{aligned}
$$

By the two previous inequality and (5.108), with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$
\begin{aligned}
&\left\|\mathrm{II} + \frac{1}{n}\mathcal{P}_{U_1}\Big(\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)(U_2\otimes U_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\right\| \\
\leqslant &\left\|U_1G_1\left((\hat{U}_2^{(0)\top}U_2)^\top\otimes(\hat{U}_3^{(0)\top}U_3)^\top\right) - U_1R_1^\top\hat{G}_1\right\|\left\|\hat{G}_1^\top\left(\hat{G}_1\hat{G}_1^\top\right)^{-1} - (R_2\otimes R_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\right\| \\
&+\left\|U_1G_1\left(R_2^\top\otimes R_3^\top\right) - U_1R_1^\top\hat{G}_1 + \frac{1}{n}\mathcal{P}_{U_1}\Big(\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)\big((U_2R_2^\top)\otimes(U_3R_3^\top)\big)\right\| \\
&\cdot\left\|(R_2\otimes R_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\right\| \\
&+\left(\left\|G_1\big((\hat{U}_2^{(0)\top}U_2-R_2)^\top\otimes(\hat{U}_3^{(0)\top}U_3)^\top\big)\right\| + \left\|G_1\big(R_2^\top\otimes(\hat{U}_3^{(0)\top}U_3-R_3)^\top\big)\right\|\right) \\
&\cdot\left\|(R_2\otimes R_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\right\| \\
\leqslant &C_2\kappa_0^2\lambda_{\min}^{-2}\left(\frac{r^2+\log(p)}{n} + \kappa_0^2\frac{p^2r^2}{n^2} + \kappa_0^2\frac{p^2r}{n^2\lambda_{\min}^2}\right) + C_2\kappa_0\left(\frac{pr}{n}+\frac{p\sqrt{r}}{n\lambda_{\min}}\right)\lambda_{\min}^{-1} + C_2\kappa_0\lambda_{\min}\left(\frac{\sqrt{p/n}}{\lambda_{\min}}\right)^2\lambda_{\min}^{-1} \\
\leqslant &C_2\left(\kappa_0\frac{pr}{n\lambda_{\min}} + \kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}\right).
\end{aligned}
\tag{5.124}
$$

For term III, by (5.123) and Lemma 5.2.10 Part 3, with probability $1-C_1p^{-3}-C_1e^{-c_1p}$ that

$$
\begin{aligned}
&\left\|\mathrm{III} - \frac{1}{n}\Big(\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)(U_2\otimes U_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\right\| \\
\leqslant &\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(0)}\otimes\hat{U}_3^{(0)})\Big(\hat{G}_1^\top\left(\hat{G}_1\hat{G}_1^\top\right)^{-1} - (R_2\otimes R_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\Big)\right\| \\
&+\left\|\frac{1}{n}\Big(\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)\big((\hat{U}_2^{(0)}R_2-U_2)\otimes(\hat{U}_3^{(0)}R_3)\big)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\right\|
\end{aligned}
$$

$$\begin{aligned}&+\Big\|\frac{1}{n}\Big(\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)\big(U_2\otimes(\hat{U}_3^{(0)}R_3-U_3)\big)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\Big\|\\ \leqslant& C_2\sqrt{\frac{pr}{n}}\Big\|\hat{G}_1^\top\big(\hat{G}_1\hat{G}_1^\top\big)^{-1}-(R_2\otimes R_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\Big\|\\ &+C_2\sqrt{\frac{pr}{n}}\|\hat{U}_2^{(0)}R_2-U_2\|\lambda_{\min}^{-1}+C_2\sqrt{\frac{pr}{n}}\|\hat{U}_3^{(0)}R_3-U_3\|\lambda_{\min}^{-1}\leqslant C_2\kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}.\end{aligned}\tag{5.125}$$

Putting (5.120), (5.122), (5.124) and (5.125) and Lemma 5.2.10 Part 3 together, we get with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$ that

$$\begin{aligned}\|\hat{U}_1-U_1R_1^\top\|\leqslant&\Big\|\frac{1}{n}\mathcal{P}_{U_1}^{\perp}\Big(\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)(U_2\otimes U_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\Big\|\\ &+C_2\kappa_0r\sqrt{\frac{p}{n}}\|\hat{U}_1-U_1R_1^\top\|+C_2\left(\kappa_0\frac{pr}{n\lambda_{\min}}+\kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}\right)\\ \leqslant& C_2\sqrt{\frac{p}{n}}\lambda_{\min}^{-1}+\frac{1}{2}\|\hat{U}_1-U_1R_1^\top\|+C_2\left(\kappa_0\frac{pr}{n\lambda_{\min}}+\kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}\right)\\ \leqslant&\frac{1}{2}\|\hat{U}_1-U_1R_1^\top\|+C_2\sqrt{\frac{p}{n}}\lambda_{\min}^{-1}.\end{aligned}$$

Therefore, with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\|\hat{U}_1-U_1R_1^\top\|\leqslant C_2\sqrt{\frac{p}{n}}\lambda_{\min}^{-1}.\tag{5.126}$$

Thus with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\begin{aligned}&\Big\|\hat{U}_1^{(0.5)}-U_1R_1^\top-\frac{1}{n}\mathcal{P}_{U_1}^{\perp}\Big(\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)(U_2\otimes U_3)G_1^\top(G_1G_1^\top)^{-1}R_1^\top\Big\|\\ \leqslant& C_2\left(\kappa_0\frac{pr}{n\lambda_{\min}}+\kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}\right)+C_2\kappa_0r\sqrt{\frac{p}{n}}\cdot\sqrt{\frac{p}{n}}\lambda_{\min}^{-1}\\ \leqslant& C_2\left(\kappa_0\frac{pr}{n\lambda_{\min}}+\kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}\right).\end{aligned}$$

Now, we continue from eq. (5.107) and prove the distribution of $\|\hat{U}_1^{(2)}\hat{U}_1^{(2)\top} - U_1U_1^\top\|_F^2$.

Step 1: bounding $\|\hat{U}_j^{(1)\top}U_j - R_j^{(1)}\|_F$ and $\|U_j - \hat{U}_j^{(1)}R_j^{(1)}\|$. Without loss of generality, we only prove the bound for $j = 1$. By definition of $\hat{U}_1^{(t+0.5)}$ in Algorithm 2, we write

$$
\begin{aligned}
&\hat{U}_1^{(t+0.5)} - U_1R_1^{(t)\top} \\
=&\Big(\mathcal{M}_1(\Delta\mathcal{T}_1^{(t+0.5)})(\hat{U}_2^{(t)} \otimes \hat{U}_3^{(t)}) - \frac{1}{n}\sum_{i=1}^{n}\langle\Delta\mathcal{T}_1^{(t+0.5)}, \mathcal{X}_i\rangle\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(t)} \otimes \hat{U}_3^{(t)})\Big)\hat{G}_1^{(t)\top}\big(\hat{G}_1^{(t)}\hat{G}_1^{(t)\top}\big)^{-1} \\
&+ \Big(U_1G_1\big((\hat{U}_2^{(t)\top}U_2)^\top \otimes (\hat{U}_3^{(t)\top}U_3)^\top\big) - U_1R_1^{(t)\top}\hat{G}_1^{(t)}\Big)\,\hat{G}_1^{(t)\top}\big(\hat{G}_1^{(t)}\hat{G}_1^{(t)\top}\big)^{-1} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(t)} \otimes \hat{U}_3^{(t)})\hat{G}_1^{(t)\top}\big(\hat{G}_1^{(t)}\hat{G}_1^{(t)\top}\big)^{-1} \\
=:&\mathfrak{J}_{U_1,1}^{(t)} + \mathfrak{J}_{U_1,2}^{(t)} + \mathfrak{J}_{U_1,3}^{(t)}.
\end{aligned}
\tag{5.127}
$$

Denote $\mathfrak{E}_1^{(t)} = \mathfrak{J}_{U_1,1}^{(t)} + \mathfrak{J}_{U_1,2}^{(t)} + \mathfrak{J}_{U_1,3}^{(t)}$. Recall that $\hat{U}_1^{(t+1)}$ are the left singular vectors of $\hat{U}_1^{(t+0.5)}$. We can also apply the spectral representation formula (Lemma 5.2.2) to investigate $\hat{U}_1^{(t+1)}$. Toward that end, we define

$$
\begin{pmatrix} 0 & \hat{U}_1^{(t+0.5)} \\ \hat{U}_1^{(t+0.5)\top} & 0 \end{pmatrix} = \begin{pmatrix} 0 & U_1R_1^{(t)\top} \\ R_1^{(t)}U_1^\top & 0 \end{pmatrix} + \begin{pmatrix} 0 & \mathfrak{E}_1^{(t)} \\ \mathfrak{E}_1^{(t)\top} & 0 \end{pmatrix}.
$$

Note that the non-zero eigenvalues of the symmetric matrix

$$
\begin{pmatrix} 0 & U_1R_1^{(t)\top} \\ R_1^{(t)}U_1^\top & 0 \end{pmatrix}
$$

are $\mu_1^{(t)} = \cdots = \mu_{r_1}^{(t)} = 1$ and $\mu_{r_1+1}^{(t)} = \cdots = \mu_{2r_1}^{(t)} = -1$, and for $1 \leqslant i \leqslant r_1$, the corresponding eigenvectors of $\mu_i^{(t)}$ and $\mu_{r_1+i}^{(t)}$ are

$$\theta_i^{(t)} = \frac{1}{\sqrt{2}} \begin{pmatrix} \bar{u}_i^{(t)} \\ e_i \end{pmatrix} \quad \text{and} \quad \theta_{r_1+i}^{(t)} = \frac{1}{\sqrt{2}} \begin{pmatrix} \bar{u}_i^{(t)} \\ -e_i \end{pmatrix},$$

where $\bar{u}_i^{(t)}$ is the i-th column of $U_1 R_1^{(t)\top}$ and e_i is the i-th canonical basis of $\mathbb{R}^{r_1}$.

Denote a $(p_1 + r_1) \times (2r_1)$ matrix

$$\Theta^{(t)} = \left(\theta_1^{(t)} \ldots \theta_{2r_1}^{(t)}\right)$$

and $\Theta_\perp^{(t)} \in \mathbb{O}_{p_1+r_1,p_1-r_1}$ such that $\left(\Theta^{(t)}\ \Theta_\perp^{(t)}\right) \in \mathbb{O}_{p_1+r_1}$. Then, we write

$$\Theta^{(t)}\Theta^{(t)\top} = \sum_{1\leqslant j\leqslant 2r_1} \theta_j^{(t)}\theta_j^{(t)\top} = \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}$$

For $k \geqslant 1$, denote

$$\left(\mathfrak{P}_1^{(t)}\right)^{-k} = \sum_{1\leqslant j\leqslant 2r_1} \frac{1}{(\mu_j^{(t)})^k}\theta_j^{(t)}\theta_j^{(t)\top} = \begin{cases} \begin{pmatrix} 0 & U_1R_1^{(t)\top} \\ R_1^{(t)}U_1^\top & 0 \end{pmatrix}, & \text{if } k \text{ is old,} \\ \begin{pmatrix} UU^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, & \text{if } k \text{ is even,} \end{cases}$$

and

$$\left(\mathfrak{P}_1^{(t)}\right)^0 = \Theta_\perp^{(t)}\Theta_\perp^{(t)\top} = \begin{pmatrix} U_{1\perp}U_{1\perp}^\top & 0 \\ 0 & 0 \end{pmatrix}.$$

Let

$$E^{(t)} = \begin{pmatrix} 0 & \mathfrak{E}_1^{(t)} \\ \mathfrak{E}_1^{(t)\top} & 0 \end{pmatrix}.$$

By Lemma 5.2.10 and together with (5.119), with probability at least $1 - C_1p^{-3} -$

$C_1 e^{-c_1 p}$,

$$\left\|E^{(0)}\right\| = \left\|\mathfrak{E}_1^{(0)}\right\| \leqslant C_2 \frac{\sqrt{p/n}}{\lambda_{\min}} + C_2 \left(\kappa_0 \frac{pr}{n\lambda_{\min}} + \kappa_0^2 \frac{p\sqrt{r}}{n\lambda_{\min}^2}\right) \leqslant C_2 \frac{\sqrt{p/n}}{\lambda_{\min}} < \frac{1}{8}. \quad (5.128)$$

By Lemma 5.2.2, with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\begin{pmatrix} \hat{U}_1^{(1)} \hat{U}_1^{(1)\top} & 0 \\ 0 & I_{r_1} \end{pmatrix} - \begin{pmatrix} U_1 U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix} = \sum_{k \geqslant 1} \mathcal{S}_{U_1^{(0)},k}\big(E^{(0)}\big) \quad (5.129)$$

where

$$\mathcal{S}_{U_1^{(t)},k}(X) = \sum_{s_1 + \cdots + s_{k+1} = k} (-1)^{1+\tau(\mathbf{s})} \cdot \big(\mathfrak{P}_1^{(t)}\big)^{-s_1} X \big(\mathfrak{P}_1^{(t)}\big)^{-s_2} X \big(\mathfrak{P}_1^{(t)}\big)^{-s_3} \cdots \big(\mathfrak{P}_1^{(t)}\big)^{-s_k} X \big(\mathfrak{P}_1^{(t)}\big)^{-s_{k+1}}$$

where $s_1, \cdots, s_{k+1}$ are non-negative integers and $\tau(\mathbf{s}) = \sum_{j=1}^{k+1} \mathbb{I}(s_j > 0)$.

Clearly, we have

$$\left\|\mathcal{S}_{U_1^{(0)},k}\big(E^{(0)}\big)\right\| \leqslant \binom{2k}{k} \|E^{(0)}\|^{k-1} \|E^{(0)}\|_F \leqslant \big(4\|E^{(0)}\|\big)^k. \quad (5.130)$$

By (5.128), (5.129) and (5.130), with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\left\|\hat{U}_1^{(1)} \hat{U}_1^{(1)\top} - U_1 U_1^\top\right\| \leqslant \sum_{k \geqslant 1} \left\|\mathcal{S}_{U_1^{(0)},k}\big(E^{(0)}\big)\right\| \leqslant C_2 \frac{\sqrt{p/n}}{\lambda_{\min}}.$$

Thus with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\left\|\hat{U}_1^{(1)\top} U_1 - R_1^{(1)}\right\| \leqslant C_2 \frac{p}{n\lambda_{\min}^2}, \quad \left\|\hat{U}_1^{(1)\top} U_1 - R_1^{(1)}\right\|_F \leqslant C_2 \frac{p\sqrt{r}}{n\lambda_{\min}^2} \quad (5.131)$$

and

$$\left\|U_1 - \hat{U}_1^{(1)} R_1^{(1)}\right\| \leqslant C_2 \frac{\sqrt{p/n}}{\lambda_{\min}}, \quad \left\|U_1 - \hat{U}_1^{(1)} R_1^{(1)}\right\|_F \leqslant C_2 \frac{\sqrt{pr/n}}{\lambda_{\min}}. \quad (5.132)$$

Step 2: representation of $\|\hat{U}_1^{(2)}\hat{U}_1^{(2)\top} - U_1U_1^\top\|_F^2$ and its first order approximation. For convenience, we denote

$$E = E^{(1)}, \quad \mathfrak{E}_1 = \mathfrak{E}_1^{(1)} \quad \text{and} \quad \mathfrak{P}_1^{-k} = \left(\mathfrak{P}_1^{(1)}\right)^{-k}.$$

Similarly to *Step 1*, we apply Lemma 5.2.2 to $\hat{U}_1^{(2)}\hat{U}_1^{(2)\top}$ and get with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$ that

$$\begin{aligned}
&\|\hat{U}_1^{(2)}\hat{U}_1^{(2)\top} - U_1U_1^\top\|_F^2 \\
=& -2\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1}\end{pmatrix}, \begin{pmatrix} \hat{U}_1^{(2)}\hat{U}_1^{(2)\top} - U_1U_1^\top & 0 \\ 0 & 0\end{pmatrix}\right\rangle \\
=& -2\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1}\end{pmatrix}, \sum_{k\geqslant 1} \mathcal{S}_{U_1^{(1)},k}(E)\right\rangle \\
=& -2\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1}\end{pmatrix}, \sum_{k\geqslant 4} \mathcal{S}_{U_1^{(1)},k}(E)\right\rangle - 2\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1}\end{pmatrix}, \mathcal{S}_{U_1^{(1)},2}(E)\right\rangle \\
&-2\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1}\end{pmatrix}, \mathcal{S}_{U_1^{(1)},3}(E)\right\rangle
\end{aligned} \tag{5.133}$$

where we use the fact $\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1}\end{pmatrix}, \mathcal{S}_{U_1^{(1)},1}(E)\right\rangle = 0$.

Similarly to (5.130), we know that with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}
\left|2\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1}\end{pmatrix}, \sum_{k\geqslant 4} \mathcal{S}_{U_1^{(1)},k}(E)\right\rangle\right| \leqslant 4r\sum_{k\geqslant 4}\left\|\mathcal{S}_{U_1^{(1)},k}(E)\right\| &\leqslant C_2 r\left(\frac{\sqrt{p/n}}{\lambda_{\min}}\right)^4 \\
&\leqslant C_2 r\frac{p^2}{n^2\lambda_{\min}^4}.
\end{aligned} \tag{5.134}$$

We now bound the third order term. Notice that

$$\mathfrak{P}_1^0 E \mathfrak{P}_1^0 = \begin{pmatrix} U_{1\perp}U_{1\perp}^\top & 0 \\ 0 & 0\end{pmatrix}\begin{pmatrix} 0 & \mathfrak{E}_1 \\ \mathfrak{E}_1^\top & 0\end{pmatrix}\begin{pmatrix} U_{1\perp}U_{1\perp}^\top & 0 \\ 0 & 0\end{pmatrix} = 0.$$

Therefore, we have

$$\begin{aligned}
&\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \mathcal{S}_{U_1^{(1)},3}(E) \right\rangle \\
=&-2\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \mathfrak{P}_1^{-1}E\mathfrak{P}_1^{0}E\mathfrak{P}_1^{0}E\mathfrak{P}_1^{-2} \right\rangle + 2\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \mathfrak{P}_1^{-1}E\mathfrak{P}_1^{-1}E\mathfrak{P}_1^{0}E\mathfrak{P}_1^{-1} \right\rangle \\
=&2\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \mathfrak{P}_1^{-1}E\mathfrak{P}_1^{-1}E\mathfrak{P}_1^{0}E\mathfrak{P}_1^{-1} \right\rangle. \qquad (5.135)
\end{aligned}$$

By simple calculation, we have

$$\mathfrak{P}_1^{-1}E\mathfrak{P}_1^{-1}E\mathfrak{P}_1^{0}E\mathfrak{P}_1^{-1} = \begin{pmatrix} U_1R_1^{(1)}\mathfrak{E}_1^\top U_1R_1^{(1)}\mathfrak{E}_1^\top \mathcal{P}_{U_1}^\perp \mathfrak{E}_1 R_1^{(1)\top}U_1^\top & 0 \\ 0 & 0 \end{pmatrix}.$$

Therefore, with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}
&\left|\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \mathfrak{P}_1^{-1}E\mathfrak{P}_1^{-1}E\mathfrak{P}_1^{0}E\mathfrak{P}_1^{-1} \right\rangle\right| \\
=&\left|\left\langle \begin{pmatrix} U_1U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \begin{pmatrix} U_1R_1^{(1)\top}\mathfrak{E}_1^\top U_1R_1^{(1)\top}\mathfrak{E}_1^\top \mathcal{P}_{U_1}^\perp \mathfrak{E}_1 R_1^{(1)}U_1^\top & 0 \\ 0 & 0 \end{pmatrix} \right\rangle\right| \\
=&\left|\operatorname{tr}\left(U_1R_1^{(1)\top}\mathfrak{E}_1^\top U_1R_1^{(1)\top}\mathfrak{E}_1^\top \mathcal{P}_{U_1}^\perp \mathfrak{E}_1 R_1^{(1)}U_1^\top\right)\right| = \left|\operatorname{tr}\left(\mathfrak{E}_1^\top U_1R_1^{(1)\top}\mathfrak{E}_1^\top \mathcal{P}_{U_1}^\perp \mathfrak{E}_1\right)\right| \\
\leqslant& r\|\mathfrak{E}_1^\top U_1\|\|\mathfrak{E}_1\|^2 \leqslant C_2\frac{pr}{n\lambda_{\min}^2}\|\mathfrak{E}_1^\top U_1\|.
\end{aligned}$$

Similarly to *Step 0.2* and by Lemma 5.2.10 and $\mathcal{P}_{U_1}^\perp U_1 = 0$, with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}
\|\mathfrak{E}_1^\top U_1\| \leqslant& \left\| \left(\frac{1}{n}\mathcal{P}_{U_1}\Big(\sum_{j=1}^{n} \xi_j \mathcal{M}_1(\mathcal{X}_j)\Big)(U_2 \otimes U_3)G_1^\top (G_1G_1^\top)^{-1}R_1^{(1)\top} \right)^\top U_1 \right\| \\
&+ C_2\left(\kappa_0\frac{pr}{n\lambda_{\min}} + \kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}\right) \leqslant C_2\left(\sqrt{\frac{r^2+\log(p)}{n\lambda_{\min}^2}} + \kappa_0\frac{pr}{n\lambda_{\min}} + \kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}\right).
\end{aligned}$$

Combining (5.135) and the above two inequalities together, we get with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\begin{aligned}\left|\left\langle \begin{pmatrix} U_1 U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \mathcal{S}_{U_1^{(1)},3}(E) \right\rangle\right| \leqslant & C_2 \frac{pr}{n\lambda_{\min}^2} \left(\sqrt{\frac{r^2 + \log(p)}{n\lambda_{\min}^2}} + \kappa_0 \frac{pr}{n\lambda_{\min}} + \kappa_0^2 \frac{p\sqrt{r}}{n\lambda_{\min}^2} \right) \\ \leqslant & C_2 \left(\frac{pr(r + \sqrt{\log(p)})}{n^{3/2}\lambda_{\min}^3} + \kappa_0 \frac{p^2 r^2}{n^2 \lambda_{\min}^3} + \kappa_0^2 \frac{p^2 r^{3/2}}{n^2 \lambda_{\min}^4} \right).\end{aligned} \tag{5.136}$$

Therefore, we conclude that with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\begin{aligned}&\left| \left\| \hat{U}_1^{(2)} \hat{U}_1^{(2)\top} - U_1 U_1^\top \right\|_F^2 + 2 \left\langle \begin{pmatrix} U_1 U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \mathcal{S}_{U_1^{(1)},2}(E) \right\rangle \right| \\ &\leqslant C_2 \left(\frac{pr(r + \sqrt{\log(p)})}{n^{3/2}\lambda_{\min}^3} + \kappa_0 \frac{p^2 r^2}{n^2 \lambda_{\min}^3} + \kappa_0^2 \frac{p^2 r^{3/2}}{n^2 \lambda_{\min}^4} \right).\end{aligned} \tag{5.137}$$

Step 3: representing the leading term of $\|\hat{U}_1^{(2)} \hat{U}_1^{(2)\top} - U_1 U_1^\top\|_F^2$**.** Recall from *Step 2*, the leading term of $\|\hat{U}_1^{(2)} \hat{U}_1^{(2)\top} - U_1 U_1^\top\|_F^2$ is

$$-2 \left\langle \begin{pmatrix} U_1 U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \mathcal{S}_{U_1^{(1)},2}(E) \right\rangle$$

In *Step 3*, we aim to approximate this leading term by a sum of independent random variables. By definition of $\mathcal{S}_{U_1^{(1)},2}(E)$, we have

$$\begin{aligned}\left\langle \begin{pmatrix} U_1 U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \mathcal{S}_{U_1^{(1)},2}(E) \right\rangle =& - \left\langle \begin{pmatrix} U_1 U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \mathfrak{P}_1^{-1} E \mathfrak{P}_1^0 E \mathfrak{P}_1^{-1} \right\rangle \\ =& - \left\langle \begin{pmatrix} U_1 U_1^\top & 0 \\ 0 & I_{r_1} \end{pmatrix}, \begin{pmatrix} U_1 R_1^{(1)\top} \mathfrak{E}_1^\top \mathcal{P}_{U_1}^\perp \mathfrak{E}_1 R_1^{(1)} U_1^\top & 0 \\ 0 & 0 \end{pmatrix} \right\rangle \\ =& - \operatorname{tr}\left(U_1 R_1^{(1)\top} \mathfrak{E}_1^\top \mathcal{P}_{U_1}^\perp \mathfrak{E}_1 R_1^{(1)} U_1^\top \right) = - \operatorname{tr}\left(\mathfrak{E}_1^\top \mathcal{P}_{U_1}^\perp \mathfrak{E}_1 \right) \\ =& - \operatorname{tr}\left(\left(\mathfrak{J}_{U_1,1}^{(1)} + \mathfrak{J}_{U_1,3}^{(1)} \right)^\top \mathcal{P}_{U_1}^\perp \left(\mathfrak{J}_{U_1,1}^{(1)} + \mathfrak{J}_{U_1,3}^{(1)} \right) \right).\end{aligned} \tag{5.138}$$

The last equation holds since $\mathfrak{J}_{U_1,2}^{(1)\top} U_{1\perp} = 0$. Therefore, we only need to bound $\operatorname{tr}\left(\mathfrak{J}_{U_1,1}^{(1)\top}\mathcal{P}_{U_1}^{\perp}\mathfrak{J}_{U_1,1}^{(1)}\right)$, $\operatorname{tr}\left(\mathfrak{J}_{U_1,1}^{(1)\top}\mathcal{P}_{U_1}^{\perp}\mathfrak{J}_{U_1,3}^{(1)}\right)$ and $\operatorname{tr}\left(\mathfrak{J}_{U_1,3}^{(1)\top}\mathcal{P}_{U_1}^{\perp}\mathfrak{J}_{U_1,3}^{(1)}\right)$, respectively.

Step 3.1: bounding $\operatorname{tr}\left(\mathfrak{J}_{U_1,1}^{(1)\top}\mathcal{P}_{U_1}^{\perp}\mathfrak{J}_{U_1,1}^{(1)}\right)$**.** Recall the definitions of $\mathfrak{J}_{U_1,1}^{(1)}$ in *Step 1*. Similarly to (5.122) and by (5.126), we get with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$ that,

$$\begin{aligned}\left\|\mathfrak{J}_{U_1,1}^{(1)}\right\| \leqslant & C_2\left(\frac{\kappa_0 pr}{n\lambda_{\min}} + \kappa_0\frac{p^{3/2}r^2}{n^{3/2}\lambda_{\min}} + \kappa_0\frac{p^{3/2}r^{3/2}}{n^{3/2}\lambda_{\min}^2}\right) + C_2\kappa_0 r\sqrt{\frac{p}{n}}\cdot\sqrt{\frac{p}{n}}\lambda_{\min}^{-1} \\ \leqslant & C_2\kappa_0\frac{pr}{n\lambda_{\min}}. \end{aligned} \tag{5.139}$$

Therefore, with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\left|\operatorname{tr}\left(\mathfrak{J}_{U_1,1}^{(1)\top}\mathcal{P}_{U_1}^{\perp}\mathfrak{J}_{U_1,1}^{(1)}\right)\right| \leqslant r\left\|\mathfrak{J}_{U_1,1}^{(1)}\right\|^2 \leqslant C_2\kappa_0^2\frac{p^2r^2}{n^2\lambda_{\min}^2}. \tag{5.140}$$

Step 3.2: bounding $\operatorname{tr}\left(\mathfrak{J}_{U_1,1}^{(1)\top}\mathcal{P}_{U_1}^{\perp}\mathfrak{J}_{U_1,3}^{(1)}\right)$**.** Denote

$$K_{U_1}^{(1)} = \left(\mathcal{M}_1(\Delta\mathcal{T}_1^{(1.5)}) - \frac{1}{n}\sum_{i=1}^{n}\langle\Delta\mathcal{T}_1^{(1.5)}, \mathcal{X}_i\rangle\mathcal{M}_1(\mathcal{X}_i)\right)(U_2\otimes U_3)$$

and

$$L_1 = \frac{1}{n}\Big(\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)(U_2\otimes U_3).$$

Lemma 5.2.10 immediately implies that

$$\mathbb{P}\left(\|L_1\| \geqslant C_2\sqrt{\frac{p}{n}}\right) \leqslant 1-p^{-3}. \tag{5.141}$$

By (5.125), with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\left\|\mathfrak{J}^{(1)}_{U_1,3} - L_1G_1^\top(G_1G_1^\top)^{-1}R_1^{(1)\top}\right\| \leqslant C_2\kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}. \tag{5.142}$$

By (5.121), (5.123), (5.126) and Lemma 5.2.9, with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}
&\left\|\mathfrak{J}^{(1)}_{U_1,1} - K^{(1)}_{U_1}G_1^\top(G_1G_1^\top)^{-1}R_1^{(1)\top}\right\| \\
\leqslant&\left\|\Big(\mathcal{M}_1(\Delta\mathcal{T}_1^{(1.5)}) - \frac{1}{n}\sum_{i=1}^n\langle\Delta\mathcal{T}_1^{(1.5)},\mathcal{X}_i\rangle\mathcal{M}_1(\mathcal{X}_i)\Big)\Big(\big(\hat{U}_2^{(1)}R_2^{(1)} - U_2\big)\otimes U_3\Big)G_1^\top(G_1G_1^\top)^{-1}R_1^{(1)\top}\right\| \\
&+\left\|\Big(\mathcal{M}_1(\Delta\mathcal{T}_1^{(1.5)}) - \frac{1}{n}\sum_{i=1}^n\langle\Delta\mathcal{T}_1^{(1.5)},\mathcal{X}_i\rangle\mathcal{M}_1(\mathcal{X}_i)\Big)\Big(\big(\hat{U}_2^{(1)}R_2^{(1)}\big)\otimes\big(\hat{U}_3^{(1)}R_3^{(1)} - U_3\big)\Big)G_1^\top(G_1G_1^\top)^{-1}R_1^{(1)\top}\right\| \\
&+\left\|\big(\mathcal{M}_1(\Delta\mathcal{T}_1^{(1.5)}) - \frac{1}{n}\sum_{i=1}^n\langle\Delta\mathcal{T}_1^{(1.5)},\mathcal{X}_i\rangle\mathcal{M}_1(\mathcal{X}_i)\big)\big(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)}\big)\right\| \\
&\qquad\qquad\times\left\|\big(R_2^{(1)}\otimes R_3^{(1)}\big)G_1^\top(G_1G_1^\top)^{-1}R_1^{(1)\top} - \hat{G}_1^\top(\hat{G}_1\hat{G}_1^\top)^{-1}\right\| \\
\leqslant&C_2\sqrt{\frac{pr}{n}}\left\|\mathcal{M}_1(\Delta\mathcal{T}_1^{(1.5)})\right\|_F\left(\left\|\hat{U}_2^{(1)}R_2^{(1)} - U_2\right\| + \left\|\hat{U}_3^{(1)}R_3^{(1)} - U_3\right\|\right)\lambda_{\min}^{-1} \\
&+C_2\sqrt{\frac{pr}{n}}\left\|\mathcal{M}_1(\Delta\mathcal{T}_1^{(1.5)})\right\|_F\kappa_0^2\lambda_{\min}^{-2}\left(\sqrt{\frac{r^2+\log(p)}{n}} + \kappa_0\frac{pr}{n} + \kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right) \\
\leqslant&C_2\sqrt{\frac{pr}{n}}\left\|\mathcal{M}_1(\Delta\mathcal{T}_1^{(1.5)})\right\|_F\kappa_0^2\sqrt{\frac{p}{n}}\lambda_{\min}^{-2} \\
\leqslant&C_2\kappa_0^2\frac{p\sqrt{r}}{n\lambda_{\min}^2}C_2\left(\kappa_0\sqrt{\frac{pr}{n}} + \kappa_0\frac{pr^{3/2}}{n} + \kappa_0\frac{pr}{n\lambda_{\min}}\right) \\
\leqslant&C_2\kappa_0^3\frac{p^{3/2}r}{n^{3/2}\lambda_{\min}^2}.
\end{aligned}$$

(5.139), (5.140), (5.141), (5.142) and the previous inequality together imply with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\left|\operatorname{tr}\left(\mathfrak{J}^{(1)\top}_{U_1,1}\mathcal{P}^\perp_{U_1}\mathfrak{J}^{(1)}_{U_1,3}\right)\right|$$

$$
\begin{aligned}
&\leqslant \left| \operatorname{tr}\left(\left(\mathfrak{J}_{U_1,1}^{(1)} - K_{U_1}^{(1)} G_1^\top (G_1 G_1^\top)^{-1} R_1^{(1)\top}\right)^\top \mathcal{P}_{U_1}^\perp \mathfrak{J}_{U_1,3}^{(1)} \right) \right| \\
&\quad + \left| \operatorname{tr}\left(\mathfrak{J}_{U_1,1}^{(1)\top} \mathcal{P}_{U_1}^\perp \left(\mathfrak{J}_{U_1,3}^{(1)} - L_1 G_1^\top (G_1 G_1^\top)^{-1} R_1^{(1)\top}\right) \right) \right| \\
&\quad + \left| \operatorname{tr}\left(R_1^{(1)} (G_1 G_1^\top)^{-1} G_1 K_{U_1}^{(1)\top} \mathcal{P}_{U_1}^\perp L_1 G_1^\top (G_1 G_1^\top)^{-1} R_1^{(1)\top} \right) \right| \\
&\leqslant r \left\| \mathfrak{J}_{U_1,1}^{(1)} - K_{U_1}^{(1)} G_1^\top (G_1 G_1^\top)^{-1} R_1^{(1)\top} \right\| \left\| \mathfrak{J}_{U_1,3}^{(1)} \right\| + r \left\| \mathfrak{J}_{U_1,1}^{(1)} \right\| \left\| \mathfrak{J}_{U_1,3}^{(1)} - L_1 G_1^\top (G_1 G_1^\top)^{-1} R_1^{(1)\top} \right\| \\
&\quad + \left| \operatorname{tr}\left(G_1^\top (G_1 G_1^\top)^{-2} G_1 K_{U_1}^{(1)\top} \mathcal{P}_{U_1}^\perp L_1 \right) \right| \\
&\leqslant C_2 r \cdot \kappa_0^3 \frac{p^{3/2} r}{n^{3/2} \lambda_{\min}^2} \cdot \frac{\sqrt{p/n}}{\lambda_{\min}} + C_2 r \cdot \kappa_0 \frac{pr}{n\lambda_{\min}} \cdot \kappa_0^2 \frac{p\sqrt{r}}{n\lambda_{\min}^2} + \left| \operatorname{tr}\left(G_1^\top (G_1 G_1^\top)^{-2} G_1 K_{U_1}^{(1)\top} \mathcal{P}_{U_1}^\perp L_1 \right) \right| \\
&\leqslant C_2 \kappa_0^3 \frac{p^2 r^{5/2}}{n^2 \lambda_{\min}^3} + \left| \operatorname{tr}\left(G_1^\top (G_1 G_1^\top)^{-2} G_1 K_{U_1}^{(1)\top} \mathcal{P}_{U_1}^\perp L_1 \right) \right|. \qquad (5.143)
\end{aligned}
$$

By (5.108) and Lemma 5.2.10, with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$
\begin{aligned}
&\left\| \mathcal{M}_1\left(\Delta \mathcal{T}_1^{(1.5)}\right) - \left((\hat{U}_1^{(1.5)} R_1^{(1)}) G_1 \left((\hat{U}_2^{(1)} R_1^{(1)})^\top \otimes (\hat{U}_3^{(1)} R_1^{(1)})^\top \right) - U_1 G_1 (U_2 \otimes U_3) \right) \right\|_F \\
= &\left\| \hat{U}_1^{(1.5)} \left(\hat{G}_1^{(1)} - R_1^{(1)} G_1 (R_2^{(1)} \otimes R_3^{(1)}) \right) (\hat{U}_2^{(1)\top} \otimes \hat{U}_3^{(1)\top}) \right\|_F \\
\leqslant &C_2 \left(\kappa_0 \frac{pr^{3/2}}{n} + \kappa_0 \frac{pr}{n\lambda_{\min}} + \sqrt{\frac{r^3 + \log(p)}{n}} \right). \qquad (5.144)
\end{aligned}
$$

Define

$$
\Delta T_{U_1,1}^{(1.5)} = (\hat{U}_1^{(1.5)} R_1^{(1)} - U_1) G_1 (U_2^\top \otimes U_3^\top), \quad \Delta T_{U_2,1}^{(1)} = U_1 G_1 \left((\hat{U}_2^{(1)} R_2^{(1)} - U_2)^\top \otimes U_3^\top \right),
$$
$$
\Delta T_{U_3,1}^{(1)} = U_1 G_1 \left(U_2^\top \otimes (\hat{U}_3^{(1)} R_3^{(1)} - U_3)^\top \right).
$$

By (5.119), (5.131), (5.132) and Lemma 5.2.10, with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$, we get

$$
\max\left\{ \|U_1 - \hat{U}_1^{(1.5)} R_1^{(1)}\|, \|U_2 - \hat{U}_2^{(1)} R_2^{(1)}\|, \|U_3 - \hat{U}_3^{(1)} R_3^{(1)}\| \right\} \leqslant C_2 \frac{\sqrt{p/n}}{\lambda_{\min}}.
$$

Therefore, with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}&\left\|(\hat{U}_1^{(1.5)}R_1^{(1)})G_1\left((\hat{U}_2^{(1)}R_1^{(1)})^\top \otimes (\hat{U}_3^{(1)}R_1^{(1)})^\top\right) - U_1G_1(U_2\otimes U_3) - \left(\Delta T_{U_1,1}^{(1.5)} + \Delta T_{U_2,1}^{(1)} + \Delta T_{U_3,1}^{(1)}\right)\right\|_F\\ \leqslant& C_2\sqrt{r}\Big(\frac{\sqrt{p/n}}{\lambda_{\min}}\Big)^2\cdot\kappa_0\lambda_{\min} = C_2\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}.\end{aligned}$$

Combining (5.144) and the above inequality, with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\left\|\mathcal{M}_1(\Delta\mathcal{T}_1^{(1.5)}) - \left(\Delta T_{U_1,1}^{(1.5)} + \Delta T_{U_2,1}^{(1)} + \Delta T_{U_3,1}^{(1)}\right)\right\|_F \leqslant C_2\Big(\kappa_0\frac{pr^{3/2}}{n} + \kappa_0\frac{pr}{n\lambda_{\min}} + \sqrt{\frac{r^3+\log(p)}{n}}\Big). \tag{5.145}$$

Define

$$\begin{aligned}S_1^{(1.5)} =& \operatorname{tr}\left(G_1^\top(G_1G_1^\top)^{-2}G_1\left\{\left[\Delta T_{U_1,1}^{(1.5)} - \frac{1}{n}\sum_{i=1}^n\langle\Delta T_{U_1,1}^{(1.5)}, \mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\right](U_2\otimes U_3)\right\}^\top\mathcal{P}_{U_1}^\perp L_1\right),\\ S_2^{(1)} =& \operatorname{tr}\left(G_1^\top(G_1G_1^\top)^{-2}G_1\left\{\left[\Delta T_{U_2,1}^{(1)} - \frac{1}{n}\sum_{i=1}^n\langle\Delta T_{U_2,1}^{(1)}, \mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\right](U_2\otimes U_3)\right\}^\top\mathcal{P}_{U_1}^\perp L_1\right),\\ S_3^{(1)} =& \operatorname{tr}\left(G_1^\top(G_1G_1^\top)^{-2}G_1\left\{\left[\Delta T_{U_3,1}^{(1)} - \frac{1}{n}\sum_{i=1}^n\langle\Delta T_{U_3,1}^{(1)}, \mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\right](U_2\otimes U_3)\right\}^\top\mathcal{P}_{U_1}^\perp L_1\right).\end{aligned}$$

By (5.145), (5.141) and Lemma 5.2.9, with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}&\left|\operatorname{tr}\left(G_1^\top(G_1G_1^\top)^{-2}G_1K_{U_1}^{(1)\top}\mathcal{P}_{U_1}^\perp L_1\right) - S_1^{(1.5)} - S_2^{(1)} - S_3^{(1)}\right|\\ \leqslant& C_2r\left\|G_1^\top(G_1G_1^\top)^{-2}G_1\right\|\|L_1\|\left(\sqrt{\frac{pr}{n}}\left\|\mathcal{M}_1\left(\Delta\mathcal{T}_1^{(1.5)}\right) - \left(\Delta T_{U_1,1}^{(1.5)} + \Delta T_{U_2,1}^{(1)} + \Delta T_{U_3,1}^{(1)}\right)\right\|_F\right)\\ \leqslant& C_2r\lambda_{\min}^{-2}\sqrt{\frac{p}{n}}\cdot\sqrt{\frac{pr}{n}}\Big(\kappa_0\frac{pr^{3/2}}{n} + \kappa_0\frac{pr}{n\lambda_{\min}} + \sqrt{\frac{r^3+\log(p)}{n}}\Big)\\ \leqslant& C_2\bigg(\kappa_0\frac{p^2r^3}{n^2\lambda_{\min}^2} + \kappa_0\frac{p^2r^{5/2}}{n^2\lambda_{\min}^3} + \frac{p\big(r^3 + r^{3/2}\sqrt{\log p}\big)}{n^{3/2}\lambda_{\min}^2}\bigg).\end{aligned} \tag{5.146}$$

Therefore, it suffices to bound $\left|S_1^{(1.5)}\right|$, $\left|S_2^{(1)}\right|$, and $\left|S_3^{(1)}\right|$, respectively.

-Step 3.2.1: bounding $\left|S_1^{(1.5)}\right|$. We consider $\left|S_1^{(1.5)}\right|$ first. The proof of this part is involved and highly non-trivial, and some decoupling techniques (e.g., De la Pena and Giné (2012)) are needed. Let

$$\mathfrak{E}_{U_1,i}^{(1)} = \mathfrak{J}_{U_1,i}^{(1)} R_1^{(1)} G_1 (U_2^\top \otimes U_3^\top), \quad \forall i \in [3].$$

By (5.127),

$$\begin{aligned}
&S_1^{(1.5)}\\
=&\operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2} G_1 \left\{\left[\mathfrak{E}_{U_1,1}^{(1)} - \frac{1}{n}\sum_{i=1}^n \langle \mathfrak{E}_{U_1,1}^{(1)}, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i)\right](U_2\otimes U_3)\right\}^\top \mathcal{P}_{U_1}^\perp L_1\right)\\
&+\operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2} G_1 \left\{\left[\mathfrak{E}_{U_1,2}^{(1)} - \frac{1}{n}\sum_{i=1}^n \langle \mathfrak{E}_{U_1,2}^{(1)}, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i)\right](U_2\otimes U_3)\right\}^\top \mathcal{P}_{U_1}^\perp L_1\right)\\
&+\operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2} G_1 \left\{\left[\mathfrak{E}_{U_1,3}^{(1)} - \frac{1}{n}\sum_{i=1}^n \langle \mathfrak{E}_{U_1,3}^{(1)}, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i)\right](U_2\otimes U_3)\right\}^\top \mathcal{P}_{U_1}^\perp L_1\right).
\end{aligned} \tag{5.147}$$

By (5.139), (5.141) and Lemma 5.2.9, we have

$$\begin{aligned}
&\left|\operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2} G_1 \left\{\left[\mathfrak{E}_{U_1,1}^{(1)} - \frac{1}{n}\sum_{i=1}^n \langle \mathfrak{E}_{U_1,1}^{(1)}, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i)\right](U_2\otimes U_3)\right\}^\top \mathcal{P}_{U_1}^\perp L_1\right)\right|\\
\leqslant& r\left\|G_1^\top(G_1G_1^\top)^{-2}G_1\right\|\|L_1\|\left(\sqrt{\frac{pr}{n}}\left\|\mathfrak{E}_{U_1,1}^{(1)}\right\|_F\right)\\
\leqslant& C_2 r \lambda_{\min}^{-2}\sqrt{\frac{p}{n}}\cdot\sqrt{\frac{pr}{n}}\sqrt{r}\left\|\mathfrak{J}_{U_1,1}^{(1)}\right\|\|G_1\| \leqslant C_2\kappa_0^2\frac{p^2r^3}{n^2\lambda_{\min}^2}.
\end{aligned} \tag{5.148}$$

In addition, by (5.108), (5.131) and Lemma 5.2.10, with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}
\left\|\mathfrak{J}_{U_1,2}^{(1)}\right\|_F =& \left\|\left(U_1G_1\left((\hat{U}_2^{(1)\top}U_2)^\top \otimes (\hat{U}_3^{(1)\top}U_3)^\top\right) - U_1R_1^{(1)\top}\hat{G}_1^{(1)}\right)\hat{G}_1^{(1)\top}\left(\hat{G}_1^{(1)}\hat{G}_1^{(1)\top}\right)^{-1}\right\|_F\\
\leqslant& C_2\lambda_{\min}^{-1}\left\|G_1\left((\hat{U}_2^{(1)\top}U_2)^\top \otimes (\hat{U}_3^{(1)\top}U_3)^\top\right) - R_1^{(1)\top}\hat{G}_1^{(1)}\right\|_F
\end{aligned}$$

$$\begin{aligned}
\leqslant & C_2\lambda_{\min}^{-1}\left\|G_1\left((\hat{U}_2^{(1)\top}U_2)^\top \otimes (\hat{U}_3^{(1)\top}U_3)^\top\right) - G_1\left(R_2^{(1)\top}\otimes R_3^{(1)\top}\right)\right\|_F \\
& + C_2\lambda_{\min}^{-1}\left\|R_1^{(1)\top}\left(\hat{G}_1^{(1)} - R_1^{(1)}G_1\left(R_2^{(1)\top}\otimes R_3^{(1)\top}\right)\right)\right\|_F \\
\leqslant & C_2\lambda_{\min}^{-1}\cdot\kappa_0\lambda_{\min}\frac{p\sqrt{r}}{n\lambda_{\min}^2} + C_2\lambda_{\min}^{-1}\Big(\kappa_0\frac{pr^{3/2}}{n} + \kappa_0\frac{pr}{n\lambda_{\min}} + \sqrt{\frac{r^3+\log p}{n}}\Big) \\
= & C_2\Big(\kappa_0\frac{pr}{n\lambda_{\min}^2} + \kappa_0\frac{pr^{3/2}}{n\lambda_{\min}} + \sqrt{\frac{r^3+\log p}{n\lambda_{\min}^2}}\Big).
\end{aligned}$$

Therefore, (5.141), Lemma 5.2.9 and the previous inequality together imply that with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}
& \left|\operatorname{tr}\left(G_1^\top(G_1G_1^\top)^{-2}G_1\Big\{\Big[\mathfrak{E}_{U_1,2}^{(1)} - \frac{1}{n}\sum_{i=1}^n\langle\mathfrak{E}_{U_1,2}^{(1)},\mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\Big](U_2\otimes U_3)\Big\}^\top\mathcal{P}_{U_1}^\perp L_1\right)\right| \\
\leqslant & r\left\|G_1^\top(G_1G_1^\top)^{-2}G_1\right\|\,\|L_1\|\Big(\sqrt{\frac{pr}{n}}\|\mathfrak{E}_{U_1,2}^{(1)}\|_F\Big) \\
\leqslant & C_2r\lambda_{\min}^{-2}\sqrt{\frac{p}{n}}\cdot\sqrt{\frac{pr}{n}}\|\mathfrak{J}_{U_1,2}^{(1)}\|_F\|G_1\| \\
\leqslant & C_2\left(\kappa_0^2\frac{p^2r^{5/2}}{n^2\lambda_{\min}^3} + \kappa_0^2\frac{p^2r^3}{n^2\lambda_{\min}^2} + \kappa_0\frac{r^3p + r^{3/2}p\sqrt{\log(p)}}{n^{3/2}\lambda_{\min}^2}\right). \qquad (5.149)
\end{aligned}$$

Let

$$J_1 = L_1G_1^\top(G_1G_1^\top)^{-1}G_1(U_2^\top\otimes U_3^\top).$$

Consider the SVD decomposition $G_1 = U_{G_1}\Lambda_1V_{G_1}^\top$, where $U_{G_1}\in\mathbb{O}_{r_1}$, $V_{G_1}\in\mathbb{O}_{r_2r_3,r_1}$ and $\Lambda_1\in\mathbb{R}^{r_1\times r_1}$ is a diagonal matrix containing all singular values of G_1. Then

$$G_1^\top(G_1G_1^\top)^{-1}G_1 = V_{G_1}V_{G_1}^\top.$$

By (5.142), with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\left|\operatorname{tr}\left(G_1^\top(G_1G_1^\top)^{-2}G_1\Big\{\Big[\mathfrak{E}_{U_1,3}^{(1)} - \frac{1}{n}\sum_{i=1}^n\langle\mathfrak{E}_{U_1,3}^{(1)},\mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\Big](U_2\otimes U_3)\Big\}^\top\mathcal{P}_{U_1}^\perp L_1\right)\right|$$

$$
\begin{aligned}
&\leqslant \left| \operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2} G_1 \left\{ \left[J_1 - \frac{1}{n}\sum_{i=1}^n \langle J_1, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i) \right] (U_2 \otimes U_3) \right\}^\top \mathcal{P}_{U_1}^\perp L_1 \right) \right| \\
&\quad + r \left\| G_1^\top (G_1G_1^\top)^{-2} G_1 \right\| \|L_1\| \left(\sqrt{\frac{pr}{n}} \left\| \mathfrak{E}_{U_1,3}^{(1)} - J_1 \right\|_F \right) \\
&\leqslant \left| \operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2} G_1 \left\{ \left[J_1 - \frac{1}{n}\sum_{i=1}^n \langle J_1, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i) \right] (U_2 \otimes U_3) \right\}^\top \mathcal{P}_{U_1}^\perp L_1 \right) \right| \\
&\quad + C_2 r \lambda_{\min}^{-2} \sqrt{\frac{p}{n}} \cdot \sqrt{\frac{pr}{n}} \sqrt{r} \left\| \mathfrak{I}_{U_1,3}^{(1)} - L_1 G_1^\top (G_1G_1^\top)^{-1} R_1^{(1)\top} \right\| \|G_1\| \\
&\leqslant \left| \operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2} G_1 \left\{ \left[J_1 - \frac{1}{n}\sum_{i=1}^n \langle J_1, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i) \right] (U_2 \otimes U_3) \right\}^\top \mathcal{P}_{U_1}^\perp L_1 \right) \right| \\
&\quad + C_2 r \lambda_{\min}^{-2} \sqrt{\frac{p}{n}} \cdot \sqrt{\frac{pr}{n}} \cdot \sqrt{r} \kappa_0^2 \frac{p\sqrt{r}}{n\lambda_{\min}^2} \cdot \kappa_0 \lambda_{\min} \\
&\leqslant \left| \operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2} G_1 \left\{ \left[J_1 - \frac{1}{n}\sum_{i=1}^n \langle J_1, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i) \right] (U_2 \otimes U_3) \right\}^\top L_1 \right) \right| \\
&\quad + \left| \operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2} G_1 \left\{ \left[J_1 - \frac{1}{n}\sum_{i=1}^n \langle J_1, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i) \right] (U_2 \otimes U_3) \right\}^\top \mathcal{P}_{U_1} L_1 \right) \right| \\
&\quad + C_2 \kappa_0^3 \frac{p^2 r^{5/2}}{n^2 \lambda_{\min}^3}. \qquad (5.150)
\end{aligned}
$$

For $i \in [n]$, let

$$Z_i = \mathcal{M}_1(\mathcal{X}_i)(U_2 \otimes U_3) V_{G_1} \in \mathbb{R}^{p_1 \times r_1},$$

and

$$Z_\xi = L_1 V_{G_1} = \sum_{i=1}^n \xi_i \mathcal{M}_1(\mathcal{X}_i)(U_2 \otimes U_3) V_{G_1} \in \mathbb{R}^{p_1 \times r_1}.$$

Then

$$Z_i \overset{\text{i.i.d.}}{\sim} N(0,1) \quad \text{and} \quad Z_\xi = \sum_{j=1}^n \xi_j Z_j.$$

Thus

$$\left| \operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2} G_1 \left\{ \left[J_1 - \frac{1}{n}\sum_{i=1}^n \langle J_1, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i) \right] (U_2 \otimes U_3) \right\}^\top L_1 \right) \right|$$

$$
\begin{aligned}
=&\left|\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\xi_j\xi_k\langle Z_i, Z_j\rangle\langle Z_i\Lambda_1^{-1}, Z_k\Lambda_1^{-1}\rangle-\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\xi_j\xi_k\langle Z_j\Lambda_1^{-1}, Z_k\Lambda_1^{-1}\rangle\right|\\
\leqslant&\left|\frac{1}{n^3}\sum_{i=1}^{n}\xi_i^2\|Z_i\|_F^2\|Z_i\Lambda_1^{-1}\|_F^2-\frac{1}{n^3}\sum_{i=1}^{n}\xi_i^2\|Z_i\Lambda_1^{-1}\|_F^2\right|\\
&+\left|\frac{1}{n^3}\sum_{i=1}^{n}\left[\sum_{j\neq i}\xi_j^2\left(\langle Z_i, Z_j\rangle\langle Z_i, Z_j\Lambda_1^{-2}\rangle-\langle Z_j, Z_j\Lambda_1^{-2}\rangle\right)\right]\right|\\
&+\left|\frac{1}{n^3}\sum_{i=1}^{n}\xi_i\|Z_i\|_F^2\sum_{k\neq i}\xi_k\langle Z_i, Z_k\Lambda_1^{-2}\rangle\right|+\left|\frac{1}{n^3}\sum_{i=1}^{n}\xi_i\,\|Z_i\Lambda_1^{-1}\|_F^2\sum_{j\neq i}\xi_j\langle Z_i, Z_j\rangle\right|\\
&+\left|\frac{2}{n^3}\sum_{i=1}^{n}\xi_i\sum_{k\neq i}\xi_k\langle Z_i, Z_k\Lambda_1^{-2}\rangle\right|\\
&+\left|\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq k\neq i}\xi_j\xi_k\left[\langle Z_j, Z_i\rangle\langle Z_i, Z_k\Lambda_1^{-2}\rangle-\langle Z_j, Z_k\Lambda_1^{-2}\rangle\right]\right|. \qquad (5.151)
\end{aligned}
$$

By (Vershynin, 2010, Corollary 5.35), for any $i \in [n]$, with probability at least $1 - e^{-c_1(p+\log(n))}$,

$$\|Z_i\| \leqslant C_2\sqrt{p+\log(n)} \quad \text{and} \quad \|Z_i\|_F \leqslant C_2\sqrt{r(p+\log(n))} \qquad (5.152)$$

and

$$\|Z_i\Lambda_1^{-1}\|_F \leqslant \|Z_i\|_F\|\Lambda_1^{-1}\| \leqslant C_2\frac{\sqrt{r(p+\log(n))}}{\lambda_{\min}}. \qquad (5.153)$$

By the union bound and Bernstein-type inequality, with probability at least $1 - e^{-C_1(pr+\log(n))}$,

$$
\begin{aligned}
\frac{1}{n^3}\sum_{i=1}^{n}\xi_i^2\|Z_i\|_F^2\|Z_i\Lambda_1^{-1}\|_F^2 \leqslant \frac{1}{n^3}\sum_{i=1}^{n}\xi_i^2\cdot C_2\frac{r^2(p+\log(n))^2}{\lambda_{\min}^2} \leqslant& C_2\frac{r^2(p+\log(n))^2}{n^3\lambda_{\min}^2}\cdot Cn\\
\leqslant& C_2\frac{p^2r^2+r^2\log^2(n)}{n^2\lambda_{\min}^2}
\end{aligned}
$$

and

$$\frac{1}{n^3}\sum_{i=1}^{n}\xi_i^2\|Z_i\Lambda_1^{-1}\|_F^2 \leqslant \frac{1}{n^3}\sum_{i=1}^{n}\xi_i^2\cdot\frac{r(p+\log n)}{\lambda_{\min}} \leqslant C_2\frac{r(p+\log n)}{n^2\lambda_{\min}^2}. \tag{5.154}$$

Therefore, with probability at least $1-e^{-c_1pr}$,

$$\left|\frac{1}{n^3}\sum_{i=1}^{n}\xi_i^2\|Z_i\|_F^2\|Z_i\Lambda_1^{-1}\|_F^2 - \frac{1}{n^3}\sum_{i=1}^{n}\xi_i^2\|Z_i\Lambda_1^{-1}\|_F^2\right| \leqslant C_2\frac{p^2r^2+r^2\log^2 n}{n^2\lambda_{\min}^2}. \tag{5.155}$$

Since Z_i and Z_j are independent for all $1\leqslant i\neq j\leqslant n$, we have $\langle Z_i, Z_j\rangle\big|Z_i \sim N(0,\|Z_i\|_F^2)$ and $\langle Z_i, Z_j\Lambda_1^2\rangle\big|Z_i \sim N(0,\|Z_i\Lambda_1^{-2}\|_F^2)$, which imply that

$$\|\langle Z_i, Z_j\rangle\|_{\psi_2}\Big|Z_i \leqslant C\|Z_i\|_F \quad\text{and}\quad \|\langle Z_i, Z_j\Lambda_1^2\rangle\|_{\psi_2}\Big|Z_i \leqslant C\|Z_i\Lambda_1^{-2}\|_F.$$

Since $\mathbb{E}\left[\langle Z_i, Z_j\rangle\langle Z_i, Z_j\Lambda_1^{-2}\rangle\right]\big|Z_i = \langle Z_i, Z_i\Lambda_1^{-2}\rangle$, by (Vershynin, 2010, Remark 5.18) and (5.229),

$$\begin{aligned}\left\|\langle Z_i, Z_j\rangle\langle Z_i, Z_j\Lambda_1^{-2}\rangle - \langle Z_j, Z_j\Lambda_1^{-2}\rangle\right\|_{\psi_1}\big|Z_i &\leqslant C\left\|\langle Z_i, Z_j\rangle\langle Z_i, Z_j\Lambda_1^{-2}\rangle\right\|_{\psi_1}\big|Z_i\\ &\leqslant C\|\langle Z_i, Z_j\rangle\|_{\psi_2}\|\langle Z_i, Z_j\Lambda_1^2\rangle\|_{\psi_2}\big|Z_i \leqslant C\|Z_i\|_F\|Z_i\Lambda_1^{-2}\|_F. \end{aligned}\tag{5.156}$$

By Bernstein-type inequality, we have

$$\begin{aligned}\mathbb{P}\Bigg(&\Big|\sum_{j\neq i}\xi_j^2\left(\langle Z_i, Z_j\rangle\langle Z_i, Z_j\Lambda_1^{-2}\rangle - \langle Z_j, Z_j\Lambda_1^{-2}\rangle\right)\Big|\\ &\geqslant C_2\|Z_i\|_F\|Z_i\Lambda_1^{-2}\|_F\Big(\sum_{j\neq i}\xi_j^4\Big)^{1/2}\log(n)\big|Z_i,\xi_1,\ldots,\xi_n\Bigg) \leqslant n^{-3}.\end{aligned}$$

The union bound and (5.152) together imply that

$$\mathbb{P}\Bigg(\Big|\frac{1}{n^3}\sum_{i=1}^{n}\Big[\sum_{j\neq i}\xi_j^2\left(\langle Z_i, Z_j\rangle\langle Z_i, Z_j\Lambda_1^{-2}\rangle - \langle Z_j, Z_j\Lambda_1^{-2}\rangle\right)\Big]\Big|$$

$$\geqslant C_2 \frac{r(p+\log(n))\log(n)}{n^2\lambda_{\min}^2} \big(\sum_{j=1}^{n} \xi_j^4 \big)^{1/2} \Big| \xi_1, \dots, \xi_n \Big) \leqslant n^{-3}.$$

Notice that

$$\mathbb{E}\xi_i^4 \leqslant \left(2 \sup_{q \geqslant 1} q^{-1/2} \left(\mathbb{E} \left| \xi_i \right|^q \right)^{1/q} \right)^4 \leqslant C \|\xi_i\|_{\psi_2}^4 \leqslant C,$$

by (Hao et al., 2020, Lemmas 7 and 8),

$$\mathbb{P}\Bigg(\sum_{j=1}^{n} \xi_j^4 - Cn \geqslant C_2 \left(\sqrt{n \log(p)} + \log^2(p) \right) \Bigg) \leqslant p^{-3}. \tag{5.157}$$

By combining the above two inequalities together, we know that with probability at least $1 - C_1 p^{-3}$,

$$\left| \frac{1}{n^3} \sum_{i=1}^{n} \left[\sum_{j \neq i} \xi_j^2 \left(\langle Z_i, Z_j \rangle \langle Z_i, Z_j \Lambda_1^{-2} \rangle - \langle Z_j, Z_j \Lambda_1^{-2} \rangle \right) \right] \right| \leqslant C_2 \frac{r(p+\log(n))\log(n)}{n^{3/2}\lambda_{\min}^2}. \tag{5.158}$$

Note that

$$\sum_{i=1}^{n} \xi_i \|Z_i\|_F^2 \sum_{k \neq i} \xi_k \langle Z_i, Z_k \Lambda_1^{-2} \rangle = \sum_{i=1}^{n} \|Z_i\|_F^2 \langle Z_i, \xi_i \sum_{k \neq i} \xi_k Z_k \Lambda_1^{-2} \rangle$$

By (De la Pena and Giné, 2012, Theorem 3.4.1), there exists a constant $C > 0$, for any $t > 0$, we have

$$\mathbb{P}\Bigg(\Big| \sum_{i=1}^{n} \xi_i \|Z_i\|_F^2 \sum_{k \neq i} \xi_k \langle Z_i, Z_k \Lambda_1^{-2} \rangle \Big| \geqslant t \Bigg) \leqslant C\mathbb{P}\Bigg(\Big| \sum_{i=1}^{n} \|Z_i^{(1)}\|_F^2 \langle Z_i^{(1)}, \xi_i \sum_{k \neq i} \xi_k Z_k^{(2)} \Lambda_1^{-2} \rangle \Big| \geqslant t/C \Bigg)$$

$$\leqslant C\mathbb{P}\Bigg(\Big| \sum_{i=1}^{n} \|Z_i^{(1)}\|_F^2 \langle Z_i^{(1)}, \xi_i \sum_{k=1}^{n} \xi_k Z_k^{(2)} \Lambda_1^{-2} \rangle \Big| \geqslant \frac{t}{2C} \Bigg) + C\mathbb{P}\Bigg(\Big| \sum_{i=1}^{n} \|Z_i^{(1)}\|_F^2 \langle Z_i^{(1)}, \xi_i^2 Z_i^{(2)} \Lambda_1^{-2} \rangle \Big| \geqslant \frac{t}{2C} \Bigg), \tag{5.159}$$

where $\{Z_1^{(1)}, \dots, Z_n^{(1)}\}$ and $\{Z_1^{(2)}, \dots, Z_n^{(2)}\}$ are two independent copies of $\{Z_1, \dots, Z_n\}$. By Lemma 3.3.3, with probability at least $1 - p^{-3}$,

$$\begin{aligned}&\Big|\sum_{i=1}^n \|Z_i^{(1)}\|_F^2 \langle Z_i^{(1)}, \xi_i \sum_{k=1}^n \xi_k Z_k^{(2)} \Lambda_1^{-2}\rangle\Big|\,\Big| \Big\{\xi_k, Z_k^{(2)}\Big\}_{k=1}^n \\ \leqslant & Cpr \Big\| \sum_{k=1}^n \xi_k Z_k^{(2)} \Lambda_1^{-2} \Big\|_F \|\vec{\xi}\|_2 \sqrt{\log(p)} \leqslant Cpr \Big\| \sum_{k=1}^n \xi_k Z_k \Big\|_F \lambda_{\min}^{-2} \|\vec{\xi}\|_2 \sqrt{\log(p)}.\end{aligned}$$

By Lemma 5.2.10, we have

$$\mathbb{P}\Big(\Big\| \sum_{k=1}^n \xi_k Z_k \Big\|_F \geqslant C_2 \sqrt{npr} \Big) \leqslant e^{-C_1 pr},$$

The previous two inequalities and (5.238) together imply that

$$\mathbb{P}\Big(\Big| \sum_{i=1}^n \|Z_i^{(1)}\|_F^2 \langle Z_i^{(1)}, \xi_i \sum_{k=1}^n \xi_k Z_k^{(2)} \Lambda_1^{-2} \rangle \Big| \geqslant C_2 n p^{3/2} r^{3/2} \sqrt{\log(p)} \lambda_{\min}^{-2} \Big) \leqslant p^{-3}. \tag{5.160}$$

Since

$$\sum_{i=1}^n \|Z_i^{(1)}\|_F^2 \langle Z_i^{(1)}, \xi_i^2 Z_i^{(2)} \Lambda_1^{-2} \rangle \Big| \Big\{\xi_k, Z_k^{(1)}\Big\}_{k=1}^n \sim N(0, \sum_{i=1}^n \xi_i^4 \|Z_i^{(1)}\|_F^4 \|Z_i^{(1)} \Lambda_1^{-2}\|_F^2),$$

we know that

$$\mathbb{P}\Big(\Big| \sum_{i=1}^n \|Z_i^{(1)}\|_F^2 \langle Z_i^{(1)}, \xi_i^2 Z_i^{(2)} \Lambda_1^{-2} \rangle \Big| \geqslant C \sqrt{\sum_{i=1}^n \xi_i^4 \|Z_i^{(1)}\|_F^4 \|Z_i^{(1)} \Lambda_1^{-2}\|_F^2 \log(p)} \Big| \Big\{\xi_k, Z_k^{(1)}\Big\}_{k=1}^n \Big) \leqslant p^{-3}.$$

By Cauchy-Schwarz inequality,

$$\sum_{i=1}^n \xi_i^4 \|Z_i^{(1)}\|_F^4 \|Z_i^{(1)} \Lambda_1^{-2}\|_F^2 \leqslant \sum_{i=1}^n \xi_i^4 \|Z_i^{(1)}\|_F^4 \|Z_i^{(1)} \Lambda_1^{-2}\|_F^2 \leqslant \sum_{i=1}^n \xi_i^4 \|Z_i^{(1)}\|_F^6 \|\Lambda_1^{-2}\|^2$$

$$\leqslant \lambda_{\min}^{-4}\Big(\sum_{i=1}^{n}\xi_i^8\Big)^{1/2}\Big(\sum_{i=1}^{n}\|Z_i^{(1)}\|_F^{12}\Big)^{1/2}.$$

Similarly to (5.157),

$$\mathbb{P}\left(\sum_{i=1}^{n}\xi_i^8 \geqslant Cn\right) \leqslant n^{-3}.$$

By (5.152) and the union bound, with probability at least $1-e^{-C_1(p+\log(n))}$,

$$\sum_{i=1}^{n}\|Z_i^{(1)}\|_F^{12} \leqslant C_2 n\left(\sqrt{r(p+\log(n))}\right)^{12} \leqslant Cnr^6\left(p+\log(n)\right)^6.$$

By combining the previous four inequalities together, we have

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n}\|Z_i^{(1)}\|_F^2\langle Z_i^{(1)}, \xi_i^2 Z_i^{(2)}\Lambda_1^{-2}\rangle\Big| \geqslant C_2\sqrt{nr^3\left(p+\log(n)\right)^3\log(p)}\lambda_{\min}^{-2}\Big) \leqslant C_1 p^{-3}.$$

By (5.159), (5.160) and the previous inequality,

$$\mathbb{P}\Big(\Big|\frac{1}{n^3}\sum_{i=1}^{n}\xi_i\|Z_i\|_F^2\sum_{k\neq i}\xi_k\langle Z_i, Z_k\Lambda_1^{-2}\rangle\Big| \geqslant C_2\frac{p^{3/2}r^{3/2}\sqrt{\log(p)}}{n^2\lambda_{\min}^2}\Big) \leqslant C_1p^{-3}. \quad (5.161)$$

Similarly, we have

$$\mathbb{P}\Big(\Big|\frac{1}{n^3}\sum_{i=1}^{n}\xi_i\|Z_i\Lambda_1^{-1}\|_F^2\sum_{k\neq i}\xi_k\langle Z_i, Z_k\rangle\Big| \geqslant C_2\frac{p^{3/2}r^{3/2}\sqrt{\log(p)}}{n^2\lambda_{\min}^2}\Big) \leqslant C_1p^{-3}. \quad (5.162)$$

By Lemma 5.2.10, with probability at least $1-e^{-C_1pr}$,

$$\Big|\sum_{i=1}^{n}\xi_i\sum_{k=1}^{n}\xi_k\langle Z_i, Z_k\Lambda_1^{-2}\rangle\Big| = \Big|\langle\sum_{i=1}^{n}\xi_i Z_i, \sum_{k=1}^{n}\xi_k Z_k\Lambda_1^{-2}\rangle\Big| \leqslant \Big\|\sum_{i=1}^{n}\xi_i Z_i\Big\|_F\Big\|\sum_{k=1}^{n}\xi_k Z_k\Lambda_1^{-2}\Big\|_F$$
$$\leqslant \Big\|\sum_{i=1}^{n}\xi_i Z_i\Big\|_F^2\|\Lambda_1^{-2}\| \leqslant C_2 npr\lambda_{\min}^{-2}.$$

The previous inequality and (5.154) together show that

$$\left|\frac{1}{n^3}\sum_{i=1}^{n}\xi_i\sum_{k\neq i}\xi_k\langle Z_i, Z_k\Lambda_1^{-2}\rangle\right| \leqslant C_2\frac{r(p+\log(n))}{n^2\lambda_{\min}^2}. \tag{5.163}$$

Now, we consider $\left|\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq k\neq i}\xi_j\xi_k\left[\langle Z_j, Z_i\rangle\langle Z_i, Z_k\Lambda_1^{-2}\rangle - \langle Z_j, Z_k\Lambda_1^{-2}\rangle\right]\right|$. By (De la Pena and Giné, 2012, Theorem 3.4.1), for any $t \geqslant 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq k\neq i}\xi_j\xi_k\left[\langle Z_j, Z_i\rangle\langle Z_i, Z_k\Lambda_1^{-2}\rangle - \langle Z_j, Z_k\Lambda_1^{-2}\rangle\right]\right| \geqslant t\right)\\
\leqslant &C\mathbb{P}\left(\left|\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq k\neq i}\xi_j\xi_k\left[\langle Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, Z_k^{(3)}\Lambda_1^{-2}\rangle - \langle Z_j^{(2)}, Z_k^{(3)}\Lambda_1^{-2}\rangle\right]\right| \geqslant \frac{t}{C}\right)\\
\leqslant &C\mathbb{P}\left(\left|\frac{1}{n^3}\sum_{i=1}^{n}\left[\langle\sum_{j\neq i}\xi_j Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle - \langle\sum_{j\neq i}\xi_j Z_j^{(2)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle\right]\right| \geqslant \frac{t}{2C}\right)\\
&+ C\mathbb{P}\left(\left|\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\xi_j^2\left[\langle Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, Z_j^{(3)}\Lambda_1^{-2}\rangle - \langle Z_j^{(2)}, Z_j^{(3)}\Lambda_1^{-2}\rangle\right]\right| \geqslant \frac{t}{2C}\right).
\end{aligned} \tag{5.164}$$

Here, $\{Z_i^{(1)}\}_{i=1}^n$, $\{Z_i^{(2)}\}_{i=1}^n$ and $\{Z_i^{(3)}\}_{i=1}^n$ are independent copies of $\{Z_i\}_{i=1}^n$. Conditioning on $\{\xi_i\}_{i=1}^n$, $\{Z_i^{(2)}\}_{i=1}^n$ and $\{Z_i^{(3)}\}_{i=1}^n$, we know that

$$\left\{\langle\sum_{j\neq i}\xi_j Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle - \langle\sum_{j\neq i}\xi_j Z_j^{(2)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle\right\}_{i=1}^{n}$$

are independent. In addition,

$$\left\langle \sum\nolimits_{j\neq i}\xi_j Z_j^{(2)}, Z_i^{(1)}\right\rangle \Big| \left\{\xi_i, Z_i^{(2)}, Z_i^{(3)}\right\}_{i=1}^{n} \sim N\big(0, \|\sum\nolimits_{j\neq i}\xi_j Z_j^{(2)}\|_F^2\big)$$

and

$$\left\langle Z_i^{(1)}, \sum\nolimits_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\right\rangle \Big| \left\{\xi_i, Z_i^{(2)}, Z_i^{(3)}\right\}_{i=1}^{n} \sim N\big(0, \|\sum\nolimits_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\|_F^2\big).$$

Note that

$$\mathbb{E}\Big(\langle\sum_{j\neq i}\xi_j Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle\Big|\Big\{\xi_i, Z_i^{(2)}, Z_i^{(3)}\Big\}_{i=1}^n\Big) = \Big\langle\sum_{j\neq i}\xi_j Z_j^{(2)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\Big\rangle$$

and

$$\begin{aligned}&\Big\|\langle\sum_{j\neq i}\xi_j Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle - \langle\sum_{j\neq i}\xi_j Z_j^{(2)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle\Big\|_{\psi_1}\Big|\Big\{\xi_i, Z_i^{(2)}, Z_i^{(3)}\Big\}_{i=1}^n\\ \leqslant& C\Big\|\langle\sum_{j\neq i}\xi_j Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle\Big\|_{\psi_1}\Big|\Big\{\xi_i, Z_i^{(2)}, Z_i^{(3)}\Big\}_{i=1}^n\\ \leqslant& C\Big\|\langle\sum_{j\neq i}\xi_j Z_j^{(2)}, Z_i^{(1)}\rangle\Big\|_{\psi_2}\Big\|\langle Z_i^{(1)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle\Big\|_{\psi_2}\Big|\Big\{\xi_i, Z_i^{(2)}, Z_i^{(3)}\Big\}_{i=1}^n\\ \leqslant& C\|\sum_{j\neq i}\xi_j Z_j^{(2)}\|_F\|\sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\|_F \leqslant C\|\sum_{j\neq i}\xi_j Z_j^{(2)}\|_F\|\sum_{k\neq i}\xi_k Z_k^{(3)}\|_F\lambda_{\min}^{-2}.\end{aligned}$$

By Bernstein-type inequality, for any $t \geqslant 0$,

$$\begin{aligned}&\mathbb{P}\Big(\Big|\sum_{i=1}^n\Big[\langle\sum_{j\neq i}\xi_j Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle - \langle\sum_{j\neq i}\xi_j Z_j^{(2)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle\Big]\Big|\\ &\qquad\qquad \geqslant \frac{t}{\lambda_{\min}^2}\Big|\Big\{\xi_i, Z_i^{(2)}, Z_i^{(3)}\Big\}_{i=1}^n\Big)\\ \leqslant& 2\exp\Big(-C_1\min\Big\{\frac{t^2}{\sum_{i=1}^n\|\sum_{j\neq i}\xi_j Z_j^{(2)}\|_F^2\|\sum_{k\neq i}\xi_k Z_k^{(3)}\|_F^2},\\ &\qquad\qquad \frac{t}{\max_{1\leqslant i\leqslant n}\|\sum_{j\neq i}\xi_j Z_j^{(2)}\|_F\|\sum_{k\neq i}\xi_k Z_k^{(3)}\|_F}\Big\}\Big).\end{aligned}$$

By Lemma 5.2.10, for any $i \in [n]$, with probability at least $1 - e^{-C_1(pr+\log(n))}$,

$$\max\Big\{\|\sum_{j\neq i}\xi_j Z_j^{(2)}\|_F, \|\sum_{k\neq i}\xi_k Z_k^{(3)}\|_F\Big\} \leqslant C_2\sqrt{n(pr+\log(n))}.$$

The union bound shows that with probability at least $1-e^{-C_1(pr+\log(n))}$,

$$\max\left\{\|\sum_{j\neq i}\xi_j Z_j^{(2)}\|_F, \|\sum_{k\neq i}\xi_k Z_k^{(3)}\|_F\right\} \leqslant C_2\sqrt{n(pr+\log(n))}, \quad \forall i\in[n].$$

Therefore, with probability at least $1-e^{-C_1(pr+\log(n))}$,

$$\sum_{i=1}^n \|\sum_{j\neq i}\xi_j Z_j^{(2)}\|_F^2 \|\sum_{k\neq i}\xi_k Z_k^{(3)}\|_F^2 \leqslant C_2 n^3\left(pr+\log(n)\right)^2,$$

and

$$\max_{1\leqslant i\leqslant n}\|\sum_{j\neq i}\xi_j Z_j^{(2)}\|_F\|\sum_{k\neq i}\xi_k Z_k^{(3)}\|_F \leqslant C_2 n(pr+\log(n)).$$

Thus

$$\begin{aligned}\mathbb{P}\Bigg(\Big|\sum_{i=1}^n\Big[\langle\sum_{j\neq i}\xi_j Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle - \langle\sum_{j\neq i}\xi_j Z_j^{(2)}, \sum_{k\neq i}\xi_k Z_k^{(3)}\Lambda_1^{-2}\rangle\Big]\Big| \\ \geqslant C_2\frac{n^{3/2}(pr+\log(n))\sqrt{\log(p)}}{\lambda_{\min}^2}\Bigg) \leqslant C_1 p^{-3}.\end{aligned} \tag{5.165}$$

Similarly to (5.156), for any $i\in[n]$, we have

$$\begin{aligned}\Big\|\langle Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, Z_j^{(3)}\Lambda_1^{-2}\rangle - \langle Z_j^{(2)}, Z_j^{(3)}\Lambda_1^{-2}\rangle\Big\|_{\psi_1}\Big| Z_i^{(1)}, \xi_j \\ \leqslant C\|Z_i^{(1)}\|_F\|Z_i^{(1)}\Lambda_1^{-2}\|_F \leqslant C\|Z_i^{(1)}\|_F^2\lambda_{\min}^{-2}.\end{aligned}$$

By Bernstein-type inequality,

$$\begin{aligned}\mathbb{P}\Bigg(\Big|\sum_{j\neq i}\xi_j^2\Big[\langle Z_j^{(2)}, Z_i^{(1)}\rangle\langle Z_i^{(1)}, Z_j^{(3)}\Lambda_1^{-2}\rangle - \langle Z_j^{(2)}, Z_j^{(3)}\Lambda_1^{-2}\rangle\Big]\Big| \\ \geqslant C_2\|Z_i^{(1)}\|_F^2\lambda_{\min}^{-2}\Big(\sum_{j\neq i}\xi_j^4\Big)^{1/2}\log(n)\Big|\{Z_i,\xi_i\}_{i=1}^n\Bigg) \leqslant n^{-3}.\end{aligned}$$

By (5.152), (5.157), the previous inequality and the union bound together show that

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n}\sum_{j\neq i}\xi_j^2\Big[\langle Z_j^{(2)},Z_i^{(1)}\rangle\langle Z_i^{(1)},Z_j^{(3)}\Lambda_1^{-2}\rangle-\langle Z_j^{(2)},Z_j^{(3)}\Lambda_1^{-2}\rangle\Big]\Big|\geqslant C_2\frac{n^{3/2}r(p+\log(n))\log(n)}{\lambda_{\min}^2}\Big)\leqslant n^{-3}.$$

By (5.164), (5.165) and the previous inequality, we have

$$\mathbb{P}\Big(\Big|\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq k\neq i}\xi_j\xi_k\Big[\langle Z_j,Z_i\rangle\langle Z_i,Z_k\Lambda_1^{-2}\rangle-\langle Z_j,Z_k\Lambda_1^{-2}\rangle\Big]\Big|\geqslant C_2\frac{r(p+\log(n))\log(n)}{n^{3/2}\lambda_{\min}^2}\Big)\leqslant C_1p^{-3}. \tag{5.166}$$

By combining (5.151), (5.155), (5.158), (5.161), (5.162), (5.163) and the previous inequality, we conclude that with probability at least $1-C_1p^{-3}$,

$$\begin{aligned}&\Big|\operatorname{tr}\Big(G_1^\top(G_1G_1^\top)^{-2}G_1\Big\{\Big[J_1-\frac{1}{n}\sum_{i=1}^{n}\langle J_1,\mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\Big](U_2\otimes U_3)\Big\}^\top L_1\Big)\Big|\\&\qquad\leqslant C_2\left(\frac{p^2r^2}{n^2\lambda_{\min}^2}+\frac{rp\log(n)}{n^{3/2}\lambda_{\min}^2}+\frac{r\log^2(n)}{n^{3/2}\lambda_{\min}^2}\right).\end{aligned} \tag{5.167}$$

By Lemma 5.2.10, with probability at least $1-p^{-3}$,

$$\|U_1^\top L_1\|=\Big\|\frac{1}{n}U_1^\top\Big(\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)\Big)(U_2\otimes U_3)\Big\|\leqslant C_2\sqrt{\frac{r^2+\log(p)}{n}}$$

and

$$\|J_1\|_F\leqslant\|L_1\|_F\|G_1^\top(G_1G_1^\top)^{-1}G_1\|=\|L_1\|_F\leqslant C_2\sqrt{\frac{p}{n}}.$$

Therefore, by Lemma 5.2.9,

$$\begin{aligned}&\Big|\operatorname{tr}\Big(G_1^\top(G_1G_1^\top)^{-2}G_1\Big\{\Big[J_1-\frac{1}{n}\sum_{i=1}^{n}\langle J_1,\mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\Big](U_2\otimes U_3)\Big\}^\top\mathcal{P}_{U_1}L_1\Big)\Big|\\&\leqslant r\|G_1^\top(G_1G_1^\top)^{-2}G_1\|\Big\|U_1^\top\Big[J_1-\frac{1}{n}\sum_{i=1}^{n}\langle J_1,\mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\Big](U_2\otimes U_3)\Big\|\|U_1^\top L_1\|\end{aligned}$$

$$\leqslant C_2\lambda_{\min}^{-2}\cdot r\sqrt{\frac{r^2+\log(p)}{n}}\cdot\sqrt{\frac{pr}{n}}\|J_1\|_F\leqslant C_2\left(\frac{r^{5/2}p}{n^{3/2}\lambda_{\min}^2}+\frac{r^{3/2}p\sqrt{\log(p)}}{n^{3/2}\lambda_{\min}^2}\right). \quad (5.168)$$

By (5.150), (5.167) and (5.168), with probability at least $1-C_1p^{-3}+C_2e^{-c_0p}$,

$$\left|\operatorname{tr}\left(G_1^\top(G_1G_1^\top)^{-2}G_1\left\{\left[\mathfrak{E}_{U_1,3}^{(1)}-\frac{1}{n}\sum_{i=1}^{n}\langle\mathfrak{E}_{U_1,3}^{(1)},\mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\right](U_2\otimes U_3)\right\}^\top\mathcal{P}_{U_1}^\perp L_1\right)\right|$$

$$\leqslant C_2\left(\kappa_0^3\frac{p^2r^{5/2}}{n^2\lambda_{\min}^3}+\frac{p^2r^2}{n^2\lambda_{\min}^2}+\frac{r^{5/2}p\log(n)}{n^{3/2}\lambda_{\min}^2}+\frac{r\log^2(n)}{n^{3/2}\lambda_{\min}^2}\right). \quad (5.169)$$

Combining (5.147), (5.148), (5.149) and (5.169) together, we have

$$\left|S_1^{(1.5)}\right|\leqslant C_2\left(\kappa_0^3\frac{p^2r^{5/2}}{n^2\lambda_{\min}^3}+\kappa_0^2\frac{p^2r^3}{n^2\lambda_{\min}^2}+\kappa_0\frac{r^3p\log(n)}{n^{3/2}\lambda_{\min}^2}+\frac{r\log^2(n)}{n^{3/2}\lambda_{\min}^2}\right). \quad (5.170)$$

-Step 3.2.2: bounding $|S_2^{(1)}|$ and $|S_3^{(1)}|$. Let $\mathfrak{E}_2^{(0)}=\hat{U}_2^{(0.5)}-U_2R_2^{(0)\top}$. For $k\geqslant 1$, denote

$$\left(\mathfrak{P}_2^{(0)}\right)^{-k}=\begin{cases}\begin{pmatrix}0 & U_2R_2^{(0)\top}\\ R_2^{(0)}U_2^\top & 0\end{pmatrix}, & \text{if } k \text{ is old},\\ \begin{pmatrix}U_2U_2^\top & 0\\ 0 & I_{r_2}\end{pmatrix}, & \text{if } k \text{ is even,}\end{cases}$$

and

$$\left(\mathfrak{P}_2^{(0)}\right)^{0}=\begin{pmatrix}U_{2\perp}U_{2\perp}^\top & 0\\ 0 & 0\end{pmatrix}.$$

Let

$$E_2^{(0)}=\begin{pmatrix}0 & \mathfrak{E}_2^{(0)}\\ \mathfrak{E}_2^{(0)\top} & 0\end{pmatrix}.$$

Similarly to (5.128), with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\|E_2^{(0)}\|\leqslant C\frac{\sqrt{p/n}}{\lambda_{\min}}.$$

By Lemma 5.2.2, with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{pmatrix} \mathcal{P}_{\hat{U}_2^{(1)}} - \mathcal{P}_{U_2} & 0 \\ 0 & 0 \end{pmatrix} = \sum_{k\geqslant 1} \mathcal{S}_{U_2^{(0)},k}\big(E_2^{(0)}\big) \tag{5.171}$$

where

$$\mathcal{S}_{U_2^{(0)},k}(X) = \sum_{s_1+\cdots+s_{k+1}=k} (-1)^{1+\tau(\mathbf{s})}\cdot\big(\mathfrak{P}_2^{(0)}\big)^{-s_1}X\big(\mathfrak{P}_2^{(0)}\big)^{-s_2}X\big(\mathfrak{P}_2^{(0)}\big)^{-s_3}\cdots\big(\mathfrak{P}_2^{(0)}\big)^{-s_k}X\big(\mathfrak{P}_2^{(0)}\big)^{-s_{k+1}}.$$

By (5.130), with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\sum_{k\geqslant 2}\left\|\mathcal{S}_{U_2^{(0)},k}\big(E_2^{(0)}\big)\right\| \leqslant \sum_{k\geqslant 2}\big(4\|E_2^{(0)}\|\big)^k = \frac{16\|E_2^{(0)}\|^2}{1-4\|E_2^{(0)}\|} \leqslant C\frac{p}{n\lambda_{\min}^2}. \tag{5.172}$$

Note that

$$\begin{aligned}
\mathcal{S}_{U_2^{(0)},1}\big(E_2^{(0)}\big) =& \big(\mathfrak{P}_2^{(0)}\big)^{-1}E_2^{(0)}\big(\mathfrak{P}_2^{(0)}\big)^{0} + \big(\mathfrak{P}_2^{(0)}\big)^{0}E_2^{(0)}\big(\mathfrak{P}_2^{(0)}\big)^{-1} \\
=& \begin{pmatrix} 0 & U_2R_2^{(0)\top} \\ R_2^{(0)}U_2^\top & 0 \end{pmatrix}\begin{pmatrix} 0 & \mathfrak{E}_2^{(0)} \\ \mathfrak{E}_2^{(0)\top} & 0 \end{pmatrix}\begin{pmatrix} U_{2\perp}U_{2\perp}^\top & 0 \\ 0 & 0 \end{pmatrix} \\
&+ \begin{pmatrix} U_{2\perp}U_{2\perp}^\top & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & \mathfrak{E}_2^{(0)} \\ \mathfrak{E}_2^{(0)\top} & 0 \end{pmatrix}\begin{pmatrix} 0 & U_2R_2^{(0)\top} \\ R_2^{(0)}U_2^\top & 0 \end{pmatrix} \\
=& \begin{pmatrix} U_2R_2^{(0)\top}\mathfrak{E}_2^{(0)\top}U_{2\perp}U_{2\perp}^\top & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} U_{2\perp}U_{2\perp}^\top\mathfrak{E}_2^{(0)}R_2^{(0)}U_2^\top & 0 \\ 0 & 0 \end{pmatrix} \\
=& \begin{pmatrix} U_2R_2^{(0)\top}\mathfrak{E}_2^{(0)\top}\mathcal{P}_{U_2}^\perp + \mathcal{P}_{U_2}^\perp\mathfrak{E}_2^{(0)}R_2^{(0)}U_2^\top & 0 \\ 0 & 0 \end{pmatrix},
\end{aligned}$$

with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}
&\left\|\Big(\mathcal{P}_{\hat{U}_2^{(1)}} - \mathcal{P}_{U_2}\Big)U_2 - \mathcal{P}_{U_2}^\perp\mathfrak{E}_2^{(0)}R_2^{(0)}\right\| \\
=&\left\|\Big(\mathcal{P}_{\hat{U}_2^{(1)}} - \mathcal{P}_{U_2} - \big(U_2R_2^{(0)\top}\mathfrak{E}_2^{(0)\top}\mathcal{P}_{U_2}^\perp + \mathcal{P}_{U_2}^\perp\mathfrak{E}_2^{(0)}R_2^{(0)}U_2^\top\big)\Big)U_2\right\|
\end{aligned}$$

$$=\left\|\begin{pmatrix}\mathcal{P}_{\hat{U}_2^{(1)}}-\mathcal{P}_{U_2} & 0\\ 0 & 0\end{pmatrix}-\mathcal{S}_{U_2^{(0)},1}\left(E_2^{(0)}\right)\right\|\leqslant\sum_{k\geqslant 2}\left\|\mathcal{S}_{U_2^{(0)},k}\left(E_2^{(0)}\right)\right\|\leqslant C\frac{p}{n\lambda_{\min}^2}.$$

In addition, with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\left\|\hat{U}_2^{(1)}R_2^{(1)}-U_2-(\mathcal{P}_{\hat{U}_2^{(1)}}-\mathcal{P}_{U_2})U_2\right\|=\left\|\hat{U}_2^{(1)}(R_2^{(1)}-\hat{U}_2^{(1)\top}U_2)\right\|\leqslant\left\|R_2^{(1)}-\hat{U}_2^{(1)\top}U_2\right\|\leqslant C\frac{p}{n\lambda_{\min}^2},$$

we know that with probability $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\left\|\hat{U}_2^{(1)}R_2^{(1)}-U_2-\mathcal{P}_{U_2}^{\perp}\mathfrak{E}_2^{(0)}R_2^{(0)}\right\|\leqslant C\frac{p}{n\lambda_{\min}^2}, \tag{5.173}$$

and

$$\begin{aligned}
&\left\|\Delta T_{U_2,1}^{(1)}-U_1G_1\Big(\big(\mathcal{P}_{U_2}^{\perp}\mathfrak{E}_2^{(0)}R_2^{(0)}\big)^{\top}\otimes U_3^{\top}\Big)\right\|_F\\
=&\left\|U_1G_1\Big(\big(\hat{U}_2^{(1)}R_2^{(1)}-U_2-\mathcal{P}_{U_2}^{\perp}\mathfrak{E}_2^{(0)}R_2^{(0)}\big)^{\top}\otimes U_3^{\top}\Big)\right\|_F\\
=&\left\|\mathcal{G}\times_1 U_1\times_2\left(\hat{U}_2^{(1)}R_2^{(1)}-U_2-\mathcal{P}_{U_2}^{\perp}\mathfrak{E}_2^{(0)}R_2^{(0)}\right)\times U_3\right\|_F\\
\leqslant&\|G_2\|\left\|\hat{U}_2^{(1)}R_2^{(1)}-U_2-\mathcal{P}_{U_2}^{\perp}\mathfrak{E}_2^{(0)}R_2^{(0)}\right\|_F\leqslant\kappa_0\lambda_{\min}\cdot C\frac{p\sqrt{r}}{n\lambda_{\min}^2}\leqslant C\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}.
\end{aligned}$$

Let

$$\widetilde{\Delta}T_{U_2,1}^{(1)}=U_1G_1\Big(\big(\mathcal{P}_{U_2}^{\perp}\mathfrak{E}_2^{(0)}R_2^{(0)}\big)^{\top}\otimes U_3^{\top}\Big)$$

and

$$\widetilde{S}_2^{(1)}=\operatorname{tr}\left(G_1^{\top}(G_1G_1^{\top})^{-2}G_1\Big\{\Big[\widetilde{\Delta}T_{U_2,1}^{(1)}-\frac{1}{n}\sum_{i=1}^{n}\langle\widetilde{\Delta}T_{U_2,1}^{(1)},\mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\Big](U_2\otimes U_3)\Big\}^{\top}\mathcal{P}_{U_1}^{\perp}L_1\right).$$

By Lemmas 5.2.9 and 5.2.10, with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\begin{aligned}
\left|S_2^{(1)}-\widetilde{S}_2^{(1)}\right|\leqslant& r\|G_1^{\top}(G_1G_1^{\top})^{-2}G_1\|\cdot\sqrt{\frac{p}{n}}\left\|\Delta T_{U_2,1}^{(1)}-\widetilde{\Delta}T_{U_2,1}^{(1)}\right\|_F\|L_1\|\\
\leqslant& C_2r\lambda_{\min}^{-2}\sqrt{\frac{p}{n}}\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\cdot\sqrt{\frac{p}{n}}\leqslant C_2\kappa_0\frac{p^2r^{3/2}}{n^2\lambda_{\min}^3}.
\end{aligned} \tag{5.174}$$

By (5.120), we have

$$\begin{aligned}\mathfrak{E}_2^{(0)} =&\left(\mathcal{M}_2(\Delta\mathcal{T}_2^{(0.5)})(\hat{U}_1^{(0)}\otimes\hat{U}_3^{(0)}) - \frac{1}{n}\sum_{i=1}^{n}\langle\Delta\mathcal{T}_2^{(0.5)},\mathcal{X}_i\rangle\mathcal{M}_2(\mathcal{X}_i)(\hat{U}_1^{(0)}\otimes\hat{U}_3^{(0)})\right)\hat{G}_2^{(0)\top}\big(\hat{G}_2^{(0)}\hat{G}_2^{(0)\top}\big)^{-1}\\ &+\Big(U_2G_2\big((\hat{U}_1^{(0)\top}U_1)^\top\otimes(\hat{U}_3^{(0)\top}U_3)^\top\big) - U_2R_2^{(0)\top}\hat{G}_2^{(0)}\Big)\hat{G}_2^{(0)\top}\big(\hat{G}_2^{(0)}\hat{G}_2^{(0)\top}\big)^{-1}\\ &+\frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_2(\mathcal{X}_i)(\hat{U}_1^{(0)}\otimes\hat{U}_3^{(0)})\hat{G}_2^{(0)\top}\big(\hat{G}_2^{(0)}\hat{G}_2^{(0)\top}\big)^{-1}\\ =:&\mathfrak{J}_{U_2,1}^{(0)} + \mathfrak{J}_{U_2,2}^{(0)} + \mathfrak{J}_{U_2,3}^{(0)}. \end{aligned} \tag{5.175}$$

Here, $\Delta\mathcal{T}_2^{(0.5)} = \hat{\mathcal{G}}^{(0)}\times_1\hat{U}_1^{(0)}\times_2\hat{U}_2^{(0.5)}\times_3\hat{U}_3^{(0)} - \mathcal{T}$. Let

$$\mathfrak{E}_{U_2,i}^{(0)} = U_1G_1\left(\left(\mathcal{P}_{U_2}^{\perp}\mathfrak{J}_{U_2,i}^{(0)}R_2^{(0)}\right)^\top\otimes U_3^\top\right),\quad \forall i\in[3].$$

Then

$$\widetilde{\Delta}T_{U_2,1}^{(1)} = \mathfrak{E}_{U_2,1}^{(0)} + \mathfrak{E}_{U_2,2}^{(0)} + \mathfrak{E}_{U_2,3}^{(0)}.$$

Similarly to (5.148) and (5.149), with probability at least $1 - C_1p^{-3} - C_1e^{-c_1p}$,

$$\begin{aligned}&\left|\operatorname{tr}\left(G_1^\top(G_1G_1^\top)^{-2}G_1\Big\{\Big[\mathfrak{E}_{U_2,1}^{(0)} - \frac{1}{n}\sum_{i=1}^{n}\langle\mathfrak{E}_{U_2,1}^{(0)},\mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\Big](U_2\otimes U_3)\Big\}^\top\mathcal{P}_{U_1}^{\perp}L_1\right)\right|\\ &\qquad\leqslant C_2\kappa_0^2\frac{p^2r^3}{n^2\lambda_{\min}^2},\end{aligned} \tag{5.176}$$

and

$$\begin{aligned}&\left|\operatorname{tr}\left(G_1^\top(G_1G_1^\top)^{-2}G_1\Big\{\Big[\mathfrak{E}_{U_2,2}^{(0)} - \frac{1}{n}\sum_{i=1}^{n}\langle\mathfrak{E}_{U_2,2}^{(0)},\mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\Big](U_2\otimes U_3)\Big\}^\top\mathcal{P}_{U_1}^{\perp}L_1\right)\right|\\ &\qquad\leqslant C_2\left(\kappa_0^2\frac{p^2r^{5/2}}{n^2\lambda_{\min}^3} + \kappa_0^2\frac{p^2r^3}{n^2\lambda_{\min}^2} + \kappa_0\frac{r^3p + r^{3/2}p\sqrt{\log(p)}}{n^{3/2}\lambda_{\min}^2}\right).\end{aligned} \tag{5.177}$$

Similarly to (5.142), with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\left\|\mathfrak{J}^{(0)}_{U_2,3} - \frac{1}{n}\sum_{i=1}^{n} \xi_i \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3) G_2^\top \left(G_2 G_2^\top\right)^{-1} R_2^{(0)\top}\right\| \leqslant C_2 \kappa_0^2 \frac{p\sqrt{r}}{n\lambda_{\min}^2} \tag{5.178}$$

and

$$\begin{aligned}
&\left\|\mathfrak{E}^{(0)}_{U_2,3} - U_1 G_1\Big(\Big(\mathcal{P}^{\perp}_{U_2}\big(\frac{1}{n}\sum_{i=1}^{n} \xi_i \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3) G_2^\top (G_2 G_2^\top)^{-1}\big)\Big)^\top \otimes U_3^\top\Big)\right\| \\
=&\left\|U_1 G_1\Big(\Big(\mathcal{P}^{\perp}_{U_2}\big(\mathfrak{J}^{(0)}_{U_2,3} - \frac{1}{n}\sum_{i=1}^{n} \xi_i \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3) G_2^\top (G_2 G_2^\top)^{-1} R_2^{(0)\top}\big) R_2^{(0)}\Big)^\top \otimes U_3^\top\Big)\right\| \\
\leqslant&\|G_1\|\left\|\mathfrak{J}^{(0)}_{U_2,3} - \frac{1}{n}\sum_{i=1}^{n} \xi_i \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3) G_2^\top \left(G_2 G_2^\top\right)^{-1} R_2^{(0)\top}\right\| \leqslant C_2 \kappa_0^3 \frac{p\sqrt{r}}{n\lambda_{\min}}.
\end{aligned} \tag{5.179}$$

Let

$$\widetilde{\mathfrak{E}}^{(0)}_{U_2,3} = U_1 G_1\Big(\Big(\mathcal{P}^{\perp}_{U_2}\big(\frac{1}{n}\sum_{i=1}^{n} \xi_i \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3) G_2^\top (G_2 G_2^\top)^{-1}\big)\Big)^\top \otimes U_3^\top\Big).$$

The same argument for proving (5.150) shows that with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\begin{aligned}
&\left|\operatorname{tr}\Big(G_1^\top (G_1 G_1^\top)^{-2} G_1 \Big\{\Big[\mathfrak{E}^{(0)}_{U_2,3} - \frac{1}{n}\sum_{i=1}^{n} \langle \mathfrak{E}^{(0)}_{U_2,3}, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i)\Big](U_2 \otimes U_3)\Big\}^\top \mathcal{P}^{\perp}_{U_1} L_1\Big)\right| \\
\leqslant&\left|\operatorname{tr}\Big(G_1^\top (G_1 G_1^\top)^{-2} G_1 \Big\{\Big[\widetilde{\mathfrak{E}}^{(0)}_{U_2,3} - \frac{1}{n}\sum_{i=1}^{n} \langle \widetilde{\mathfrak{E}}^{(0)}_{U_2,3}, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i)\Big](U_2 \otimes U_3)\Big\}^\top \mathcal{P}^{\perp}_{U_1} L_1\Big)\right| \\
&+ C_2 \kappa_0^3 \frac{p^2 r^{5/2}}{n^2 \lambda_{\min}^3}.
\end{aligned} \tag{5.180}$$

Let

$$G_2 = U_{G_2} \Lambda_2 V_{G_2}^\top$$

be the SVD decomposition of G_2. Let $W_i = U_{1\perp}^\top \mathcal{M}_1(\mathcal{X}_i)(U_2 \otimes U_3)V_{G_1} \in \mathbb{R}^{(p_1-r_1)\times r_1}$ and $\widetilde{W}_i = U_{2\perp}^\top \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3)V_{G_2} \in \mathbb{R}^{(p_2-r_2)\times r_2}$. Then $W_i \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $\widetilde{W}_i \overset{\text{i.i.d.}}{\sim} N(0,1)$. In addition, since $\mathcal{X}_i \times_1 [U_1\ U_{1\perp}] \times_2 [U_2\ U_{2\perp}] \times_3 [U_3\ U_{3\perp}] \overset{\text{i.i.d.}}{\sim} N(0,1)$, $U_{1\perp}^\top \mathcal{M}_1(\mathcal{X}_i)(U_2 \otimes U_3)$ and $U_{2\perp}^\top \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3)$ are independent. Therefore, W_i and $\widetilde{W}_i$ are independent.

Note that $\widetilde{\mathfrak{E}}^{(0)\top}_{U_2,3}\mathcal{P}^\perp_{U_1} = 0$ and

$$
\begin{aligned}
&\langle \widetilde{\mathfrak{E}}^{(0)}_{U_2,3}, \mathcal{M}_1(\mathcal{X}_i)\rangle \\
=&\left\langle U_1 G_1\Big(\Big(\mathcal{P}^\perp_{U_2}\big(\frac{1}{n}\sum_{j=1}^n \xi_j \mathcal{M}_2(\mathcal{X}_j)(U_1 \otimes U_3)G_2^\top (G_2G_2^\top)^{-1}\big)\Big)^\top \otimes U_3^\top\Big), \mathcal{M}_1(\mathcal{X}_i)\right\rangle \\
=&\left\langle \mathcal{G} \times_1 U_1 \times_2 \Big(\mathcal{P}^\perp_{U_2}\Big(\frac{1}{n}\sum_{j=1}^n \xi_j \mathcal{M}_2(\mathcal{X}_j)(U_1 \otimes U_3)G_2^\top (G_2G_2^\top)^{-1}\Big)\Big) \times_3 U_3, \mathcal{X}_i\right\rangle \\
=&\left\langle \mathcal{P}^\perp_{U_2}\frac{1}{n}\sum_{j=1}^n \xi_j \mathcal{M}_2(\mathcal{X}_j)(U_1 \otimes U_3)G_2^\top\big(G_2G_2^\top\big)^{-1} G_2\big(U_1^\top \otimes U_3^\top\big), \mathcal{M}_2(\mathcal{X}_i)\right\rangle \\
=&\left\langle \frac{1}{n}\sum_{j=1}^n \xi_j U_{2\perp}^\top \mathcal{M}_2(\mathcal{X}_j)(U_1 \otimes U_3)V_{G_2}, U_{2\perp}^\top \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3)V_{G_2}\right\rangle \\
=&\left\langle \frac{1}{n}\sum_{j=1}^n \xi_j \widetilde{W}_j, \widetilde{W}_i\right\rangle,
\end{aligned}
$$

we have

$$
\begin{aligned}
&\left|\operatorname{tr}\left(G_1^\top (G_1G_1^\top)^{-2}G_1\Big\{\Big[\widetilde{\mathfrak{E}}^{(0)}_{U_2,3} - \frac{1}{n}\sum_{i=1}^n \langle \widetilde{\mathfrak{E}}^{(0)}_{U_2,3}, \mathcal{M}_1(\mathcal{X}_i)\rangle \mathcal{M}_1(\mathcal{X}_i)\Big](U_2 \otimes U_3)\Big\}^\top \mathcal{P}^\perp_{U_1} L_1\right)\right| \\
=&\left|\operatorname{tr}\left(\left(\frac{1}{n}\sum_{i=1}^n \Big\langle \frac{1}{n}\sum_{j=1}^n \xi_j \widetilde{W}_j, \widetilde{W}_i\Big\rangle U_{1\perp}\mathcal{M}_1(\mathcal{X}_i)(U_2 \otimes U_3)V_{G_1}\right)^\top U_{1\perp} L_1 V_{G_1}\Lambda_1^{-2}\right)\right| \\
=&\left|\frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{k=1}^n \xi_j\xi_k \langle \widetilde{W}_i, \widetilde{W}_j\rangle\langle W_i, W_k\Lambda_1^{-2}\rangle\right| \\
\leqslant&\left|\frac{1}{n^3}\sum_{i=1}^n \xi_i^2 \|\widetilde{W}_i\|_F^2\, \|W_i\Lambda_1^{-2}\|_F^2\right| + \left|\frac{1}{n^3}\sum_{i=1}^n\sum_{k\neq i} \xi_i\xi_k \|\widetilde{W}_i\|_F^2 \langle W_i, W_k\Lambda_1^{-2}\rangle\right|
\end{aligned}
$$

$$+\left|\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\xi_i\xi_j\langle\widetilde{W}_i,\widetilde{W}_j\rangle\left\|W_i\Lambda_1^{-1}\right\|_F^2\right|+\left|\frac{1}{n^3}\sum_{i=1}^{n}\sum_{j\neq i}\xi_j^2\langle\widetilde{W}_i,\widetilde{W}_j\rangle\langle W_i,W_j\Lambda_1^{-2}\rangle\right|$$
$$+\left|\frac{1}{n^3}\sum_{i\neq j\neq k}\xi_j\xi_k\langle\widetilde{W}_i,\widetilde{W}_j\rangle\langle W_i,W_k\Lambda_1^{-2}\rangle\right|.$$

Similarly to (5.167), with probability at least $1-C_1p^{-3}$,

$$\left|\operatorname{tr}\left(G_1^\top(G_1G_1^\top)^{-2}G_1\Big\{\Big[\widetilde{\mathfrak{E}}^{(0)}_{U_{2,3}}-\frac{1}{n}\sum_{i=1}^{n}\langle\widetilde{\mathfrak{E}}^{(0)}_{U_{2,3}},\mathcal{M}_1(\mathcal{X}_i)\rangle\mathcal{M}_1(\mathcal{X}_i)\Big](U_2\otimes U_3)\Big\}^\top\mathcal{P}^\perp_{U_1}L_1\right)\right|$$
$$\leqslant C_2\left(\frac{p^2r^2}{n^2\lambda_{\min}^2}+\frac{rp\log(n)}{n^{3/2}\lambda_{\min}^2}+\frac{r\log^2(n)}{n^{3/2}\lambda_{\min}^2}\right).$$

By (5.174), (5.176), (5.177), (5.180) and the above inequality, with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\left|S_2^{(1)}\right|\leqslant C_2\left(\kappa_0^3\frac{p^2r^{5/2}}{n^2\lambda_{\min}^3}+\kappa_0^2\frac{p^2r^3}{n^2\lambda_{\min}^2}+\kappa_0\frac{r^3p\log(n)}{n^{3/2}\lambda_{\min}^2}+\frac{r\log^2(n)}{n^{3/2}\lambda_{\min}^2}\right).\tag{5.181}$$

Similarly, with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\left|S_3^{(1)}\right|\leqslant C_2\left(\kappa_0^3\frac{p^2r^{5/2}}{n^2\lambda_{\min}^3}+\kappa_0^2\frac{p^2r^3}{n^2\lambda_{\min}^2}+\kappa_0\frac{r^3p\log(n)}{n^{3/2}\lambda_{\min}^2}+\frac{r\log^2(n)}{n^{3/2}\lambda_{\min}^2}\right).\tag{5.182}$$

Putting (5.143), (5.146), (5.170), (5.181) and (5.182) together, we get with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$ that

$$\left|\operatorname{tr}\left(\mathfrak{J}^{(1)\top}_{U_1,1}\mathcal{P}^\perp_{U_1}\mathfrak{J}^{(1)}_{U_1,3}\right)\right|\leqslant C_2\left(\kappa_0^3\frac{p^2r^{5/2}}{n^2\lambda_{\min}^3}+\kappa_0^2\frac{p^2r^3}{n^2\lambda_{\min}^2}+\kappa_0\frac{r^3p\log(n)}{n^{3/2}\lambda_{\min}^2}+\frac{r\log^2(n)}{n^{3/2}\lambda_{\min}^2}\right).\tag{5.183}$$

Final step: characterizing the distribution of $\left|\operatorname{tr}\left(\mathfrak{J}_{U_1,3}^{(1)\top}\mathcal{P}_{U_1}^{\perp}\mathfrak{J}_{U_1,3}^{(1)}\right)\right|$. By (5.123), with probability at least $1-p^3-C_1e^{-c_1p}$,

$$
\begin{aligned}
&\left\|\mathfrak{J}_{U_1,3}^{(1)}-\frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)\Big((\hat{U}_2^{(1)}R_2^{(1)})\otimes(\hat{U}_3^{(1)}R_3^{(1)})\Big)G_1^\top(G_1G_1^\top)^{-1}R_1^{(1)\top}\right\|\\
=&\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\Big(\hat{G}_1^{(1)\top}\left(\hat{G}_1^{(1)}\hat{G}_1^{(1)\top}\right)^{-1}-(R_2^{(1)}\otimes R_3^{(1)})G_1^\top(G_1G_1^\top)^{-1}R_1^{(1)\top}\Big)\right\|\\
\leqslant&C_2\kappa_0^2\lambda_{\min}^{-2}\left(\sqrt{\frac{r^2+\log(p)}{n}}+\kappa_0\frac{pr}{n}+\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right)\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right\|\\
\leqslant&C_2\kappa_0^2\lambda_{\min}^{-2}\left(\sqrt{\frac{r^2+\log(p)}{n}}+\kappa_0\frac{pr}{n}+\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right)\\
&\cdot\left(\sqrt{\frac{p}{n}}+\sqrt{\frac{pr^2}{n}}\left\|\hat{U}_2^{(1)}-U_2R_2^{(1)\top}\right\|+\sqrt{\frac{pr^2}{n}}\left\|\hat{U}_3^{(1)}-U_3R_3^{(1)\top}\right\|\right)\\
\leqslant&C_2\kappa_0^2\lambda_{\min}^{-2}\left(\sqrt{\frac{r^2+\log(p)}{n}}+\kappa_0\frac{pr}{n}+\kappa_0\frac{p\sqrt{r}}{n\lambda_{\min}}\right)\sqrt{\frac{p}{n}}\\
\leqslant&C_2\left(\kappa_0^2\frac{\sqrt{pr^2+p\log(p)}}{n\lambda_{\min}^2}+\kappa_0^3\frac{p^{3/2}r}{n^{3/2}\lambda_{\min}^2}+\kappa_0^3\frac{p^{3/2}r^{1/2}}{n^{3/2}\lambda_{\min}^3}\right).
\end{aligned}
$$

Moreover, by (5.173) and Lemma 5.2.10, with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$
\begin{aligned}
&\left\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)\left((\hat{U}_2^{(1)}R_2^{(1)})\otimes(\hat{U}_3^{(1)}R_3^{(1)})-(U_2+\mathcal{P}_{U_2}^{\perp}\mathfrak{E}_2^{(0)}R_2^{(0)})\otimes(U_3+\mathcal{P}_{U_3}^{\perp}\mathfrak{E}_3^{(0)}R_3^{(0)})\right)\right.\\
&\qquad\left.\cdot G_1^\top(G_1G_1^\top)^{-1}\right\|\\
\leqslant&C_2\sqrt{\frac{pr}{n}}\|\hat{U}_2^{(1)}R_2^{(1)}-U_2-\mathcal{P}_{U_2}^{\perp}\mathfrak{E}_2^{(0)}R_2^{(0)}\|\|G_1^\top(G_1G_1^\top)^{-1}\|\\
&+C_2\sqrt{\frac{pr}{n}}\|\hat{U}_3^{(1)}R_3^{(1)}-U_3-\mathcal{P}_{U_3}^{\perp}\mathfrak{E}_3^{(0)}R_3^{(0)}\|\|G_1^\top(G_1G_1^\top)^{-1}\|\\
&+C_2\sqrt{\frac{pr}{n}}\|\hat{U}_2^{(1)}R_2^{(1)}-U_2-\mathcal{P}_{U_2}^{\perp}\mathfrak{E}_2^{(0)}R_2^{(0)}\|\|\hat{U}_3^{(1)}R_3^{(1)}-U_3-\mathcal{P}_{U_3}^{\perp}\mathfrak{E}_3^{(0)}R_3^{(0)}\|_F\|G_1^\top(G_1G_1^\top)^{-1}\|
\end{aligned}
$$

$$\leqslant C_2 \frac{p^{3/2} r^{1/2}}{n^{3/2} \lambda_{\min}^3}.$$

Similarly to (5.139), with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\left\| \mathfrak{I}^{(0)}_{U_2,1} \right\| \leqslant C_2 \kappa_0 \frac{pr}{n\lambda_{\min}}. \tag{5.184}$$

Note that $\mathcal{P}^{\perp}_{U_2} \mathfrak{I}^{(0)}_{U_2,2} = 0$, by (5.178) and (5.184), with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\begin{aligned}
&\left\| \mathcal{P}^{\perp}_{U_2} \mathfrak{E}^{(0)}_2 R^{(0)}_2 - \frac{1}{n} \sum_{i=1}^{n} \xi_i \mathcal{P}^{\perp}_{U_2} \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3) G_2^{\top} (G_2 G_2^{\top})^{-1} \right\| \\
\leqslant & \left\| \mathfrak{I}^{(0)}_{U_2,1} \right\| + \left\| \mathfrak{I}^{(0)}_{U_2,3} - \frac{1}{n} \sum_{i=1}^{n} \xi_i \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3) G_2^{\top} (G_2 G_2^{\top})^{-1} R^{(0)\top}_2 \right\| \\
\leqslant & C_2 \left(\kappa_0 \frac{pr}{n\lambda_{\min}} + \kappa_0^2 \frac{p\sqrt{r}}{n\lambda^2_{\min}} \right).
\end{aligned} \tag{5.185}$$

Similarly, with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\left\| \mathcal{P}^{\perp}_{U_3} \mathfrak{E}^{(0)}_3 R^{(0)}_3 - \frac{1}{n} \sum_{i=1}^{n} \xi_i \mathcal{P}^{\perp}_{U_3} \mathcal{M}_3(\mathcal{X}_i)(U_1 \otimes U_2) G_3^{\top} (G_3 G_3^{\top})^{-1} \right\| \leqslant C_2 \left(\kappa_0 \frac{pr}{n\lambda_{\min}} + \kappa_0^2 \frac{p\sqrt{r}}{n\lambda^2_{\min}} \right). \tag{5.186}$$

In addition, with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\| \mathcal{P}^{\perp}_{U_2} \mathfrak{E}^{(0)}_2 R^{(0)}_2 \| \| \mathcal{P}^{\perp}_{U_3} \mathfrak{E}^{(0)}_3 R^{(0)}_3 \| \leqslant \| \mathfrak{E}^{(0)}_2 \| \| \mathfrak{E}^{(0)}_3 \| \leqslant C_2 \frac{p}{n\lambda^2_{\min}}.$$

Let

$$Q_2 = \frac{1}{n} \sum_{i=1}^{n} \xi_i \mathcal{P}^{\perp}_{U_2} \mathcal{M}_2(\mathcal{X}_i)(U_1 \otimes U_3) G_2^{\top} \left(G_2 G_2^{\top} \right)^{-1}$$

and

$$Q_3 = \frac{1}{n} \sum_{i=1}^{n} \xi_i \mathcal{P}^{\perp}_{U_3} \mathcal{M}_3(\mathcal{X}_i)(U_1 \otimes U_2) G_3^{\top} \left(G_3 G_3^{\top} \right)^{-1}.$$

Combining the previous six inequalities together, with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\begin{aligned}&\left\|\mathfrak{J}^{(1)}_{U_1,3} - \frac{1}{n}\sum_{i=1}^{n} \xi_i \mathcal{M}_1(\mathcal{X}_i)\left(U_2 \otimes U_3 + Q_2 \otimes U_3 + U_2 \otimes Q_3\right) G_1^\top (G_1 G_1^\top)^{-1} R_1^{(1)\top}\right\| \\ \leqslant & C_2\left(\kappa_0^2 \frac{\sqrt{pr^2 + p\log(p)}}{n\lambda_{\min}^2} + \kappa_0^3 \frac{p^{3/2} r^{3/2}}{n^{3/2}\lambda_{\min}^2} + \kappa_0^3 \frac{p^{3/2} r}{n^{3/2}\lambda_{\min}^3}\right). \qquad (5.187)\end{aligned}$$

Let $\bar{\mathfrak{J}}^{(1)}_{U_1,3} = \frac{1}{n}\sum_{i=1}^{n} \xi_i \mathcal{M}_1(\mathcal{X}_i)\left(U_2 \otimes U_3 + Q_2 \otimes U_3 + U_2 \otimes Q_3\right) G_1^\top (G_1 G_1^\top)^{-1}$. By Lemma 5.2.10, with probability at least $1 - e^{-C_1 p}$,

$$\left\|\bar{\mathfrak{J}}^{(1)}_{U_1,3}\right\| \leqslant C_2 \sqrt{\frac{p}{n}} \lambda_{\min}^{-1} + C_2 \sqrt{\frac{pr}{n}} \sqrt{\frac{p}{n}} \lambda_{\min}^{-2} \leqslant C_2 \sqrt{\frac{p}{n}} \lambda_{\min}^{-1}.$$

Therefore, with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\begin{aligned}&\left|\operatorname{tr}\left(\mathfrak{J}^{(1)\top}_{U_1,3} \mathcal{P}^{\perp}_{U_1} \mathfrak{J}^{(1)}_{U_1,3}\right) - \operatorname{tr}\left(\bar{\mathfrak{J}}^{(1)\top}_{U_1,3} \mathcal{P}^{\perp}_{U_1} \bar{\mathfrak{J}}^{(1)}_{U_1,3}\right)\right| \\ \leqslant & 2\left|\operatorname{tr}\left(\left(\mathfrak{J}^{(1)}_{U_1,3} - \bar{\mathfrak{J}}^{(1)}_{U_1,3} R_1^{(1)\top}\right)^\top \mathcal{P}^{\perp}_{U_1} \bar{\mathfrak{J}}^{(1)}_{U_1,3} R_1^{(1)\top}\right)\right| + \left|\operatorname{tr}\left(\left(\mathfrak{J}^{(1)}_{U_1,3} - \bar{\mathfrak{J}}^{(1)}_{U_1,3} R_1^{(1)\top}\right)^\top \mathcal{P}^{\perp}_{U_1}\left(\mathfrak{J}^{(1)}_{U_1,3} - \bar{\mathfrak{J}}^{(1)}_{U_1,3} R_1^{(1)\top}\right)\right)\right| \\ \leqslant & r\left\|\mathfrak{J}^{(1)}_{U_1,3} - \bar{\mathfrak{J}}^{(1)}_{U_1,3} R_1^{(1)\top}\right\|\left\|\bar{\mathfrak{J}}^{(1)}_{U_1,3}\right\| + r\left\|\mathfrak{J}^{(1)}_{U_1,3} - \bar{\mathfrak{J}}^{(1)}_{U_1,3} R_1^{(1)\top}\right\|^2 \\ \leqslant & C_2\left(\kappa_0^2 \frac{pr^2 + pr\sqrt{\log(p)}}{n^{3/2}\lambda_{\min}^3} + \kappa_0^3 \frac{p^2 r^{5/2}}{n^2\lambda_{\min}^3} + \kappa_0^3 \frac{p^2 r^2}{n^2 \lambda_{\min}^4}\right). \qquad (5.188)\end{aligned}$$

Let

$$\widetilde{Z}_j = U_{1\perp}^\top \mathcal{M}_1(\mathcal{X}_j)(U_2 \otimes U_3) V_{G_1} \in \mathbb{R}^{(p_1 - r_1) \times r_1}.$$

Then $\widetilde{Z}_j \overset{\text{i.i.d.}}{\sim} N(0,1)$. With probability at least $1 - e^{-C_1 p}$,

$$\begin{aligned}&\left|\operatorname{tr}\left(\bar{\mathfrak{J}}^{(1)\top}_{U_1,3} \mathcal{P}^{\perp}_{U_1} \bar{\mathfrak{J}}^{(1)}_{U_1,3}\right) - \frac{1}{n^2}\left\|\sum_{i=1}^{n} \xi_i \widetilde{Z}_i \Lambda_1^{-1}\right\|_F^2\right| \\ \leqslant & 2\left|\left\langle \mathcal{P}^{\perp}_{U_1} \frac{1}{n}\sum_{j=1}^{n} \xi_j \mathcal{M}_1(\mathcal{X}_j)(U_2 \otimes U_3) V_{G_1} \Lambda_1^{-2}, \frac{1}{n}\sum_{i=1}^{n} \xi_i \mathcal{M}_1(\mathcal{X}_i)\left(Q_2 \otimes U_3\right) V_{G_1}\right\rangle\right|\end{aligned}$$

$$
\begin{aligned}
&+2\left|\left\langle \mathcal{P}^{\perp}_{U_1}\frac{1}{n}\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)(U_2\otimes U_3)V_{G_1}\Lambda_1^{-2}, \frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)\left(U_2\otimes Q_3\right)V_{G_1}\right\rangle\right| \\
&+C_2 r\left(\sqrt{\frac{pr}{n}}\sqrt{\frac{p}{n}}\lambda_{\min}^{-2}\right)^2. \qquad (5.189)
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\left|\left\langle \mathcal{P}^{\perp}_{U_1}\frac{1}{n}\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)(U_2\otimes U_3)V_{G_1}\Lambda_1^{-2}, \frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)\left(Q_2\otimes U_3\right)V_{G_1}\right\rangle\right| \\
=&\Bigg|\Bigg\langle \left(U_{1\perp}^{\top}\frac{1}{n}\sum_{k=1}^{n}\xi_k\mathcal{M}_1(\mathcal{X}_k)(U_{2\perp}\otimes U_3)\right)^{\top}\left(U_{1\perp}^{\top}\frac{1}{n}\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)(U_2\otimes U_3)V_{G_1}\right)\Lambda_1^{-2}, \\
&\left(\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i U_{2\perp}^{\top}\mathcal{M}_2(\mathcal{X}_i)(U_1\otimes U_3)G_2^{\top}\left(G_2G_2^{\top}\right)^{-1}\right)\otimes I_{r_3}\right)V_{G_1}\Bigg\rangle\Bigg| \\
=&\left|\left\langle\left(\frac{1}{n}\sum_{j=1}^{n}\xi_j W_j V_{G_1}\Lambda_1^{-2}\right)\left[\left(\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i\widetilde{W}_i V_{G_2}\Lambda_2^{-1}\right)\otimes I_{r_3}\right)V_{G_1}\right]^{\top}, \frac{1}{n}\sum_{k=1}^{n}\xi_k\bar{W}_k\right\rangle\right|.
\end{aligned}
$$

Here, $\bar{W}_i = U_{1\perp}^{\top}\mathcal{M}_1(\mathcal{X}_i)(U_{2\perp}\otimes U_3)\in\mathbb{R}^{(p_1-r_1)\times((p_2-r_2)r_3)}, W_i = U_{1\perp}^{\top}\mathcal{M}_1(\mathcal{X}_i)(U_2\otimes U_3)\in\mathbb{R}^{(p_1-r_1)\times r_2r_3}$ and $\widetilde{W}_i = U_{2\perp}^{\top}\mathcal{M}_2(\mathcal{X}_i)(U_1\otimes U_3)\in\mathbb{R}^{(p_2-r_2)\times r_1r_3}$. Since $\mathcal{X}_i\times_1[U_1\ U_1^{\top}]\times_2[U_2\ U_2^{\top}]\times_3[U_3\ U_3^{\top}]\overset{\text{i.i.d.}}{\sim}N(0,1)$, we know that $\bar{W}_i\overset{\text{i.i.d.}}{\sim}N(0,1), W_i\overset{\text{i.i.d.}}{\sim}N(0,1), \widetilde{W}_i\overset{\text{i.i.d.}}{\sim}N(0,1)$, and $\bar{W}_i, W_i$ and $\widetilde{W}_i$ are independent. Therefore,

$$
\begin{aligned}
&\left\langle \mathcal{P}^{\perp}_{U_1}\frac{1}{n}\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)(U_2\otimes U_3)V_{G_1}\Lambda_1^{-2}, \frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)\left(Q_2\otimes U_3\right)V_{G_1}\right\rangle\Bigg|\left\{W_k,\widetilde{W}_k,\xi_k\right\}_{k=1}^{n} \\
\sim&N\left(0,\left\|\left(\frac{1}{n}\sum_{j=1}^{n}\xi_jW_jV_{G_1}\Lambda_1^{-2}\right)\left[\Big((\frac{1}{n}\sum_{i=1}^{n}\xi_i\widetilde{W}_iV_{G_2}\Lambda_2^{-1})\otimes I_{r_3}\Big)V_{G_1}\right]^{\top}\right\|_F^2\frac{\|\vec{\xi}\|_2^2}{n^2}\right).
\end{aligned}
$$

Note that with probability at least $1-e^{-C_1p}$,

$$
\left\|\left(\frac{1}{n}\sum_{j=1}^{n}\xi_jW_jV_{G_1}\Lambda_1^{-2}\right)\left[\Big((\frac{1}{n}\sum_{i=1}^{n}\xi_i\widetilde{W}_iV_{G_2}\Lambda_2^{-1})\otimes I_{r_3}\Big)V_{G_1}\right]^{\top}\right\|_F
$$

$$\begin{aligned}&\leqslant\sqrt{r}\Big\|\frac{1}{n}\sum_{j=1}^{n}\xi_j W_j V_{G_1}\Lambda_1^{-2}\Big\|\Big\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\widetilde{W}_i V_{G_2}\Lambda_2^{-1}\Big\| \leqslant C_2\sqrt{r}\left(\sqrt{\frac{p}{n}}\lambda_{\min}^{-2}\right)\left(\sqrt{\frac{p}{n}}\lambda_{\min}^{-1}\right)\\ &\leqslant C_2\frac{pr^{1/2}}{n\lambda_{\min}^3},\end{aligned}$$

and as a result

$$\begin{aligned}&\mathbb{P}\left(\left|\left\langle \mathcal{P}_{U_1}^{\perp}\frac{1}{n}\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)(U_2\otimes U_3)V_{G_1}\Lambda_1^{-2},\frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)\left(Q_2\otimes U_3\right)V_{G_1}\right\rangle\right|\geqslant C_2\frac{\sqrt{r}p\sqrt{\log(p)}}{n^{3/2}\lambda_{\min}^3}\right)\\ &\leqslant \frac{C_1}{p^3}.\end{aligned}$$

Similarly,

$$\begin{aligned}&\mathbb{P}\left(\left|\left\langle \mathcal{P}_{U_1}^{\perp}\frac{1}{n}\sum_{j=1}^{n}\xi_j\mathcal{M}_1(\mathcal{X}_j)(U_2\otimes U_3)V_{G_1}\Lambda_1^{-2},\frac{1}{n}\sum_{i=1}^{n}\xi_i\mathcal{M}_1(\mathcal{X}_i)\left(U_2\otimes Q_3\right)V_{G_1}\right\rangle\right|\geqslant C_2\frac{\sqrt{r}p\sqrt{\log(p)}}{n^{3/2}\lambda_{\min}^3}\right)\\ &\leqslant \frac{C_1}{p^3}.\end{aligned}$$

By (5.188), (5.189) and the two inequalities above, with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$,

$$\left|\mathrm{tr}\left(\mathfrak{J}_{U_{1,3}}^{(1)\top}\mathcal{P}_{U_1}^{\perp}\mathfrak{J}_{U_{1,3}}^{(1)}\right)-\frac{1}{n^2}\Big\|\sum_{i=1}^{n}\xi_i\widetilde{Z}_i\Lambda_1^{-1}\Big\|_F^2\right|\leqslant C_2\left(\kappa_0^2\frac{pr^2+pr\sqrt{\log(p)}}{n^{3/2}\lambda_{\min}^3}+\kappa_0^3\frac{p^2r^{5/2}}{n^2\lambda_{\min}^3}+\kappa_0^3\frac{p^2r^2}{n^2\lambda_{\min}^4}\right).$$

By (5.138), (5.140), (5.183) and the previous inequality, with probability at least $1-C_1p^{-3}-C_1e^{-c_1p}$, we have proved

$$\begin{aligned}&\left|\left\langle\begin{pmatrix}U_1U_1^{\top} & 0\\ 0 & I_{r_1}\end{pmatrix},\mathcal{S}_{U_1^{(1)},2}\left(E\right)\right\rangle+\frac{1}{n^2}\left\|\sum_{i=1}^{n}\xi_i\widetilde{Z}_i\Lambda_1^{-1}\right\|_F^2\right|\\ &\leqslant C_2\left(\kappa_0^2\frac{pr^2+pr\sqrt{\log(p)}}{n^{3/2}\lambda_{\min}^3}+\kappa_0^3\frac{p^2r^{5/2}}{n^2\lambda_{\min}^3}+\kappa_0^3\frac{p^2r^2}{n^2\lambda_{\min}^4}+\kappa_0^2\frac{p^2r^3}{n^2\lambda_{\min}^2}+\kappa_0\frac{r^3p\log(n)}{n^{3/2}\lambda_{\min}^2}+\frac{r\log^2(n)}{n^{3/2}\lambda_{\min}^2}\right).\end{aligned}\tag{5.190}$$

By (5.133), (5.137) and (5.190), with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\left| \left\| \hat{U}_1^{(2)} \hat{U}_1^{(2)\top} - U_1 U_1^\top \right\|_F^2 - \frac{2}{n^2} \left\| \sum_{i=1}^n \xi_i \widetilde{Z}_i \Lambda_1^{-1} \right\|_F^2 \right|$$
$$\leqslant C_2 \left(\kappa_0^2 \frac{pr^2 + pr\sqrt{\log(p)}}{n^{3/2} \lambda_{\min}^3} + \kappa_0^3 \frac{p^2 r^{5/2}}{n^2 \lambda_{\min}^3} + \kappa_0^3 \frac{p^2 r^2}{n^2 \lambda_{\min}^4} + \kappa_0^2 \frac{p^2 r^3}{n^2 \lambda_{\min}^2} + \kappa_0 \frac{r^3 p \log(n)}{n^{3/2} \lambda_{\min}^2} + \frac{r \log^2(n)}{n^{3/2} \lambda_{\min}^2} \right). \tag{5.191}$$

Recall that $\widetilde{Z}_i = U_{1\perp}^\top \mathcal{M}_1(\mathcal{X}_i)(U_2 \otimes U_3) V_{G_1} \in \mathbb{R}^{(p_1 - r_1) \times r_1}$ so that $\widetilde{Z}_i \overset{\text{i.i.d.}}{\sim} N(0,1)$. For fixed $a = (a_1, \dots, a_n) \in \mathbb{R}^n$, the rows of $(\sum_{i=1}^n a_i \widetilde{Z}_i) \Lambda_1^{-1} \overset{\text{i.i.d.}}{\sim} N\left(0, \|a\|_2^2 \Lambda_1^{-2}\right)$. Therefore, for any $a \in \mathbb{R}^n$,

$$\frac{1}{\|a\|_2^2} \left\| \Big(\sum_{i=1}^n a_i \widetilde{Z}_i \Big) \Lambda_1^{-1} \right\|_F^2 \overset{\text{d.}}{=} \frac{1}{n} \left\| \sum_{i=1}^n \widetilde{Z}_i \Lambda_1^{-1} \right\|_F^2$$

which means that

$$\frac{1}{\|\vec{\xi}\|_2^2} \left\| \Big(\sum_{i=1}^n \xi_i \widetilde{Z}_i \Big) \Lambda_1^{-1} \right\|_F^2 \overset{\text{d.}}{=} \frac{1}{n} \left\| \sum_{i=1}^n \widetilde{Z}_i \Lambda_1^{-1} \right\|_F^2.$$

For any $1 \leqslant i \leqslant p_1 - r_1$,

$$\mathbb{E} \left\| \Big(\sum_{j=1}^n (\widetilde{Z}_j)_{[i,:]} \Big) \Lambda_1^{-1} \right\|_2^2 = n \|\Lambda_1^{-1}\|_F^2,$$

$$\mathrm{Var} \left(\left\| \Big(\sum_{j=1}^n (\widetilde{Z}_j)_{[i,:]} \Big) \Lambda_1^{-1} \right\|_2^2 \right) = 2n^2 \|\Lambda_1^{-2}\|_F^2.$$

and

$$\mathbb{E} \left\| \Big(\sum_{j=1}^n (\widetilde{Z}_j)_{[i,:]} \Big) \Lambda_1^{-1} \right\|_2^6 \leqslant C_3 n^3 \sum_{j_1, j_2, j_3 = 1}^{r_1} \frac{1}{\lambda_{j_1}^{(1)2} \lambda_{j_2}^{(1)2} \lambda_{j_3}^{(1)2}} \leqslant C_3 n^3 \|\Lambda_1^{-1}\|_F^6.$$

By Berry-Esseen theorem, we have

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\Big(\frac{\frac{2}{n\|\vec{\xi}\|_2^2}\big\|(\sum_{j=1}^n \xi_j\widetilde{Z}_j)\Lambda_1^{-1}\big\|_F^2-2(p_1-r_1)\|\Lambda_1^{-1}\|_F^2/n}{\sqrt{8(p_1-r_1)}\|\Lambda_1^{-2}\|_F/n}\leqslant x\Big)-\Phi(x)\right|\leqslant C\Big(\frac{\|\Lambda_1^{-1}\|_F^4}{\|\Lambda_1^{-2}\|_F^2}\Big)^{3/2}\frac{1}{\sqrt{p}}. \tag{5.192}$$

By Lemma 5.2.10, with probability $1-e^{-C_1p}$,

$$\big\|\sum_{j=1}^n\xi_j\widetilde{Z}_j\big\|\leqslant C_2\sqrt{np}.$$

Therefore, with probability $1-e^{-C_1p}$,

$$\Big\|(\sum_{j=1}^n\xi_j\widetilde{Z}_j)\Lambda_1^{-1}\Big\|_F^2\leqslant r_1\Big\|\sum_{j=1}^n\xi_j\widetilde{Z}_j\Big\|^2\|\Lambda_1^{-1}\|^2\leqslant C_2rnp\lambda_{\min}^{-2}.$$

By Bernstein-type inequality ((Vershynin, 2010, Proposition 5.16)),

$$\mathbb{P}\left(|\|\vec{\xi}\|_2^2-n|\geqslant C_2\sqrt{n\log(p)}\right)\leqslant 2\exp\left[-C_1\min\left\{\frac{n\log(p)}{n},\sqrt{n\log(p)}\right\}\right]\leqslant p^{-C_1}. \tag{5.193}$$

Therefore, with probability at least $1-C_1p^{-3}$,

$$\left|\frac{2}{n\|\vec{\xi}\|_2^2}-\frac{2}{n^2}\right|\leqslant\frac{2|\|\vec{\xi}\|_2^2-n|}{n^2\|\vec{\xi}\|_2^2}\leqslant C_2\frac{\sqrt{n\log(p)}}{n^3}. \tag{5.194}$$

By (5.192), (5.194) and the previous inequality and the similar proof in Theorem 3.3.1, we have

$$\begin{aligned}&\sup_{x\in\mathbb{R}}\left|\mathbb{P}\Big(\frac{\|\hat{U}_1^{(2)}\hat{U}_1^{(2)\top}-U_1U_1^\top\|_F^2-\frac{2p_1}{n}\|\Lambda_1^{-1}\|_F^2}{\frac{\sqrt{8p_1}}{n}\|\Lambda_1^{-2}\|_F}\leqslant x\Big)-\Phi(x)\right|\\ \leqslant&C_2\left(\kappa_0^4\sqrt{\frac{pr^3+pr\log(p)}{n\lambda_{\min}^2}}+\kappa_0^5\frac{p^{3/2}r^2}{n\lambda_{\min}}+\kappa_0^5\frac{p^{3/2}r^{3/2}}{n\lambda_{\min}^2}+\kappa_0^4\frac{p^{3/2}r^{5/2}}{n}+\kappa_0^3\sqrt{\frac{r^5p\log^2(n)}{n}}\right)\end{aligned}$$

$$+ C_3\frac{r^{3/2}}{\sqrt{p}} + C_1 e^{-c_1 p}$$

where we use the fact that $C_2\frac{\kappa_0^2\sqrt{r}\log^2(n)}{\sqrt{pn}} \leqslant C_3\frac{r^{3/2}}{\sqrt{p}}$.

5.2.5 Proof of Theorem 3.3.4

Without loss of generality, we assume $\sigma = 1$. By (5.108) and Lemma 5.2.10, with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\max_{1\leqslant i\leqslant r_1}\left|\lambda_i - \hat{\lambda}_i^{(1)}\right| \leqslant \left\|\hat{G}_1 - R_1 G_1(R_2^\top \otimes R_3^\top)\right\| \leqslant C_2\left(\kappa_0\frac{pr}{n} + \kappa_0\frac{p\sqrt{r}}{n\lambda_{\mathsf{min}}} + \sqrt{\frac{r^2+\log(p)}{n}}\right).$$

Therefore, with probability at least $1 - C_1 p^{-3} - C_1 e^{-c_1 p}$,

$$\left|\left\|\Lambda_1^{-1}\right\|_F^2 - \left\|\left(\hat{\Lambda}_1^{(1)}\right)^{-1}\right\|_F^2\right| \leqslant r\max_{1\leqslant i\leqslant r_1}\frac{\left|\lambda_i^2 - \hat{\lambda}_i^2\right|}{\lambda_i^2\hat{\lambda}_i^2} \leqslant C_2 r\left(\kappa_0\frac{pr}{n} + \kappa_0\frac{p\sqrt{r}}{n\lambda_{\mathsf{min}}} + \sqrt{\frac{r^2+\log(p)}{n}}\right)\lambda_{\mathsf{min}}^{-3},$$

and

$$\frac{\left|\left\|\Lambda_1^{-1}\right\|_F^2 - \left\|\left(\hat{\Lambda}_1^{(1)}\right)^{-1}\right\|_F^2\right|}{\left\|\Lambda_1^{-2}\right\|_F} \leqslant \frac{\max_{1\leqslant i\leqslant r_1}\frac{\left|\lambda_i^2 - \hat{\lambda}_i^2\right|}{\lambda_i^2\hat{\lambda}_i^2}}{\kappa_0^{-2}\lambda_{\mathsf{min}}^{-2}} \leqslant C_2\kappa_0^2\left(\kappa_0\frac{pr}{n} + \kappa_0\frac{p\sqrt{r}}{n\lambda_{\mathsf{min}}} + \sqrt{\frac{r^2+\log(p)}{n}}\right)\lambda_{\mathsf{min}}^{-1}.$$

The rest of the proof is essentially the same as the proof of Theorem 3.3.2.

5.2.6 Proof of Theorem 3.4.1

Without loss of generality, we assume $\sigma = 1$, $\pi(j) = j$ and only prove the normal approximation for $\langle\hat{u}_1, u_1\rangle^2$. We denote $\tilde{U}_1 = (u_2, \cdots, u_r) \in \mathbb{O}_{p_1, r-1}$, $\tilde{V}_1 = (v_2, \cdots, v_r) \in \mathbb{O}_{p_2, r-1}$ and $\tilde{W}_1 = (w_2, \cdots, w_r) \in \mathbb{O}_{p_3, r-1}$. Denote the $(r-1) \times (r-1) \times (r-1)$ diagonal tensor $\tilde{\boldsymbol{\Lambda}} = \mathrm{diag}(\lambda_2, \cdots, \lambda_r)$, and $\tilde{\Lambda}_1 = \mathcal{M}_1(\tilde{\boldsymbol{\Lambda}}), \tilde{\Lambda}_2 = \mathcal{M}_2(\tilde{\boldsymbol{\Lambda}}), \tilde{\Lambda}_3 = \mathcal{M}_3(\tilde{\boldsymbol{\Lambda}})$.

By definition, $\hat{u}_1$ is the left singular vector of $\mathcal{A} \times_2 \hat{v}_1^{(1)\top} \times_3 \hat{w}_1^{(1)\top}$, for which we write

$$\begin{aligned}\mathcal{A}\times_2\hat{v}_1^{(1)\top}\times_3\hat{w}_1^{(1)\top}=&\lambda_1\langle\hat{v}_1^{(1)},v_1\rangle\langle\hat{w}_1^{(1)},w_1\rangle u_1+\tilde{U}_1\tilde{\Lambda}_1(\tilde{V}_1\otimes\tilde{W}_1)^\top(\hat{v}_1^{(1)}\otimes\hat{w}_1^{(1)})+\underbrace{\mathcal{Z}\times_2\hat{v}_1^{(1)\top}\times_3\hat{w}_1^{(1)\top}}_{\hat{z}_1^{(1)}}\\=:&\tilde{\lambda}_1u_1+\hat{E}_1^{(1)}\end{aligned}$$

where we define $\tilde{\lambda}_1=\lambda_1\langle\hat{v}_1^{(1)},v_1\rangle\langle\hat{w}_1^{(1)},w_1\rangle$ and $\hat{E}_1^{(1)}=\tilde{U}_1\tilde{\Lambda}_1(\tilde{V}_1\otimes\tilde{W}_1)^\top(\hat{v}_1^{(1)}\otimes\hat{w}_1^{(1)})+\hat{z}_1^{(1)}$.

Similarly to the proof of Theorem 3.3.1, the following bounds hold with probability at least $1-C_1e^{-c_1p}$,

$$\max\left\{\|\tilde{V}_1^\top\hat{v}_1^{(1)}\|_2,\|\tilde{W}_1^\top\hat{w}_1^{(1)}\|_2\right\}\leqslant C_2\sqrt{p}/\lambda_{\min}\quad\text{and}\quad\|\hat{z}_1^{(1)}\|_2\leqslant C_3(\sqrt{p}+p\lambda_{\min}^{-1}).$$

As a result, with the same probability, $\|\hat{E}_1^{(1)}\|\leqslant C_3(\sqrt{p}+\kappa_0p\lambda_{\min}^{-1})$. Then, we write

$$\begin{pmatrix}0&\tilde{\lambda}_1u_1+\hat{E}_1^{(1)}\\\tilde{\lambda}_1u_1^\top+\hat{E}_1^{(1)\top}&0\end{pmatrix}=\tilde{\lambda}_1\begin{pmatrix}0&u_1\\u_1^\top&0\end{pmatrix}+\begin{pmatrix}0&\hat{E}_1^{(1)}\\\hat{E}_1^{(1)\top}&0\end{pmatrix}.$$

We now apply Lemma 5.2.2 and represent $\langle u_1u_1^\top,\hat{u}_1\hat{u}_1^\top-u_1u_1^\top\rangle$. Similarly to the proof of Theorem 3.3.1, the 1st-order term does not matter, and the 4-th and higher-order terms can be simply bounded. Therefore, we obtain,

$$\langle u_1u_1^\top,\hat{u}_1\hat{u}_1^\top-u_1u_1^\top\rangle=-\frac{1}{\tilde{\lambda}_1^2}\operatorname{tr}\left(\hat{E}_1^{(1)\top}U_{1\perp}U_{1\perp}^\top\hat{E}_1^{(1)}\right)+\frac{2}{\tilde{\lambda}_1^3}\operatorname{tr}\left(\hat{E}_1^{(1)\top}u_1\hat{E}_1^{(1)\top}U_{1\perp}U_{1\perp}^\top\hat{E}_1^{(1)}\right)+\tilde{R}_1^{(1)}$$

where $(u_1,U_{1\perp})\in\mathbb{O}_{p_1}$, and $\|\tilde{R}_1^{(1)}\|\leqslant C_3\kappa_0^4p^2/\lambda_{\min}^4$ with probability at least $1-C_1e^{-c_1p}$.

Similarly to the proof of Theorem 3.3.1, we have $|\hat{E}_1^{(1)\top}u_1|\leqslant C_1(p\lambda_{\min}^{-1}+\sqrt{\log p})$ which holds with probability at least $1-p^{-3}$. Therefore, with probability at least $1-2p^{-3}$, $|\operatorname{tr}(\hat{E}_1^{(1)\top}u_1\hat{E}_1^{(1)\top}U_{1\perp}U_{1\perp}^\top\hat{E}_1^{(1)})|\leqslant C_2\kappa_0^2p(p\lambda_{\min}^{-1}+\sqrt{\log p})$. Therefore, we conclude with probability at least $1 - 3p^{-3}$ that

$$\left| \langle u_1 u_1^\top, \hat{u}_1 \hat{u}_1^\top - u_1 u_1^\top \rangle + \frac{1}{\tilde{\lambda}_1^2} \operatorname{tr}\left(\hat{E}_1^{(1)\top} U_{1\perp} U_{1\perp}^\top \hat{E}_1^{(1)} \right) \right| \leqslant C_2 \Big(\frac{\kappa_0^2 p \sqrt{\log p}}{\lambda_{\min}^3} + \frac{\kappa_0^4 p^2}{\lambda_{\min}^4} \Big).$$

Similarly to the proof of Theorem 3.3.1, we have $|\tilde{\lambda}_1^2 - \lambda_1^2| \leqslant C_2 \kappa_0^2 p$ with probability at least $1 - C_1 e^{-c_1 p}$.

Therefore, with probability at least $1 - 4p^{-3}$ that

$$\left| \langle u_1 u_1^\top, \hat{u}_1 \hat{u}_1^\top - u_1 u_1^\top \rangle + \frac{1}{\lambda_1^2} \operatorname{tr}\left(\hat{E}_1^{(1)\top} U_{1\perp} U_{1\perp}^\top \hat{E}_1^{(1)} \right) \right| \leqslant C_2 \Big(\frac{\kappa_0^2 p \sqrt{\log p}}{\lambda_{\min}^3} + \frac{\kappa_0^4 p^2}{\lambda_{\min}^4} \Big).$$

It then suffices to prove the normal approximation of $\operatorname{tr}(\hat{E}_1^{(1)\top} U_{1\perp} U_{1\perp}^\top \hat{E}_1^{(1)})$. Recall that $\hat{E}_1^{(1)} = \tilde{U}_1 \tilde{\Lambda}_1 (\tilde{V}_1 \otimes \tilde{W}_1)^\top (\hat{v}_1^{(1)} \otimes \hat{w}_1^{(1)}) + \hat{z}_1^{(1)}$. Note that

$$\|\tilde{U}_1 \tilde{\Lambda}_1 (\tilde{V}_1 \otimes \tilde{W}_1)^\top (\hat{v}_1^{(1)} \otimes \hat{w}_1^{(1)})\|_2 \leqslant C_2 \frac{\kappa_0 p}{\lambda_{\min}}$$

implying that, with probability at least $1 - C_1 e^{-c_1 p}$,

$$\left| \operatorname{tr}\Big(\big(\tilde{U}_1 \tilde{\Lambda}_1 (\tilde{V}_1 \otimes \tilde{W}_1)^\top (\hat{v}_1^{(1)} \otimes \hat{w}_1^{(1)})\big)^\top U_{1\perp} U_{1\perp}^\top \big(\tilde{U}_1 \tilde{\Lambda}_1 (\tilde{V}_1 \otimes \tilde{W}_1)^\top (\hat{v}_1^{(1)} \otimes \hat{w}_1^{(1)})\big) \Big) \right| \leqslant C_3 \frac{\kappa_0^2 p^2}{\lambda_{\min}^2}.$$

Now, we consider the cross term (recall $z_1 = \mathcal{Z} \times_2 v_1^\top \times_3 w_1^\top$) and conclude with probability at least $1 - p^{-3}$,

$$\begin{aligned}
&\left| \operatorname{tr}\Big(\hat{z}_1^{(1)\top} U_{1\perp} U_{1\perp}^\top \big(\tilde{U}_1 \tilde{\Lambda}_1 (\tilde{V}_1 \otimes \tilde{W}_1)^\top (\hat{v}_1^{(1)} \otimes \hat{w}_1^{(1)})\big) \Big) \right| = \left| \operatorname{tr}\Big(\hat{z}_1^{(1)\top} \big(\tilde{U}_1 \tilde{\Lambda}_1 (\tilde{V}_1 \otimes \tilde{W}_1)^\top (\hat{v}_1^{(1)} \otimes \hat{w}_1^{(1)})\big) \Big) \right| \\
&\quad \leqslant \left| \operatorname{tr}\Big(z_1^\top \big(\tilde{U}_1 \tilde{\Lambda}_1 (\tilde{V}_1 \otimes \tilde{W}_1)^\top (\hat{v}_1^{(1)} \otimes \hat{w}_1^{(1)})\big) \Big) \right| + \left| \operatorname{tr}\Big((\hat{z}_1^{(1)} - z_1)^\top \big(\tilde{U}_1 \tilde{\Lambda}_1 (\tilde{V}_1 \otimes \tilde{W}_1)^\top (\hat{v}_1^{(1)} \otimes \hat{w}_1^{(1)})\big) \Big) \right| \\
&\quad \leqslant C_2 \sqrt{r \log p} \cdot \kappa_0 p / \lambda_{\min} + C_3 \kappa_0 p^2 / \lambda_{\min}^2
\end{aligned}$$

where the first term is due to $\|z_1^\top \tilde{U}_1\| = O(\sqrt{r \log p})$ with probability at least $1 - p^{-3}/2$, and the second term is similar as the proof of Lemma 2.

To this end, we obtain with probability at least $1-5p^{-3}$ that

$$\left|\langle u_1u_1^\top,\hat{u}_1\hat{u}_1^\top-u_1u_1^\top\rangle+\frac{1}{\lambda_1^2}\operatorname{tr}\big(\hat{z}_1^{(1)\top}U_{1\perp}U_{1\perp}^\top\hat{z}_1^{(1)}\big)\right|\leqslant C_2\Big(\frac{\kappa_0^2p\sqrt{\log p}+\kappa_0p\sqrt{r\log p}}{\lambda_{\min}^3}+\frac{\kappa_0^4p^2}{\lambda_{\min}^4}\Big). \tag{5.195}$$

Now, we investigate the main term $\operatorname{tr}\big(\hat{z}_1^{(1)\top}U_{1\perp}U_{1\perp}^\top\hat{z}_1^{(1)}\big)$ for which we write

$$\begin{aligned}
&\operatorname{tr}\big(\hat{z}_1^{(1)\top}U_{1\perp}U_{1\perp}^\top\hat{z}_1^{(1)}\big)=\operatorname{tr}\big(U_{1\perp}U_{1\perp}^\top\hat{z}_1^{(1)}\hat{z}_1^{(1)\top}\big)=\operatorname{tr}\big(U_{1\perp}U_{1\perp}^\top Z_1\big((\hat{v}_1^{(1)}\hat{v}_1^{(1)\top})\otimes(\hat{w}_1^{(1)}\hat{w}_1^{(1)\top})\big)Z_1^\top\big)\\
=&\operatorname{tr}\big(U_{1\perp}U_{1\perp}^\top Z_1\big((v_1v_1^\top)\otimes(w_1w_1^\top)\big)Z_1^\top\big)+\operatorname{tr}\big(U_{1\perp}U_{1\perp}^\top Z_1\big((\hat{v}_1^{(1)}\hat{v}_1^{(1)\top}-v_1v_1^\top)\otimes(w_1w_1^\top)\big)Z_1^\top\big)\\
&+\operatorname{tr}\big(U_{1\perp}U_{1\perp}^\top Z_1\big((v_1v_1^\top)\otimes(\hat{w}_1^{(1)}\hat{w}_1^{(1)\top}-w_1w_1^\top)\big)Z_1^\top\big)\\
&+\operatorname{tr}\big(U_{1\perp}U_{1\perp}^\top Z_1\big((\hat{v}_1^{(1)}\hat{v}_1^{(1)\top}-v_1v_1^\top)\otimes(\hat{w}_1^{(1)}\hat{w}_1^{(1)\top}-w_1w_1^\top)\big)Z_1^\top\big)
\end{aligned}$$

where we denote $Z_1=\mathcal{M}_1(\mathcal{Z})$, and the last term can be simply bounded by $C_4p^2/\lambda_{\min}^2$ with probability at least $1-C_1e^{-c_1p}$.

The idea of bounding the term $\operatorname{tr}\big(U_{1\perp}U_{1\perp}^\top,Z_1\big((\hat{v}_1^{(1)}\hat{v}_1^{(1)\top}-v_1v_1^\top)\otimes(w_1w_1^\top)\big)Z_1^\top\big)$ is the same as the Step 4 in the proof of Theorem 3.3.1. Indeed, we shall recall that $\hat{v}_1^{(1)}$ is the singular vector of

$$\begin{aligned}
\hat{v}_1^{(0.5)}=&\mathcal{A}\times_1\hat{u}_1^{(0)}\times_3\hat{w}_1^{(0)}\\
=&\lambda_1\langle u_1,\hat{u}_1^{(0)}\rangle\langle w_1,\hat{w}_1^{(0)}\rangle v_1+\tilde{V}_1\tilde{\Lambda}_2(\tilde{U}_1\otimes\tilde{W}_1)^\top(\hat{u}_1^{(0)}\otimes\hat{w}_1^{(0)})+\mathcal{Z}\times_1\hat{u}_1^{(0)}\times_3\hat{w}_1^{(0)}\\
=&\tilde{\lambda}_1^{(0)}v_1+\hat{E}_2^{(0)}.
\end{aligned}$$

Similarly to the proof of Theorem 3.3.1, it suffices to consider the 1st-order term in $\hat{v}_1^{(1)}\hat{v}_1^{(1)\top}-v_1v_1^\top$. It is then easy to show that $\big|\operatorname{tr}\big(U_{1\perp}U_{1\perp}^\top,Z_1\big((\hat{v}_1^{(1)}\hat{v}_1^{(1)\top}-v_1v_1^\top)\otimes(w_1w_1^\top)\big)Z_1^\top\big)\big|\leqslant C_4\kappa_0p\sqrt{\log p}/\lambda_{\min}+C_5\kappa_0^4p^2/\lambda_{\min}^2$ with probability at least $1-p^{-3}$. Together with (5.195), we conclude with probability at least $1-6p^{-3}$ that

$$\left|\langle u_1u_1^\top,\hat{u}_1\hat{u}_1^\top-u_1u_1^\top\rangle+\frac{1}{\lambda_1^2}\operatorname{tr}\big(z_1^\top U_{1\perp}U_{1\perp}^\top z_1\big)\right|\leqslant C_2\Big(\frac{\kappa_0^2p\sqrt{\log p}+k_0p\sqrt{r\log p}}{\lambda_{\min}^3}+\frac{\kappa_0^4p^2}{\lambda_{\min}^4}\Big).$$

The rest of the proof is identical to the final step in the proof of Theorem 3.3.1.

5.2.7 Proof of Theorem 3.5.1

Without loss of generality, we assume $\sigma = 1$. For random variables (or vectors) A and B, we use $A \overset{d.}{\to}$ (resp. $\overset{p.}{\to}, \overset{a.s.}{\to}$)$B$ as a shorthand for $A \overset{d.}{\to}$ (resp. $\overset{p.}{\to}, \overset{a.s.}{\to}$)$B$ as $p \to \infty$. By (Zhang and Xia, 2018, Theorem 1), with probability at least $1 - C_1 e^{-c_1 p}$,

$$\|\hat{u} - u\|_2, \|\hat{v} - v\|_2, \|\hat{w} - w\|_2 \leqslant C_2 \frac{\sqrt{p}}{\lambda}.$$

Let $U_\perp \in \mathbb{O}_{p_1, p_1 - 1}, V_\perp \in \mathbb{O}_{p_2, p_2 - 1}, W_\perp \in \mathbb{O}_{p_3, p_3 - 1}$ be orthogonal complements of u, v and w, respectively. For any $O_i \in \mathbb{O}_{p_i - 1}, i \in [3]$, let

$$\tilde{O}_1 = uu^\top + U_\perp O_1 U_\perp^\top \in \mathbb{O}_{p_1}, \quad \tilde{O}_2 = vv^\top + V_\perp O_2 V_\perp^\top \in \mathbb{O}_{p_2}, \quad \tilde{O}_3 = ww^\top + W_\perp O_3 W_\perp^\top \in \mathbb{O}_{p_3}.$$

Let $\tilde{\mathcal{A}} = \mathcal{A} \times_1 \tilde{O}_1^\top \times_2 \tilde{O}_2^\top \times_3 \tilde{O}_3^\top$. Notice that $\tilde{O}_1 u = u, \tilde{O}_2 v = v, \tilde{O}_3 w = w$, we have

$$\begin{aligned}\tilde{\mathcal{A}} =& \mathcal{A} \times_1 \tilde{O}_1^\top \times_2 \tilde{O}_2^\top \times_3 \tilde{O}_3^\top = \mathcal{T} \times_1 \tilde{O}_1^\top \times_2 \tilde{O}_2^\top \times_3 \tilde{O}_3^\top + \mathcal{Z} \times_1 \tilde{O}_1^\top \times_2 \tilde{O}_2^\top \times_3 \tilde{O}_3^\top \\ =& \mathcal{T} + \mathcal{Z} \times_1 \tilde{O}_1^\top \times_2 \tilde{O}_2^\top \times_3 \tilde{O}_3^\top.\end{aligned}$$

Since $O_i \in \mathbb{O}_{p_{i-1}}$, the entries of $\mathcal{Z} \times_1 \tilde{O}_1^\top \times_2 \tilde{O}_2^\top \times_3 \tilde{O}_3^\top \overset{i.i.d.}{\sim} N(0, 1)$. Consequently,

$$\tilde{\mathcal{A}} \overset{d.}{=} \mathcal{A}. \tag{5.196}$$

Let $\tilde{u}, \tilde{v}, \tilde{w}$ be the outputs of Algorithm 4 after $t_{\max}$ iterations. Then we have

$$\tilde{u} = \tilde{O}_1^\top \hat{u}, \quad \tilde{v} = \tilde{O}_2^\top \hat{v}, \quad \tilde{w} = \tilde{O}_3^\top \hat{w}.$$

In addition, we have

$$\langle u, \hat{u} \rangle = \langle \tilde{O}_1 u, \hat{u} \rangle = \langle u, \tilde{O}_1^\top \hat{u} \rangle = \langle u, \tilde{u} \rangle.$$

Similarly, we have $\langle v, \hat{v}\rangle = \langle v, \tilde{v}\rangle$ and $\langle w, \hat{w}\rangle = \langle w, \tilde{w}\rangle$. By (5.196), for any $O_i \in \mathbb{O}_{p_i-1}, i \in [3]$,

$$\begin{aligned}(\hat{u}^\top U_\perp, \hat{v}^\top V_\perp, \hat{w}^\top W_\perp)\Big|(\langle u,\hat{u}\rangle, \langle v,\hat{v}\rangle, \langle w,\hat{w}\rangle) \overset{d.}{=}&(\tilde{u}^\top U_\perp, \tilde{v}^\top V_\perp, \tilde{w}^\top W_\perp)\Big|(\langle u,\tilde{u}\rangle, \langle v,\tilde{v}\rangle, \langle w,\tilde{w}\rangle)\\ =&(\hat{u}^\top \tilde{O}_1 U_\perp, \hat{v}^\top \tilde{O}_2 V_\perp, \hat{w}^\top \tilde{O}_3 W_\perp)\Big|(\langle u,\hat{u}\rangle, \langle v,\hat{v}\rangle, \langle w,\hat{w}\rangle)\\ =&(\hat{u}^\top U_\perp O_1, \hat{v}^\top V_\perp O_2, \hat{w}^\top W_\perp O_3)\Big|(\langle u,\hat{u}\rangle, \langle v,\hat{v}\rangle, \langle w,\hat{w}\rangle).\end{aligned}$$

Therefore, for any $O_i \in \mathbb{O}_{p_i-1}, i \in [3]$,

$$\begin{aligned}&\left(\frac{\hat{u}^\top U_\perp}{\|U_\perp^\top \hat{u}\|_2}, \frac{\hat{v}^\top V_\perp}{\|V_\perp^\top \hat{v}\|_2}, \frac{\hat{w}^\top W_\perp}{\|W_\perp^\top \hat{w}\|_2}\right)\Bigg|(\langle u,\hat{u}\rangle, \langle v,\hat{v}\rangle, \langle w,\hat{w}\rangle)\\ \overset{d.}{=}&\left(\frac{\hat{u}^\top U_\perp O_1}{\|(U_\perp O_1)^\top \hat{u}\|_2}, \frac{\hat{v}^\top V_\perp O_2}{\|(V_\perp O_2)^\top \hat{v}\|_2}, \frac{\hat{w}^\top W_\perp O_3}{\|(W_\perp O_3)^\top \hat{w}\|_2}\right)\Bigg|(\langle u,\hat{u}\rangle, \langle v,\hat{v}\rangle, \langle w,\hat{w}\rangle)\\ =&\left(\frac{\hat{u}^\top U_\perp}{\|U_\perp^\top \hat{u}\|_2}O_1, \frac{\hat{v}^\top V_\perp}{\|V_\perp^\top \hat{v}\|_2}O_2, \frac{\hat{w}^\top W_\perp}{\|W_\perp^\top \hat{w}\|_2}O_3\right)\Bigg|(\langle u,\hat{u}\rangle, \langle v,\hat{v}\rangle, \langle w,\hat{w}\rangle).\end{aligned}$$

Set $O_1 = I_{p_1-1}$, for any $v_1, v_2 \in \mathbb{S}^{p_2-2} = \{x \in \mathbb{R}^{p_2-1} : \|x\|_2 = 1\}$ and $w_1, w_2 \in \mathbb{S}^{p_3-2}$, we know that there exists $O_2 \in \mathbb{O}_{p_2-1}$ and $O_3 \in \mathbb{O}_{p_3-1}$ such that $v_2 = O_2 v_1, w_2 = O_3 v_2$. Then for any Borel set $A \subseteq \mathbb{S}^{p_1-2}$,

$$\begin{aligned}&\mathbb{P}\left(\frac{U_\perp^\top \hat{u}}{\|U_\perp^\top \hat{u}\|_2} \in A\middle|\frac{V_\perp^\top \hat{v}}{\|V_\perp^\top \hat{v}\|_2} = v_1, \frac{W_\perp^\top \hat{w}}{\|W_\perp^\top \hat{w}\|_2} = w_1\right)\\ =&\mathbb{P}\left(\frac{U_\perp^\top \hat{u}}{\|U_\perp^\top \hat{u}\|_2} \in A\middle|O_2^\top\frac{V_\perp^\top \hat{v}}{\|V_\perp^\top \hat{v}\|_2} = v_1, O_3^\top\frac{W_\perp^\top \hat{w}}{\|W_\perp^\top \hat{w}\|_2} = w_1\right)\\ =&\mathbb{P}\left(\frac{U_\perp^\top \hat{u}}{\|U_\perp^\top \hat{u}\|_2} \in A\middle|\frac{V_\perp^\top \hat{v}}{\|V_\perp^\top \hat{v}\|_2} = v_2, \frac{W_\perp^\top \hat{w}}{\|W_\perp^\top \hat{w}\|_2} = w_2\right).\end{aligned}$$

Therefore, $\frac{U_\perp^\top \hat{u}}{\|U_\perp^\top \hat{u}\|_2}$ and $\left(\frac{V_\perp^\top \hat{v}}{\|V_\perp^\top \hat{v}\|_2}, \frac{W_\perp^\top \hat{w}}{\|W_\perp^\top \hat{w}\|_2}\right)$ are independent. Similarly, we know that $\frac{U_\perp^\top \hat{u}}{\|U_\perp^\top \hat{u}\|_2}, \frac{V_\perp^\top \hat{v}}{\|V_\perp^\top \hat{v}\|_2}$ and $\frac{W_\perp^\top \hat{w}}{\|W_\perp^\top \hat{w}\|_2}$ are independent. In addition, by using the property of Haar measure, we know that $\left(\frac{\hat{u}^\top U_\perp}{\|U_\perp^\top \hat{u}\|_2}, \frac{\hat{v}^\top V_\perp}{\|V_\perp^\top \hat{v}\|_2}, \frac{\hat{w}^\top W_\perp}{\|W_\perp^\top \hat{w}\|_2}\right)^\top$ and $(\langle u,\hat{u}\rangle, \langle v,\hat{v}\rangle, \langle w,\hat{w}\rangle)^\top$

are independent, and

$$\frac{U_\perp^\top \hat{u}}{\|U_\perp^\top \hat{u}\|_2} \overset{\text{d.}}{=} \left(\frac{g_1^{(1)}}{\sqrt{\sum_{i=1}^{p_1-1} g_i^{(1)2}}}, \ldots, \frac{g_{p_1-1}^{(1)}}{\sqrt{\sum_{i=1}^{p_1-1} g_i^{(1)2}}} \right)^\top,$$

$$\frac{V_\perp^\top \hat{v}}{\|V_\perp^\top \hat{v}\|_2} \overset{\text{d.}}{=} \left(\frac{g_1^{(2)}}{\sqrt{\sum_{i=1}^{p_2-1} g_i^{(2)2}}}, \ldots, \frac{g_{p_2-1}^{(2)}}{\sqrt{\sum_{i=1}^{p_2-1} g_i^{(2)2}}} \right)^\top,$$

$$\frac{W_\perp^\top \hat{w}}{\|W_\perp^\top \hat{w}\|_2} \overset{\text{d.}}{=} \left(\frac{g_1^{(3)}}{\sqrt{\sum_{i=1}^{p_3-1} g_i^{(3)2}}}, \ldots, \frac{g_{p_3-1}^{(3)}}{\sqrt{\sum_{i=1}^{p_3-1} g_i^{(3)2}}} \right)^\top$$

where $g^{(1)} = (g_1^{(1)}, \ldots, g_{p_1-1}^{(1)})^\top, g^{(2)} = (g_1^{(2)}, \ldots, g_{p_2-1}^{(2)})^\top, g^{(3)} = (g_1^{(3)}, \ldots, g_{p_3-1}^{(3)})^\top$ are independent standard Gaussian random vectors. Moreover, by SLLN,

$$\frac{1}{p_1} \sum_{i=1}^{p_1-1} g_i^{(1)2} \overset{\text{a.s.}}{\to} 1, \quad \frac{1}{p_2} \sum_{i=1}^{p_2-1} g_i^{(2)2} \overset{\text{a.s.}}{\to} 1, \quad \frac{1}{p_3} \sum_{i=1}^{p_3-1} g_i^{(3)2} \overset{\text{a.s.}}{\to} 1.$$

For any fixed $f_1 \in \mathbb{S}^{p_1-2}, f_2 \in \mathbb{S}^{p_2-2}, f_3 \in \mathbb{S}^{p_3-2}$, notice that $(f_1^\top g_1, f_2^\top g_2, f_3^\top g_3)^\top \sim N(0, I_3)$, we have

$$\begin{aligned} &\left(\sqrt{p_1} \frac{\hat{u}^\top U_\perp}{\|U_\perp^\top \hat{u}\|_2} f_1, \sqrt{p_2} \frac{\hat{v}^\top V_\perp}{\|V_\perp^\top \hat{v}\|_2} f_2, \sqrt{p_3} \frac{\hat{w}^\top W_\perp}{\|W_\perp^\top \hat{w}\|_2} f_3 \right)^\top \\ &\overset{\text{d.}}{=} \operatorname{diag} \left(\sqrt{\frac{p_1}{\sum_{i=1}^{p_1-1} g_i^{(1)2}}}, \sqrt{\frac{p_2}{\sum_{i=1}^{p_2-1} g_i^{(2)2}}}, \sqrt{\frac{p_3}{\sum_{i=1}^{p_3-1} g_i^{(3)2}}} \right) \cdot (f_1^\top g_1, f_2^\top g_2, f_3^\top g_3)^\top \\ &\overset{\text{d.}}{\to} N(0, I_3). \end{aligned} \tag{5.197}$$

By Theorem 3.3.1,

$$\left(\frac{\langle u, \hat{u} \rangle^2 - (1 - p_1 \lambda^{-2})}{\sqrt{2p_1} \lambda^{-2}}, \frac{\langle v, \hat{v} \rangle^2 - (1 - p_2 \lambda^{-2})}{\sqrt{2p_2} \lambda^{-2}}, \frac{\langle w, \hat{w} \rangle^2 - (1 - p_3 \lambda^{-2})}{\sqrt{2p_3} \lambda^{-2}} \right)^\top \overset{\text{d.}}{\to} N(0, I_3). \tag{5.198}$$

The delta method and the fact that $1 - p_i\lambda^{-2} \overset{\text{a.s.}}{\to} 1$ for $i \in [3]$ together show that

$$\left(\frac{\langle u, \hat{u} \rangle - \sqrt{1 - p_1\lambda^{-2}}}{\sqrt{p_1/2}\lambda^{-2}}, \frac{\langle v, \hat{v} \rangle - \sqrt{1 - p_2\lambda^{-2}}}{\sqrt{p_2/2}\lambda^{-2}}, \frac{\langle w, \hat{w} \rangle - \sqrt{1 - p_3\lambda^{-2}}}{\sqrt{p_3/2}\lambda^{-2}} \right)^\top \overset{\text{d.}}{\to} N(0, I_3). \tag{5.199}$$

Also note that

$$1 - p_i\lambda^{-2}/2 - \sqrt{1 - p_i\lambda^{-2}} = \frac{p_i^2\lambda^{-4}}{4(\sqrt{1 - p_i\lambda^{-2}} + 1 - p_i\lambda^{-2}/2)} \asymp \frac{p_i^2}{\lambda^4} \ll \frac{\sqrt{p_i}}{\lambda^2}$$

and $\left(\frac{\hat{u}^\top U_\perp}{\|U_\perp^\top \hat{u}\|_2}, \frac{\hat{v}^\top V_\perp}{\|V_\perp^\top \hat{v}\|_2}, \frac{\hat{w}^\top W_\perp}{\|W_\perp^\top \hat{w}\|_2} \right)^\top$ is independent of $(\langle u, \hat{u} \rangle, \langle v, \hat{v} \rangle, \langle w, \hat{w} \rangle)^\top$, by (5.197) and (5.199), for any fixed $f_1 \in \mathbb{S}^{p_1-2}, f_2 \in \mathbb{S}^{p_2-2}, f_3 \in \mathbb{S}^{p_3-2}$,

$$\begin{pmatrix} \sqrt{p_1}\frac{\hat{u}^\top U_\perp}{\|U_\perp^\top \hat{u}\|_2} f_1 \\ \sqrt{p_2}\frac{\hat{v}^\top V_\perp}{\|V_\perp^\top \hat{v}\|_2} f_2 \\ \sqrt{p_3}\frac{\hat{w}^\top W_\perp}{\|W_\perp^\top \hat{w}\|_2} f_3 \\ \frac{\langle u,\hat{u} \rangle - (1 - p_1\lambda^{-2}/2)}{\sqrt{p_1/2}\lambda^{-2}} \\ \frac{\langle v,\hat{v} \rangle - (1 - p_2\lambda^{-2}/2)}{\sqrt{p_2/2}\lambda^{-2}} \\ \frac{\langle w,\hat{w} \rangle - (1 - p_3\lambda^{-2}/2)}{\sqrt{p_3/2}\lambda^{-2}} \end{pmatrix} \overset{\text{d.}}{\to} N(0, I_6). \tag{5.200}$$

By (5.198),

$$\frac{\|U_\perp^\top \hat{u}\|_2^2}{p_1/\lambda^2} = \frac{1 - \langle \hat{u}, u \rangle^2}{p_1/\lambda^2} \overset{\text{p.}}{\to} 1.$$

By (5.200), for any fixed $f_1 \in \mathbb{S}^{p_1-2}, f_2 \in \mathbb{S}^{p_2-2}, f_3 \in \mathbb{S}^{p_3-2}$,

$$\Bigg(\lambda\hat{u}^\top U_\perp f_1, \lambda\hat{v}^\top V_\perp f_2, \lambda\hat{w}^\top W_\perp f_3, \tag{5.201}$$
$$\frac{\langle u, \hat{u} \rangle - (1 - p_1\lambda^{-2}/2)}{\sqrt{p_1/2}\lambda^{-2}}, \frac{\langle v, \hat{v} \rangle - (1 - p_2\lambda^{-2}/2)}{\sqrt{p_2/2}\lambda^{-2}}, \frac{\langle w, \hat{w} \rangle - (1 - p_3\lambda^{-2}/2)}{\sqrt{p_3/2}\lambda^{-2}} \Bigg)^\top$$

$$
= \mathrm{diag}\left(\frac{\|U_\perp^\top \hat{u}\|_2}{\sqrt{p_1}/\lambda}, \frac{\|V_\perp^\top \hat{v}\|_2}{\sqrt{p_2}/\lambda}, \frac{\|W_\perp^\top \hat{w}\|_2}{\sqrt{p_3}/\lambda}, 1, 1, 1\right) \begin{pmatrix} \sqrt{p_1}\frac{\hat{u}^\top U_\perp}{\|U_\perp^\top \hat{u}\|_2} f_1 \\ \sqrt{p_2}\frac{\hat{v}^\top V_\perp}{\|V_\perp^\top \hat{v}\|_2} f_2 \\ \sqrt{p_3}\frac{\hat{w}^\top W_\perp}{\|W_\perp^\top \hat{w}\|_2} f_3 \\ \frac{\langle u, \hat{u}\rangle - (1 - p_1\lambda^{-2}/2)}{\sqrt{p_1/2}\lambda^{-2}} \\ \frac{\langle v, \hat{v}\rangle - (1 - p_2\lambda^{-2}/2)}{\sqrt{p_2/2}\lambda^{-2}} \\ \frac{\langle w, \hat{w}\rangle^2 - (1 - p_3\lambda^{-2}/2)}{\sqrt{p_3/2}\lambda^{-2}} \end{pmatrix}
$$

$$
\xrightarrow{d.} N(0, I_6). \tag{5.202}
$$

For simplicity, let $q_i = q_i^{(p_i)}$ for $i \in [3]$. Note that

$$
\langle \hat{u}, q_1 \rangle = \langle \hat{u}, \mathcal{P}_u q_1 \rangle + \langle \hat{u}, \mathcal{P}_u^\perp q_1 \rangle = (q_1^\top u)\hat{u}^\top u + (U_\perp^\top q_1)^\top U_\perp^\top \hat{u}. \tag{5.203}
$$

If $q_1 \neq \pm u, q_2 \neq \pm v, q_3 \neq \pm w$ for $i \in [3]$, since $U_\perp^\top q_1, V_\perp^\top q_2, W_\perp^\top q_3$ are fixed vectors, by (5.201), we have

$$
\left(\lambda\frac{(U_\perp^\top q_1)^\top}{\|U_\perp^\top q_1\|_2} U_\perp^\top \hat{u}, \frac{\langle u, \hat{u}\rangle - (1 - p_1\lambda^{-2}/2)}{\sqrt{p_1/2}\lambda^{-2}}, \lambda\frac{(V_\perp^\top q_2)^\top}{\|V_\perp^\top q_2\|_2} V_\perp^\top \hat{v}, \frac{\langle v, \hat{v}\rangle - (1 - p_2\lambda^{-2}/2)}{\sqrt{p_2/2}\lambda^{-2}}, \lambda\frac{(W_\perp^\top q_3)^\top}{\|W_\perp^\top q_3\|_2} W_\perp^\top \hat{w}, \frac{\langle w, \hat{w}\rangle - (1 - p_3\lambda^{-2}/2)}{\sqrt{p_3/2}\lambda^{-2}}\right)^\top
$$

$$
\xrightarrow{d.} N(0, I_6). \tag{5.204}
$$

Since

$$
\frac{\langle q_1, \hat{u} - u\rangle + \frac{p_1\langle q_1, u\rangle}{2\lambda^2}}{\sqrt{\frac{p_1\langle q_1, u\rangle^2}{2\lambda^4} + \frac{1-\langle q_1, u\rangle^2}{\lambda^2}}} = \left(\frac{\frac{\sqrt{1-\langle u, q_1\rangle^2}}{\lambda}}{\sqrt{\frac{1-\langle u, q_1\rangle^2}{\lambda^2} + \frac{p_1\langle u, q_1\rangle^2}{2\lambda^2}}}, \frac{\frac{\sqrt{p_1/2}\langle u, q_1\rangle}{\lambda^2}}{\sqrt{\frac{1-\langle u, q_1\rangle^2}{\lambda^2} + \frac{p_1\langle u, q_1\rangle^2}{2\lambda^2}}}\right)
$$

$$
\cdot \left(\lambda\frac{(U_\perp^\top q_1)^\top}{\|U_\perp^\top q_1\|_2} U_\perp^\top \hat{u}, \frac{\langle u, \hat{u}\rangle - (1 - p_1\lambda^{-2}/2)}{\sqrt{p_1/2}\lambda^{-2}}\right)^\top
$$

where $\left(\frac{\frac{\sqrt{1-\langle u,q_1\rangle^2}}{\lambda}}{\sqrt{\frac{1-\langle u,q_1\rangle^2}{\lambda^2}+\frac{p_1\langle u,q_1\rangle^2}{2\lambda^2}}}, \frac{\frac{\sqrt{p_1/2}\langle u,q_1\rangle}{\lambda^2}}{\sqrt{\frac{1-\langle u,q_1\rangle^2}{\lambda^2}+\frac{p_1\langle u,q_1\rangle^2}{2\lambda^2}}}\right)^\top$ is a fixed unit vector, by Lemma 5.2.11, we have

$$\left(\frac{\langle q_1, \hat{u}-u\rangle + \frac{p_1\langle q_1,u\rangle}{2\lambda^2}}{\sqrt{\frac{p_1\langle q_1,u\rangle^2}{2\lambda^4}+\frac{1-\langle q_1,u\rangle^2}{\lambda^2}}}, \frac{\langle q_2, \hat{v}-v\rangle + \frac{p_2\langle q_2,v\rangle}{2\lambda^2}}{\sqrt{\frac{p_2\langle q_2,v\rangle^2}{2\lambda^4}+\frac{1-\langle q_2,v\rangle^2}{\lambda^2}}}, \frac{\langle q_3, \hat{w}-w\rangle + \frac{p_3\langle q_3,w\rangle}{2\lambda^2}}{\sqrt{\frac{p_3\langle q_3,w\rangle^2}{2\lambda^4}+\frac{1-\langle q_3,w\rangle^2}{\lambda^2}}}\right)^\top \xrightarrow{d.} N(0, I_3). \tag{5.205}$$

If $q_1 = \pm u$, $q_2 = \pm v$ or $q_3 = \pm w$, by (5.198), we still have (5.205).

Specifically, if if $|u_i|, |v_j|, |w_k| \ll \min\{\lambda/p, 1\}$ for some $i \in [p_1], j \in [p_2], k \in [p_3]$, by setting $q_1 = e_i, q_2 = e_j, q_3 = e_k$ and noticing that $\frac{p_1 u_i^2}{\lambda^4} \ll \frac{p_1^2 u_i^2}{\lambda^4} \ll \frac{1}{\lambda^2}$ and $u_i \xrightarrow{a.s.} 0$, we know that (3.17) holds.

Given $\lambda^{-1} \ll |u_i|, |v_j|, |w_k| \ll \min\{\lambda/p, 1/\sqrt{\log(p)}\}$, immediately we have $\frac{\hat{u}_i}{u_i} \xrightarrow{p.} 1$, $\frac{\hat{v}_j}{v_j} \xrightarrow{p.} 1$, $\frac{\hat{w}_k}{w_k} \xrightarrow{p.} 1$. Then

$$\begin{pmatrix} \lambda\frac{\hat{u}_i\hat{v}_j\hat{w}_k - u_i\hat{v}_j\hat{w}_k}{v_j w_k} \\ \lambda\frac{u_i\hat{v}_j\hat{w}_k - u_i v_j\hat{w}_k}{u_i w_k} \\ \lambda\frac{u_i v_j\hat{w}_k - u_i v_j w_k}{u_i v_j} \end{pmatrix} = \begin{pmatrix} \frac{\hat{v}_j\hat{w}_k}{v_j w_k} & 0 & 0 \\ 0 & \frac{\hat{w}_k}{w_k} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda(\hat{u}_i - u_i) \\ \lambda(\hat{v}_j - v_j) \\ \lambda(\hat{w}_k - w_k) \end{pmatrix} \xrightarrow{d.} N(0, I_3).$$

By Lemma 5.2.11, we have

$$\lambda\frac{\hat{u}_i\hat{v}_j\hat{w}_k - u_i v_j w_k}{\sqrt{u_i^2 v_j^2 + v_j^2 w_k^2 + w_k^2 u_i^2}} = \begin{pmatrix} \frac{v_j w_k}{\sqrt{u_i^2 v_j^2 + v_j^2 w_k^2 + w_k^2 u_i^2}} \\ \frac{w_k u_i}{\sqrt{u_i^2 v_j^2 + v_j^2 w_k^2 + w_k^2 u_i^2}} \\ \frac{u_i v_j}{\sqrt{u_i^2 v_j^2 + v_j^2 w_k^2 + w_k^2 u_i^2}} \end{pmatrix}^\top \begin{pmatrix} \lambda\frac{\hat{u}_i\hat{v}_j\hat{w}_k - u_i\hat{v}_j\hat{w}_k}{v_j w_k} \\ \lambda\frac{u_i\hat{v}_j\hat{w}_k - u_i v_j\hat{w}_k}{u_i w_k} \\ \lambda\frac{u_i v_j\hat{w}_k - u_i v_j w_k}{u_i v_j} \end{pmatrix} \xrightarrow{d.} N(0, 1). \tag{5.206}$$

Notice that $(\hat{u}_i^2\hat{v}_j^2 + \hat{v}_j^2\hat{w}_k^2 + \hat{w}_k^2\hat{u}_i^2)/(u_i^2 v_j^2 + v_j^2 w_k^2 + w_k^2 u_i^2) \xrightarrow{p.} 1$, we have

$$\lambda\frac{\hat{u}_i\hat{v}_j\hat{w}_k - u_i v_j w_k}{\sqrt{\hat{u}_i^2\hat{v}_j^2 + \hat{v}_j^2\hat{w}_k^2 + \hat{w}_k^2\hat{u}_i^2}} \xrightarrow{d.} N(0, 1).$$

Finally, by (5.105), with probability at least $1 - Cp^{-3}$,

$$\big|(\hat{\lambda}-\lambda)\frac{\hat{u}_i\hat{v}_j\hat{w}_k}{\sqrt{\hat{u}_i^2\hat{v}_j^2+\hat{v}_j^2\hat{w}_k^2+\hat{w}_k^2\hat{u}_i^2}}\big| \leqslant C_2(\frac{p}{\lambda}+\sqrt{\log(p)})|w_k| \ll C_2, \tag{5.207}$$

i.e.,

$$\left|(\hat{\lambda}-\lambda)\frac{\hat{u}_i\hat{v}_j\hat{w}_k}{\sqrt{\hat{u}_i^2\hat{v}_j^2+\hat{v}_j^2\hat{w}_k^2+\hat{w}_k^2\hat{u}_i^2}}\right| \xrightarrow{p.} 0.$$

Therefore, we conclude that

$$\frac{\hat{\mathcal{T}}_{ijk}-\mathcal{T}_{ijk}}{\sqrt{\hat{u}_i^2\hat{v}_j^2+\hat{v}_j^2\hat{w}_k^2+\hat{w}_k^2\hat{u}_i^2}} \xrightarrow{d.} N(0,1).$$

5.2.8 Proof of Theorem 3.5.2

Without loss of generality, we assume $\sigma = 1$. We discuss in four scenarios:

(1). $|u_i|, |v_j|, |w_k| \geqslant (\log(p))^{1/8}\lambda^{-1}$. By Theorem 3.5.1,

$$\lim_{p\to\infty}\mathbb{P}\left(|\mathcal{T}_{ijk}-T_{ijk}| \leqslant z_{\alpha/2}\sqrt{\hat{u}_i^2\hat{v}_j^2+\hat{v}_j^2\hat{w}_k^2+\hat{w}_k^2\hat{u}_i^2}\right) = 1-\alpha.$$

Therefore (3.20) holds.

(2). Exactly two of $|u_i|, |v_j|, |w_k| \geqslant (\log(p))^{1/8}\lambda^{-1}$. Without loss of generality, we assume $|v_j|, |w_k| \geqslant (\log(p))^{1/8}\lambda^{-1}$. By the essentially same proof of (5.206), we have

$$\lambda\frac{\hat{u}_i\hat{v}_j\hat{w}_k - u_iv_jw_k}{\sqrt{u_i^2v_j^2+v_j^2w_k^2+w_k^2u_i^2}} \xrightarrow{d.} N(0,1).$$

(If $u_i = 0$, then immediately we have $\lambda\frac{\hat{u}_i\hat{v}_j\hat{w}_k - u_iv_jw_k}{\sqrt{u_i^2v_j^2+v_j^2w_k^2+w_k^2u_i^2}} = \frac{\hat{v}_j\hat{w}_k}{v_jw_k}\cdot\lambda\hat{u}_i \xrightarrow{d.} N(0,1)$.)
If $|u_i| \geqslant (\log(p))^{1/16}\lambda^{-1}$, then by (3.17), $\hat{u}_i/u_i, \hat{v}_j/v_j, \hat{w}_k/v_k \xrightarrow{p.} 1$. Therefore, $(\hat{u}_i^2\hat{v}_j^2+\hat{v}_j^2\hat{w}_k^2+\hat{w}_k^2\hat{u}_i^2)/(u_i^2v_j^2+v_j^2w_k^2+w_k^2u_i^2) \xrightarrow{p.} 1$. If $|u_i| < (\log(p))^{1/16}\lambda^{-1}$,

then (3.17) shows that $\hat{u}_i/((\log(p))^{1/16}\lambda^{-1}) \xrightarrow{p.} 0$. Thus

$$\frac{\hat{u}_i^2\hat{v}_j^2+\hat{v}_j^2\hat{w}_k^2+\hat{w}_k^2\hat{u}_i^2}{u_i^2v_j^2+v_j^2w_k^2+w_k^2u_i^2}=\frac{\hat{u}_i^2\hat{v}_j^2+\hat{v}_j^2\hat{w}_k^2+\hat{w}_k^2\hat{u}_i^2}{\hat{v}_j^2\hat{w}_k^2}\cdot\frac{v_j^2w_k^2}{u_i^2v_j^2+v_j^2w_k^2+w_k^2u_i^2}\cdot\frac{\hat{v}_j^2\hat{w}_k^2}{v_j^2w_k^2}\xrightarrow{p.} 1.$$

As a consequence,

$$\lambda\frac{\hat{u}_i\hat{v}_j\hat{w}_k-u_iv_jw_k}{\sqrt{\hat{u}_i^2\hat{v}_j^2+\hat{v}_j^2\hat{w}_k^2+\hat{w}_k^2\hat{u}_i^2}}\xrightarrow{d.} N(0,1). \tag{5.208}$$

By combining (5.207) and (5.208) together, we have

$$\frac{\hat{\mathcal{T}}_{ijk}-\mathcal{T}_{ijk}}{\sqrt{\hat{u}_i^2\hat{v}_j^2+\hat{v}_j^2\hat{w}_k^2+\hat{w}_k^2\hat{u}_i^2}}\xrightarrow{d.} N(0,1),$$

which indicates that (3.20) holds.

(3). At least two of $|u_i|, |v_j|, |w_k| < (\log(p))^{1/8}\lambda^{-1}$. Without loss of generality, we assume $|v_j|, |w_k| < (\log(p))^{1/8}\lambda^{-1}$. By (3.17),

$$\frac{|\hat{v}_j|}{(\log(p))^{1/6}\lambda^{-1}}\xrightarrow{p.} 0, \quad \frac{|\hat{w}_k|}{(\log(p))^{1/6}\lambda^{-1}}\xrightarrow{p.} 0.$$

By (5.105), we have $\hat{\lambda}/\lambda \xrightarrow{p.} 1$. Then

$$\begin{aligned}&\left|\frac{\hat{\mathcal{T}}_{ijk}}{\sqrt{s(\hat{u}_i^2)s(\hat{v}_j^2)+s(\hat{v}_j^2)s(\hat{w}_k^2)+s(\hat{w}_k^2)s(\hat{u}_i^2)}}\right|\\ =&\frac{|\hat{\lambda}\hat{u}_i\hat{v}_j\hat{w}_k|}{\sqrt{s(\hat{u}_i^2)s(\hat{v}_j^2)+s(\hat{v}_j^2)s(\hat{w}_k^2)+s(\hat{w}_k^2)s(\hat{u}_i^2)}}\\ \leqslant&\hat{\lambda}|\hat{w}_k|\frac{|\hat{v}_k|}{\sqrt{s(\hat{v}_k^2)}}\leqslant(\log(p))^{-1/6}\frac{\hat{\lambda}}{\lambda}\cdot\frac{|\hat{v}_k|}{(\log(p))^{1/6}\lambda^{-1}}\frac{|\hat{w}_k|}{(\log(p))^{1/6}\lambda^{-1}}\xrightarrow{p.} 0,\end{aligned}$$

and

$$
\begin{aligned}
&\left|\frac{\mathcal{T}_{ijk}}{\sqrt{s(\hat{u}_i^2)s(\hat{v}_j^2)+s(\hat{v}_j^2)s(\hat{w}_k^2)+s(\hat{w}_k^2)s(\hat{u}_i^2)}}\right| \\
=&\frac{|\lambda u_i v_j w_k|}{\sqrt{s(\hat{u}_i^2)s(\hat{v}_j^2)+s(\hat{v}_j^2)s(\hat{w}_k^2)+s(\hat{w}_k^2)s(\hat{u}_i^2)}} \\
\leqslant&\lambda|w_k|\frac{|v_k|}{\sqrt{s(\hat{v}_k^2)}} \leqslant \lambda\frac{((\log(p))^{1/8}\lambda^{-1})^2}{\sqrt{\log(p)}\hat{\lambda}^{-1}} \xrightarrow{p.} 0.
\end{aligned}
$$

Therefore,

$$
\lim_{p\to\infty}\mathbb{P}(\mathcal{T}_{ijk}\in\widetilde{\mathrm{CI}}_{\alpha}(\hat{\mathcal{T}}_{ijk})) = \lim_{p\to\infty}\mathbb{P}\left(\frac{|\hat{\mathcal{T}}_{ijk}-\mathcal{T}_{ijk}|}{\sqrt{s(\hat{u}_i^2)s(\hat{v}_j^2)+s(\hat{v}_j^2)s(\hat{w}_k^2)+s(\hat{w}_k^2)s(\hat{u}_i^2)}}\leqslant z_{\alpha/2}\right) = 1.
$$

In conclusion, we have proved (3.20).

5.2.9 Proof of supporting lemmas

Lemma 5.2.4. *For any* $0 \leqslant \delta \leqslant 1$, *if either of the following inequality holds, (1)* $\|\mathcal{M}_j(\hat{\mathcal{T}}^{(0)} - \mathcal{T})\| \leqslant \delta\lambda_{\min}/2$; *(2)* $\|\hat{U}_j^{(0)\top}\hat{U}_j^{(0)\top} - U_jU_j^\top\| \leqslant \delta$; *(3)* $\|\hat{U}_j^{(0)\top}\hat{U}_j^{(0)\top} - U_jU_j^\top\|_F \leqslant \sqrt{2}\delta$; *(4)* $\|\hat{U}_j^{(0)\top}U_j\| \geqslant \sqrt{1-\delta^2}$; *(5)* $\|\hat{U}_j^{(0)\top}U_j\|_F \geqslant \sqrt{r_j - \delta^2}$, *we have* $\|\sin\Theta(\hat{U}_j^{(0)}, U_j)\| \leqslant \delta$.

Proof of Lemma 5.2.4. For simplicity, let T_j and $\hat{T}_j^{(0)}$ denote $\mathcal{M}_j(\mathcal{T})$ and $\mathcal{M}_j(\hat{\mathcal{T}}^{(0)})$, respectively. Suppose $\|\hat{T}_j^{(0)} - T_j\| \leqslant \delta\lambda_{\min}/2$. By (Zhang and Xia, 2018, Lemma 6), we have

$$\|\hat{U}_{j\perp}^{(0)}T_j\| \leqslant 2\|\hat{T}_j^{(0)} - T_j\| \leqslant \delta\lambda_{\min}$$

and consequently,

$$\big\|\sin\Theta(\hat{U}_j^{(0)}, U_j)\big\| = \|\hat{U}_{j\perp}^{(0)\top}U_j\| \leqslant \frac{\|\hat{U}_{j\perp}^{(0)\top}U_jU_j^\top T_j\|}{\sigma_{\min}(U_j^\top T_j)} = \frac{\|\hat{U}_{j\perp}^{(0)\top}T_j\|}{\lambda_{\min}} \leqslant \delta.$$

In addition, by (Cai and Zhang, 2018, Lemma 1), we have

$$\big\|\sin\Theta(\hat{U}_j^{(0)}, U_j)\big\| = \sqrt{1 - \|\hat{U}_j^{(0)\top}U_j\|^2} \leqslant \|\hat{U}_j^{(0)}\hat{U}_j^{(0)\top} - U_jU_j^\top\|$$

and

$$\big\|\sin\Theta(\hat{U}_j^{(0)}, U_j)\big\| \leqslant \big\|\sin\Theta(\hat{U}_j^{(0)}, U_j)\big\|_F = \sqrt{r_j - \|\hat{U}_j^{(0)\top}U_j\|_F^2} = \|\hat{U}_j^{(0)}\hat{U}_j^{(0)\top} - U_jU_j^\top\|_F/\sqrt{2},$$

which have finished the proof of Lemma 5.2.4. □

Proof of Lemma 3.3.1. Notice that

$$\|\mathcal{A} - \mathcal{A}\times_1\mathcal{P}_{\hat{U}_1}\times_2\mathcal{P}_{\hat{U}_2}\times_3\mathcal{P}_{\hat{U}_3}\|_F = \|\mathcal{Z} - \mathcal{Z}\times_1\mathcal{P}_{\hat{U}_1}\times_2\mathcal{P}_{\hat{U}_2}\times_3\mathcal{P}_{\hat{U}_3} + \mathcal{T} - \mathcal{T}\times_1\mathcal{P}_{\hat{U}_1}\times_2\mathcal{P}_{\hat{U}_2}\times_3\mathcal{P}_{\hat{U}_3}\|_F,$$

we have

$$\Big|\|\mathcal{A} - \mathcal{A}\times_1\mathcal{P}_{\hat{U}_1}\times_2\mathcal{P}_{\hat{U}_2}\times_3\mathcal{P}_{\hat{U}_3}\|_F - \|\mathcal{Z}\|_F\Big|$$

$$\leqslant \|\mathcal{Z} \times_1 \mathcal{P}_{\hat{U}_1} \times_2 \mathcal{P}_{\hat{U}_2} \times_3 \mathcal{P}_{\hat{U}_3}\|_F + \|\mathcal{T} - \mathcal{T} \times_1 \mathcal{P}_{\hat{U}_1} \times_2 \mathcal{P}_{\hat{U}_2} \times_3 \mathcal{P}_{\hat{U}_3}\|_F.$$

By (5.214) and (5.215), with probability at least $1 - C_1 e^{-c_1 p}$,

$$\|\mathcal{Z} \times_1 \mathcal{P}_{\hat{U}_1} \times_2 \mathcal{P}_{\hat{U}_2} \times_3 \mathcal{P}_{\hat{U}_3}\|_F = \|\hat{U}_1^\top Z_1(\hat{U}_2 \otimes \hat{U}_3)\|_F \leqslant \sqrt{r_1}\|Z_1(\hat{U}_2 \otimes \hat{U}_3)\| \leqslant C_2 \sigma \sqrt{pr}.$$

In addition, we have

$$\begin{aligned}
&\|\mathcal{T} - \mathcal{T} \times_1 \mathcal{P}_{\hat{U}_1} \times_2 \mathcal{P}_{\hat{U}_2} \times_3 \mathcal{P}_{\hat{U}_3}\|_F \\
=&\|\mathcal{T} \times_1 \mathcal{P}_{U_1} \times_2 \mathcal{P}_{U_2} \times_3 \mathcal{P}_{U_3} - \mathcal{T} \times_1 \mathcal{P}_{\hat{U}_1} \times_2 \mathcal{P}_{\hat{U}_2} \times_3 \mathcal{P}_{\hat{U}_3}\|_F \\
\leqslant&\|(\mathcal{P}_{U_1} - \mathcal{P}_{\hat{U}_1}) T_1 (\mathcal{P}_{U_2} \otimes \mathcal{P}_{U_3})\|_F + \|(\mathcal{P}_{U_2} - \mathcal{P}_{\hat{U}_2}) T_2 (\mathcal{P}_{\hat{U}_1} \otimes \mathcal{P}_{U_3})\|_F + \|(\mathcal{P}_{U_3} - \mathcal{P}_{\hat{U}_3}) T_3 (\mathcal{P}_{\hat{U}_1} \otimes \mathcal{P}_{U_2})\|_F \\
\leqslant&\big(\|\mathcal{P}_{U_1} - \mathcal{P}_{\hat{U}_1}\| + \|\mathcal{P}_{U_2} - \mathcal{P}_{\hat{U}_2}\| + \|\mathcal{P}_{U_2} - \mathcal{P}_{\hat{U}_2}\|\big)\|\mathcal{T}\|_F \\
\leqslant&C_2 \frac{\sqrt{p}\sigma}{\lambda_{\min}} \cdot \sqrt{r}\kappa_0 \lambda_{\min} = C_2 \kappa_0 \sigma \sqrt{pr}
\end{aligned}$$

with probability at least $1 - C_1 e^{-c_1 p}$. Therefore, with probability at least $1 - C_1 e^{-c_1 p}$, we have

$$\Big| \|\mathcal{A} - \mathcal{A} \times_1 \mathcal{P}_{\hat{U}_1} \times_2 \mathcal{P}_{\hat{U}_2} \times_3 \mathcal{P}_{\hat{U}_3}\|_F - \|\mathcal{Z}\|_F \Big| \leqslant C_2 \kappa_0 \sigma \sqrt{pr}. \tag{5.209}$$

By (Laurent and Massart, 2000, Lemma 1),

$$\mathbb{P}\left(\left| \frac{\|\mathcal{Z}\|_F^2}{\sigma^2} - p_1 p_2 p_3 \right| \geqslant C_2 (\sqrt{p_1 p_2 p_3} \sqrt{\log(p)} + \log(p)) \right) \leqslant p^{-3}.$$

As a consequence, with probability at least $1 - p^{-3}$,

$$\Big| \|\mathcal{Z}\|_F - \sqrt{p_1 p_2 p_3}\sigma \Big| \leqslant C_2 \sqrt{\log(p)}\sigma.$$

Combing (5.209) and the previous inequality together, we know that with probability at least $1 - C_1 p^{-3}$,

$$|\hat{\sigma}/\sigma - 1| \leqslant C_2(\kappa_0 \sqrt{r} p^{-1} + p^{-3/4}\sqrt{\log(p)})$$

and

$$|\hat\sigma^2/\sigma^2{-}1| = |\hat\sigma/\sigma{-}1||\hat\sigma/\sigma{+}1| \leqslant 2|\hat\sigma/\sigma{-}1|{+}|\hat\sigma/\sigma{-}1|^2 \leqslant C_2(\kappa_0\sqrt{r}p^{-1}{+}p^{-3/4}\sqrt{\log(p)}).$$

□

Proof of Lemma 5.2.1. By definition, $\|\mathfrak{E}_1\| \leqslant \|\mathfrak{J}_1\| + \|\mathfrak{J}_2\| + \|\mathfrak{J}_3\| + \|\mathfrak{J}_4\|$.

$$\|\mathfrak{E}_1\| \leqslant \|\mathfrak{J}_1\| + \|\mathfrak{J}_2\| + \|\mathfrak{J}_3\| + \|\mathfrak{J}_4\|. \tag{5.210}$$

We first proved the upper bound for $\|\mathfrak{J}_1\|$. By the definition of $\|\mathfrak{J}_1\|$,

$$\|\mathfrak{J}_1\| \leqslant \left\|T_1(\mathcal{P}_{\hat{U}_2^{(1)}} \otimes \mathcal{P}_{\hat{U}_3^{(1)}})Z_1^\top\right\| \leqslant \left\|T_1\right\|\|(\mathcal{P}_{\hat{U}_2^{(1)}} \otimes \mathcal{P}_{\hat{U}_3^{(1)}})Z_1^\top\right\| \leqslant \kappa_0\lambda_{\min}\left\|Z_1(\hat{U}_2^{(1)} \otimes \hat{U}_3^{(1)})\right\|. \tag{5.211}$$

For any fixed matrices $X \in \mathbb{R}^{p_2\times r_2}, Y \in \mathbb{R}^{p_3 \times r_3}$ satisfying $\|X\|, \|Y\| \leqslant 1$,

$$\mathbb{P}\left(\|Z_1(X \otimes Y)\| \geqslant C_2\sqrt{pr}\right) \leqslant C_1 e^{-c_1 pr}. \tag{5.212}$$

Let $\mathcal{X}_{p_k,r_k} = \{X \in \mathbb{R}^{p_k\times r_k} : \|X\| \leqslant 1\}$. By (Zhang and Xia, 2018, Lemma 7), there exists an 1/4-net $\bar{\mathcal{X}}_{p_k,r_k}$ with cardinality at most $9^{p_k r_k}$ for $\mathcal{X}_{p_k,r_k}$. That is, for any $X \in \mathcal{X}_{p_k,r_k}$, there exists $X' \in \mathcal{X}_{p_k,r_k}$ such that $\|X'-X\| \leqslant 1/4$. For any $X \in \mathcal{X}_{p_2,r_2}$ and $Y \in \mathcal{X}_{p_3,r_3}$, let $X' \in \bar{\mathcal{X}}_{p_2,r_2}$ and $Y' \in \bar{\mathcal{X}}_{p_3,r_3}$ satisfying $\|X - X'\| \leqslant 1/4, \|Y - Y'\| \leqslant 1/4$. Then

$$\begin{aligned}
&\|Z_1(X \otimes Y)\| \\
\leqslant&\|Z_1(X' \otimes Y')\| + \|Z_1((X - X') \otimes Y)\| + \|Z_1(X \otimes (Y - Y'))\| + \|Z_1((X - X') \otimes (Y - Y'))\| \\
\leqslant&\|Z_1(X' \otimes Y')\| + \frac{3}{4}\sup_{\substack{X\in\mathbb{R}^{p_2\times r_2}, Y\in\mathbb{R}^{p_3\times r_3}\\ \|X\|,\|Y\|\leqslant 1}}\|Z_1(X \otimes Y)\|.
\end{aligned}$$

By taking the supremum over any $X \in \mathcal{X}_{p_2,r_2}$ and $Y \in \mathcal{X}_{p_3,r_3}$, we have

$$\sup_{\substack{X\in\mathbb{R}^{p_2\times r_2}, Y\in\mathbb{R}^{p_3\times r_3}\\ \|X\|,\|Y\|\leqslant 1}} \|Z_1(X\otimes Y)\| \leqslant 4 \sup_{X'\in\bar{\mathcal{X}}_{p_2,r_2}, Y'\in\bar{\mathcal{X}}_{p_3,r_3}} \|Z_1(X'\otimes Y')\|.$$

The union bound shows that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\substack{X\in\mathbb{R}^{p_2\times r_2}, Y\in\mathbb{R}^{p_3\times r_3}\\ \|X\|,\|Y\|\leqslant 1}} \|Z_1(X\otimes Y)\| \geqslant C_2\sqrt{pr}\right)\\
\leqslant&\mathbb{P}\left(\sup_{X'\in\bar{\mathcal{X}}_{p_2,r_2}, Y'\in\bar{\mathcal{X}}_{p_3,r_3}} \|Z_1(X'\otimes Y')\| \geqslant C_2\sqrt{pr}\right)\\
\leqslant&\sum_{X'\in\bar{\mathcal{X}}_{p_2,r_2}, Y'\in\bar{\mathcal{X}}_{p_3,r_3}} \mathbb{P}\left(\|Z_1(X'\otimes Y')\| \geqslant C_2\sqrt{pr}\right)\\
\leqslant&C_1e^{-c_1pr}.
\end{aligned}$$

By (Cai and Zhang, 2018, Lemma 1), with probability at least $1 - C_1e^{-c_1p}$,

$$\|\mathcal{P}_{U_k}^{\perp}\hat{U}_k^{(1)}\| = \|U_{k\perp}^{\top}\hat{U}_k^{(1)}\| \leqslant \|\hat{U}_k^{(1)}\hat{U}_k^{(1)\top} - U_kU_k^{\top}\| \leqslant C_2\sqrt{p}/\lambda_{\min}, \quad 1\leqslant k\leqslant 3. \tag{5.213}$$

Therefore, with probability at least $1 - C_1e^{-c_1p}$,

$$\begin{aligned}
&\left\|Z_1(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right\|\\
=&\left\|Z_1(\mathcal{P}_{U_2\otimes U_3}+\mathcal{P}_{U_{2\perp}\otimes U_3}+\mathcal{P}_{U_2\otimes U_{3\perp}}+\mathcal{P}_{U_{2\perp}\otimes U_{3\perp}})(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right\|\\
\leqslant&\left\|Z_1\mathcal{P}_{U_2\otimes U_3}(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right\|+\left\|Z_1\mathcal{P}_{U_{2\perp}\otimes U_3}(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right\|\\
&+\left\|Z_1\mathcal{P}_{U_2\otimes U_{3\perp}}(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right\|+\left\|Z_1\mathcal{P}_{U_{2\perp}\otimes U_{3\perp}}(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right\|\\
=&\left\|Z_1(U_2\otimes U_3)(U_2\otimes U_3)^{\top}(\hat{U}_2^{(1)}\otimes\hat{U}_3^{(1)})\right\|+\left\|Z_1\left((\mathcal{P}_{U_2}^{\perp}\hat{U}_2^{(1)})\otimes(\mathcal{P}_{U_3}\hat{U}_3^{(1)})\right)\right\|\\
&+\left\|Z_1\left((\mathcal{P}_{U_2}\hat{U}_2^{(1)})\otimes(\mathcal{P}_{U_3}^{\perp}\hat{U}_3^{(1)})\right)\right\|+\left\|Z_1\left((\mathcal{P}_{U_2}^{\perp}\hat{U}_2^{(1)})\otimes(\mathcal{P}_{U_3}^{\perp}\hat{U}_3^{(1)})\right)\right\|\\
\leqslant&\|Z_1(U_2\otimes U_3)\|+C_2\sqrt{pr}\left\|\mathcal{P}_{U_2}^{\perp}\hat{U}_2^{(1)}\right\|\left\|\mathcal{P}_{U_3}\hat{U}_3^{(1)}\right\|
\end{aligned}$$

$$
\begin{aligned}
&+ C_2\sqrt{pr}\left\|\mathcal{P}_{U_2}\hat{U}_2^{(1)}\right\|\left\|\mathcal{P}_{U_3}^{\perp}\hat{U}_3^{(1)}\right\| + C_2\sqrt{pr}\left\|\mathcal{P}_{U_2}^{\perp}\hat{U}_2^{(1)}\right\|\left\|\mathcal{P}_{U_3}^{\perp}\hat{U}_3^{(1)}\right\| \\
\leqslant\ & \|Z_1(U_2 \otimes U_3)\| + C_2\sqrt{pr}\sqrt{p}\lambda_{\min}^{-1} + C_2\sqrt{pr}\sqrt{p}\lambda_{\min}^{-1} + C_2\sqrt{pr}p\lambda_{\min}^{-2} \\
\leqslant\ & \|Z_1(U_2 \otimes U_3)\| + C_2\sqrt{pr}\sqrt{p}\lambda_{\min}^{-1} \\
\leqslant\ & \|Z_1(U_2 \otimes U_3)\| + C_2\sqrt{p}. \qquad (5.214)
\end{aligned}
$$

By the Gaussian concentration inequality,

$$\mathbb{P}\left(\|Z_1(U_2 \otimes U_3)\| \geqslant C_3\sqrt{p}\right) \leqslant C_1 e^{-c_1 p}. \qquad (5.215)$$

(5.211), (5.214) and (5.215) together imply that

$$\mathbb{P}\left(\|\mathfrak{J}_1\| \geqslant C_2\kappa_0\lambda_{\min}\sqrt{p}\right) \leqslant C_1 e^{-c_1 p}. \qquad (5.216)$$

Since $\mathfrak{J}_2 = \mathfrak{J}_1^{\top}$, we also have

$$\mathbb{P}\left(\|\mathfrak{J}_2\| \geqslant C_2\kappa_0\lambda_{\min}\sqrt{p}\right) \leqslant C_1 e^{-c_1 p}. \qquad (5.217)$$

For $\mathfrak{J}_3$, by definition,

$$\|\mathfrak{J}_3\| = \|Z_1(\hat{U}_2^{(1)} \otimes \hat{U}_3^{(1)})\|^2. \qquad (5.218)$$

Combining (5.218), (5.214) and (5.215) together, we have

$$\mathbb{P}\left(\|\mathfrak{J}_3\| \geqslant C_2 p\right) \leqslant C_1 e^{-c_1 p}. \qquad (5.219)$$

Then, we consider $\mathfrak{J}_4$. By (5.213), with probability at least $1 - C_1 e^{-c_1 p}$,

$$
\begin{aligned}
\|\mathfrak{J}_4 \leqslant\ & \left\|T_1((\mathcal{P}_{\hat{U}_2^{(1)}} - \mathcal{P}_{U_2}) \otimes \mathcal{P}_{\hat{U}_3^{(1)}})T_1^{\top}\right\| + \left\|T_1(\mathcal{P}_{U_2} \otimes (\mathcal{P}_{\hat{U}_3^{(1)}} - \mathcal{P}_{U_3})T_1^{\top}\right\| \\
=\ & \left\|U_1 G_1((U_2^{\top}(\mathcal{P}_{\hat{U}_2^{(1)}} - \mathcal{P}_{U_2})U_2) \otimes (U_3^{\top}\mathcal{P}_{\hat{U}_3^{(1)}}U_3))G_1^{\top}U_1^{\top}\right\| \\
& + \left\|U_1 G_1(\mathcal{P}_{U_2} \otimes (U_3^{\top}(\mathcal{P}_{\hat{U}_3^{(1)}} - \mathcal{P}_{U_3})U_3))G_1^{\top}U_1^{\top}\right\|
\end{aligned}
$$

$$
\begin{aligned}
=&\left\|U_1G_1((U_2^\top\mathcal{P}^\perp_{\hat{U}_2^{(1)}}U_2)\otimes(U_3^\top\mathcal{P}_{\hat{U}_3^{(1)}}U_3))G_1^\top U_1^\top\right\| \\
&+\left\|U_1G_1(\mathcal{P}_{U_2}\otimes(U_3^\top\mathcal{P}^\perp_{\hat{U}_3^{(1)}}U_3))G_1^\top U_1^\top\right\| \\
\leqslant&\|G_1\|^2\left\|U_2^\top\hat{U}_{2\perp}^{(1)}\right\|^2+\|G_1\|^2\left\|U_3^\top\hat{U}_{3\perp}^{(1)}\right\|^2 \\
\leqslant&C_2\kappa_0^2\lambda_{\min}^2(\sqrt{p}/\lambda_{\min})^2 \\
=&C_2\kappa_0^2p.
\end{aligned}
$$

Therefore, by (5.210), (5.216), (5.217) and (5.219) and notice that $\lambda_{\min}\geqslant C_2\kappa_0\sqrt{p}$, we conclude with

$$\mathbb{P}\big(\|\mathfrak{E}_1\|\geqslant C_2\kappa_0\lambda_{\min}\sqrt{p}\big)\leqslant C_1e^{-c_1p}.$$

□

Proof of Lemma 5.2.3. Consider the SVD decomposition $\hat{U}^\top U=LSW^\top$, where $L,W\in\mathbb{O}_r$, and $S\in\mathbb{R}^{r\times r}=\operatorname{diag}(s_1,\dots,s_r)$ is a diagonal matrix with diagonal entries $1\geqslant s_1\geqslant\cdots\geqslant s_r\geqslant 0$. By setting $R=LW^\top$, we have $\hat{U}^\top U-R=L(S-I_r)W^\top$. Therefore, $\left\|\hat{U}^\top U-R\right\|=\|S-I_r\|$. Since $|x-1|\leqslant|x^2-1|$ for all $x\geqslant 0$, we have

$$\left\|\hat{U}^\top U-R\right\|\leqslant\left\|S^2-I_r\right\|=\left\|\hat{U}^\top UU^\top\hat{U}-I_r\right\|=\left\|\hat{U}^\top U_\perp U_\perp^\top\hat{U}\right\|=\left\|U_\perp^\top\hat{U}\right\|^2. \tag{5.220}$$

For $\left\|\hat{U}^\top U-R\right\|_F$, we have

$$\left\|\hat{U}^\top U-R\right\|_F\leqslant\sqrt{r}\left\|\hat{U}^\top U-R\right\|\leqslant\sqrt{r}\left\|U_\perp^\top\hat{U}\right\|^2 \tag{5.221}$$

and

$$\left\|\hat{U}^\top U-R\right\|_F\leqslant\left\|S^2-I_r\right\|_F=\left\|\hat{U}^\top UU^\top\hat{U}-I_r\right\|_F=\left\|\hat{U}^\top U_\perp U_\perp^\top\hat{U}\right\|_F\leqslant\left\|U_\perp^\top\hat{U}\right\|_F^2, \tag{5.222}$$

which have finish the proof of Lemma 5.2.3. □

Lemma 5.2.5. *Under tensor regression model (3.2) with* $\mathcal{X}(i_1,i_2,i_3)\overset{i.i.d.}{\sim}N(0,1)$, $\operatorname{Var}(\xi_i)=\sigma^2$ *and* $\|\xi_i\|_{\psi_2}\leqslant C\sigma$ *for some constant* $C>0$, *if* $\|\tilde{\mathcal{T}}-\mathcal{T}\|_F^2\leqslant C_2pr_{\max}\sigma^2/n$, $n(\lambda_{\min}/\sigma)^2\geqslant$

$C_0(p^{3/2} \vee \kappa_0^4 p r_{\max}^2)$ *and* $n \geqslant C_0(p^{3/2} \vee \kappa_0^2 p r_{\max}^3)$ *for some constants* $C_0, C_2 > 0$*, then there exists some constants* $C_1, c_1, C_3 > 0$ *such that with probability at least* $1 - C_1 e^{-c_1 p}$*,*

$$\left\|\sin\Theta(\hat{U}_j^{(1)}, U_j)\right\| \leqslant C_3\sqrt{p/n}\sigma/\lambda_{\min}, \quad \forall j = 1, 2, 3,$$

where $U_j^{(1)}$ *is the one-step alternating minimization defined in Algorithm 2.*

Proof of Lemma 5.2.5. Without loss of generality, we assume $\sigma = 1$. By Assumption 3.3.2 and (Zhang and Xia, 2018, Lemma 6), with probability $1 - C_1 e^{-c_1 p}$,

$$\|\hat{U}_{1\perp}^{(0)\top} T_1\|_F \leqslant 2\|T_1 - \hat{T}_1^{(0)}\|_F \leqslant C\frac{pr}{n}.$$

By (Cai and Zhang, 2018, Lemma 1), we get with probability $1 - C_1 e^{-c_1 p}$ that

$$\inf_{O \in \mathbb{O}_{r_1}} \|\hat{U}_1^{(0)} - U_1 O\|_F \leqslant \|\hat{U}_1^{(0)}\hat{U}_1^{(0)\top} - U_1 U_1^\top\|_F = \sqrt{2}\|\sin\Theta(\hat{U}_1^{(0)}, U_1)\|_F = \sqrt{2}\|\hat{U}_{1\perp}^{(0)\top} U_1\|_F$$
$$\leqslant \sqrt{2}\frac{\|\hat{U}_{1\perp}^{(0)\top} T_1\|_F}{\|G_1(U_2^\top \otimes U_3^\top)\|} \leqslant C\frac{\sqrt{pr/n}}{\|G_1\|} \leqslant C\frac{\sqrt{pr/n}}{\lambda_{\min}}.$$

Similarly, with the same probability, we get

$$\|\hat{U}_2^{(0)}\hat{U}_2^{(0)\top} - U_2 U_2^\top\|_F, \|\hat{U}_3^{(0)}\hat{U}_3^{(0)\top} - U_3 U_3^\top\|_F \leqslant C\frac{\sqrt{pr/n}}{\lambda_{\min}}.$$

Based on the two equations above, we can prove Lemma 5.2.5 by similar proof of (5.119). □

Lemma 5.2.6. *There exists an* ϵ*-net* $\bar{\mathbb{O}}_{p,r} = \{U^{(j)} \in \mathbb{O}_{p,r}, 1 \leqslant j \leqslant N\}$ *in* $\|\cdot\|$ *norm with cardinality* $N \leqslant ((4+\epsilon)/\epsilon)^{pr}$ *for* $\mathbb{O}_{p,r}$*. That is, for any* $U \in \mathbb{O}_{p,r}$*, there exists* $j \in [N]$ *such that* $\|U - U^{(j)}\| \leqslant \epsilon$*.*

Proof of Lemma 5.2.6. By (Zhang and Xia, 2018, Lemma 7), for $\mathcal{U}_{p,r} = \{U \in \mathbb{R}^{p\times r}, \|U\| \leqslant 1\}$, there exists an $\epsilon/2$-net $\bar{\mathcal{U}}_{p,r} = \{\bar{U}^{(j)} \in \mathbb{R}^{p\times r}, \|\bar{U}^{(j)}\| \leqslant 1, 1 \leqslant j \leqslant N\}$ in $\|\cdot\|$ norm with $N \leqslant ((4+\epsilon)/\epsilon)^{pr}$ for $\mathcal{U}_{p,r}$. Let $U^{(j)} \in \arg\min_{U \in \mathbb{O}_{p,r}} \|\bar{U}^{(j)} - U\|$,

$1 \leqslant j \leqslant N$. For any $U \in \mathbb{O}_{p,r}$, there exists $\bar{U}^{(j)}$ such that $\|\bar{U}^{(j)} - U\| \leqslant \epsilon/2$. Then $\|U^{(j)} - U\| \leqslant \|\bar{U}^{(j)} - U\| + \|\bar{U}^{(j)} - U^{(j)}\| \leqslant 2\|\bar{U}^{(j)} - U\| \leqslant \epsilon$. □

Lemma 5.2.7. *Suppose $Z \in \mathbb{R}^{p\times q}$ is a matrix with independent zero-mean σ-sub-Gaussian entries. $A \in \mathbb{R}^{m\times p}, B \in \mathbb{R}^{q\times n}$ satisfy $\|A\|, \|B\| \leqslant 1$, $m \leqslant p, n \leqslant q$. Then*

$$\mathbb{P}\left(\|AZB\| \geqslant 2\sigma\sqrt{m+t}\right) \leqslant 2\cdot 5^n \exp\left[-c\min\left(\frac{t^2}{m}, t\right)\right]. \tag{5.223}$$

$$\mathbb{P}\left(\|AZB\|_F \geqslant \sigma\sqrt{mn+t}\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{mn}, t\right)\right]. \tag{5.224}$$

Proof of Lemma 5.2.7. Without loss of generality, assume $\sigma = 1$. For fixed $x \in \mathbb{R}^n$ satisfying $\|x\|_2 = 1$, we have $AZBx = \mathrm{vec}(AZBx) = (x^\top B^\top \otimes A)\mathrm{vec}(Z)$. Since Z_{ij} is 1-sub-Gaussian, we know that $\mathrm{Var}(Z_{ij}) \leqslant 1$. In addition,

$$\begin{aligned}
\mathbb{E}\|(x^\top B^\top \otimes A)\mathrm{vec}(Z)\|_2^2 =& \mathbb{E}\left[\mathrm{trace}\left(\mathrm{vec}(Z)^\top (x^\top B^\top \otimes A)^\top (x^\top B^\top \otimes A)\mathrm{vec}(Z)\right)\right]\\
=&\mathrm{trace}\left[\mathbb{E}\left((x^\top B^\top \otimes A)^\top (x^\top B^\top \otimes A)\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)\right]\\
=&\mathrm{trace}\left[(x^\top B^\top \otimes A)^\top (x^\top B^\top \otimes A)\mathbb{E}\left(\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)\right]\\
\leqslant&\mathrm{trace}\left((x^\top B^\top \otimes A)^\top (x^\top B^\top \otimes A)\right)\\
=&\left\|x^\top B^\top \otimes A\right\|_F^2 = \|Bx\|_2^2\|A\|_F^2 \leqslant \|x\|_2^2\|A\|_F^2\\
\leqslant& m.
\end{aligned} \tag{5.225}$$

The first inequality holds since $\mathbb{E}\left(\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)$ is a diagonal matrix with diagonal entries $\mathrm{Var}(Z_{ij}) \leqslant 1$; the last inequality is due to $\|A\|_F \leqslant \min\{m, p\}\|A\|_2 \leqslant m$. By Hanson-Wright inequality, we have

$$\mathbb{P}\left(\|AZBx\|_2^2 - m \geqslant t\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{\|(Bxx^\top B^\top)\otimes(A^\top A)\|_F^2}, \frac{t}{\|(Bxx^\top B^\top)\otimes(A^\top A)\|}\right)\right].$$

Since $\|x\|_2 = 1$ and $\|A\|, \|B\| \leqslant 1$,

$$\begin{aligned}\|(Bxx^\top B^\top) \otimes (A^\top A)\|_F^2 =& \|Bxx^\top B^\top\|_F^2 \|A^\top A\|_F^2 = (x^\top B^\top Bx)^2 \|A^\top A\|_F^2 \\ \leqslant& (x^\top x)^2 \|A^\top A\|_F^2 = \sum_{i=1}^{\min\{m,p\}} \sigma_i^4(A) \leqslant m,\end{aligned}$$

$$\|(Bxx^\top B^\top) \otimes (A^\top A)\| \leqslant \|Bxx^\top B^\top\| \|A^\top A\| \leqslant \|xx^\top\| \|A^\top A\| \leqslant 1.$$

Thus, for fixed x satisfying $\|x\|_2 = 1$, we have

$$\mathbb{P}\left(\|AZBx\|_2^2 \geqslant m + t\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{m}, t\right)\right]. \tag{5.226}$$

By Vershynin (2010)[Lemma 5.2], there exists $\mathcal{N}_{1/2}$, a 1/2-net of $\{x \in \mathbb{R}^n : \|x\|_2 = 1\}$, such that $\left|\mathcal{N}_{1/2}\right| \leqslant 5^n$. The union bound, Vershynin (2010)[Lemma 5.2] and (5.226) together imply that

$$\mathbb{P}\left(\|AZB\| \geqslant 2\sqrt{m+t}\right) \leqslant \mathbb{P}\left(\max_{x \in \mathcal{N}_{1/2}} \|AZBx\|_2 \geqslant \sqrt{m+t}\right) \leqslant 2 \cdot 5^n \exp\left[-c\min\left(\frac{t^2}{m}, t\right)\right].$$

For $\|AZB\|_F$, note that $AZB = (B^\top \otimes A)\mathrm{vec}(Z)$, Similarly to (5.225), we have

$$\begin{aligned}\mathbb{E}\|(B^\top \otimes A)\mathrm{vec}(Z)\|_2^2 =& \mathbb{E}\left[\mathrm{vec}(Z)^\top (B^\top \otimes A)^\top (B^\top \otimes A)\mathrm{vec}(Z)\right] \\ =& \mathbb{E}\mathrm{trace}\left[\mathrm{vec}(Z)^\top (B^\top \otimes A)^\top (B^\top \otimes A)\mathrm{vec}(Z)\right] \\ =& \mathrm{trace}\mathbb{E}\left[(B^\top \otimes A)^\top (B^\top \otimes A)\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right] \\ =& \mathrm{trace}\left[(B^\top \otimes A)^\top (B^\top \otimes A)\mathbb{E}\left(\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)\right] \\ \leqslant& \mathrm{trace}\left[(B^\top \otimes A)^\top (B^\top \otimes A)\right] \\ =& \|B^\top \otimes A\|_F^2 = \|B\|_F^2 \|A\|_F^2 \\ \leqslant& mn.\end{aligned}$$

By Hanson-Wright inequality, we have

$$\mathbb{P}\left(\|AZB\|_F^2 - mn \geqslant t\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{\|(BB^\top)\otimes(A^\top A)\|_F^2}, \frac{t}{\|(BB^\top)\otimes(A^\top A)\|}\right)\right].$$

Since $\|A\|, \|B\| \leqslant 1$, we have

$$\|(BB^\top)\otimes(A^\top A)\|_F = \sqrt{\|A^\top A\|_F^2\|BB^\top\|_F^2} = \sqrt{\sum_{i=1}^{\min\{m,p\}}\sigma_i^4(A)\sum_{i=1}^{\min\{q,n\}}\sigma_i^4(B)} \leqslant \sqrt{mn},$$

$$\|(BB^\top)\otimes(A^\top A)\| \leqslant 1.$$

Therefore,

$$\mathbb{P}\left(\|AZB\|_F^2 \geqslant mn + t\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{mn}, t\right)\right].$$

□

Lemma 5.2.8. *For the class of low-rank tensors under the Frobenius norm* $\mathcal{X}_{\mathbf{p},\mathbf{r}} = \{\mathcal{A} \in \mathbb{R}^{p_1\times p_2\times p_3} : \text{rank}(\mathcal{M}_i(\mathcal{A})) \leqslant r_i, i \in [3], \|\mathcal{A}\|_F \leqslant 1\}$, *there exists*

$$\bar{\mathcal{X}}_{\mathbf{p},\mathbf{r}} = \{\mathcal{A}^{(1)}, \dots, \mathcal{A}^{(N)}\} \tag{5.227}$$

with $N \leqslant ((8+\epsilon)/\epsilon)^{r_1r_2r_3+\sum_{i=1}^3 p_ir_i}$ *satisfying* $\mathcal{A}^{(i)} \in \mathbb{R}^{p_1\times p_2\times p_3} : \|\mathcal{A}^{(i)}\|_F \leqslant 1$, *such that for all* $\mathcal{A} \in \mathcal{X}_{\mathbf{p},\mathbf{r}}$, *there exists* $i \in [N]$ *satisfying* $\|\mathcal{A}^{(i)} - \mathcal{A}\|_F \leqslant \epsilon$.

Proof of Lemma 5.2.8. By (Zhang and Xia, 2018, Lemma 7), there exist $\epsilon/4$-nets $\bar{\mathcal{X}}_{p_i,r_i}$ for $\mathcal{X}_{p_i,r_i} = \{U \in \mathbb{R}^{p_i\times r_i} : \|U\| \leqslant 1\}$ under the spectral norm with cardinality at most $((8+\epsilon)/\epsilon)^{p_ir_i}$, $i \in [3]$, and $\bar{\mathcal{X}}_{r_1,r_2r_3}$ for $\mathcal{X}_{r_1,r_2r_3} = \{B \in \mathbb{R}^{r_1\times(r_2r_3)} : \|B\|_F \leqslant 1\}$ under the Frobenius norm with cardinality at most $((8+\epsilon)/\epsilon)^{r_1r_2r_3}$. Let

$$\bar{\mathcal{X}}_{\mathbf{p},\mathbf{r}} = \{\mathcal{B} \times_1 U_1 \times_2 U_2 \times_3 U_3 : U_i \in \bar{\mathcal{X}}_{p_i,r_i}, \mathcal{M}_1(\mathcal{B}) \in \bar{\mathcal{X}}_{r_1,r_2r_3}\}.$$

For any $\mathcal{A} \in \mathcal{X}_{\mathbf{p},\mathbf{r}}$, there exist $U_i \in \mathcal{X}_{p_i,r_i}$ and $D_1 \in \mathcal{X}_{r_1,r_2r_3}$ such that $\mathcal{M}_1(\mathcal{A}) = V_1D_1(V_2^\top \otimes V_3^\top)$. Then we can find $U_i^* \in \bar{\mathcal{X}}_{p_i,r_i}$ and $\mathcal{B}^* \in \mathbb{R}^{r_1\times r_2\times r_3}$, $B_1^* = \mathcal{M}_1(\mathcal{B}^*) \in$

$\bar{\mathcal{X}}_{r_1,r_2r_3}$, and $\mathcal{B}^* \times_1 U_1^* \times_2 U_2^* \times_3 U_3^* \in \bar{\mathcal{X}}_{\mathbf{p},\mathbf{r}}$ satisfying

$$\begin{aligned}
&\|\mathcal{A} - \mathcal{B}^* \times_1 U_1^* \times_2 U_2^* \times_3 U_3^*\|_F \\
=&\|V_1 D_1 (V_2^\top \otimes V_3^\top) - U_1^* B_1^* (U_2^{*\top} \otimes U_3^{*\top})\|_F \\
\leqslant&\|(V_1 - U_1^*) D_1 (V_2^\top \otimes V_3^\top)\|_F + \|U_1^*(D_1 - B_1^*)(V_2^\top \otimes V_3^\top)\|_F + \|U_1^* B_1^*((V_2 - U_2^*)^\top \otimes V_3^\top)\|_F \\
&+ \|U_1^* B_1^* (U_2^{*\top} \otimes (V_3 - U_3^*)^\top)\|_F \\
\leqslant&\|V_1 - U_1^*\|\|V_2\|\|V_3\|\|D_1\|_F + \|U_1^*\|\|V_2\|\|V_3\|\|D_1 - B_1^*\|_F + \|U_1^*\|\|V_2 - U_2^*\|\|V_3\|\|B_1^*\|_F \\
&+ \|U_1^*\|\|U_2^*\|\|V_3 - U_3^*\|\|B_1^*\|_F \\
\leqslant&\frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} + \frac{\epsilon}{4} = \epsilon.
\end{aligned}$$

Notice that $|\bar{\mathcal{X}}_{\mathbf{p},\mathbf{r}}| \leqslant |\bar{\mathcal{X}}_{p_1,r_1}||\bar{\mathcal{X}}_{p_2,r_2}||\bar{\mathcal{X}}_{p_3,r_3}||\bar{\mathcal{X}}_{r_1,r_2r_3}| \leqslant ((8+\epsilon)/\epsilon)^{r_1r_2r_3 + \sum_{i=1}^3 p_i r_i}$, we have finished the proof of Lemma 5.2.8. □

Lemma 5.2.9. *(1) Suppose* $X \in \mathbb{R}^{p_1 \times p_2}, X(i_1, i_2) \overset{\text{i.i.d.}}{\sim} N(0,1)$ *and* $X_1, \dots, X_n$ *are i.i.d. copies of* X*. Then there exist two universal constants* $C, C_1 > 0$ *such that for any fixed* $U \in \mathbb{O}_{p_1,r_1}, V \in \mathbb{O}_{p_2,r_2}$ *and* $\Delta \in \mathbb{R}^{p_1 \times p_2}$,

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n \langle X_i, \Delta\rangle U^\top X_i V - U^\top \Delta V\right\| \geqslant C\|\Delta\|_F t\right) \leqslant 2 \cdot 7^{r_1+r_2} e^{-C_1 \min\{nt^2, nt\}}. \tag{5.228}$$

(2) Suppose $\mathcal{X} \in \mathbb{R}^{p_1 \times p_2 \times p_3}, \mathcal{X}(i_1, i_2, i_3) \overset{\text{i.i.d.}}{\sim} N(0,1)$ *and* $\mathcal{X}_1, \dots, \mathcal{X}_n$ *are i.i.d. copies of* $\mathcal{X}$*. Then*

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\substack{U_i \in \mathbb{R}^{p_i \times r_i}, \|U_i\| \leqslant 1 \\ \mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times p_3}, \|\mathcal{A}\|_F \leqslant 1 \\ \text{rank}(\mathcal{A}) \leqslant (\bar{r}_1, \bar{r}_2, \bar{r}_3)}} \left\|\frac{1}{n}\sum_{i=1}^n \langle \mathcal{X}_i, \mathcal{A}\rangle \mathcal{M}_1(\mathcal{X}_i)(U_2 \otimes U_3) - \mathcal{M}_1(\mathcal{A})(U_2 \otimes U_3)\right\| \geqslant Ct\right) \\
\leqslant& 2 \cdot 7^{p_1 + r_2 r_3} 9^{p_2 r_2 + p_3 r_3} 33^{\bar{r}_1 \bar{r}_2 \bar{r}_3 + \sum_{i=1}^3 p_i \bar{r}_i} e^{-C_1 \min\{nt^2, nt\}}.
\end{aligned}$$

Proof of Lemma (5.2.9). (1) We only need to show that (5.228) holds for any fixed $U \in \mathbb{O}_{p_1,r_1}, V \in \mathbb{O}_{p_2,r_2}$. For any fixed $a \in \mathbb{R}^{r_1}, b \in \mathbb{R}^{r_2}$ satisfying $\|a\|_2 =$

$1, \|b\|_2 = 1$, notice that $\mathbb{E}[\langle X_i, \Delta\rangle X_i] = \Delta$ for $i \in [n]$, we have

$$\mathbb{E}[\langle X_i, \Delta\rangle a^\top U^\top X_i V b] = a^\top U^\top \Delta V b, \quad \forall i \in [n].$$

For any random variable Y_1 and Y_2, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}\|Y_1 Y_2\|_{\psi_1} \leqslant & C \sup_{q \geqslant 1} \frac{1}{q} \left(\mathbb{E}|Y_1 Y_2|^q\right)^{1/q} \leqslant C \left[\sup_{q \geqslant 1} \frac{1}{\sqrt{2q}} \left(\mathbb{E}|Y_1|^{2q}\right)^{\frac{1}{2q}}\right] \left[\sup_{q \geqslant 1} \frac{1}{\sqrt{2q}} \left(\mathbb{E}|Y_1|^{2q}\right)^{\frac{1}{2q}}\right] \\ \leqslant & C \|Y_1\|_{\psi_2} \|Y_2\|_{\psi_2}.\end{aligned} \tag{5.229}$$

Since $\langle X_i, \Delta\rangle \sim N(0, \|\Delta\|_F^2)$ and $a^\top U^\top X_i V b \sim N(0,1)$, by (Vershynin, 2010, Remark 5.18) and the above inequality, we have

$$\begin{aligned} & \left\|\langle X_i, \Delta\rangle a^\top U^\top X_i V b - a^\top U^\top \Delta V b\right\|_{\psi_1} \\ \leqslant & C \left\|\langle X_i, \Delta\rangle a^\top U^\top X_i V b\right\|_{\psi_1} \\ \leqslant & C \|\langle X_i, \Delta\rangle\|_{\psi_2} \|a^\top U^\top X_i V b\|_{\psi_2} \\ \leqslant & C \|\Delta\|_F. \end{aligned}$$

By Bernstein-type inequality, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n \frac{1}{n}\left(\langle X_i, \Delta\rangle a^\top U^\top X_i V b - a^\top U^\top \Delta V b\right)\right| \geqslant C\|\Delta\|_F t\right) \leqslant 2\exp\left[-C_1 \min\{nt^2, nt\}\right].$$

By (Vershynin, 2010, Lemma 5.2), there exist a 1/3-net $\mathcal{N}_1$ for $S^{r_1 - 1} = \{x : x \in \mathbb{R}^{r_1}, \|x\|_2 = 1\}$ with cardinality at most 7^{r_1} and a 1/3-net $\mathcal{N}_2$ for $S^{r_2-1} = \{x : x \in \mathbb{R}^{r_2}, \|x\|_2 = 1\}$ with cardinality at most 7^{r_2}. By the union bound, we have

$$\mathbb{P}\left(\sup_{a \in \mathcal{N}_1, b \in \mathcal{N}_2} \left|a^\top \left[\sum_{i=1}^n \frac{1}{n}\left(\langle X_i, \Delta\rangle U^\top X_i V - U^\top \Delta V\right)\right] b\right| \geqslant C\|\Delta\|_F t\right) \leqslant 2 \cdot 7^{r_1 + r_2} e^{-C_1 \min\{nt^2, nt\}}. \tag{5.230}$$

Let $a^* \in S^{r_1-1}$ and $b^* \in S^{r_2-1}$ satisfy

$$\begin{aligned}
&\left| a^{*\top} \left[\sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right] b^* \right| \\
&= \sup_{a \in S^{r_1-1}, b \in S^{r_2-1}} \left| a^\top \left[\sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right] b \right|.
\end{aligned}$$

Then there exist $\tilde{a} \in \mathcal{N}_1$ and $\tilde{b} \in \mathcal{N}_2$ such that $\|\tilde{a} - a^*\| \leqslant \frac{1}{3}$ and $\left\|\tilde{b} - b^*\right\| \leqslant \frac{1}{3}$. Therefore,

$$\begin{aligned}
&\left\| \sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right\| \\
=& \left| a^{*\top} \left[\sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right] b^* \right| \\
\leqslant& \left| \tilde{a}^\top \left[\sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right] \tilde{b} \right| + \left| (a^* - \tilde{a})^\top \left[\sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right] \tilde{b}^* \right| \\
&+ \left| a^{*\top} \left[\sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right] \left(b^* - \tilde{b} \right) \right| \\
\leqslant& \sup_{a \in \mathcal{N}_1, b \in \mathcal{N}_2} \left| a^\top \left[\sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right] b \right| + \frac{2}{3} \left\| \sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right\|,
\end{aligned}$$

which means that

$$\left\| \sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right\| \leqslant 3 \sup_{a \in \mathcal{N}_1, b \in \mathcal{N}_2} \left| a^\top \left[\sum_{i=1}^{n} \frac{1}{n} \left(\langle X_i, \Delta \rangle U^\top X_i V - U^\top \Delta V \right) \right] b \right|. \tag{5.231}$$

Combining the previous inequality and (5.230), we have proved the first part.

(2) For fixed $U_i \in \mathbb{R}^{p_i r_i}$ and $\mathcal{A} \in \mathbb{R}^{p_1 \times p_2 \times p_3}$ satisfying $\|U_i\| \leqslant 1$ and $\|\mathcal{A}\|_F \leqslant 1$, by

(5.228), we have

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\langle\mathcal{X}_i,\mathcal{A}\rangle\mathcal{M}_1(\mathcal{X}_i)(U_2\otimes U_3)-\mathcal{M}_1(\mathcal{A})(U_2\otimes U_3)\right\|\geqslant Ct\right)\leqslant 2{\cdot}7^{p_1+r_2r_3}e^{-C_1\min\{nt^2,nt\}}.$$

By (Zhang and Xia, 2018, Lemma 7), there exist 1/4-nets $\bar{\mathcal{X}}_{p_i,r_i}$ for $\mathcal{X}_{p_i,r_i}=\{U\in\mathbb{R}^{p_i\times r_i}:\|U\|\leqslant 1\}$ with cardinality at most $9^{p_ir_i}$. Therefore, by the union bound, we have

$$\begin{aligned}&\mathbb{P}\left(\sup_{\substack{U_i\in\bar{\mathcal{X}}_{p_i,r_i}\\ \mathcal{A}\in\bar{\mathcal{X}}_{\mathbf{p},\bar{\mathbf{r}}}}}\left\|\frac{1}{n}\sum_{i=1}^{n}\langle\mathcal{X}_i,\mathcal{A}\rangle U_1^\top\mathcal{M}_1(\mathcal{X}_i)(U_2\otimes U_3)-U_1^\top\mathcal{M}_1(\mathcal{A})(U_2\otimes U_3)\right\|\geqslant Ct\right)\\ \leqslant&2\cdot 7^{p_1+r_2r_3}9^{p_2r_2+p_3r_3}33^{\bar{r}_1\bar{r}_2\bar{r}_3+\sum_{i=1}^{3}p_i\bar{r}_i}e^{-C_1\min\{nt^2,nt\}},\end{aligned}$$

where $\bar{\mathcal{X}}_{\mathbf{p},\bar{\mathbf{r}}}$ is defined in (5.227) with $\epsilon=1/5$.
Similarly to (5.231), we have

$$\begin{aligned}&\sup_{\substack{U_i\in\mathbb{R}^{p_i\times r_i},\|U_i\|\leqslant 1\\ \mathcal{A}\in\mathbb{R}^{p_1\times p_2\times p_3},\|\mathcal{A}\|_F\leqslant 1\\ \operatorname{rank}(\mathcal{A})\leqslant(\bar{r}_1,\bar{r}_2,\bar{r}_3)}}\left\|\frac{1}{n}\sum_{i=1}^{n}\langle\mathcal{X}_i,\mathcal{A}\rangle U_1^\top\mathcal{M}_1(\mathcal{X}_i)(U_2\otimes U_3)-U_1^\top\mathcal{M}_1(\mathcal{A})(U_2\otimes U_3)\right\|\\ \leqslant&4\sup_{\substack{U_i\in\bar{\mathcal{X}}_{p_i,r_i}\\ \mathcal{A}\in\bar{\mathcal{X}}_{\mathbf{p},\mathbf{r}}}}\left\|\frac{1}{n}\sum_{i=1}^{n}\langle\mathcal{X}_i,\mathcal{A}\rangle U_1^\top\mathcal{M}_1(\mathcal{X}_i)(U_2\otimes U_3)-U_1^\top\mathcal{M}_1(\mathcal{A})(U_2\otimes U_3)\right\|.\end{aligned}$$

By combining the two previous inequalities together, we have finished the proof of the second part.

□

Lemma 5.2.10. *(1) Suppose* $X\in\mathbb{R}^{p_1\times p_2}$ *is a matrix with independent entries satisfying* $\mathbb{E}X_{ij}=0,\operatorname{Var}(X_{ij})=1,\|X_{ij}\|_{\psi_2}\leqslant C$, *and* $X_1,\ldots,X_n$ *are i.i.d. copies of* X. *Suppose* $\xi_1,\ldots,\xi_n$ *are independent zero-mean* $C\sigma$*-sub-Gaussian random variables. For any* *fixed* $U \in \mathbb{O}_{p_1,r_1}$ *and* $V \in \mathbb{O}_{p_2,r_2}$*, we have*

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \xi_i U^\top X_i V\right\| \geqslant C_2\sqrt{n}\sqrt{r_1+r_2+x}\sigma\right) \leqslant e^{-C_1 x} + e^{-c_1 n}.$$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \xi_i U^\top X_i V\right\|_F \geqslant C_2\sqrt{n}\sqrt{r_1 r_2+x}\sigma\right) \leqslant e^{-C_1 x} + e^{-c_1 n}.$$

(2) *Suppose* $\mathcal{X} \in \mathbb{R}^{p_1\times p_2\times p_3}$ *is a tensor with independent entries satisfying* $\mathbb{E}\mathcal{X}_{ijk} = 0, \mathrm{Var}(\mathcal{X}_{ijk}) = 1, \|\mathcal{X}_{ijk}\|_{\psi_2} \leqslant C$*, and* $\mathcal{X}_1,\dots,\mathcal{X}_n$ *are i.i.d. copies of* $\mathcal{X}$*. Suppose* $\xi_1,\dots,\xi_n$ *are independent zero-mean* $C\sigma$*-sub-Gaussian random variables. Let* $p = \max_{j=1,2,3} p_j$ *and* $r = \max\{r_1,r_2,r_3\}$*. Suppose* $r \leqslant \sqrt{p}$*. Then for fixed* $V_i \in \mathbb{O}_{p_i,r_i}$*,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \xi_i V_1^\top \mathcal{M}_1(\mathcal{X}_i)(V_2\otimes V_3)\right\| \geqslant C_2\sqrt{n}\sqrt{r_1+r_2r_3+\log(p)}\sigma\right) \leqslant p^{-C_1}+e^{-c_1 n}. \tag{5.232}$$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \xi_i V_1^\top \mathcal{M}_1(\mathcal{X}_i)(V_2\otimes V_3)\right\|_F \geqslant C_2\sqrt{n}\sqrt{r_1r_2r_3+\log(p)}\sigma\right) \leqslant p^{-C_1}+e^{-c_1 n}. \tag{5.233}$$

Moreover,

$$\mathbb{P}\left(\sup_{V_i\in\mathbb{O}_{p_i,r_i}}\left\|\sum_{i=1}^{n} \xi_i V_1^\top \mathcal{M}_1(\mathcal{X}_i)(V_2\otimes V_3)\right\| \geqslant C_2\sqrt{npr}\sigma\right) \leqslant e^{-C_1pr}+e^{-C_1 n}. \tag{5.234}$$

$$\mathbb{P}\left(\sup_{V_i\in\mathbb{O}_{p_i,r_i}}\left\|\sum_{i=1}^{n} \xi_i V_1^\top \mathcal{M}_1(\mathcal{X}_i)(V_2\otimes V_3)\right\|_F \geqslant C_2\sqrt{npr}\sigma\right) \leqslant e^{-C_1pr}+e^{-C_1 n}. \tag{5.235}$$

(3) *Suppose the conditions in (2) holds and* $p_1 \geqslant r_2r_3 \geqslant r_1$*. Then for fixed* $V_2 \in \mathbb{O}_{p_2,r_2}$

and $V_3 \in \mathbb{O}_{p_3,r_3}$ *and* $V_4 \in \mathbb{O}_{r_2r_3,r_1}$,

$$\mathbb{P}\left(\sup_{V_i \in \mathbb{O}_{p_i,r_i}} \left\| \sum_{i=1}^{n} \xi_i \mathcal{M}_1(\mathcal{X}_i)(V_2 \otimes V_3)V_4 \right\| \geqslant C_2\sqrt{np}\sigma\right) \leqslant e^{-C_1p} + e^{-C_1n}.$$

$$\mathbb{P}\left(\sup_{V_i \in \mathbb{O}_{p_i,r_i}} \left\| \sum_{i=1}^{n} \xi_i \mathcal{M}_1(\mathcal{X}_i)(V_2 \otimes V_3)V_4 \right\|_F \geqslant C_2\sqrt{npr}\sigma\right) \leqslant e^{-C_1pr} + e^{-C_1n}.$$

In addition,

$$\mathbb{P}\left(\sup_{\substack{V_i \in \mathbb{O}_{p_i,r_i}, i=2,3 \\ V_4 \in \mathbb{O}_{r_2r_3,r_1}}} \left\| \sum_{i=1}^{n} \mathcal{M}_1(\mathcal{X}_i)(V_2 \otimes V_3)V_4 \right\| \geqslant C_2\sqrt{npr}\sigma\right) \leqslant e^{-C_1pr} + e^{-C_1n}.$$

$$\mathbb{P}\left(\sup_{\substack{V_i \in \mathbb{O}_{p_i,r_i}, i=2,3 \\ V_4 \in \mathbb{O}_{r_2r_3,r_1}}} \left\| \sum_{i=1}^{n} \mathcal{M}_1(\mathcal{X}_i)(V_2 \otimes V_3)V_4 \right\|_F \geqslant C_2\sqrt{npr}\sigma\right) \leqslant e^{-C_1pr} + e^{-C_1n}.$$

Proof of Lemma 5.2.10. Without loss of generality, we assume $\sigma = 1$. For any fixed $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, noting that the entries of $\sum_{i=1}^n a_i X_i$ are independent $C\|a\|_2$-sub-Gaussian random variables with mean 0 and variance $\|a\|_2^2$. By Lemma 5.2.7, for any $x \geqslant 0$,

$$\mathbb{P}\left(\left\| \sum_{i=1}^{n} a_i U^\top X_i V \right\| \geqslant C\|a\|_2\sqrt{r_1 + r_2 + x}\right) \leqslant e^{-C_1x}.$$

$$\mathbb{P}\left(\left\| \sum_{i=1}^{n} a_i U^\top X_i V \right\|_F \geqslant C\|a\|_2\sqrt{r_1r_2 + x}\right) \leqslant e^{-C_1x}.$$

Therefore,

$$\mathbb{P}\left(\left\| \sum_{i=1}^{n} \xi_i U^\top X_i V \right\| \geqslant C\|\xi\|_2\sqrt{r_1 + r_2 + x}\,\middle|\, \xi_1, \ldots, \xi_n\right) \leqslant e^{-C_1x}. \tag{5.236}$$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \xi_i U^\top X_i V\right\|_F \geqslant C\|\xi\|_2\sqrt{r_1 r_2 + x}\middle|\xi_1, \ldots, \xi_n\right) \leqslant e^{-C_1 x}. \tag{5.237}$$

By Bernstein-type inequality ((Vershynin, 2010, Proposition 5.16)),

$$\mathbb{P}\left(\|\xi\|_2 \geqslant C\sqrt{n}\right) \leqslant e^{-C_1 n}. \tag{5.238}$$

Thus we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \xi_i U^\top X_i V\right\| \geqslant C\sqrt{n}\sqrt{r_1 + r_2 + x}\right) \leqslant e^{-C_1 x} + e^{-C_1 n}.$$

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} \xi_i U^\top X_i V\right\|_F \geqslant C\sqrt{n}\sqrt{r_1 r_2 r_3 + x}\right) \leqslant e^{-C_1 x} + e^{-C_1 n}.$$

By setting $x = \log(p), U = U_1, X_i = \mathcal{M}(\mathcal{X}_i)$ and $V = U_2 \otimes U_3$ in the previous two inequalities, we have proved (5.232) and (5.233), respectively.

By setting $t = C\sqrt{pr}$ in (5.236) and (5.237) and using Lemma 5.2.6, (5.238) and the ϵ-net argument, we have (5.234) and (5.235).

Similarly, we can prove the inequalities in Part 3. □

Proof of Lemma 3.3.2. By the definition of ψ_α-norm,

$$\mathbb{E}\left[\exp\left(\left|\frac{X_i}{K_i}\right|^{\alpha_i}\right)\right] \leqslant 2, \quad \forall i \in [n].$$

By the weighted AM-GM inequality, we have

$$\left|\frac{\prod_{i=1}^{n} X_i}{\prod_{i=1}^{n} K_i}\right|^{\frac{1}{\sum_{i=1}^{n}\frac{1}{\alpha_i}}} \leqslant \sum_{i=1}^{n} \frac{\frac{1}{\alpha_i}}{\sum_{i=1}^{n}\frac{1}{\alpha_i}} \left|\frac{X_i}{K_i}\right|^{\alpha_i},$$

and

$$\exp\left(\left|\frac{\prod_{i=1}^n X_i}{\prod_{i=1}^n K_i}\right|^{\frac{1}{\sum_{i=1}^n \frac{1}{\alpha_i}}}\right) \leqslant \exp\left(\sum_{i=1}^n \frac{\frac{1}{\alpha_i}}{\sum_{i=1}^n \frac{1}{\alpha_i}}\left|\frac{X_i}{K_i}\right|^{\alpha_i}\right)$$
$$= \prod_{i=1}^n \exp\left(\frac{\frac{1}{\alpha_i}}{\sum_{j=1}^n \frac{1}{\alpha_j}}\left|\frac{X_i}{K_i}\right|^{\alpha_i}\right)$$
$$\leqslant \sum_{i=1}^n \frac{\frac{1}{\alpha_i}}{\sum_{j=1}^n \frac{1}{\alpha_j}} \exp\left(\left|\frac{X_i}{K_i}\right|^{\alpha_i}\right).$$

By taking expectations on both sides, we have

$$\mathbb{E}\left[\exp\left(\left|\frac{\prod_{i=1}^n X_i}{\prod_{i=1}^n K_i}\right|^{\frac{1}{\sum_{i=1}^n \frac{1}{\alpha_i}}}\right)\right] \leqslant \sum_{i=1}^n \frac{\frac{1}{\alpha_i}}{\sum_{j=1}^n \frac{1}{\alpha_j}}\mathbb{E}\left[\exp\left(\left|\frac{X_i}{K_i}\right|^{\alpha_i}\right)\right] \leqslant 2,$$

which has finished the proof of Lemma (3.3.2) □

Proof of Lemma 3.3.3. Since $\frac{\langle Z_i, M_i\rangle}{\|M_i\|_F} \sim N(0,1)$, $\|\frac{1}{\|M_i\|_F}\langle Z_i, M_i\rangle\|_{\psi_2} \leqslant C$. Notice that $\|Z_i\|_F^2 \sim \chi^2_{pr}$ by (Wainwright, 2019, Example 2.8), we have

$$\mathbb{E}\left(e^{t(\|Z_i\|_F^2 - pr)}\right) \leqslant e^{2prt^2}, \quad \forall |t| < \frac{1}{4}.$$

Set $t = \sqrt{\frac{\log(2)}{2pr}} \leqslant 1/4$, we have

$$\mathbb{E}\left(e^{\frac{\|Z_i\|_F^2 - pr}{\sqrt{2pr/\log(2)}}}\right) \leqslant 2.$$

Therefore, $\left\|\|Z_i\|_F^2 - pr\right\|_{\psi_1} \leqslant C\sqrt{pr}$. By Lemma 3.3.2,

$$\left\|(\|Z_i\|_F^2 - pr)\frac{\langle Z_i, M_i\rangle}{\|M_i\|_F}\right\|_{\psi_{2/3}} \leqslant \left\|\frac{1}{\|M_i\|_F}\langle Z_i, M_i\rangle\right\|_{\psi_2} \left\|\|Z_i\|_F^2 - pr\right\|_{\psi_1} \leqslant C\sqrt{pr}.$$

Since

$$\mathbb{E}\left(\sum_{i=1}^{n}(\|Z_i\|_F^2 - pr)\langle Z_i, M_i\rangle\right) = 0,$$

by (Hao et al., 2020, Lemma 8), with probability at least $1 - p^{-C_1}$,

$$\left|\sum_{i=1}^{n}(\|Z_i\|_F^2 - pr)\langle Z_i, M_i\rangle\right| \leqslant C\sqrt{pr}\left(\left(\sum_{i=1}^{n}\|M_i\|_F^2\right)^{1/2}\sqrt{\log(p)} + \left(\max_{1\leqslant i\leqslant n}\|M_i\|_F\right)(\log(p))^{3/2}\right).$$

In addition, since $\sum_{i=1}^{n}\langle Z_i, M_i\rangle \sim N(0, \sum_{i=1}^{n}\|M_i\|_F^2)$,

$$\mathbb{P}\left(\left|pr\sum_{i=1}^{n}\langle Z_i, M_i\rangle\right| \geqslant Cpr\left(\sum_{i=1}^{n}\|M_i\|_F^2\right)^{1/2}\sqrt{\log(p)}\right) \leqslant p^{-C_1}.$$

By combining the previous two inequalities, we have finished the proof of Lemma 3.3.3. □

Lemma 5.2.11. *Suppose* k *and* d *are fixed numbers satisfying* $1 \leqslant k \leqslant d$. $Y_p \overset{d.}{\to} N(0, I_d)$ *as* $p \to \infty$. *Then for any deterministic matrix array* $\{A_p\}_{p=1}^{\infty}$ *satisfying* $A_p \in \mathbb{O}_{d,k}$, *we have*

$$A_p^\top Y_p \overset{d.}{\to} N(0, I_k) \quad \text{as} \quad p \to \infty.$$

Proof of Lemma 5.2.11. By Skorohod's theorem (Shao, 2003, Theorem 1.8), there exist random vectors $\{Z_p\}, Z$ defined on a common probability space such that $Z \sim N(0, I_3)$, $Y_p \overset{d.}{=} Z_p$ and $Z_p \overset{a.s.}{\to} Z$ as $p \to \infty$. Notice that

$$A_p^\top Z \sim N(0, 1)$$

and

$$\|A_p^\top(Z_p - Z)\|_2 \leqslant \|A_p\|\|Z_p - Z\|_2 = \|Z_p - Z\|_2 \overset{a.s.}{\to} 0 \quad \text{as} \quad p \to \infty,$$

we have

$$A_p^\top Y_p \overset{d.}{\to} A_p^\top Z_p = A_p^\top Z + A_p^\top (Z_p - Z) \overset{d.}{\to} N(0, I_k) \quad \text{as} \quad p \to \infty.$$

□

5.3 Appendix to Chapter 4

We collect all technical proofs of Chapter 4 in this section.

5.3.1 Proof of Theorem 4.1

For convenience, let $\widehat{U}_i$, $\widehat{V}_i$, R_i and $\widetilde{R}_i$ denote $\widehat{U}_i^{(0)}$, $\widehat{V}_i^{(1)}$, $R_i^{(0)}$ and $\widetilde{R}_i^{(0)}$, respectively. By Lemma 4.3.1 and

$$\begin{aligned}
&I_{p_2\cdots p_d} - P_{(\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3 \otimes I_{p_2})\widehat{V}_2} \\
=&P_{(\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3 \otimes I_{p_2})\widehat{V}_{2\perp}} + P_{(\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_4 \otimes I_{p_2 p_3})(\widehat{V}_{3\perp} \otimes I_{p_2})} + \cdots + P_{\widehat{V}_{d\perp} \otimes I_{p_2\cdots p_{d-1}}},
\end{aligned}$$

we have

$$\begin{aligned}
&\left\| \widehat{\mathcal{X}}^{(1)} - \mathcal{X} \right\|_F^2 \\
=&\left\| \left[[\mathcal{Y}]_1 (\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}}) \cdots (\widehat{V}_3 \otimes I_{p_2}) \widehat{V}_2 \right] \widehat{V}_2^\top (\widehat{V}_3^\top \otimes I_{p_2}) \cdots (\widehat{V}_d^\top \otimes I_{p_2\cdots p_{d-1}}) - [\mathcal{X}]_1 \right\|_F^2 \\
=&\left\| [\mathcal{Z}]_1 P_{(\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3 \otimes I_{p_2})\widehat{V}_2} + [\mathcal{X}]_1 P_{(\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3 \otimes I_{p_2})\widehat{V}_2} - [\mathcal{X}]_1 \right\|_F^2 \\
\leqslant&C\Bigg(\left\| [\mathcal{Z}]_1 P_{(\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3 \otimes I_{p_2})\widehat{V}_2} \right\|_F^2 + \left\| [\mathcal{X}]_1 P_{(\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_3 \otimes I_{p_2})\widehat{V}_{2\perp}} \right\|_F^2 \\
&+ \left\| [\mathcal{X}]_1 P_{(\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}})\cdots(\widehat{V}_4 \otimes I_{p_2 p_3})(\widehat{V}_{3\perp} \otimes I_{p_2})} \right\|_F^2 + \cdots + \left\| [\mathcal{X}]_1 P_{\widehat{V}_{d\perp} \otimes I_{p_2\cdots p_{d-1}}} \right\|_F^2 \Bigg) \\
\leqslant&C\Bigg(\left\| [\mathcal{Z}]_1 (\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}}) \cdots (\widehat{V}_3 \otimes I_{p_2}) \widehat{V}_2 \right\|_F^2 + \left\| [\mathcal{X}]_1 (\widehat{V}_d \otimes I_{p_2\cdots p_{d-1}}) \cdots (\widehat{V}_3 \otimes I_{p_2}) \widehat{V}_{2\perp} \right\|_F^2
\end{aligned}$$

$$+\left\|[\mathcal{X}]_1(\widehat{V}_d \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_4 \otimes I_{p_2p_3})(\widehat{V}_{3\perp} \otimes I_{p_2})\right\|_F^2 + \cdots + \left\|[\mathcal{X}]_1(\widehat{V}_{d\perp} \otimes I_{p_2 \dots p_{d-1}})\right\|_F^2 \Bigg). \tag{5.239}$$

To prove (4.9), we only need to show that for all $2 \leqslant k \leqslant d$,

$$\begin{aligned}&\left\|[\mathcal{X}]_1(\widehat{V}_d \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_2 \cdots p_k})(\widehat{V}_{k\perp} \otimes I_{p_2 \cdots p_{k-1}})\right\|_F \\ \leqslant & C\left\|\widehat{U}_{k-1}^\top(I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{Z}]_{k-1}(\widehat{V}_d \otimes I_{p_k \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\right\|_F,\end{aligned} \tag{5.240}$$

where

$$[\mathcal{X}]_1(\widehat{V}_d \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_2 \cdots p_k})(\widehat{V}_{k\perp} \otimes I_{p_2 \cdots p_{k-1}}) = [\mathcal{X}]_1(\widehat{V}_d \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_3 \otimes I_{p_2})\widehat{V}_{2\perp}$$

if $k = 2$ and

$$[\mathcal{X}]_1(\widehat{V}_d \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_2 \cdots p_k})(\widehat{V}_{k\perp} \otimes I_{p_2 \cdots p_{k-1}}) = [\mathcal{X}]_1(\widehat{V}_{d\perp} \otimes I_{p_2 \dots p_{d-1}})$$

if $k = d$.
By Lemma 4.3.2, we have

$$\begin{aligned}&\left\|[\mathcal{X}]_1(\widehat{V}_d \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_2 \cdots p_k})(\widehat{V}_{k\perp} \otimes I_{p_2 \cdots p_{k-1}})\right\|_F \\ =&\left\|[A^{(p_2 \cdots p_{k-1}, p_1)}]^\top([\mathcal{X}]_{k-1} \otimes I_{p_2 \cdots p_{k-1}})(\widehat{V}_d \otimes I_{p_2 \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_2 \cdots p_k})(\widehat{V}_{k\perp} \otimes I_{p_2 \cdots p_{k-1}})\right\|_F \\ =&\left\|[A^{(p_2 \cdots p_{k-1}, p_1)}]^\top\left(\left([\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right) \otimes I_{p_2 \cdots p_{k-1}}\right)\right\|_F \\ =&\left\|[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k \dots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_F.\end{aligned} \tag{5.241}$$

The third equation holds since the realignment doesn't change the Frobenious norm.
Moreover, recall that $U_1 \in \mathbb{R}^{p_1 \times r_1}$ is the left singular space of $[\mathcal{X}]_1$, and $\widetilde{U}_j \in \mathbb{R}^{p_j r_{j-1} \times r_j}$ is the left singular space of $(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j$

for $2 \leqslant j \leqslant d-1$, by Lemma 4.3.2, for any $2 \leqslant k \leqslant d-1$,

$$\begin{aligned}
[\mathcal{X}]_k =& (I_{p_2\cdots p_k} \otimes [\mathcal{X}]_1) A^{(p_2\cdots p_k, p_{k+1}\cdots p_d)} \\
=& (I_{p_2\cdots p_k} \otimes P_{U_1}[\mathcal{X}]_1) A^{(p_2\cdots p_k, p_{k+1}\cdots p_d)} \\
=& (I_{p_2\cdots p_k} \otimes P_{U_1})(I_{p_2\cdots p_k} \otimes [\mathcal{X}]_1) A^{(p_2\cdots p_k, p_{k+1}\cdots p_d)} \\
=& (I_{p_2\cdots p_k} \otimes P_{U_1})[\mathcal{X}]_k,
\end{aligned} \tag{5.242}$$

and for any $2 \leqslant j < k$,

$$\begin{aligned}
& (I_{p_j\cdots p_k} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}\cdots p_k} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2\cdots p_k} \otimes \widehat{U}_1^\top)[\mathcal{X}]_k \\
=& (I_{p_j\cdots p_k} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}\cdots p_k} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2\cdots p_k} \otimes \widehat{U}_1^\top)(I_{p_{j+1}\cdots p_k} \otimes [\mathcal{X}]_j) A^{(p_{j+1}\cdots p_k, p_{k+1}\cdots p_d)} \\
=& \left(I_{p_{j+1}\cdots p_k} \otimes \left[(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2\cdots p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right]\right) A^{(p_{j+1}\cdots p_k, p_{k+1}\cdots p_d)} \\
=& \left(I_{p_{j+1}\cdots p_k} \otimes \left[P_{\widetilde{U}_j}(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2\cdots p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right]\right) A^{(p_{j+1}\cdots p_k, p_{k+1}\cdots p_d)} \\
=& (I_{p_{j+1}\cdots p_k} \otimes P_{\widetilde{U}_j})(I_{p_j\cdots p_k} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}\cdots p_k} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2\cdots p_k} \otimes \widehat{U}_1^\top)(I_{p_{j+1}\cdots p_k} \otimes [\mathcal{X}]_j) \\
& \quad \cdot A^{(p_{j+1}\cdots p_k, p_{k+1}\cdots p_d)} \\
=& (I_{p_{j+1}\cdots p_k} \otimes P_{\widetilde{U}_j})(I_{p_j\cdots p_k} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}\cdots p_k} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2\cdots p_k} \otimes \widehat{U}_1^\top)[\mathcal{X}]_k,
\end{aligned} \tag{5.243}$$

where $A^{(i,j)}$ is defined in (4.5) for any $i, j > 0$.
Therefore, by (5.242),

$$\begin{aligned}
& \left\| [\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp} \right\|_F \\
=& \left\| (I_{p_2\cdots p_{k-1}} \otimes P_{U_1})[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp} \right\|_F \\
=& \left\| (I_{p_2\cdots p_{k-1}} \otimes U_1^\top)[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp} \right\|_F \\
\leqslant & \left\| (I_{p_2\cdots p_{k-1}} \otimes \widehat{U}_1^\top)(I_{p_2\cdots p_{k-1}} \otimes U_1)(I_{p_2\cdots p_{k-1}} \otimes U_1^\top)[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp} \right\|_F \\
& \cdot s_{\min}^{-1}\left((I_{p_2\cdots p_{k-1}} \otimes \widehat{U}_1^\top)(I_{p_2\cdots p_{k-1}} \otimes U_1)\right)
\end{aligned}$$

$$= \left\|(I_{p_2\cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_F \cdot s_{\min}^{-1}(\widehat{U}_1^\top U_1). \tag{5.244}$$

The inequality holds since $\|B\|_F \leqslant \|AB\|_F \cdot s_{\min}^{-1}(A)$ for any invertible matrix $A \in \mathbb{R}^{m_1 \times m_1}$ and $B \in \mathbb{R}^{m_1 \times m_2}$; in the last step, we used $(I_{p_2\cdots p_{k-1}} \otimes U_1)(I_{p_2\cdots p_{k-1}} \otimes U_1^\top)[\mathcal{X}]_{k-1} = (I_{p_2\cdots p_{k-1}} \otimes P_{U_1})[\mathcal{X}]_{k-1} = [\mathcal{X}]_{k-1}$. Similarly to (5.244), by (5.243), for $1 \leqslant j \leqslant k-2$,

$$\begin{aligned}
&\left\|(I_{p_{j+1}\cdots p_{k-1}} \otimes \widehat{U}_j^\top)\cdots(I_{p_2\cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_F \\
=&\Big\|(I_{p_{j+2}\cdots p_{k-1}} \otimes P_{\widetilde{U}_{j+1}})(I_{p_{j+1}\cdots p_{k-1}} \otimes \widehat{U}_j^\top)\cdots(I_{p_2\cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{k-1} \\
&\cdot(\widehat{V}_d \otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_F \\
=&\Big\|(I_{p_{j+2}\cdots p_{k-1}} \otimes \widetilde{U}_{j+1}^\top)(I_{p_{j+1}\cdots p_{k-1}} \otimes \widehat{U}_j^\top)\cdots(I_{p_2\cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{k-1} \\
&\cdot(\widehat{V}_d \otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_F \\
\leqslant&\Big\|(I_{p_{j+2}\cdots p_{k-1}} \otimes \widehat{U}_{j+1}^\top)(I_{p_{j+1}\cdots p_{k-1}} \otimes \widehat{U}_j^\top)\cdots(I_{p_2\cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{k-1} \\
&\cdot(\widehat{V}_d \otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\Big\|_F \cdot s_{\min}^{-1}(\widehat{U}_{j+1}^\top \widetilde{U}_{j+1}).
\end{aligned} \tag{5.245}$$

By (5.244) and (5.245),

$$\begin{aligned}
&\left\|[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_F \\
\leqslant&\left\|(I_{p_2\cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_F s_{\min}^{-1}(\widehat{U}_1^\top U_1) \\
\leqslant&\left\|(I_{p_3\cdots p_{k-1}} \otimes \widehat{U}_2^\top)(I_{p_2\cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_F \\
&\cdot s_{\min}^{-1}(U_1^\top \widehat{U}_1) s_{\min}^{-1}(\widetilde{U}_2^\top \widehat{U}_2) \\
\leqslant&\cdots \\
\leqslant&\left\|\widehat{U}_{k-1}^\top(I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top)\cdots(I_{p_2\cdots p_{k-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{k-1}(\widehat{V}_d \otimes I_{p_k\dots p_{d-1}})\cdots(\widehat{V}_{k+1} \otimes I_{p_k})\widehat{V}_{k\perp}\right\|_F
\end{aligned}$$

$$
\begin{aligned}
&\cdot s_{\min}^{-1}(U_1^\top \widehat{U}_1) s_{\min}^{-1}(\widetilde{U}_2^\top \widehat{U}_2) \cdots s_{\min}^{-1}(\widetilde{U}_{k-1}^\top \widehat{U}_{k-1}) \\
\leqslant & \left\| \widehat{U}_{k-1}^\top (I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top) [\mathcal{X}]_{k-1} (\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k}) \widehat{V}_{k\perp} \right\|_F \\
&\cdot \left(\frac{1}{\sqrt{1-c_0^2}} \right)^{k-1} \\
\leqslant & C \left\| \widehat{U}_{k-1}^\top (I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top) [\mathcal{X}]_{k-1} (\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k}) \widehat{V}_{k\perp} \right\|_F .
\end{aligned}
\tag{5.246}
$$

By the definition of $\widehat{V}_k \in \mathbb{R}^{(p_k r_k) \times r_{k-1}}$ and Lemma 4.3.3, we know that $\widehat{V}_k$ is the right singular space of

$$
\begin{aligned}
& \widehat{U}_{k-1}^\top (I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top) [\mathcal{Y}]_{k-1} (\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k}) \\
= & \widehat{U}_{k-1}^\top (I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top) [\mathcal{X}]_{k-1} (\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k}) \\
& + \widehat{U}_{k-1}^\top (I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top) [\mathcal{Z}]_{k-1} (\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k}),
\end{aligned}
$$

Lemma 5.3.3 shows that

$$
\begin{aligned}
& \left\| \widehat{U}_{k-1}^\top (I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top) [\mathcal{X}]_{k-1} (\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k}) \widehat{V}_{k\perp} \right\|_F \\
\leqslant & 2 \left\| \widehat{U}_{k-1}^\top (I_{p_{k-1}} \otimes \widehat{U}_{k-2}^\top) \cdots (I_{p_2 \cdots p_{k-1}} \otimes \widehat{U}_1^\top) [\mathcal{Z}]_{k-1} (\widehat{V}_d \otimes I_{p_k \ldots p_{d-1}}) \cdots (\widehat{V}_{k+1} \otimes I_{p_k}) \right\|_F .
\end{aligned}
\tag{5.247}
$$

Combine (5.241), (5.246) and (5.247) together, we know that (5.240) holds for all $2 \leqslant k \leqslant d$, which has finished the proof of Theorem 4.1.

5.3.2 Proof of Theorem 4.2

For $i \geqslant 1$, by the definition of $\mathcal{X}^{(2i)}$ and Lemma 4.3.1, we have

$$
\begin{aligned}
\left\| \mathcal{Y} - \widehat{\mathcal{X}}^{(2i)} \right\|_F^2 &= \left\| \left(I_{p_1 \cdots p_{d-1}} - P_{(I_{p_2 \cdots p_{d-1}} \otimes \widehat{U}_1^{(2i)}) \cdots (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^{(2i)}) \widehat{U}_{d-1}^{(2i)}} \right) [\mathcal{Y}]_{d-1} \right\|_F^2 \\
&= \| [\mathcal{Y}]_{d-1} \|_F^2 - \left\| P_{(I_{p_2 \cdots p_{d-1}} \otimes \widehat{U}_1^{(2i)}) \cdots (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^{(2i)}) \widehat{U}_{d-1}^{(2i)}} [\mathcal{Y}]_{d-1} \right\|_F^2
\end{aligned}
$$

$$= \|\mathcal{Y}\|_F^2 - \left\|\widehat{\mathcal{X}}^{(2i)}\right\|_F^2.$$

Similarly, we have

$$\left\|\mathcal{Y} - \widehat{\mathcal{X}}^{(2i-1)}\right\|_F^2 = \|\mathcal{Y}\|_F^2 - \left\|\widehat{\mathcal{X}}^{(2i-1)}\right\|_F^2.$$

In addition, we have

$$\begin{aligned}
\left\|\mathcal{Y} - \widehat{\mathcal{X}}^{(2i)}\right\|_F^2 &= \|[\mathcal{Y}]_{d-1}\|_F^2 - \left\|P_{(I_{p_2\cdots p_{d-1}} \otimes \hat{U}_1^{(2i)})\cdots(I_{p_{d-1}} \otimes \hat{U}_{d-2}^{(2i)})\hat{U}_{d-1}^{(2i)}}[\mathcal{Y}]_{d-1}\right\|_F^2 \\
&= \|[\mathcal{Y}]_{d-1}\|_F^2 - \left\|\widehat{U}_{d-1}^{(2i)\top}(I_{p_{d-1}} \otimes \widehat{U}_{d-2}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}} \otimes \widehat{U}_1^{(2i)\top})[\mathcal{Y}]_{d-1}\right\|_F^2 \\
&= \|[\mathcal{Y}]_1\|_F^2 - \left\|\widehat{U}_{d-1}^{(2i)\top}(I_{p_{d-1}} \otimes \widehat{U}_{d-2}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}} \otimes \widehat{U}_1^{(2i)\top})[\mathcal{Y}]_{d-1}\widehat{V}_d^{(2i-1)}\right\|_F^2 \\
&\quad - \left\|\widehat{U}_{d-1}^{(2i)\top}(I_{p_{d-1}} \otimes \widehat{U}_{d-2}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}} \otimes \widehat{U}_1^{(2i)\top})[\mathcal{Y}]_{d-1}\widehat{V}_{d\perp}^{(2i-1)}\right\|_F^2 \\
&\leqslant \|[\mathcal{Y}]_1\|_F^2 - \left\|\widehat{U}_{d-1}^{(2i)\top}(I_{p_{d-1}} \otimes \widehat{U}_{d-2}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}} \otimes \widehat{U}_1^{(2i)\top})[\mathcal{Y}]_{d-1}\widehat{V}_d^{(2i-1)}\right\|_F^2 \\
&= \|[\mathcal{Y}]_1\|_F^2 - \left\|(I_{p_{d-1}} \otimes \widehat{U}_{d-2}^{(2i)\top})(I_{p_{d-2}p_{d-1}} \otimes \widehat{U}_{d-3}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}} \otimes \widehat{U}_1^{(2i)\top})[\mathcal{Y}]_{d-1}\widehat{V}_d^{(2i-1)}\right\|_F^2.
\end{aligned}$$

The last equation holds since $\widehat{U}_{d-1}^{(2i)}$ is the left singular space of $(I_{p_{d-1}} \otimes \widehat{U}_{d-2}^{(2i)\top})(I_{p_{d-2}p_{d-1}} \otimes \widehat{U}_{d-3}^{(2i)\top})\cdots(I_{p_2\cdots p_{d-1}} \otimes \widehat{U}_1^{(2i)\top})[\mathcal{Y}]_{d-1}\widehat{V}_d^{(2i-1)}$.

For any $B \in \mathbb{R}^{n\times r}$ and $1 \leqslant l \leqslant r$, we can check that the l-th columns of $A^{(m,n)}B$ and $(I_m \otimes B \otimes I_m)A^{(m,r)}$ are equal:

$$(A^{(m,n)}B)_{[:,l]} = \sum_{j=1}^{n} B_{j,l} \sum_{k=1}^{m} e_{(k-1)mn+(j-1)m+k}^{(m^2n)} = ((I_m \otimes B \otimes I_m)A^{(m,r)})_{[:,l]}$$

where $e_{(k-1)mn+(j-1)m+k}^{(m^2n)}$ is the $((k-1)mn + (j-1)m + k)$-th canonical basis of $\mathbb{R}^{m^2n}$ and $A^{(i,j)}$ is defined in (4.5). Therefore,

$$A^{(m,n)}B = (I_m \otimes B \otimes I_m)A^{(m,r)}.$$

By the last equation and Lemma 4.3.2, we have

$$\begin{aligned}
&(I_{p_{d-1}} \otimes \widehat{U}_{d-2}^{(2i)\top})(I_{p_{d-2}p_{d-1}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-1}} \otimes \widehat{U}_{1}^{(2i)\top})[\mathcal{Y}]_{d-1}\widehat{V}_d^{(2i-1)} \\
=&(I_{p_{d-1}} \otimes \widehat{U}_{d-2}^{(2i)\top})(I_{p_{d-2}p_{d-1}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-1}} \otimes \widehat{U}_{1}^{(2i)\top})(I_{p_{d-1}} \otimes [\mathcal{Y}]_{d-2})A^{(p_{d-1},p_d)}\widehat{V}_d^{(2i-1)} \\
=&\Big(I_{p_{d-1}} \otimes \Big(\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-2}} \otimes \widehat{U}_{1}^{(2i)\top})[\mathcal{Y}]_{d-2}\Big)\Big)\Big(I_{p_{d-1}} \otimes (\widehat{V}_d^{(2i-1)} \otimes I_{p_{d-1}})\Big) \\
&\cdot A^{(p_{d-1},r_{d-1})} \\
=&\Big(I_{p_{d-1}} \otimes \Big(\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-2}} \otimes \widehat{U}_{1}^{(2i)\top})[\mathcal{Y}]_{d-2}(\widehat{V}_d^{(2i-1)} \otimes I_{p_{d-1}})\Big)\Big) A^{(p_{d-1},r_{d-1})} \\
=&\text{Reshape}\big(\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-2}} \otimes \widehat{U}_{1}^{(2i)\top})[\mathcal{Y}]_{d-2}(\widehat{V}_d^{(2i-1)} \otimes I_{p_{d-1}}), r_{d-2}p_{d-1}, r_{d-1}\big).
\end{aligned}$$

Since the realignment does not change the Frobenius norm, we have

$$\left\|\mathcal{Y} - \widehat{\mathfrak{X}}^{(2i)}\right\|_F^2 \leqslant \|[\mathcal{Y}]_1\|_F^2 - \left\|\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-2}} \otimes \widehat{U}_{1}^{(2i)\top})[\mathcal{Y}]_{d-2}(\widehat{V}_d^{(2i-1)} \otimes I_{p_{d-1}})\right\|_F^2. \tag{5.248}$$

By similar proof of (5.248), we have

$$\begin{aligned}
&\left\|\mathcal{Y} - \widehat{\mathfrak{X}}^{(2i)}\right\|_F^2 \\
\leqslant& \|[\mathcal{Y}]_1\|_F^2 - \left\|\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-2}} \otimes \widehat{U}_{1}^{(2i)\top})[\mathcal{Y}]_{d-2}(\widehat{V}_d^{(2i-1)} \otimes I_{p_{d-1}})\right\|_F^2 \\
=& \|[\mathcal{Y}]_1\|_F^2 - \left\|\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-2}} \otimes \widehat{U}_{1}^{(2i)\top})[\mathcal{Y}]_{d-2}(\widehat{V}_d^{(2i-1)} \otimes I_{p_{d-1}})\widehat{V}_{d-1}^{(2i-1)}\right\|_F^2 \\
&- \left\|\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-2}} \otimes \widehat{U}_{1}^{(2i)\top})[\mathcal{Y}]_{d-2}(\widehat{V}_d^{(2i-1)} \otimes I_{p_{d-1}})\widehat{V}_{d-1\perp}^{(2i-1)}\right\|_F^2 \\
\leqslant& \|[\mathcal{Y}]_1\|_F^2 - \left\|\widehat{U}_{d-2}^{(2i)\top}(I_{p_{d-2}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-2}} \otimes \widehat{U}_{1}^{(2i)\top})[\mathcal{Y}]_{d-2}(\widehat{V}_d^{(2i-1)} \otimes I_{p_{d-1}})\widehat{V}_{d-1}^{(2i-1)}\right\|_F^2 \\
=& \|[\mathcal{Y}]_1\|_F^2 - \left\|(I_{p_{d-2}} \otimes \widehat{U}_{d-3}^{(2i)\top}) \cdots (I_{p_2\cdots p_{d-2}} \otimes \widehat{U}_{1}^{(2i)\top})[\mathcal{Y}]_{d-2}(\widehat{V}_d^{(2i-1)} \otimes I_{p_{d-1}})\widehat{V}_{d-1}^{(2i-1)}\right\|_F^2 \\
\leqslant& \cdots \\
\leqslant& \|[\mathcal{Y}]_1\|_F^2 - \left\|[\mathcal{Y}]_1(\widehat{V}_d^{(2i-1)} \otimes I_{p_2 \ldots p_{d-1}}) \cdots (\widehat{V}_3^{(2i-1)} \otimes I_{p_2})\widehat{V}_2^{(2i-1)}\right\|_F^2 \\
=& \left\|[\mathcal{Y}]_1 \left(I_{p_2\cdots p_d} - P_{(\widehat{V}_d^{(2i-1)} \otimes I_{p_2 \ldots p_{d-1}}) \cdots (\widehat{V}_3^{(2i-1)} \otimes I_{p_2})\widehat{V}_2^{(2i-1)}}\right)\right\|_F^2
\end{aligned}$$

$$=\left\|\mathcal{Y}-\widehat{\mathcal{X}}^{(2i-1)}\right\|_F^2.$$

Similarly, we can prove (4.11) holds for $k = 2i, i \geqslant 0$.

5.3.3 Proof of Theorem 4.3

Without loss of generality, we assume $\sigma^2 = 1$. We still let $\widehat{U}_i$, $\widehat{V}_i$, R_i and $\widetilde{R}_i$ denote $\widehat{U}_i^{(0)}$, $\widehat{V}_i^{(1)}$, $R_i^{(0)}$ and $\widetilde{R}_i^{(0)}$, respectively.

Lemma 5.3.2 Part 4 immediately shows that (4.15) holds with probability at least $1 - Ce^{-cp}$. Next, we show that with probability at least $1 - Ce^{-cp}$,

$$\left\|\sin\Theta(\widehat{U}_k, \widetilde{U}_k)\right\| \leqslant C\frac{\sqrt{\sum_{i=1}^{k-1} p_i r_{i-1} r_i} + \sqrt{p_k r_{k-1}} + \sqrt{p_{k+1}\cdots p_d}}{\lambda_k} \leqslant \frac{1}{2}, \quad \forall 1 \leqslant k \leqslant d-1. \tag{5.249}$$

Recall that

$$\widehat{U}_1 = \mathrm{SVD}^L_{r_1}([\mathcal{Y}]_1), \quad [\mathcal{Y}]_1 = [\mathcal{X}]_1 + [\mathcal{Z}]_1,$$

where $[\mathcal{X}]_1 \in \mathbb{R}^{p_1 \times p_{-1}}$ satisfying $\mathrm{rank}([\mathcal{X}]_1) = r_1$, $[\mathcal{Z}]_1 \in \mathbb{R}^{p_1 \times p_{-1}}$, by Lemmas 5.3.3 and 5.3.2, with probability $1 - Ce^{-cp}$, we have

$$\|\widehat{U}_{1\perp}^\top [\mathcal{X}]_1\| \leqslant 2\|[\mathcal{Z}]_1\| \leqslant C(p_1^{1/2} + (p_2\cdots p_d)^{1/2}).$$

Therefore, with probability at least $1 - Ce^{-cp}$,

$$\left\|\sin\Theta(\widehat{U}_1, U_1)\right\| \leqslant \frac{\left\|\widehat{U}_{1\perp}^\top U_1 U_1^\top [\mathcal{X}]_1\right\|}{s_{r_1}(U_1^\top [\mathcal{X}]_1)} = \frac{\left\|\widehat{U}_{1\perp}^\top [\mathcal{X}]_1\right\|}{s_{r_1}([\mathcal{X}]_1)} \leqslant C\frac{\sqrt{p_1} + \sqrt{p_2\cdots p_d}}{\lambda_1}.$$

For $2 \leqslant i \leqslant j \leqslant d-1$, by the definition of $\widetilde{U}_i$ and Lemma 4.3.2, we have

$$\begin{aligned}[\mathcal{X}]_j =&(I_{p_2\cdots p_j} \otimes [\mathcal{X}]_1)A^{(p_2\cdots p_j, p_{j+1}\cdots p_d)} = (I_{p_2\cdots p_j} \otimes (P_{U_1}[\mathcal{X}]_1))A^{(p_2\cdots p_j, p_{j+1}\cdots p_d)}\\ =&(I_{p_2\cdots p_j} \otimes P_{U_1})(I_{p_2\cdots p_j} \otimes [\mathcal{X}]_1)A^{(p_2\cdots p_j, p_{j+1}\cdots p_d)} = (I_{p_2\cdots p_j} \otimes U_1)(I_{p_2\cdots p_j} \otimes U_1^\top)[\mathcal{X}]_j\end{aligned} \tag{5.250}$$

and

$$
\begin{aligned}
&\left(I_{p_i\cdots p_j}\otimes\widehat{U}_{i-1}^{\top}\right)\cdots\left(I_{p_2\cdots p_j}\otimes\widehat{U}_1^{\top}\right)[\mathcal{X}]_j\\
=&\left(I_{p_{i+1}\cdots p_j}\otimes(I_{p_i}\otimes\widehat{U}_{i-1}^{\top})\right)\cdots\left(I_{p_{i+1}\cdots p_j}\otimes(I_{p_2\cdots p_i}\otimes\widehat{U}_1^{\top})\right)(I_{p_{i+1}\cdots p_j}\otimes[\mathcal{X}]_i)A^{(p_{i+1}\cdots p_j,p_{j+1}\cdots p_d)}\\
=&\left(I_{p_{i+1}\cdots p_j}\otimes\left((I_{p_i}\otimes\widehat{U}_{i-1}^{\top})\cdots(I_{p_2\cdots p_i}\otimes\widehat{U}_1^{\top})[\mathcal{X}]_i\right)\right)A^{(p_{i+1}\cdots p_j,p_{j+1}\cdots p_d)}\\
=&\left(I_{p_{i+1}\cdots p_j}\otimes\left(P_{\widetilde{U}_i}(I_{p_i}\otimes\widehat{U}_{i-1}^{\top})\cdots(I_{p_2\cdots p_i}\otimes\widehat{U}_1^{\top})[\mathcal{X}]_i\right)\right)A^{(p_{i+1}\cdots p_j,p_{j+1}\cdots p_d)}\\
=&\left(I_{p_{i+1}\cdots p_j}\otimes P_{\widetilde{U}_i}\right)\left(I_{p_{i+1}\cdots p_j}\otimes\left((I_{p_i}\otimes\widehat{U}_{i-1}^{\top})\cdots(I_{p_2\cdots p_i}\otimes\widehat{U}_1^{\top})[\mathcal{X}]_i\right)\right)A^{(p_{i+1}\cdots p_j,p_{j+1}\cdots p_d)}\\
=&\left(I_{p_{i+1}\cdots p_j}\otimes\widetilde{U}_i\right)\left(I_{p_{i+1}\cdots p_j}\otimes\widetilde{U}_i^{\top}\right)\left(I_{p_i\cdots p_j}\otimes\widehat{U}_{i-1}^{\top}\right)\cdots\left(I_{p_2\cdots p_j}\otimes\widehat{U}_1^{\top}\right)[\mathcal{X}]_j,
\end{aligned}
\tag{5.251}
$$

where $I_{p_{i+1}\cdots p_j}=1$ if $i=j$. Let

$$
L_k=\left\|\sin\Theta\left(\widetilde{U}_k,\widehat{U}_k\right)\right\|,\quad 2\leqslant k\leqslant d-1.
$$

For $k=2$, by (5.250) and Lemma 5.3.1, with probability at least $1-Ce^{-cp}$,

$$
\begin{aligned}
s_{r_2}\left((I_{p_2}\otimes\widehat{U}_1^{\top})[\mathcal{X}]_2\right)\geqslant&s_{\min}\left((I_{p_2}\otimes\widehat{U}_1^{\top})(I_{p_2}\otimes U_1)\right)s_{r_2}([\mathcal{X}]_2)\\
=&s_{\min}(\widehat{U}_1^{\top}U_1)\lambda_2\\
=&\sqrt{1-\|\sin\Theta(\widehat{U}_1,U_1)\|^2}\lambda_2\\
\geqslant&\sqrt{\frac{3}{4}}\lambda_2.
\end{aligned}
$$

Since $\widehat{U}_2=\mathrm{SVD}_{r_2}^{L}((I_{p_2}\otimes\widehat{U}_1^{\top})[\mathcal{Y}]_2)$, and $(I_{p_2}\otimes\widehat{U}_1^{\top})[\mathcal{Y}]_2=(I_{p_2}\otimes\widehat{U}_1^{\top})[\mathcal{X}]_2+(I_{p_2}\otimes\widehat{U}_1^{\top})[\mathcal{Z}]_2$, by Lemma 5.3.3 and Lemma 5.3.1, we know that with probability at least $1-Ce^{-cpr}$,

$$
\|\widehat{U}_{2\perp}^{\top}(I_{p_2}\otimes\widehat{U}_1^{\top})[\mathcal{X}]_2\|\leqslant2\|(I_{p_2}\otimes\widehat{U}_1^{\top})[\mathcal{Z}]_2\|\leqslant C(\sqrt{p_2r_1}+(p_3\cdots p_d)^{1/2}+\sqrt{p_1r_1}).
$$

Combine the two previous inequalities together and recall that $\widetilde{U}_2$ is the left singular space of $(I_{p_2} \otimes \widehat{U}_1^\top)[\mathcal{X}]_2$, we have

$$\begin{aligned}\left\|\sin\Theta\left(\widehat{U}_2, \widetilde{U}_2\right)\right\| \leqslant & \frac{\|\widehat{U}_{2\perp}^\top \widetilde{U}_2 \widetilde{U}_2^\top (I_{p_2} \otimes \widehat{U}_1^\top)[\mathcal{X}]_2\|}{s_{r_2}\left(\widetilde{U}_2^\top (I_{p_2} \otimes \widehat{U}_1^\top)[\mathcal{X}]_2\right)} \\ = & \frac{\|\widehat{U}_{2\perp}^\top (I_{p_2} \otimes \widehat{U}_1^\top)[\mathcal{X}]_2\|}{s_{r_2}\left((I_{p_2} \otimes \widehat{U}_1^\top)[\mathcal{X}]_2\right)} \\ \leqslant & C\frac{\sqrt{p_1 r_1} + \sqrt{p_2 r_1} + (p_3 \cdots p_d)^{1/2}}{\lambda_2}\end{aligned}$$

with probability at least $1 - Ce^{-cp}$.

Assume that (5.249) holds for $k \leqslant j-1$ with probability $1 - Ce^{-cp}$. For $k = j$, by Lemma 5.3.1 and (5.251), with probability at $1 - Ce^{-cp}$, we have

$$\begin{aligned}& s_{r_j}\left((I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right) \\ \geqslant & s_{\min}\left((I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_j} \otimes \widetilde{U}_{j-1})\right) s_{r_j}\left((I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right) \\ = & s_{\min}\left(\widehat{U}_{j-1}^\top \widetilde{U}_{j-1}\right) s_{r_j}\left((I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right) \\ \geqslant & s_{\min}\left(\widehat{U}_{j-1}^\top \widetilde{U}_{j-1}\right) s_{\min}\left((I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)(I_{p_{j-1}p_j} \otimes \widetilde{U}_{j-2})\right) \\ & \cdot s_{r_j}\left((I_{p_{j-2}p_{j-1}p_j} \otimes \widehat{U}_{j-3}^\top) \cdots (I_{p_2 \cdots p_{j-1} p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right) \\ \geqslant & \cdots \\ \geqslant & s_{\min}\left(\widehat{U}_{j-1}^\top \widetilde{U}_{j-1}\right) \cdots s_{\min}\left(\widehat{U}_1^\top \widetilde{U}_1\right) s_{r_j}([\mathcal{X}]_j) \\ = & \sqrt{1 - L_{j-1}^2} \cdots \sqrt{1 - L_1^2} \lambda_j \\ \geqslant & (\sqrt{3/4})^{j-1} \lambda_j \geqslant c\lambda_j.\end{aligned} \tag{5.252}$$

In the last inequality, we used the fact that d is a fixed number and $(\sqrt{3/4})^{j-1} \geqslant (\sqrt{3/4})^{d-1} \geqslant c$.

By the definition of $\widehat{U}_j$ and Lemma 4.3.3, we have

$$\widehat{U}_j = \mathrm{SVD}^L_{r_j}\left((I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_{j-1}p_j} \otimes \widehat{U}_1^\top)[\mathcal{Y}]_j\right).$$

Note that

$$\begin{aligned}&(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_{j-1}p_j} \otimes \widehat{U}_1^\top)[\mathcal{Y}]_j\\ =&(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_{j-1}p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\\ &+(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_{j-1}p_j} \otimes \widehat{U}_1^\top)[\mathcal{Z}]_j,\end{aligned}$$

by Lemma 5.3.3, with probability at least $1 - e^{-cpr^2}$,

$$\begin{aligned}&\left\|\widehat{U}_{j\perp}^\top(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_{j-1}p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right\|\\ \leqslant&2\left\|(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_{j-1}p_j} \otimes \widehat{U}_1^\top)[\mathcal{Z}]_j\right\|\\ \leqslant&C\left(\left(\sum_{i=1}^{j-1} p_i r_{i-1} r_i\right)^{1/2} + (p_j r_{j-1})^{1/2} + (p_{j+1}\cdots p_d)^{1/2}\right).\end{aligned}$$

Therefore, with probability at least $1 - Ce^{-cp}$,

$$\begin{aligned}\left\|\sin\Theta\left(\widehat{U}_j, \widetilde{U}_j\right)\right\| \leqslant&\frac{\left\|\widehat{U}_{j\perp}^\top\widetilde{U}_j\widetilde{U}_j^\top(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_{j-1}p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right\|}{s_{r_j}\left(\widetilde{U}_j^\top(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_{j-1}p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right)}\\ =&\frac{\left\|\widehat{U}_{j\perp}^\top(I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_{j-1}p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right\|}{s_{r_j}\left((I_{p_j} \otimes \widehat{U}_{j-1}^\top)(I_{p_{j-1}p_j} \otimes \widehat{U}_{j-2}^\top)\cdots(I_{p_2\cdots p_{j-1}p_j} \otimes \widehat{U}_1^\top)[\mathcal{X}]_j\right)}\\ \leqslant&C\frac{\left(\sum_{i=1}^{j-1} p_i r_{i-1} r_i\right)^{1/2} + (p_j r_{j-1})^{1/2} + (p_{j+1}\cdots p_d)^{1/2}}{\lambda_j}.\end{aligned}$$

Therefore, (4.13) holds with probability $1 - Ce^{-cp}$.

Finally, we consider (4.14). Let $\mathcal{E}_0 = \{(4.13) \text{ and } (4.15) \text{ hold}\}$. Without loss of generality, we only show that under $\mathcal{E}_0$,

$$\left\|\sin\Theta\left(\widehat{V}_k, \widetilde{V}_k\right)\right\| \leqslant C\frac{\sqrt{\sum_{i=1}^{d} p_i r_{i-1} r_i}}{\lambda_{k-1}} \leqslant \frac{1}{2}, \quad \forall 2 \leqslant k \leqslant d. \tag{5.253}$$

In fact, (5.253) can be proved by induction. Let $V_d \in \mathbb{R}^{p_d \times r_{d-1}}$ be the right singular space of $[\mathcal{X}]_{d-1}$. Then there exists an orthogonal matrix $\widetilde{Q}_{d-1} \in \mathbb{O}_{r_{d-1}}$ such that

$$V_d \widetilde{Q}_{d-1} = \mathrm{SVD}^R\big(\widehat{U}_{d-1}^\top (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^\top) \cdots (I_{p_{d-1} \dots p_2} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{d-1}\big).$$

Similarly to (5.252), under $\mathcal{E}_0$,

$$s_{r_{d-1}}\left(\widehat{U}_{d-1}^\top (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^\top) \cdots (I_{p_{d-1} \dots p_2} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{d-1}\right) \geqslant \left(\sqrt{3/4}\right)^{d-1} \lambda_{d-1} \geqslant c\lambda_{d-1}.$$

Therefore, by Lemma 5.3.3, under $\mathcal{E}_0$,

$$\begin{aligned}
\left\|\sin\Theta\left(\widehat{V}_d, V_d\right)\right\| &= \left\|\sin\Theta\left(\widehat{V}_d, V_d \widetilde{Q}_{d-1}\right)\right\| \\
&\leqslant \frac{\left\|\widehat{U}_{d-1}^\top (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^\top) \cdots (I_{p_{d-1} \dots p_2} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{d-1} \widehat{V}_{d\perp}^\top\right\|}{s_{r_{d-1}}\left(\widehat{U}_{d-1}^\top (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^\top) \cdots (I_{p_{d-1} \dots p_2} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{d-1}\right)} \\
&\leqslant \frac{2\left\|\widehat{U}_{d-1}^\top (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^\top) \cdots (I_{p_{d-1} \dots p_2} \otimes \widehat{U}_1^\top)[\mathcal{Z}]_{d-1}\right\|}{s_{r_{d-1}}\left(\widehat{U}_{d-1}^\top (I_{p_{d-1}} \otimes \widehat{U}_{d-2}^\top) \cdots (I_{p_{d-1} \dots p_2} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{d-1}\right)} \\
&\leqslant C\frac{\sqrt{\sum_{i=1}^{d} p_i r_{i-1} r_i}}{\lambda_{d-1}}.
\end{aligned}$$

Suppose (5.253) holds for $j+1 \leqslant k \leqslant d$. For $k = j$, since $\widetilde{V}_j$ is the right singular space of $[\mathcal{X}]_{j-1}(\widehat{V}_d \otimes I_{p_j \dots p_{d-1}}) \cdots (\widehat{V}_{j+1} \otimes I_{p_j})$, there exists $\widetilde{Q}_{j-1} \in \mathbb{O}_{r_{j-1}}$ such that

$$\widetilde{V}_j \widetilde{Q}_{j-1} = \mathrm{SVD}^R\big(\widehat{U}_{j-1}^\top (I_{p_{j-1}} \otimes \widehat{U}_{j-2}^\top) \cdots (I_{p_2 \cdots p_{j-1}} \otimes \widehat{U}_1^\top)[\mathcal{X}]_{j-1}(\widehat{V}_d \otimes I_{p_j \dots p_{d-1}}) \cdots (\widehat{V}_{j+1} \otimes I_{p_j})\big).$$

By Lemma 5.3.1, (5.250), (5.251) and (5.252), under $\mathcal{E}_0$,

$$\begin{aligned}
&s_{r_{j-1}}\left(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\right)\\
\geqslant & s_{r_{j-1}}\left(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+2}\otimes I_{p_jp_{j+1}})(\widetilde{V}_{j+1}\otimes I_{p_j})\right)\\
&\cdot s_{\min}\left((\widetilde{V}_{j+1}^{\top}\otimes I_{p_j})(\widehat{V}_{j+1}\otimes I_{p_j})\right)\\
= & s_{r_{j-1}}\left(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+2}\otimes I_{p_jp_{j+1}})\right)\\
&\cdot s_{\min}(\widetilde{V}_{j+1}^{\top}\widehat{V}_{j+1})\\
\geqslant & \cdots\\
\geqslant & s_{r_{j-1}}\left(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}\right)s_{\min}(\widetilde{V}_d^{\top}\widehat{V}_d)\cdots s_{\min}(\widetilde{V}_{j+1}^{\top}\widehat{V}_{j+1})\\
\geqslant & s_{\min}(\widehat{U}_{j-1}^{\top}\widetilde{U}_{j-1})s_{r_{j-1}}\left((I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}\right)s_{\min}(\widetilde{V}_d^{\top}\widehat{V}_d)\cdots s_{\min}(\widetilde{V}_{j+1}^{\top}\widehat{V}_{j+1})\\
\geqslant & \left(\sqrt{\frac{3}{4}}\right)^{j-1}\lambda_{j-1}\cdot\left(\sqrt{\frac{3}{4}}\right)^{d-j}\geqslant c\lambda_{j-1}.
\end{aligned}$$

Note that $\widehat{V}_j\in\mathbb{O}_{p_jr_j,r_{j-1}}$ is the right singular space of $\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{Y}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})$ and

$$\begin{aligned}
&\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{Y}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\\
=&\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\\
&+\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{Z}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j}),
\end{aligned}$$

By Lemma 5.3.3, under $\mathcal{E}_0$,

$$\begin{aligned}
&\left\|\sin\Theta\left(\widehat{V}_j,\widetilde{V}_j\right)\right\|=\left\|\sin\Theta\left(\widehat{V}_j,\widetilde{V}_j\widetilde{Q}_{j-1}\right)\right\|\\
\leqslant&\frac{\left\|\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\widehat{V}_{j\perp}\right\|}{s_{r_{j-1}}\left(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\right)}\\
\leqslant&\frac{2\left\|\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{Z}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\right\|}{s_{r_{j-1}}\left(\widehat{U}_{j-1}^{\top}(I_{p_{j-1}}\otimes\widehat{U}_{j-2}^{\top})\cdots(I_{p_2\cdots p_{j-1}}\otimes\widehat{U}_1^{\top})[\boldsymbol{\mathcal{X}}]_{j-1}(\widehat{V}_d\otimes I_{p_j\ldots p_{d-1}})\cdots(\widehat{V}_{j+1}\otimes I_{p_j})\right)}
\end{aligned}$$

$$\leqslant C\frac{\left(\sum_{i=1}^{d} p_i r_i r_{i-1}\right)^{1/2}}{\lambda_{j-1}}.$$

Therefore, under $\mathcal{E}_0$, (5.253) holds.

Thus, we have finished the proof of Theorem 4.3.

5.3.4 Proof of Corollary 4.4.1

Let $Q = \{(4.15), (5.249) \text{ hold}\}$, then $\mathbb{P}(Q^c) \leqslant C\exp(-cp)$ and

$$\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\|_F^2 \leqslant C\sum_{i=1}^{d} p_i r_i r_{i-1} \quad \text{under } Q.$$

Under Q^c, due to the property of projection matrices, we know that

$$\left\|\widehat{\mathcal{X}}^{(t)}\right\|_F \leqslant \|\mathcal{Y}\|_F \leqslant \|\mathcal{X}\|_F + \|\mathcal{Z}\|_F.$$

Moreover,

$$\begin{aligned}
&\mathbb{E}\left\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\right\|_F^4 \leqslant C\left(\mathbb{E}\left\|\widehat{\mathcal{X}}^{(t)}\right\|_F^4 + \|\mathcal{X}\|_F^4\right) \leqslant C\|\mathcal{X}\|_F^4 + C\mathbb{E}\|\mathcal{Z}\|_F^4\\
\leqslant& C\exp(4c_0p) + C\mathbb{E}\left(\chi^2_{p_1\cdots p_d}\right)^2 \leqslant C\exp(4c_0p) + C(p_1\cdots p_d)^2\\
\leqslant& C\exp(4c_0p) + C\exp\left(2c_0p\right) \leqslant C\exp(4c_0p).
\end{aligned}$$

Therefore, we have the following upper bound for the Frobenius norm risk of $\widehat{\mathcal{X}}$:

$$\begin{aligned}
&\mathbb{E}\left\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\right\|_F^2 = \mathbb{E}\left\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\right\|_F^2 1_Q + \mathbb{E}\left\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\right\|_F^2 1_{Q^c}\\
\leqslant& C\sum_{i=1}^{d} p_i r_i r_{i-1} + \sqrt{\mathbb{E}\left\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\right\|_F^4 \cdot \mathbb{P}(Q^c)}\\
\leqslant& C\sum_{i=1}^{d} p_i r_i r_{i-1} + C\exp\left((4c_0 - c)p/2\right).
\end{aligned}$$

By selecting $c_0 < c/4$, we have

$$\mathbb{E}\left\|\widehat{\mathcal{X}}^{(t)} - \mathcal{X}\right\|_F^2 \leqslant C\sum_{i=1}^{d} p_i r_i r_{i-1}.$$

Therefore, we have finished the proof of Corollary 4.4.1.

5.3.5 Proof of Theorem 4.4

Without loss of generality, assume $\sigma^2 = 1$. Since d is a fixed number, we only need to show that for any $1 \leqslant i \leqslant d$,

$$\inf_{\widehat{\mathcal{X}}} \sup_{\mathcal{X}\in\mathcal{F}_{\mathbf{p},\mathbf{r}}(\lambda)} \mathbb{E}\left\|\widehat{\mathcal{X}} - \mathcal{X}\right\|_F^2 \geqslant c p_i r_i r_{i-1}. \tag{5.254}$$

Suppose $\mathcal{X}$ can be written as (4.1), $U_j \in \mathbb{R}^{(p_j r_{j-1})\times r_j}$ and $V_j \in \mathbb{R}^{(p_j r_j)\times r_{j-1}}$ are reshaped from $\mathcal{G}_j \in \mathbb{R}^{r_{j-1}\times p_j\times r_j}$, $G_1 = U_1$, $G_d = V_d$. For any $1 \leqslant i \leqslant d-1$, by Lemma 4.3.1, we have

$$[\mathcal{X}]_i = (I_{p_2\cdots p_i} \otimes U_1)\cdots(I_{p_i}\otimes U_{i-1})U_i V_{i+1}^\top \left(V_{i+2}^\top \otimes I_{p_{i+1}}\right)\cdots\left(V_d^\top \otimes I_{p_{i+1}\cdots p_{d-1}}\right). \tag{5.255}$$

For all $j \neq i, 1 \leqslant j \leqslant d-1$, let $U_j \overset{iid}{\sim} N(0,1), V_d \overset{iid}{\sim} N(0,1)$ and $U_1,\ldots,U_{i-1},U_{i+1},\ldots,U_{d-1},V_d$ are all independent. By Lemma 5.3.1, for any $1 \leqslant j \leqslant d-1$, we have

$$s_{r_j}\left((I_{p_2\cdots p_j}\otimes U_1)\cdots(I_{p_j}\otimes U_{j-1})U_j\right) \geqslant s_{\min}\left(I_{p_2\cdots p_j}\otimes U_1\right)\cdots s_{\min}(U_j) = s_{r_1}(U_1)\cdots s_{r_j}(U_j),$$

Similarly,

$$s_{r_j}\left(V_{j+1}^\top\left(V_{j+2}^\top\otimes I_{p_{j+1}}\right)\cdots\left(V_d^\top\otimes I_{p_{j+1}\cdots p_{d-1}}\right)\right) \geqslant s_{r_j}(V_{j+1})\cdots s_{r_{d-1}}(V_d).$$

Moreover, Lemma 5.3.1 Part 1 tells us

$$s_{r_j}\left((I_{p_2\cdots p_j}\otimes U_1)\cdots(I_{p_j}\otimes U_{j-1})U_j V_{j+1}^\top\left(V_{j+2}^\top\otimes I_{p_{j+1}}\right)\cdots\left(V_d^\top\otimes I_{p_{j+1}\cdots p_{d-1}}\right)\right)$$

$$\begin{aligned}&\geqslant s_{r_j}\left(\left(I_{p_2\cdots p_j}\otimes U_1\right)\cdots\left(I_{p_j}\otimes U_{j-1}\right)U_j\right)s_{r_j}\left(V_{j+1}^{\top}\left(V_{j+2}^{\top}\otimes I_{p_{j+1}}\right)\cdots\left(V_d^{\top}\otimes I_{p_{j+1}\cdots p_{d-1}}\right)\right)\\&\geqslant s_{r_1}(U_1)\cdots s_{r_j}(U_j)s_{r_j}(V_{j+1})\cdots s_{r_{d-1}}(V_d).\end{aligned}\tag{5.256}$$

Recall that V_j is reshaped from U_j for all $1\leqslant j\leqslant d-1$, by Vershynin (2010)[Corollary 5.35], we know that with probability at least $1-Ce^{-cp}$, for all $1\leqslant j\leqslant d-1, j\neq i$,

$$\begin{aligned}\frac{\sqrt{p_jr_{j-1}}}{4}&\leqslant\sqrt{p_jr_{j-1}}-\sqrt{r_j}-\frac{\sqrt{p_jr_{j-1}}}{25}\leqslant s_{r_j}(U_j)\leqslant s_1(U_j)\\&\leqslant\sqrt{p_jr_{j-1}}+\sqrt{r_j}+\frac{\sqrt{p_jr_{j-1}}}{25}\leqslant 2\sqrt{p_jr_{j-1}},\\\frac{\sqrt{p_jr_j}}{4}&\leqslant s_{r_{j-1}}(V_j)\leqslant s_1(V_j)\leqslant 2\sqrt{p_jr_j},\quad\text{and}\quad\frac{\sqrt{p_d}}{4}\leqslant s_{r_{d-1}}(V_d)\leqslant s_{r_1}(V_d)\leqslant 2\sqrt{p_d}.\end{aligned}\tag{5.257}$$

For a fixed $U_0\in\mathbb{O}_{p_ir_{i-1},r_i}$, define the following ball with radius $\varepsilon>0$,

$$B(U_0,\varepsilon)=\left\{U'\in\mathbb{O}_{p_ir_{i-1},r_i}:\|\sin\Theta(U',U_0)\|_F\leqslant\varepsilon\right\}.$$

By Lemma 1 in Cai et al. (2013), for $0<\alpha<1$ and $0<\varepsilon\leqslant 1$, there exist $\widetilde{U}_i^{(1)'},\ldots,\widetilde{U}_i^{(m)'}\subseteq B(U_0,\varepsilon)$ such that

$$m\geqslant\left(\frac{c_0}{\alpha}\right)^{r_i(p_ir_{i-1}-r_i)},\quad\min_{1\leqslant j\neq k\leqslant m}\left\|\sin\Theta\left(\widetilde{U}_i^{(j)'},\widetilde{U}_i^{(k)'}\right)\right\|_F\geqslant\alpha\varepsilon.$$

By Lemma 1 in Cai and Zhang (2018), one can find a rotation matrix $O_k\in\mathbb{O}_{r_i}$ such that

$$\|U_0-\widetilde{U}_i^{(k)'}O_k\|_F\leqslant\sqrt{2}\left\|\sin\Theta\left(U_0,\widetilde{U}_i^{(k)'}\right)\right\|_F\leqslant\sqrt{2}\varepsilon.$$

Let $\widetilde{U}_i^{(k)}=\widetilde{U}_i^{(k)'}O_k$, we have

$$\left\|\widetilde{U}_i^{(k)}-U_0\right\|_F\leqslant\sqrt{2}\varepsilon,\quad\left\|\sin\Theta\left(\widetilde{U}_i^{(j)},\widetilde{U}_i^{(k)}\right)\right\|_F\geqslant\alpha\varepsilon,\quad 1\leqslant j<k\leqslant m.$$

Let $U_i^{(k)} = S+\widetilde{U}_i^{(k)}$, where $S \overset{iid}{\sim} N(0,\tau^2)$. Set $\tau \geqslant 8/\sqrt{p_i}$, Vershynin (2010)[Corollary 5.35] shows that with probability at least $1-Ce^{-cp}$,

$$\begin{aligned}\frac{\tau\sqrt{p_i r_{i-1}}}{8} &\leqslant \tau\left(\sqrt{p_i r_{i-1}} - \sqrt{r_i} - \frac{\sqrt{p_i r_{i-1}}}{25}\right) - 1 \leqslant s_{r_i}\left(S\right) - s_1\left(\widetilde{U}_i^{(k)}\right) \leqslant s_{r_i}\left(U_i^{(k)}\right) \\ \leqslant & s_1\left(U_i^{(k)}\right) \leqslant s_1\left(S\right) + s_1\left(\widetilde{U}_i^{(k)}\right) \leqslant \tau\left(\sqrt{p_i r_{i-1}} + \sqrt{r_i} + \frac{\sqrt{p_i r_{i-1}}}{25}\right) + 1 \leqslant 2\tau\sqrt{p_i r_{i-1}}.\end{aligned} \tag{5.258}$$

If $2 \leqslant i \leqslant d-1$, since $V_i^{(k)}$ is reshaped from $U_i^{(k)}$, we know that $V_i^{(k)} = T + \widetilde{V}_i^{(k)}$, where $T \overset{iid}{\sim} N(0,\tau^2)$, and $\widetilde{V}_i^{(k)}$ is realigned from $\widetilde{U}_i^{(k)}$. Notice that

$$s_1(\widetilde{V}_i^{(k)}) = \|\widetilde{V}_i^{(k)}\| \leqslant \|\widetilde{V}_i^{(k)}\|_F = \|\widetilde{U}_i^{(k)}\|_F = r_i,$$

Since $\tau \geqslant 8/\sqrt{p_i}$, by Vershynin (2010)[Corollary 5.35], with probability at least $1-Ce^{-cp_i r_i}$,

$$\begin{aligned}\frac{\tau\sqrt{p_i r_i}}{8} &\leqslant \tau\left(\sqrt{p_i r_i} - \sqrt{r_{i-1}} - \frac{\sqrt{p_i r_i}}{25}\right) - \sqrt{r_i} \leqslant s_{r_i}\left(T\right) - s_1\left(\widetilde{V}_i^{(k)}\right) \leqslant s_{r_i}\left(V_i^{(k)}\right) \\ \leqslant & s_1\left(V_i^{(k)}\right) \leqslant s_1\left(T\right) + s_1\left(\widetilde{V}_i^{(k)}\right) \leqslant \tau\left(\sqrt{p_i r_i} + \sqrt{r_{i-1}} + \frac{\sqrt{p_i r_i}}{25}\right) + \sqrt{r_i} \leqslant 2\tau\sqrt{p_i r_i}.\end{aligned} \tag{5.259}$$

Choose fixed $U_1, \cdots, U_{i-1}, V_{i+1}, \cdots, V_d, S$ such that (5.257), (5.258) and (5.259) hold. Let

$$[\mathcal{X}^{(k)}]_i = (I_{p_2\cdots p_i} \otimes U_1)\cdots(I_{p_i} \otimes U_{i-1})U_i^{(k)}V_{i+1}^\top\left(V_{i+2}^\top \otimes I_{p_{i+1}}\right)\cdots\left(V_d^\top \otimes I_{p_{i+1}\cdots p_{d-1}}\right) \tag{5.260}$$

and $\mathcal{X}^{(k)} \in \mathbb{R}^{p_1\times\cdots\times p_d}$ is the corresponding tensor. (5.256), (5.257), (5.258) and (5.259) together show that

$$\sigma_{r_j}([\mathcal{X}^{(k)}]_j) \geqslant \tau\prod_{k=1}^{j}\frac{\sqrt{p_k r_{k-1}}}{8}\prod_{k=j+1}^{d}\frac{\sqrt{p_k r_k}}{8} = \tau\frac{\sqrt{p_1\cdots p_d r_1\cdots r_{d-1}}}{C\sqrt{r_j}}. \tag{5.261}$$

By setting $\tau = \frac{C \max_{1\leqslant i\leqslant d-1} \lambda_i \max_{1\leqslant j\leqslant d-1} \sqrt{r_j}}{\sqrt{p_1 \cdots p_d r_1 \cdots r_{d-1}}} \vee 8 \max_{1\leqslant i\leqslant d-1} \sqrt{1/p_i}$, we have

$$\sigma_{r_j}\left([X^{(k)}]_j\right) \geqslant \lambda_j, \quad \forall 1 \leqslant j \leqslant d-1.$$

For $1 \leqslant k < j \leqslant m$,

$$\begin{aligned}
&\|\mathcal{X}^{(k)} - \mathcal{X}^{(j)}\|_F^2 \\
=&\left\|(I_{p_2\cdots p_i} \otimes U_1) \cdots (I_{p_i} \otimes U_{i-1}) \left(U_i^{(k)} - U_i^{(j)}\right) V_{i+1}^\top \left(V_{i+2}^\top \otimes I_{p_{i+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{i+1}\cdots p_{d-1}}\right)\right\|_F^2 \\
\geqslant& s_{\min}^2\left((I_{p_2\cdots p_i} \otimes U_1) \cdots (I_{p_i} \otimes U_{i-1})\right) \left\|\left(U_i^{(k)} - U_i^{(j)}\right) V_{i+1}^\top \left(V_{i+2}^\top \otimes I_{p_{i+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{i+1}\cdots p_{d-1}}\right)\right\|_F^2 \\
=& s_{r_{i-1}}^2\left(\left(I_{p_2\cdots p_{i-1}} \otimes U_1\right) \cdots U_{i-1}\right) s_{r_i}^2\left(V_{i+1}^\top \left(V_{i+2}^\top \otimes I_{p_{i+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{i+1}\cdots p_{d-1}}\right)\right) \left\|U_i^{(k)} - U_i^{(j)}\right\|_F^2 \\
=& s_{r_{i-1}}^2\left(\left(I_{p_2\cdots p_{i-1}} \otimes U_1\right) \cdots U_{i-1}\right) s_{r_i}^2\left(V_{i+1}^\top \left(V_{i+2}^\top \otimes I_{p_{i+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{i+1}\cdots p_{d-1}}\right)\right) \left\|\widetilde{U}_i^{(k)} - \widetilde{U}_i^{(j)}\right\|_F^2 \\
\geqslant& s_{r_1}^2(U_1) \cdots s_{r_{i-1}}^2(U_{i-1}) s_{r_i}^2(V_{i+1}) \cdots s_{r_{d-1}}^2(V_d) \min_{O \in \mathbb{O}_{r_i}} \left\|\widetilde{U}_i^{(k)} - \widetilde{U}_i^{(j)} O\right\|_F^2 \\
\geqslant& \prod_{h=1}^{i-1} \frac{p_h r_{h-1}}{16} \prod_{l=i+1}^{d} \frac{p_l r_l}{16} \min_{O \in \mathbb{O}_{r_i}} \left\|\widetilde{U}_i^{(k)} - \widetilde{U}_i^{(j)} O\right\|_F^2 \\
\geqslant& \prod_{h=1}^{i-1} \frac{p_h r_{h-1}}{16} \prod_{l=i+1}^{d} \frac{p_l r_l}{16} \left\|\sin\Theta\left(\widetilde{U}_i^{(k)}, \widetilde{U}_i^{(j)}\right)\right\|_F^2 \\
\geqslant& c\left(\prod_{h=1}^{i-1} p_h r_{h-1} \prod_{l=i+1}^{d} p_l r_l\right) \alpha^2 \varepsilon^2. \qquad (5.262)
\end{aligned}$$

In addition, let $\mathcal{Y}^{(k)} = \mathcal{X}^{(k)} + \mathcal{Z}^{(k)}$ and $\mathcal{Z}^{(k)} \overset{iid}{\sim} N(0,1)$. The KL-divergence between distributions $\mathcal{Y}^{(k)}$ and $\mathcal{Y}^{(j)}$ is

$$\begin{aligned}
&D_{KL}\left(\mathcal{Y}^{(k)} \| \mathcal{Y}^{(j)}\right) = \frac{1}{2}\|\mathcal{X}^{(k)} - \mathcal{X}^{(j)}\|_F^2 \\
=&\frac{1}{2}\left\|(I_{p_2\cdots p_i} \otimes U_1) \cdots (I_{p_i} \otimes U_{i-1}) \left(U_i^{(k)} - U_i^{(j)}\right) V_{i+1}^\top \left(V_{i+2}^\top \otimes I_{p_{i+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{i+1}\cdots p_{d-1}}\right)\right\|_F^2 \\
\leqslant&\frac{1}{2}\|(I_{p_2\cdots p_i} \otimes U_1) \cdots (I_{p_i} \otimes U_{i-1})\|^2 \left\|V_{i+1}^\top \left(V_{i+2}^\top \otimes I_{p_{i+1}}\right) \cdots \left(V_d^\top \otimes I_{p_{i+1}\cdots p_{d-1}}\right)\right\|^2 \left\|U_i^{(k)} - U_i^{(j)}\right\|_F^2
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \frac{1}{2} s_1^2(U_1)\cdots s_1^2(U_{i-1}) s_1^2(V_{i+1})\cdots s_1^2(V_d)\left\|U_i^{(k)}-U_i^{(j)}\right\|_F^2\\
&\leqslant \frac{1}{2}\prod_{h=1}^{i-1}(4p_h r_{h-1})\prod_{l=i+1}^{d}(4p_l r_l)\left(\left\|U_i^{(k)}-U_0\right\|_F+\left\|U_i^{(k)}-U_0\right\|_F\right)^2\\
&\leqslant C\left(\prod_{h=1}^{i-1}(p_h r_{h-1})\prod_{l=i+1}^{d}(p_l r_l)\right)\varepsilon^2. \qquad (5.263)
\end{aligned}
$$

By generalized Fano's Lemma,

$$
\begin{aligned}
&\inf_{\widehat{\mathcal{X}}}\sup_{\mathcal{X}\in\{\mathcal{X}^{(k)}\}_{k=1}^m}\mathbb{E}\left\|\widehat{\mathcal{X}}-\mathcal{X}\right\|_F\\
&\geqslant c\sqrt{\prod_{h=1}^{i-1}p_h r_{h-1}\prod_{l=i+1}^{d}p_l r_l}\,\alpha\varepsilon\left(1-\frac{C\left(\prod_{h=1}^{i-1}(p_h r_{h-1})\prod_{l=i+1}^{d}(p_l r_l)\right)\varepsilon^2+\log 2}{r_i(p_i r_{i-1}-r_i)\log(c_0/\alpha)}\right)
\end{aligned}
$$

By setting $\varepsilon = c'\sqrt{\frac{r_i(p_i r_{i-1}-r_i)}{C\prod_{h=1}^{i-1}(p_h r_{h-1})\prod_{l=i+1}^{d}(p_l r_l)}} \leqslant \frac{1}{2}$, $\alpha = (c_0\wedge 1)/8$, we know that for any $1\leqslant i\leqslant d-1$,

$$
\inf_{\widehat{\mathcal{X}}}\sup_{\mathcal{X}\in\mathcal{F}_{\mathbf{p},\mathbf{r}}(\lambda)}\mathbb{E}\left\|\widehat{\mathcal{X}}-\mathcal{X}\right\|_F^2\geqslant\left(\inf_{\widehat{\mathcal{X}}}\sup_{\mathcal{X}\in\{\mathcal{X}^{(k)}\}_{k=1}^m}\mathbb{E}\left\|\widehat{\mathcal{X}}-\mathcal{X}\right\|_F\right)^2\geqslant c_1 r_i p_i r_{i-1}.
$$

For $i = d$, similarly to the case $i = 1$, we have

$$
\inf_{\widehat{\mathcal{X}}}\sup_{\mathcal{X}\in\mathcal{F}_{\mathbf{p},\mathbf{r}}(\lambda)}\mathbb{E}\left\|\widehat{\mathcal{X}}-\mathcal{X}\right\|_F^2\geqslant c_1 p_d r_{d-1}.
$$

Therefore, we have proved Theorem 4.4.

5.3.6 Proof of Proposition 4.5.1

Define $\tilde{G}_1 \in \mathbb{R}^{p\times r_1}, \tilde{\mathcal{G}}_k \in \mathbb{R}^{r_{k-1}\times p\times r_k}$, $\tilde{G}_d \in \mathbb{R}^{p\times r_{d-1}}$ such that

$$\begin{aligned}
\tilde{G}_{1,[i,l]} &= (G_1(i))_l, \quad \forall i \in [p], l \in [r_1],\\
\tilde{\mathcal{G}}_{k,[j,i,l]} &= \left(G_k(i, e_j^{(r_{k-1})})\right)_l, \quad \forall i \in [p], j \in [r_{k-1}], l \in [r_k], 2 \leqslant k \leqslant d-1,\\
\tilde{G}_{d,[i,l]} &= G_d(i, e_l^{(r_{d-1})}), \forall i \in [p], l \in [r_{d-1}]
\end{aligned}$$

where $e_i^{(k)}$ is the i-th canonical basis of $\mathbb{R}^k$. Then

$$\tilde{P}_1(X_{t+1}) = \tilde{G}^\top_{1,[X_{t+1},:]} \in \mathbb{R}^{r_1},$$

$$\begin{aligned}
\tilde{P}_2(X_{t+1}, X_{t+2}) =& G_2\left(X_{t+2}, \tilde{P}_1(X_{t+1})\right)\\
&\overset{\text{linear map}}{=} \sum_{j=1}^{r_1} G_2(X_{t+2}, e_j^{(r_1)}) \left(\tilde{P}_1(X_{t+1})\right)_j = \left(\tilde{G}_{1,[X_{t+1},:]}\tilde{\mathcal{G}}_{2,[:,X_{t+2},:]}\right)^\top,
\end{aligned}$$

By induction, for any $2 \leqslant k \leqslant d-1$,

$$\begin{aligned}
\tilde{P}_k(X_{t+1}, \ldots, X_{t+k}) &= G_k(X_{t+k}, \tilde{P}_{k-1}(X_{t+1}, \ldots, X_{t+k-1}))\\
&\overset{\text{linear map}}{=} \sum_{j=1}^{r_{k-1}} G_k(X_{t+k}, e_j^{(r_{k-1})}) \left(\tilde{P}_{k-1}(X_{t+1}, \ldots, X_{t+k-1})\right)_j\\
&= \tilde{\mathcal{G}}^\top_{k,[:,X_{t+k},:]} \tilde{P}_{k-1}(X_{t+1}, \ldots, X_{t+k-1})\\
&= \left(\tilde{G}_{1,[X_{t+1},:]}\tilde{\mathcal{G}}_{2,[:,X_{t+2},:]} \cdots \tilde{\mathcal{G}}_{k,[:,X_{t+k},:]}\right)^\top
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}\left(X_{t+d}|X_{t+1}, \ldots, X_{t+d-1}\right) =& G_d(X_{t+d}, \tilde{P}_{d-1}(X_{t+1}, \ldots, X_{t+d-1}))\\
=& \tilde{P}^\top_{d-1}(X_{t+1}, \ldots, X_{t+d-1}) \tilde{G}^\top_{d,[X_{t+d},:]}\\
=& \tilde{G}_{1,[X_{t+1},:]} \widetilde{\mathcal{G}}_{2,[:,X_{t+2},:]} \cdots \tilde{\mathcal{G}}_{d-1,[:,X_{t+d-1},:]} \tilde{G}^\top_{d,[X_{t+d},:]}.
\end{aligned}$$

Therefore,

$$\mathcal{P} = [\![\tilde{G}_1, \tilde{\mathcal{G}}_2, \ldots, \tilde{\mathcal{G}}_{d-1}, \tilde{G}_d]\!]$$

and has TT-rank $(r_1, \ldots, r_{d-1})$.

5.3.7 Proof of Proposition 4.5.2

Let $\mathcal{Z} = \widehat{\mathcal{P}}^{\text{emp}} - \mathcal{P}$, then $\mathbb{E}\mathcal{Z} = 0$. Let

$$\mathcal{T}^{(k)}_{i_1,\ldots,i_d} = 1_{\{X(i_1,\ldots,i_{d-1};k)=i_d\}}, \quad \forall 1 \leqslant k \leqslant n; 1 \leqslant i_1, \ldots, i_d \leqslant p$$

and

$$\mathcal{Z}^{(k)}_{i_1,\ldots,i_d} = \mathcal{T}^{(k)}_{i_1,\ldots,i_d} - \mathbb{P}\left(i_d | i_1, \ldots, i_{d-1}\right), \quad \forall 1 \leqslant k \leqslant n; 1 \leqslant i_1, \ldots, i_d \leqslant p.$$

Then $\mathbb{E}\mathcal{Z}^{(k)} = 0$. Moreover, by definition, for any $1 \leqslant j \leqslant d-1$, the rows of $\left[\mathcal{Z}^{(k)}\right]_j \in \mathbb{R}^{p^j \times p^{d-j}}$ are independent, and there exists a partition $\{\Omega^{(j)}_1, \ldots, \Omega^{(j)}_{p^{d-j-1}}\}$ of $\{1, \ldots, p^{d-j}\}$ satisfying $\left|\Omega^{(j)}_1\right| = \cdots = \left|\Omega^{(j)}_{p^{d-j-1}}\right| = p$, such that $\left(\left[\mathcal{Z}^{(k)}\right]_j\right)_{[:,\Omega^{(j)}_1]}, \ldots,$ $\left(\left[\mathcal{Z}^{(k)}\right]_j\right)_{[:,\Omega^{(j)}_{p^{d-j-1}}]}$ are independent and

$$\sum_{l \in \Omega^{(j)}_i} \left(\left[\mathcal{T}^{(k)}\right]_j\right)_{m,l} = 1, \quad \forall 1 \leqslant m \leqslant p^j, 1 \leqslant k \leqslant n.$$

Therefore,

$$\sum_{l \in \Omega^{(j)}_i} \left|\left(\left[\mathcal{Z}^{(k)}\right]_j\right)_{m,l}\right| \leqslant \sum_{l \in \Omega^{(j)}_i} \left(\left[\mathcal{T}^{(k)}\right]_j\right)_{m,l} + \mathbb{E} \sum_{l \in \Omega^{(j)}_i} \left(\left[\mathcal{T}^{(k)}\right]_j\right)_{m,l} = 2, \quad \forall 1 \leqslant m \leqslant p^j, 1 \leqslant k \leqslant n.$$

For any fixed $x_1 \in \mathbb{R}^{p^j}$ and $x_2 \in \mathbb{R}^{p^{d-j}}$ satisfying $\|x_1\|_2 = 1$ and $\|x\|_2 = 1$, we have

$$\left|\sum_{l \in \Omega^{(j)}_i} \left(\left[\mathcal{Z}^{(k)}\right]_j\right)_{m,l} (x_2)_l\right| \leqslant \max_{l \in \Omega^{(j)}_i} (x_2)_l \sum_{l \in \Omega^{(j)}_i} \left|\left(\left[\mathcal{Z}^{(k)}\right]_j\right)_{m,l}\right| \leqslant 2 \max_{l \in \Omega^{(j)}_i} (x_2)_l \leqslant 2 \left\|(x_2)_{\Omega^{(j)}_i}\right\|_2.$$

By (Wainwright, 2019, Exercise 2.4), $\sum_{l\in\Omega_i^{(j)}}\left(\left[\mathcal{Z}^{(k)}\right]_j\right)_{m,l}(x_2)_l$ is $2\left\|(x_2)_{\Omega_i^{(j)}}\right\|_2$-sub-Gaussian. Therefore,

$$x_1^\top\left[\mathcal{Z}^{(k)}\right]_j x_2=\sum_{m=1}^{p^j}(x_1)_m\sum_{i=1}^{p^{d-j-1}}\left(\sum_{l\in\Omega_i^{(j)}}\left(\left[\mathcal{Z}^{(k)}\right]_j\right)_{m,l}(x_2)_l\right)$$

is $\left(\sum_{m=1}^{p^j}(x_1)_m^2\sum_{i=1}^{p^{d-j-1}}4\left\|(x_2)_{\Omega_i^{(j)}}\right\|_2^2\right)^{1/2}=2\|x_1\|_2\|x_2\|_2=2$-sub-Gaussian. Notice that $\mathcal{Z}=\frac{1}{n}\sum_{k=1}^n\mathcal{Z}^{(k)}$, the Hoeffding bound (Wainwright, 2019, Proposition 2.5) shows that

$$\mathbb{P}\left(\left|x_1^\top[\mathcal{Z}]_j x_2\right|\geqslant t\right)\leqslant 2\exp\left(-\frac{nt^2}{8}\right),\quad\forall t\geqslant 0.$$

Therefore, for any fixed $U\in\mathbb{O}_{p^j,r_j}, V\in\mathbb{O}_{p^{d-j},pr_{j+1}}, x\in\mathbb{R}^{r_j}, y\in\mathbb{R}^{pr_{j+1}}$ with $\|x\|_2=1$ and $\|y\|_2=1$,

$$\mathbb{P}\left(\left|x^\top U^\top[\mathcal{Z}]_j V^\top y\right|\geqslant t\right)\leqslant 2\exp\left(-\frac{nt^2}{8}\right),\quad\forall t\geqslant 0.$$

Similarly to the proof of (5.267), with probability at least $1-Ce^{-cp}$, for all $1\leqslant k\leqslant d-1$,

$$\left\|\widehat{U}_k^{(0)\top}(I_p\otimes\widehat{U}_{k-1}^{(0)\top})\cdots(I_{p^{k-1}}\otimes\widehat{U}_1^{(0)\top})[\mathcal{Z}]_k(\widehat{V}_d^{(1)}\otimes I_{p^{d-k-1}})\cdots(\widehat{V}_{k+2}^{(1)}\otimes I_p)\right\|\leqslant C\sqrt{\frac{\sum_{i=1}^d p_i r_i r_{i-1}}{n}}.$$

Similarly, with probability at least $1-Ce^{-cp}$,

$$\left\|[\mathcal{Z}]_1(\widehat{V}_d^{(1)}\otimes I_{p^{d-2}})\cdots(\widehat{V}_3^{(1)}\otimes I_p)\widehat{V}_2^{(1)}\right\|\leqslant C\sqrt{\frac{\sum_{i=1}^d p_i r_i r_{i-1}}{n}}.$$

Notice that $\|X\|_F \leqslant \sqrt{r}\|X\|$ if $\operatorname{rank}(X) = r$, by the previous two inequalities and Theorem 4.1, we know that with probability at least $1 - Ce^{-cp}$,

$$\left\|\widehat{\boldsymbol{\mathcal{P}}}^{(1)} - \boldsymbol{\mathcal{P}}\right\|_F^2 \leqslant C\left(\max_{1\leqslant i\leqslant d-1} r_i\right)\frac{\sum_{i=1}^d p_i r_i r_{i-1}}{n}.$$

Finally, by the definition of $\widehat{\boldsymbol{\mathcal{P}}}$, we have

$$\left\|\widehat{\boldsymbol{\mathcal{P}}} - \boldsymbol{\mathcal{P}}\right\|_F \leqslant \left\|\widehat{\boldsymbol{\mathcal{P}}}^{(1)} - \boldsymbol{\mathcal{P}}\right\|_F + \left\|\widehat{\boldsymbol{\mathcal{P}}}^{(1)} - \widehat{\boldsymbol{\mathcal{P}}}\right\|_F \leqslant 2\left\|\widehat{\boldsymbol{\mathcal{P}}}^{(1)} - \boldsymbol{\mathcal{P}}\right\|_F,$$

which has finished the proof of Theorem 4.5.2.

5.3.8 Proof of Lemma 4.3.3

By symmetry, we only need to prove (4.6). By definition, (4.6) holds for $k = 1$. Suppose it holds for $k = j$. For $k = j+1$, since $S_{j+1} \in \mathbb{R}^{(r_j p_{j+1})\times(p_{j+2}\cdots p_d)}$ is realigned from $\widetilde{S}_j = M_j^\top S_j \in \mathbb{R}^{r_j\times(p_{j+1}\cdots p_d)}$, Lemma 4.3.2 that $S_{j+1} = (I_{p_{j+1}} \otimes \widetilde{S}_j)A^{(p_{j+1},p_{j+2}\cdots p_d)}$, where the realignment matrix $A^{(i,j)}$ is defined in (4.5). Therefore,

$$\begin{aligned}
S_{j+1} &= \left(I_{p_{j+1}} \otimes \widetilde{S}_j\right) A^{(p_{j+1},p_{j+2}\cdots p_d)}\\
&= \left(I_{p_{j+1}} \otimes M_j^\top S_j\right) A^{(p_{j+1},p_{j+2}\cdots p_d)}\\
&= \left(I_{p_{j+1}} \otimes M_j^\top\right)\left(I_{p_{j+1}} \otimes S_j\right) A^{(p_{j+1},p_{j+2}\cdots p_d)}\\
&= \left(I_{p_{j+1}} \otimes M_j^\top\right)\left(I_{p_{j+1}} \otimes \left((I_{p_j} \otimes M_{j-1}^\top)\cdots(I_{p_2\cdots p_j} \otimes M_1^\top)[\boldsymbol{\mathcal{T}}]_j\right)\right) A^{(p_{j+1},p_{j+2}\cdots p_d)}\\
&= \left(I_{p_{j+1}} \otimes M_j^\top\right)\left(I_{p_{j+1}} \otimes (I_{p_j} \otimes M_{j-1}^\top)\right)\cdots\left(I_{p_{j+1}} \otimes (I_{p_2\cdots p_j} \otimes M_1^\top)\right)\left(I_{p_{j+1}} \otimes [\boldsymbol{\mathcal{T}}]_j\right) A^{(p_{j+1},p_{j+2}\cdots p_d)}\\
&= \left(I_{p_{j+1}} \otimes M_j^\top\right)\left(I_{p_j p_{j+1}} \otimes M_{j-1}^\top\right)\cdots\left(I_{p_2\cdots p_{j+1}} \otimes M_1^\top\right)[\boldsymbol{\mathcal{T}}]_{j+1}.
\end{aligned}$$

The third equation and the fifth equation hold since $(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$; the last equation holds since $Y_{j+1} = \left(I_{p_{j+1}} \otimes Y_j\right) A^{(p_{j+1},p_{j+2}\cdots p_d)}$ and $A \otimes (B \otimes C) = (A \otimes B) \otimes C$.

Also notice that $\widetilde{S}_k = M_k^\top S_k$, we have finished the proof of (4.6).

5.3.9 Technical Lemmas

We collect the additional technical lemmas in this section.

Lemma 5.3.1.

(1) Suppose $A \in \mathbb{R}^{m_1 \times m_2}, B \in \mathbb{R}^{m_2 \times m_3}$*, where* $m_1 \geqslant m_2$*. Then*

$$s_{\min\{m_2,m_3\}}(AB) \geqslant s_{m_2}(A)s_{\min\{m_2,m_3\}}(B).$$

(2) Suppose $A \in \mathbb{R}^{m \times p_1}, B \in \mathbb{R}^{n \times p_2}, X \in \mathbb{R}^{p_1 \times p_2}$*, rank*$(X) = r, p_1 \geqslant m, p_2 \geqslant n$*. If* $X = U_1 M V_1^\top$*, where* $U_1 \in \mathbb{O}_{p_1,m}$ *and* $V_1 \in \mathbb{O}_{p_2,n}$*, then*

$$\sigma_r(AXB) \geqslant s_{\min}(AU_1)\sigma_r(X)s_{\min}(V_1^\top B).$$

Proof of Lemma 5.3.1. (1) Consider the SVD decomposition $A = U_A \Sigma_A V_A^\top, B = U_B \Sigma_B V_B^\top$, where $U_A \in \mathbb{O}_{m_1,m_2}, V_A \in \mathbb{O}_{m_2}, U_B \in \mathbb{O}_{m_2,\min\{m_2,m_3\}}, V_B \in \mathbb{O}_{\min\{m_2,m_3\},m_3}$, $\Sigma_A = \mathrm{diag}(\sigma_1(A), \ldots, s_{m_2}(A))$ and $\Sigma_B = \mathrm{diag}(s_1(B), \ldots, s_{\min\{m_2,m_3\}}(B))$ are diagonal matrices with nonnegative diagonal entries. Then

$$s_{\min\{m_2,m_3\}}(AB) = s_{\min\{m_2,m_3\}}(U_A \Sigma_A V_A^\top U_B \Sigma_B V_B^\top) = s_{\min\{m_2,m_3\}}(\Sigma_A V_A^\top U_B \Sigma_B). \tag{5.264}$$

For any $x \in \mathbb{R}^{\min\{m_2,m_3\}}$ satisfying $\|x\|_2 = 1$, we have

$$\|\Sigma_A V_A^\top U_B \Sigma_B x\|_2 \geqslant s_{m_2}(A)\|V_A^\top U_B \Sigma_B x\|_2 = s_{m_2}(A)\|\Sigma_B x\|_2 \geqslant s_{m_2}(A)s_{\min\{m_2,m_3\}}(B).$$

Therefore

$$s_{\min\{m_2,m_3\}}(AB) = s_{\min\{m_2,m_3\}}(\Sigma_A V_A^\top U_B \Sigma_B) \geqslant s_{m_2}(A)s_{\min\{m_2,m_3\}}(B).$$

(2) Consider the SVD decomposition $X = U\Sigma V^\top$, where $U \in \mathbb{O}_{p_1,r}, V \in \mathbb{O}_{p_2,r}$ and Σ is a diagonal matrix. Then we know that there exist two matrices $L \in \mathbb{R}^{m \times r}$ and

$R \in \mathbb{R}^{n\times r}$ satisfying $U = U_1 L$ and $V = V_1 R$. Moreover,

$$L^\top L = L^\top U_1^\top U_1 L = U^\top U = I_r, \quad R^\top R = R^\top V_1^\top V_1 R = V^\top V = I_r.$$

Therefore,

$$\sigma_r(AXB) = \sigma_r(AU_1 L\Sigma R^\top V_1^\top B) \geqslant s_{\min}(AU_1)\sigma_r(L\Sigma R^\top)s_{\min}(V_1^\top B) = s_{\min}(AU_1)\sigma_r(X)s_{\min}(V_1^\top B).$$

□

Lemma 5.3.2. *Suppose* Z *is a matrix with independent zero-mean* σ*-sub-Gaussian entries,* d *is a fixed number,* $r_0 = r_d = 1$.
(1) Suppose $Z \in \mathbb{R}^{p\times q}$, $A \in \mathbb{R}^{m\times p}, B \in \mathbb{R}^{q\times n}$ *satisfy* $\|A\|, \|B\| \leqslant 1$, $m \leqslant p, n \leqslant q$. *Then*

$$\mathbb{P}\left(\|AZB\| \geqslant 2\sigma\sqrt{m+t}\right) \leqslant 2\cdot 5^n \exp\left[-c\min\left(\frac{t^2}{m}, t\right)\right]. \tag{5.265}$$

$$\mathbb{P}\left(\|AZB\|_F \geqslant \sigma\sqrt{mn+t}\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{mn}, t\right)\right]. \tag{5.266}$$

(2) Suppose $Z \in \mathbb{R}^{(p_1\cdots p_k)\times m}$, $2 \leqslant k \leqslant d-1$. *Then*

$$\max_{\substack{U_i\in\mathbb{R}^{(p_i r_{i-1})\times r_i}\\ \|U_i\|\leqslant 1}} \left\|(I_{p_k}\otimes U_{k-1}^\top)\cdots(I_{p_2\cdots p_k}\otimes U_1^\top)Z\right\| \geqslant C\sigma\sqrt{\sum_{i=1}^{k-1} p_i r_{i-1} r_i + p_k r_{k-1} + m}. \tag{5.267}$$

with probability at least $1 - C\exp(-c(\sum_{i=1}^{k-1} p_i r_{i-1} r_i + p_k r_{k-1} + m))$.
(3) Suppose $Z \in \mathbb{R}^{(p_1\cdots p_k)\times(p_{k+1}\cdots p_d)}$, $2 \leqslant k \leqslant d-2$. *Then*

$$\begin{aligned}&\max_{(U_1,\dots,V_d)\in\mathcal{A}} \left\|U_k^\top(I_{p_k}\otimes U_{k-1}^\top)\cdots(I_{p_2\cdots p_k}\otimes U_1^\top)Z(V_d\otimes I_{p_{k+1}\dots p_{d-1}})\cdots(V_{k+2}\otimes I_{p_{k+1}})\right\|\\ \geqslant &C\sigma\sqrt{\sum_{i=1}^{d} p_i r_{i-1} r_i}\end{aligned} \tag{5.268}$$

with probability at least $1 - C\exp(-c\sum_{i=1}^{d} p_i r_{i-1} r_i)$. *Here,*

$$\mathcal{A} = \{(U_1, \dots, U_k, V_{k+2}, \dots, V_d) : U_i \in \mathbb{R}^{(p_i r_{i-1}) \times r_i}, \|U_i\| \leqslant 1, V_j \in \mathbb{R}^{(p_i r_i) \times r_{i-1}}, \|V_j\| \leqslant 1\}. \tag{5.269}$$

(4) Suppose $Z \in \mathbb{R}^{(p_1 \cdots p_{d-1}) \times p_d}$. *Then with probability at least* $1 - C\exp(-c\sum_{i=1}^{d} p_i r_{i-1} r_i)$,

$$\max_{U_i \in \mathbb{R}^{(p_i r_{i-1}) \times r_i}, \|U_i\| \leqslant 1} \left\|U_{d-1}^\top (I_{p_{d-1}} \otimes U_{d-2}^\top) \cdots (I_{p_2 \cdots p_{d-1}} \otimes U_1^\top) Z\right\|_F \geqslant C\sigma \sqrt{\sum_{i=1}^{d} p_i r_{i-1} r_i}. \tag{5.270}$$

(5) Suppose $Z \in \mathbb{R}^{(p_1 \cdots p_k) \times (p_{k+1} \cdots p_d)}$, $2 \leqslant k \leqslant d-2$. *Then*

$$\begin{aligned} &\max_{(U_1, \dots, V_d) \in \mathcal{A}} \left\|U_k^\top (I_{p_k} \otimes U_{k-1}^\top) \cdots (I_{p_2 \cdots p_k} \otimes U_1^\top) Z (V_d \otimes I_{p_{k+1} \dots p_{d-1}}) \cdots (V_{k+2} \otimes I_{p_{k+1}})\right\|_F \\ \geqslant & C\sigma \sqrt{\sum_{i=1}^{d} p_i r_{i-1} r_i} \end{aligned} \tag{5.271}$$

with probability at least $1 - C\exp(-c\sum_{i=1}^{d} p_i r_{i-1} r_i)$. *Here,* $\mathcal{A}$ *is defined in* (5.269).

Proof of Lemma 5.3.2. W.O.L.G., assume $\sigma = 1$.
(1) For fixed $x \in \mathbb{R}^n$ satisfying $\|x\|_2 = 1$, we have $AZBx = (x^\top B^\top \otimes A)\text{vec}(Z)$. Since

Z_{ij} is 1-sub-Gaussian, we know that $\mathrm{Var}(Z_{ij}) \leqslant 1$. In addition,

$$\begin{aligned}
\mathbb{E}\|(x^\top B^\top \otimes A)\mathrm{vec}(Z)\|_2^2 =& \mathbb{E}\left[\mathrm{trace}\left(\mathrm{vec}(Z)^\top (x^\top B^\top \otimes A)^\top (x^\top B^\top \otimes A)\mathrm{vec}(Z)\right)\right] \\
=& \mathrm{trace}\left[\mathbb{E}\left((x^\top B^\top \otimes A)^\top (x^\top B^\top \otimes A)\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)\right] \\
=& \mathrm{trace}\left[(x^\top B^\top \otimes A)^\top (x^\top B^\top \otimes A)\mathbb{E}\left(\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)\right] \\
\leqslant& \mathrm{trace}\left((x^\top B^\top \otimes A)^\top (x^\top B^\top \otimes A)\right) \\
=& \left\|x^\top B^\top \otimes A\right\|_F^2 = \|Bx\|_2^2 \|A\|_F^2 \leqslant \|x\|_2^2 \|A\|_F^2 \\
\leqslant& m.
\end{aligned} \tag{5.272}$$

The first inequality holds since $\mathbb{E}\left(\mathrm{vec}(Z)\mathrm{vec}(Z)^\top\right)$ is a diagonal matrix with diagonal entries $\mathrm{Var}(Z_{ij}) \leqslant 1$; the last inequality is due to $\|A\|_F \leqslant \min\{m, p\}\|A\|_2 \leqslant m$. By Hanson-Wright inequality, we have

$$\mathbb{P}\left(\|AZBx\|_2^2 - m \geqslant t\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{\|(Bxx^\top B^\top) \otimes (A^\top A)\|_F^2}, \frac{t}{\|(Bxx^\top B^\top) \otimes (A^\top A)\|}\right)\right].$$

Since $\|x\|_2 = 1$ and $\|A\|, \|B\| \leqslant 1$,

$$\begin{aligned}
\|(Bxx^\top B^\top) \otimes (A^\top A)\|_F^2 =& \|Bxx^\top B^\top\|_F^2 \|A^\top A\|_F^2 = (x^\top B^\top Bx)^2 \|A^\top A\|_F^2 \\
\leqslant& (x^\top x)^2 \|A^\top A\|_F^2 = \sum_{i=1}^{\min\{m,p\}} \sigma_i^4(A) \leqslant m,
\end{aligned}$$

$$\|(Bxx^\top B^\top) \otimes (A^\top A)\| \leqslant \|Bxx^\top B^\top\| \|A^\top A\| \leqslant \|xx^\top\| \|A^\top A\| \leqslant 1.$$

Thus, for fixed x satisfying $\|x\|_2 = 1$, we have

$$\mathbb{P}\left(\|AZBx\|_2^2 \geqslant m + t\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{m}, t\right)\right]. \tag{5.273}$$

By Vershynin (2010)[Lemma 5.2], there exists $\mathcal{N}_{1/2}$, a 1/2-net of $\{x \in \mathbb{R}^n : \|x\|_2 = 1\}$, such that $\left|\mathcal{N}_{1/2}\right| \leqslant 5^n$. The union bound, Vershynin (2010)[Lemma 5.2] and (5.273) together imply that

$$\mathbb{P}\left(\|AZB\| \geqslant 2\sqrt{m+t}\right) \leqslant \mathbb{P}\left(\max_{x\in\mathcal{N}_{1/2}} \|AZBx\|_2 \geqslant \sqrt{m+t}\right) \leqslant 2\cdot 5^n \exp\left[-c\min\left(\frac{t^2}{m}, t\right)\right].$$

For $\|AZB\|_F$, note that $AZB = (B^\top \otimes A)\text{vec}(Z)$, Similarly to (5.272), we have

$$\begin{aligned}\mathbb{E}\|(B^\top \otimes A)\text{vec}(Z)\|_2^2 =& \mathbb{E}\left[\text{vec}(Z)^\top (B^\top \otimes A)^\top (B^\top \otimes A)\text{vec}(Z)\right] \\ =& \mathbb{E}\text{trace}\left[\text{vec}(Z)^\top (B^\top \otimes A)^\top (B^\top \otimes A)\text{vec}(Z)\right] \\ =& \text{trace}\mathbb{E}\left[(B^\top \otimes A)^\top (B^\top \otimes A)\text{vec}(Z)\text{vec}(Z)^\top\right] \\ =& \text{trace}\left[(B^\top \otimes A)^\top (B^\top \otimes A)\mathbb{E}\left(\text{vec}(Z)\text{vec}(Z)^\top\right)\right] \\ \leqslant& \text{trace}\left[(B^\top \otimes A)^\top (B^\top \otimes A)\right] \\ =& \|B^\top \otimes A\|_F^2 = \|B\|_F^2\|A\|_F^2 \\ \leqslant& mn.\end{aligned}$$

By Hanson-Wright inequality, we have

$$\mathbb{P}\left(\|AZB\|_F^2 - mn \geqslant t\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{\|(BB^\top)\otimes(A^\top A)\|_F^2}, \frac{t}{\|(BB^\top)\otimes(A^\top A)\|}\right)\right].$$

Since $\|A\|, \|B\| \leqslant 1$, we have

$$\|(BB^\top)\otimes(A^\top A)\|_F = \sqrt{\|A^\top A\|_F^2\|BB^\top\|_F^2} = \sqrt{\sum_{i=1}^{\min\{m,p\}} \sigma_i^4(A) \sum_{i=1}^{\min\{q,n\}} \sigma_i^4(B)} \leqslant \sqrt{mn},$$

$$\|(BB^\top)\otimes(A^\top A)\| \leqslant 1.$$

Therefore,

$$\mathbb{P}\left(\|AZB\|_F^2 \geqslant mn + t\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{mn}, t\right)\right].$$

(2) For fixed $x \in \mathbb{R}^m$ and $A \in \mathbb{R}^{(p_k r_{k-1})\times(p_1\cdots p_k)}$ satisfying $\|x\|_2 = 1$ and $\|A\| \leqslant 1$,

by (5.265) with $B = I_m$, we have

$$\mathbb{P}\left(\|AZ\| \geqslant 2\sqrt{p_k r_{k-1} + t}\right) \leqslant 2 \cdot 5^m \exp\left[-c \min\left(\frac{t^2}{p_k r_{k-1}}, t\right)\right]. \tag{5.274}$$

By Zhang and Xia (2018)[Lemma 7], for $1 \leqslant i \leqslant k-1$, there exist ε-nets: $U_i^{(1)}, \ldots, U_i^{(N_i)} \in \mathbb{R}^{(p_i r_{i-1}) \times r_i}$ (here $r_0 = 1$), $N_i \leqslant ((2+\varepsilon)/\varepsilon)^{(p_i r_{i-1}) \times r_i}$, such that

$$\forall U \in \mathbb{R}^{(p_i r_{i-1}) \times r_i} \text{ satisfying } \|U\| \leqslant 1, \exists 1 \leqslant j \leqslant N_i \text{ such that } \|U_i^{(j)} - U\| \leqslant \varepsilon.$$

Therefore,

$$\begin{aligned} &\mathbb{P}\left(\max_{i_1,\ldots,i_{k-1}} \left\|(I_{p_k} \otimes U_{k-1}^{(i_{k-1})\top}) \cdots (I_{p_2 \cdots p_k} \otimes U_1^{(i_1)\top}) Z\right\| \geqslant 2\sqrt{p_k r_{k-1} + t}\right) \\ \leqslant& 2((2+\varepsilon)/\varepsilon)^{\sum_{i=1}^{k-1} p_i r_{i-1} r_i} 5^m \exp\left[-c \min\left(\frac{t^2}{p_k r_{k-1}}, t\right)\right]. \end{aligned} \tag{5.275}$$

Let

$$\begin{aligned} U_1^*, \ldots, U_{k-1}^* \in& \underset{\substack{U_i \in \mathbb{R}^{(p_i r_{i-1}) \times r_i}, 1 \leqslant i \leqslant k-1 \\ \|U_i\| \leqslant 1, \quad 1 \leqslant i \leqslant k-1}}{\arg\max} \left\|(I_{p_k} \otimes U_{k-1}^\top) \cdots (I_{p_2 \cdots p_k} \otimes U_1^\top) Z\right\|, \\ M =& \max_{\substack{U_i \in \mathbb{R}^{(p_i r_{i-1}) \times r_i}, 1 \leqslant i \leqslant k-1 \\ \|U_i\| \leqslant 1, \quad 1 \leqslant i \leqslant k-1}} \left\|(I_{p_k} \otimes U_{k-1}^\top) \cdots (I_{p_2 \cdots p_k} \otimes U_1^\top) Z\right\|. \end{aligned}$$

Then for any $1 \leqslant i \leqslant k-1$, there exists $1 \leqslant j_i \leqslant N_i$, such that $\|U_i^{(j_i)} - U_i^*\| \leqslant \varepsilon$.

Then

$$\begin{aligned} M =& \left\|(I_{p_k} \otimes U_{k-1}^{*\top}) \cdots (I_{p_2 \cdots p_k} \otimes U_1^{*\top}) Z\right\| \\ \leqslant& \left\|(I_{p_k} \otimes U_{k-1}^{(j_{k-1})\top}) \cdots (I_{p_2 \cdots p_k} \otimes U_1^{(j_1)\top}) Z\right\| \\ &+ \left\|\left(I_{p_k} \otimes (U_{k-1}^{*} - U_{k-1}^{(j_{k-1})})\right)^{\top} (I_{p_{k-1}p_k} \otimes U_{k-2}^{(j_{k-2})\top}) \cdots (I_{p_2 \cdots p_k} \otimes U_1^{(j_1)\top}) Z\right\| \\ &+ \cdots + \left\|(I_{p_k} \otimes U_{k-1}^{*\top}) \cdots (I_{p_3 \cdots p_k} \otimes U_2^{*\top}) \left(I_{p_2 \cdots p_k} \otimes (U_1^{*} - U_1^{(j_1)})^{\top}\right) Z\right\| \\ \leqslant& \left\|(I_{p_k} \otimes U_{k-1}^{(j_{k-1})\top}) \cdots (I_{p_2 \cdots p_k} \otimes U_1^{(j_1)\top}) Z\right\| + \varepsilon(k-1)M. \end{aligned} \tag{5.276}$$

Combine (5.275) and the previous inequality together, we have

$$\begin{aligned} &\mathbb{P}\left(M \geqslant \frac{2\sqrt{p_k r_{k-1} + t}}{1 - (k-1)\varepsilon}\right) \\ \leqslant& 2((2+\varepsilon)/\varepsilon)^{\sum_{i=1}^{k-1} p_i r_{i-1} r_i} 5^m \exp\left[-c \min\left(\frac{t^2}{p_k r_{k-1}}, t\right)\right]. \end{aligned} \tag{5.277}$$

By setting $\varepsilon = \frac{1}{2(k-1)}$ and $t = C\sqrt{\sum_{i=1}^{k-1} p_i r_{i-1} r_i + p_k r_{k-1} + m}$, we have proved (5.267).

(3) For fixed $A \in \mathbb{R}^{r_k \times (p_1 \cdots p_k)}, B \in \mathbb{R}^{(p_{k+1} \cdots p_d) \times (p_{k+1} r_{k+1})}$ satisfying $\|A\| \leqslant 1, \|B\| \leqslant 1$, by (5.265), we have

$$\mathbb{P}\left(\|AZB\| \geqslant 2\sqrt{r_k + t}\right) \leqslant 2 \cdot 5^{p_{k+1} r_{k+1}} \exp\left[-c \min\left(\frac{t^2}{r_k}, t\right)\right].$$

Let

$$\begin{aligned} &M \\ =& \max_{\substack{U_i \in \mathbb{R}^{(p_i r_{i-1}) \times r_i}, \|U_i\| \leqslant 1, 1 \leqslant i \leqslant k \\ V_i \in \mathbb{R}^{(p_i r_i) \times r_{i-1}}, \|V_i\| \leqslant 1, k+2 \leqslant i \leqslant d}} \left\|U_k^{\top}(I_{p_k} \otimes U_{k-1}^{\top}) \cdots (I_{p_2 \cdots p_k} \otimes U_1^{\top}) Z (V_d \otimes I_{p_{k+1} \dots p_{d-1}}) \cdots (V_{k+2} \otimes I_{p_{k+1}})\right\|, \end{aligned}$$

By similar arguments as (5.277), one has

$$\mathbb{P}\left(M \geqslant \frac{2\sqrt{r_k + t}}{1-(d-1)\varepsilon}\right) \leqslant 2((2+\varepsilon)/\varepsilon)^{\sum_{1\leqslant i\leqslant d, i\neq k+1} p_i r_{i-1} r_i} 5^{p_{k+1}r_{k+1}} \exp\left[-c\min\left(\frac{t^2}{r_k}, t\right)\right]$$

for any $0 < \epsilon < \frac{1}{d}$. By setting $\varepsilon = \frac{1}{2(d-1)}$ and $t = C\sum_{i=1}^{d} p_i r_{i-1} r_i$, we have proved the third part of Lemma 5.3.2.

(4) For fixed $U_1, \ldots, U_{d-1}$ satisfying $\|U_i\| \leqslant 1$, let $A = U_{d-1}^\top (I_{p_{d-1}} \otimes U_{d-2}^\top) \cdots (I_{p_2 \cdots p_{d-1}} \otimes U_1^\top) \in \mathbb{R}^{r_{d-1}\times(p_1\cdots p_{d-1})}$, then $\|A\| \leqslant 1$. By (5.266) with $B = I_{p_d}$, we have

$$\mathbb{P}\left(\|AZ\|_F^2 \geqslant p_d r_{d-1} + t\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{p_d r_{d-1}}, t\right)\right].$$

Let

$$M = \max_{U_i \in \mathbb{R}^{(p_i r_{i-1})\times r_i}, \|U_i\|\leqslant 1} \|U_{d-1}^\top (I_{p_{d-1}} \otimes U_{d-2}^\top) \cdots (I_{p_2\cdots p_{d-1}} \otimes U_1^\top) Z\|_F.$$

The similar proof of (5.277) leads us to

$$\mathbb{P}\left(M^2 \geqslant \frac{r_{d-1}p_d + t}{(1-\varepsilon(d-1))^2}\right) \leqslant 2\left((2+\varepsilon)/\varepsilon\right)^{\sum_{k=1}^{d-1} p_k r_{k-1} r_k} \exp\left[-c\min\left(\frac{t^2}{p_d r_{d-1}}, t\right)\right]. \tag{5.278}$$

for $0 < \varepsilon < \frac{1}{d-1}$. By setting $\varepsilon = \frac{1}{2(d-1)}$ and $t = C\sum_{k=1}^{d} p_k r_{k-1} r_k$, we have arrived at (5.270).

(5) For fixed $A \in \mathbb{R}^{r_k \times (p_1\cdots p_k)}$, $B \in \mathbb{R}^{(p_{k+1}\cdots p_d)\times(p_{k+1}r_{k+1})}$, $\|A\| \leqslant 1, \|B\| \leqslant 1$, by (5.266), we have

$$\mathbb{P}\left(\|AZB\|_F^2 \geqslant p_{k+1}r_{k+1}r_k + t\right) \leqslant 2\exp\left[-c\min\left(\frac{t^2}{p_{k+1}r_{k+1}r_k}, t\right)\right].$$

Let

$$M = \max_{\substack{U_i \in \mathbb{R}^{(p_i r_{i-1}) \times r_i}, \|U_i\| \leqslant 1 \\ V_i \in \mathbb{R}^{(p_i r_i) \times r_{i-1}}, \|V_i\| \leqslant 1}} \left\| U_k^\top (I_{p_k} \otimes U_{k-1}^\top) \cdots (I_{p_2 \cdots p_k} \otimes U_1^\top) Z (V_d \otimes I_{p_{k+1} \dots p_{d-1}}) \cdots (V_{k+2} \otimes I_{p_{k+1}}) \right\|_F,$$

Similarly to (5.277), for any $0 < \varepsilon < \frac{1}{d-1}$, we have

$$\mathbb{P}\left(M \geqslant \frac{\sqrt{p_{k+1} r_{k+1} r_k} + t}{1 - (d-1)\varepsilon} \right) \leqslant 2((2+\varepsilon)/\varepsilon)^{\sum_{1 \leqslant i \leqslant d, i \neq k+1} p_i r_{i-1} r_i} \exp\left[-c \min\left(\frac{t^2}{p_{k+1} r_{k+1} r_k}, t \right) \right]. \tag{5.279}$$

By setting $\varepsilon = \frac{1}{2(d-1)}$ and $t = C \sum_{i=1}^{d} p_i r_{i-1} r_i$, we have proved (5.271). □

Lemma 5.3.3. *Suppose* $X, Z \in \mathbb{R}^{p_1 \times p_2}$, *rank*$(X) = r$. *Let* $Y = X + Z$, $\widehat{U} = SVD_r^L(Y)$, $\widehat{V} = SVD_r^R(Y)$. *Then we have*

$$\max\{\|\widehat{U}_\perp^\top X\|, \|X\widehat{V}_\perp\|\} \leqslant 2\|Z\|, \quad \max\{\|\widehat{U}_\perp^\top X\|_F, \|X\widehat{V}_\perp\|_F\} \leqslant 2\min\{\|Z\|_F, \sqrt{r}\|Z\|\}.$$

Proof of Lemma 5.3.3. See (Zhang and Xia, 2018, Lemma 6) and (Luo and Zhang, 2020b, Theorem 1). □

references

Ahsen, M Eren, and Mathukumalli Vidyasagar. 2017. Error bounds for compressed sensing algorithms with group sparsity: A unified approach. *Applied and Computational Harmonic Analysis* 43(2):212–232.

Allahyar, Amin, and Jeroen De Ridder. 2015. Feral: network-based classifier with application to breast cancer outcome prediction. *Bioinformatics* 31(12):i311–i319.

Allen, Genevera. 2012a. Sparse higher-order principal components analysis. In *Artificial intelligence and statistics*, 27–36.

Allen, Genevera I. 2012b. Regularized tensor factorizations and higher-order principal components analysis. *arXiv preprint arXiv:1202.2476*.

Anandkumar, Anima, Yuan Deng, Rong Ge, and Hossein Mobahi. 2016. Homotopy analysis for tensor pca. *arXiv preprint arXiv:1610.09322*.

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

Anandkumar, Animashree, Rong Ge, and Majid Janzamin. 2014b. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*.

Anandkumar, Animashree, Daniel Hsu, and Sham M Kakade. 2012. A method of moments for mixture models and hidden markov models. In *Conference on learning theory*, 33–1.

Arous, Gerard Ben, Song Mei, Andrea Montanari, and Mihai Nica. 2019. The landscape of the spiked tensor model. *Communications on Pure and Applied Mathematics* 72(11):2282–2330.

Auddy, Arnab, and Ming Yuan. 2020. Perturbation bounds for orthogonally decomposable tensors and their applications in high dimensional data analysis. *arXiv preprint arXiv:2007.09024*.

Bao, Zhigang, Xiucai Ding, and Ke Wang. 2018. Singular vector and singular subspace distribution for the matrix denoising model. *arXiv preprint arXiv:1809.10476*.

Barak, Boaz, and Ankur Moitra. 2016. Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on learning theory*, 417–445.

Belkin, Mikhail, Luis Rademacher, and James Voss. 2018. Eigenvectors of orthogonally decomposable functions. *SIAM Journal on Computing* 47(2):547–615.

Bengua, Johann A, Ho N Phien, Hoang Duong Tuan, and Minh N Do. 2017. Efficient tensor completion for color image and video recovery: Low-rank tensor train. *IEEE Transactions on Image Processing* 26(5):2466–2479.

Benson, Austin R, David F Gleich, and Lek-Heng Lim. 2017. The spacey random walk: A stochastic process for higher-order data. *SIAM Review* 59(2):321–345.

Berchtold, André, and Adrian E Raftery. 2002. The mixture transition distribution model for high-order markov chains and non-gaussian time series. *Statistical Science* 328–356.

Berry, Andrew C. 1941. The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society* 49(1):122–136.

Bertsekas, Dimitri, and Angelia Nedic. 2003. Convex analysis and optimization (conservative).

Bhattacharya, Anirban, and David B Dunson. 2012. Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association* 107(497):362–377.

Bi, Xuan, Annie Qu, and Xiaotong Shen. 2018. Multilayer tensor factorization with applications to recommender systems. *The Annals of Statistics* 46(6B):3308–3333.

Bickel, Peter J, YaâĽ™acov Ritov, and Alexandre B Tsybakov. 2009. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* 37(4):1705–1732.

Bigoni, Daniele, Allan P Engsig-Karup, and Youssef M Marzouk. 2016. Spectral tensor-train decomposition. *SIAM Journal on Scientific Computing* 38(4):A2405–A2439.

Bravyi, Sergey, David Gosset, and Ramis Movassagh. 2019. Classical algorithms for quantum mean values. *arXiv preprint arXiv:1909.11485*.

Brennan, Matthew, and Guy Bresler. 2020. Reducibility and statistical-computational gaps from secret leakage. *arXiv preprint arXiv:2005.08099*.

Bunea, Florentina, Johannes Lederer, and Yiyuan She. 2013. The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Transactions on Information Theory* 60(2):1313–1325.

Cai, Changxiao, Gen Li, H Vincent Poor, and Yuxin Chen. 2019a. Nonconvex low-rank tensor completion from noisy data. In *Advances in neural information processing systems*, 1863–1874.

Cai, Changxiao, H Vincent Poor, and Yuxin Chen. 2020. Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. *arXiv preprint arXiv:2006.08580*.

Cai, Jian-Feng, Emmanuel J Candès, and Zuowei Shen. 2010. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization* 20(4): 1956–1982.

Cai, T Tony, and Zijian Guo. 2017. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics* 45(2):615–646.

Cai, T. Tony, Xiaodong Li, and Zongming Ma. 2016a. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics* 44:2221–2251.

Cai, T Tony, Tengyuan Liang, and Alexander Rakhlin. 2016b. Geometric inference for general high-dimensional linear inverse problems. *The Annals of Statistics* 44(4): 1536–1563.

Cai, T Tony, Zongming Ma, and Yihong Wu. 2013. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics* 41(6):3074–3110.

Cai, T Tony, and Anru Zhang. 2013a. Compressed sensing and affine rank minimization under restricted isometry. *IEEE Transactions on Signal Processing* 61(13): 3279–3290.

———. 2013b. Sharp rip bound for sparse signal and low-rank matrix recovery. *Applied and Computational Harmonic Analysis* 35(1):74–93.

———. 2014. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE transactions on information theory* 60(1):122–132.

———. 2018. Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *The Annals of Statistics* 46(1):60–89.

Cai, T Tony, Anru Zhang, and Yuchen Zhou. 2019b. Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *arXiv preprint arXiv:1909.09851*.

Calvi, Giuseppe G, Ahmad Moniri, Mahmoud Mahfouz, Zeyang Yu, Qibin Zhao, and Danilo P Mandic. 2019. Tucker tensor layer in fully connected neural networks. *arXiv preprint arXiv:1903.06133*.

Candes, EJ, J Romberg, and T Tao. 2006. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory* 52(2):489–509.

Candes, EJ, and T Tao. 2005. Decoding by linear programming. *IEEE Transactions on Information Theory* 51(12):4203–4215.

Candes, Emmanuel, and Terence Tao. 2007. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* 35(6):2313–2351.

Candès, Emmanuel J, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? *Journal of the ACM (JACM)* 58(3):11.

Candes, Emmanuel J, and Yaniv Plan. 2010. Matrix completion with noise. *Proceedings of the IEEE* 98(6):925–936.

———. 2011. A probabilistic and ripless theory of compressed sensing. *IEEE Transactions on Information Theory* 57(11):7235–7254.

Candès, Emmanuel J, and Benjamin Recht. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9(6):717.

Candes, Emmanuel J, Carlos A Sing-Long, and Joshua D Trzasko. 2013a. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE transactions on signal processing* 61(19):4643–4657.

Candes, Emmanuel J, Thomas Strohmer, and Vladislav Voroninski. 2013b. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics* 66(8): 1241–1274.

Carpentier, Alexandra, Jens Eisert, David Gross, and Richard Nickl. 2019. Uncertainty quantification for matrix compressed sensing and quantum tomography problems. In *High dimensional probability viii*, 385–430. Springer.

Chaplot, Devendra Singh, Pushpak Bhattacharyya, and Ashwin Paranjape. 2015. Unsupervised word sense disambiguation using markov random field and dependency parser. In *Twenty-ninth aaai conference on artificial intelligence*.

Chatterjee, Soumyadeep, Karsten Steinhaeuser, Arindam Banerjee, Snigdhansu Chatterjee, and Auroop Ganguly. 2012. Sparse group lasso: Consistency and climate applications. In *Proceedings of the 2012 siam international conference on data mining*, 47–58. SIAM.

Chatterjee, Sourav. 2015. Matrix estimation by universal singular value thresholding. *The Annals of Statistics* 43(1):177–214.

Chen, Han, Garvesh Raskutti, and Ming Yuan. 2019a. Non-convex projected gradient descent for generalized low-rank tensor regression. *The Journal of Machine Learning Research* 20(1):172–208.

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

Chen, Wei-Kuo. 2019. Phase transition in the spiked random tensor with rademacher prior. *The Annals of Statistics* 47(5):2734–2756.

Chen, Yuxin, Jianqing Fan, Cong Ma, and Yuling Yan. 2019b. Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences* 116(46):22931–22937.

Chi, Eric C, Brian R Gaines, Will Wei Sun, Hua Zhou, and Jian Yang. 2018. Provable convex co-clustering of tensors. *arXiv preprint arXiv:1803.06518*.

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

Colombo, Nicolo, and Nikos Vlassis. 2016. Tensor decomposition via joint matrix schur decomposition. In *International conference on machine learning*, 2820–2828.

De Domenico, Manlio, Vincenzo Nicosia, Alexandre Arenas, and Vito Latora. 2015. Structural reducibility of multilayer networks. *Nature communications* 6(1):1–9.

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

———. 2000b. On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications* 21(4):1324–1342.

De Silva, Vin, and Lek-Heng Lim. 2008. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications* 30(3):1084–1127.

Dolgov, Sergey V, and Dmitry V Savostyanov. 2014. Alternating minimal energy methods for linear systems in higher dimensions. *SIAM Journal on Scientific Computing* 36(5):A2248–A2271.

Donoho, David, and Matan Gavish. 2014. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics* 42(6):2413–2440.

Du, Zhe, Necmiye Ozay, and Laura Balzano. 2019. Mode clustering for markov jump systems. In *2019 ieee 8th international workshop on computational advances in multi-sensor adaptive processing (camsap)*, 126–130. IEEE.

Duan, Yaqi, Mengdi Wang, Zaiwen Wen, and Yaxiang Yuan. 2020. Adaptive low-nonnegative-rank approximation for state aggregation of markov chains. *SIAM Journal on Matrix Analysis and Applications* 41(1):244–278.

Duchi, John, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. 2008. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on machine learning*, 272–279.

Dudeja, Rishabh, and Daniel Hsu. 2020. Statistical query lower bounds for tensor pca. *arXiv preprint arXiv:2008.04101*.

Dunson, David B, and Chuanhua Xing. 2009. Nonparametric bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* 104(487): 1042–1051.

Eckart, Carl, and Gale Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* 1(3):211–218.

Esseen, Carl-Gustaf. 1942. On the liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys.* 28:1–19.

Fan, Jianqing, Wenyan Gong, and Ziwei Zhu. 2019. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of econometrics* 212(1): 177–202.

Fannes, Mark, Bruno Nachtergaele, and Reinhard F Werner. 1992. Finitely correlated states on quantum spin chains. *Communications in mathematical physics* 144(3): 443–490.

Feizi, Soheil, Hamid Javadi, and David Tse. 2017. Tensor biclustering. In *Advances in neural information processing systems*, 1311–1320.

Foucart, Simon, Deanna Needell, Yaniv Plan, and Mary Wootters. 2017. De-biasing low-rank projection for matrix completion. In *Wavelets and sparsity xvii*, vol. 10394, 1039417. International Society for Optics and Photonics.

Foucart, Simon, and Holger Rauhut. 2013. *A mathematical introduction to compressive sensing*, vol. 1. Birkhäuser Basel.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*.

Ganguly, Arnab, Tatjana Petrov, and Heinz Koeppl. 2014. Markov chain aggregation and its applications to combinatorial reaction networks. *Journal of mathematical biology* 69(3):767–797.

Van de Geer, Sara, Peter Bühlmann, YaâŁ™acov Ritov, and Ruben Dezeure. 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 42(3):1166–1202.

Grasedyck, Lars, Melanie Kluge, and Sebastian Kramer. 2015. Variants of alternating least squares tensor completion in the tensor train format. *SIAM Journal on Scientific Computing* 37(5):A2424–A2450.

Gross, David. 2011. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory* 57(3):1548–1566.

Han, Rungang, Rebecca Willett, and Anru Zhang. 2020. An optimal statistical and computational framework for generalized tensor estimation. *arXiv preprint arXiv:2002.11255*.

Hao, Botao, Anru Zhang, and Guang Cheng. 2018. Sparse and low-rank tensor estimation via cubic sketchings. *arXiv preprint arXiv:1801.09326*.

———. 2020. Sparse and low-rank tensor estimation via cubic sketchings. *IEEE Transactions on Information Theory* 66:5927–5964.

Hillar, Christopher J, and Lek-Heng Lim. 2013. Most tensor problems are np-hard. *Journal of the ACM (JACM)* 60(6):1–39.

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

Hsu, Daniel, Sham Kakade, and Tong Zhang. 2012. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability* 17.

Huang, Jiaoyang, Daniel Z. Huang, Qing Yang, and Guang Cheng. 2020. Power iteration for tensor pca. *arXiv preprint arXiv: 2012.13669*.

Jaganathan, Kishore, Samet Oymak, and Babak Hassibi. 2013. Sparse phase retrieval: Convex algorithms and limitations. In *2013 ieee international symposium on information theory*, 1022–1026. IEEE.

Jalali, Amin, and Maryam Fazel. 2013. A convex method for learning d-valued models. In *2013 ieee global conference on signal and information processing*, 1123–1126. IEEE.

Jalali, Amin, Adel Javanmard, and Maryam Fazel. 2019. New computational and statistical aspects of regularized regression with application to rare feature selection and aggregation. *arXiv preprint arXiv:1904.05338*.

Javanmard, Adel, and Andrea Montanari. 2014. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* 15(1):2869–2909.

———. 2018. Debiasing the lasso: Optimal sample size for gaussian designs. *The Annals of Statistics* 46(6A):2593–2622.

Jing, Bing-Yi, Ting Li, Zhongyuan Lyu, and Dong Xia. 2020. Community detection on mixture multi-layer networks via regularized tensor decomposition. *arXiv preprint arXiv:2002.04457*.

Johnstone, Iain M, and Arthur Yu Lu. 2009. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* 104(486):682–693.

Karatzoglou, Alexandros, Xavier Amatriain, Linas Baltrunas, and Nuria Oliver. 2010. Multiverse recommendation: n-dimensional tensor factorization for contextaware collaborative filtering. In *Proceedings of the fourth acm conference on recommender systems*, 79–86.

Kearns, Michael J, and Satinder P Singh. 1999. Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in neural information processing systems*, 996–1002.

Klopp, Olga. 2015. Matrix completion by singular value thresholding: sharp bounds. *Electronic journal of statistics* 9(2):2348–2369.

Kolda, Tamara G. 2001. Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications* 23(1):243–255.

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

Koltchinskii, Vladimir, Karim Lounici, and Alexandre B Tsybakov. 2011. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics* 39(5):2302–2329.

Koltchinskii, Vladimir, and Dong Xia. 2015. Optimal estimation of low rank density matrices. *Journal of Machine Learning Research* 16(53):1757–1792.

Laurent, Beatrice, and Pascal Massart. 2000. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 1302–1338.

Lesieur, Thibault, Léo Miolane, Marc Lelarge, Florent Krzakala, and Lenka Zdeborová. 2017. Statistical and computational phase transitions in spiked tensor estimation. In *2017 ieee international symposium on information theory (isit)*, 511–515. IEEE.

Leurgans, Sue E, Robert T Ross, and Rebecca B Abel. 1993. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications* 14(4):1064–1083.

Li, Lingjie, Wenjian Yu, and Kim Batselier. 2019a. Faster tensor train decomposition for sparse data. *arXiv preprint arXiv:1908.02721*.

Li, Nan, and Baoxin Li. 2010. Tensor completion for on-board compression of hyperspectral images. In *2010 ieee international conference on image processing*, 517–520. IEEE.

Li, Stan Z. 2009. *Markov random field modeling in image analysis*. Springer Science & Business Media.

Li, Xiaodong, Shuyang Ling, Thomas Strohmer, and Ke Wei. 2019b. Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Applied and computational harmonic analysis* 47(3):893–934.

Li, Xiaoshan, Da Xu, Hua Zhou, and Lexin Li. 2018. Tucker tensor regression and neuroimaging analysis. *Statistics in Biosciences* 10(3):520–545.

Li, Yanming, Bin Nan, and Ji Zhu. 2015. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics* 71(2):354–363.

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

Liu, Yipeng, Longxi Chen, and Ce Zhu. 2018. Improved robust tensor principal component analysis via low-rank core matrix. *IEEE Journal of Selected Topics in Signal Processing* 12(6):1378–1389.

Liu, Yu, Fahui Wang, Yu Xiao, and Song Gao. 2012. Urban land uses and traffic âĽ˜source-sink areasâĽ™: Evidence from gps-enabled taxi data in shanghai. *Landscape and Urban Planning* 106(1):73–87.

Lounici, K, M Pontil, AB Tsybakov, and SA Van De Geer. 2009. Taking advantage of sparsity in multi-task learning. In *Colt 2009-the 22nd conference on learning theory*.

Lounici, Karim, Massimiliano Pontil, Sara Van De Geer, and Alexandre B Tsybakov. 2011. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics* 39(4):2164–2204.

Lozano, Aurelie C, and Grzegorz Swirszcz. 2012. Multi-level lasso for sparse multi-task regression. In *Proceedings of the 29th international conference on international conference on machine learning*, 595–602. Omnipress.

Lu, Canyi, Jiashi Feng, Yudong Chen, Wei Liu, Zhouchen Lin, and Shuicheng Yan. 2016. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 5249–5257.

———. 2019. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE transactions on pattern analysis and machine intelligence* 42(4): 925–938.

Lubich, Christian, Thorsten Rohwedder, Reinhold Schneider, and Bart Vandereycken. 2013. Dynamical approximation by hierarchical tucker and tensor-train tensors. *SIAM Journal on Matrix Analysis and Applications* 34(2):470–494.

Luo, Yuetian, Garvesh Raskutti, Ming Yuan, and Anru R Zhang. 2020. A sharp blockwise tensor perturbation bound for orthogonal iteration. *arXiv preprint arXiv:2008.02437*.

Luo, Yuetian, and Anru R Zhang. 2020a. Open problem: Average-case hardness of hypergraphic planted clique detection. *Conference of Learning Theory (COLT)* 125: 3852–3856.

———. 2020b. A schatten-q matrix perturbation theory via perturbation projection error bound. *arXiv preprint arXiv:2008.01312*.

———. 2020c. Tensor clustering with planted structures: Statistical optimality and computational limits. *arXiv preprint arXiv:2005.10743*.

Lynch, RE, John R Rice, and Donald H Thomas. 1964. Tensor product analysis of partial difference equations. *Bulletin of the American Mathematical Society* 70(3): 378–384.

Ma, Zongming. 2013. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics* 41(2):772–801.

Mitra, Ritwik, and Cun-Hui Zhang. 2016. The benefit of group sparsity in group inference with de-biased scaled group lasso. *Electronic Journal of Statistics* 10(2): 1829–1873.

Mondelli, Marco, and Andrea Montanari. 2019. On the connection between learning two-layer neural networks and tensor decomposition. In *The 22nd international conference on artificial intelligence and statistics*, 1051–1060.

Montanari, Andrea, and Nike Sun. 2018. Spectral algorithms for tensor completion. *Communications on Pure and Applied Mathematics* 71(11):2381–2425.

Nadler, Boaz. 2008. Finite sample approximation results for principal component analysis: A matrix perturbation approach. *The Annals of Statistics* 36(6):2791–2817.

Nasiri, Mahdi, M Rezghi, and B Minaei. 2014. Fuzzy dynamic tensor decomposition algorithm for recommender system. *UCT Journal of Research in Science, Engineering and Technology* 2(2):52–55.

Negahban, Sahand N, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. 2012. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* 27(4):538–557.

Neumann, John von. 1936. The uniqueness of Haar's measure. *Rec. Math. [Mat. Sbornik]* 1(5):721–734.

Novikov, Alexander, Pavel Izmailov, Valentin Khrulkov, Michael Figurnov, and Ivan V Oseledets. 2020. Tensor train decomposition on tensorflow (t3f). *Journal of Machine Learning Research* 21(30):1–7.

Novikov, Alexander, Dmitrii Podoprikhin, Anton Osokin, and Dmitry P Vetrov. 2015. Tensorizing neural networks. In *Advances in neural information processing systems*, 442–450.

Novikov, Alexander, Anton Rodomanov, Anton Osokin, and Dmitry Vetrov. 2014. Putting mrfs on a tensor train. In *International conference on machine learning*, 811–819.

Orús, Román. 2019. Tensor networks for complex quantum systems. *Nature Reviews Physics* 1(9):538–550.

Oseledets, IV. 2009. A new tensor decomposition. In *Doklady mathematics*, vol. 80, 495–496. Pleiades Publishing, Ltd.

Oseledets, IV, and EE Tyrtyshnikov. 2009a. Recursive decomposition of multidimensional tensors. In *Doklady mathematics*, vol. 80, 460–462. Springer.

Oseledets, Ivan, and Eugene Tyrtyshnikov. 2010. Tt-cross approximation for multidimensional arrays. *Linear Algebra and its Applications* 432(1):70–88.

Oseledets, Ivan V. 2011. Tensor-train decomposition. *SIAM Journal on Scientific Computing* 33(5):2295–2317.

Oseledets, Ivan V, and Eugene E Tyrtyshnikov. 2009b. Breaking the curse of dimensionality, or how to use svd in many dimensions. *SIAM Journal on Scientific Computing* 31(5):3744–3759.

Oymak, Samet, Amin Jalali, Maryam Fazel, Yonina C Eldar, and Babak Hassibi. 2015. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory* 61(5):2886–2908.

Oymak, Samet, Amin Jalali, Maryam Fazel, and Babak Hassibi. 2013. Noisy estimation of simultaneously structured models: Limitations of convex relaxation. In *Decision and control (cdc), 2013 ieee 52nd annual conference on*, 6019–6024. IEEE.

De la Pena, Victor, and Evarist Giné. 2012. *Decoupling: from dependence to independence*. Springer Science & Business Media.

Perry, Amelia, Alexander S Wein, and Afonso S Bandeira. 2020. Statistical limits of spiked tensor models. In *Annales de l'institut henri poincaré, probabilités et statistiques*, vol. 56, 230–264. Institut Henri Poincaré.

Poignard, Benjamin. 2018. Asymptotic theory of the adaptive sparse group lasso. *Annals of the Institute of Statistical Mathematics* 1–32.

Puterman, Martin L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Raftery, Adrian E. 1985. A model for high-order markov chains. *Journal of the Royal Statistical Society: Series B (Methodological)* 47(3):528–539.

Rajih, Myriam, Pierre Comon, and Richard A Harshman. 2008. Enhanced line search: A novel method to accelerate parafac. *SIAM journal on matrix analysis and applications* 30(3):1128–1147.

Rakhuba, Maxim, and Ivan Oseledets. 2016. Calculating vibrational spectra of molecules using tensor train decomposition. *The Journal of Chemical Physics* 145(12): 124101.

Rao, Nikhil, Christopher Cox, Rob Nowak, and Timothy T Rogers. 2013. Sparse overlapping sets lasso for multitask learning and its application to fmri analysis. In *Advances in neural information processing systems*, 2202–2210.

Rao, Nikhil, Robert Nowak, Christopher Cox, and Timothy Rogers. 2015. Classification with the sparse group lasso. *IEEE Transactions on Signal Processing* 64(2): 448–463.

Raskutti, Garvesh, Ming Yuan, and Han Chen. 2019. Convex regularization for high-dimensional multiresponse tensor regression. *The Annals of Statistics* 47(3): 1554–1584.

Rauhut, Holger, Reinhold Schneider, and Zeljka Stojanac. 2017. Low rank tensor recovery via iterative hard thresholding. *Linear Algebra and its Applications* 523: 220–262.

Richard, Emile, and Andrea Montanari. 2014. A statistical model for tensor pca. In *Advances in neural information processing systems*, 2897–2905.

Rinaldo, Alessandro. 2009. Properties and refinements of the fused lasso. *The Annals of Statistics* 37(5B):2922–2952.

Robeva, Elina. 2016. Orthogonal decomposition of symmetric tensors. *SIAM Journal on Matrix Analysis and Applications* 37(1):86–102.

Rudelson, Mark, and Roman Vershynin. 2013. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability* 18.

Sanders, Jaron, Alexandre Proutière, and Se-Young Yun. 2020. Clustering in block markov chains. *The Annals of Statistics* to appear.

Schollwöck, Ulrich. 2011. The density-matrix renormalization group in the age of matrix product states. *Annals of physics* 326(1):96–192.

Shah, Devavrat, and Christina Lee Yu. 2019. Iterative collaborative filtering for sparse noisy tensor estimation. In *2019 ieee international symposium on information theory (isit)*, 41–45. IEEE.

Shao, Jun. 2003. Mathematical statistics.

Sharan, Vatsal, and Gregory Valiant. 2017. Orthogonalized als: A theoretically principled tensor decomposition algorithm for practical use. In *International conference on machine learning*, 3095–3104.

Shechtman, Yoav, Amir Beck, and Yonina C Eldar. 2014. Gespar: Efficient phase retrieval of sparse signals. *IEEE transactions on signal processing* 62(4):928–938.

Silver, Matt, Peng Chen, Ruoying Li, Ching-Yu Cheng, Tien-Yin Wong, E-Shyong Tai, Yik-Ying Teo, and Giovanni Montana. 2013. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two asian cohorts. *PLoS genetics* 9(11):e1003939.

Simon, Noah, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2013. A sparse-group lasso. *Journal of Computational and Graphical Statistics* 22(2):231–245.

Singh, Satinder P, Tommi Jaakkola, and Michael I Jordan. 1995. Reinforcement learning with soft state aggregation. In *Advances in neural information processing systems*, 361–368.

Song, Zhao, David P Woodruff, and Peilin Zhong. 2017. Relative error tensor low rank approximation. *arXiv preprint arXiv:1704.08246*.

Sprechmann, Pablo, Ignacio Ramirez, Guillermo Sapiro, and Yonina Eldar. 2010. Collaborative hierarchical sparse modeling. In *2010 44th annual conference on information sciences and systems (ciss)*, 1–6. IEEE.

Steinlechner, Michael Maximilian. 2016. Riemannian optimization for solving high-dimensional problems with low-rank tensor structure. Tech. Rep., EPFL.

Stojnic, Mihailo, Weiyu Xu, and Babak Hassibi. 2008. Compressed sensing-probabilistic analysis of a null-space characterization. In *2008 ieee international conference on acoustics, speech and signal processing*, 3377–3380. IEEE.

Stoudenmire, Edwin, and David J Schwab. 2016. Supervised learning with tensor networks. In *Advances in neural information processing systems*, 4799–4807.

Sun, Tingni, and Cun-Hui Zhang. 2012. Scaled sparse linear regression. *Biometrika* 99(4):879–898.

Sun, Will Wei, and Lexin Li. 2017. Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research* 18(1):4908–4944.

———. 2019. Dynamic tensor clustering. *Journal of the American Statistical Association* 114(528):1894–1907.

Sun, Will Wei, Junwei Lu, Han Liu, and Guang Cheng. 2017. Provable sparse tensor decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 3(79):899–916.

Sutton, Richard S, Andrew G Barto, et al. 1998. *Introduction to reinforcement learning*, vol. 135. MIT press Cambridge.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–288.

Tibshirani, Robert, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(1):91–108.

Tomioka, Ryota, and Taiji Suzuki. 2013. Convex tensor decomposition via structured schatten norm regularization. In *Advances in neural information processing systems*, 1331–1339.

Tsay, Ruey S. 2005. *Analysis of financial time series*, vol. 543. John wiley & sons.

Tucker, Ledyard R. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika* 31(3):279–311.

Vannieuwenhoven, Nick, Raf Vandebril, and Karl Meerbergen. 2012. A new truncation strategy for the higher-order singular value decomposition. *SIAM Journal on Scientific Computing* 34(2):A1027–A1052.

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

Verzelen, Nicolas. 2012. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics* 6:38–90.

Vidyasagar, Mathukumalli. 2014. Machine learning methods in the computational biology of cancer. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 470(2167):20140081.

Wainwright, Martin J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press.

Wainwright, Martin J, and Michael I Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1(1–2):1–305.

Wang, Miaoyan, and Lexin Li. 2018. Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *arXiv preprint arXiv:1811.05076*.

Wang, Miaoyan, and Yuchen Zeng. 2019. Multiway clustering via tensor block models. In *Advances in neural information processing systems*, 715–725.

Wang, Weiguang, Yingbin Liang, and Eric Xing. 2013. Block regularized lasso for multivariate multi-response linear regression. In *Artificial intelligence and statistics*, 608–617.

Wei, Zhi, and Hongzhe Li. 2007. A markov random field model for network-based analysis of genomic data. *Bioinformatics* 23(12):1537–1544.

Weil, André. 1940. L'intégration dans les groupes topologiques et ses applications. *Hermann et Cie.*

Wein, Alexander S, Ahmed El Alaoui, and Cristopher Moore. 2019. The kikuchi hierarchy and tensor pca. In *2019 ieee 60th annual symposium on foundations of computer science (focs)*, 1446–1468. IEEE.

Wilmoth, J. R., and V. Shkolnikov. 2006. Human mortality database, available at: http://www.mortality.org.

Wozniak, Jeffrey R, Linda Krach, Erin Ward, Bryon A Mueller, Ryan Muetzel, Sarah Schnoebelen, Andrew Kiragu, and Kelvin O Lim. 2007. Neurocognitive and neuroimaging correlates of pediatric traumatic brain injury: a diffusion tensor imaging (dti) study. *Archives of Clinical Neuropsychology* 22(5):555–568.

Wu, Tao, Austin R Benson, and David F Gleich. 2016. General tensor spectral co-clustering for higher-order data. In *Advances in neural information processing systems*, 2559–2567.

Xia, Dong. 2019a. Confidence region of singular subspaces for low-rank matrix regression. *IEEE Transactions on Information Theory* 65(11):7437–7459.

———. 2019b. Normal approximation and confidence region of singular subspaces. *arXiv preprint arXiv:1901.00304*.

Xia, Dong, and Ming Yuan. 2019. On polynomial time methods for exact low-rank tensor completion. *Foundations of Computational Mathematics* 19(6):1265–1313.

———. 2020. Statistical inferences of linear forms for noisy matrix completion. *Journal of the Royal Statistical Society Series B* to appear.

Xia, Dong, Anru R Zhang, and Yuchen Zhou. 2020. Inference for low-rank tensors–no need to debias. *arXiv preprint arXiv:2012.14844*.

Yuan, Ming, and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(1):49–67.

Yuan, Ming, and Cun-Hui Zhang. 2016. On tensor completion via nuclear norm minimization. *Foundations of Computational Mathematics* 16(4):1031–1068.

Zhang, Anru. 2019. Cross: Efficient low-rank tensor completion. *The Annals of Statistics* 47(2):936–964.

Zhang, Anru, and Rungang Han. 2018. Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association*.

———. 2019. Optimal sparse singular value decomposition for high-dimensional high-order data. *Journal of the American Statistical Association* 1–34.

Zhang, Anru, and Mengdi Wang. 2020. Spectral state compression of markov processes. *IEEE Transactions on Information Theory* 66(5):3202–3231.

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

Zhang, Anru R, Yuetian Luo, Garvesh Raskutti, and Ming Yuan. 2020a. Islet: Fast and optimal low-rank tensor regression via importance sketching. *SIAM Journal on Mathematics of Data Science* 2(2):444–479.

Zhang, Chenyu, Rungang Han, Anru R Zhang, and Paul M Voyles. 2020b. Denoising atomic resolution 4d scanning transmission electron microscopy data with tensor singular value decomposition. *Ultramicroscopy* 219:113123.

Zhang, Cun-Hui, and Jian Huang. 2008. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* 36(4):1567–1594.

Zhang, Cun-Hui, and Stephanie S Zhang. 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1):217–242.

Zhang, H, LZ Cheng, and W Zhu. 2011. A lower bound guaranteeing exact matrix completion via singular value thresholding algorithm. *Applied and Computational Harmonic Analysis* 31(3):454–459.

Zhang, Tong, and Gene H Golub. 2001. Rank-one approximation to high order tensors. *SIAM Journal on Matrix Analysis and Applications* 23(2):534–550.

Zhang, Yongyue, Michael Brady, and Stephen Smith. 2001. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging* 20(1):45–57.

Zhao, Jing, and Shiliang Sun. 2016. High-order gaussian process dynamical models for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems* 17(7):2014–2019.

Zheng, Qinqing, and Ryota Tomioka. 2015. Interpolating convex and non-convex tensor decompositions via the subspace norm. In *Advances in neural information processing systems*, 3106–3113.

Zhong, Kai, Zhao Song, and Inderjit S Dhillon. 2017. Learning non-overlapping convolutional neural networks with multiple kernels. *arXiv preprint arXiv:1711.03440*.

Zhou, Hao Henry, Yilin Zhang, Vamsi K Ithapu, Sterling C Johnson, and Vikas Singh. 2017. When can multi-site datasets be pooled for regression? hypothesis tests, ℓ_2-consistency and neuroscience applications. In *Proceedings of the 34th international conference on machine learning-volume 70*, 4170–4179. JMLR. org.

Zhou, Hua, Lexin Li, and Hongtu Zhu. 2013. Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association* 108(502): 540–552.

Zhou, Pan, and Jiashi Feng. 2017. Outlier-robust tensor pca. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 2263–2271.

Zhou, Yuchen, Anru R Zhang, Lili Zheng, and Yazhen Wang. 2020. Optimal high-order tensor svd via tensor-train orthogonal iteration. *arXiv preprint arXiv:2010.02482*.

Zhu, Ziwei, Xudong Li, Mengdi Wang, and Anru Zhang. 2019. Learning markov models via low-rank optimization. *arXiv preprint arXiv:1907.00113*.